Bayesian Inference in Large-scale Problems

by

James E. Johndrow

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Sayan Mukherjee

_____
Robert Wolpert

_____
Jonathan Mattingly

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2016

Abstract

Bayesian Inference in Large-scale Problems

by

James E. Johndrow

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Sayan Mukherjee

_____
Robert Wolpert

_____
Jonathan Mattingly

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2016

Abstract

Many modern applications fall into the category of "large-scale" statistical problems, in which both the number of observations n and the number of features or parameters p may be large. Many existing methods focus on point estimation, despite the continued relevance of uncertainty quantification in the sciences, where the number of parameters to estimate often exceeds the sample size, despite huge increases in the value of n typically seen in many fields. Thus, the tendency in some areas of industry to dispense with traditional statistical analysis on the basis that "n=all" is of little relevance outside of certain narrow applications. The main result of the Big Data revolution in most fields has instead been to make computation much harder without reducing the importance of uncertainty quantification. Bayesian methods excel at uncertainty quantification, but often scale poorly relative to alternatives. This conflict between the statistical advantages of Bayesian procedures and their substantial computational disadvantages is perhaps the greatest challenge facing modern Bayesian statistics, and is the primary motivation for the work presented here.

Two general strategies for scaling Bayesian inference are considered. The first is the development of methods that lend themselves to faster computation, and the second is design and characterization of computational algorithms that scale better in n or p. In the first instance, the focus is on joint inference outside of the standard problem of multivariate continuous data that has been a major focus of previous theoretical work in this area. In the second area, we pursue strategies for improving the

speed of Markov chain Monte Carlo algorithms, and characterizing their performance in large-scale settings. Throughout, the focus is on rigorous theoretical evaluation combined with empirical demonstrations of performance and concordance with the theory.

One topic we consider is modeling the joint distribution of multivariate categorical data, often summarized in a contingency table. Contingency table analysis routinely relies on log-linear models, with latent structure analysis providing a common alternative. Latent structure models lead to a reduced rank tensor factorization of the probability mass function for multivariate categorical data, while log-linear models achieve dimensionality reduction through sparsity. Little is known about the relationship between these notions of dimensionality reduction in the two paradigms. In Chapter 2, we derive several results relating the support of a log-linear model to nonnegative ranks of the associated probability tensor. Motivated by these findings, we propose a new collapsed Tucker class of tensor decompositions, which bridge existing PARAFAC and Tucker decompositions, providing a more flexible framework for parsimoniously characterizing multivariate categorical data. Taking a Bayesian approach to inference, we illustrate empirical advantages of the new decompositions.

Latent class models for the joint distribution of multivariate categorical, such as the PARAFAC decomposition, data play an important role in the analysis of population structure. In this context, the number of latent classes is interpreted as the number of genetically distinct subpopulations of an organism, an important factor in the analysis of evolutionary processes and conservation status. Existing methods focus on point estimates of the number of subpopulations, and lack robust uncertainty quantification. Moreover, whether the number of latent classes in these models is even an identified parameter is an open question. In Chapter 3, we show that when the model is properly specified, the correct number of subpopulations can be recovered almost surely. We then propose an alternative method for estimating

the number of latent subpopulations that provides good quantification of uncertainty, and provide a simple procedure for verifying that the proposed method is consistent for the number of subpopulations. The performance of the model in estimating the number of subpopulations and other common population structure inference problems is assessed in simulations and a real data application.

In contingency table analysis, sparse data is frequently encountered for even modest numbers of variables, resulting in non-existence of maximum likelihood estimates. A common solution is to obtain regularized estimates of the parameters of a log-linear model. Bayesian methods provide a coherent approach to regularization, but are often computationally intensive. Conjugate priors ease computational demands, but the conjugate Diaconis–Ylvisaker priors for the parameters of log-linear models do not give rise to closed form credible regions, complicating posterior inference. In Chapter 4 we derive the optimal Gaussian approximation to the posterior for log-linear models with Diaconis–Ylvisaker priors, and provide convergence rate and finite-sample bounds for the Kullback-Leibler divergence between the exact posterior and the optimal Gaussian approximation. We demonstrate empirically in simulations and a real data application that the approximation is highly accurate, even in relatively small samples. The proposed approximation provides a computationally scalable and principled approach to regularized estimation and approximate Bayesian inference for log-linear models.

Another challenging and somewhat non-standard joint modeling problem is inference on tail dependence in stochastic processes. In applications where extreme dependence is of interest, data are almost always time-indexed. Existing methods for inference and modeling in this setting often cluster extreme events or choose window sizes with the goal of preserving temporal information. In Chapter 5, we propose an alternative paradigm for inference on tail dependence in stochastic processes with arbitrary temporal dependence structure in the extremes, based on the

idea that the information on strength of tail dependence and the temporal structure in this dependence are both encoded in waiting times between exceedances of high thresholds. We construct a class of time-indexed stochastic processes with tail dependence obtained by endowing the support points in de Haan's spectral representation of max-stable processes with velocities and lifetimes. We extend Smith's model to these max-stable velocity processes and obtain the distribution of waiting times between extreme events at multiple locations. Motivated by this result, a new definition of tail dependence is proposed that is a function of the distribution of waiting times between threshold exceedances, and an inferential framework is constructed for estimating the strength of extremal dependence and quantifying uncertainty in this paradigm. The method is applied to climatological, financial, and electrophysiology data.

The remainder of this thesis focuses on posterior computation by Markov chain Monte Carlo. The Markov Chain Monte Carlo method is the dominant paradigm for posterior computation in Bayesian analysis. It has long been common to control computation time by making approximations to the Markov transition kernel. Comparatively little attention has been paid to convergence and estimation error in these approximating Markov Chains. In chapter 6, we propose a framework for assessing when to use approximations in MCMC algorithms, and how much error in the transition kernel should be tolerated to obtain optimal estimation performance with respect to a specified loss function and computational budget. The results require only ergodicity of the exact kernel and control of the kernel approximation accuracy. The theoretical framework is applied to approximations based on random subsets of data, low-rank approximations of Gaussian processes, and a novel approximating Markov chain for discrete mixture models.

Data augmentation Gibbs samplers are arguably the most popular class of algorithm for approximately sampling from the posterior distribution for the parameters

of generalized linear models. The truncated Normal and Pólya-Gamma data augmentation samplers are standard examples for probit and logit links, respectively. Motivated by an important problem in quantitative advertising, in Chapter 7 we consider the application of these algorithms to modeling rare events. We show that when the sample size is large but the observed number of successes is small, these data augmentation samplers mix very slowly, with a spectral gap that converges to zero at a rate at least proportional to the reciprocal of the square root of the sample size up to a log factor. In simulation studies, moderate sample sizes result in high autocorrelations and small effective sample sizes. Similar empirical results are observed for related data augmentation samplers for multinomial logit and probit models. When applied to a real quantitative advertising dataset, the data augmentation samplers mix very poorly. Conversely, Hamiltonian Monte Carlo and a type of independence chain Metropolis algorithm show good mixing on the same dataset.

For my bear

Contents

List of Tables

List of Figures

Acknowledgements

I offer my sincerest thanks to my advisor, David Dunson, whose guidance and support were critical to the body of work represented herein. He is among the most dynamic, creative, and independent thinkers that I have known, and his advice and his example have been instrumental in my development as a scholar and researcher. Perhaps most importantly, he offered me complete freedom to choose research topics and collaborate with others but always provided feedback and comments regardless of the topic or direction I chose. Despite serving as a mentor to numerous students and postdocs and running a relatively large research group, he is extremely accessible and generous with his time, and always sets meetings and provides feedback quickly. The importance of this cannot be understated, nor can its rarity. David is also exceptionally supportive of me within the statistics community. Finally, we share great appreciation and enthusiasm for outdoor sports, and David has been one of my regular trail running buddies throughout graduate school. Several of the chapters in this dissertation benefitted substantially from conversations between me and David that took place during runs.

I am lucky to have several mentors and collaborators in addition to David. During the second half of my first year – which, through an interesting series of events took place one year after the first half, see below – I took Robert Wolpert's "stochastic processes" course. The course was mostly about diffusions and Lévy processes. Toward the end, he spent a few weeks on multivariate extremes and tail dependence,

a topic that had interested me as early as 2007, when the financial crisis began to set in. Those of us working in the large and relatively dispersed community of financial analysts had a sense then that risk had been underestimated, and this was among the catalysts for my pursuit of a Ph.D. in statistics. Two topics considered in that course – waiting times between threshold exceedances in Brownian motion and the limiting distribution of maxima of stochastic processes – gave rise in my mind to the idea of studying tail dependence via waiting times between peaks over thresholds. After some initial failed attempts to develop a general class of stochastic processes in which such "tail waiting times" were tractable, Robert hit upon the idea of endowing the support points in de Haan's spectral characterization of the max-stable process with velocities and lifetimes, and the work presented in Chapter 5 took shape.

Around the time that Robert and I began working on multivariate extremes, I had some initial conversations with Aniraban Bhattacharya on the general topic of latent class models for multivariate categorical data. We were both intersted in the induced prior on the parameters of the log-linear model. At the time there was empirical evidence that stick-breaking priors on the weights in these mixture models induced shrinkage of the parameters of the log-linear model, particularly for higher-order interactions, but these properties were not understood theoretically. We eventually became curious about the more fundamental issue of whether sparse log-linear models had low Tucker or PARAFAC tensor rank. Initially this seemed doubtful, given the absence of any relationship between sparse graphical models and low-rank covariance matrices in the continuous setting, but during JSM in the summer of 2012, we discovered that in the special case of probability matrices there was in fact a relationship between the PARAFAC rank and sparsity of a log-linear model. This eventually grew into the work in Chapter 2, and our continued collaboration led us to the work in Chapter 4 as well as some ongoing projects not represented in this dissertation.

By the end of 2013, I had developed somewhat of a niche in the area of high-dimensional contingency tables through my work with David and Anirban in that area. During the preceding year, I had collaborated with Andrew Cron at MaxPoint on the analysis of count data on transitions between pairs of websites. David and I had developed an approach for scaling up Gibbs sampling in latent class models for multivariate categorical data, but I was interested in the more general topic of the behavior of MCMC algorithms that use approximations to the exact transition kernel. I related this during a casual chat with Sayan Mukherjee at JSM during the summer of 2013, and he asked if I would "like some help on that." This was the genesis of another major research direction in my Ph.D. work. Within a couple of months, Jonathan Mattingly had also become involved. The project grew into something much more than I had initially envisioned, and I ended up learning the theory of Markov chains at a more fundamental level than I had anticipated, thanks to Jonathan's and Sayan's mentorship. It was a transformative experience. A portion of this work is represented in Chapter 6, and other threads are ongoing.

While working on the MaxPoint project, David and I noticed that when the MLEs are close to zero or one, data-augmentation Gibbs sampling algorithms for categorical and binary outcomes mixed very poorly. We temporarily shelved this topic, but at a conference in 2014, David mentioned it to Natesh Pillai, and we began collaborating on a project to understand theoretically the basis of this behavior. Natesh involed Aaron Smith, who, like Jonathan, is an outstanding probabilist with a deep understanding of Markov chain theory. The result of this work is partially represented by Chapter 7.

My time as a Ph.D. student was a transitional period in my life and in many ways represented the culmination of a path that I started down at age 18. I began undergrad as a math major, then later switched to Chemistry and was planning on medical school. I worked for several years in biology laboratories but was not

sufficiently interested in doing experiments to continue. I had, I think, a vague sense that too much p-value fishing was occurring in experimental biology, though I would not have articulated it in that way. I then spent four years at NERA, an economic consulting firm, basically functioning as a data scientist and, later, a manager of data scientists, though at the time no one used that term. I benefitted from some great colleagues there, especially David Harrison and Bernie Reddy. My experience at NERA did not lead me immediately to statistics. To that I owe a debt to Alex Lenkoski, a college friend who preceded me at NERA and had gone on to a Ph.D. in statistics at the University of Washington. He was the first to suggest the idea of graduate school in statistics to me, and provided my first exposure to statistics research. He also nudged me toward Duke, where I ended up in the fall of 2010.

Almost as soon as I arrived, I left Duke for a year. I met my wife, Kristian Lum, when she was finishing her Ph.D. at Duke and I was starting mine. She had already taken a postdoc in Brazil and I followed her, taking a leave of absence from graduate school for the entire calendar year of 2011. This could have ended my pursuit of a Ph.D. before it really started, were it not for Mike West, who supported my leave request with the dean, and, I think, a majority of the faculty, despite what I am certain was a fair bit of skepticism regarding the likelihood of my return and the potential that I would accomplish anything even if I did. In many ways, the leave was beneficial. I arrived in graduate school with enough mathematics background and aptitude to succeed, but quickly realized that I knew almost no statistics or probability. Toward the end of 2011, Kristian put me on a two-month crash course program, and when I returned in the winter of 2012 I was much more successful. I am also grateful to David for taking me on as a student again. If he had lost confidence in me after my abrupt departure from graduate school, no one would have blamed him.

After first year I did not yet have all of the tools necessary to do the work you see

here. In part, my growth in the second year was a result of my work with David and Anirban. Another critical piece was "multivariate statistics" with Robert Wolpert during the spring of my second year. When add/drop ended there were only three of us enrolled, and no one was especially eager to do multivariate normal theory for the fifth time. We prevailed upon Robert to change the format of the course and create the rigorous statistical inference course that at the time seemed not to exist in the department. This was critical in development of my research program and in gaining a larger context in which to view my work that transcends the Bayesian/non-Bayesian divide.

I have benefitted from a number of great colleagues and collaborators in addition to those mentioned already. David Banks has been very supportive in inviting me for conferences and arranging consulting work. Kristian and I have developed a productive collaboration on methods for population estimation, particularly related to their use in human rights applications. In this area I also have had the pleasure of working with Daniel Manrique-Vallier at the University of Indiana and Patrick Ball at Human Rights Data Analysis Group (HRDAG). My collaboration with Irene Liu on population genetics and songbird behavior has lead to a number of interesting research directions, one of which is represented in Chapter 3. I have also benefitted from numerous conversations with some very talented students and postdocs that have passed through the department during my time here, including Xiangyu (Sam) Wang, Yun Yang, Joshua Vogelstein, Sanvesh Srivastava, Stanislav Minsker, Debdeep Pati, Zoey Zhao, and numerous others.

Earning a Ph.D. is hard on many people. It requires independence and self-confidence, and it is arguably the case that these traits are tested more than intelligence and creativity. One's support network is critical to persevering through the ups and downs of graduate school, and academia in general. I owe as much to my family and close friends as I do to my professional network. Above all, I thank

Kristian for her support and encouragement – both emotional and practical – and her willingness to compromise in various ways to enable me to thrive as a graduate student. I thank my parents, Judy Mortellaro and David Johndrow, for endowing me with the intellectual curiosity that above all drives my research. A few close friendships have also been critical. Galen Reeves and Mary Knox have been good friends since their arrival in Durham over two years ago, and we have enjoyed many climbing and running adventures. Alex Lenkoski and Disa Thorarinsdottir have remained close friends from across the pond, and have been welcoming hosts on several European vacations, most of which involved rock climbing and bouldering. During the two years that Kristian worked at Virginia Tech and I commuted between Duke and Blacksburg, we had a wonderful friendship network there, including Jennifer Chang, Caitlin Rivers, Isaac Yeaton, Emily Hairfield, and Adam Walker. We have been lucky to have a similarly exemplary group of friends in Durham, including Jake Stauch, Tatiana Birgisson, and Ying Shi. In addition to being an advisor and mentor, David and his wife Amy Herring have been good friends to us. I have also benefitted from the diverse group of graduate, professional, and undergraduate scholars in the University Scholars Program and from a great relationship with Tori Lodewick, the program's director.

Finally, I would like to acknowledge those who funded my Ph.D.: the University Scholars program, the J.B. Duke graduate fellowships, the National Institutes of Health, and MaxPoint interactive. I also received travel grants from ISBA and the graduate school.

1

Introduction

A major focus of modern statistical science is the development of theory and methods for "large-scale" problems. The term "large-scale" refers to settings in which both the number of parameters p and the number of observations n may be large. This focus is driven by the availability of huge quantities of data, a phenomenon that is transforming science, medicine, and industry. The full potential of these data is still limited by the lack of algorithms and methods that provide accurate *uncertainty quantification* for large-scale problems. Many existing methods, particularly in large n settings, obtain point estimates, despite the importance of uncertainty quantification in the sciences, where the large n, huge p paradigm is common. In such instances, the availability of large samples has not lifted the curse of dimensionality. Outside of limited "$n = $ all" industrial applications, the $p \gg n$ paradigm remains very relevant for statistical methodology; the main effect of Big Data is to make computation orders of magnitude harder without removing the need for robust uncertainty quantification. Fully Bayesian methods are well-suited to uncertainty quantification in complex scientific problems, but generally scale poorly relative to alternatives. Therefore, continued relevance of Bayesian inference for large-scale

problems will require alleviating this scaling problem.

Here, we focus on two general strategies for improving the scalability of Bayesian inference. The first is development of methods that lend themselves to faster computation, and on theoretical analysis of the properties of these methods relative to canonical, less scalable methods. Our focus in this area is joint inference, which provides many examples of large-scale problems due to the rapid growth of the dimension of the parameter space in the number of variables and the necessity of obtaining relatively large samples to achieve acceptable levels of uncertainty in parameter estimates. The second general strategy is development of more scalable algorithms for posterior computation. Here the focus is Markov chain Monte Carlo, which has long been the dominant computational strategy for fully Bayesian inference but is often unacceptably slow on large datasets. The remainder of this chapter provides some general background on these two areas, then describes the specific research questions and contributions found in later chapters.

1.1 Background and motivation

This section provides general background to motivate the specific research topics that follow.

1.1.1 Large scale joint inference

The scaling problem is particularly acute in joint modeling, where the number of parameters tends to grow very rapidly in the number of variables. Even simple joint models for p variables, such as the multivariate Gaussian model, have order p^2 parameters. It is well-known that when a sequence of problem sizes $p_n > n$, traditional statistical estimators fail to even be consistent. As a result, in the large p setting, it is common to impose *parsimony* in a variety of ways to make consistent estimation possible. In the continuous data setting, common alternatives for imposing parsi-

mony include sparse graphical models and factor models, which are examples of the more general sparsity and low-rank paradigms.

Much of the literature focuses on the case where the variables of interest $y_1, \ldots, y_p$ are continuous random variables. An arguably more challenging problem that has received somewhat less focus in the high-dimensional context is where $y_1, \ldots, y_p$ are categorical, with each y_j taking d_j possible values. In this case, the joint distribution of $y_1, \ldots, y_p$ can be described by a $\prod_j d_j$ probability tensor π, i.e. a nonnegative tensor with entries adding to one given by

$$\pi_{i_1,\ldots,i_p} = \Pr(y_1 = i_1, \ldots, y_p = i_p), \tag{1.1}$$

for any $\boldsymbol{i} = (i_1, \ldots, i_p) \in \times_{j=1}^{p} \{1, \ldots, d_j\}$, where $\times$ is the Cartesian product. It is common to summarize the data in a contingency table, another $\prod_j d_j$ nonnegative tensor with entries $n(\boldsymbol{i})$ the number of observations of $(y_1 = i_1, \ldots, y_p = i_p)$. It is worth noting that, if the variables $y_1, \ldots, y_p$ are latent variables in a hierarchical model, this structure can be used as a general nonparametric approach to modeling the joint distribution of mixed-scale variables. Thus, all of the discussion that follows applies to a more general class of methods for joint modeling, but to simplify the exposition we will assume that $y_1, \ldots, y_p$ are observed.

The low-rank and sparsity paradigms are both relevant approaches for imposing parsimony in this setting. The canonical approach for imposing sparsity is to express π as a log-linear model

$$\log\left(\frac{\pi_{\boldsymbol{i}}}{\pi_{\boldsymbol{0}}}\right) = \sum_{E \subset V} \theta_E(\boldsymbol{i}_E),$$

where $V = \{1, \ldots, p\}$ and $\theta_E(\boldsymbol{i}_E)$ is a parameter corresponding to the variables in E taking the values $\{i_j, j \in E\}$. A common Bayesian approach to impose sparsity in this setting is to place a prior on the space of all sparse log-linear models, and an independent, conjugate prior on the parameters given the model. This results in

a Bayesian model averaging posterior, which can be expressed as a set of posterior probabilities for each (sparse) model and a posterior distribution for parameters given the model. Since there are $D = \prod_j d_j - 1$ free parameters in θ, there are $2^D \gg 2^p$ possible sparsity patterns in θ. Even if the model space is restricted to graphical models, the dimension of the model space still grows exponentially in p. As a result, typical computational algorithms for Bayesian inference on sparse log-linear models scale very poorly in the number of variables. Even with sophisticated stochastic search methods, the largest scale demonstration of computation for Bayesian model averaging with log-linear models has $p = 16$, which certainly does not fit the modern definition of large p.

One means of imposing a low-rank structure on π is to express π in PARAFAC tensor expansion

$$\pi_{i_1,\ldots,i_p} = \sum_{h=1}^{K} \nu_h \prod_{j=1}^{p} \lambda_{hi_j}^{(j)}, \tag{1.2}$$

where $\lambda_h^{(j)}$ and ν are probability vectors, i.e. nonnegative vectors with entries adding to 1. Low-rank structure can be encouraged by choosing a prior that favors a relatively small number of mixture components K in (1.2). Another common approach is to fix K at some modest value, then use a stick-breaking or Dirichlet$(1/K, \ldots, 1/K)$ prior on ν; the latter has been shown to asymptotically recover the correct number of mixture components in other settings. Posterior computation under these priors can be performed using a straightforward Gibbs sampler that scales linearly in p. Several demonstrations of related models in the literature have performed inference for p into the thousands. Thus, the fully Bayesian sparse paradigm for log-linear models is not scalable, whereas the low-rank paradigm scales well in p.

An important application for the model in (1.2) is inference on population structure from multilocus genotype data. While in many cases the objective of the model

in (1.2) is to obtain a parsimonious representation of the joint distribution of multivariate categorical data, in this setting a focus of inference is estimation of the rank, K. The dominant methods in this application area provide poor or no uncertainty quantification, despite the scientific nature of the application, where hypothesis testing is considered critical. A more fundamental issue is whether K is an identified parameter, and whether consistent estimators of K even exist. These questions relate to an interesting area of algebraic statistics that contains a number of open problems.

In cases where n is not too small relative to D, good performance in estimation of π and θ can sometimes be achieved through shrinkage, without utilizing priors that place nonzero mass on sparse θ. Shrinkage priors generally offer much faster computation than sparsity-inducing priors. In such settings, shrinkage often remains beneficial even when n is relatively large, since low probability configurations of $y_1, \ldots, y_p$ are often unobserved, leading to non-existence of classical maximum likelihood estimates. The Bayesian paradigm offers a coherent way to induce shrinkage through the prior. Conjugate priors offer a particularly efficient alternative, since the posterior is available in closed form so that the only computation necessary in the exponential family is the calculation of sufficient statistics. When the number of observations of $y_1, \ldots, y_p$ is considered fixed at N, the result is the Multinomial(N, π) likelihood for the counts $n(\boldsymbol{i})$. Since the multinomial is in the exponential family, there exists a conjugate prior of the Diaconis-Ylvisaker class given by

$$dq(\theta; N_0, n_0) = e^{N_0 n_0^T \theta - N_0 M(\theta)}, \qquad N_0 \in \mathbb{R}, n_0 \in \mathbb{R}^D. \tag{1.3}$$

On observing data n with entries summing to N and sufficient statistics $\bar{n}$, the posterior is then also Diaconis–Ylvisaker, with parameters $N_0 + N, n_0 + \bar{n}$, i.e. $dq(\theta \mid x) = dp(\theta; N_0 + N, n_0 + \bar{n})$. Since the data are often provided as the counts $n(\boldsymbol{i})$ for $\{\boldsymbol{i} : n(\boldsymbol{i}) > 0\}$, this approach to inducing shrinkage is essentially computation-free. The main drawback is that q is a non-standard distribution, which complicates

posterior inference, computation of credible intervals, and hypothesis testing. As a consequence, it is more common to rely on Gaussian priors, with computation by data augmentation Gibbs sampling. Although computation is much more intensive, the resulting MCMC sample paths can be used directly for approximate posterior inference.

Another joint inference problem in which the large n, large p setting is commonly encountered is that of modeling dependence in the extremes of stochastic processes. Tail dependence is a topic of growing interest in many areas. Following the financial crisis of 2008-2009, many blamed the widespread use of the Gaussian copula, which lacks tail dependence, to model risk in credit derivative portfolios as a contributing factor. Growing concerns about the long-term effects of climate change have induced a fundamental re-examination of how we quantify dependence in catastrophic events. Ultimately, existential risk stems from many "unlikely" events occurring simultaneously, and therefore it is critical to understand whether "worst case" events are likely to co-occur.

The standard model of extremes of stochastic processes is the max-stable process (De Haan (1984), Beirlant et al. (2006), Coles et al. (2001)). A process $Y(x), x \in \mathcal{X}$ for an index set $\mathcal{X}$ is max-stable if there exist sequences $a_n(x), b_n(x)$ and a process $w(x), x \in \mathcal{X}$ such that for every finite collection of points x, we have

$$Y(x) = \lim_{n\to\infty} \frac{[\bigvee_{i=1}^{n} w_i(x)] - a_n(x)}{b_n(x)}, \tag{1.4}$$

where $\{w_i\}_{i\leq n}$ are independent copies of $w(x)$ (De Haan and Ferreira (2007), Beirlant et al. (2006), Schlather (2002)). In the spatial or spatiotemporal setting, one usually takes $\mathcal{X} = \mathbb{R}^d$ for some integer d. This model is quite general in the sense that *if* there exist $\mathcal{X}$-indexed sequences $a_n(x)$ and $b_n(x)$ such that normalized maxima of the form (5.1) converge, then the limit must be a max-stable process De Haan and Ferreira (2007). Fitting max-stable processes to data is inherently a high-dimensional

problem because there is no finite-dimensional parametric model of a max-stable process. This contrasts with other stochastic processes commonly encountered in statistics, such as the Gaussian process, where the entire process can be modeled in terms of a positive-definite kernel function with only a few parameters.

One way to construct max-stable processes is to utilize de Haan's spectral characterization. Define a stochastic process Z by

$$Z(x) = \sup_j u_j k(x, \xi_j) \tag{1.5}$$

where $k : \mathcal{X} \times X \to \mathbb{R}$ is a nonnegative kernel, $p(u) \propto u^{-2}$, and ξ are points of a homogeneous Poisson process. This process has a max-stable distribution (Schlather (2002)). A common choice for k is the isotropic Gaussian kernel $k(x, x') = \frac{1}{\sqrt{2\pi}} \exp(-(x - x')^2/2)$, which was initially proposed by Smith. This choice leads to tractable joint distributions for the process at pairs of locations. Despite this, fitting the process to data is challenging. To begin with, data are usually transformed prior to model fitting by taking maxima over windows or discarding data that do not exceed some high threshold, since only the extreme observations are well-approximated by a max-stable process, and data near the center of the distribution would tend to swamp the extreme observations in any likelihood-based (e.g. Bayesian) modeling approach. Thus, one generally needs to start with large n to have sufficient data left after taking maxima or thresholding. Moreover, even with the Gaussian choice for k, computation remains a major challenge, because of the infinite number of unknown support points in the Poisson process and the intractability of the likelihood. As a result, fully Bayesian inference on tail dependence is practically infeasible, and the literature has focused on approximations and optimization-based computation.

1.1.2 Bayesian computation and MCMC

Thus far we have focused on scalability from a methodological perspective, using the challenging context of joint inference as a canonical example. From this perspective, some methods are more amenable to computation than others, and thus scaling Bayesian inference to large-scale problems is associated with proposing methods that lend themselves to faster computation. When utilizing this strategy, quantifying theoretically the tradeoffs in choosing a method because of the availability of a scalable algorithm rather than purely based on its statistical properties is of fundamental importance. Another perspective from which to view this problem is to propose new computational algorithms for Bayesian computation that make more methods tractable for large-scale problems. This approach has the advantage of avoiding the tradeoff resulting from choosing a method that is in some sense sub-optimal for the question of inferential interest simply for its computational tractability.

Arguably the dominant method for posterior computation is Markov chain Monte Carlo (MCMC). While a number of alternatives, such as variational approximations, are generally more scalable, they tend to be inferior to MCMC in uncertainty quantification. Since robust quantification of uncertainty is an important motivation for using Bayesian methods, MCMC has remained very popular among practitioners. MCMC proceeds by constructing a transition kernel $\mathcal{P}$ of an ergodic Markov chain $\theta_1, \ldots, \theta_T$ with invariant measure the posterior $\Pi(\theta \mid y)$, then collects sample paths from the chain. Expectations of functionals are then approximated from their ergodic averages $\Pi(\theta \mid y)(f) \approx \frac{1}{T}\sum_{t=0}^{T} f(\theta_t)$. Other posterior quantities of interest can also be approximated based on samples from the chain, such as approximating quantiles of the posterior by sample quantiles of the chain, or the full posterior by the empirical measure of the sample path.

The computational complexity of MCMC can be understood in terms of three

quantities: (1) the computational complexity of taking a single step from the transition kernel $\mathcal{P}$; (2) the number of steps necessary for the measure of the next step of the chain to be "close" to the invariant measure; and (3) the number of steps required at or near the stationary measure to achieve acceptable simulation error in approximating posterior quantities of interest. The first can be studied via computational complexity analysis common in computer science, while the second and the third are associated with the *convergence rate* and *autocorrelation function* of the Markov chain. The overall computational complexity of the chain scales like the product of (1) and the least efficient of (2) and (3). For example, if the computational complexity of a single step is $\mathcal{O}(np)$, $\mathcal{O}\big(n\sqrt{p}\big)$ steps are required to get "close" to the invariant measure, and $\mathcal{O}(n^2p)$ samples are required at stationarity to achieve the desired simulation error, then the overall computational complexity of the MCMC algorithm is $\mathcal{O}(n^3p^2)$.

One approach to improving the computational scalability of MCMC is to reduce the computational complexity of taking a single step from $\mathcal{P}$ by making approximations. In particular, suppose that $\mathcal{P}_\epsilon$ is a transition kernel that provides a uniform approximation to $\mathcal{P}$, i.e. that satisfies

$$\sup_{\theta,\theta' \in \Theta} ||\mathcal{P}(\theta;\cdot) - \mathcal{P}(\theta';\cdot)||_{\text{TV}} < \epsilon, \tag{1.6}$$

where $||P-Q||_{\text{TV}}$ is the total variation distance between probability measures P and Q given by

$$||P - Q||_{\text{TV}} = \sup_{A \subset \Theta} |P(A) - Q(A)|,$$

and Θ is the parameter space, which is also the state space of the Markov chain. Then, if $\mathcal{P}_\epsilon$ has lower computational complexity per step than $\mathcal{P}$, a more scalable MCMC algorithm can potentially be constructed by substituting $\mathcal{P}_\epsilon$ for $\mathcal{P}$. This practice – largely unacknowledged – has long been common in applied Bayesian

statistics, but has very little theoretical support. Critically, $\mathcal{P}_\epsilon$ will not have invariant measure Π, and thus any quantities computed based on sample paths from $\mathcal{P}_\epsilon$ will be biased, possibly negating the computational advantage of using $\mathcal{P}_\epsilon$ in the first place. On the other hand, $\mathcal{P}_\epsilon$ may have a different convergence rate and autocorrelation function than $\mathcal{P}$, which could either improve or erode the computational advantage of $\mathcal{P}_\epsilon$. It is therefore of substantial interest to develop a better theoretical framework for assessing the performance of such transition kernel approximations in Bayesian computation.

Understanding the computational complexity of MCMC is an important problem beyond the context of approximating transition kernels. Often, MCMC algorithms that perform well with moderate p and n can become unacceptably inefficient in large n or large p problems. One way to address this is to study the limiting behavior of MCMC algorithms as either n or p approach ∞. While the computational complexity of taking a single step from $\mathcal{P}$ is usually straightforward to compute, understanding the overall scaling behavior of the algorithm requires studying the scaling behavior of the convergence rate and the autocorrelation function, which is often mathematically challenging. These types of studies are quite common in the probability and computational physics literature, but relatively rare in the statistics literature. However, if MCMC is to remain the principal computational paradigm in Bayesian statistics, more such investigations will need to be done, particularly for the MCMC algorithms most commonly used by applied statisticians. This would likely result in the proposal of alternative MCMC algorithms when commonly used algorithms are found to exhibit poor limiting behavior. Of particular interest is what should be considered "scalable" in this context. The traditional notion in computer science is that exponential time algorithms are unacceptably slow, while polynomial time algorithms are often "good enough." However, some have suggested that in the large-scale setting, polynomial time is often unacceptable, making it necessary

to search for quasilinear time algorithms for most problems. Which of these notions is closer to the truth remains somewhat of an open question.

1.2 Research topics and principal contributions

1.2.1 Tensor decompositions and sparse log-linear models

The discussion in the previous section regarding contingency table analysis makes clear that, when the data generating process corresponds to a relatively low-rank probability tensor π, Bayesian inference for latent structure models is highly scalable. On the other hand, while sparse log-linear models can be very parsimonious, Bayesian computation for these models scales very poorly in the number of parameters p. While one option is to simply choose the latent structure model, sparse log-linear models are often highly appealing because of the interpretation of sparsity in θ as corresponding to conditional independence relationships between the variables $y_1, \ldots, y_p$. Therefore, an appealing possibility is to perform computation in the latent structure parametrization, then use the posterior for π to perform inference in the log-linear parametrization.

A basic requirement for this approach to be feasible is a relationship between the sparse and low-rank notions of parsimony for π. In particular, if the true joint probability distribution π corresponds to a sparse log-linear model, there must exist a relatively low-rank PARAFAC representation for π. This is the basic question that we address in Chapter 2, by attempting to upper bound the PARAFAC rank of π as a function of the sparsity of θ. We show that such a relationship does exist, and that certain classes of sparse log-linear models are associated with low-rank tensors, while other classes may have tensor ranks that grow much more rapidly than the number of nonzero entries of π. Motivated by these findings, we propose a completely new family of latent class models, corresponding to a new type of tensor factorization, and show that these models offer parsimonious representations of a larger family of sparse

log-linear models without sacrificing computational scalability. We show in a series of simulation studies that the parameters of a large class of sparse log-linear models can be recovered from the posterior of the new latent class models with minimal estimation loss, providing support for the feasibility of this strategy. In addition, the techniques used in showing the bounds are likely to be of substantial independent interest.

1.2.2 Estimation of tensor rank and its application in population structure analysis

When the variables $y_1, \ldots, y_p$ in (1.1) correspond to alleles of genes at p loci, the model in (1.2) is often employed for inference on the genetic structure of a population. Of particular interest is estimation of K, the PARAFAC tensor rank of π. The currently dominant method in biology offers no uncertainty quantification, and a number of papers in the statistics literature raise questions about whether it is possible to reliably estimate K. In Chapter 3, we address these issues by proposing an alternative prior on K that allows for computation by a split-merge Gibbs sampler and produces a full posterior distribution for K that can be employed to quantify uncertainty in this parameter. We also address the issue of whether K can be estimated from data by showing that under very genereal conditions, when the true π is generated by a random PARAFAC decomposition of rank K, then there exists a lower-rank representation of π with probability zero. In conjunction with existing results in algebraic statistics, this provides theoretical support for estimation of K. We also provide a numeric procedure for checking formal identifiability of K for a problem size (value of p and $d_j, j = 1, \ldots, p$) of interest.

1.2.3 Optimal Gaussian approximations to the posterior for log-linear models with conjugate priors

The posterior under the Diaconis-Ylvisaker prior in (4.7) provides essentially computation-free shrinkage estimation for log-linear models, but the resulting posterior is non-

standard, complicating inference and hypothesis testing. A more tractable approximation to the posterior under the Diaconis-Ylvisaker prior is therefore of substantial interest. In Chapter 4, we propose a Gaussian approximation to the Diaconis-Ylvisaker posterior, and show that it is the optimal Gaussian approximation in the Kullback-Leibler sense. Empirically, the proposed approximation provides a better approximation to the posterior than the more commonly used Laplace approximation. This observation is of broader interest, since the Laplace approximation is commonly used in exponential families. If the result that the Laplace approximation is not optimal extends beyond the case of the Multinomial likelihood considered here, then it suggests an alternative class of default Gaussian approximations for exponential families.

1.2.4 Tail waiting times and extremes of stochastic processes

Inference on dependence in the extremes of stochastic processes focuses on the case where some transformation of the data – e.g. maxima over windows or the data that exceed some threshold – are thought to be approximately realizations of the max-stable process defined in (1.4). In applications, it is almost always the case that one has a set of observations $w(\boldsymbol{x}, t) = (w(x_1, t), \ldots, w(x_n, t))$ on a stochastic process $\{W(x, t)\}$ at a collection of points $x_1, \ldots, x_n$ and times $t_1, \ldots, t_p$. These observations could represent hourly precipitation, maximum daily wind speed, or, if we treat the spatial index set $\mathcal{X}$ as a latent coordinate in an attribute space, essentially any multivariate time series, such as daily stock prices. Incorporating temporal dependence within existing methods for inference in this setting is often cumbersome, with many of the proposed approaches requiring multi-stage estimation procedures that sacrifice efficiency. In this context, data are almost always temporally dependent, so this is a substantial shortcoming of current approaches. Moreover, the max-stable process itself does not typically include a time-indexed dimension, so that

if time is introduced as an added dimension in $\mathcal{X}$, dependence across time would be governed by the same process as dependence across space.

The spectral characterization of the max-stable process provides motivation for how one might develop a class of heavy-tailed stochastic processes better suited to modeling temporal and spatially dependent extreme events. A useful physical heuristic for the expression in (1.5) takes the support points ξ_j to be the centers of storms, the kernels k their shapes, and the magnitudes u_j their intensities. In this rubric, the value of Z can be taken to be extreme weather events, caused by the most intense storm to occur at any location over some period of time. This heuristic suggests some shortcomings of the generic max-stable process for modeling weather events and other physical phenomena, since the "storms" are static in space and time, unlike real storms. This motivates the development of a novel class of heavy-tailed stochastic processes that we refer to as *max-stable velocity processes*, discussed in Chapter 5. This process explicitly includes a time-indexed domain with different dynamics than the spatial margins of the process. Informally, a max-stable velocity process arises when the "storms" in the spectral characterization of the max-stable process are endowed with finite lifetimes and nonzero velocities.

Direct inference on the parameters of a max-stable velocity process is computationally intensive. To circumvent this, while preserving temporal and dependence information, we utilize waiting times between threshold exceedances at pairs of points as data. We obtain several results showing that the waiting time distribution in a max-stable velocity process can be well approximated by finite mixtures of exponentials. We also propose a method to utilize the posterior distribution under a finite exponential mixture model to perform inference on the strength of extreme dependence, and hypothesis testing for tail independence. The method is computationally efficient and simple.

We then apply this method to analysis of four real datasets: (1) daily precipi-

tation data for U.S. cities; (2) exchange rates; (3) daily prices of the 30 stocks that comprise the Dow Jones Industrial Average; and (4) electrical potential at single neurons in the brain of a mouse exploring a maze. The method compares favorably with existing methods when extremes tend to occur simultaneously, while detecting tail dependence that occurs at time lags that other methods miss, and providing a quantitative measure of the strength of tail dependence that is meaningful in both cases. The use of waiting times between exceedances represents a new paradigm in statistical inference on tail dependence.

1.2.5 "Approximate" Markov chains in large-scale Bayesian inference.

Substituting a transition kernel $\mathcal{P}_\epsilon$ satisfying the uniform approximation error condition in (1.6) for a transition kernel $\mathcal{P}$ with invariant measure Π can potentially lead to improved computational speed and scaling in n and p. However, the extent of this advantage depends on several characteristics of both $\mathcal{P}$ and $\mathcal{P}_\epsilon$, as well as on the inferential goals and loss function chosen. In Chapter 6, we show a number of theoretical results relating the approximation error ϵ and the convergence rate of the original kernel $\mathcal{P}$ to estimation error for ergodic averages of sample paths from $\mathcal{P}$ and $\mathcal{P}_\epsilon$. We then introduce an optimality concept that provides a decision rule for whether to use $\mathcal{P}_\epsilon$ or $\mathcal{P}$ given a loss function measuring the cost of incorrectly estimating Πf and a computational budget τ. This concept is referred to as *compminimax*. The compminimax decision rule is computed by thresholding a relative efficiency function. This function depends on ϵ and τ as well as the *speedup* s_ϵ of the kernel $\mathcal{P}_\epsilon$, which is the number of steps that can be taken with $\mathcal{P}_\epsilon$ in the time required to take one step from $\mathcal{P}$. The conclusion is that for relatively small computational budgets, aMCMC is compminimax optimal. As the computational budget increases, the advantage of aMCMC diminishes and eventually exact MCMC is compminimax optimal.

Using the theortical results and optimality concept described above, we analyze three non-trivial aMCMC algorithms: Gibbs sampling for logistic regression using random minibatches of data; low-rank approximations to Gaussian process regression, and a novel approximate algorithm we propose for computation of the tensor decomposition model described in Chapter 2. An important conclusion is that to achieve control of the approximation error to the posterior, many aMCMC algorithms must adapt to the current state of the Markov chain. Empirical studies showed that in all three cases, it is possible to achieve significant speedup with minimal loss of estimation accuracy *for certain parameters and functionals*. Thus, another important conclusion is that tradeoffs between estimation accuracy and computation time must be assessed in the context of inferential goals.

1.2.6 Scaling behavior of MCMC samplers for generalized linear models.

The scaling properties of MCMC algorithms as n or p grow without bound are useful in understanding the performance of these algorithms in large-scale problems. This generally must be addressed on a case-by-case basis for different MCMC algorithms. As yet, little work has been done to characterize these properties for MCMC algorithms for generalized linear models, despite the importance of these models in applied statistics. The most common MCMC algorithms for generalized linear models utilize data augmentation Gibbs sampling. Motivated by a challenging dataset from quantitative advertising, we show in Chapter 7 that these algorithms exhibit poor mixing and convergence properties in large samples. The cause of this deficit is not specific to a particular algorithm, but rather a feature of data augmentation that is difficult to avoid. Additionally, the results imply that the usual paradigm in machine learning and computer science of characterizing polynomial time algorithms as "scalable" or "fast" is not always relevant to MCMC. The algorithms we study have convergence times that grow polynomially in n, but are practically useless in

large samples, providing support that quasilinear time algorithms may be required to ensure practical performance in large-scale problems.

2

Tensor Decompositions and Sparse Log-linear Models

2.1 Introduction

Parsimonious models for contingency tables are of growing interest due to the routine collection of data on moderate to large numbers of categorical variables. We study the relationship between two paradigms for inference in contingency tables: the log-linear model (Fienberg and Rinaldo (2007), Bishop et al. (2007), Agresti (2002)) and latent structure models (Stouffer et al. (1950), Gibson (1955), Lazarsfeld and Henry (1968), Anderson (1954), Madansky (1960), Haberman (1974), Goodman (1974)) that induce a tensor decomposition of the joint probability mass function (Dunson and Xing (2009), Bhattacharya and Dunson (2012)). We aim to understand situations where the joint probability corresponding to a sparse log-linear model has a low rank tensor factorization. Connecting the seemingly distinct notions of parsimony in the two parameterizations can motivate the use of factorizations having a combination of computational tractability and flexibility.

Let $V = \{1, \ldots, p\}$ denote a set of p categorical variables. We use $(y_j, j \in V)$

to denote variables, with $y_j \in \mathcal{I}_j$ having $d_j = |\mathcal{I}_j|$ levels. Without loss of generality, we assume $\mathcal{I}_j = \{1, \ldots, d_j\}$. Let $\mathcal{I}_V = \times_{j \in V} \mathcal{I}_j$. Elements of $\mathcal{I}_V$ are referred to as cells of the contingency table; there are $\prod_{j=1}^{p} d_j$ cells in total. We generically denote a cell by $\boldsymbol{i}$, with $\boldsymbol{i} = (i_1, \ldots, i_p) \in \mathcal{I}_V$. The joint probability mass function of $\boldsymbol{y} = (y_1, \ldots, y_p)$ is denoted by π, with

$$\pi_{i_1 \ldots i_p} = Pr(y_1 = i_1, \ldots y_p = i_p), \quad \boldsymbol{i} \in \mathcal{I}_V. \tag{2.1}$$

A p-way tensor $M \in \mathbb{R}^{d_1 \times \ldots \times d_p}$ is a multiway-array which generalizes matrices to higher dimensions Kolda and Bader (2009). Two common forms of tensor decomposition which extend the matrix singular value decomposition are the PARAFAC Harshman (1970) and Tucker Tucker (1966); De Lathauwer et al. (2000a,b) decompositions. Note that $\pi = (\pi_{i_1 \ldots i_p})_{\boldsymbol{i} \in \mathcal{I}_V}$ can be identified with a $\mathbb{R}^{d_1 \times \ldots \times d_p}$-*probability tensor*, which is a non-negative tensor with entries summing to one. Given n i.i.d. replicates of $\boldsymbol{y}$, let $\boldsymbol{n}(\boldsymbol{i})$ denote the cell-count of cell $\boldsymbol{i}$. We assume the cell counts are multinomially distributed according to the probabilities in π.

Inference for contingency tables often employs log-linear models that express the logarithms of the entries in π as a linear function of parameters related to the index of each cell. Most of these parameters relate to interactions between the variables Agresti (2002). A saturated log-linear model has as many parameters as π has cells. To reduce dimensionality, it is common to assume a large subset of the interaction parameters are zero, and estimate the model using L_1 regularization Roth and Fischer (2008); Nardi and Rinaldo (2012), decomposition approaches Dahinden et al. (2010), or Bayesian model averaging Dobra and Lenkoski (2011); Dobra et al. (2004); Massam et al. (2009). Zero interaction terms are easily interpreted in terms of conditional and marginal independence relationships among the variables. A significant literature exists on Bayesian inference for log-linear models, focusing mainly on the development of novel conjugate priors Dawid and Lauritzen (1993); Massam et al.

(2009), model selection/averaging Letac and Massam (2012); Hu et al. (2009), and stochastic search algorithms to explore the model space (e.g. Dobra and Massam (2010)).

An alternative approach is to assume that the p variables are conditionally independent given one or more discrete latent class indices, with dependence induced upon marginalization over the latent variable(s). The attractiveness of such latent class models arises partly from easy model fitting using data-augmentation, with a Bayesian nonparametric formulation allowing the number of latent classes to be learned from the data Dunson and Xing (2009). Dunson and Xing (2009) showed that a single latent class model is equivalent to a reduced-rank non-negative PARAFAC decomposition of the joint probability tensor π, while the multiple latent class model in Bhattacharya and Dunson (2012) implied a Tucker decomposition. See also Zhou et al. (2013) and Kunihama and Dunson (2012) for extensions of these models to more complex settings.

Latent class models and log-linear models can be unified within a larger class of graphical models with observed and unobserved variables (see e.g. Lauritzen (1996); Humphreys and Titterington (2003)). In particular, Geiger et al. (2001) describes relationships between the number of components in a PARAFAC expansion of π and the topological structure of the corresponding parameter space of a log-linear model, with consequences for estimation and selection in latent structure models. Others have established additional connections between latent structure models and the algebraic topology of the log-linear model Settimi and Smith (1998); Smith and Croft (2003); Rusakov and Geiger (2002); Garcia et al. (2005); Fienberg et al. (2007); Letac and Massam (2012).

These two classes of models impose sparsity (or parsimony) in seemingly different ways, and to best of our knowledge, no connection has been established yet in this regard. The class of sparse log-linear models is often considered a desirable data

generating class in high dimensional settings for flexibility and ease of interpretation, and it is important to determine whether there exist low rank expansions for probability tensors corresponding to sparse log-linear models. Determining whether a nontrivial relationship exists is a major focus of the paper. Working with a class of weakly hierarchical log-linear models, we provide precise bounds on the tensor ranks of sparse log-linear models. There are limited results on ranks of higher-order tensors, and the techniques developed here may be of independent interest.

The complementary goal of this work is to leverage insights from our theoretical study to develop improved classes of factorization models that provide computationally tractable alternatives to sparse log-linear models. Sparse log-linear models are appealing in terms of interpretation and flexibility but unfortunately cannot be implemented practically in high dimensions. Motivated by our theoretical results that usual latent class models require many extra parameters to characterize sparse log-linear models, we propose a new class of collapsed Tucker (c-Tucker) factorizations. These factorizations can parsimoniously characterize complex interactions in categorical data, including data generated from sparse log-linear models. We propose Bayesian methods for analyzing data under c-Tucker models, demonstrating advantages over usual PARAFAC-type latent class models.

This paper is organized as follows. Section 2 introduces notation and provides background relevant to log-linear models and latent structure models. Section 3 provides our main theoretical results on the rank of probability tensors corresponding to sparse log-linear models, and defines classes of sparse log-linear models corresponding to relatively low rank probability tensors. Section 4 introduces and motivates the proposed collapsed Tucker model. Section 5 presents a numerical study of the Bayesian collapsed Tucker model, focusing on its performance in estimation of π and the parameters of a log-linear model; we also show close agreement to an alternative method on a real data example. Section 6 gives further discussion of results and

implications.

2.2 Notation and background

We introduce some notation and background on log-linear models and tensor decompositions. Additional notation will be introduced in Section 3.

2.2.1 Log-linear models

A standard approach to contingency table analysis parametrizes π as a log-linear model satisfying certain constraints. For a subset of variables $E \subset V$, we adopt the notation of Massam et al. (2009) to denote by $\boldsymbol{i}_E$ the cells in the marginal E-table, so that $\boldsymbol{i}_E \in \mathcal{I}_E := \times_{j \in E} \mathcal{I}_j$. Let $\theta_E(\boldsymbol{i}_E)$ denote the interaction among the variables in E corresponding to the levels in $\boldsymbol{i}_E$. With this notation, a log-linear model assumes the form

$$\log(\pi_{\boldsymbol{i}}) = \sum_{E \subset V} \theta_E(\boldsymbol{i}_E). \tag{2.2}$$

As a convention, $\theta_\emptyset$ corresponds to $E = \emptyset$. To identify the model we choose the corner parameterization Agresti (2002); Massam et al. (2009), which sets $\theta_E(\boldsymbol{i}_E) = 0$ if there exists $j \in E$ such that $i_j = 1$. In the binary setting ($d_j = 2$ for all j) with corner parametrization, any E for which $\theta_E(\boldsymbol{i}_E) \neq 0$ must have every element of $\boldsymbol{i}_E$ equal to 2. In this case we will represent $\theta_E(\boldsymbol{i}_E)$ as θ_E since there is no ambiguity. When $d > 2$, the notation θ_E refers to the collection of parameters $\{\theta_E(\boldsymbol{i}_E) : \boldsymbol{i}_E \in \mathcal{I}_E\}$, and $\theta_E = 0$ indicates $\theta_E(\boldsymbol{i}_E) = 0$ for all $\boldsymbol{i}_E \in \mathcal{I}_E$.

Let $\boldsymbol{\theta} = \{\theta_E(\boldsymbol{i}_E) : i_\gamma \neq 1, \forall\, \gamma \in E\}$ denote the collection of free model parameters and S_θ denote the collection of nonzero elements of $\boldsymbol{\theta}$. A saturated model includes all free model parameters, so that $|S_\theta| = \prod_j d_j - 1$. Although any model that is not saturated is technically sparse, when we refer to sparse log-linear models we have in mind settings where $|S_\theta| \ll \prod_j d_j - 1$. We will be primarily concerned with how the

degree and structure of sparsity affects the nonnegative tensor rank of π.

An attractive feature of log-linear models is that the parameters are interpretable as defining conditional and marginal independence relationships between the y_j's. A log-linear model is hierarchical Massam et al. (2009); Dellaportas and Forster (1999); Darroch et al. (1980) if for every $E \subset V$ for which $\theta_E = 0$, we have $\theta_F = 0$ for all $F \supseteq E$. Here we work with a more general class of log-linear models that contains hierarchical models. We refer to this class as weakly hierarchical.

Definition 2.2.1. A log-linear model is weakly hierarchical when the following condition is satisfied: if $\theta_E(\boldsymbol{i}_E) = 0$ for $E \subset V$ and $\boldsymbol{i}_E \in \mathcal{I}_E$, then $\theta_F(\boldsymbol{i}'_F) = 0$ for every $F \supseteq E$ and $\boldsymbol{i}'_F \in \mathcal{I}_F$ such that $i'_j = i_j$ for all $j \in E$.

When $d_j = 2$ for all j, weakly hierarchical models and hierarchical models define identical subsets of log-linear models, but if any $d_j > 2$, the collection of hierarchical models is a proper subset of the collection of weakly hierarchical models. To see this, suppose a model is weakly hierarchical. Assume $\theta_E = 0$. Then, $\theta_E(\boldsymbol{i}_E) = 0$ for all $\boldsymbol{i}_E \in \mathcal{I}_E$. Let $F \supseteq E$. For any $\boldsymbol{i}'_F \in \mathcal{I}_F$, $\theta_F(\boldsymbol{i}'_F) = 0$ by weak hierarchicality, since $\theta_E(\boldsymbol{i}'_E) = 0$. Since $\boldsymbol{i}'_F$ is arbitrary, we must have $\theta_F = 0$, proving hierarchicality.

The essential difference between hierarchical and weakly hierarchical models is illustrated by the following example. Let $V = \{1, 2, 3\}$ and $d_1 = d_2 = d_3 = 4$. Suppose

$$S_\theta = \left\{\theta_{\{1\}}(2), \theta_{\{2\}}(2), \theta_{\{3\}}(2), \theta_{\{1,2\}}(2,2), \theta_{\{1,3\}}(2,2), \theta_{\{2,3\}}(2,2), \theta_{\{1,2,3\}}(2,2,2)\right\}.$$

In other words, any interactions that correspond to all variables in E taking level 2 are nonzero, and all others are zero. This model is weakly hierarchical but not hierarchical. For a model to be hierarchical, the collection of nonzero parameters must be uniquely specified by a generator – a collection of subsets of V. For weakly hierarchical models, some interactions corresponding to a single subset E may be zero and others nonzero, so long as Definition 2.2.1 is satisfied.

2.2.2 Tensor Factorization Models

An alternative to log-linear models is latent structure analysis (Stouffer et al. (1950); Gibson (1955); Lazarsfeld and Henry (1968); Anderson (1954); Madansky (1960); Haberman (1974); Goodman (1974)), which assumes the $y_1, \ldots, y_p$ are conditionally independent given one or more latent class variables. In marginalizing out the latent class variables, one obtains a tensor decomposition of π. Latent structure models inducing PARAFAC and Tucker decompositions are briefly reviewed below.

PARAFAC models

An m-component non-negative PARAFAC decomposition Harshman (1970) of a probability tensor π is given by

$$\pi = \sum_{h=1}^{m} \nu_h \lambda_h^{(1)} \otimes \ldots \otimes \lambda_h^{(p)} = \sum_{h=1}^{m} \nu_h \bigotimes_{j=1}^{p} \lambda_h^{(j)}, \tag{2.3}$$

where $\otimes$ denotes an outer product[1], each $\lambda_h^{(j)} \in \Delta^{(d_j-1)}$ is an element of the $(d_j - 1)$ dimensional simplex[2], and $\nu \in \Delta^{(m-1)}$. Element wise, $\pi_{i_1 \ldots i_p} = \sum_{h=1}^{m} \nu_h \prod_{j=1}^{p} \lambda_{hi_j}^{(j)}$. By constraining ν and the $\lambda_h^{(j)}$s to be probability vectors, it is ensured that the entries of π are non-negative and sum to one. The vectors $\lambda_h^{(j)}$ are referred to as the arms of the tensor decomposition.

A probabilistic PARAFAC decomposition (Dunson and Xing (2009)) of π can be induced by a single index latent class model

$$\begin{aligned} y_j \mid z &\overset{\text{ind.}}{\sim} \text{Multi}(\{1, \ldots, d_j\}, \lambda_{z1}^{(j)}, \ldots, \lambda_{zd_j}^{(j)}), \\ \Pr(z = h) &= \nu_h, h = 1, \ldots, m. \end{aligned} \tag{2.4}$$

Marginalizing over the latent variable z, we obtain expression (2.3).

[1] $\{\bigotimes_{j=1}^{p} \lambda_h^{(j)}\}_{i_1,\ldots i_p} = \prod_{j=1}^{p} \lambda_{hij}^{(j)}$

[2] $\Delta^{(r-1)} = \{x \in \mathbb{R}^r : x_j \geq 0 \, \forall \, j, \sum_{j=1}^{r} x_j = 1\}$

Unlike matrices, there is no unambiguous definition of the rank of a tensor. A notion of tensor rank is derived restricting attention to PARAFAC expansions. The nonnegative PARAFAC rank of a nonnegative tensor M is the minimal value of m for which there exist nonnegative vectors $\tilde{\lambda}_h^{(j)}$ such that

$$M = \sum_{h=1}^{m} \bigotimes_{j=1}^{p} \tilde{\lambda}_h^{(j)}. \tag{2.5}$$

We will denote the nonnegative PARAFAC rank of a tensor M as $\text{rnk}_P^+(M)$. In the case of probability tensors, the definition in (2.5) is equivalent to the minimum m such that (2.3) holds, since the weights ν_h can be absorbed into the arms $\lambda_h^{(j)}$. For probability tensors, we can always write a trivial PARAFAC expansion exploiting the probabilistic structure as

$$\begin{aligned} \pi_{i_1 \ldots i_p} &= Pr(y_1 = i_1 \mid y_2 = i_2, \ldots, y_p = i_p) Pr(y_2 = i_2, \ldots, y_p = i_p) \\ &= \sum_{c_2 \in \mathcal{I}_2} \cdots \sum_{c_p \in \mathcal{I}_p} Pr(y_1 = i_1 \mid y_2 = c_2, \ldots, y_p = c_p) \mathbb{1}_{(c_2 = i_2, \ldots, c_p = i_p)} \\ &\times Pr(y_2 = c_2, \ldots, y_p = c_p). \end{aligned} \tag{2.6}$$

To see the correspondence with (2.3), introduce one level of h for each distinct value of the multi-index $(c_2, \ldots c_p)$ so that $m = \prod_{j=2}^{p} d_j$, and set $\nu_h = Pr(y_2 = c_2, \ldots, y_p = c_p)$, $\lambda_{hi_1}^{(1)} = Pr(y_1 = i_1 \mid y_2 = c_2, \ldots, y_p = c_p)$ and $\lambda_{hi_j}^{(j)} = \mathbb{1}_{(i_j = c_j)}$ for $j = 2, \ldots, p$. As a consequence, we obtain an upper bound of d^{p-1} on the nonnegative PARAFAC rank $\text{rnk}_P^+(\pi)$ when $d_j = d$ for all j. Thus, every nonnegative tensor has finite nonnegative PARAFAC rank, and the single latent class model has full support.

Tucker models

An m-component nonnegative Tucker decomposition Tucker (1966); De Lathauwer et al. (2000a) alternatively expresses the entries in π as

$$\pi_{c_1 \ldots c_p} = \sum_{h_1=1}^{m} \cdots \sum_{h_p=1}^{m} \phi_{h_1 \ldots h_p} \prod_{j=1}^{p} \lambda_{h_j c_j}^{(j)}, \tag{2.7}$$

where ϕ is an m^p *core* probability tensor and $\lambda_h^{(j)} \in \Delta^{d_j-1}$ for every h and j. The Tucker decomposition can be thought of as a weighted sum of m^p tensors each having PARAFAC rank one with weights given by the entries in ϕ; conversely, the PARAFAC is a special case of the Tucker decomposition where the core is an $m \times 1$ probability vector.

A probabilistic Tucker expansion of a probability tensor π can be induced by a latent class model with a vector of latent class indicators $z = (z_1, \ldots, z_p)$,

$$y_j \mid z \overset{\text{ind.}}{\sim} \text{Multi}(\{1, \ldots, d_j\}, \lambda_{z_j 1}^{(j)}, \ldots, \lambda_{z_j d_j}^{(j)}),$$
$$\Pr(z_1 = h_1, \ldots, z_p = h_p) = \phi_{h_1 \ldots h_p}. \tag{2.8}$$

From this, it is clear that ϕ parametrizes the joint distribution of the latent variables $z_1, \ldots, z_p$. See Bhattacharya and Dunson (2012) for a class of hierarchical models that induce a structured Tucker decomposition of a probability tensor.

The Tucker decomposition gives rise to an alternative definition of the nonnegative tensor rank of a tensor M as the minimal value of m such that M can be expressed exactly by an expansion of the form in (2.7). We will denote the non-negative Tucker rank of a tensor M as $\text{rnk}_T^+(M)$. In the case where $d_j = d$ for all j, an argument similar to the one in (2.6) shows that for probability tensors π, $\text{rnk}_T^+(\pi) \leq d$. The scale of Tucker ranks is quite different from that of PARAFAC ranks because the core itself has dimension m^p. Therefore, in modeling it is common to choose a parsimonious representation of the core, an issue we revisit in Section 2.4.

2.3 Main results: PARAFAC rank of sparse log-linear models

2.3.1 PARAFAC rank result for general p and d

We now provide bounds on the non-negative PARAFAC rank of joint probability tensors. There are few results on ranks of tensors beyond three dimensions and

the techniques developed here are likely to be of independent interest. All proofs are deferred to the Appendix. In addition to the bounds developed in this section based on probabilistic arguments, we provide algebraic constructions in the two-dimensional case in a supplementary document (see Johndrow et al. (2014a)).

In the results that follow, we exploit the fact that a PARAFAC expansion of a probability tensor has a dual representation as a latent variable model (2.4), and the PARAFAC rank of a probability tensor can be defined in terms of the support of the corresponding latent class variable. Remark 2.3.1 re-expresses an observation from Lim and Comon (2009) that formalizes this relationship. For a nonnegative integer-valued random variable w, denote $\text{spt}(w) = \{h : Pr(w = h) > 0\}$.

Remark 2.3.1. Suppose π is a $\prod_{j=1}^{p} d_j$ probability tensor, and let $y_1, \ldots, y_p$ be categorical random variables with joint distribution defined by π. Then $\text{rnk}_P^+(\pi) = \bigwedge_{z \in \mathcal{Z}} |\text{spt}(z)|$, where $\mathcal{Z}$ is the collection of all finitely-supported, discrete latent variables z such that

$$Pr(y_1 = i_1, \ldots, y_p = i_p | z = h) = \prod_{j=1}^{p} Pr(y_j = i_j | z = h), \tag{2.9}$$

for all $h \in \text{spt}(z)$ and $\boldsymbol{i} \in \mathcal{I}_V$.

Therefore, if a latent variable z satisfying (2.9) can be constructed, then the rank of π can be at most $|\text{spt}(z)|$. Our recipe to create such discrete random variables z is to partition the probability space $\mathcal{Y}$ on which $(y_1, \ldots, y_p)$ is defined and assign z a constant value on each partition set. Since $\boldsymbol{y}$ is a mapping from $\mathcal{Y}$ to $\mathcal{I}_V$, for any partition of $\mathcal{I}_V$, the inverse images of the partition sets under the mapping $\boldsymbol{y}$ define a partition of $\mathcal{Y}$. We shall restrict our attention to such partitions of $\mathcal{Y}$. As a convention to simplify notations, we shall continue to use Pr to denote probabilities under the probability measure induced on $\mathcal{I}_V$ via the measurable map $\boldsymbol{y}$. For subsets $B_j \subset \mathcal{I}_j$, it follows from a standard property that $Pr(y_1 \in B_1, \ldots, y_p \in B_p) = Pr(\bigtimes_{j=1}^{p} B_j)$,

with the first probability defined on the σ-algebra on $\mathcal{Y}$ and the second on the product σ-algebra on $\mathcal{I}_V$. We shall henceforth identify the event $\{y_1 \in B_1, \ldots, y_p \in B_p\}$ in $\mathcal{Y}$ with the event $\times_{j=1}^{p} B_j$ in $\mathcal{I}_V$. For a set $A \in \mathcal{I}_V$, $Pr(y_1 \in B_1, \ldots, y_p \in B_p \mid A)$ is defined as $Pr[(\times_{j=1}^{p} B_j) \cap A]/Pr(A)$.

We now elaborate on the construction of z. For a partition $\mathcal{P}$ of $\mathcal{I}_V$, with $\{A_1, \ldots, A_{|\mathcal{P}|}\}$ denoting an (arbitrary) enumeration of the sets in $\mathcal{P}$, we define a discrete random variable $z = z_{\mathcal{P}}$ on $\mathcal{Y}$ corresponding to $\mathcal{P}$ as

$$z = h \mathbb{1}_{A_h}(\boldsymbol{y}),\ h = 1, \ldots, |\mathcal{P}|. \tag{2.10}$$

In particular, for partitions $\mathcal{P}_j$ of $\mathcal{I}_j$, we can define the product partition $\mathcal{P}$ as

$$\mathcal{P} = \bigtimes_{j=1}^{p} \mathcal{P}_j := \left\{ \bigtimes_{j=1}^{p} B_j : B_j \in \mathcal{P}_j \right\}. \tag{2.11}$$

It follows from properties of the Cartesian product that $\mathcal{P}$ indeed forms a partition of $\mathcal{I}_V$ and $|\mathcal{P}| = \prod_{j=1}^{p} |\mathcal{P}_j|$.

Clearly, for any z as in (2.10), (2.9) is equivalent to

$$Pr(y_1 = i_1, \ldots, y_p = i_p | A_h) = \prod_{j=1}^{p} Pr(y_j = i_j | A_h), \tag{2.12}$$

for all $h = 1, \ldots, |\mathcal{P}|$ and $\boldsymbol{i} \in \mathcal{I}_V$. We now proceed to create partitions $\mathcal{P}$ satisfying (2.12). First, observe that the trivial PARAFAC expansion in (2.6) corresponds to the product partition (2.11) with $\mathcal{P}_1 = \mathcal{I}_1$ and $\mathcal{P}_j = \{\{c_j\} : c_j \in \mathcal{I}_j\}$ for $j \geq 2$, so that the event $\{z = h\}$ for each h designates an event of the form $\mathcal{I}_1 \times \{c_2\} \times \ldots \times \{c_p\}$. Clearly, $|\mathcal{P}| = d^{p-1}$; the trivial upper bound. Our main target is to show that much tighter bounds can be achieved under the assumption of weak hierarchicality.

We introduce some additional notation here. For a variable $j \in V$, let $C_\theta^{(j)}$ denote the levels of variable j that share a non-zero two-way or higher order interaction with at least one other variable. For weakly hierarchical models, it is sufficient

to only search over the non-zero two-way interactions, so that $C_\theta^{(j)} = \{c_j \in \mathcal{I}_j :$ there exists $j' \neq j$ and $c_{j'} \in \mathcal{I}_{j'}$ such that $\theta_{\{j,j'\}}(c_j, c_{j'}) \neq 0\}$. For any $\boldsymbol{\theta}$, let $C_\theta := \{(E, \boldsymbol{i}_E) : |E| \geq 2, \theta_E(\boldsymbol{i}_E) \neq 0\}$ and $C_{\theta,2} := \{(E, \boldsymbol{i}_E) : |E| = 2, \theta_E(\boldsymbol{i}_E) \neq 0\}$. Note that C_θ is not the collection of non-zero second or higher order interactions; elements of C_θ are tuples $(E, \boldsymbol{i}_E)$ such that there is a non-zero interaction among variables in E corresponding to the levels in $\boldsymbol{i}_E$. $C_{\theta,2}$ is constructed similarly for the non-zero two-way interactions only.

If the model is weakly hierarchical, it follows from Definition 2.2.1 that for any subset C' of $(C_\theta^{(j)})^c$, $y_j \mathbb{1}_{C'}(y_j) \perp y_{[-j]}$, where $y_{[-j]} = (y_1, \ldots y_{j-1}, y_{j+1}, \ldots, y_p)$ and for random variables x_1, x_2, $x_1 \perp x_2$ indicates marginal independence. Thus, instead of having to let the levels of z vary over all events of the form $\{\{c_2\} \cap \ldots \cap \{c_p\}\}$, one can coarsen the partition $\mathcal{P}$ in (2.11) by pooling together all the levels in $(C_\theta^{(j)})^c$ to form a single element of $\mathcal{P}_j$. Further improvement can be achieved by scanning through the variables in a particular order and only considering the subset of $C_\theta^{(j)}$ that correspond to non-zero two-way interactions with variables that appear later in the ordering. We formalize this observation in Theorem 2.3.1 below.

Theorem 2.3.1. *Suppose π is a d^p probability tensor corresponding to a weakly hierarchical log-linear model. Let σ be a permutation on V. For each $j = 1, \ldots, p-1$, denote $G_\sigma^{(j)} = \{\sigma(j+1), \ldots, \sigma(p)\}$ and define $B_{\sigma(j)}$ to be the following subset of $C_\theta^{(j)}$:*

$$B_{\sigma(j)} = \{i_{\sigma(j)} \in \mathcal{I}_{\sigma(j)} : \ \exists\ f \in G_\sigma^{(j)} \text{ and } i_f \in \mathcal{I}_f \text{ s.t. } \theta_{\{\sigma(j),f\}}(i_{\sigma(j)}, i_f) \neq 0\}.$$

Then, the PARAFAC rank $rnk_P^+(\pi)$ of π is at most

$$\bigwedge_\sigma \prod_{j=1}^{p-1} \left(\mid B_{\sigma(j)} \mid +1\right).$$

The bound in Theorem 2.3.1 gives the correct upper bound d^{p-1} when the model is saturated, since then for any permutation σ we have $|B_{\sigma(j)}| = (d-1)$ for $j =$

$1, \ldots, p-1$. More importantly, it is easy to compute and provides a useful estimate of the order of the PARAFAC rank in d and/or p when the interactions are *uniformly* spread. However, if the interactions are highly structured, Theorem 2.3.1 may yield the trivial upper bound irrespective of the true rank, as seen in Example 2.3.3 below.

Our next result provides sharper bounds on the PARAFAC rank. In the first part of Theorem 2.3.2, we provide a "dimension-free" upper bound that is unaffected by increasing d as long as the true PARAFAC rank is constant. We then present a tight upper bound in the second part of Theorem 2.3.2 which cannot be globally improved in the class of weakly hierarchical log-linear models.

Theorem 2.3.2. *Suppose π is a probability tensor corresponding to a weakly hierarchical log-linear model. Let $H = \{H_1, \ldots, H_p\}$ denote collections of sets of indices, where each $H_j \subset \mathcal{I}_j$. Given H, define $T_{(C_\theta, H)} = \{(E, \boldsymbol{i}_E) \in C_\theta : i_j \in H_j \text{ for some } j \in E\}$ and let*

$$\mathscr{H} = \{H : T_{(C_\theta, H)} = C_\theta\}. \tag{2.13}$$

Assume $C_\theta^{(j)} \neq \emptyset$ for all j. Then,

$$rnk_P^+(\pi) \leq \bigwedge_{H \in \mathscr{H}} \left(\prod_{j \in V} (\mid H_j \mid +1) \right). \tag{2.14}$$

For any $l \in V$, set $W_l = \{j \in V \setminus \{l\} \ : \ \mid H_j \mid = d-1\}$ and $\bar{W}_l = V \setminus W_l$. Then, a tight upper bound on $rnk_P^+(\pi)$ is

$$\bigwedge_{H \in \mathscr{H}} \bigwedge_{l \in V} \left(\prod_{j \in V} (\mid H_j \mid +1) - \Big[\prod_{j \in W_l} (\mid H_j \mid +1) \Big] \Big[\prod_{j \in \bar{W}_l} \mid H_j \mid \Big] \right). \tag{2.15}$$

The full proof of Theorem 2.3.2 is provided in Appendix A.1; Example 2.3.4 illustrates the main ideas of the proof.

Remark 2.3.2. By definition, $T_{(C_\theta, H)} \subset C_\theta$, so the condition $T_{(C_\theta, H)} = C_\theta$ in the definition of $\mathscr{H}$ in (2.13) equivalently requires that for every $(E, \boldsymbol{i}_E) \in C_\theta$, $i_j \in H_j$ for

some $j \in E$. Moreover, for weakly hierarchical models, $T_{(C_\theta,H)} = C_\theta \Leftrightarrow T_{(C_{\theta,2},H)} = C_{\theta,2}$.

Remark 2.3.3. Theorem 2.3.2 assumes $C_\theta^{(j)} \neq \emptyset$ for all j, i.e., every variable shares at least one second order interaction. Clearly, the set of variables which do not satisfy the condition are marginally independent of all other variables and do not contribute to the rank. Letting $U = \{j : C_\theta^{(j)} = \emptyset\}$, the statement of Theorem 2.3.2 will continue to hold without this assumption as long as we replace all instances of V by $V^* = V \setminus U$.

2.3.2 Illustrative Examples

In this subsection, we present two examples to highlight the refinement of the bounds in Theorem 2.3.2 over Theorem 2.3.1 and illustrate the main ideas behind the proof of Theorem 2.3.2.

In the setting of Example 2.3.3 below, the expressions in (2.14) and (2.15) can be explicitly calculated to illustrate the improvement over Theorem 2.3.1.

Example 2.3.3. Suppose $p = 2$ and $d_1 = d_2 = d$. Assume $\theta_{\{1,2\}}(2, c_2) \neq 0$ for all $c_2 \geq 2$, $\theta_{\{1,2\}}(c_1, 2) \neq 0$ for all $c_1 \geq 2$ and $\theta_{\{1,2\}}(c_1, c_2) = 0$ otherwise. Thus, level 2 of variable 1 interacts with all levels except 1 of variable 2, and similarly, level 2 of variable 2 interacts with all levels except 1 of variable 1. In addition, for convenience of illustration, also assume that all main effects are zero[3], so that

$$\log \pi_{i_1 i_2} = \theta_0 + \theta_{\{1,2\}}(i_1, i_2)\mathbb{1}_{(i_i=2, i_2\geq 2)} + \theta_{\{1,2\}}(i_1, i_2)\mathbb{1}_{(i_i\geq 2, i_2=2)}.$$

Letting J_d denote the $d \times d$ matrix given by $v_1 \otimes v_2$, where $\{v_1\}_{i_1} = \mathbb{1}_{i_1 \neq 2}$ and $\{v_2\}_{i_2} = \mathbb{1}_{i_2 \neq 2}$, we can write $\pi = e^{\theta_0} J_d + \tilde{\pi}$, where $\tilde{\pi}$ is a $d \times d$ non-negative matrix

[3] Here and in several later examples, we assume that the main effects $\{\theta_E(\boldsymbol{i}_E) : |E| = 1\}$ are zero for notational brevity. While formally these models are not weakly hierarchical, the inclusion of nonzero main effects do not influence the PARAFAC rank and hence this assumption can be made without loss of generality.

with entries

$$\tilde{\pi}_{i_1 i_2} = e^{\theta_0 + \theta_{\{1,2\}}(i_1,2)\mathbb{1}_{(i_2=2)} + \theta_{\{1,2\}}(2,i_2)\mathbb{1}_{(i_1=2)}} \mathbb{1}_{(i_i=2 \text{ or } i_2=2)}.$$

Note that $\tilde{\pi}$ is everywhere zero except in the second row and column. In case of non-negative matrices, $\text{rnk}_P^+(A)$ equals the ordinary matrix rank $\text{rnk}(A)$ when $\text{rnk}(A) \leq 2$ (see Gregory and Pullman (1983)). It is easy to see that the ordinary matrix rank of $\tilde{\pi}$ is 2, since there are at most two linearly independent columns. Hence, $\text{rnk}_P^+(\tilde{\pi}) = 2$ and applying Lemma A.1.1 in the Appendix, we conclude $\text{rnk}_P^+(\pi) \leq 1 + \text{rnk}_P^+(\tilde{\pi}) \leq 3$. Barring pathological cases, the ordinary rank $\text{rnk}(\pi)$ will always be 3, and since $\text{rnk}_P^+(A) \geq \text{rnk}(A)$ for matrices (Cohen and Rothblum (1993)), $\text{rnk}_P^+(\pi)$ will also be exactly 3.

In applying Theorem 2.3.1, we have $|B_1| = |B_2| = d - 1$, so that we always get the trivial upper bound d irrespective of the choice of σ.

Next, apply Theorem 2.3.2. Observe that $H = \{\{2\}, \{2\}\} \in \mathscr{H}$, since all of the interaction terms have either $c_1 = 2$ or $c_2 = 2$, and hence the upper bound in (2.14) is reduced to 4 irrespective of the value of d. With this choice of H, the expression inside the minimum in (2.15) becomes $(|H_1|+1)(|H_2|+1) - |H_1||H_2| = 4 - 1 = 3$, which returns the exact rank.

As in case of Theorem 2.3.1, the main strategy of proving Theorem 2.3.2 is to carefully construct a partition $\mathcal{P}$ of $\mathcal{I}_V$ and define z as in (2.10). In this case generate a partition utilizing the sets H_j and establish the conditional independence (2.12) exploiting the definition of $\mathscr{H}$. Let $\bar{H}_j = \mathcal{I}_j \backslash H_j$ and let $\mathcal{P}_{H,j}$ denote the partition of $\mathcal{I}_j$ consisting of the singleton sets $\{i_j\}$ for $i_j \in H_j$ and the set $\bar{H}_j$. Define a partition $\mathcal{P}_H^0$ of $\mathcal{I}_V$ as the Cartesian product (2.11) of the partitions $\mathcal{P}_{H,j}$. It is then immediate that $|\mathcal{P}_j| = |H_j|+1$ and hence $|\mathcal{P}| = \prod_{j=1}^{p}(|H_j|+1)$. The non-trivial aspect of the proof of (2.14) is to show that for any $H \in \mathscr{H}$, $y_1, \ldots, y_p$ are conditionally independent given any set A in $\mathcal{P}_H^0$. The tight upper bound in (2.15) of Theorem

2.3.2 exploits that certain sets in $\mathcal{P}_H^0$ can be merged without sacrificing conditional independence. Although detailed proofs of these facts are provided in Appendix A.1, we highlight the salient features in Example 2.3.4, which is an extension of Example 2.3.3 to higher dimensions with a more complicated interaction structure.

Example 2.3.4. Let $p = 5$ with $d \geq 4$ and suppose S_θ is given by

$$\begin{aligned}
&\theta_{\{1,2\}}(2, c_2) \neq 0 \text{ for } c_2 \geq 2 \qquad \theta_{\{2,3\}}(2, c_3) \quad \neq 0 \text{ for } c_3 \geq 2 \\
&\theta_{\{3,4\}}(2, c_4) \neq 0 \text{ for } c_4 \geq 2 \qquad \theta_{\{4,5\}}(2, c_5) \quad \neq 0 \text{ for } c_5 \geq 2 \\
&\theta_{\{1,5\}}(c_1, 2) \neq 0 \text{ for } c_1 \geq 2 \qquad \theta_{\{2,4\}}(2, c_4) \quad \neq 0 \text{ for } c_4 \geq 2 \\
&\theta_{\{1,4\}}(2, c_4) \neq 0 \text{ for } c_4 \geq 2 \qquad \theta_{\{1,2,4\}}(2, 2, 4) \neq 0 \\
&\theta_{\{2,5\}}(2, c_5) \neq 0 \text{ for } c_5 \geq 2 \qquad \theta_{\{1,5\}}(2, c_5) \quad \neq 0 \text{ for } c_5 \geq 2 \\
&\theta_{\{1,2,5\}}(2, 2, 4) \neq 0,
\end{aligned}$$

so there are two nonzero three-way interactions. It is not difficult to see that Theorem 3.1 gives the trivial bound of d^4 for all $5! = 120$ permutations. Now, let $H_j = \{2\}$ for each j, so that $H = \{\{2\}, \{2\}, \{2\}, \{2\}, \{2\}\}$. From (2.3.2), we can verify that $H \in \mathscr{H}$. Hence, the conclusion of (2.14) holds and $\text{rnk}_P^+(\pi) \leq 2^5 = 32$, a massive reduction.

As an illustration of the proof technique, we now show that

1. (2.12) holds with a specific $A \in \mathcal{P}_H^0$ and a specific cell $\boldsymbol{i} \in A$, providing intuition for the proof of (2.14);

2. (2.12) continues to hold when two example sets in $\mathcal{P}_H^0$ that have $(|V|-1)$ identical coordinate projections that are singleton sets are merged, providing intuition for the proof of (2.15); and,

3. that (2.12) fails when two example sets in $\mathcal{P}_H^0$ that do not have $(|V|-1)$ identical coordinate projections that are singleton sets are merged, providing a heuristic for the tightness of (2.15).

Since $H_j = \{2\}$, $\bar{H}_j = \{1, 3, \ldots, d\}$; we shall denote this by $\{\neq 2\}$ for brevity. The partition $\mathcal{P}_{H,j}$ of $\mathcal{I}_j$ therefore consist of the two sets $\{2\}$ and $\{\neq 2\}$ for each $j = 1, \ldots, 5$ and the partition $\mathcal{P}_H^0$ has 32 elements.

Part 1

Consider the event $A = \{2\} \times \{2\} \times \{2\} \times \{\neq 2\} \times \{\neq 2\} \in \mathcal{P}_H^0$ and the cell $\boldsymbol{i} = (2, 2, 2, 4, 4)$. We show that (2.12) holds with A and $\boldsymbol{i}$, i.e., if A^* denotes the event $\{\boldsymbol{y} = \boldsymbol{i}\}$ then

$$\begin{aligned} Pr(A^* \mid A) =& Pr(y_1 = 2 \mid A) Pr(y_2 = 2 \mid A) Pr(y_3 = 2 \mid A) \\ & \times Pr(y_4 = 4 \mid A) Pr(y_5 = 4 \mid A) \\ =& 1 \times 1 \times 1 \times Pr(y_4 = 4 \mid A) Pr(y_5 = 5 \mid A). \end{aligned} \tag{2.16}$$

Now notice that

$$Pr(y_4 = 4|A) = \sum_{c_5 \neq 2} \frac{\pi_{2224c_5}}{Pr(A)} = Pr(A^*|A) \sum_{c_5 \neq 2} \frac{\pi_{2224c_5}}{\pi_{22244}}$$

and similarly

$$Pr(y_5 = 4|A) = \sum_{c_4 \neq 2} \frac{\pi_{222c_44}}{Pr(A)} = Pr(A^*|A) \sum_{c_4 \neq 2} \frac{\pi_{222c_44}}{\pi_{22244}}.$$

So (2.16) is equivalent to showing

$$\frac{Pr(A)}{\pi_{22244}} = \sum_{c_4 \neq 2} \sum_{c_5 \neq 2} \frac{\pi_{2224c_5}}{\pi_{22244}} \frac{\pi_{222c_44}}{\pi_{22244}}.$$

Since

$$\frac{Pr(A)}{\pi_{22244}} = \sum_{c_4 \neq 2} \sum_{c_5 \neq 2} \frac{\pi_{222c_4c_5}}{\pi_{22244}} = \sum_{c_4 \neq 2} \sum_{c_5 \neq 2} \frac{\pi_{222c_4c_5}}{\pi_{222c_44}} \frac{\pi_{222c_44}}{\pi_{22244}},$$

we need to show that

$$\frac{\pi_{222c_4c_5}}{\pi_{222c_44}} = \frac{\pi_{2224c_5}}{\pi_{22244}}. \tag{2.17}$$

All main effects and interactions that correspond to variables $y_1, \ldots, y_4$ will be eliminated in the ratios on both sides, so we focus only on those involving y_5. This gives us that the LHS of (2.17) – assuming $c_5 \neq 4$ – is

$$\exp\left(\theta_{\{5\}}(c_5) - \theta_{\{5\}}(4) + \theta_{\{1,5\}}(2, c_5) - \theta_{\{1,5\}}(2, 4) + \theta_{\{2,5\}}(2, c_5)\right.$$
$$\left. -\theta_{\{2,5\}}(2, 4) - \theta_{\{1,2,5\}}(2, 2, 4)\right).$$

The RHS differs only in the value of y_4, but since there are no $\{4, 5\}$ interactions at these levels of the variables and the level of y_4 is the same in the numerator and denominator on the RHS, equality holds in (2.17), despite the fact that there are nonzero three-way interactions. Note that $\theta_{\{1,2,4\}}(2, 2, 4)$ cancelled on the RHS and was either zero or cancelled on the LHS as well (the latter occuring when $c_4 = 4$).

Part 2

Fix $l = 5$. The partition $\mathcal{P}_H^0$ contains the sets

$$A^\delta = \{\{2\} \times \{2\} \times \{2\} \times \{2\} \times \{2\}\}$$
$$A^\beta = \{\{2\} \times \{2\} \times \{2\} \times \{2\} \times \{\neq 2\}\}.$$

These sets share $|V|-1 = 4$ coordinate projections that are singleton sets, e.g. the set $\{2\}$ corresponding to variables 1 through 4. Now set

$$A^\epsilon = A^\delta \cup A^\beta = \{\{2\} \times \{2\} \times \{2\} \times \{2\} \times \mathcal{I}_5\}$$

and put $\mathcal{P}_H^1 = (\mathcal{P}_H^0 \setminus A^\beta, A^\delta) + A^\epsilon$. Following the argument in the display after (2.11), we have conditional independence given A^ϵ. Since this is the only set in $\mathcal{P}_H^1$ that is not in $\mathcal{P}_H^0$, $\mathcal{P}_H^1$ satisfies (2.12). Therefore, we see that it is possible to merge two sets that have $(|V|-1)$ identical coordinate projections that are singleton sets to create a coarser partition that continues to satisfy (2.12). Though we do not show it rigorously in this example, it is only possible to merge two sets of this form in $\mathcal{P}_H^0$ while maintaining conditional independence, giving us the upper bound $\text{rnk}_P^+(\pi) = 2^5 - 1 = 31$. The same value is given by (2.15).

Part 3

We now utilize the same setup to demonstrate the key argument as to why (2.15) is tight. This principle can be described succinctly as the failure of conditional independence upon replacing sets in the partition $\mathcal{A}_H^0$ with their union when these sets do not have in common at least $|V|-1$ identical singleton events. Let A^β and A^* be as in Parts 2 and 1, respectively, and set

$$A^\gamma = \{\{2\}, \{2\}, \{2\}, \{\neq 2\}, \{\neq 2\}\}$$

and note that A^γ and A^β share $3 = |V|-2$ identical coordinate projections which are singleton events. Then

$$A^\gamma \cup A^\beta = \{\{2\}, \{2\}, \{2\}, \mathcal{I}_4, \{\neq 2\}\}.$$

Since A^γ and A^β share only $|V|-2$ singleton coordinate projections, (2.12) should fail if we merge these sets. So we want to show that

$$Pr(A^* \mid A) \neq Pr(\mathcal{I}_4 \mid A) Pr(\{\neq 2\} \mid A).$$

This will be true iff

$$\frac{\pi_{222c_4c_5}}{\pi_{222c_44}} \neq \frac{\pi_{2224c_5}}{\pi_{22244}} \tag{2.18}$$

for one or more values of $c_4 \in A_4, c_5 \in A_5$. Here, unlike our previous example using this setup, c_4 can take any value in $\mathcal{I}_4$, *including the value 2.* However, $\theta_{\{4,5\}}(2, c_5) \neq 0$ for any $c_5 \geq 2$. So now on the LHS of (2.18) we get

$$\exp\{\theta_{\{5\}}(c_5) - \theta_{\{5\}}(4) + \theta_{\{1,5\}}(2, c_5) - \theta_{\{1,5\}}(2, 4) + \theta_{\{2,5\}}(2, c_5) -$$

$$\theta_{\{2,5\}}(2, 4) - \theta_{\{1,2,5\}}(2, 2, 4) + \theta_{\{4,5\}}(2, c_5) - \theta_{\{4,5\}}(2, 4)\}.$$

when $c_4 = 2$ and $c_5 \neq 4$. But on the RHS we still get

$$\exp\{\theta_{\{5\}}(c_5) - \theta_{\{5\}}(4) + \theta_{\{1,5\}}(2, c_5) - \theta_{\{1,5\}}(2, 4) +$$

$$\theta_{\{2,5\}}(2, c_5) - \theta_{\{2,5\}}(2, 4) - \theta_{\{1,2,5\}}(2, 2, 4)\}$$

always, so there are events contained in A where the equality fails, and therefore conditional independence does not hold.

As a concluding comment, in all the examples where we could calculate the exact rank explicitly, the bound in (2.15) produced the exact rank. However, to show that (2.15) provides the exact rank, we need an additional condition; see Remark 2.3.4 below (a proof is provided in the supplement Johndrow et al. (2014a)).

Remark 2.3.4. Suppose for every $H \in \mathscr{H}$ for which there exists $H^* \in \mathscr{H}$ such that $H_j^* \subseteq H_j$ for every j, the smallest partition $\mathcal{P}_H^{\inf}$ satisfying (2.9) that can be formed from unions of the events in $\mathcal{P}_H^0$ satisfies $|\mathcal{P}_H^{\inf}| \geq |\mathcal{P}_{H^*}^{\inf}|$. Then (2.15) gives the exact value of $\text{rnk}_P^+(\pi)$.

2.3.3 Practical consequences of rank results

We provide corollaries to Theorem 2.3.1 that give insight into cases where a relatively low PARAFAC rank can be expected. These corollaries motivate subsequent analysis of the statistical properties of latent class models. The number of parameters in a PARAFAC decomposition is given by $(k-1) + k\sum_{j=1}^{p} d_j$, where $k = \text{rnk}_P^+(\pi)$. Hence, the PARAFAC rank determines precisely the parameter complexity of the related latent class model, and bounding the rank is sufficient to bound parameter complexity.

The results in this section make some additional assumptions about the support of the log-linear model. As a basis for comparison across the different cases, we will consider the order of the PARAFAC rank as a function of p or d under different scenarios for the support of the log-linear model. This provides a rough indication of the extent of dimension reduction that is achievable with PARAFAC decompositions in different cases.

Corollary 2.3.5 shows that when the maximum number of interacting levels of all variables is small relative to d the rank will be substantially reduced.

Corollary 2.3.5. *If* $| C_\theta^{(j)} | < \eta - 1$ *for all* j, $rnk_P^+(\pi) < \eta^{p-1}$.

Proof. This follows immediately from Theorem 2.3.1 by noting that the condition $| C_\theta^{(j)} | < \eta - 1$ implies that $| B_{\sigma(j)} | < \eta - 1$ for every permutation σ and every j. $\square$

In the case where $\eta \ll d$, the condition in Corollary 2.3.5 reduces the PARAFAC rank by a factor of $(d/\eta)^{p-1}$. However, the PARAFAC rank is still exponential in p, so this assumption is unhelpful in controlling the PARAFAC rank as a function of p. By Theorem 2.3.2, the exact rank is also exponential in p, so in general the order of the exact PARAFAC rank as a function of p is the same as that given by Corollary 2.3.5, which relies on Theorem 2.3.1.

If we also assume that certain types of conditional independence exist, useful bounds on the PARAFAC rank as a function of both d and p can be obtained. Corollary 2.3.6 gives one such result.

Corollary 2.3.6. *Suppose that the conditions in Corollary 2.3.5 hold and for $J \subset V$, set $y_{(J)} = \{y_j : j \in J\}$. Then if $y_{(J^c)}$ are independent given the variables $y_{(J)}$, $rnk_P^+(\pi) \leq \eta^{|J|}$.*

The simplest such case is represented by the graphical log-linear model in Ex. 1 of Figure 2.1: a single star-graph, where y_7 is the hub variable.[4] More generally, if we consider the special case of graphical models, the setting in Corollary 2.3.6 has a graphical representation where all edges involve at least one of the variables in J. The PARAFAC rank is then exponential in $|J|$, not p. With $\eta \leq \log d$ and $|J| \leq \log p$, we obtain $\mathrm{rnk}_P^+(\pi) \leq (\log d)^{\log p}$, so the rank becomes at most exponential in $\log p$.

Similar bounds can be obtained when marginal independence exists, which is represented by the graphical model in Ex. 2 in Figure 2.1 and formalized for general weakly hierarchical models in Corollary 2.3.7.

[4] While we use graphical representations to simplify exposition, none of the results presented in this section require that the log-linear model is graphical; it is sufficient that it be weakly hierarchical.

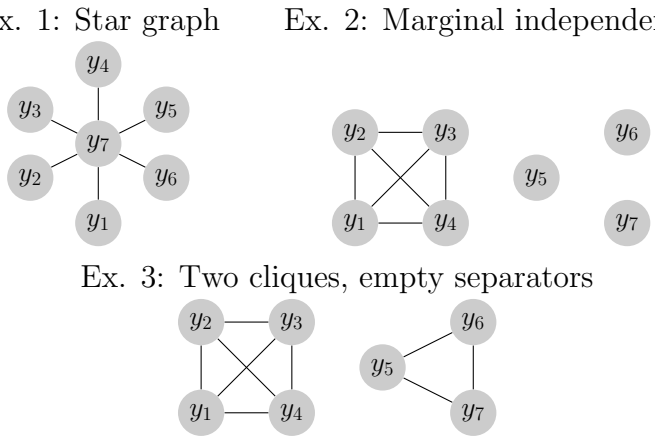


FIGURE 2.1*: Graphical representations of certain sparse log-linear models. Ex. 1 and Ex. 2 are graphs associated with sparse weakly hierarchical log-linear models that have low PARAFAC rank. The model need not be graphical for the rank to be low; any weakly hierarchical log-linear model with these dependence graphs will have low rank relative to the maximal rank. Ex. 1 is a canonical example of extensive conditional independence, which, by Corollary 2.3.6 leads to low PARAFAC rank. Ex. 2 has extensive marginal independence, as discussed in Corollary 2.3.7. Ex. 3 corresponds to a sparse log-linear model that has high PARAFAC rank (one half of the maximal rank).*

Corollary 2.3.7. *Suppose the conditions of Corollary 2.3.5 hold, and suppose there exists $J \subset V$ with the property that $j \in J^c \Rightarrow y_j \perp y_{[-j]}$. Then $rnk_P^+(\pi) \leq \eta^{(|J|)}$.*

Thus, in this case the PARAFAC rank will depend only on the number of variables that are not marginally independent; the same result that we obtained in Corollary 2.3.6 with conditional independence. It follows we can also achieve the $(\log d)^{\log p}$ order of the PARAFAC rank in p and d with the same assumptions on η and $|J|$.

The previous results in this section were corollaries to Theorem 2.3.1, which provides a relatively easy way to calculate bounds on the PARAFAC rank and allows us to clarify cases in which the PARAFAC rank of weakly hierarchical log-linear models will be small. However, this bound is not tight, as illustrated in Example 2.3.3, and thus when a specific weakly hierarchical interaction structure or class of structures is under consideration, it is necessary to utilize Theorem 2.3.2 to obtain a tight bound on the rank. We illustrate below through a concrete example that the conclusion of Theorem 2.3.2 is not simply of theoretical importance, the posterior

distribution on the number of components indeed increasingly concentrates on the upper bound implied by Theorem 2.3.2 as sample size increases.

Example 2.3.8. Set $p = 5$ and $d_j = d = 5$, so that we have a $5^5 = 3125$ cell tensor. Let $\boldsymbol{n} \sim \text{Multinomial}(N, \pi_0)$, where π_0 corresponds to the weakly hierarchical log-linear model with all main effects nonzero and

$$\theta_{\{1,2\}}(2, c_2) \neq 0 \text{ for all } c_2 \geq 2, \qquad \theta_{\{1,2\}}(c_1, 2) \neq 0 \text{ for all } c_1 \geq 2,$$

$$\theta_{\{1,3\}}(2, c_3) \neq 0 \text{ for all } c_3 \geq 2, \qquad \theta_{\{2,3\}}(2, c_3) \neq 0 \text{ for all } c_3 \geq 2,$$

$$\theta_{\{1,2,3\}}(2, 2, c_3) \neq 0 \text{ for all } c_3 \geq 2,$$

with all other interaction terms identically zero and $\theta_{\{\emptyset\}} = 0$ for identification. It can be verified that the minimal H for this model is $\{\{2\}, \{2\}, \emptyset, \emptyset, \emptyset\}$, so the PARAFAC rank is at most 4.

A simulation study was performed to assess performance of the Bayes PARAFAC model when the data are generated by the sparse weakly hierarchical log-linear model in Example 2.3.8. The nonzero entries of $\boldsymbol{\theta}$ were sampled from $N(0, 1)$, truncated to lie in the set $(-\infty, -0.2] \cup [0.2, \infty)$. The sampling of the $\boldsymbol{\theta}$ parameters was repeated ten times, and for each sample of the log-linear model parameters, $\boldsymbol{n}$ was sampled independently for $N = 1000, 5000$, and $10,000$ – sample sizes that range from about one third of the number of cells in the table to about three times the number of cells. We then performed MCMC computation for the Bayes PARAFAC model using the Gibbs sampling algorithm in Dunson and Xing (2009). For comparison, we also fit a regularized log-linear model using Lasso with ten-fold cross-validation to select the penalty, as implemented in the `glmnet` package for `R`, and the oracle model – i.e. a log-linear model for only the nonzero entries of $\boldsymbol{\theta}$ – by maximum likelihood. These comparison methods are used in all subsequent simulation examples.

Figure 2.2 shows, on the left, a boxplot of the cumulative sum for the largest ten class probabilities (for the class probabilities in descending order of magnitude). The

first five class probabilities nearly sum to one in every simulation, with the first four summing to at least 0.95 in each case. Thus, the posterior for the PARAFAC rank concentrates around the theoretical rank of 4. Figure 2.3 summarizes performance in estimation of $\boldsymbol{\theta}$ and π. Specifically, in this and all subsequent simulation examples, we use the samples of the PARAFAC parameters to obtain samples of π and of $\boldsymbol{\theta}$ – the latter by way of the Möbius transformation (see Massam et al. (2009)) – then use the ergodic average and median as point estimates for $\boldsymbol{\theta}$ and π, respectively. Normalized root mean squared error ($\text{RMSE}(\hat{\boldsymbol{\theta}})/\text{sd}(\boldsymbol{\theta})$) for estimation of $\boldsymbol{\theta}$, as well as the L_1 loss for estimation of π, are shown in Figure 2.3. Here, $\text{sd}(\boldsymbol{\theta})$ is the true standard deviation of the entries of $\boldsymbol{\theta}$ in the simulations. Also shown for comparison are the identical quantities for the Lasso estimator and the oracle MLE. PARAFAC is seen to perform competitively with Lasso for estimating $\boldsymbol{\theta}$ and is superior for estimation of π, despite the fact that generating data from a sparse log-linear model seemingly favors Lasso, which also benefits from cross-validation. There are clear problems with identification for the oracle estimator in the smaller sample sizes resulting from sparsity of the sampled table.

2.4 Collapsed Tucker decompositions

Corollaries 2.3.6 and 2.3.7 demonstrate the main ways in which exponential scaling of the PARAFAC rank in p can be avoided. However, these settings correspond to special cases of conditional independence mediated by a few variables or extensive marginal independence. More generally, Theorem 2.3.2 shows that low PARAFAC rank requires that all of the interactions can be accounted for by a small number of levels of the variables, as is the case in Example 2.3.8. Outside this relatively limited class, PARAFAC rank, and therefore parameter complexity, scales unfavorably in the dimension of the contingency table. As such, statistical efficiency relative to the log-linear model is expected to degrade as dimensions increase. This is likely most

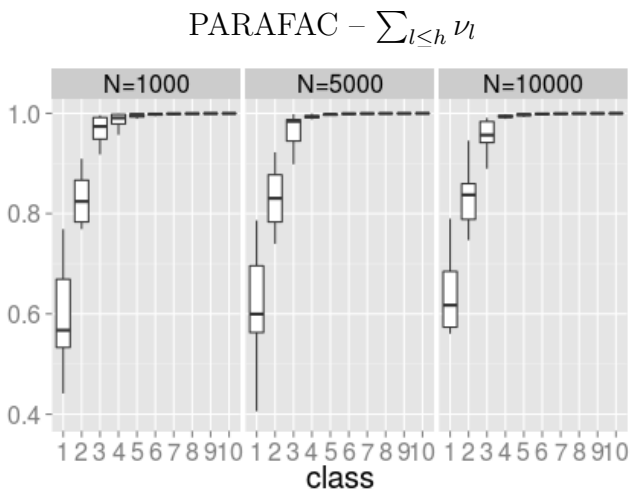


FIGURE 2.2*: Boxplot of posterior mean of cumulative sum of largest h class probabilities for $h = 1, \ldots, 10$ from PARAFAC model estimated on data generated from ten replicate simulations from the log-linear model in Example 2.3.8. The boxes within each panel are the posterior means for $\sum_{l\leq h}\nu_l$ for $h = 1, \ldots, 10$ and the different panels represent sample sizes $N = 1000$ (left), $N = 5000$ (center) and $N = 10000$ (right).*

evident in poor recovery of log-linear model parameters, as there may exist low rank expansions that well-approximate π but have quite different values of $\boldsymbol{\theta}$. We show several simulation examples in the sequel in which this degradation of statistical performance of the PARAFAC model occurs, particularly for estimation of $\boldsymbol{\theta}$.

As p grows, the number of classes in the PARAFAC model must grow rapidly to represent complex dependence among the variables. The Tucker decomposition, on the other hand, has p latent class variables and thus the number of latent classes does not depend on p at all, as shown in the following corollary to Theorem 2.3.2.

Corollary 2.4.1. *If π is a probability tensor corresponding to a sparse log-linear model then the Tucker rank*

$$rnk_T^+(\pi) \leq \bigwedge_{H\in\mathscr{H}} \bigvee_{j\in V} (|H_j|+1),$$

where $\mathscr{H}$ is the collection defined in the statement of Theorem 2.3.2.

The parsimony gained in the Tucker model by requiring few latent classes is offset to varying degrees by the need to model the dependence between the p latent

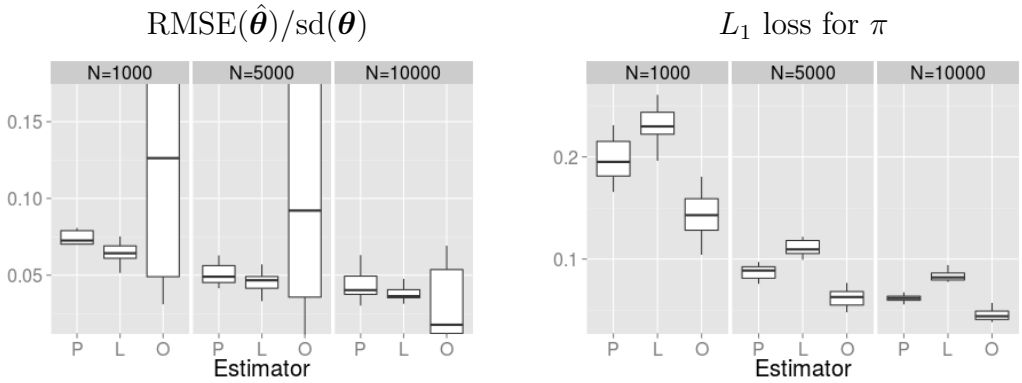


FIGURE 2.3*: Left figure: Boxplot of RMSE($\hat{\boldsymbol{\theta}}$)/sd($\boldsymbol{\theta}$) for PARAFAC (P), lasso (L), and oracle MLE (O) estimated on data generated from ten replicate simulations from the sparse log-linear model in Example 2.3.8. The three sub-panels of the figure show results for three different sample sizes $N = 1000, 5000, 10,000$. Right figure: identical arrangement, but here the plotted values are the L_1 loss for estimation of π.*

categorical variables through the $[\text{rnk}_T^+(\pi)]^p$ core tensor – the parameter ϕ in (2.7). Clearly, unless $\text{rnk}_T^+(\pi) \ll \max_j d_j$, the core is nearly as large as π. Therefore, while PARAFAC rank is an appropriate measure of parameter complexity in single latent class models, the Tucker rank is less meaningful unless d_j is large for most j. When p is even modest in size, parsimony and effective number of parameters in a Tucker model is mainly a function of *how the core is parametrized*. As a result, it becomes critical to count parameters in hierarchical models that induce Tucker decompositions of π rather than simply relying on the rank. For example, Bhattacharya and Dunson (2012) used a hierarchical random effects model to borrow information across the entries in the core tensor, greatly reducing parameter complexity relative to having an unstructured prior on the entries of ϕ.

In what follows, we motivate and develop a meta-family of tensor decompositions obtained by allowing the dimension of the core tensor to be any value between 1 (the PARAFAC) and p (the Tucker). We refer to these as collapsed Tucker (c-Tucker) decompositions. These decompositions can be induced by hierarchical latent class models where the number of latent class variables is between 1 and p. To control

parameter complexity, we choose to model the core through a latent PARAFAC decomposition. This is a modeling choice, and is not required to induce a c-Tucker decomposition. For example, one could instead choose an analogue of the random effects model of Bhattacharya and Dunson (2012) to model the core. To illustrate the advantages of c-Tucker factorizations, we focus on data generated from sparse log-linear models with groups of variables in which there is arbitrary dependence for variables within a group but independence or structured dependence across groups.

2.4.1 Independent PARAFACs

To motivate the c-Tucker decomposition, we first show how a variation of the PARAFAC decomposition can eliminate the exponential factor of $\log(p)$ that appears in Corollary 2.3.7 in cases where there are multiple groups of variables that are marginally independent of all the other groups. An example of a graphical model with this dependence structure is shown in Ex. 3 in Figure 2.1: two cliques with empty separators.

Divide $y_1, \ldots, y_p$ into k groups, and let s_j indicate the group membership of variable j. For each $s \in \{1, \ldots, k\}$ define a PARAFAC expansion for the marginal probability tensor corresponding to $\pi^{(s)} = Pr(\{y_j : s_j = s\})$, as

$$\pi^{(s)} = \sum_{h=1}^{m_s} \nu_{sh} \bigotimes_{j:s_j=s} \lambda_h^{(j)}.$$

We define the joint distribution of $y_1, \ldots, y_p$ as

$$\pi_{c_1,\ldots,c_p} = \prod_{s=1}^{k} \prod_{j:s_j=s} \pi_{c_j}^{(s)}.$$

This model can be described succinctly as k independent PARAFACs. This is a generalization of the sparse PARAFAC (sp-PARAFAC) model of Zhou et al. (2013) to the case of more than two groups, and gives much stronger control over parameter

growth than PARAFAC when the truth consists of marginally independent groups of variables. This is shown formally for the special case of graphical models with empty separators in Theorem 2.4.2.

Theorem 2.4.2. *Consider a graphical log-linear model for binary data defined by parameters* $\boldsymbol{\theta}$*. Let* $\mathscr{F}$ *be the collection of all cliques, and suppose* $|\ \mathscr{F}\ | = \mathcal{O}(k)$*. Then if* $\bigvee_{F\in\mathscr{F}}|F| = \mathcal{O}(\log_2(p))$ *and all separators are empty, the tensor* π *can be expressed by* k *independent tensors* $\pi^{(1)}, \ldots, \pi^{(k)}$ *with* $\sum_{s=1}^{k} rnk_P^+(\pi^{(s)}) = \mathcal{O}(kp)$*.*

Remark 2.4.1. In the special case where $\log_2(p)$ is an integer and all cliques have identical cardinality, we obtain $\sum_{s=1}^{k} \mathrm{rnk}_P^+(\pi^{(s)}) = \mathcal{O}(p^2/\log_2(p))$.

Remark 2.4.2. The result in Theorem 2.4.2 also holds for any weakly hierarchical log-linear model with the same dependence structure, since the graphical model has the maximum number of nonzero interaction terms for any set of dependence/independence relationships.

It follows that where marginally independent sets of variables exist, grouping variables and performing independent PARAFAC decompositions for each of the marginal probability tensors corresponding to the groups can reduce the effective number of parameters drastically. Although Theorem 2.4.2 is stated for the special case of binary outcomes, conceptually it applies for general d_j and the advantage is borne out empirically, as we show with the following example.

Example 2.4.3. Let π be a d^5 probability tensor corresponding to a sparse log-linear model where all main effects are nonzero and in addition

$$\theta_{\{1,2\}}(2, c_2) \neq 0 \text{ for } c_2 \geq 2 \qquad \theta_{\{3,4\}}(c_3, 2) \neq 0 \text{ for } c_3 \geq 2$$

$$\theta_{\{1,2\}}(c_1, 2) \neq 0 \text{ for } c_1 \geq 2 \qquad \theta_{\{3,5\}}(c_3, 2) \neq 0 \text{ for } c_3 \geq 2$$

$$\theta_{\{3,4\}}(2, c_4) \neq 0 \text{ for } c_4 \geq 2 \qquad \theta_{\{4,5\}}(2, c_5) \neq 0 \text{ for } c_4 \geq 2$$

$$\theta_{\{3,5\}}(2, c_5) \neq 0 \text{ for } c_5 \geq 2 \qquad \theta_{\{4,5\}}(c_4, 2) \neq 0 \text{ for } c_4 \geq 2,$$

with all other interaction terms equal to zero. Letting $H = \{\{2\}, \{2\}, \{2\}, \{2\}, \{2\}\} \in \mathscr{H}$, we know $\text{rnk}_P^+(\pi) \leq 2^5 = 32$, so the PARAFAC decomposition has approximately $31 + 32 \times 20 = 674$ parameters. The structure of sparsity guarantees that $y_1, y_2 \perp y_3, y_4, y_5$. As a result, the number of parameters in two independent PARAFAC decompositions is only $(3 + 4 \times 8) + (7 + 8 \times 12) = 138$.

We simulated data from the model in Example 2.4.3 with $d = 5$ using the same distribution for the nonzero log-linear model parameters as in the simulation study for Example 2.3.8. We performed computation by MCMC for the PARAFAC model as well as the independent PARAFAC model with two variable groups: y_1, y_2 and y_3, y_4, y_5. Figure 2.4 shows normalized RMSE for estimation of $\boldsymbol{\theta}$ and L_1 loss for estimation of π for PARAFAC, independent PARAFAC, Lasso, and the oracle MLE. PARAFAC performs poorly relative to Lasso in estimation of $\boldsymbol{\theta}$ but is comparable to Lasso for estimation of π, suggesting that the posterior concentrates around a lower-rank tensor with entries that are very similar to π but for which the equivalent log-linear model has a rather different value of $\boldsymbol{\theta}$. This is probably a consequence of the fact that the true PARAFAC rank in Example 2.4.3 is much larger than the PARAFAC rank in Example 2.3.8, so that the exact expansion has high parameter complexity. In contrast, the independent PARAFAC performs slightly better than Lasso for estimation of $\boldsymbol{\theta}$ and substantially better for estimation of π, despite the fact that the data generating model is a sparse log-linear model.

The approach outlined above is limited to cases in which the variable groups are marginally independent, which in the special case of graphical models corresponds to empty separators. However, additional flexibility can be gained by introducing another set of parameters to control dependence between the groups. This is the essence of the collapsed Tucker model, where we project p dimensional $\boldsymbol{y}$ to $k \ll p$

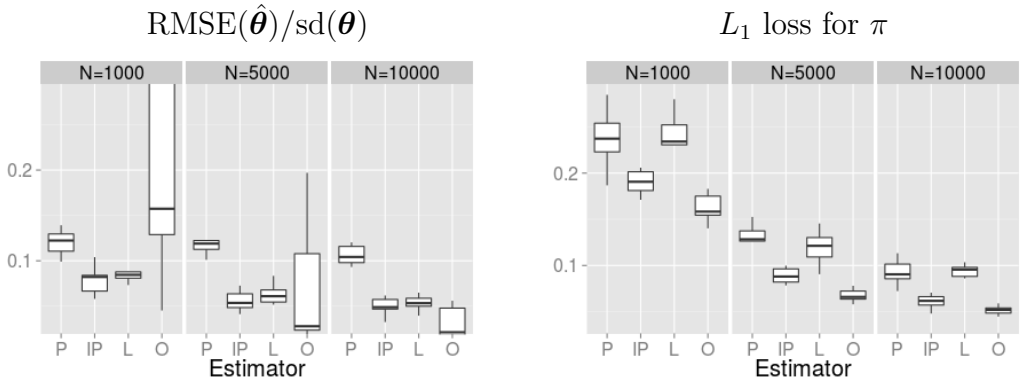


FIGURE 2.4*: Left figure: Boxplot of RMSE($\hat{\boldsymbol{\theta}}$)/sd($\boldsymbol{\theta}$) for PARAFAC (P), independent PARAFAC (IP), lasso (L), and oracle MLE (O) estimated on data generated from ten replicate simulations from the log-linear model in Example 2.4.3. The three sub-panels of the figure show results for three different sample sizes $N = 1000, 5000, 10,000$. Right figure: identical arrangement, but here the plotted values are the L_1 loss for estimation of π*

dimensional $\boldsymbol{z}$ and model the joint p.m.f. of $\boldsymbol{z}$ via a PARAFAC.

2.4.2 Latent class models inducing collapsed Tucker decompositions

We now define c-Tucker decompositions. Specifically, let

$$\pi_{c_1 \ldots c_p} = \sum_{h_1=1}^{m} \cdots \sum_{h_k=1}^{m} \phi_{h_1 \ldots h_k} \prod_{j=1}^{p} \lambda^{(j)}_{h^*_j c_j}, \tag{2.19}$$

where $h^*_j = h_{s_j}$ with $s_j \in \{1, \ldots, k\}$ for $j = 1, \ldots, p$ and $k \ll p$ when p is moderate to large. The s_j's are group indices for $\{y_j : j \in V\}$, with $s_j = \rho$ denoting that y_j is allocated to group ρ. For a particular configuration of the s_j's, the p variables are assigned to k groups, and $s_j = s_{j'}$ indicates that y_j and $y_{j'}$ belong to the same group. We refer to (2.19) as a m-component collapsed Tucker (c-Tucker) factorization.

c-Tucker is a latent class model with k latent class indices. Letting $\boldsymbol{z} = (z_1, \ldots, z_k)^{\mathrm{T}}$ denote a vector of group indices, the c-Tucker model in (2.19) has a hierarchical representation where given $\boldsymbol{z}$, $y_1, \ldots, y_p$ are conditionally independent with $\Pr(y_j = c_j \mid \boldsymbol{z}, s_j) = \lambda^{(j)}_{z_{s_j} c_j}$. The parameter ϕ is a m^k nonnegative core tensor; it is a probability tensor that parametrizes the joint distribution of the latent categorical variables $z_1, \ldots, z_k$. Clearly, for $k = 1$ we recover the PARAFAC decomposition and for $k = p$

we obtain the Tucker decomposition. Graphical representations of dependence between observed and latent variables in PARAFAC, Tucker, and c-Tucker models are shown in Figure A.1 in the Appendix.

The number of parameters in the core ϕ grows exponentially in k, so superficially the problem of rapidly growing parameter complexity remains. To control this, we model ϕ using a PARAFAC decomposition

$$\phi_{h_1 \ldots h_k} = \sum_{l=1}^{r} \xi_l \prod_{s=1}^{k} \psi_{lh_s}^{(s)}, \tag{2.20}$$

where $\xi = \{\xi_l\}$ is a vector of probabilities, $\psi_l^{(s)} = \{\psi_{lh}^{(s)}\}$ are probability vectors of dimension m for $s = \{1, \ldots, k\}$, and $1 < k < p$. If $r = 1$ we obtain a k-group independent PARAFAC model as in section 2.4.1. Under (2.20), the number of free parameters[5] in a c-Tucker expansion scales as

$$r - 1 + r(m-1)k + m \sum_{j=1}^{p} (d_j - 1). \tag{2.21}$$

This effective parameter count depends not only on the number of components m but also on r and k, suggesting that unlike in the PARAFAC case the *rank is not useful by itself* as a measure of parsimony. Hence, we focus on parameter count (2.21) and rank of the core r instead of m in what follows. The first two terms in (2.21) are specific to the choice of a PARAFAC factorization for the core ϕ, while the term $m \sum_{j=1}^{p} (d_j - 1)$ appears in the parameter count for any c-Tucker factorization.

We can obtain insight into what types of log-linear models might be parsimonious in the c-Tucker representation but not the PARAFAC representation by considering the setup in Theorem 2.4.2: binary variables consisting of k independent groups

[5] These are upper bounds rather than exact expressions. That the effective dimension of the parameter space for a PARAFAC model can in some cases be smaller than the nominal number of parameters in the expansion has been well documented, see e.g. Geiger et al. (2001) and Fienberg et al. (2007).

each with at most $\log_2(p)$ members. In general, if k marginally independent groups of variables exist and all outcomes are binary, the PARAFAC rank will be of the order 2^{p-k}. The proof of this is straightforward and is omitted. Therefore, Theorem 2.4.2 gives conditions under which the ordinary PARAFAC rank is approximately 2^{p-k}, with parameter complexity $2^{p-k} - 1 + p2^{p-k}$. Under the same conditions, c-Tucker has parameter complexity of approximately $kp - k + p^2$, obtained from (2.21) with $r = 1$, $m = p$, and $d_j = 2$ for all j.[6] This is quadratic in p instead of exponential.

2.5 Estimation and applications for c-Tucker models

We present an algorithm for inference and computation for c-Tucker models in the Bayesian paradigm. The model is illustrated in simulation studies and an application to the functional disability data from the national long term care survey (NLTCS).

2.5.1 Bayesian inference for c-Tucker models

Bayesian inference for c-Tucker models requires priors on the parameters of the core, arms, and the group memberships of the variables. We choose conjugate Dirichlet priors on the arms $\lambda^{(j)}_{h^*_j c_j}$. We specify truncated stick-breaking priors Ishwaran and James (2001) on the latent class probabilities $Pr(z_{is} = h)$ and fix the maximum number of latent classes. A similar approach is used for the arms $\{\zeta^{(s)}_h\}$ in the PARAFAC expansion of the core. When group memberships are inferred, we use a Dirichlet$(1/k, \ldots, 1/k)$ prior on variable group membership probabilities.

The Bayesian c-Tucker model can be expressed in hierarchical form as

$$y_{ij} \mid z_{i1}, \ldots, z_{ik}, \boldsymbol{\lambda}^{(j)} \sim \text{Multi}(\{1, \ldots, d_j\}, \lambda^{(j)}_{z_{ih_{s_j}}1}, \ldots, \lambda^{(j)}_{z_{ih_{s_j}}d_j}),$$

$$\boldsymbol{\lambda}^{(j)}_h \sim \text{Diri}(a_{h1}, \ldots, a_{hd_j}),$$

$$z_{is} \mid w_i, \boldsymbol{\psi}^{(s)} \sim \text{Multi}(\{1, \ldots, m\}, \psi^{(s)}_{w_i 1}, \ldots, \psi^{(s)}_{w_i m})$$

[6] the difference between this expression and that in Theorem 2.4.2 is simply a result of the latter being the sum of PARAFAC ranks and the former a count of free parameters.

$$\text{pr}(w_i = l) = \nu_l^* \prod_{t<l}(1-\nu_t^*),\ \nu_l^* \sim \text{beta}(1,\beta)$$

$$\psi_{lh}^{(s)} = \zeta_{lh}^{(s)} \prod_{h'<h}(1-\zeta_{lh'}^{(s)}),\ \zeta_{lh}^{(s)} \sim \text{beta}(1,\delta_s)$$

$$s_1,\ldots,s_p \sim \text{Multi}(\{1,\ldots,k\},\xi_1,\ldots,\xi_k)$$

$$\xi \sim \text{Dirichlet}(1/k,\ldots,1/k),$$

where the index $i = 1,\ldots,n$ is a scalar subject index – not the multi-index $\boldsymbol{i}$ of a cell of the corresponding contingency table – and y_i is a p-vector of categorical observations for the ith subject. Bayesian computation for this model can be performed using a straightforward Gibbs sampler. Details of the computation are given in the supplement Johndrow et al. (2014a).

2.5.2 Simulation studies and application for c-Tucker model

We revisit Example 2.4.3 to illustrate the performance of the c-Tucker model in the case of marginally independent variable groups. Using data from the simulation procedure in section 2.4.1, we performed computation for the c-Tucker model by MCMC using the algorithm in Johndrow et al. (2014a), first by fixing two variable groups (y_1, y_2 and y_3, y_4, y_5) and letting the algorithm learn the rank of the core, and then by setting the number of groups to be two and allowing the algorithm to learn both the groups and the rank of the core. In the latter case, the group membership was initialized by performing agglomerative clustering using one minus the pairwise Cramér's V statistic as a dissimilarity matrix for the variables.

Figure 2.5 shows boxplots of $\sum_{l\leq h}\nu_l$ (for ν in descending magnitude order) for the PARAFAC and c-Tucker model with fixed groups. When $N = 1000$, the PARAFAC has approximate posterior rank five – judged by counting the minimal number of classes such that the cumulative class probability is at least 0.99 – as does the c-Tucker core tensor, ϕ. However, as the sample size increases, the approximate

PARAFAC rank grows, whereas the rank of the c-Tucker core ϕ decreases. With $N = 10,000$, the approximate rank of the c-Tucker core decreases to three, with most of the weight on the largest class, whereas the approximate PARAFAC rank increases to seven, and the weight on the largest class decreases. Recalling that the PARAFAC rank in this example is 32, while the rank of the c-Tucker core is one, this result is consistent with the ranks converging toward their true values as the sample size grows.

Figure 2.6 shows performance of PARAFAC, independent PARAFAC, c-Tucker with fixed groups, and c-Tucker with learned groups in estimation of $\boldsymbol{\theta}$ and π (the results for PARAFAC and independent PARAFAC are identical to those in figure 2.4 but are shown for ease of comparison). The performance of c-Tucker is seen to be virtually identical to that of independent PARAFAC at each sample size, regardless of whether groups are fixed or learned, showing that the enhanced flexibility of c-Tucker need not result in loss of performance when the truth is exactly an independent PARAFAC. Recalling that independent PARAFAC is superior to Lasso in this simulation on these loss functions, this indicates better performance for c-Tucker as well. PARAFAC performs poorly relative to methods that incorporate variable grouping, which is as expected for the reasons described in section 2.4.1. The superior performance of c-Tucker is consistent with the theoretical results in sections 2.3 and 2.4. In this example, the effective posterior parameter complexity in the PARAFAC and c-Tucker models – computed using (2.21) – is roughly equivalent, as shown in Figure S.2 in Johndrow et al. (2014a). Thus, c-Tucker provides lower estimation error with similar parameter complexity.

A final simulation illustrates the c-Tucker model in the more challenging case when there are no marginally independent groups of variables, based on Example 2.3.4. The nonzero entries of $\boldsymbol{\theta}$ were sampled as described in section 2.4.1, and ten replicates of each simulation were performed for sample sizes $N = 1000, 5000$, and

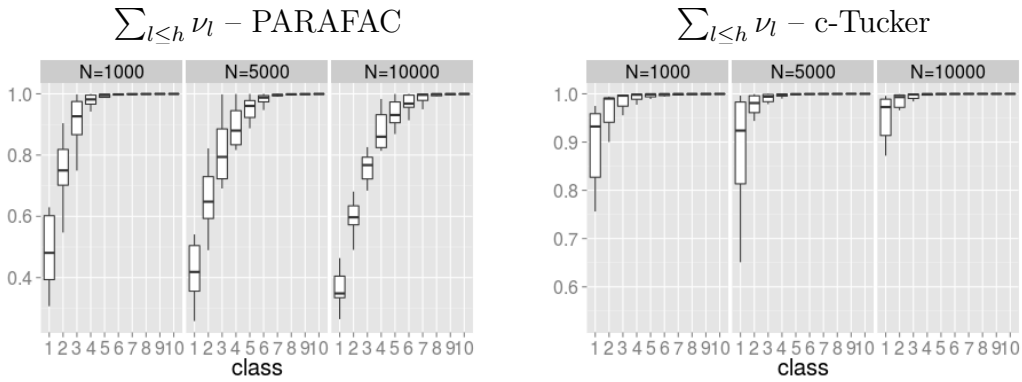


FIGURE 2.5*: Left figure: Boxplots of posterior mean of* $\sum_{l \le h} \nu_l$ *for* $h = 1, \ldots, 10$ *for the PARAFAC model estimated on data from over ten replicate simulations from the log-linear model in Example 2.4.3. The boxes within each panel are the posterior means of* $\sum_{l \le h} \nu_l$ *for* $h = 1, \ldots, 10$*, and the three panels correspond to sample sizes* $N = 1000$ *(left),* $N = 5000$ *(center), and* $N = 10,000$ *(right). Right figure: the same posterior summary shown for the collapsed Tucker model with fixed groups; here* ν_l *are the component weights in the PARAFAC expansion of the core tensor* ϕ*.*

$10,000$. We perform computation by MCMC for both PARAFAC and c-Tucker, and in the latter, we set the number of variable groups to three, allowing learning of the group memberships. Normalized RMSE for estimation of $\boldsymbol{\theta}$ and L_1 loss for estimation of π are shown in Figure 2.7. c-Tucker outperforms PARAFAC with respect to MSE for estimating $\boldsymbol{\theta}$, while showing similar performance for estimation of π. Lasso is superior for estimation of $\boldsymbol{\theta}$, but similar to PARAFAC and c-Tucker for estimation of π. That PARAFAC performs similarly to Lasso on either metric is surprising given that π corresponds to a sparse log-linear model (only 57 nonzero parameters of 3125), whereas the PARAFAC rank is relatively high (a rank of 32, corresponding to 671 free parameters).[7] This is consistent with the result for example 2.3.8 and merits a similar interpretation.

We apply the c-Tucker model with learned groups to analysis of functional disability data from the national long term care survey (NLTCS). The data take the form of a 2^{16} contingency table, and are extensively described in Dobra and Lenkoski

[7] The effective dimension may be smaller than this – see Geiger et al. (2001).

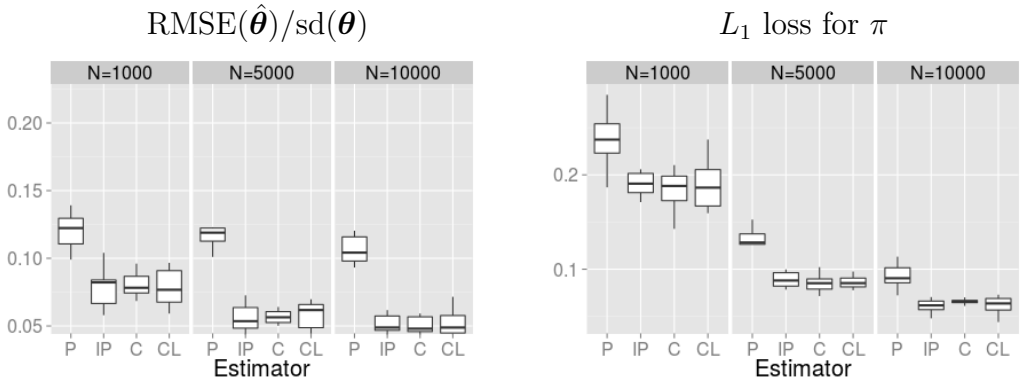


FIGURE 2.6: *Left figure: Boxplot of RMSE($\hat{\boldsymbol{\theta}}$)/sd($\boldsymbol{\theta}$) for PARAFAC (P), independent PARAFAC (IP), c-Tucker with fixed groups (C), and c-Tucker with learned groups (CL) estimated on data from ten replicate simulations from the log-linear model in Example 2.4.3. The three sub-panels of the figure show results for three different sample sizes* $N = 1000, 5000, 10,000$. *Right figure: identical arrangement, but here the plotted values are the* L_1 *loss for estimation of* π

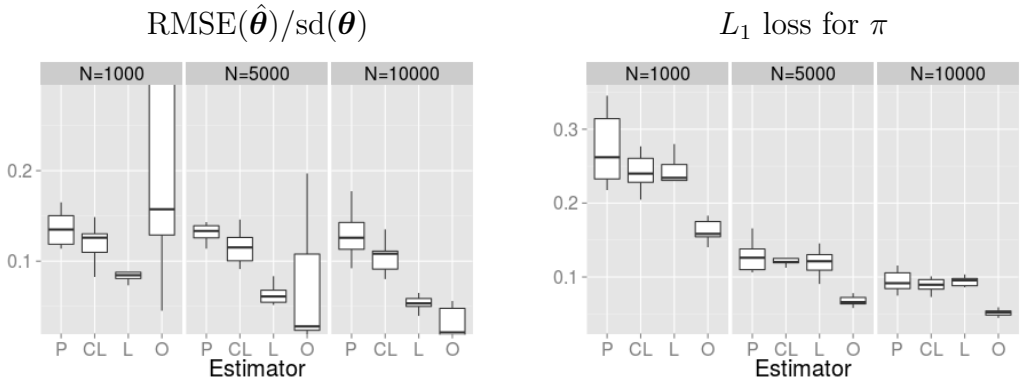


FIGURE 2.7: *Left figure: Boxplot of RMSE($\hat{\boldsymbol{\theta}}$)/sd($\boldsymbol{\theta}$) for PARAFAC (P), c-Tucker with learned groups (CL), Lasso (L), and oracle MLE (O) estiamted on data from ten replicate simulations from the log-linear model in Example 2.3.4. The three sub-panels of the figure show results for three different sample sizes* $N = 1000, 5000, 10,000$. *Right figure: identical arrangement, but here the plotted values are the* L_1 *loss for estimation of* π

(2011), who applied a novel copula Gaussian graphical model. Their model is extremely flexible while favoring parsimony, but has the primary disadvantage of being highly computationally intensive, lacking scalability beyond relatively small tables. Our interest here is in assessing whether the much more computationally efficient c-Tucker model can perform comparably to the Dobra and Lenkoski (2011) approach for these data. We performed posterior computation using the MCMC algorithm

described in Johndrow et al. (2014a). Table S.1 in Johndrow et al. (2014a) shows the posterior means of pairwise Cramér's V and $Pr(H_{1,\rho}|\boldsymbol{y})$, where $H_{1,\rho} = \mathbb{1}(\rho > 0.1)$ and ρ is the pairwise Cramér's V. For comparison, we reproduce the same results based on posterior samples for the copula Gaussian graphical model from Dobra and Lenkoski (2011) in Table S.2 in Johndrow et al. (2014a). Our results demonstrate close agreement with Dobra and Lenkoski (2011).

2.6 Conclusion

The relationship between the sparsity of a log-linear model and the rank of the associated probability tensor derived here makes clear that a large class of very sparse log-linear models nonetheless has high PARAFAC tensor rank. The statistical consequence of this result is that estimation performance for single latent class models for the joint distribution of multivariate categorical data will tend to degrade as the number of variables grows large, unless dependence in the true model can be accounted for by a small number of levels of the variables, as is the case when marginal independence or highly structured conditional independence exists.

This motivates development of more flexible tensor factorizations that can parsimoniously characterize a broader class of interactions in multivariate categorical data. Tucker factorizations are promising in this regard, and we obtain theory on parameter complexity of Tucker factorizations of sparse log-linear models. These results lead naturally to a novel meta-class of tensor decompositions we refer to as collapsed Tucker. These decompositions are considerably more flexible than either Tucker or PARAFAC, and are highly promising in broad applications. We illustrate some of this promise in simulation examples and an application to real data showing similar results to those obtained with sophisticated graphical modeling methods, which are much more computationally intensive. In fact, computational algorithms for estimation in tensor factorization models in the classes we consider are vastly more

scalable to high-dimensional data than algorithms for estimation of sparse log-linear models, so the theoretical results and methods developed here are of substantial practical consequence for high-dimensional statistics with discrete data.

3

Rank Identifiability of Mixture Models

3.1 Introduction

In many areas of biology and psychology, estimation of the number of latent subpopulations of subjects is an important inferential goal. For example, in population genetics, there is often interest in estimating the number of genetically distinct subpopulations of an organism. This can allow researchers to test hypotheses about the evolutionary history of the organism, and is important in assessing the conservation status of extant populations. Thus, whether the number of latent classes of certain mixture models can be reliably estimated is of fundamental importance in these application areas.

Arguably the most common methods used in population genetics for estimating genetically distinct populations were proposed by Pritchard et al. (2000) and implemted in the popular `STUCTURE` software package. Two basic models were proposed, one for situations in which the subpopulations are thought to be *admixed*, and one for populations without admixture. The term admixture refers to the setting in which individuals may have mixed class memberships, or, in biological terms, breed-

ing takes place between the genetic subpopulations. Here, we focus on the model for non-admixed populations. The motivation for this choice is that estimating the number of subpopulations is necessarily easier without admixture, and thus if reliable estimation is not possible in this context, it is unlikely to be possible in the presence of admixture.

A slight variation on the model of Pritchard et al. (2000) without admixture is given by

$$\mathbb{P}\left[y_{i1} = c_1, \ldots, y_{ip} = c_p \mid z_i = h\right] = \prod_{j=1}^{p} \lambda_h^{(j)} \tag{3.1}$$

$$\lambda_h^{(j)} \sim \text{Dirichlet}(a_1^{(j)}, \ldots, a_{d_j}^{(j)}), \quad \mathbb{P}\left[z_i = h\right] = \nu_h, \quad \nu_h \sim \text{Dirichlet}(\alpha, \ldots, \alpha) \tag{3.2}$$

Here, there are assumed to be K subpopulations, $y_{i1}, \ldots, y_{ip}$ are observed alleles at p loci in the diploid genome of individual i, and $\lambda_h^{(j)}$ is the population-level allele frequency at locus j in subpopulation h. One difference between the model in (3.10)-(3.11) and that of Pritchard et al. (2000) is that we do not treat the mixture weights as known. This certainly makes sense, since in general we do not know the weights *a priori* and should attempt to learn them from the data. The other difference between the model in (3.10)-(3.11) and the original model is that the latter restricts the allele frequencies for both copies of each gene in diploid genomes to be equal. The choice in Pritchard et al. (2000) is clearly correct in the context of this application, and later when we apply this model for inference on population structure we do the same. However, the theory and simulation studies we present first apply to models of the form in (3.10)-(3.11) more generally, so to simplify exposition and make clear the broader relevance of the results in most of this paper, we keep with this general form of the model until the genetics application in Section 3.5.

The procedure suggested in Pritchard et al. (2000) for estimation of K is to perform computation for the model with different values of K via Markov chain

Monte Carlo (MCMC), then use these approximate posterior samples under fixed K to estimate the posterior evidence $p(K \mid y) \propto p(y \mid K)p(K)$. This procedure is rather *ad hoc* and lacks theoretical support, a point made more than once in Pritchard et al. (2000). As a result, the authors suggest using this procedure only to provide a rough guide to which values of K are most consistent with the observed data, and not for rigorous statistical hypothesis testing and inference. Nonetheless, researchers routinely used this procedure to obtain estimates of K until the publication of Evanno et al. (2005), which suggests an alternative procedure. That procedure computes a statistic, ΔK, from MCMC sample paths obtained for different values of K, and uses this statistic to make a point estimate of K. This procedure also lacks theory support and the justification was entirely based on simulation results. Moreover, it is purely a point estimation procedure and provides no quantification of uncertainty in the estimate of K. That Pritchard et al. (2000) and Evanno et al. (2005) have been cited 15,645 and 7,219 times, respectively, provides some measure of the level of scientific interest in this problem and the associated statistical methods.

Estimating the number of mixture components using Bayesian methods has been investigated in the more general context by a number of authors. The approach taken in Pritchard et al. (2000) and Evanno et al. (2005) is unusual in relying on parameter estimates obtained under a range of fixed values of K. In particular, this method assumes that the negative log-likelihood is well-approximated by a Gaussian distribution, which in the relatively small samples often encountered in genetics is unlikely to be the case. Since the procedure in Evanno et al. (2005) relies upon this approximation, the same concerns apply. A more common approach in the statistics literature is to use the posterior for the number of clusters in a Dirichlet process mixture (DPM) to make inferences about the true number of subpopulations (see Huelsenbeck and Andolfatto (2007), Medvedovic and Sivaganesan (2002), Otranto and Gallo (2002), Xing et al. (2006), Fearnhead (2004)). However, a others have

observed that when the data are generated by a finite mixture, the posterior for the number of components in a DPM tends to overestimate the true number of components (see West and Escobar (1993) and Onogi et al. (2011) for examples). More recently, Miller and Harrison (2014) showed that under very general conditions, when the truth is a finite mixture, the posterior under a DPM is not even consistent for the number of components.

An alternative to the Dirichlet process when the data are thought to originate from a finite mixture with an unknown number of components is a variable-dimension mixture. A Bayesian approach to variable-dimension mixtures is to simply place a prior on the number of components, an approach for which there are numerous precedents (see Nobile (1994), Richardson and Green (1997), Green and Richardson (2001), and Nobile and Fearnside (2007)). We refer to this approach as a mixture of finite mixtures (MFM). Miller (2014) shows that when the model satisfies a *mixture identifiability* condition that is weaker than the usual notion of identifiability, then the posterior under an MFM model is consistent for the number of components, assuming the data are *iid* from a finite mixture model. Therefore, when the number of mixture components is the target of inference, MFMs provide theoretical support for good performance, at least asymptotically.

The mixture identifiability condition used in Miller (2014) does not hold trivially, and thus must be checked on a case-by-case basis. One case in which this condition fails to hold for all K that is relevant to our setting is for the model in (3.10)-(3.11). Thus, the approach in Miller (2014) cannot be used directly to prove consistency for the number of subpopulations using a MFM prior for this model. The main theoretical contributions of this work are the following. First, we characterize the nature of the mixture identifiability problem for finite mixtures of the form in (3.10)-(3.11), and show by counterexample why the condition does not hold for every K. We then show a weaker condition that provides a first step to showing consistency for

the number of mixture components using MFM models in the general case, without truncating the prior on K. We then show how it is possible to check whether the condition holds for a problem size of interest, thus allowing verification of mixture identifiability, and therefore consistency for K, on a case-by-case basis.

Taken together, the theoretical work suggests that MFM is a more reliable alternative to the methods of Pritchard et al. (2000) and Evanno et al. (2005) for estimating K. We provide empirical evidence that MFM can recover the true value of K in simulation studies, while also providing the uncertainty quantification that is lacking in existing methods. We then apply MFM to a real dataset on redwing blackbird populations, and show that the posterior mode for K is the same as the value of K selected using the method of Evanno et al. (2005). However, the full posterior distribution for K indicates that there is substantial uncertainty in the actual value of K. We also provide a discussion of prior choice and some empirical studies of sensitivity of the posterior for K to the prior on the λ parameters.

3.2 Background

In this section we provide further development of the model in (3.10)-(3.11), then provide some necessary algebraic preliminaries for the result that follows.

3.2.1 Mixture models and nonnegative tensor decompositions

N observations on $y_1, \ldots, y_p$ in (3.10)-(3.11) can equivalently be represented as a *contingency table*, and it is in this context that we develop the theoretical results. Let $V = \{1, \ldots, p\}$ be an index set for p categorical variables $y_j, j \in V$. For each $j \in V$, $\text{spt}(y_j) = \{1, \ldots, d_j\}$, where $\text{spt}(y) = \{c : \mathbb{P}(y = c) > 0\}$. Let $\mathcal{I}_j = \{1, \ldots, d_j\}$ and $\mathcal{I}_V = \times_{j \in V} \mathcal{I}_j$. We will use $\boldsymbol{i}$ to represent a generic element of $\mathcal{I}_V$, and π to represent a $\prod_j d_j$ dimensional probability tensor – i.e. such that the entries of π are nonnegative and sum to one. For a sample from Multinomial(N, π), let $n(\boldsymbol{i})$ be the observed count

for cell $\boldsymbol{i}$, and $\boldsymbol{\eta}$ a random variable with distribution Multinomial(N, π). The model in (3.10)-(3.11) represents π in nonnegative rank K PARAFAC expansion, i.e.

$$\pi = \sum_{h=1}^{k} \nu_h \bigotimes_{j=1}^{p} \lambda_h^{(j)}, \tag{3.3}$$

where $\nu \in \Delta^{k-1}$ and $\lambda_h^{(j)} \in \Delta^{d_j - 1}$, and Δ^s is the s-dimensional simplex. Explicitly incorporating the constraints on $\lambda_h^{(j)}$ and ν, we can express (3.3) as

$$\begin{aligned}\pi_{c_1,\ldots,c_p} = \sum_{h=1}^{K-1} \nu_h \left(\prod_{j:c_j<d_j} \lambda_{hc_j}^{(j)} \right) \left(\prod_{j:c_j=d_j} \left(1 - \sum_{c<d_j} \lambda_{hc}^{(j)} \right) \right) \\ + \left(1 - \sum_{h<K} \nu_h \right) \left(\prod_{j:c_j<d_j} \lambda_{hc_j}^{(j)} \right) \left(\prod_{j:c_j=d_j} \left(1 - \sum_{c<d_j} \lambda_{hc}^{(j)} \right) \right),\end{aligned}$$

or in the equivalent multinomial logit parametrization,

$$\begin{aligned}\pi_{c_1,\ldots,c_p} = \sum_{h=1}^{K-1} \left(\frac{e^{\gamma_h}}{1 + \sum_{l<K} e^{\gamma_l}} \right) \left(\prod_{j:c_j<d_j} \frac{e^{\theta_{hc_j}^{(j)}}}{1 + \sum_{c<d_j} e^{\theta_{hc}^{(j)}}} \right) \left(\prod_{j:c_j=d_j} \frac{1}{1 + \sum_{c<d_j} e^{\theta_{hc}^{(j)}}} \right) \\ + \left(\frac{1}{1 + \sum_{l<K} e^{\gamma_l}} \right) \left(\prod_{j:c_j<d_j} \frac{e^{\theta_{hc_j}^{(j)}}}{1 + \sum_{c<d_j} e^{\theta_{hc}^{(j)}}} \right) \left(\prod_{j:c_j=d_j} \frac{1}{1 + \sum_{c<d_j} e^{\theta_{hc}^{(j)}}} \right),\end{aligned}$$

which is more convenient for computing derivatives. These expressions define transformations $g_1 : (\nu, \lambda) \to \pi$ and $g_2 : (\gamma, \theta) \to \pi$. It turns out that the mixture identifiability condition in Miller (2014) is equivalent to the requirement that the Jacobian matrix J_{g_1} of the transformation g_1 (or, equivalently, J_{g_2}) be full-rank. When the Jacobian for a K component expansion is full-rank, we say that the model has the *expected dimension*, which is given by the parameter count

$$C(K) = \left(K - 1 + K \sum_{j=1}^{p} (d_j - 1) \right) \wedge \left(\prod_j d_j - 1 \right). \tag{3.4}$$

The minimum ensures that the model is not trivially overparametrized, which occurs when the number of parameters in the expansion exceeds the number of free entries in π. Thus, mixture identifiability would require that for every finite p, collection of finite integers $\{d_j\}_{j=1,\ldots,p}$, and every K for which $C(K) < \left(\prod_j d_j - 1\right)$, the Jacobian is full-rank. Unfortunately, there are several well-known counterexamples, the simplest of which is the following.

Example 3.2.1 (Rank-deficient PARAFAC expansion)**.** Let $p = 4$, $K = 3$, and $d_j = 2$ for every j. Then $C(K) = 2 + 3(4) = 14$ but $\text{rnk}(J_{g_2}) = \text{rnk}(J_{g_1}) = 13$, where $\text{rnk}(J)$ is the rank of the matrix J. Thus J_{g_2} is rank-deficient, and the mixture identifiability condition does not hold in this case.

It is interesting to note that all such examples that we are aware of occur for values of K for which $C(K+1) = \prod_j d_j - 1$. For the case above, $3 + 4(4) = 19 > 2^4 - 1$, so $C(4) = 2^4$. Some other examples are given Fienberg et al. (2007), all of which share this characteristic. It is tempting to conjecture that such dimension defects only occur for values of K satisfying $C(K+1) = \prod_j d_j - 1$, but we are aware of no such proof nor is there an obvious path to such a result.

3.2.2 Algebraic preliminaries

We now review some algebraic concepts which allow us to prove a result that effectively limits how large dimension defects such as that in Example 3.2.1 can ever be. The discussion in this section follows chapter 4 of Drton et al. (2009). Let $\mathbb{K}$ be a field and consider the affine varieties $V, W \subset \mathbb{K}^k$. The *join* of V and W is the affine algebraic variety

$$\mathcal{J}(V, W) := \overline{\{\lambda v + (1-\lambda)w : v \in V, w \in W, \lambda \in \mathbb{K}\}}, \tag{3.5}$$

where for any set A, $\bar{A}$ is the *Zariski closure* of A, given in Definition 3.2.2.

Definition 3.2.2 (Zariski closed sets)**.** Let $\mathcal{F}$ be a subset of the polynomial ring in n indeterminants over an infinite field k. Let $\mathbb{A}^n$ denote affine n-space over k, so a point of $\mathbb{A}^n$ is an n-tuple $(a_1, \ldots, a_n)$ with each $a_i \in k$. Then a set of the form

$$V(\mathcal{F}) = \{x \in \mathbb{A}^n : f(x) = 0 \ \forall \ f \in \mathcal{F}\}$$

is a closed set in the Zariski topology.

Therefore, the join defined in 3.5 is the Zariski closure of the set of all points lying on lines spanned by a point in V and a point in W. If $W = V$ this is the *secant* variety of V, which we write as $\text{Sec}^2(V) = \mathcal{J}(V, V)$. The s-th secant variety is defined by the recursion

$$\text{Sec}^1(V) = V, \qquad \text{Sec}^s(V) = \mathcal{J}(\text{Sec}^{s-1}(V), V). \tag{3.6}$$

Often in statistics, the parameter spaces of models are semi-algebraic sets, not alegraic varieties.

Definition 3.2.3 (Semi-algebraic set)**.** Let V be a semi-algebraic set, and $f_i, i = 1, \ldots, m_1$ and $h_j, j = 1, \ldots, m_2$ collections of polynomials. Then V can be represented as a finite union of sets of the form

$$\{x \in \mathbb{R}^n : f_i(x) > 0, h_j(x) = 0\}. \tag{3.7}$$

In many statistical models, the sets V, W will be subsets of $\mathbb{R}^k$. In particular, we are interested in the case where V corresponds to the set of all d^p nonnegative tensors of nonnegative PARAFAC rank 1. The following remark ensures that V is a semi-algebraic set.

Remark 3.2.1. V is a semi-algebraic set.

In the previous section, we described the relationship between the Jacobian of the map g_1 and mixture identifiability. The following Theorem from Geiger et al.

(2001) shows that the Jacobian rank is equal to the dimension of a semialgebraic set, providing the algebraic context for studying mixture identifiability for models of the form (3.3).

Theorem 3.2.4 (Geiger et al. 2001, Theorem 10)**.** *Let $g : A \subseteq \mathbb{R}^m \to \mathbb{R}^n$ be a polynomial mapping where A is a semialgebraic open set. Let $J(x) = \partial g/\partial x$ be the Jacobian matrix at x. Then the maximal rank of $J(x)$ is equal to the dimension of $g(A)$.*

The model in (3.10)-(3.11) is a mixture model, which also has an algebraic equivalent. The *mixture* of V and W is the set

$$\text{Mixt}(V, W) = \{\nu v + (1-\nu)w, v \in V, w \in W, \nu \in [0,1]\},$$

with higher order mixtures defined by

$$\text{Mixt}^1(V) = V, \qquad \text{Mixt}^s(V) = \text{Mixt}(\text{Mixt}^{s-1}(V), V),$$

which can be written explicitly as

$$\text{Mixt}^s(V) = \left\{ \sum_{h=1}^{K} \nu_h v_h, v_h \in V, \nu \in \Delta^{K-1} \right\}. \tag{3.8}$$

Note the correspondence between (3.8) and (3.3). For large K, this operation stabilizes, resulting in the convex hull of V. This gives an algebraic intuition for the minimum in the expression for $C(K)$ in (3.4); beyond a certain point, every value of K results in the convex hull of the set of rank-1 nonnegative probability tensors.

The following result provides a relationship between mixtures and secant varieties.

Proposition 3.2.5. *If V is a semi-algebraic set, then the secant variety* $\text{Sec}^s(\bar{V})$ *is the Zariski closure of the mixture* $\text{Mixt}^s(V)$*.*

We now give a result that says that while dimension defects may exist, and thus models of the form (3.3) are not always mixture identifiable, they are "rank

identifiable" in the sense that if the model is properly specified and data are generated from the prior with K mixture components, then the resulting parameter π will have rank K almost surely.

Theorem 3.2.6 (Rank identifiability of PARAFAC models)**.** *Let V be the set of all $\prod_j d_j$ probability tensors with nonnegative PARAFAC rank ≤ 1. Let V_K be the space of $\prod_{j=1}^p d_j$ probability tensors defined by*

$$V_K := \left\{ \pi : \pi = \sum_{h=1}^{K} \nu_h \bigotimes_{j=1}^{p} \lambda_h^{(j)} \right\},$$

where $\nu \in \Delta_{K-1}$ and $\lambda_h^{(j)} \in \Delta_{d_j-1}$ for each h and j. Equivalently, $V_K = \text{Mixt}^R(V)$. Assume $K + (K+1)\sum_j (d_j - 1) < \prod_j d_j - 1$, and that $K < \min_l \prod_{j \neq l} d_j$. Let $\nu \sim \mu_\nu$ and $\lambda_h^{(j)} \sim \mu_{\lambda_h^{(j)}}$ be probability measures on Δ_{K-1} and Δ_{d_j-1}, respectively, that are jointly absolutely continuous with respect to Lebesgue measure. Let μ be the probability measure on $\text{Mixt}^R(V)$ induced by assigning the measures $\mu_\nu, \{\mu_{\lambda_h^{(j)}}\}_{j \leq p, h \leq K}$ independently to the model parameters. Then

$$\mu(\text{Mixt}^{K-1}(V)) = 0.$$

Although we do not prove it here, this result should be sufficient to show that the posterior concentrates on the correct number of components in large samples.

3.2.3 Checking mixture identifiability

The result in Theorem 3.2.4 provides a practical means to check for dimension deficiency in mixture models of the form (3.3) when combined with the following lemma

Lemma 3.2.1 (Geiger et al. 2001, Lemma 9)**.** *Let $g : \mathbb{R}^m \to \mathbb{R}^n$ be a polynomial mapping. Let $J(x) = \partial g / \partial dx$ be the Jacobian matrix at x. Then the rank of $J(x)$ equals the maximal rank almost everywhere.*

Thus, one can compute the rank of the Jacobian by calculating its entries – this can be done analytically for the map g_2 – then computing its rank at an arbitrary point. This will give the maximal rank with probability 1. An algorithm of this sort is proposed in Geiger et al. (2001).

This suggests an approach for manually checking the mixture identifiability conditions prior to computation for a particular dataset using the following procedure

1. Compute p and d_j for each j.
2. Calculate the maximal value of K for which $C(K) < \prod_j d_j$. Call this $K_{\max}$. Compute $C(K)$ for every $K < K_{\max}$.
3. Compute analytically the entries of $J_{g_2}(K)$ for each K.
4. For every $K < K_{\max}$, compute the rank of $J_{g_2}(K)$ at an arbitrary point and compare it to $C(K)$.

If every $C(K)$ is equal to the rank of $J_{g_2}(K)$, then this particular model is mixture identifiable. Note that because the model is equivalent for every $K > K_{\max}$, there are always finitely many ranks to compute. When $K_{\max}$ is large, we generally expect the true value of K to be small relative to $K_{\max}$, and thus it may be warranted to restrict the prior to values of $K \ll K_{\max}$, further reducing the computational demand of computing Jacobian ranks. Nonetheless, this procedure is only practical when $C(K)$ for the largest value of K assigned any prior mass is not too large; otherwise, numeric methods for computing the rank of J_{g_2} will be both numerically unstable and very computationally intensive. The main implication of this is that in many cases, one can actually verify directly the mixture identifiability condition, and therefore consistency for K, on a case-by-case basis using a relatively straightforward procedure. As such, the model in (3.10)-(3.11) with the MFM prior we describe in Section 3.3 will probably be consistent for K in most cases, and the consistency

condition can actually be checked using the procedure described above in conjunction with the results in Miller (2014).

3.3 A MFM model for inferring population structure

The result in Theorem and the procedure described in Section 3.2.3 suggest that (1) MFM models may well be consistent for the number of classes K for models of the form (3.3) and (2) it is possible to check whether there is any cause for concern about consistency using a relatively simple procedure, at least for cases in which p and d_j are not too large. Given the lack of any theoretical support for the methods of Pritchard et al. (2000) and Evanno et al. (2005) and the demonstrated inconsistency of DPMs for the number of components, MFM models would seem to offer a far superior approach to inference on population structure. Here we develop one such model.

To begin, we modify the model in (3.10)-(3.11) to place a prior on K, $K \sim p(K)$. Although in principle any probability mass function can be chosen for $p(K)$, we follow the recommendation in Miller (2014) and put $\log p(K) \propto \log(1/10) + (K - 1)\log(9/10)$. With a proper prior on K, we obtain a proper posterior distribution for K that offers Bayes point estimates as well as uncertainty quantification via the posterior distribution. A fully Bayesian specification of the model is completed by choosing values for the prior hyperparameters $\lambda_h^{(j)}$ and ν. We consider two choices for $a^{(j)}$: $a_c^{(j)} = 1$ and $a_c^{(j)} = 1/d_j$. The latter has the advantage of being dimension-free, unlike the Dirichlet$(1, 1, \ldots, 1)$ prior, which corresponds to increasing numbers of "prior observations" with increasing d_j. On the other hand, the Dirichlet$(1, 1, \ldots, 1)$ is the default choice of Pritchard et al. (2000), so it has probably been used in the vast majority of previous analyses with `STRUCTURE` software; moreover, it is a common "default" Dirichlet prior in similar models, see e.g. Dunson and Xing

(2009). The choice of $a_c^{(j)} = 1$ corresponds to prior belief that the allele frequencies are roughly equal in the population, whereas $a_c^{(j)} = 1/d_j$ corresponds to the prior belief that there are a few very common alleles and the rest are relatively rare. Ultimately, where some prior knowledge exists to guide this choice, it is always preferred to the use of a "default" prior that may be informative, as is the case with both choices we consider. For ν, we are restricted to choose a symmetric Dirichlet prior to make inference computationally feasible (see Miller (2014)). We elect to put $\nu = (1, 1, \ldots, 1)$, favoring subpopulations of roughly equal sizes. If one had special prior knowledge that there are likely to exist one or more very small subpopulations, it would make sense to modify this choice. Regardless, any choice here is preferred to the approach in Pritchard et al. (2000) of fixing these parameters at $1/K$, since we are allowing the data to inform about them.

Computation for this model is performed by a split-merge Gibbs sampler. The sampler is virtually identical to that outlined in Miller and Harrison (2015). The marginal likelihoods $m(y)$ appearing in various steps of the sampler are the Dirichlet-Multinomial likelihoods arising from marginalizing over $\lambda_h^{(j)}$ in (3.11), given by

$$\log m(y \mid a_h^{(j)}) = \log \Gamma\left(\sum_{c_j=1}^{d_j} a_{hc_j}^{(j)}\right) - \log \Gamma\left(\sum_{i=1}^{N} \mathbb{1}_{\{z_i=h\}} + \sum_{c_j=1}^{d_j} a_{hc_j}^{(j)}\right) + \sum_{c_j=1}^{d_j} \left[\log \Gamma\left(a_{hc_j}^{(j)} + \sum_{i:z_i=h} \mathbb{1}_{\{y_i^{(j)}=c_j\}}\right) - \log \Gamma\left(a_{hc_j}^{(j)}\right)\right]. \tag{3.9}$$

3.4 Simulation studies

We conduct a simulation study to assess the performance of the MFM prior on K for the model in (3.10)-(3.11) in estimating K when the model is properly specified. Data are simulated using the following procedure

1. Fix p, K, and d_j for each $j = 1, \ldots, p$. Sample $\lambda_h^{(j)} \sim \text{Dirichlet}(a, \ldots, a)$ for

$j = 1, \ldots, p$, $h = 1, \ldots, K$.

2. Fix ν and N and sample $z_i \sim \text{Categorical}(\nu)$ for $i = 1, \ldots, N$.

3. Sample $y_{ij} \mid z_i \sim \text{Categorical}(\lambda^{(j)}_{hz_i})$ for $i = 1, \ldots, N$, $j = 1, \ldots, p_0$.

In all of the simulations, we simulate data with $p = 10$, $K = 4$, $d_j = 4$ for every j, and $\nu = (0.5, 0.2, 0.1, 0.1)$. We choose a to be either 1 or $1/p$, and conduct simulations with $N = 100$, $N = 1000$, and $N = 10,000$. Note that the dimension of π is $4^{10} = 2^{20} \approx 10^6$, so that the sample size is always much smaller than the dimension of the full parameter space. However, the true value of $C(K) = 3 + 4(3)(10) = 123$, so that the expected dimension of the model is comparable to the sample size when $N = 100$ and smaller than the the sample size for the other values of N.

After data are simulated, computation is performed for the model in (3.10)-(3.11) using the split-merge Gibbs sampler described in Miller and Harrison (2015) and Section 3.3. We retain 1000 samples for inference from 4500 samples taken after a burn-in period of 500 iterations, for a thinning factor of 9. In each case, we set the prior hyperparameter $a^{(j)}_h$ to the same value used in simulating data. For comparison, we also estimate a DPM. This model can be defined using the stick-breaking representation of the Dirichlet process by

$$\mathbb{P}\left[y_{i1} = c_1, \ldots, y_{ip} = c_p \mid z_i\right] = \prod_j \lambda^{(j)}_{z_i c_j}$$

$$L_1, L_2, \ldots \overset{iid}{\sim} \bigtimes_{j=1}^{p} \text{Dirichlet}(a^{(j)}_1, \ldots, a^{(j)}_{d_j})$$

$$z_1, \ldots, z_N \overset{iid}{\sim} \nu, \quad \nu = (\nu_1, \nu_2, \ldots), \quad \nu_h = \nu^*_h \prod_{l=1}^{h-1}(1 - \nu^*_l)$$

$$\nu^*_1, \nu^*_2, \ldots \overset{iid}{\sim} \text{Beta}(1, \alpha),$$

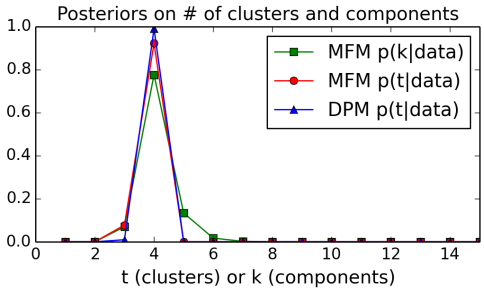
where here $L_h = \{\lambda_h^{(j)}\}_{j=1,\ldots,p}$ and $\times$ denotes a product measure. We use the hyper-prior $p(\alpha) \propto e^{-\alpha}$ on the concentration parameter α. Computation for the DPM is performed using the computational algorithm in Miller and Harrison (2015), which is almost identical to the computational algorithm used for the MFM prior.

We focus exclusively on estimation of K, which is always 4. The posterior under the MFM prior gives explicitly a posterior distribution $K \mid y_1, \ldots, y_p = p(k \mid y_1, \ldots, y_p)$ for the parameter K. This is denoted by $p(k \mid \text{data})$ in the figures; some authors refer to K as the number of components and $p(k \mid \text{data})$ as the posterior distribution on the number of components. In the DPM, $K = \infty$; there are infinitely many components so K is not a parameter of the model. Thus, to estimate the true value of K when the data are thought to originate using a finite mixture, it is common to use an estimate of the posterior distribution for the number of occupied clusters as though it is a posterior distribution for K. We denote this as $p(t \mid \text{data})$ in the figures, and it is shown for both the DPM and MFM for comparison. For a detailed discussion of the distinction between components and clusters and how $p(t \mid \text{data})$ and $p(k \mid \text{data})$ are computed, see Miller and Harrison (2015).

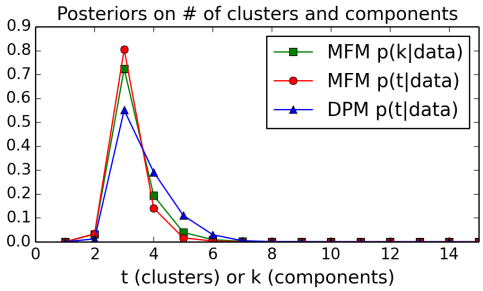
Figure 3.1 shows results for the six simulations (two values of a and three values of n). Shown are the posterior distribution on the number of components $p(k \mid \text{data})$ for the MFM and the posterior on the number of clusters $p(t \mid \text{data})$ for both the DPM and MFM. When $a = 1/p$, the distribution $p(t \mid \text{data})$ actually performs slightly better than either $p(k \mid \text{data})$ or $p(t \mid \text{data})$ when using the posterior mode of either distribution as a point estimate of K. It is possible, however, that this distribution understates the uncertainty in K. With $a = 1/p$, for $N = 1000$ and $N = 10,000$, the posterior is highly concentrated on the true value of $K = 4$; only for $N = 100$ is there meaningful uncertainty, and then only under the MFM model. The results for $a = 1$ are quite different. In this case, for $N = 100$, there is noticeable uncertainty in

all three distributions. Although all three have a mode at the true value of $K = 4$, $p(k \mid \text{data})$ and $p(t \mid \text{data})$ under the MFM are more concentrated around this value than $p(t \mid \text{data})$. As N increases, the MFM posterior becomes more concentrated on the true value of K, while the DPM continues to place substantial mass on larger numbers of clusters. This behavior of the DPM is consistent with the theoretical results in Miller and Harrison (2014).

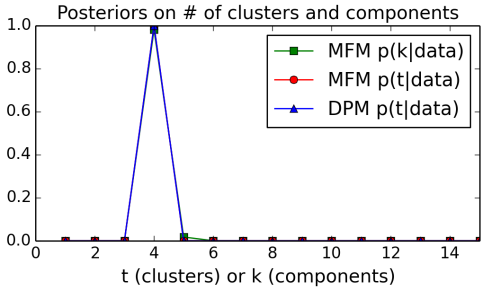
$N = 100, a_c^{(j)} = 1/d_j$



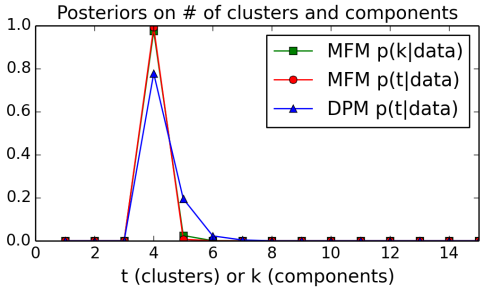
$N = 100, a_c^{(j)} = 1$



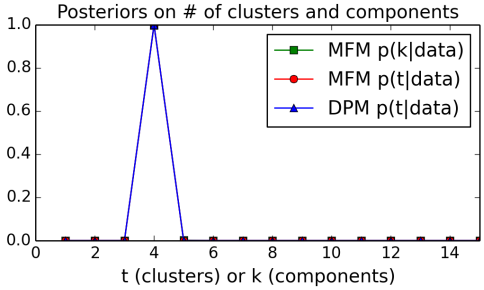
$N = 1000, a_c^{(j)} = 1/d_j$



$N = 1000, a_c^{(j)} = 1$



$N = 10,000, a_c^{(j)} = 1/d_j$



$N = 10,000, a_c^{(j)} = 1$

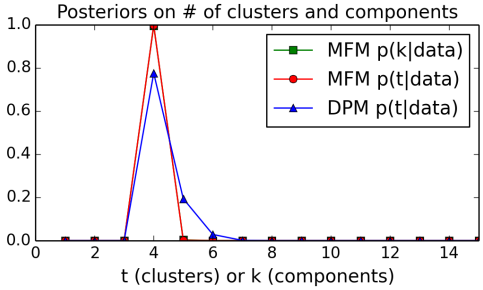


FIGURE 3.1: *Plots showing the posterior distribution on the number of components* $p(k \mid \text{data})$ *for the MFM and the posterior distribution on the number of clusters* $p(t \mid \text{data})$ *for both the DPM and MFM.*

3.5 Application to red-winged blackbird data

3.5.1 Model

We now apply the model in (3.10)-(3.11) to estimate the number of distinct genetic populations of red-winged blackbirds. To begin, we modify the model to the specific application of multilocus genotype data from a diploid genome, consistent with Pritchard et al. (2000), i.e.

$$\mathbb{P}\left[y_{i1}^{(1)} = c_{11}, y_{i1}^{(2)} = c_{12}, \ldots, y_{ip}^{(1)} = c_{p1}, y_{ip}^{(2)} = c_{p2} \mid z_i = h\right] = \prod_{j=1}^{p}\prod_{r=1}^{2} \lambda_{hc_{jr}}^{(j)} \tag{3.10}$$

$$\lambda_h^{(j)} \sim \text{Dirichlet}(a_1^{(j)}, \ldots, a_{d_j}^{(j)}), \quad \mathbb{P}\left[z_i = h\right] = \nu_h, \quad \nu_h \sim \text{Dirichlet}(\alpha, \ldots, \alpha). \tag{3.11}$$

Here, $y_{ij}^{(1)}, y_{ij}^{(2)}$ are the two alleles of gene j in the diploid genome of subject i. Inference and computation for this model with MFM prior, and the analogous Dirichlet process mixture, is very similar to that for the generic model in (3.10)-(3.11). The main practical difference is that the sufficient statistics used in computing the log marginal likelihood in (3.9) replace the expression $\sum_{i:z_i=h} \mathbb{1}_{\{y_i^{(j)}=c_j\}}$ with $\sum_{i:z_i=h} \mathbb{1}_{\{y_{i1}^{(j)}=c_j\}} + \mathbb{1}_{\{y_{i2}^{(j)}=c_j\}}$, to account for the fact that each individual now contributes two observations at each locus. From a modeling standpoint, the effect of this change is to appropriately incorporate the knowledge that the population-level allele frequencies are the same for $y_{ij}^{(1)}$ and $y_{ij}^{(2)}$. Marginalizing over the latent class variable z, we can represent the joint distribution of the y variables as a (self) outer product of a PARAFAC tensor factorization. In other words, if $\pi_{c_{11},c_{12},\ldots,c_{p1},c_{p2}} = \mathbb{P}\left[y_{i1}^{(1)} = c_{11}, y_{i1}^{(2)} = c_{12}, \ldots, y_{ip}^{(1)} = c_{p1}, y_{ip}^{(2)} = c_{p2}\right]$, then π is a $\left(\prod_j d_j\right)^2$ probability tensor given by

$$\pi = \pi^* \otimes \pi^*, \quad \pi^* = \sum_{h=1}^{K} \nu_h \bigotimes_{j=1}^{p} \lambda_h^{(j)}.$$

3.5.2 Data and results

The data are multilocus genotype data at ten loci from red-winged blackbirds in eight locations: Kentucky, New York, Michigan, Ontario, Wisconsin, Washington, Pennsylvania, and the Bahamas. The data are described extensively in Liu et al. (2015). The largest number of unique alleles at any locus observed in the data is 68; the smallest is 2. The corresponding contingency table, with parameter π given in (4.1), has $(2.9 \times 10^{14})^2$ cells, so this application is relatively high-dimensional. There are a total of $N = 306$ individuals in the data. The largest number of individuals from any geographic region is 66 from the Bahamas, the smallest is 13 from Ontario. When applied to these data, the method of Evanno et al. (2005) gives a point estimate of $K = 2$ unique subpopulations. We applied the MFM and corresponding DPM models described in Section 3.5.1 to estimation of K in these data. Our main interest is in assessing (1) agreement with the method of Evanno et al. (2005) in point estimation and (2) uncertainty quantification for K using the posterior distribution under the MFM and DPM models.

To test sensitivity to the prior distribution on $\lambda_h^{(j)}$, we consider two choices for $a^{(j)}$: $a_c^{(j)} = 1$, and $a_c^{(j)} = 1/d_j$. The former is the default choice in Pritchard et al. (2000) and in the `STRUCTURE` software commonly used in biology. The latter is a the unit information prior and a common symmetric alternative. Figure 3.2 shows the approximate posterior distribution of K based on 1000 samples retained from 19,000 samples taken after a burn-in period of 1000 samples. Regardless of the model, when $a_c^{(j)} = 1$, the posterior mode for K is consistent with the point estimate of $K = 2$ obtained using the Evanno et al. (2005) method, the posterior probability that $K \neq 2$ is less than 0.01. However, when $a_c^{(j)} = 1/d_j$, the mode actually occurs at $K = 3$ with both the MFM and DPM models, and the MFM posterior $p(k \mid \text{data})$ places significant mass on $K = 2, 3, 4$. Thus, estimation of K is highly sensitive to the prior

choice in this dataset.

$a_c^{(j)} = 1/d_j$ $\qquad$ $a_c^{(j)} = 1$

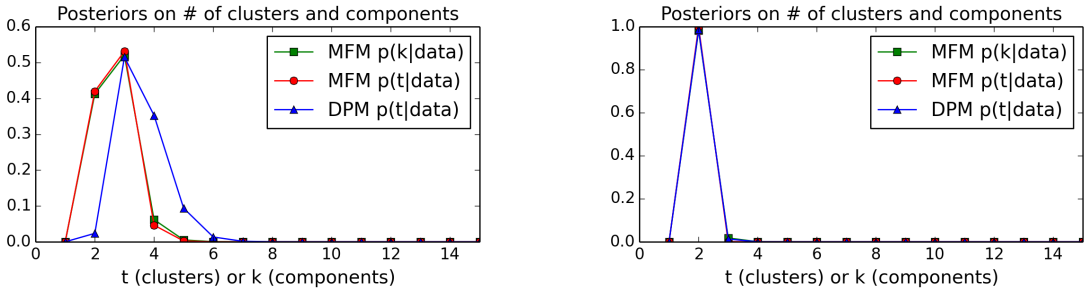


FIGURE 3.2*: Plots showing the posterior distribution on the number of components* $p(k \mid \text{data})$ *for the MFM and the posterior distribution on the number of clusters* $p(t \mid \text{data})$ *for both the DPM and MFM for* $a_h^{(j)} = 1/d_j$ *and* $a_h^{(j)} = 1$*. Models are estimated on red-winged blackbird multilocus genotype data.*

Additional results from the model suggest that uncertainty in K may not be biologically significant. We use the MCMC output from the MFM model with either choice for $a_c^{(j)}$ to assess how the latent population structure relates to the geographic subpopulations in the data. One question of interest is whether the Bahamas population is genetically distinct from the continental populations, and whether there are any distinct subpopulations within the continental populations. The analysis in Liu et al. (2015) using the method of Pritchard et al. (2000) suggested that when $K = 2$, the Bahamas population is genetically distinct from the continental populations. To assess this structure, we compute the average (across individuals) posterior probability of membership in each latent class for each geographic subpopulation. To identify the latent classes, at every iteration we assign the latent classes a numeric index corresponding to their relative size, with the largest class assigned the numeric index 1. Table 3.1 shows results for $a_c^{(j)} = 1/d_j$ and Table 3.2 shows results for $a^{(j)}c = 1$. In both cases, the largest class consists of almost all of the continental individuals with high probability, and the second largest class consists of exclusively individuals from the Bahamas, and in most cases, it consists of all of the Bahamas

individuals. When $a_c^{(j)} = 1$, no other classes appear in the 1000 samples taken. When $a_c^{(j)} = 1/d_j$, two other classes appear, but these classes have very low weight and are often unoccupied. The largest share of either of classes 3 or 4 of the posterior weight of any geographic subpopulation is 0.02 for New York. Thus, while the posterior for K suggests that there may in fact be more than two distinct genetic subpopulations if we have strong prior preference for $a_c^{(j)} = 1/d_j$, these genetic subpopulations are very small and do not correlate strongly with any of the geographic subpopulations. As such, they are unlikely to be biologically meaningful.

Table 3.1: Approximate average posterior probability of class membership within the eight geographic populations with $a_h^{(j)} = 1/d_j$.

	Bah	KY	MI	NY	Ont	PA	WA	WI
1	0.0004	1.0000	1.0000	0.9789	1.0000	1.0000	0.9995	0.9996
2	0.9996	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000	0.0200	0.0000	0.0000	0.0003	0.0003
4	0.0000	0.0000	0.0000	0.0011	0.0000	0.0000	0.0002	0.0001

Table 3.2: Approximate average posterior probability of class membership within the eight geographic populations with $a_h^{(j)} = 1$.

	Bah	KY	MI	NY	Ont	PA	WA	WI
1	0.0007	0.9645	0.9930	0.9893	0.9995	0.9980	0.9972	0.9389
2	0.9993	0.0355	0.0070	0.0107	0.0005	0.0019	0.0028	0.0611

3.6 Discussion

The analysis of population structure using multilocus genotype data relies heavily on latent class models for multivariate categorical data. In this context, the number of latent classes is interpreted as the number of genetically distinct subpopulations of an organism, which is relevant to assessing evolutionary history and conservation status. Here we have described some weaknesses in the popular methods of Pritchard et al. (2000) and Evanno et al. (2005) for estimating the number of latent classes.

The weaknesses in these methods are plainly acknowledged in the original paper of Pritchard et al. (2000), but the fact that Evanno et al. (2005) has been cited over 7,000 times and deals only with estimation of K based on the model of Pritchard et al. (2000) suggests that researchers in population genetics rely heavily on these methods for inference on K.

Here we have proposed an alternative approach for inference on K. Like the methods of Pritchard et al. (2000), our approach is fully Bayesian. However, we do not rely on approximations to marginal likelihoods from multiple MCMC sample paths to assess posterior evidence for different values of K, but instead place a prior on K and perform joint inference on K and the other model paramters. We provide some theoretical support for the proposed method and characterize its performance empirically. We also give a numerical method for checking on a case-by-case basis whether the model we propose is mixture identifiable. Since mixture identifiability implies consistency for K via the approach in Miller and Harrison (2015), the consistency of the method for estimation of K can be checked on a case-by-case basis. That the MFM prior structure we recommend has strong theoretical support provides a greater degree of confidence in our approach. Most importantly, uncertainty quantification is a critical aspect of science, and the methods of Pritchard et al. (2000) and Evanno et al. (2005) do not provide robust quantification of uncertainty in K, whereas our method does. As such, we suggest the use of our method to provide quantification of uncertainty in K, minimally as a check on the current default methods of Pritchard et al. (2000) and Evanno et al. (2005).

4

Optimal Gaussian approximations to the posterior for log-linear models with Diaconis–Ylvisaker priors

4.1 Introduction

Contingency table analysis routinely relies on log-linear models, which represent the logarithm of cell probabilities as an additive model (Agresti, 2002). With the standard choice of Multinomial or Poisson likelihood, these are exponential family models, and are routinely fit through maximum likelihood estimation (Fienberg and Rinaldo, 2007). However, sparsity in the observed cell counts often makes maximum likelihood estimation infeasible (see Haberman (1974) and Bishop et al. (2007)) in practical applications. In such cases, regularization is often used to obtain unique parameter estimates (Park and Hastie, 2007; Zou and Hastie, 2005).

A common Bayesian approach to inference in high-dimensional contingency tables is to place a conjugate prior on the parameters of a graphical or hierarchical log-linear model, and an independent prior over the space of all such models (see e.g. Massam et al. (2009)). This leads to a standard model-averaged posterior (Hoeting et al., 1998), where all possible sparse log-linear models in the chosen class are

weighted by their posterior evidence. Use of non-conjugate (e.g. Gaussian) priors with computation by Markov chain Monte Carlo (Gelfand and Smith, 1990) has also been proposed (Dellaportas and Forster, 1999). Although model averaging is generally considered ideal in high dimensional settings, computational algorithms for posterior inference scale exceedingly poorly in p. Since the smallest contingency table corresponding to cross-classification of p categorical variables has 2^p cells, the corresponding log-linear model has $2^p - 1$ free parameters, so the model space grows super-exponentially in p. Accordingly, posterior computation is essentially infeasible for $p > 15$, the largest case demonstrated to date in the literature (Dobra and Massam, 2010) to the best of our knowledge.

Alternatively, one can place a Gaussian prior on the parameters of a saturated log-linear model to induce Tikhonov type regularization, and then perform computation by Markov chain Monte Carlo. This approach is well-suited to situations in which the sample size is not tiny relative to the table dimension, but where zero counts nonetheless exist in some cells. In this case, data augmentation Gibbs samplers such as that proposed by Polson et al. (2013) provide for conditionally conjugate updates. However, this by itself is computationally intensive relative to alternatives such as elastic net (Zou and Hastie, 2005), and can suffer from poor mixing. In principle, a more scalable Bayesian approach for producing Tikhonov regularized point estimates would be to utilize the Diaconis–Ylvisaker conjugate prior (Diaconis and Ylvisaker, 1979) on the parameters of the log-linear model, which is essentially computation free. The main drawback is that the resulting posterior distribution is difficult to work with, lacking closed form expressions for even marginal credible intervals or fast algorithms for sampling from the posterior. An accurate and more tractable approximation to this posterior is therefore of practical interest.

Approximations to the posterior distribution have a long history in Bayesian statistics, with the Laplace approximation perhaps the most common and simple

alternative (Tierney and Kadane, 1986; Shun and McCullagh, 1995). More sophisticated approximations, such as those obtained using variational methods (Attias, 1999) may in some cases be more accurate but require computation similar to that for generic EM algorithms. Moreover, there exist no theoretical guarantees of the approximation error in finite samples, and these approximations are known to be inadequate in relatively simple models (Wang and Titterington, 2004, 2005).

In this article, we propose a Gaussian approximation to the posterior for log-linear models with Diaconis–Ylvisaker priors. The approximation is shown to be the optimal Gaussian approximation to the posterior in the Kullback–Leibler divergence, and convergence rates to the exact posterior and a finite-sample Kullback–Leibler error bound are provided. The approximation is shown empirically to be accurate even for modest sample sizes; effectively, the empirical results suggest that the approximation is accurate enough to be used in place of the exact posterior within the range of sample sizes for which the posterior is sufficiently concentrated to be statistically useful. We also show how the approximation can be used to perform model selection using the penalized credible region method (Bondell and Reich, 2012). In a real data application, the method performs favorably in model selection for graphical log-linear models compared to methods requiring vastly greater computational resources.

4.2 Background

We first provide a brief review of exponential families. We then describe the family of conjugate priors for the natural parameter of an exponential family, referred to as Diaconis–Ylvisaker priors. We then provide more detailed background on log-linear models for Multinomial likelihoods and the associated Diaconis–Ylvisaker prior.

4.2.1 Exponential families

Following Diaconis and Ylvisaker (1979), let μ be a σ-finite measure defined on $(\mathbb{R}^p, \mathcal{B})$, where $\mathcal{B}$ denotes all Borel sets on $\mathbb{R}^p$. Let $\text{supp}(\mu) = \{y \in \mathbb{R}^p : d\mu(y) > 0\}$ be the support of μ, and define $\mathscr{Y}$ as the interior of the convex hull of $\text{supp}(\mu)$. For $\theta \in \mathbb{R}^p$, define $M(\theta) = \log \int_{\mathscr{Y}} e^{\theta^T y} d\mu(y)$, and let $\Theta = \{\theta \in \mathbb{R}^p : M(\theta) < \infty\}$, which we assume is an open set. We refer to Θ as the natural parameter space. The exponential family of probability measures $\{P(\cdot;\theta)\}$ indexed by a parameter $\theta \in \Theta$ is defined by

$$dP(y;\theta) = e^{\theta^T y - M(\theta)} d\mu(y), \qquad \theta \in \Theta. \tag{4.1}$$

This family includes many of the probability distributions commonly used as sampling models in likelihood-based statistics. Diaconis and Ylvisaker (1979) develop the family of conjugate priors for the parameter θ of regular exponential family likelihoods. These Diaconis–Ylvisaker priors are given by

$$d\pi(\theta; n_0, y_0) = e^{n_0 y_0^T \theta - n_0 M(\theta)}, \qquad n_0 \in \mathbb{R}, y_0 \in \mathbb{R}^d. \tag{4.2}$$

On observing data y consisting of n observations with sufficient statistics $\bar{y}$, the posterior is then also Diaconis–Ylvisaker, with parameters $n_0 + n, y_0 + \bar{y}$, i.e. $d\pi(\theta \mid y) = d\pi(\theta; n_0 + n, y_0 + \bar{y})$. In the sequel we focus on one member of the exponential family, the multinomial. In the natural parametrization, the ultinomial likelihood gives rise to the log-linear model and the closely related multinomial logit model, which we now describe.

4.2.2 Log-linear models

Let $\mathcal{S}^d = \{(x_1, \ldots, x_d) \in [0,1]^d : \sum_{j=1}^d x_j \leq 1\}$ denote the d-dimensional unit simplex. Consider N independent samples from a categorical variable with $(d+1)$ levels. We denote the levels of the variable by $0, 1, \ldots d$, without loss of generality. Let y_j denote the number of times the jth level is observed in the N samples and

set $y = (y_0, y_1, \ldots, y_d)^{\mathrm{T}}$; clearly $\sum_{j=0}^{d} y_j = N$. The joint distribution of y is given by a multinomial distribution, denoted $y \sim \text{Multinomial}\,(N, \pi)$, which is parametrized by $\pi = (\pi_1, \ldots, \pi_d)^{\mathrm{T}} \in \mathcal{S}^d$, where π_j is the probability of observing the jth level for $j = 1, \ldots, d$.

The log-linear model is a generalized linear model for multinomial likelihoods obtained by choosing the logistic link function, which also results in the natural exponential family parametrization. Define the logistic transformation $\ell : \mathbb{R}^d \to \mathcal{S}^d$ and its inverse log ratio transformation $\ell^{-1} : \mathcal{S}^d \to \mathbb{R}^d$ as

$$\pi_j = \frac{e^{\theta_j}}{1 + \sum_{l=1}^{d} e^{\theta_l}}, \quad \theta_j = \log(\pi_j/\pi_0), \quad (j = 1, \ldots, d), \tag{4.3}$$

where $\pi_0 = 1 - \sum_{j=1}^{d} \pi_j$, and $\theta_0 = 0$. We shall write $\pi = \ell(\theta)$ and $\theta = \ell^{-1}(\pi) = \log(\pi/\pi_0)$, respectively, to denote the transformations in (4.3). Using (4.3), the multinomial likelihood in the log-linear parameterization can be expressed as

$$f(y \mid \theta) \;\propto\; \frac{\exp\left(\sum_{j=1}^{d} y_j \theta_j\right)}{\left(1 + \sum_{l=1}^{d} e^{\theta_l}\right)^N}. \tag{4.4}$$

An important motivating case is when $y = \text{vec}(\mathbf{n})$, with $\mathbf{n}$ a contingency table arising from cross-classification of N independent observations on p categorical variables $y_1, \ldots, y_p$. Suppose that the vth variable y_v has d_v many levels, so that the contingency table has $\prod_{v=1}^{p} d_v$ many *cells*, and y is a $(d+1)$-dimensional vector of counts with $d = \prod_{v=1}^{p} d_v - 1$. We refer to the parametrization $\theta = \log(\pi/\pi_0)$ in the contingency table setting as the *identity* parametrization. Also of particular interest in this setting are reparametrizations of (4.3) that represent $\log \pi/\pi_0$ as an additive model involving parameters that correspond to interactions among $y_1, \ldots, y_p$. Every identified parametrization of the log-linear model for the multinomial likelihood can be represented by

$$\log(\pi/\pi_0) = X\theta^*, \tag{4.5}$$

where X is a d by d non-singular binary matrix and $\theta^* \in \mathbb{R}^d$. In the simulations and application, we make a specific choice for X that corresponds to the *corner parametrization* of the log-linear model (Massam et al., 2009). We illustrate the identity and corner parameterizations through a 2^3 contingency table in Example 4.2.1 below. Details for the general case can be found in the Appendix.

Example 4.2.1. Consider three binary variables y_1, y_2, y_3, with $y_v \in \{0, 1\}$ for $v = 1, 2, 3$, and let

$$\psi_{i_1 i_2 i_3} = \text{pr}(y_1 = i_1, y_2 = i_2, y_3 = i_3), \quad (i_1, i_2, i_3) \in \{0, 1\}^3.$$

A 2^3 contingency table $\mathbf{n} = (n_{i_1 i_2 i_3})$ is obtained from the cross-classification of N independent observations on y_1, y_2, y_3, with $n_{i_1 i_2 i_3}$ denoting the cell count for the cell (i_1, i_2, i_3). Let $y = \text{vec}(\mathbf{n}) = (n_{000}, \ldots, n_{111})^{\text{T}}$ be the vectorized cell counts with $d = 7$. In the *identity* parametrization, the vector of log-linear parameters $\theta \in \mathbb{R}^7$ is given by

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \end{pmatrix} = \log \begin{pmatrix} \pi_1/\pi_0 \\ \pi_2/\pi_0 \\ \pi_3/\pi_0 \\ \pi_4/\pi_0 \\ \pi_5/\pi_0 \\ \pi_6/\pi_0 \\ \pi_7/\pi_0 \end{pmatrix} = \log \begin{pmatrix} \psi_{001}/\psi_{000} \\ \psi_{010}/\psi_{000} \\ \psi_{011}/\psi_{000} \\ \psi_{100}/\psi_{000} \\ \psi_{101}/\psi_{000} \\ \psi_{110}/\psi_{000} \\ \psi_{111}/\psi_{000} \end{pmatrix}.$$

On the other hand, in the *corner* parametrization, we express

$$\theta = \log \begin{pmatrix} \psi_{001}/\psi_{000} \\ \psi_{010}/\psi_{000} \\ \psi_{011}/\psi_{000} \\ \psi_{100}/\psi_{000} \\ \psi_{101}/\psi_{000} \\ \psi_{110}/\psi_{000} \\ \psi_{111}/\psi_{000} \end{pmatrix} = \begin{pmatrix} \theta^*_{001} \\ \theta^*_{010} \\ \theta^*_{001} + \theta^*_{010} + \theta^*_{011} \\ \theta^*_{100} \\ \theta^*_{001} + \theta^*_{100} + \theta^*_{101} \\ \theta^*_{010} + \theta^*_{100} + \theta^*_{110} \\ \theta^*_{001} + \theta^*_{010} + \theta^*_{100} + \theta^*_{011} + \theta^*_{101} + \theta^*_{110} + \theta^*_{111} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} \theta^*_{001} \\ \theta^*_{010} \\ \theta^*_{011} \\ \theta^*_{100} \\ \theta^*_{101} \\ \theta^*_{110} \\ \theta^*_{111} \end{pmatrix}$$

$$= X\theta^*.$$

The indexing of the elements of θ^* by binary indices is for ease of interpretation. Indeed, entries of θ^* with a single 1 in the binary index are main effects, those with two 1's are two-way interactions and θ^*_{111} is a three-way interaction term. The matrix X can be easily verified to be non-singular, so that the θ and θ^* parametrizations are equivalent, with $d = 7$ free parameters in either case.

4.2.3 Conjugate priors for log-linear models

We now present the Diaconis–Ylvisaker prior for the multinomial likelihood (4.4) and derive an optimal Gaussian approximation to the corresponding posterior in Kullback–Leibler divergence. Extensions to log-linear models with a non-identity parametrization (i.e., $X \neq \mathrm{I}_d$ in (4.5)) is straightforward by invariance properties of the Kullback–Leibler divergence and are discussed subsequently. All proofs are deferred to the Appendix.

For the multinomial likelihood (4.4), the Diaconis–Ylvisaker prior is obtained by applying the inverse logistic transformation ℓ^{-1} to a Dirichlet distribution, which not surprisingly is the conjugate prior for π. Recall that $\pi_0 = 1 - \sum_{j=1}^d \pi_j$. The Dirichlet distribution $\mathcal{D}(\alpha)$ on $\mathcal{S}^d$ with parameter vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_d)^{\mathrm{T}}$ has density

$$q(\pi; \alpha) = \frac{\Gamma(\sum_{j=0}^d \alpha_j)}{\prod_{j=0}^d \Gamma(\alpha_j)} \prod_{j=0}^d \pi_j^{\alpha_j - 1}, \quad \pi \in \mathcal{S}^d, \tag{4.6}$$

and corresponding probability measure $\mathcal{Q}(\cdot, \alpha)$ with $\mathcal{Q}(A, \alpha) = \int_A q(\pi; \alpha) d\pi$ for Borel subsets A of $\mathcal{S}^d$.

Proposition 4.2.2. *Suppose* $\pi \sim \mathcal{D}(\alpha)$ *and let* $\theta = \log(\pi/\pi_0) \in \mathbb{R}^d$. *Define* $A = \sum_{j=0}^{d} \alpha_j$. *Then* θ *has a density on* $\mathbb{R}^d$ *given by*

$$p(\theta;\alpha) = \frac{\Gamma(\sum_{j=0}^{d}\alpha_j)}{\prod_{j=0}^{d}\Gamma(\alpha_j)} \frac{\exp(\sum_{j=1}^{d}\alpha_j\theta_j)}{(1+\sum_{l=1}^{d}e^{\theta_l})^A}. \tag{4.7}$$

We write $\theta \sim \mathcal{LD}(\alpha)$ and use $\mathcal{P}(\cdot;\alpha)$ to denote the probability measure associated with the density (4.7), with $\mathcal{P}(B;\alpha) = \int_B p(\theta;\alpha)d\theta$ for Borel subsets B of $\mathbb{R}^d$. If a non-identity parametrization $\theta = X\theta^*$ as in (4.5) is employed, then we denote the induced distribution on $\theta^* = X^{-1}\theta$ by $\mathcal{P}_X(\cdot;\alpha)$ and the density by $p_X(\theta;\alpha)$.

It is immediate that $\mathcal{LD}(\alpha)$ is a conjugate family of prior distributions for the likelihood (4.4), with the posterior $\theta \mid y \sim \mathcal{LD}(\alpha + y)$. To obtain some preliminary insight into the distribution family $\mathcal{LD}(\alpha)$, we derive the mean and covariance in Proposition 2 below.

Proposition 4.2.3. *Let* $\theta \sim \mathcal{LD}(\beta)$, *with* $\beta = (\beta_0, \beta_1, \ldots, \beta_d)^{\mathrm{T}}$ *and* $\beta_j > 0$ *for all* j. *Then,*

$$E(\theta_j) = \psi(\beta_j) - \psi(\beta_0), \quad (j = 1, \ldots, d)$$
$$\mathrm{cov}(\theta_j, \theta_{j'}) = \psi'(\beta_j)\delta_{jj'} + \psi'(\beta_0), \quad (j, j' = 1, \ldots, d)$$

where ψ *and* ψ' *are the digamma and trigamma functions, respectively, and* $\delta_{jj'} = 0$ *if* $j \neq j'$ *and* $\delta_{jj'} = 1$ *otherwise.*

The proof of Proposition 4.2.3 is established within the proof of Theorem 4.3.1 in the Appendix. Assume the data y is generated from a Multinomial (N, π^0) distribution and let $\theta^0 = \log(\pi^0/\pi_0^0)$ be the true log-linear parameter, where $\pi_0^0 = 1 - \sum_{j=1}^{d}\pi_j^0$. If a $\mathcal{LD}(\alpha)$ prior is placed on θ, one can use Proposition 4.2.3 to show that the posterior mean $E(\theta \mid y)$ converges almost surely to θ^0 with increasing sample size, and the posterior covariance $\mathrm{cov}(\theta \mid y)$ converges to the inverse Fisher information matrix as long as the entries of the prior hyperparameter α are suitably bounded. In

fact, a Bernstein–von Mises type result can be established, showing that the posterior distribution approaches a Gaussian distribution, centered at the true parameter value and having covariance the inverse Fisher information matrix, in the total variation metric. We do not pursue such frequentist asymptotic validations further in this paper. Our goal rather is to provide a Gaussian approximation to the posterior distribution that can be used in practice, and provide finite sample bounds to the approximation error.

4.3 Main results

In this section, we provide an optimal Gaussian approximation to a $\mathcal{LD}(\beta)$ distribution (4.7) in the Kullback–Leibler divergence, i.e., we exhibit a vector $\mu^* \in \mathbb{R}^d$ and a positive definite matrix Σ^* such that the Kullback–Leibler divergence between $\mathcal{LD}(\beta)$ and $\mathcal{N}(\mu^*, \Sigma^*)$ is the minimum among all Gaussian distributions. This result provides a readily available Gaussian approximation to the posterior distribution $\mathcal{LD}(\beta = \alpha + y)$ of the log-linear parameter θ in (4.4) with a Diaconis–Ylvisaker prior $\mathcal{LD}(\alpha)$. We also provide a non-asymptotic error bound for the Kullback–Leibler approximation. Using Pinsker's inequality, the approximation error in the total variation distance can be bounded in finite samples.

For two probability measures $\nu \ll \nu^*$, we write

$$D(\nu \mid\mid \nu^*) = E_{\nu^*} \log d\nu/d\nu^*$$

to denote the Kullback–Leibler divergence between ν and ν^*.

Theorem 4.3.1. *Given $\beta_j > 0, j = 0, 1, \ldots, d$, let $\beta = (\beta_0, \ldots, \beta_d)^{\mathrm{T}}$, and define*

$$\mu_j^* = \psi(\beta_j) - \psi(\beta_0), \quad \sigma_{jj'}^* = \psi'(\beta_j)\delta_{jj'} + \psi'(\beta_0), \tag{4.8}$$

where ψ and ψ' denote the digamma and trigamma functions respectively. Define $\mu^* = (\mu_j^*) \in \mathbb{R}^d$ and $\Sigma^* = (\sigma_{jj'}^*) \in \mathbb{R}^{d\times d}$. Then,

$$D\Big\{\mathcal{LD}(\beta) \mid\mid \mathcal{N}(\mu^*, \Sigma^*)\Big\} = \inf_{\mu,\Sigma} D\Big\{\mathcal{LD}(\beta) \mid\mid \mathcal{N}(\mu, \Sigma)\Big\}, \tag{4.9}$$

where the infimum is over all $\mu \in \mathbb{R}^d$ *and all* $\Sigma \succ 0 \in \mathbb{R}^{d\times d}$. *Further, if* $\beta_j > 1/2$ *for all* $j = 0, 1, \ldots, d$, *then*

$$D\Big\{\mathcal{LD}(\beta) \;||\; \mathcal{N}(\mu^*, \Sigma^*)\Big\} < \frac{1}{2}\sum_{j=0}^{d}\frac{1}{\beta_j} + \frac{1}{6B}, \tag{4.10}$$

where $B = \sum_{j=0}^{d}\beta_j$.

The matrix Σ^* has a compound-symmetry structure and is therefore positive-definite. From Proposition 4.2.3, the parameters of the optimal Gaussian approximation μ^* and Σ^* are indeed the mean and covariance matrix of the $\mathcal{LD}(\beta)$ distribution. Equation (4.10) provides an upper-bound to the approximation error. In the posterior, $\beta_j = \alpha_j + y_j$ and $B = \sum_{j=0}^{d}\alpha_j + N$. The condition $\beta_j \geq 1/2$ is therefore satisfied whenever every category has at least one observation. Since

$$\mathbb{E}_y[\alpha_j + y_j] = \alpha_j + N\pi_j^0,$$

the approximation error is approximately in the order of $\sum_{j=0}^{d}(\pi_j^0 N)^{-1}$, where as before π_j^0 denotes the true probability of category j. In the best case where all the categories receive approximately equal probability, i.e., $\pi_j^0 \asymp (d+1)^{-1}$, the approximation error is $\mathcal{O}(d^2/N)$. However, the convergence rate in N can be slower if some of the π_j^0s are very small. In other words, the higher the entropy of the data generating distribution, the worse the approximation is, although our simulations suggest that the approximation is practicable even for moderate sample sizes and unbalanced category probabilities. When one considers that the eigenvalues of the covariance matrix enter into the constant in Berry-Esséen convergence rates, and that here the covariance of the data is given by $\mathrm{diag}(\pi^0) - \pi^0(\pi^0)^{\mathrm{T}}$, it appears that a similar phenomenon is at work here.

The main idea behind our proof is to exploit the invariance of the Kullback–Leibler divergence under bijective transformations and transfer the domain of the

problem from $\mathbb{R}^d$ to $\mathcal{S}^d$. Since an $\mathcal{LD}(\beta)$ distribution is obtained from a Dirichlet $\mathcal{D}(\beta)$ distribution via the inverse log-ratio transform ℓ^{-1}, the problem of finding the best Gaussian approximation to $\mathcal{LD}(\beta)$ is equivalent to finding the best approximation to $\mathcal{D}(\beta)$ among a class of distributions obtained by applying the logistic transform to Gaussian random variables. If $\theta \sim N(\mu, \Sigma)$, the induced distribution on $\pi = \ell(\theta)$ is called a logistic normal distribution – denoted $\mathcal{L}(\mu, \Sigma)$ – and has density on $\mathcal{S}^d$ given by

$$\begin{aligned}\widetilde{q}(\pi; \mu, \Sigma) =& (2\pi)^{-d/2}|\Sigma|^{-1/2}\left(\prod_{j=0}^{d}\pi_j\right)^{-1} \\ & \times \exp\left[-\frac{1}{2}\{\log(\pi/\pi_0) - \mu\}^{\mathrm{T}}\Sigma^{-1}\{\log(\pi/\pi_0) - \mu\}\right]. \end{aligned} \tag{4.11}$$

The problem therefore boils down to calculating the Kullback–Leibler divergence between a Dirichlet density $q(\cdot; \beta)$ and a logistic normal density $\widetilde{q}(\cdot; \mu, \Sigma)$ and optimizing the expression with respect to μ and Σ. The details are deferred to the Appendix.

Once the approximation is derived in the identity parametrization, we appeal to the invariance of the Kullback–Leibler divergence under one-to-one transformations to obtain the corresponding approximation in a non-identity parameterization $\theta = X\theta^*$ as in (4.5) for any non-singular X. The result is stated below.

Corollary 4.3.2. *If $\theta \sim \mathcal{LD}(\beta)$ then*

$$D\left(\mathcal{P}_X(\cdot; \beta) \;||\; \mathcal{N}(\cdot; X\mu^*, X^T\Sigma^*X)\right) = \inf_{\mu,\Sigma} D\left(\mathcal{P}_X(\cdot; \beta) \;||\; \mathcal{N}(\cdot; \mu, \Sigma)\right) \tag{4.12}$$

for any full-rank d by d matrix X. Moreover, the bound on the KL divergence as a function of β in (4.10) is attained for $D\left(\mathcal{P}_X(\cdot; \beta) \;||\; \mathcal{N}(\cdot; \mu^*, \Sigma^*)\right)$

Thus, the best Gaussian approximation to the posterior (in the Kullback–Leibler sense) under the Diaconis–Ylviaker prior is given by $N(X\mu^*, X'\Sigma^*X)$ for any one-

to-one linear transformation X. We refer to this as the optimal Normal (oN) approximation.

4.4 Simulations

We conducted several simulation studies to assess the performance of the approximation in Theorem 4.3.1 and Corollary 4.3.2. In each study, we simulated 100 realizations from

$$\pi \sim \mathcal{D}(a, \ldots, a), \quad y \sim \text{Multinomial}\,(N, \pi)\,, \tag{4.13}$$

with the posterior of π under a Dirichlet $\mathcal{D}(a, \ldots, a)$ prior being $\mathcal{D}(y_1 + a, \ldots, y_d + a)$. We chose the dimension d to be 2^8, corresponding to a $p = 8$-way contingency table for binary variables. To obtain a simulation-based approximation to the posterior for $\theta = \log(\pi/\pi_0)$ under the Diaconis–Ylvisaker prior, we sampled mc many π values from the $\mathcal{D}(y_1 + a, \ldots, y_d + a)$ posterior and then transformed to $\theta = \ell^{-1}(\pi)$ to obtain posterior samples of θ; we refer to this procedure as the Monte Carlo approximation. We also computed a Laplace approximation to the posterior under the Diaconis–Ylvisaker prior, which is given by Normal $\left(\hat{\theta}_{MAP}, \mathcal{I}(\hat{\theta}_{MAP})^{-1}\right)$, where $\hat{\theta}_{MAP}$ is the *maximum a-posteriori* estimate of θ and $\mathcal{I}(\theta)$ is the Fisher information matrix evaluated at θ. The maximum a-posteriori estimate $\hat{\theta}_{MAP}$ was computed by the Newton–Raphson method.

We compare the accuracy of the proposed Gaussian approximation to the Monte Carlo procedure and the Laplace approximation. In addition to the identity parameterization, i.e., $X = \mathrm{I}_d$ in (4.5), we also consider the corner parameterization given by $\log(\pi/\pi_0) = X\theta^*$ for an appropriate X matrix; see Appendix for more details. For the Monte Carlo samples, each sample of θ is transformed to θ^* via $X^{-1}\theta = \theta^*$. For the normal approximations $\theta \sim \text{Normal}\,(\mu, \Sigma)$, the corresponding approximate posterior is given by $\theta^* \sim \text{Normal}\,(X^{-1}\mu, X^{-1}\Sigma X^{-1})$.

We conduct simulations for different values of N (250, 1000, and 10,000) and a (1 and $1/d$). We then assess performance in several ways.

- Proportion of variation unexplained, measured by $\sqrt{\sum_{j=1}^{d}(\theta-\theta_0)^2}/\mathrm{sd}(\theta_0)$, where θ_0 is the true value of θ (or θ^*, as appropriate).
- Coverage of 95 percent posterior credible intervals for θ or θ^*.
- The standardized loss in the Frobenius norm for estimates of Σ, the posterior covariance, given by $||\widehat{\Sigma}-\Sigma||_F/||\Sigma||_F$, where $||S||_F$ is the Frobenius norm of S. Note that the covariance in Theorem 4.3.1 is exactly the posterior covariance, so this measure is computed only for the simulation and Laplace approximations.
- The value of the Kolmogorov-Smirnov statistic for comparing the Monte Carlo empirical measure $\frac{1}{mc}\sum_{t=1}^{mc}\delta_{\theta_t}$ to the normal approximation from Theorem 4.3.1, Normal (μ,Σ).
- The computation time required to compute each posterior approximation.

Table 4.1 shows unexplained variation for the Laplace approximation, the Monte Carlo approximation for $mc = 10^3, 10^4, 10^5$, and 10^6, and the optimal normal approximation. As expected, the optimal normal approximation outperforms the Laplace approximation. Moreover, it is comparable to the Monte Carlo approximation at every sample size and for all of the values of mc considered. Performance for all approximations is noticeably better in the corner parametrization than the identity parametrization.

Table 4.2 shows coverage of approximate 95 percent credible intervals for the Laplace approximation, optimal Normal approximation, and the Monte Carlo approximation. The intervals derived using the Laplace approximation are universally too wide. Nominal coverage for the Monte Carlo approximation is insensitive to the

Table 4.1: $\sqrt{\sum_{j=1}^{d}(\theta-\theta_0)^2}/sd(\theta_0)$ *for different values of mc, different sample sizes, and two parametrizations. Results are averaged over 100 replicate simulations for each sample size.*

	Laplace	$mc = 10^3$	$mc = 10^4$	$mc = 10^5$	$mc = 10^6$	oN
identity, N=250	1.08	0.98	0.98	0.98	0.98	0.98
corner, N=250	0.85	0.81	0.81	0.81	0.81	0.81
identity, N=1000	0.84	0.77	0.77	0.77	0.77	0.77
corner, N=1000	0.67	0.61	0.61	0.61	0.61	0.61
identity, N=10,000	0.40	0.35	0.35	0.35	0.35	0.35
corner, N=10,000	0.31	0.27	0.27	0.27	0.27	0.27

value of mc in the range tested, and is slightly high at the two smaller sample sizes. The optimal normal approximation has the best coverage; in all cases it is between 0.94 and 0.96 and for $N = 10,000$ the coverage is 0.95 in both parametrizations.

Table 4.2: coverage of 95% posterior credible intervals

	Laplace	$mc = 10^3$	$mc = 10^4$	$mc = 10^5$	$mc = 10^6$	oN
identity, N=250	0.95	0.97	0.97	0.97	0.97	0.96
corner, N=250	1.00	0.96	0.96	0.96	0.96	0.96
identity, N=1000	0.98	0.96	0.96	0.96	0.96	0.96
corner, N=1000	1.00	0.94	0.94	0.94	0.94	0.94
identity, N=10,000	1.00	0.95	0.95	0.95	0.95	0.95
corner, N=10,000	1.00	0.95	0.95	0.95	0.95	0.95

Table 4.3 shows dependence of $||\widehat{\Sigma} - \Sigma||_F/||\Sigma||_F$ on mc for the two different parametrizations and three sample sizes considered. Note that Σ is known exactly since $\Sigma = \Sigma^*$, the posterior covariance under the DY prior. The main point of this table is to demonstrate the relatively large number of Monte Carlo samples required to obtain reasonably small error in estimation of the posterior covariance. Even with 10^5 samples the relative error is on the 1 percent range. Thus, compound linear hypothesis testing and computation of credible regions is very inefficient using the Monte Carlo method.

Table 4.4 shows the computation time in seconds for each of the three approximations. The Laplace approximation is fast, requiring about 0.03-0.04 seconds to

Table 4.3: $||\widehat{\Sigma} - \Sigma||_F / ||\Sigma||_F$ *for different sample sizes and values of mc*

	$mc = 10^3$	$mc = 10^4$	$mc = 10^5$	$mc = 10^6$
identity, N=250	0.0982	0.0328	0.0093	0.0032
corner, N=250	0.0923	0.0290	0.0086	0.0029
identity, N=1000	0.1045	0.0330	0.0103	0.0035
corner, N=1000	0.0882	0.0277	0.0087	0.0029
identity, N=10,000	0.1231	0.0397	0.0118	0.0040
corner, N=10,000	0.0861	0.0280	0.0084	0.0027

compute at all sample sizes. The optimal normal approximation is about an order of magnitude faster, with the computation time arising mainly in computing the polygamma functions and matrix multiplications. The Monte Carlo approximation is about four orders of magnitude slower than the optimal Normal approximation. Here, only $mc = 10^6$ is considered because of the non-negligible error in the posterior covariance for smaller samples; the algorithm scales linearly in mc so for $mc = 10^5$ the required time would be approximately 3 seconds. Only about 100 samples could be obtained in the 0.003 seconds required to compute the optimal normal approximation.

Table 4.4: Average time (seconds) to compute each approximation, averaged over 100 replicate simulations for each sample size.

	Laplace	$mc = 10^6$	oN
N=250	0.037	32.652	0.003
N=1000	0.031	31.980	0.003
N=10,000	0.035	32.338	0.003

Results in the previous tables make clear that the optimal normal approximation is superior to the other approximations considered in terms of point estimation, estimation of 95 percent credible intervals, covariance estimation, and computation time. However, it is possible that differences between the optimal normal approximation and the exact posterior exist in the tails of the distribution. To assess this, we compare the empirical measure of the Monte Carlo approximation using $mc = 10^6$ samples to the optimal normal approximation by computing the Kolmogorov-Smirnov

(KS) statistic for the marginal distributions of 20 randomly selected entries of θ. The entries considered were re-selected for each of the 100 replicate simulations and for each of the three sample sizes. Shown in Figure 4.1 are histograms of these KS statistics in the corner and identity parametrizations. Most are less than 0.02, and none are greater than 0.07. Considering that the KS statistic is a point estimate of the total variation distance between distributions, this indicates that the optimal normal approximation is an excellent approximation to the posterior marginals. Moreover, we cannot rule out the possibility of residual Monte Carlo error in the marginals from the Monte Carlo approximation, which may account for part of the observed discrepancy.

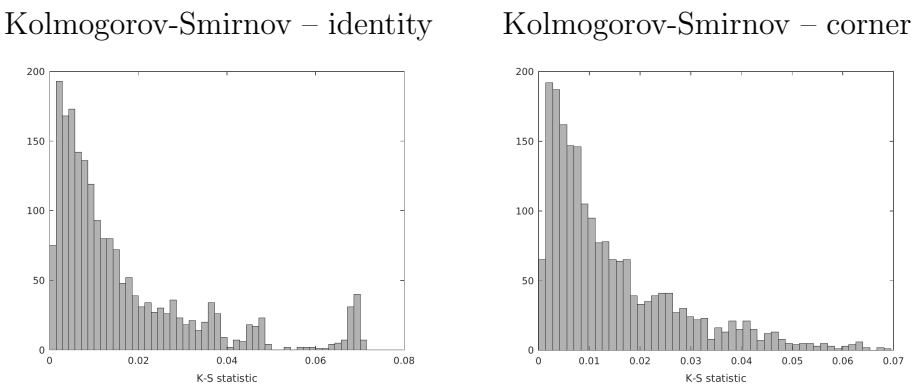


FIGURE 4.1*: Distribution of Kolmogorov-Smirnov statistics comparing* $\frac{1}{mc}\sum_{t=1}^{mc}\delta_{\theta_t}$ *to the oN approximation for 20 randomly selected entries of* θ *and over 100 replicate simulations (entries of* θ *were re-selected for each replicate).*

4.5 Real Data Example

We consider the Rochdale data, a 2^8 contingency table with $N = 665$ that is over 50 percent sparse, and for which the top ten cell counts all exceed 20. This dataset is described at length in Dobra and Lenkoski (2011). We first assess the accuracy of the approximation to the full posterior under the Diaconis–Ylvisaker prior in the same manner as in §4.4, by comparing marginal posteriors computed using the approxi-

mation to those obtained from large Monte Carlo samples from the exact Dirichlet posterior transformed to the log-linear parametrization. For the log-linear model in the corner parametrization, the distribution of Kolmogorov-Smirnov statistics computed for the 255 entries of θ^* obtained by comparing 10^6 Monte Carlo samples from the exact posterior to the optimal Gaussian approximation is shown in Fig. 4.2. The distribution is very similar to that observed for the simulations in §4.4.

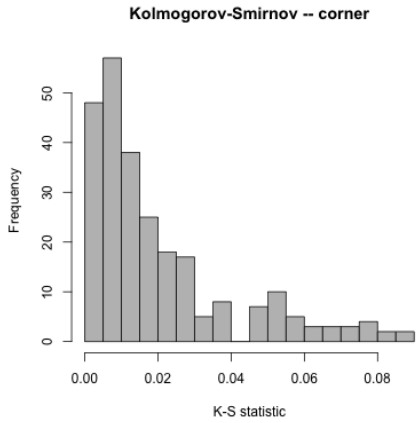


FIGURE 4.2: *Histogram of Kolmogorov-Smirnov statistics for the comparison of* 10^6 *Monte Carlo samples from the exact Dirichlet posterior, transformed to* θ^**, to the optimal Gaussian approximation to the posterior for* θ^* *under the Diaconis–Ylvisaker prior.*

Undoubtedly, the Diaconis–Ylvisaker prior is less well-suited to inference on important variable interactions in this dataset than the more sophisticated methods of Dobra and Lenkoski (2011) and Bhattacharya and Dunson (2012). However, our approximation has the advantage of being essentially computation-free, whereas the methods of Dobra and Lenkoski (2011) and Bhattacharya and Dunson (2012) are computationally intensive even at this small scale. In many settings, particularly with modern large-scale problems, some loss of performance may be acceptable in order to obtain useful inferences instantaneously. Thus, we are interested in the extent to which our method can replicate the results of Dobra and Lenkoski (2011), which were similar to those of Bhattacharya and Dunson (2012) in many respects.

We analyze performance in testing conditional independence hypotheses (i.e. learning an interaction graph).

Sparse θ^* is a set of measure zero with respect to the posterior under the Diaconis–Ylvisaker prior. To obtain a sparse point estimate of the interaction graph, we employ the penalized credible region approach of Bondell and Reich (2012). This method produces a point estimate by finding the sparsest θ^* within a $1-\alpha$ credible region for θ^*. Although the exact solution to this problem is intractable, Bondell and Reich (2012) show that it can be approximated using a lasso path, and provide software in the `BayesPen R` package (Wilson et al., 2015). Using the resulting lasso path from `BayesPen`, the selected model corresponding to any value of $\alpha \in (0, 1)$ can be obtained as follows.

1. For the selected value of α, find the $1-\alpha$ quantile of a χ^2 distribution with $d-1$ degrees of freedom. Label this $\delta_{\max}$.

2. For each model θ_0 in the Lasso path, compute the Mahalanobis distance $\delta(\theta_0) = (\theta^* - \theta_0)^T(\Sigma^*)^{-1}(\theta^* - \theta_0)$.

3. Find the sparsest model in the lasso path having $\delta(\theta_0) \leq \delta_{\max}$. This is the sparse point estimate.

With 256 cells and 665 observations, the posterior under the saturated model with Diaconis–Ylvisaker prior is very diffuse. To make a reasonable comparison, we obtain the posterior under the Diaconis–Ylvisaker prior for the marginal tables corresponding to all $\binom{8}{4} = 70$ unique subsets of four variables. For each of these marginal tables, we then utilize the penalized credible region procedure of Bondell and Reich (2012) to obtain a sparse model. For comparison, we utilize the median probability graphical model from Dobra and Lenkoski (2011), which is shown in Table 4.5. Specifically, for every subset of four variables, we obtain the marginal graph corresponding to

Table 4.5: Left, titled CGGM Results: Marginal posterior inclusion probabilities of edges (above the main diagonal) and indicator of edge inclusion in the median probability model (below the main diagonal) from copula Gaussian graphical model estimated on Rochdale data in Dobra and Lenkoski (2011). Rows and columns correspond to the eight binary variables, which are labeled a-h. Right, titled Comparison to oN: table of edge classifications for all marginal tables of size 2^4 from copula Gaussian graphical model median probability model (columns, labeled CGGM) and penalized credible region for Gaussian approximation to posterior under the DY prior (rows, labeled oN-PCR).

CGGM Results

	a	b	c	d	e	f	g	h
a	–	0.93	0.67	0.92	0.32	0.42	1	0.26
b	1	–	0.27	1	0.88	0.29	0.70	0.96
c	1	0	–	0.29	0.91	0.35	0.85	0.25
d	1	1	0	–	0.37	0.59	0.66	0.50
e	0	1	1	0	–	0.98	0.58	0.17
f	0	0	0	1	1	–	0.82	0.22
g	1	1	1	1	1	1	–	0.32
h	0	1	0	1	0	0	0	–

Comparison to oN

		CGGM	
		0	1
oN-PCR	0	4	74
	1	7	335

the median probability model of Dobra and Lenkoski (2011) by removing the complement of the subset of nodes under consideration and moralizing, i.e. placing an edge between nodes that (1) have an edge between them in the full graph or (2) are connected solely by a path through nodes that were removed. We treat the graph obtained in this way as the standard for assessing performance of the penalized credible region applied to our Gaussian posterior approximation.

We compute the true (false) negative and positive counts for the penalized credible region procedure applied to our posterior Gausian approximation to all 70 marginal graphs, treating the corresponding marginal median probability graph from Dobra and Lenkoski (2011) as the truth. This produces a total of $70\binom{4}{2} = 420$ dependent pseudo hypothesis tests. The results for $\alpha = 0.1$ in the penalized credible region procedure are shown in Table 4.5. We obtain a false discovery rate of 0.02, and an F_1 score of 0.89, indicating that for marginal tables of size 2^4, the posterior approximation is useful for model selection on the Rochdale data.

4.6 Discussion

Outside of linear models, conjugate priors are often non-standard or their multivariate generalizations are difficult to work with. This hampers uncertainty quantification because it is difficult to obtain posterior credible regions for parameters under such priors. Given that automatic and coherent quantification of uncertainty through the posterior is one of the chief advantages of a fully Bayesian approach, this limitation is a significant problem. The optimal Gaussian approximation to the posterior for log-linear models with Dianconis-Ylvisaker conjugate priors derived here offers a highly accurate and essentially computation-free approximation to posterior credible regions for this important class of models. Interestingly, this Gaussian approximation is not the Laplace approximation, and it is faster to compute and offers a better approximation to the posterior than the Laplace approximation. If similar results could be obtained for the posterior in other models, it suggests that the Laplace approximation may not be an appropriate default Gaussian approximation to the posterior. The theoretical result provided here can be easily extended to cases where some categories cannot co-occur, i.e. cases of structural zeros in contingency tables. Extensions to model selection using our approximation are also available by the penalized credible region approach. It seems reasonable that the strategy used here to obtain optimality and convergence rate guarantees could be extended to a larger class of generalized linear models by studying the properties of multivariate Gaussian distributions under inverse link transformations. This may also present a strategy for obtaining approximate credible intervals for parameters in the Bayesian model averaging context for generalized linear models with conjugate priors.

5

Tail waiting times and the extremes of stochastic processes

5.1 Introduction

The standard model of extremes of stochastic processes is the max-stable process (De Haan (1984), Beirlant et al. (2006), Coles et al. (2001)). A process $Y(x), x \in \mathcal{X}$ for an index set $\mathcal{X}$ is max-stable if there exist sequences $a_n(x), b_n(x)$ and a process $w(x), x \in \mathcal{X}$ such that for every finite collection of points x, we have

$$Y(x) = \lim_{n\to\infty} \frac{[\bigvee_{i=1}^{n} w_i(x)] - a_n(x)}{b_n(x)}, \tag{5.1}$$

where $\{w_i\}_{i\leq n}$ are independent copies of $w(x)$ (De Haan and Ferreira (2007), Beirlant et al. (2006), Schlather (2002)). In the spatial or spatiotemporal setting, one usually takes $\mathcal{X} = \mathbb{R}^d$ for some integer d. This model is quite general in the sense that *if* there exist $\mathcal{X}$-indexed sequences $a_n(x)$ and $b_n(x)$ such that normalized maxima of the form (5.1) converge, then the limit must be a max-stable process De Haan and Ferreira (2007).

In applications where multivariate or spatial extremes are of interest, it is almost always the case that one has a set of observations $w(\boldsymbol{x}, t) = (w(x_1, t), \ldots, w(x_n, t))$ on

a stochastic process $\{W(x,t)\}$ at a collection of points $x_1, \ldots, x_n$ and times $t_1, \ldots, t_p$. These observations could represent hourly precipitation, maximum daily wind speed, or, if we treat the spatial index set $\mathcal{X}$ as a latent coordinate in an attribute space, essentially any multivariate time series, such as daily stock prices.

There are two competing paradigms for statistical inference on max-stable processes based on multivariate time series. The first is the max-over-windows (MOW) approach, which collapses observations $w(\boldsymbol{x},t)$ on a space/time-indexed process $\{W(x,t)\}$ to observations that are (approximately) from a max-stable process $Y(x)$ by selecting a time window Δt, setting $s_0 = t_0, s_j = t_j + \Delta t$, and putting $y_j(\boldsymbol{x}) = \max_{s_j \leq t < s_{j+1}} w(\boldsymbol{x},t)$ (Tawn (1988), Tawn (1990)). The data $y_j(\boldsymbol{x}), j \in 1, \ldots, \lfloor t_p/\Delta t \rfloor$ are then used for estimation and inference. An alternative is the peaks-over-thresholds (POT) approach (Davison and Smith (1990), Smith (1984)), which keeps only observations $w(\boldsymbol{x},t)$ that exceed some threshold, then treats these observations as realizations of a max-stable process. Various ways of thresholding have been proposed, including: keeping observations at times t where $\max_{i=1}^{n} w(x_i,t)$ exceeds a pre-specified threshold (Rootzén and Tajvidi (2006), Buishand et al. (2008)); fixing a specific component (say, x_1) and keeping observations at times t where $w(x_1,t)$ exceeds a threshold (Heffernan and Tawn (2004), Heffernan and Resnick (2007), Das et al. (2011), Balkema and Embrechts (2007)); and, keeping all observations at times t where the norm of the vector $||w(\boldsymbol{x},t)||$ exceeds a threshold (Coles and Tawn (1991), Ballani and Schlather (2011)).

Extreme events often cluster temporally; they tend to occur nearby in time but not always at identical times. Temporal structure can be accounted for in the context of both approaches in a variety of ways. Perhaps the most common way is choosing window size in a MOW approach such that dependence across consecutive time windows is probably small. Often, clustering of extreme observations is used to select window lengths (Zhang and Smith (2010), Meinguet (2012)). A common model for

temporally dependent spatial extremes that fits broadly into the MOW context is the maxima of moving maxima (M4) model (Smith and Weissman (1996),Heffernan et al. (2007)). However, MOW methods have the drawback that they tend to discard a larger portion of the data than POT and artifacts can arise from the arbitrary location of the time window boundaries. Estimation procedures for these models also tend to be complex and multi-stage, resulting in potential loss of efficiency.

Using a POT approach, if extremes at two locations are dependent but temporally lagged, temporal structure must be introduced into the model (see e.g. Smith (1989), Méndez et al. (2006), Katz et al. (2002), Beguería and Vicente-Serrano (2006), Beguería et al. (2011)). With a generic POT approach that treats the observations which exceed a threshold as approximately independent realizations of a max-stable process, if exceedances do not occur contemporaneously (or at time lags much shorter than the sampling rate), then dependence information is lost. An illustrative example is shown in Figure 5.1: extreme values of the variable y_1 are always followed by extreme values at y_2 exactly two time units later (y_2 also has some extreme events that occur randomly). Two settings in which this aspect of POT methods is obviously a problem are when the exceedance at x_1 is causal, or when x_1 and x_2 are arranged spatially such that extreme events tend to occur at x_1 and then the phenomenon causing the exceedance (e.g. a weather event) subsequently moves to x_2. While this can be addressed by introducing temporal dependence in parameters, this results in more complex modeling, places more demands on usually limited data, and can complicate selection of an appropriate threshold.

The main contribution of this work is to propose an alternative paradigm for inference on tail dependence between spatial locations with arbitrary temporal structure. The basic motivation is that waiting times between threshold exceedances at different sites contain information on extremal dependence across both space and time. To provide a canonical setting in which to study properties of this approach,

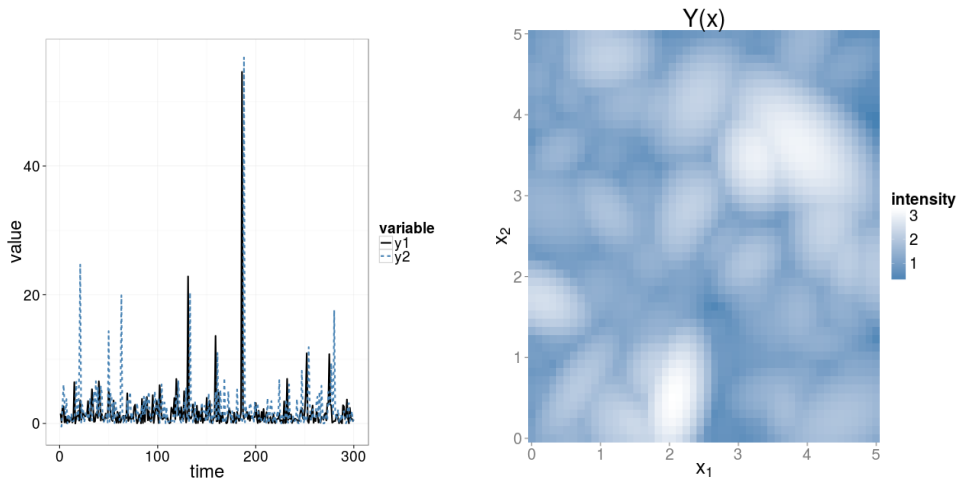


FIGURE 5.1*: Left panel: example of two time series in which large values of one (y_1) are followed by large values of the other (y_2) two time units later. Right panel: a realization of a Gaussian max stable process; blue indicates small values, white large values.*

we propose a heavy-tailed stochastic process with spatiotemporal dynamics, so that temporal dependence in extreme events can be accommodated naturally. To provide intuition for our proposed model, recall that the process $Z(x) = \sup_j u_j k(x, \xi_j)$ where $k : \mathcal{X} \times X \to \mathbb{R}$ is a nonnegative kernel, $p(u) \propto u^{-2}$, and ξ are points of a homogeneous Poisson process, is a max-stable process (Schlather (2002)). A single realization from $Z(x)$ where $k(x, \xi)$ is a bivariate Gaussian kernel is shown in the right panel of Figure 5.1: kernels associated with large values of u_j are clearly discernable. The new stochastic process proposed here endows the points ξ_j with velocities and lifetimes, making the process dynamic and inducing a temporal dependence strucuture. As a consequence, time-indexed multivariate data in this paradigm are viewed as a single realization from a *max-stable velocity process*, rather than multiple (possibly dependent) realizations of a max-stable process. The resulting process can be visualized as a dynamic version of the sample realization in Figure 5.1 in which the elliptical light regions appear at some time, move around the space, and then disappear.

Fitting the max-stable velocity process to data is challenging and computation-

ally intensive. The other main contribution of this work is to propose a method for inference on extreme dependence that does not require modeling the process explicitly. Instead, we model waiting times between thresholds exceedances at pairs of locations, so that inference can be done using only the exceedance time data. We show that if the data are generated by a max-stable velocity process, then the waiting times between exceedances are well-approximated by mixtures of exponential random variables and an atom at zero. Further, the waiting time distribution under independence is also a mixture of exponentials and a point mass, but with parameters that depend on the marginal waiting times at each location. By fitting these two models to data and performing model comparison, we obtain a measure of extreme dependence that naturally captures time lags and exposes differences in the temporal structure of dependence across space. While we focus on inference in the setting where the data are thought to originate from a max-stable velocity process, the use of waiting times to model extreme dependence is general and can be applied in settings where the max-stable velocity process is considered inadequate by choosing a fully nonparametric model for the waiting times.

The remainder of this paper is organized as follows. Section 2 introduces a stochastic process with explicit time indexing that is conditionally max-stable and derives a number of its properties, including the distribution of waiting times between exceedances of a threshold. In section 3 a new tail dependence measure is introduced, and an approach to inference based on waiting times between peaks over thresholds suggested. Section 4 provides a model for the waiting times between exceedances in the process presented in section 3, as well as an algorithm for computation. Section 5 presents a simulation study motivated by weather extremes. Section 6 applies the method to four real datasets. Section 7 concludes.

5.2 Model

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $\mathcal{F}_t, t \in T$ a filtration; without loss of generality we will take $T = \mathbb{R}$. Let $W(x,t)$ be a stochastic process indexed by $\mathcal{X}$ adapted to $\mathcal{F}_t$ taking values in a Polish space $(\mathcal{W}, d_W)$, and $w(\boldsymbol{x}, t), t \in t_0, \ldots, t_p$ a collection of observations on W at points $\boldsymbol{x} \in \mathcal{X}^n$. We will assume that $W(x,t)$ is in the max domain-of-attraction of a max-stable process $Y(x)$, i.e. that there exist sequences $a_n(x), b_n(x)$ and a max-stable process $Y(x)$ for which (5.1) holds for $W(x,t_0)$ for any $t_0 \in T$. Further, we assume that $W(x,t)$ is a Markov process, so that for $\mu(W^{-1})$-measurable sets A, $\mathbb{P}[W(x,t) \in A \mid \mathcal{F}_s] = \mathbb{P}[W(x,t) \in A \mid W(x,s)]$.

Because W is in the max domain-of-attraction of Y, observations $w(\boldsymbol{x}, t)$ that are "extreme" in some sense should be approximately observations from Y, and so should $\max_{0<s<t} w(\boldsymbol{x}, s)$. However, $w(\boldsymbol{x}, t)$ are not *i.i.d.* observations from a stochastic process $W(x)$, but dependent observations from a Markov process $W(x,t)$. As such, extreme observations of $w(\boldsymbol{x}, t)$ and $\max_{0<s<t} w(\boldsymbol{x}, s)$ are approximately observations from a Markov process $Y(x,t)$ with the property that for any t_0, $Y(x,t_0)$ is max-stable. In this section we construct the max-stable velocity process, which fits this requirement, and thus provides a model for the extremes of a Markov process.

5.2.1 Max-stable velocity processes

The max-stable velocity process is constructed from the spectral characterization originally proposed in De Haan (1984), which is given in Theorem 5.2.1.

Theorem 5.2.1 (de Haan)**.** *Let $\{(u_j, \xi_j)\}_{j \geq 1}$ be the points of a Poisson process on $(0, \infty] \times \mathbb{R}^d$ with intensity measure $u^{-2}\, du\, \nu(d\xi)$ for some σ-finite measure ν on $\mathbb{R}^d$. Let $\{Y(x)\}_{x \in \mathbb{R}^d}$ be a max-stable process with unit Fréchet margins and continuous sample paths. Then there exist nonnegative continuous functions $\{k(x,y) : x, y \in \mathbb{R}^d\}$*

such that

$$\int_{\mathbb{R}^d} k(x,y)\nu(dy) = 1 \qquad \forall x \in \mathbb{R}^d$$

for which

$$\{Y(x)\}_{x\in\mathbb{R}^d} \stackrel{\mathcal{D}}{=} \left\{\sup_{j\geq 1} u_j k(x,\xi_j)\right\}_{x\in\mathbb{R}^d}, \tag{5.2}$$

where $\stackrel{\mathcal{D}}{=}$ *denotes equality in distribution. Moreover, any process defined by the right side of (5.2) is max-stable.*

A useful heuristic for de Haan's spectral characterization is that of weather extremes, where the locations of the support points are taken to be storm centers, the kernel functions $k(x,.)$ describe the shape of the storm, and the marks u the severity. In this context, the process realization is the maximum over some period of time of a climatological quantity, such as precipitation or temperature. To create a time-indexed process, we endow the points ξ_j with lifetimes and velocities; this approach has the advantage of easily extending the physical interpretation of the points ξ_j as storms or, more generally "events." Now, the storms will move and have finite lifespans.

Specifically, let $\mathcal{N} \sim \text{Po}(\beta u^{-2} du\ d\xi\ ds\ \pi(da))$ be a Poisson random field on $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ \times \mathcal{A}$ for some probability measure $\pi(da)$ on an "attribute space" $\mathcal{A}$, and let $k:\ \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}_+$ be a positive-definite function $k(x,\xi;a,s,t)$ satisfying an integrability condition discussed below. Using the climatological heuristic, the support points (ξ_j, s_j, u_j, a_j) of $\mathcal{N}(du\, d\xi\, ds\, da)$ can be thought of as storms of magnitude $u_j > 0$, initiating at time $s_j \in \mathbb{R}$ and location $\xi_j \in \mathbb{R}^d$, with attribute vector a_j that may include a velocity v_j, shape Λ_j, duration τ_j, or other features.

For any location $x \in \mathbb{R}^d$, define

$$Y(x,t) := \sup_j \{u_j k(x,\xi_j;a_j,s_j,t)\} \qquad t \in \mathbb{R} \tag{5.3}$$

$$Y^*(x,t) := \sup_{0\le s\le t} Y(x,t), \qquad\qquad t \ge 0.$$

We refer to (5.3) as a max-stable velocity process.

In the sequel, we will assume that $a_j = (v_j, \Lambda_j, \tau_j)$, and that the intensity measure for the process is given by $\beta u^{-2} du\, d\xi\, ds\, \delta e^{-\delta\tau}\, \pi(dv\, d\Lambda)$, so that the "storms" have exponential lifetimes, the birth rate is time-invariant, and $\nu(d\xi)$ is Lebesgue measure. We also assume v, Λ have finite expectation with respect to the probability measure π. The initiation time s_j, the time index t, and the velocity v_j will enter the kernel as the location of the point j at time t, giving $Y(x,t) := \sup_j u_j k(x, \xi_j + v_j(t-s_j); \Lambda_j) \mathbb{1}_{\{t-\tau_j<s_j<t\}}$. We also require the kernel to be stationary so that $k(x,y) = k(0, x-y)$ with $\int_{\xi\in\mathbb{R}} k^*(0,\xi;\Lambda)d\xi = 1$. Many of the results that follow would actually hold for other choices of $\nu(d\xi)$ and larger families of kernels, but the required integrability condition is more complex. To aid interpretation, we provide some parallel results in the special case where k is the isotropic Gaussian kernel, $k(x,\xi) = |\Lambda/2\pi|^{1/2}\exp\{-(x-\xi)'\Lambda(x-\xi)/2\}$, which was originally proposed as a model for the max-stable process by Smith (1990).

5.2.2 Main results

Since the magnitudes u_j have the same distribution in the max-stable velocity process as in (5.1), tail dependence remains intact. This is formalized in Theorem 5.2.2. Proofs are found in Appendix D.

Theorem 5.2.2 (Max-stable velocity process)**.** *Let $\{Y(x,t)\}_{x\in\mathbb{R}^d, t\in\mathbb{R}}$ be a max-stable velocity process. Then, for any $B \subset \mathbb{R}^d$, the number of points that $\mathcal{N}(d\xi\, dt\, du\, da)$ places in B at time t is a spatial Poisson process with time-invariant intensity $\frac{\beta}{\delta}$. Furthermore, when $\pi(d\Lambda)$ is an atom, then for any $t \in \mathbb{R}$, $\{Y(x,t)\}_{x\in\mathbb{R}^d}$ is max-stable and the corresponding Poisson process has intensity measure $\frac{\beta}{\delta} d\xi u^{-2} du$. If additionally, $\pi(dv) = \delta_0$, then $Y^*(x,t)$ is max-stable, and the corresponding Poisson*

process representation has intensity measure $\beta\left[t+\frac{1}{\delta}\right]d\xi u^{-2}du$.

Theorem 5.2.2 implies that for any fixed t, the distribution of the process $Y(x,t)$ does not depend on velocity. The requirement that $\pi(dv)=\delta_0$ to show that $Y^*(x,t)$ is max-stable suggests that $Y^*(x,t)$ does depend on $\pi(dv)$, an intuitive conclusion that will shortly become clear.

In addition to the marginal and joint distribution of the max-stable velocity process, we are also interested in waiting times, as these will eventually form the basis for inference. For two arbitrary points $(x_1,x_2)\in\mathcal{X}^2$, fix a threshold vector (y_1,y_2). Now define the random variables

$$\kappa_i(y_i)=\inf_{t>0}\{t:Y(x_i,t)>y_i\},\qquad i=1,2$$

$$\kappa_{(1,2)}(y_1,y_2)=|\kappa_1(y_1)-\kappa_2(y_2)|. \tag{5.4}$$

Then for locations $(x_1,x_2)\in\mathcal{X}\times\mathcal{X}$, $\kappa_{(1,2)}(y_1,y_2)$ is the waiting time between first exceedances of y_1 at x_1 and y_2 at x_2. Theorem 5.2.3 gives the marginal distribution of a max-stable velocity process and that of the waiting times until first exceedance.

Theorem 5.2.3 (Marginals and waiting times)**.** *The max-stable velocity process has Fréchet marginals, with distribution function* $\mathbb{P}[Y(x,0)<y]=e^{-\beta/(\delta y)}$. *When* $\pi(dv)=\delta_0$, *the waiting time until first exceedance* $\kappa(y)$ *has survival function* $\mathbb{P}[\kappa(y)>t]=e^{-\beta(t+\delta^{-1})/y}$. *When* $\pi(dv)\neq\delta_0$, $\mathbb{P}[\kappa(y)>t]=e^{-\beta(\delta^{-1}+t+f(t))/y}$ *for a positive, monotone nondecreasing function* $f(t)$. *For isotropic* k, $f(t)<t\mathbb{E}_{\pi(dv,d\Lambda)}\left[\frac{c_d(\Lambda)}{c_{d-1}(\Lambda)}|v|\right]$ *by Lemma D.1.1, where* $\frac{c_d(\Lambda)}{c_{d-1}(\Lambda)}$ *are functions of* Λ *depending on the spatial dimension* d.

So $\kappa(y)$ is zero with probability $\mathbb{P}[Y(x,0)>y]$ given by theorem 5.2.3, and otherwise it is stochastically dominated by an exponential random variable with rate

β/y. The weight assigned to the atom is given by $1 - e^{-\beta/y\delta}$, which approaches zero as $y \to \infty$. Notably, the waiting time distribution, like the distribution of $Y^*(x,t)$ but unlike the distribution of $Y(x,t)$ for fixed t, depends on velocity and shape. Theorem 5.2.4 gives the joint distribution of the max-stable velocity process at two points and different times t_1, t_2.

Theorem 5.2.4 (Joint distribution)**.** *For any two points x_1, x_2 and times t_1, t_2 the joint distribution of the max-stable velocity process is given by $\mathbb{P}[Y(x_1,t_1) \leq y_1,\ Y(x_2,t_2) \leq y_2] = \exp(-|B|)$ with*

$$|B| = \frac{\beta/\delta}{y_1}\left[1 - e^{-\delta|t_2-t_1|}\int F(\Lambda, \Delta(v); y_1, y_2)\,\pi(dvd\Lambda)\right] + \frac{\beta/\delta}{y_2}\left[1 - e^{-\delta|t_2-t_1|}\int F(\Lambda, \bar{\Delta}(v); y_2, y_1)\,\pi(dvd\Lambda)\right]. \tag{5.5}$$

where F is a function satisfying $0 \leq F \leq 1$, $\Delta(v) = x_2 - x_1 - (t_2 - t_1)v$, and $\bar{\Delta}(v) = x_1 - x_2 - (t_1 - t_2)v$. For the Gaussian max-stable velocity process, F is given by

$$F(\Lambda, \Delta(v); y_1, y_2) = \Phi\left(-\frac{\sigma(\Delta(v),\Lambda)}{2} + \frac{1}{\sigma(\Delta(v),\Lambda)}\log\frac{y_1}{y_2}\right), \tag{5.6}$$

where $\sigma^2(\Delta(v),\Lambda) = \Delta(v)'\Lambda\Delta(v)$ and $\Phi(\cdot)$ is the standard Gaussian distribution function.

The expression in (5.6) is similar to that in equation 3.1 of Smith (1990), with the additional term $e^{-\delta|t_2-t_2|}$ arising from the support point lifetimes and σ an explicit function of velocity and shape.

Theorem 5.2.5 gives an expression for the waiting times between first exceedances under independence, which can be thought of as the limit of the waiting time distribution at two points for fixed y_1, y_2 as the distance between points goes to infinity.

Theorem 5.2.5 (Waiting times between exceedances under independence)**.** *Suppose* $\kappa_1(y_1) \perp \kappa_2(y_2)$*, and set* $\lambda_1 = e^{-\beta/(y_1\delta)}$*,* $\lambda_2 = e^{-\beta/(y_2\delta)}$*. Then if* $\pi(dv) = \delta_0$

$$\begin{aligned}\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) = 0] =& (1-\lambda_1)(1-\lambda_2)\\ \mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t] =& e^{-\beta t/y_2}\left(\lambda_2(1-\lambda_1) + \frac{y_1}{y_1+y_2}\lambda_1\lambda_2\right)\\ &+ e^{-\beta t/y_1}\left(\lambda_1(1-\lambda_2) + \frac{y_2}{y_1+y_2}\lambda_1\lambda_2\right), \qquad (5.7)\end{aligned}$$

a mixture of a point mass at zero and exponential distributions with rates β/y_1 *and* β/y_2*. If* $\pi(dv) \neq \delta_0$*, then*

$$\begin{aligned}\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) = 0] &= (1 - e^{-\beta/(y_1\delta)})(1 - e^{-\beta/(y_2\delta)})\\ \mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t] &= e^{-\beta(t+f_2(t,y_2))/y_2}(1 - e^{-\beta/(y_1\delta)})e^{-\beta/(y_2\delta)}\\ &\quad + e^{-\beta(t+f_1(t,y_1))/y_1}(1 - e^{-\beta/(y_2\delta)})e^{-\beta/(y_1\delta)}\\ &\quad + \left[1 - \left(\int_{\kappa_2=t}^{\infty}(e^{-\beta(\kappa_2-t+f_1(\kappa_2-t,y_1))/y_1} - e^{-\beta(\kappa_2+t+f_1(\kappa_2+t,y_1))/y_1})\mu_2(d\kappa_2)\right.\right.\\ &\quad \left.\left. + \int_{\kappa_2=0}^{\infty}(1 - e^{-\beta(\kappa_2+t+f_1(\kappa_2+t,y_1))/y_1})\mu_2(d\kappa_2)\right)\right] \times e^{-\beta/(y_1\delta)}e^{-\beta/(y_2\delta)}\end{aligned}$$

where f_1*,* f_2 *are positive, monotone nondecreasing functions satisfying*

$$f_j(t) < t\mathbb{E}_{\pi(dv,d\Lambda)}\left[\frac{c_d(\Lambda)}{c_{d-1}(\Lambda)}|v|\right]$$

when k *is isotropic, and* $\mu_2(d\kappa_2)$ *is the measure induced by* κ_2*.*

The distribution of $\kappa_{(1,2)}(y_1, y_2)$ when $\pi(dv) \neq 0$ is therefore a mixture of four components, one of which is an atom, and two of which are dominated by an exponential random variable when k is isotropic. The last component is more opaque, but the general result in Theorem 5.2.7 shows that this component can be described by a mixture of two random variables that are stochastically dominated by exponentials. On the other hand, the distribution of $\kappa_{(1,2)}(y_1, y_2)$ under independence when

$v = 0$ almost surely is a simple mixture of two exponentials and an atom at zero. If $y_1 = y_2 = y$ and $v = 0$ almost surely, (5.7) reduces to a mixture of an atom at zero and a single exponential with rate β/y. As $y_1, y_2 \to \infty$, $\lambda_1, \lambda_2 \to 1$ in Theorem (5.7, and the distribution in (5.7) converges to the distribution with survival function given by $\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) = 0] = 0$ and

$$\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t] = e^{-\beta t/y_2} \frac{y_1}{y_1 + y_2} + e^{-\beta t/y_1} \frac{y_2}{y_1 + y_2},$$

so the atom vanishes asymptotically. The distribution of $\kappa_{(1,2)}(y_1, y_2)$ in the general model is complicated, so to aid interpretation we first obtain its distribution in the special case of $v = 0$, which we provide in Theorem 5.2.6.

Theorem 5.2.6 (waiting times between exceedances when $v = 0$)**.** *Suppose $v = 0$ with probability one. Then the distribution of $\kappa_{(1,2)}(y_1, y_2)$ is given by*

$$\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) = 0] = 1 - e^{-\beta\lambda_0/\delta} + e^{-\beta\lambda_0/\delta}(1 - e^{-\beta\lambda_1/\delta})(1 - e^{-\beta\lambda_2/\delta}) + \frac{\lambda_0 e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta}}{\lambda_0 + \lambda_1 + \lambda_2} \tag{5.8a}$$

$$\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t] = e^{-t\beta(\lambda_0+\lambda_1)} \left(e^{-\frac{\beta(\lambda_0+\lambda_1)}{\delta}} (1 - e^{-\frac{\beta\lambda_2}{\delta}}) + \frac{\lambda_2 e^{-\frac{\beta(\lambda_0+\lambda_1+\lambda_2)}{\delta}}}{\lambda_0 + \lambda_1 + \lambda_2} \right) + e^{-t\beta(\lambda_0+\lambda_2)} \left(e^{-\frac{\beta(\lambda_2+\lambda_0)}{\delta}} (1 - e^{-\frac{\beta\lambda_1}{\delta}}) + \frac{\lambda_1 e^{-\frac{\beta(\lambda_0+\lambda_1+\lambda_2)}{\delta}}}{\lambda_0 + \lambda_1 + \lambda_2} \right) \tag{5.8b}$$

where

$$\lambda_0 = \int_{\Lambda \in \mathcal{L}} \left[\frac{1}{y_1} F(\Lambda, \Delta(0); y_1, y_2) + \frac{1}{y_2} F(\Lambda, \bar{\Delta}(0); y_2, y_1) \right] \pi(d\Lambda)$$

$$\lambda_1 = \int_{\Lambda \in \mathcal{L}} \left[\frac{1}{y_1} [1 - F(\Lambda, \Delta(0); y_1, y_2)] - \frac{1}{y_2} F(\Lambda, \bar{\Delta}(0); y_2, y_1) \right] \pi(d\Lambda)$$

$$\lambda_2 = \int_{\Lambda \in \mathcal{L}} \left[\frac{1}{y_2} [1 - F(\Lambda, \bar{\Delta}(0); y_2, y_1)] - \frac{1}{y_1} F(\Lambda, \Delta(0); y_1, y_2) \right] \pi(d\Lambda),$$

with F and $\Delta(v)$ as defined in Theorem 5.2.4. The result for the Gaussian model uses the definition of F given in Theorem 5.2.4.

In the case where $y_1 = y_2 = y$ and $\lambda_1 = \lambda_2$, (5.8b), like (5.7), reduces to an atom at zero and a single exponential distribution. However, unlike (5.7), the weight assigned to the atom does not vanish as $y_1, y_2 \to \infty$; instead, the distribution approaches

$$\begin{aligned}\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) = 0] =& \frac{\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2} e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta} \\ \mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t] =& e^{-t\beta(\lambda_0+\lambda_1)} \frac{\lambda_2}{\lambda_0 + \lambda_1 + \lambda_2} + e^{-t\beta(\lambda_0+\lambda_2)} \frac{\lambda_1}{\lambda_0 + \lambda_1 + \lambda_2},\end{aligned}$$

so the weights stabilize asymptotically.

The result in Theorem 5.2.6 is relevant to the case with nonzero velocity in two ways. First, since contemporaneous exceedances occur at the birth time of the kernel almost surely even when $v \neq 0$, $\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) = 0]$ is given by (5.8a) for any velocity distribution with finite expectation. This fact is used extensively in obtaining the result in Theorem 5.2.7. Second, since all of the exceedances in the $v = 0$ case occur at birth time, nonzero velocity can only shorten the waiting time until exceedance of any threshold. As a result, (5.8b) is a stochastic lower bound for the distribution of $\kappa_{(1,2)}(y_1, y_2)$ with nonzero velocity – i.e. the survival function of $\kappa_{(1,2)}(y_1, y_2)$ is bounded above by (5.8b). This is clear from the result for the general case in Theorem 5.2.7.

Theorem 5.2.7 (Survival function with nonzero velocity)**.** *Suppose $Y(x, t)$ is a max-stable velocity process. Then for any two points x_1, x_2 and thresholds y_1, y_2, $\kappa_{(1,2)}(y_1, y_2)$ is a mixture given by*

$$\begin{aligned}\mathbb{P}[\kappa_{(1,2)} = 0] &= 1 - e^{-\beta\lambda_0/\delta} + e^{-\beta\lambda_0/\delta}(1 - e^{-\beta\lambda_1/\delta})(1 - e^{-\beta\lambda_1/\delta}) + e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta} p_0, \\ \mathbb{P}[\kappa_{(1,2)} > t] &= e^{-\beta(\lambda_0+\lambda_2)/\delta} e^{-\beta(t/y_2 + g_{12}(t,y_1,y_2) + h_{12}^{(0)}(t,y_1,y_2))} \\ &\quad + e^{-\beta(\lambda_0+\lambda_1)/\delta} e^{-\beta(t/y_1 + g_{21}(t,y_1,y_2) + h_{21}^{(0)}(t,y_1,y_2))} \\ &\quad + e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta} p_1 e^{-\beta(t/y_1 + g_{12}(t,y_1,y_2))} \left(\int_{T=0}^{\infty} e^{-h_{12}^{(+)}(t,y_1,y_2,T)} \mu_1(dT) \right)\end{aligned}$$

$$+ e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta} p_2 e^{-\beta(t/y_2+g_{21}(t,y_1,y_2))} \left(\int_{T=0}^{\infty} e^{-h_{21}^{(+)}(t,y_1,y_2,T)} \mu_2(dT) \right)$$

for positive functions $g_{12}(t,y_1,y_2), g_{21}(t,y_1,y_2)$ *satisfying* $g_{jj'}(t) < t\mathbb{E}_{\pi(dv,d\Lambda)}\left[\frac{c_d(\Lambda)}{c_{d-1}(\Lambda)}|v|\right]$ *for isotropic* k, $p_0,p_1,p_2 > 0$ *with* $p_0+p_1+p_2 = 1$, λ_j *as defined in Theorem 5.2.6, probability measures* $\mu_1(dT), \mu_2(dT)$ *depending on* x_1, x_2, y_1, y_2, *and positive, monotone nondecreasing functions* $h_{21}^{(0)}(t,y_1,y_2), h_{12}^{(0)}(t,y_1,y_2), h_{12}^{(+)}(t,y_1,y_2,T)$, $h_{21}^{(+)}(t,y_1,y_2,T)$ *that are bounded by constants depending on* y_1, y_2.

Theorem 5.2.7 shows that the waiting time distribution in the general case is given by a mixture of an atom at zero and four components. Two of the four components have survival functions that are bounded by a constant times an exponential survival function. The final two components are stochastically dominated by exponentials. Consequently, mixture models of a point mass at zero with exponential distributions are a reasonable choice for modeling waiting times in the general max-stable velocity process model.

5.3 Inference: Tail waiting times

While fitting the Gaussian max-stable velocity process to data would be computationally intensive and require complex algorithms, the waiting times between exceedances in the model are well-approximated by simple mixtures of exponential distributions. Thus, we propose to use the waiting times between exceedances as data to perform inference on tail dependence. In this section we propose a novel tail dependence index that is a function of waiting times. We then propose a Bayesian approach to inference on this quantity.

5.3.1 Waiting times as a measure of extremal dependence

Classically, the tail dependence index of a stochastic process $Y(x)$ at points x_1, x_2 is defined as $\gamma = \lim_{y\to\infty} \mathbb{P}[Y(x_1) > y | Y(x_2) > y]$, the asymptotic limit of the condi-

tional survival function. A conceptually similar approach is possible for *tail waiting times* $\kappa_{(1,2)}(y_1, y_2)$ as defined in (5.4). Let $\mu_{(1,2)}(\cdot, y_1, y_2)$ be the family of probability measures on $\mathbb{R}$ induced by the collection of random variables $\kappa_{(1,2)}(y_1, y_2)$, so that for any μ-measurable set A and any (y_1, y_2) we have $\mu_{(1,2)}(A, y_1, y_2) = \mathbb{P}[\kappa_{(1,2)}(y_1, y_2) \in A]$. Similarly, let $\mu^{\perp}_{(1,2)}(\cdot, y_1, y_2)$ be the measure induced by $\kappa_{(1,2)}(y_1, y_2)$ under the assumption $\kappa_1 \perp \kappa_2$; this can be derived from the marginal distributions of $\kappa_1(y_1)$ and $\kappa_2(y_2)$. For any metric $d(\cdot, \cdot)$ on the space of probability measures, define a tail dependence index $\gamma_d(y_1, y_2)$ based on tail waiting times as

$$\gamma_d(y_1, y_2) := d\left(\mu_{(1,2)}(\cdot, y_1, y_2), \mu^{\perp}_{(1,2)}(\cdot, y_1, y_2)\right), \qquad \gamma_d := \lim_{y_1, y_2 \to \infty} \gamma_d(y_1, y_2),$$

In this framework, a value of $\gamma_d = 0$ corresponds to convergence as $y_1, y_2 \to \infty$ of the distribution – in the topology induced by d – of $\kappa_{(1,2)}(y_1, y_2)$ to its distribution under independence, whereas nonzero γ_d corresponds to discrepancy between the measures that persists asymptotically.

Inference on $\gamma_d(y_1, y_2)$ requires specification of a metric d. We primarily use a form of the Wasserstein metric characterized in detail in Minsker et al. (2014) that is easy to estimate from samples, even in high dimensions.

Consider the space of functions $\mathcal{F}_\varphi := \{f : \Theta \to \mathbb{R}, \sqrt{\langle f, f \rangle_{\mathbb{H}}} \leq 1\}$, where $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ is a reproducing kernel Hilbert space with reproducing kernel $\varphi : \Theta \times \Theta \to \mathbb{R}$. Then for probability measures μ_1, μ_2 satisfying $\int_\Theta \varphi(\theta, \theta')^q d\mu(\theta) < \infty$, the Wasserstein-1 distance with respect to the metric φ is

$$W_{1,\varphi}(\mu_1, \mu_2) = \left\| \int_\Theta \varphi(\theta, \cdot) d(\mu_1 - \mu_2)(\theta) \right\|_{\mathbb{H}}.$$

For certain choices of φ, including the Gaussian kernel $\varphi(\theta, \theta') = \tau \exp\left(-\phi ||\theta - \theta'||_2^2\right)$, the metric satisfies $W_{1,\varphi}(P, Q) = 0$ if and only if $P = Q$ and $0 \leq W_{1,\varphi} \leq \tau$. So, choosing $\tau = 1$ gives a metric supported on the unit interval. It is also easy to compute a finite-sample estimate using equation 2.12 in Minsker et al. (2014).

Moreover, conditions similar to those required for convergence of the posterior in the Hellinger metric – as defined in Ghosal et al. (2000) – imply convergence in $W_{1,\varphi}$, at the same rate Minsker et al. (2014). For simplicity, in the sequel we refer to $W_{1,\varphi}$ as the "Wasserstein distance," and all of the estimates presented utilize the isotropic Gaussian kernel with $\tau, \phi = 1$.

Another metric used mainly as a point of comparison for $W_{1,\varphi}$ is the total variation distance (TV), given by $||\mu - \tilde{\mu}||_{\mathrm{TV}} = \sup_A |\mu(A) - \tilde{\mu}(A)|$, where the supremum is taken over sets A measurable with respect to μ and $\tilde{\mu}$. For one dimensional distributions, a simple point estimate of $||\mu_1 - \mu_2||_{TV}$ is provided by the Kolmogorov-Smirnov statistic.

5.3.2 Calculating waiting times from observed data

We assume that the data $w(\boldsymbol{x}, t)$ are transformed to identical marginals, which without loss of generality are taken to be unit Fréchet. For each $x_i : i \in 1, \ldots, n$, select a threshold y_i corresponding to a high quantile of the Fréchet distribution. Compute the marginal waiting times until first exceedance $\kappa_i(y_i)$ by:

(1) Find the first observed time s_1 for which $w(x_i, s_1)$ exceeds y_i, this is the first observed value of $\kappa_i(y_i)$;

(2) Find $s^* = \min\{t_j > s_1 : w(x_i, t_j) < y_i\}$;

(3) Find $s_2^* = \min\{t_j > s^* : w(x_i, t_j) > y_i\}$ and set $s_2 = s_2^* - s^*$; this is the next observed value of $\kappa_i(y_i)$;

(4) Repeat (2)-(3) until the set $\{t_j > s^* : w(x_i, t_j) > y_i\}$ is empty.

To compute the observed waiting times between exceedances $\kappa_{(1,2)}(y_1, y_2)$ at locations (x_1, x_2), for each observed value s of $\kappa_1(y_1)$ do

(1) Find $s_1^+ = \min\{t_j \geq s_1 : w(x_2, t_j) > y_2\}$ and $s_1^- = \max\{t_j \leq s_1 : w(x_2, t_j) > y_2\}$;

(2) Set $s_1 = |s - s_1^+| \wedge |s - s_1^-|$.

The resulting data are n vectors $\kappa_i(y_i)$ of marginal waiting times until first exceedances and $n(n-1)/2$ vectors of waiting times $\kappa_{(i,i')}(y_i, y_{i'})$ between first exceedances at pairs of locations $(x_i, x_{i'})$. Clearly, the larger y_i, the fewer observations of first exceedances exist. The effect of the threshold on sample size and the necessity of choosing y_i large enough that values exceeding this threshold have approximately a limiting max-stable velocity distribution are the two opposing considerations in choosing a threshold. Our goal here is mainly to illustrate the potential of the method, so we take the simple approach of choosing thresholds such that $\min_i ||\kappa_i(y_i)||_0 \geq 100$, where for a vector κ, $||\kappa||_0$ is the number of elements of κ.

5.3.3 Estimation of $\gamma_d(y_1, y_2)$

We take a model-based approach to estimating $\mu^{\perp}_{(i,i')}(y_i, y_{i'})$ and $\mu_{(i,i')}(y_i, y_{i'})$, guided by Theorems 5.2.5 and 5.2.7. If the data originate from a max-stable velocity process, then the marginal waiting time distribution is given by Theorem 5.2.3, and the distribution of waiting times between exceedances is given by Theorem 5.2.7. These results suggest that a mixture of an atom at zero and several exponential distributions provides a good approximation to the waiting time distributions. The distribution of $\kappa_i(y_i)$ may be well-approximated using only one exponential component, though depending on the shape of the function $f(t, y)$, several components may be required. The waiting times between exceedance may require four or more components, both when $\kappa_i \perp \kappa_{i'}$ and in the general case. When $v = 0$ almost surely, both of these distributions is exactly a mixture of two exponential distributions and an atom at zero. As such, we model $\kappa_i(y_i)$ and $\kappa_{(i,i')}(y_i, y_{i'})$ using a mixture of several exponential components and an atom, e.g.

$$\kappa \sim \eta_0 \delta_0 + \sum_{j=1}^{K-1} \eta_j \text{Exponential}(\lambda_j). \tag{5.11}$$

We take a Bayesian approach to inference in this model by specifying the priors $\eta \sim \text{Dirichlet}(1/K, \ldots, 1/K)$, and $\lambda_j \sim \text{Gamma}\left((,1),1\right)$. Computation for this model is via a straightforward Gibbs sampler, which is described in Appendix D. All of the empirical results in the sequel were obtained using $K = 11$ components, and point estimates $\widehat{\eta}$ always correspond to the ergodic average from sample paths obtained from the Gibbs sampler.

The motivation for a Bayesian approach is threefold. First, censoring of κ is very common in applications because data are often sampled at regular, discrete intervals. This is easy to handle in a Bayesian model by imputing uncensored waiting times. Second, by setting K to be relatively large, the posterior is consistent for the true number of mixture components, and the extraneous components tend to empty (Rousseau and Mengersen (2011)). Thus, the Bayesian approach provides an easy way to estimate the number of mixture components and assess model fit. When $v = 0$ almost surely, the theoretical number of exponential components is exact, and this scenario may be a good approximation in some cases where velocity tends to be small relative to the distance between points. For general $\pi(dv)$, Theorems 5.2.3, 5.2.6 and 5.2.7 suggest that mixtures of between one and four exponential components and an atom can provide a reasonable approximation. The main interest is in estimating how this compares to the number of components necessary to fit observed waiting times. Finally, a Bayesian approach provides estimates of uncertainty in $\mu_{(i,i')}(\cdot, y_i, y_{i'})$ and $\mu^{\perp}_{(i,i')}(\cdot, y_i, y_{i'})$ without the use of additional procedures.

Estimating models of the form in (5.11) by MCMC yields approximate posterior samples $(\eta_i, \lambda_i \mid \kappa_i(y_i))$, $(\eta_{i'}, \lambda_{i'} \mid \kappa_{i'}(y_{i'}))$, and $(\eta_{(i,i')}, \lambda_{(i,i')} \mid \kappa_{(i,i')}(y_i, y_{i'}))$. To obtain approximate samples from

$$\Pi(\gamma_d(y_1, y_2) \mid \kappa) = \Pi(\gamma_d(y_i, y_{i'}) \mid \kappa_i(y_i), \kappa_{i'}(y_{i'}), \kappa_{(i,i')}(y_i, y_{i'})), \tag{5.12}$$

where Π is the posterior distribution, follow the procedure:

(1) For each sampled value of $\{(\eta_i, \lambda_i), (\eta_{i'}, \lambda_{i'})\}$, take M samples from the posterior distribution $\Pi_0(\kappa^{\perp}_{(i,i')}(y_i, y_{i'}) \mid (\eta_i, \lambda_i), (\eta_{i'}, \lambda_{i'}))$, e.g. by sampling independently from the posterior distributions for κ_i and $\kappa_{i'}$ and computing $|\kappa_i - \kappa_{i'}|$.

(2) For each sampled value of $(\eta_{(i,i')}, \lambda_{(i,i')})$, take M samples from the posterior distribution $\Pi(\kappa_{(i,i')}(y_i, y_{i'}) \mid (\eta, \lambda))$.

(3) Compute an estimate of $d(\mu_{(i,i')}(\cdot, y_i, y_{i'}), \mu^{\perp}_{(i,i')}(\cdot, y_i, y_{i'}))$ based on the M samples taken in the previous two steps.

(4) Repeat (1)-(3) for each retained sample of $\{(\lambda_i, p_i), (\lambda_{i'}, p_{i'}), (\eta, \mu)\}$.

The resulting samples are approximately samples from the posterior distribution of $\gamma_d(y_i, y_{i'})$. All point estimates $\hat{\gamma}_d(y_1, y_2)$ in the empirical results were obtained from ergodic averages of $\gamma_d(y_i, y_{i'})$.

5.3.4 Posterior Inference

We now suggest approach for characterizing the strength of tail dependence given the posterior distribution in (5.12). Let $\theta = (\eta, \lambda)$ be the parameters of the mixture model in (5.11). For any θ, let $\kappa_{(i,i')}(y_i, y_{i'} \mid \theta)$ be a random variable with distribution given by (5.11), and let $\mu_{(i,i')}(y_i, y_{i'} \mid \theta)$ be the measure induced by $\kappa_{(i,i')}(y_i, y_{i'} \mid \theta)$. Consider the random variable d^* given by $d^*(\theta_1, \theta_2) = d(\mu_{(i,i')}(y_i, y_{i'} \mid \theta_1), \mu_{(i,i')}(y_i, y_{i'} \mid \theta_2))$, where $\theta_1, \theta_2 \overset{\perp}{\sim} \Pi(\theta \mid \kappa_{(i,i')}(y_i, y_{i'}))$. The distribution of d^* is the posterior distribution of distances between measures induced by the likelihood in (5.11), and reflects uncertainty about the parameters θ after having observed the waiting times $\kappa_{(i,i')}(y_i, y_{i'})$. If there is strong tail dependence between sites x_i and $x_{i'}$, we would expect $\hat{\gamma}_d(y_1, y_2)$ to be large relative to d^*. Thus, as a basic measure of the relative magnitude of $\gamma_d(y_1, y_2)$, we estimate

$$p_d = \mathbb{P}[\hat{\gamma}_d(y_1, y_2) > d^* \mid \kappa] = \mathbb{E}_{\Pi \times \Pi, \Pi_0}\left[\mathbb{1}_{\{d^*(\theta_1,\theta_2) < \gamma_d(\mu_{(i,i')}(y_i, y_{i'} \mid \theta_1), \mu^{\perp}_{(i,i')}(y_i, y_{i'} \mid \theta_0))\}}\right];$$

if p_d is near 1, there is strong evidence for tail dependence.

5.4 Simulation

In this section, a simulation study is constructed to illustrate the method. The simulation is motivated by extreme weather events, where basic scientific knowledge allows informative choices. Data are simulated from a Gaussian max-stable velocity process with attribute distribution

$$\pi(a) = \pi(d\Lambda dr d\theta) \;\propto\; |\Lambda|^{(\nu-k-1)/2} e^{-\mathrm{tr}(\Psi^{-1}\Lambda)/2}\; r^{-3/2} e^{-(\phi(r-m)^2)/(2m^2 r)} \mathbb{1}_{\{r>0\}}$$
$$\times \prod_{h=1}^{k-1} \left[\frac{1}{2} q e^{-q\theta} \mathbb{1}_{\{\theta>0\}} + \frac{1}{2} q e^{-q\theta} \mathbb{1}_{\{\theta<0\}} \right],$$

where $(r, \theta_1, \ldots, \theta_{k-1})$ is the polar parametrization of the velocity v. This corresponds to a Wishart distribution for the kernel shape Λ with degrees of freedom ν and shape Ψ, an inverse Gaussian distribution for the magnitude of the velocity r with parameters m, ϕ, and wrapped double exponential distributions with parameter q for the angles $(\theta_1, \ldots, \theta_{k-1})$ defining the direction of the velocity in $\mathbb{R}^k$. As previously specified, the storm lifetimes $\tau \sim \mathrm{Exp}(\delta)$ and the support points ξ follow a homogeneous spatial Poisson process with rate $\beta d\xi$. The intensity measure in the specification of the process $u \;\propto\; u^{-2}$ is improper. For the simulation, we put $u \sim \mathrm{Pareto}(u_{\min}, 1)$, which results from truncating u from below at $u_{\min}$. Hyperparameters for the simulation are shown in Table 5.1.

Table 5.1: Hyperparameter choices for simulations

	β	$u_{\min}$	δ	Ψ	ν	m	ϕ	q
value	1/600	1	1/120	I_2	7	1/10	1/2	1/2

The data are simulated on a 10×10 box B. In order to inform hyperparameter choices, this box is taken to roughly represent a 5000km^2 area containing the continental United States and southern Canada. As a result, the distributions of velocity

and lifetimes of storms are chosen to approximate the behavior of weather events in this geographic region. To set the time scale and allow easier interpretation of results, one unit of time in the simulation is considered one hour. Support points of the marginal process $\mathcal{N}(d\xi ds)$ are sampled on $B \times [0, T]$, with $T = 50 \times 365 \times 24$, so that the number of support points of $\mathcal{N}(d\xi ds)$ are Poisson distributed with mean $\beta \times T \times 100$. The choice of $\beta = 1/600$ gives an average of four storms a day forming in the region. Storm lifetimes τ_j are sampled for each support point from Ex(1/120), which gives an average storm lifetime of five days. Intensities, shapes, and velocities are sampled from the specified distributions with the hyperparameters given in Table 5.1. The values $m = 0.1, \phi = 1/2$ for the inverse Gaussian distribution on r gives an average speed of about 30 miles per hour. The parameter $q = 0.5$ places most of the mass on easterly storm tracks.

The value of the process $Y(x, t)$ is recorded at one million homogeneously spaced time points from $[0, T]$ at the five locations $\{x_1 = (5, 5), x_2 = (5, 5.5), x_3 = (1, 1), x_4 = (8, 8), x_5 = (3, 5)\}$, as well as twenty additional locations sampled uniformly on B. The five fixed points should result in the process at some pairs of locations being highly tail dependent, some pairs weakly dependent, and some nearly independent. After simulation, waiting times between exceedances of $y = \hat{F}_i^{-1}(0.99)$ and $y = \hat{F}_i^{-1}(0.999)$, where $\hat{F}_i^{-1}(\cdot)$ is the empirical distribution function of y_i, are calculated at the every unique pair of points. Mixture models of the form in (5.11) are then fit to the waiting times until first exceedance and the waiting times between exceedances at all pairs of locations.

Figure 5.2 shows results; some additional results are provided in Figure D.2 in Appendix D. The posterior concentrates around a single exponential component in the models for $\kappa_i(y_i)$, which is consistent with the marginal distribution when $v = 0$ almost surely. Since for most pairs of points, the velocity is small relative to distance between the points and the expected lifetimes, this result is sensible.

Two components, and occasionally three, are required to model $\kappa_{(i,i')}(y_i, y_{i'})$. We remarked earlier that Theorems 5.2.7 and 5.2.6 suggest that $\kappa_{(i,i')}(y_i, y_{i'})$ should be well-approximated by mixtures of between two and four exponential components and a point mass at zero; the empirical evidence supports this. At both thresholds, $\hat{\gamma}_d(y_1, y_2)$ has a prominent mode near 0.05., and regardless of the choice of d, $\hat{\gamma}_d(y_1, y_2)$ decreases with the distance between points. When $d = \varphi$, for distances greater than about 2000 kilometers, $\hat{\gamma}_d(y_1, y_2)$ is approximately 0.05. Notably, $\hat{\gamma}_d(y_1, y_2)$ decays more slowly and exhibits substantially more variation when $d = TV$, suggesting possibly lower power.

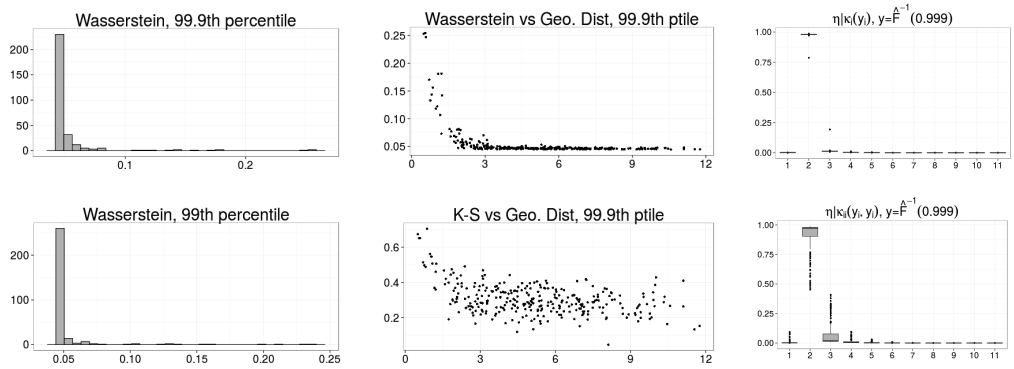


FIGURE 5.2: *Left: histograms of $\hat{\gamma}_d(y_1, y_2)$ for all pairs of locations with $d = W_{1,\varphi}$ for thresholds $y_i = \hat{F}_i^{-1}(0.999)$ and $\hat{F}_i^{-1}(0.99)$. Center: plots of $\hat{\gamma}_d(y_1, y_2)$ for $d = W_{1,\varphi}$ and $d = TV$ for threshold $y_i = \hat{F}_i^{-1}(0.999)$ versus Euclidean distance between points. Right: Posterior estimates of $\eta_i \mid \kappa_i(y_i)$ and $\eta_{(i,i')} \mid \kappa_{(i,i')}(y_i, y_{i'})$ for $y = F^{-1}(0.999)$. Boxplots are over all recorded points or all pairs of points.*

5.5 Applications

The method is applied to four real data sets: (1) Daily precipitation data for 25 weather stations in the United States for the period 1940-2014; (2) Daily exchange rates for 12 currencies for the period 1986-1996; (3) Daily prices for the period 2000-2014 of the 30 stocks that made up the Dow Jones Industrial Index as of January 2015; and (4) Electrical potential at 62 single neurons in the brain of a

mouse exploring a maze, sampled at 500 Hz.

The first dataset is similar to the simulation study and the physical heuristic that we introduced in Section 5.2. The points in the max-stable velocity process can be thought of as storms or pressure systems in this context. The fourth – the mouse electrophysiology data – retains an explicit spatial component, as the physical distance between neurons is related to the speed at which signals can be propogated between them. The second and third applications are financial. One way to interpret the max stable velocity model in financial settings is to think of assets as embedded in latent space, with the distances between them reflecting similarity in the factors that determine their price. The support points of the process can be thought of as market sentiment, and their movement reflects the spread of sentiment through the asset space. Larger values of u reflect panicks – sentiment that affects many asset classes simultaneously – and large velocity reflects rapidly spreading sentiment. Finally, in financial settings it is usually extremes in the left tail that are of particular interest, and thus it is useful to think of the max-stable velocity process as the negative of the observed data, in which case exceedance of the negative of a low threshold is the relevant event.

In choosing thresholds for analysis, the heuristic used is that at least 100 first exceedances of the highest threshold must be observed at all points. Thresholds correspond to empirical quantiles of the observed data, and at least two thresholds are analyzed for every dataset. The consequence of choosing thresholds in this way is that for some datasets, the empirical quantile of the chosen threshold is much more extreme than for others. For example, the electrophysiology data has nearly two million observations, so we can choose a threshold of $y_i = \hat{F}_i^{-1}(0.998)$ and $y_i = \hat{F}_i^{-1}(0.99)$ for analysis; the exchange rate data has only about two thousand observations, so the thresholds chosen are $y_i = \hat{F}_i^{-1}(0.05)$ and $y_i = \hat{F}_i^{-1}(0.10)$. For the Dow Jones data, the thresholds are $y_i = \hat{F}_i^{-1}(0.025)$ and $\hat{F}_i^{-1}(0.05)$, and for the

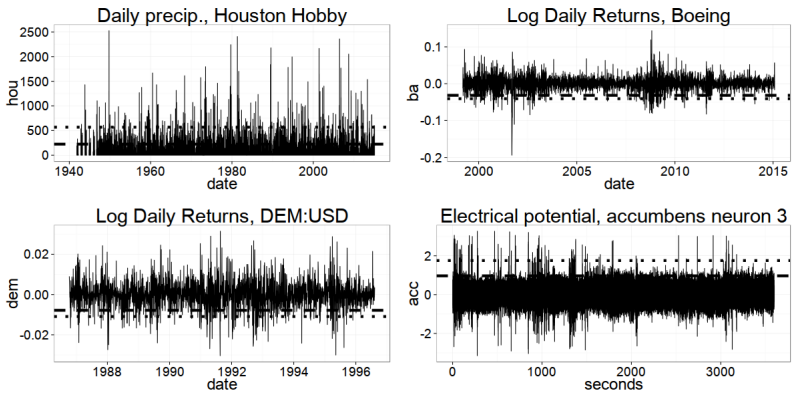


FIGURE 5.3: *Examples of raw data $w(x,t)$. Dotted lines: more extreme threshold; dashed lines: less extreme threshold for computing $\gamma_d(y_1, y_2)$.*

precipitation data, the thresholds are $y_i = \hat{F}_i^{-1}(0.99)$ and $y_i = \hat{F}_i^{-1}(0.95)$. Examples of single components of the four datasets are shown in Figure 5.3. In the case of the two financial datasets, the displayed series is the log daily returns $\log[w(x,t)/w(x,t-1)]$, which, following standard practice in finance, is the data used for analysis. The other two show the raw data plotted in the time domain.

For comparison, we use the method described in Engelke et al. (2014). When W is in the max domain-of-attraction of a Brown-Resnick process, Engelke et al. (2014) that if the process is transformed to exponential margins, correlations in the increments of the process relative to its value any one location, conditional on threshold exceedance at the reference location, have approximately a Gaussian distribution. Practically, this approach proceeds by first transforming the data to have exponential margins. Label the transformed data $\tilde{w}(x,t)$. Then choose a reference location and label it location 1. Compute the increments $\tilde{w}(\boldsymbol{x}_{[-1]}, t) - \tilde{w}(x_1, t)$ relative to location 1, where $\tilde{w}(\boldsymbol{x}_{[-1]}, t)$ is the vector of process realizations at time t at locations other than location 1. Retain only data points where $\tilde{w}(x_1, t)$ exceeds a high threshold. Now compute the correlations in $\tilde{w}(\boldsymbol{x}_{[-1]}, t) - \tilde{w}(x_1, t)$. The magnitude of the correlation provides inference on the strength of tail dependence; the closer the absolute correlation is to 1 or -1, the stronger the tail dependence. We refer to this

as the conditional correlation method.

5.5.1 Precipitation

Results for analysis of the precipitation data are summarized in Figure 5.4. A map showing the location of each station can be found in the Appendix. Clear geographic structure is evident in the estimated values of $\gamma_d(y_1, y_2)$. Similar geographic structure is seen in tail waiting times and conditional correlations. Overall, tail dependence is evident at nearby sites but decays with distance; for distances greater than about 500 km the estimated values of $\gamma_d(y_1, y_2)$ are all very similar. Notably, the number of exponential components in the models for $\kappa_{(i,i')}(y_i, y_{i'})$ and $\kappa_i(y_i)$ is almost always two, and in many cases the weight on the second component is relatively small. This shows relatively close agreement with the simulation study, suggesting that the magnitude of velocity is generally small relative to the distance between points and storm lifetimes. In a few cases, four exponential components are required; these probably correspond to points that are relatively close, so that velocity has a substantial impact on waiting times.

5.5.2 Dow Jones components

Figure 5.5 shows results for analysis of log daily returns $-\log[w(x,t)/w(x,t-1)]$ for the DJIA data; the reference asset for the conditional correlations method is axp (American express). Similar dependence structure is evident using the conditional correlation and tail waiting times methods. Most values of $\hat{\gamma}_d(y_1, y_2)$ cluster around 0.2 at both thresholds, with a long right tail in the posterior point estimates. Here, unlike in the precipitation data, the posterior for the parameters of the mixture models for both $\kappa_{(i,i')}(y_i, y_{i'})$ and $\kappa_i(y_i)$ concentrated around four exponential components, suggesting that "velocity" – which in this case can be thought of as the rate at which market sentiment diffuses through asset space – is an important factor

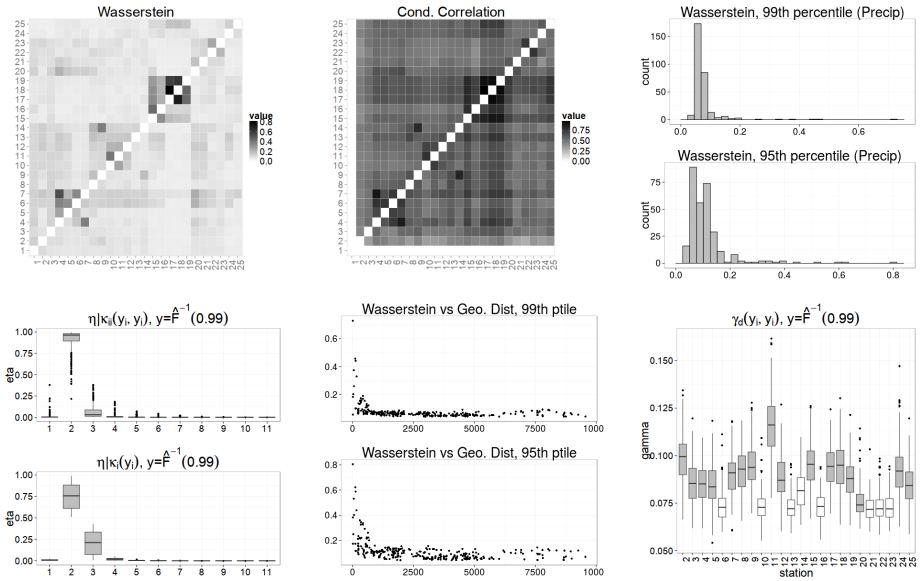


FIGURE 5.4: *Results for daily precipitation data, $d = W_{1,\varphi}$. Top Left: $\hat{\gamma}_d(y_1, y_2)$ for thresholds $y_i = \hat{F}_i^{-1}(0.99)$ (below the main diagonal) and $y_i = \hat{F}_i^{-1}(0.95)$ (above the main diagonal). The sites are arranged by geographic distance. Top center: conditional correlations relative to site 1 $y_i = \hat{F}_i^{-1}(0.99)$ (below the main diagonal) and $y_i = \hat{F}_i^{-1}(0.95)$ (above the main diagonal). Top right: histograms of $\hat{\gamma}_d(y_1, y_2)$ for both thresholds. Bottom left: boxplots of posterior means of $\eta_{(i,i')} \mid \kappa_{(i,i')}(y_i, y_{i'})$ (top) and $\eta_i \mid \kappa_i(y_i)$ for $y_i = \hat{F}_i^{-1}(0.99)$. Bottom center: plots of $\hat{\gamma}_d(y_1, y_2)$ versus geographic distance for both thresholds. Bottom right: boxplot of posterior samples of $\gamma_d(y_1, y_2)$ for $y_i = -\hat{F}_i^{-1}(0.99)$ for pairs that include station 1, gray boxes indicate $p_d > 0.95$.*

in determining waiting times. About half the assets pairs involving axp have values of $p_d > 0.95$. Notably, the largest values of $\hat{\gamma}_d(y_1, y_2)$ for pairs involving axp are for jpm (JP Morgan Chase) and v (Visa), two other large credit card issuers. There is considerable dispersion in imputed waiting times, with a median of around eight days. A table showing the identity of the stock corresponding to each row/column in the colormaps in Figure 5.5 can be found in Appendix D.

5.5.3 Exchange Rates

Exchange rate data for twelve currencies is described in Harrison and West (1999) and Prado and West (2010). Similar to the Dow Jones data, these series are transformed to negative log daily returns ($-\log[w(x,t)/w(x,t-1)]$) and analyzed at two

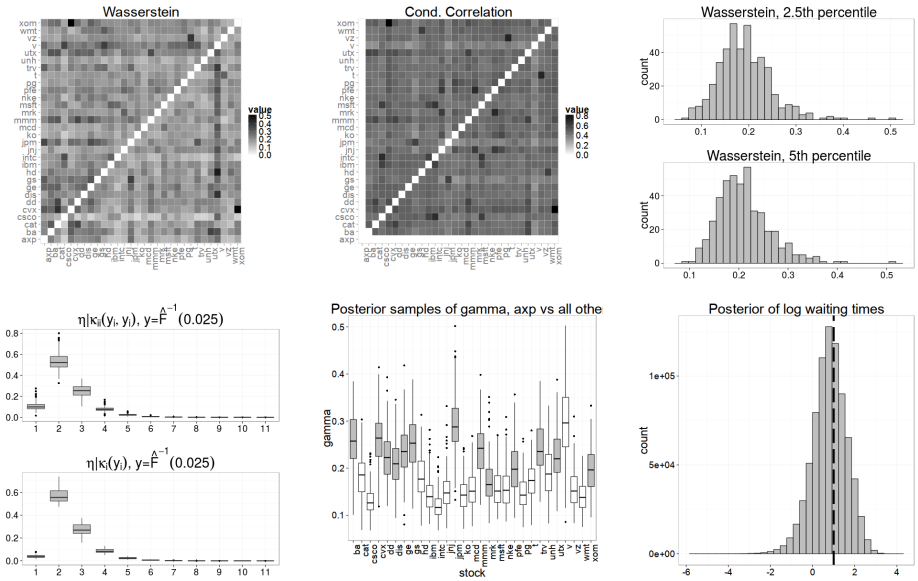


FIGURE 5.5: *Results for DJIA components. Top Left:* $\hat{\gamma}_d(y_1, y_2)$ *for thresholds* $y_i = \hat{F}_i^{-1}(0.025)$ *(below the main diagonal) and* $y_i = \hat{F}_i^{-1}(0.05)$ *(above the main diagonal). Top center: correlations conditional on exceedance of* $y_i = \hat{F}_i^{-1}(0.025)$ *(below the main diagonal) and* $y_i = \hat{F}_i^{-1}(0.05)$ *(above the main diagonal) by axp. Top right: histograms of* $\hat{\gamma}_d(y_1, y_2)$ *for both thresholds. Bottom left: posterior means of* $\eta_{(i,i')} \mid \kappa_{(i,i')}(y_i, y_{i'})$ *(top) and* $\eta_i \mid \kappa_i(y_i)$ *for* $y_i = \hat{F}_i^{-1}(0.99)$*. Bottom center: posterior samples of* $\gamma_d(y_1, y_2)$ *for* $y_i = -\hat{F}_i^{-1}(0.025)$ *for pairs that include axp; gray boxes indicate* $p_d > 0.95$*. Bottom right: samples of log imputed posterior waiting times (all pairs pooled) for exceedance of* $y_i = -\hat{F}_i^{-1}(0.025)$*; vertical line at ten days.*

thresholds: $y_i = -\hat{F}_i^{-1}(0.05)$ and $y_i = -\hat{F}_i^{-1}(0.10)$. A table giving the full name of each currency and its corresponding row/column in the colormap is provided in the Appendix. Figure 5.6 shows some results. Here, there is very clear structure in the pattern of pairwise dependence, with the European currencies (BEF, FRF, DEM, NLG, ESP, SEK, CHF, and GBP) showing strong evidence of dependence wile the other four currencies (AUD, CAD, JPY, and NZD) show little evidence of tail dependence among themselves or with the European currencies. Results for the conditional correlation method were similar (not shown). The posterior intervals for $\gamma_d(y_1, y_2)$ – again, colored gray if $p_d > 0.95$ – show the same pattern, though notably all of the currencies have $p_d > 0.95$ for at least one pair. These results are

broadly consistent with previous analysis of dependence at the mean in this dataset (see Prado and West (2010)).

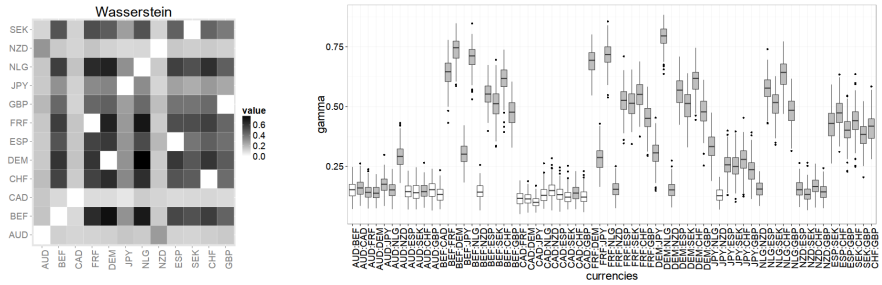


FIGURE 5.6: *Results for exchange rate data. Left:* $\hat{\gamma}_d(y_1, y_2)$ *for thresholds of (the negative of)* $y_i = -\hat{F}_i^{-1}(0.05)$ *(below the main diagonal) and* $y_i = -\hat{F}_i^{-1}(0.10)$ *(above the main diagonal). Right: boxplot of posterior samples of* $\gamma_d(y_1, y_2)$ *for* $y_i = -\hat{F}_i^{-1}(0.025)$ *for all pairs of currencies. White boxes indicate* $p_d > 0.95$

5.5.4 Electrophysiology

Potential data recorded at single neurons in the brain of a mouse interacting with a maze are analyzed using two thresholds: $y_i = \hat{F}_i^{-1}(0.998)$ and $y_i = \hat{F}_i^{-1}(0.99)$. In electrophysiology, interest lies in modeling dependence between neuron "spikes" at different locations in the brain. Neuronal voltage spikes indicate transmission of signals along axonal pathways, and large potentials tend to cluster together in small time windows. These events are referred to as "spike trains." Thus, in this data one expects to see extensive tail dependence, but the waiting times between spikes at different neurons is relevant, since it informs about the pathway that the signal takes through the brain.

The elecrophysiology data used here are from single neurons in the brain of a mouse exploring a maze, and are described in Dzirasa et al. (2010). The neurons are assigned to regions of the brain, which are shown in some of the subimages in Figure 5.7. There is ample evidence of strong tail dependence for most pairs of neurons. Four neurons evidence a markedly different pattern of dependence from the others; at times, electrodes are faulty, resulting in data quality problems, which is

the most likely explanation in this case. This anomaly aside, strong dependence is evident both from waiting times and conditional correlations approach. However, a pattern revealed by the waiting times method that is not as clear with conditional correlations is that $\gamma_d(y_1, y_2)$ increases markedly with y for almost every neuron pair. Moreover, clear differences in dependence between and within regions is evident using the waiting times approach, but less clear with the conditional correlation method. There is again evidence that velocity is an important factor, with the models for $\kappa_i(y_i)$ requiring four to five exponential components and those for $\kappa_{(i,i')}(y_i, y_{i'})$ between three and four. Here, we can interpret velocity roughly as communication between different brain regions, as opposed to simultaneous stimulus of multiple regions. This is consistent with the basic mechanism by which neurons transmit information, by propagation of electrical impulses along axons and transmission of signals across the synaptic cleft to other neurons. Posterior credible intervals for $\gamma_d(y_1, y_2)$ at $y = \hat{F}_i^{-1}(0.998)$ have $p_d > 0.95$ for all pairs involving the first Accumbens region neuron; this was largely the case for all 1891 pairs of neurons, indicating strong evidence of tail dependence across all brain regions.

The electrophysiology data are overall most similar to the Dow Jones data: values of $\gamma_d(y_1, y_2)$ are mostly large and increase as the threshold becomes more extreme. However, in these data the shift toward one in $\gamma_d(y_1, y_2)$ as y increases is much more prominent, with $\gamma_d(y_1, y_2)$ appearing almost uniformly distributed at $y_i = \hat{F}_i^{-1}(0.99)$ but massed near 1 at $y_i = \hat{F}_i^{-1}(0.999)$. Again, this shift is only evident using the waiting times method; it is not shown by conditional correlations. That the imputed waiting times are rather dispersed suggests a similar explanation for this phenomenon to the similar pattern seen for the Dow Jones data.

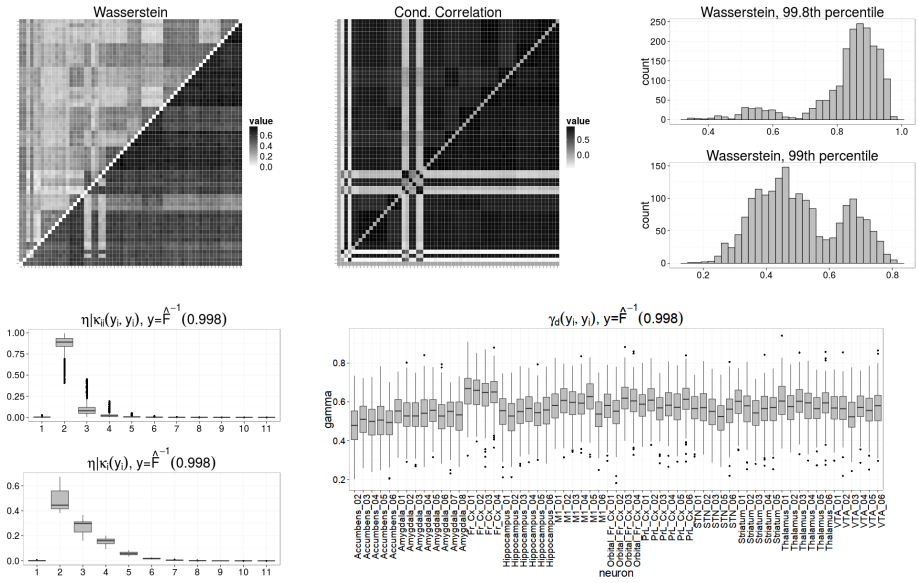


FIGURE 5.7: *Results for electrophysiology data with $d = W_{1,\varphi}$. Top left: $\hat{\gamma}_d(y_1, y_2)$ for thresholds $y_i = \hat{F}_i^{-1}(0.998)$ (below the main diagonal) and $y_i = \hat{F}_i^{-1}(0.99)$ (above the main diagonal). Top center: correlations in increments relative to Accumbens 1, conditional on exceedance of $y_i = \hat{F}_i^{-1}(0.998)$ (below the main diagonal) and $y_i = \hat{F}_i^{-1}(0.99)$ (above the main diagonal) at Accumbens 1. Top right: histograms of $\gamma_d(y_1, y_2)$ for both thresholds across all pairs of neurons. Bottom left: $\eta_{(i,i')} \mid \kappa_{(i,i')}(y_i, y_{i'})$ and $\eta_i \mid \kappa_i(y_i)$ for $y = \hat{F}_i^{-1}(0.998)$. Bottom right: posterior samples of $\gamma_d(y_1, y_2)$ for $y_i = \hat{F}_i^{-1}(0.998)$ for all pairs that include Accumbens 1.*

5.6 Discussion

Characterizing tail dependence based on waiting times between peaks over thresholds has the advantage of greater flexibility and generality than many existing approaches for performing inferences on extremal depenence. The method relies strictly on the waiting times and inference on the parameter $\gamma_d(y_1, y_2)$ is relatively simple and computationally scalable, particularly when closed-form expressions are available for $d(\mu_{(1,2)}(\cdot, y_1, y_2), \mu^{\perp}_{(1,2)}(\cdot, y_1, y_2))$. The proposed method has strong theoretical grounding in the class of max-stable velocity processes described here, which is sufficiently general to apply to various distinct application areas where extremal dependence is of interest. Broader classes of processes with non-trivial tail waiting times – and for which $\lim_{y_1, y_2 \to \infty} \gamma_d(y_1, y_2) \neq 0$ – certainly exist. Simply modifying

the max-stable velocity process to include time-varying magnitudes and velocities would create a much richer class of processes that retains the basic characteristics of the simpler process described here.

Like other peaks-over-thresholds methods, this approach requires the choice of appropriate thresholds. Substantial work has been done on threshold choice for standard peaks-over-thresholds methods. It is unclear whether this will translate directly to threshold choice in this novel context. Here, we have taken the simpler approach of choosing multiple thresholds and attempting to infer a general pattern as the threshold changes. Further work on threshold choice is undoubtedly called for. Additionally, we have thus far modeled the pairwise waiting times entirely independently; clear gains in estimation efficiency would result from sharing information across all pairs. These are promising areas for future work.

6

Approximations of Markov Chains and Bayesian Inference

6.1 Introduction

The fundamental entity in Bayesian statistics is the posterior distribution

$$\Pi(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int_\theta p(x \mid \theta)p(\theta)}, \tag{6.1}$$

the conditional distribution of the model parameters θ given the data x. The integral in the denominator of (7.1) is typically not available in closed form. A common approach constructs an ergodic Markov chain with invariant distribution $\Pi(\theta \mid x)$, and then collect finite sample paths $\theta_1, \ldots, \theta_t$ from the chain. Statistical inference then relies on properties of the ergodic measure $\frac{1}{t}\sum_{k=0}^{t-1} \delta_{\theta_k}$, associated ergodic averages $\frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k)$ for functions f, and other pathwise quantities. This is referred to as Markov Chain Monte Carlo (Robert and Casella (2004), Gamerman and Lopes (2006)) or MCMC.

We consider Markov chains that result from approximating the transition kernel $\mathcal{P}(\theta, \cdot)$ by another kernel $\mathcal{P}_\epsilon(\theta, \cdot)$ satisfying $||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}} < \epsilon$. The use of approximate kernels – often without showing such an error bound – is common

practice in Bayesian analysis, and is usually computationally motivated, i.e. obtaining samples from $\mathcal{P}_\epsilon(\theta, \cdot)$ requires less computation than sampling from $\mathcal{P}(\theta, \cdot)$. Our main contributions are as follows. We provide results showing bounds in total variation and expected L_2 estimation error for finite-time ergodic averages, under simple assumptions on the original chain and the approximating kernel. We also provide a general result on the computational advantage and approximation error tradeoff, providing an explicit criterion for the level of error to tolerate in the approximate transition kernel. We include an illustration to three approximate MCMC (aMCMC) algorithms in which we verify the approximation error assumption, and show practical performance.

While being arguably the dominant algorithm for Bayesian inference, MCMC is computationally demanding in high-dimensional settings, e.g. where either p (the dimension of θ) or n (the number of observations) is large. To more easily apply MCMC in such settings, it is common to approximate $\mathcal{P}(\theta, \cdot)$ with a kernel that is simpler or faster to sample from. One example is inference for Gaussian process models, bypassing $O(n^3)$ matrix inversion through approximations (Banerjee et al. (2008); Banerjee et al. (2013); Hughes and Haran (2013)). Another prevalent example is the use of Laplace or Gaussian approximations to conditional distributions. Guhaniyogi et al. (2014) proposes an algorithm that replaces some sampling steps with point estimates. Korattikara et al. (2013) approximate Metropolis-Hastings acceptance decisions using subsets of the data. It is also common to approximate intractable full conditionals by simpler distributions, with Bhattacharya and Dunson (2010) using a beta approximation, O'brien and Dunson (2004) replacing the logistic with a t distribution, and Ritter and Tanner (1992) discretizing.

While approximating $\mathcal{P}(\theta, \cdot)$ by $\mathcal{P}_\epsilon(\theta, \cdot)$ is common, literature addressing convergence and approximation error of these algorithms is recent. Pillai and Smith (2014) present perhaps the most complete treatment to date, utilizing the theoretical foun-

dation in Joulin et al. (2010) to show error bounds in the Wasserstein topology under fairly general conditions. Their results are applied to the algorithm in Korattikara et al. (2013) and similar subsampling based algorithms. Rudolf and Schweizer (2015) show results under effectively the same conditions, but use Lyapunov functions to eliminate exit probability terms from the bounds in Pillai and Smith (2014) that grow with t; the most recent version of Pillai and Smith (2014) uses a similar approach. Alquier et al. (2014) provide results in a similar context, but focusing on bounding the error between the ergodic measures of the approximate and exact chains. Earlier references show error bounds for perturbations of uniformly ergodic chains (Mitrophanov (2005)) and geometrically ergodic chains (Ferré et al. (2013), **?**; the latter focuses on perturbation resulting from numerical imprecision). Among these references, Pillai and Smith (2014) has a substantial focus on implications of the theoretical results for parameter estimation.

Our work differs from the precedents in several ways. Our results focus on estimation error and on interpretation of the error bounds, and the entire theoretical framework is constructed from a statistical perspective, i.e. with the view that sample paths from the Markov chain will form the basis of estimation via the empirical measure. All of our bounds improve with the MCMC sample path length at the expected rate in t. We provide explicit criteria for determining the optimal level of approximation error given a speedup function quantifying the computational advantage of the approximation and a discrepancy measure quantifying the statistical performance of the approximate algorithm. This is perhaps the most unique aspect of our work, as precedents have not directly addressed the question of when an approximate chain is superior to an exact chain from the point of view of estimation, which is of critical relevance in applications. We further verify the usefulness of the results by applying them to three approximate samplers constructed from common MCMC algorithms for standard Bayesian models: one that employs random subsets

of data, another for Gaussian process models using a low-rank covariance approximation, and a novel algorithm for mixture models for high-dimensional contingency tables. Thus, we consider a broader variety of approximate MCMC algorithms than precedents, which have focused almost exclusively on subsets of data for large n.

6.2 Ergodicity and Approximation Error

This section provides error bounds for statistical estimators constructed from approximate MCMC chains. In particular, we provide bounds in total variation and expected L_2 loss for posterior functions using sample paths from approximate chains. These bounds are then compared to similar bounds for exact chains to illustrate the relative computational efficiency of the approximate chain as a function of computational clock time. Because the bounds obtained for the exact kernel are tight, the comparison of the bounds leads naturally to a novel notion of computational optimality that we refer to as compminimax. Under this optimality criterion, approximate chains are optimal for surprisingly long computation times, though the advantage relative to the exact chain diminishes with computational time.

6.2.1 Approximate MCMC

Consider a family of likelihoods $p(x \mid \theta)$ parametrized by $\theta \in \Theta$. We assume that $X \sim p(x \mid \theta)$ takes values in a Polish space $\mathcal{X}$. In general, the spaces Θ of interest will be equipped with a dominating measure $m^*(\cdot)$. We are concerned with Markov chain Monte Carlo algorithms, which obtain samples from the posterior distribution in (7.1) by constructing an ergodic Markov chain with invariant distribution $\Pi(\theta \mid x)$. To obtain useful bounds on the error from use of an approximate kernel, we require the original Markov chain to satisfy minimal convergence and mixing properties. Our main condition is given in Assumption 6.2.1.

Assumption 6.2.1 (Doeblin condition for exact chain)**.** *There exists a constant* $0 < \alpha < 1$ *such that*

$$\sup_{\theta,\theta^* \in \Theta \times \Theta} ||\mathcal{P}(\theta, \cdot) - \mathcal{P}(\theta^*, \cdot)||_{\mathrm{TV}} < 1 - \alpha \tag{6.2}$$

where $||P - Q||_{\mathrm{TV}}$ *is the total variation distance between probability measures* P *and* Q*. When a kernel* $\mathcal{P}$ *satisfies this condition we write* $d(\mathcal{P}) = \alpha$*.*

Assumption 6.2.1 implies uniform ergodicity. An immediate implication of Assumption 6.2.1 is that $||\Pi - \nu P^t||_{\mathrm{TV}} \leq (1 - \alpha)^t \, ||\nu - \Pi||_{\mathrm{TV}}$.

Although related results can be obtained under a weaker geometric ergodicity condition, the resulting bounds are more complex (e.g. Pillai and Smith (2014), Rudolf and Schweizer (2015)). The Doeblin condition has the advantage of leading to a simple characterization of the approximation accuracy and computational time tradeoff. In practice, the condition can be shown in a variety of cases involving compact state spaces. Compactness is not an overly restrictive assumption in practice, as choosing priors with bounded support is justified in most applications.

Consider a family of alternative transition kernels $\mathscr{P}_\epsilon(\mathcal{P})$, whose members approximate $\mathcal{P}(\theta, \cdot)$. We will require the condition on $\mathscr{P}_\epsilon(P)$ given in Assumption 6.2.2.

Assumption 6.2.2 (Condition on the approximating kernel)**.** *There exists a constant* $0 < \epsilon < \alpha/2$ *such that*

$$\sup_{\theta \in \Theta} ||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}} < \epsilon \tag{6.3}$$

for every $\mathcal{P}_\epsilon(\theta, \cdot) \in \mathscr{P}_\epsilon(\mathcal{P})$*.*

Although we state most results in terms of $\mathcal{P}_\epsilon(\theta, \cdot)$, a generic element of $\mathscr{P}_\epsilon(\mathcal{P})$, they should be understood to hold for every member of $\mathscr{P}_\epsilon(\mathcal{P})$, and apply to chains constructed using an arbitrary sequence of members of $\mathscr{P}_\epsilon(\mathcal{P})$; this simplification is

made for brevity and notational convenience. Assumption 6.2.2 can be weakened; for example, requiring that the approximation error bound hold only on a subset of the parameter space and some structure, such as a Foster-Lyapunov function, which ensures return to that subset. However, we prefer to keep the assumptions and resulting bounds simple and transparent.

6.2.2 Main results

The main results of this section relate the convergence properties of the original chain and the approximation error of the kernel $\mathcal{P}_\epsilon(\theta, \cdot)$ to the approximation error for $\Pi(\theta \mid x)$. First, define

$$\widehat{\Pi}^t_{\mathcal{P}} f = \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k), \qquad \widehat{\Pi}^t_{\mathcal{P}_\epsilon} f = \frac{1}{t}\sum_{k=0}^{t-1} f(\theta^\epsilon_k), \qquad \Pi f = \int_\Theta f(\theta)\Pi(d\theta \mid x).$$

We often omit the subscripts $\mathcal{P}$ and $\mathcal{P}_\epsilon$ when considering transition kernels with a particular invariant measure Π.

Our focus is on the computational efficiency of statistical estimators constructed from sample paths of approximating kernels. To this end, consider any $\mathcal{P}$ corresponding to a MCMC algorithm and a *discrepancy measure* D that quantifies the statistical value of sample paths of length t from $\mathcal{P}_\epsilon \in \mathscr{P}_\epsilon(\mathcal{P})$. Two natural choices for D that we consider here are

$$D_{TV}(\Pi, \mathcal{P}_\epsilon, t) = \left\| \Pi - \frac{1}{t}\sum_{k=0}^{t-1} \nu \mathcal{P}^k_\epsilon \right\|_{\mathrm{TV}} \tag{6.4}$$

$$D_{L_2}(\Pi, \mathcal{P}_\epsilon, t) = \sup_{f:|f|<1} \mathbb{E}\left[\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta^\epsilon_k) \right)^2 \right], \tag{6.5}$$

where in (6.5), the expectation is taken with respect to the distribution of the first t steps of the Markov chain.

The potential advantage of aMCMC is that longer sample paths can be obtained in equal computational (wall clock) time. For any transition kernel $\mathcal{P}$, let $\tau_{\mathcal{P}}(t)$ be

the computational time required to obtain a sample path of length t. Define the *speedup function* $s(\epsilon)$ of a class of approximations $\mathscr{P}_\epsilon(P)$ by

$$s(\epsilon) = \frac{\tau_{\mathcal{P}}(t)}{\inf_{\mathcal{P}_\epsilon \in \mathscr{P}_\epsilon(\mathcal{P})} \tau_{\mathcal{P}_\epsilon}(t)}, \tag{6.6}$$

which we assume is constant as a function of t. Since we focus on cases where aMCMC provides a computational advantage, it makes sense to restrict attention to speedup functions that are monotone nondecreasing in ϵ on the interval $0 < \epsilon < \alpha/2$, and satisfy $s(0) = 1$. For simplicity, we assume that every member of $\mathcal{P}_\epsilon(\mathcal{P})$ having approximation error ϵ_0 has speedup $s(\epsilon_0)$, so that in the sequel the infimum in the denominator of (6.6) is redundant. Without loss of generality, we also take $\tau_{\mathcal{P}}(t) = t$ so that speedup can be interpreted as the number of samples obtained from $\mathcal{P}_\epsilon$ in the time required to obtain one sample from $\mathcal{P}$.

When $s(\epsilon)$ is not constant, there exists the potential that for finite computational budgets, some member of $\mathscr{P}_\epsilon(\mathcal{P})$ will be superior to $\mathcal{P}$ with respect to a discrepancy measure D, because the longer sample paths obtained from $\mathcal{P}_\epsilon$ might more than compensate for any bias and difference in convergence/mixing properties. To make this rigorous, we define a notion of statistical optimality that we refer to as "computational minimax" (compminimax) due to its conceptual similarity to minimax estimators.

Definition 6.2.3 (Definition: Compminimax)**.** Fix a computational budget $\tau_{\max}$ and a discrepancy measure D. An approximation error $\epsilon_c(\tau_{\max})$ is compminimax if

$$\epsilon_c(\tau_{\max}) = \operatorname*{arginf}_{\epsilon < \alpha/2} \sup_{\mathcal{P}_\epsilon \in \mathscr{P}_\epsilon(\mathcal{P})} D(\Pi, \mathcal{P}_\epsilon, \max_t \{t : \tau_{\mathcal{P}_\epsilon}(t) < \tau_{\max}\}) \tag{6.7}$$

With the assumption that $\tau(t) = t$, we have $\max_t \{t : \tau_{\mathcal{P}_\epsilon}(t) < \tau_{\max}\} = \lfloor s(\epsilon)\tau_{\max} \rfloor$.

The definition of compminimax effectively gives a decision rule that assures optimal performance in the worst case scenario when the available information is the

value of α, $s(\epsilon)$, and a computational budget. Using only assumptions 6.2.1 and 6.2.2, we obtain the simple estimation error results in Theorem 6.2.4, which allow evaluation of minimax computational efficiency of aMCMC with respect to the discrepancy measures in (6.4) and (6.5).

Theorem 6.2.4 (Estimation error for aMCMC ergodic averages)**.** *Suppose $\mathcal{P}$ satisfies Assumption 6.2.1, $\mathcal{P}_\epsilon$ satisfies 6.2.2 and f is bounded. Let $\theta_0, \theta_0^\epsilon \sim \nu$ for any probability measure ν on $(\Theta, \mathcal{F}_0)$, where $\mathcal{F}_0$ is the σ-field generated by Θ. Then*

$$\left|\left|\Pi - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}^k\right|\right|_{\mathrm{TV}} \leq \frac{1-(1-\alpha)^t\,||\Pi-\nu||_{\mathrm{TV}}}{\alpha t}, \tag{6.8}$$

$$\left|\left|\Pi - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}_\epsilon^k\right|\right|_{\mathrm{TV}} \leq \frac{\epsilon}{\alpha} + \frac{(1-(1-\alpha_\epsilon)^t)\,||\Pi_\epsilon-\nu||_{\mathrm{TV}}}{\alpha_\epsilon t}; \tag{6.9}$$

and

$$\frac{\mathbb{E}\left[\left(\Pi f - \widehat{\Pi}^t f\right)^2\right]}{||f||_*^2} \leq 4\left(\frac{(1-(1-\alpha)^t)\,||\Pi-\nu||_{\mathrm{TV}}}{\alpha t}\right)^2 + S(t,\alpha), \tag{6.10}$$

$$\frac{\mathbb{E}\left[\left(\Pi f - \widehat{\Pi}_\epsilon^t f\right)^2\right]}{||f||_*^2} \leq 4\left(\frac{\alpha}{\epsilon} + \frac{(1-(1-\alpha_\epsilon)^t)\,||\Pi_\epsilon-\nu||_{\mathrm{TV}}}{\alpha_\epsilon t}\right)^2 + S(t,\alpha_\epsilon) \tag{6.11}$$

where $\alpha_\epsilon = \alpha - 2\epsilon$, $||f||_* = \inf_{c\in\mathbb{R}}||f-c||_\infty$, and

$$S(t,\alpha) = \left(\frac{2}{\alpha t} + \frac{2}{\alpha t^2} + \frac{2(1-\alpha)^{t+1}}{\alpha^2 t^2} - \frac{1}{t} - \frac{2}{\alpha^2 t^2}\right).$$

Expectations are taken with respect to the measure of the first t steps of the Markov chain.

Proofs of the results in this section are provided in the Appendix. Remark 6.2.1 characterizes sharpness of the bounds in Theorem 6.2.4.

Remark 6.2.1 (Sharpness of bounds)**.** The bound in (6.8) is sharp; that is, for every α, there exists a transition kernel $\mathcal{P}$ satisfying the Doeblin condition with $d(\mathcal{P}) = \alpha$ for which equality holds in (6.8) for every t. In addition, for every α, there exists a perturbation $\mathcal{P}_\epsilon$ of a Markov kernel $\mathcal{P}$ with $d(\mathcal{P}) = \alpha$ that satisfies Assumption 6.2.2 for which

$$||\Pi - \Pi_\epsilon||_{\mathrm{TV}} = \frac{\epsilon}{\alpha},$$

and a distinct perturbation $\tilde{\mathcal{P}}_\epsilon$ of $\mathcal{P}$ that achieves

$$\left|\left|\Pi - \frac{1}{t}\sum_{k=0}^{t-1} \nu \tilde{\mathcal{P}}_\epsilon^k \right|\right|_{\mathrm{TV}} = \frac{\left(1 - (1-\alpha_\epsilon)^t\right) ||\Pi_\epsilon - \nu||_{\mathrm{TV}}}{\alpha_\epsilon t}$$

with $\alpha_\epsilon = \alpha - 2\epsilon$. Finally, under the additional technical conditions that the operator

$$Fg(\theta) := \int g(\theta')\mathcal{P}(\theta, \theta')d\theta' \tag{6.12}$$

is self-adjoint and compact, then for every α there exists a function and a Markov chain such that

$$\frac{1}{||f||_*^2}\mathbb{E}\left[\left(\Pi f - \widehat{\Pi}^t f\right)^2\right] \geq S(t, \alpha),$$

is achieved for every t, and a (possibly distinct) Markov chain and function such that

$$\frac{1}{||f||_*^2}\mathbb{E}\left[\left(\Pi f - \widehat{\Pi}^t f\right)^2\right] = 4\left(\frac{\left(1 - (1-\alpha)^t\right) ||\Pi - \nu||_{\mathrm{TV}}}{\alpha t}\right)^2.$$

These conditions would hold, for example, for a reversible Markov chain on a finite state space.

Thus, the total variation bound for the exact kernel is sharp, and the two components of the L_2 bound for the exact kernel are achieved up to a factor of two in the first term in some special cases. Although the bounds for the approximate chain may not be sharp, the two components of the total variation bound are both

achieved. Thus, the bounds provide reasonable estimates of the supremum in (6.7) when $D = D_{TV}$ or D_{L_2}.

Both bounds for approximate kernels in Theorem 6.2.4 contain an asymptotic bias term. The relative performance is governed by the magnitude of this bias, differences in the worst-case convergence rate, and the speedup. For D_{L_2}, the bias is $\frac{4\epsilon^2}{\alpha^2}$ (obtained from expanding the squared term in (6.11), while for D_{TV}, it is ϵ/α. The D_{L_2} bound has additional terms involving ϵ that disappear in the infinite-time limit, which also arise from expanding the squared term. The convergence rate and worst-case autocorrelations for the approximation can be worse than that of the exact chain, since $\alpha_\epsilon < \alpha$; however, this will not always be the case, and in some cases approximate algorithms will have better mixing properties. As a result, the results that follow may understate the benefits of using $\mathcal{P}_\epsilon$. Similar results could be obtained for f with $||f||_\infty = \infty$ using either concentration and tail assumptions or moment assumptions, but the convergence rate in t and the scale of the bias would not change, so we retain the boundedness assumption throughout this section.

An important interpretation is that for relatively short path lengths, the bounds in (6.9) and (6.11) are dominated by terms related to the mixing/convergence properties of the chain, assuming ϵ is small relative to α. For (6.9), this is the term $\frac{(1-(1-\alpha_\epsilon)^t)||\Pi_\epsilon - \nu||_{\text{TV}}}{t\alpha_\epsilon}$, which is similar in magnitude to the bound in (6.8) when $\epsilon \ll \alpha$. This term decays with t, so that eventually the bias term ϵ/α becomes dominant. So for relatively short path lengths, there should exist a range of ϵ values for which aMCMC offers better performance in the compminimax sense. For longer path lengths, the values of ϵ for which an advantage persists will tend to be small relative to α.

A similar analysis applies to (6.11), where the leading term is $S(t, \alpha_\epsilon)$. This is effectively a variance term that is bounded by the covariances for worst case functions. For shorter path lengths, the variance term will dominate the overall estimation error

and aMCMC will offer better performance. For longer path lengths, the bias term $\frac{4\epsilon^2}{\alpha^2}$ is more important. One factor that is clear from (6.11) but not revealed by (6.9) is that aMCMC can still have a significant advantage when a burn-in period is used and the first t_b samples are discarded. Although this results in the term $||\Pi - \nu||_{\mathrm{TV}}$ being small – since we would now replace ν by νP^{t_b} – the leading term $S(t, \alpha)$ is unaffected. In other words, burn-in cannot cure the problem of high autocorrelations in a chain with small α.

6.2.3 Analysis of compminimax approximation error

We now apply Theorem 6.2.4 to analysis of the compminimax approximation error. In light of Remark 6.2.1, the bounds in (6.9) and (6.11) provide useful estimates of the supremum in (6.7) when $D = D_{TV}$ or $D = D_{L_2}$. In the sequel, we focus on these discrepancy measures, and approximate $\epsilon_c(\tau_{\max})$ for different values of $\tau_{\max}$ and functional forms for $s(\epsilon)$ by minimizing the upper bounds in (6.9) and (6.11). Since the bounds are tight for $\epsilon = 0$, when the analysis suggests that $\epsilon_c(\tau_{\max}) > 0$, it will always be the case that some $\epsilon > 0$ is compminimax; however, the optimal value may actually be larger than that computed by minimizing the (possibly loose) upper bounds for $\mathcal{P}_\epsilon$.

Empirical analysis of ϵ_c requires choices of α and $s(\epsilon)$. We consider values between $\alpha = 0.1$ and $\alpha = 10^{-4}$. These values are chosen by considering the upper bound on the δ-mixing time t_δ of the chain

$$t_\delta = \inf\{t : \left|\left|\nu P^t - \Pi\right|\right|_{\mathrm{TV}} < \delta\} \tag{6.13}$$

A corollary of Assumption 6.2.1 is that (6.13) is upper bounded by $\log(\delta)/\log(1-\alpha)$ when $||\nu - \Pi||_{\mathrm{TV}} = 1$. The corresponding worst-case δ-mixing times for a few values of δ and the four values of α considered are given in Table 6.1. This range of α values gives mixing times between about 45 and $92,000$ for $\delta \in (10^{-2}, 10^{-4})$, which reflects the empirical performance of many MCMC algorithms. In particular, a very rapidly

mixing MCMC algorithm may reach apparent stationarity in only a few iterations. On the other hand, it is not uncommon that MCMC algorithms for complex models may require a burn-in period of tens of thousands of iterations.[1]

Table 6.1: δ-mixing times for kernels with $d(\mathcal{P}) = \alpha$ for different values of α and δ.

	$\delta = 0.01$	$\delta = 0.001$	$\delta = 0.0001$
$\alpha = 0.1$	44	66	87
$\alpha = 0.01$	458	687	916
$\alpha = 0.001$	4,603	6,904	9,206
$\alpha = 0.0001$	46,049	69,074	92,099

We consider four functional forms for $s(\epsilon)$: logarithmic, linear, quadratic, and exponential. Constants are chosen such that $s(0) = 1$ and $s(\alpha/2) = 100$. Plots of the four functions for $\alpha = 10^{-4}$ are shown in Figure 6.1.

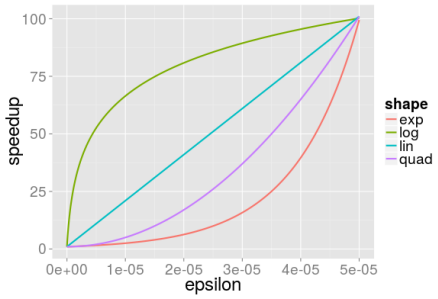


FIGURE 6.1*: Speedup functions used in analysis of compminimax.*

For each choice of $s(\epsilon)$ and a grid of values of $\tau_{\max} \in [1, 10^5]$, we approximate $\epsilon_c(\tau_{\max})$ by minimizing the upper bound in (6.9) with $t = s_\epsilon \tau_{\max}$, corresponding to our standing assuption that $\tau_{\mathcal{P}}(t) = t$. Results are summarized in Figure 6.2. The top two panels show results for D_{TV}. When $D = D_{TV}$, it is clear that over a range of values of $\tau_{\max}$ substantially larger than the mixing times, the optimal value of ϵ is

[1] We acknowledge that the criteria used to select burn-in times can result in burn-in periods that do not correspond to a mixing time, particularly when the posterior is strongly multimodal and the transition kernel has small conductance. However, comparing mixing times and burn-in periods still provides a useful heuristic, and in most cases violation of the criteria used to select a burn-in period is sufficient to guarantee that the chain has *not* mixed.

nonzero, regardless of the form of $s(\epsilon)$. As $\tau_{\max}$ increases, the (approximate) optimal value of ϵ decreases.

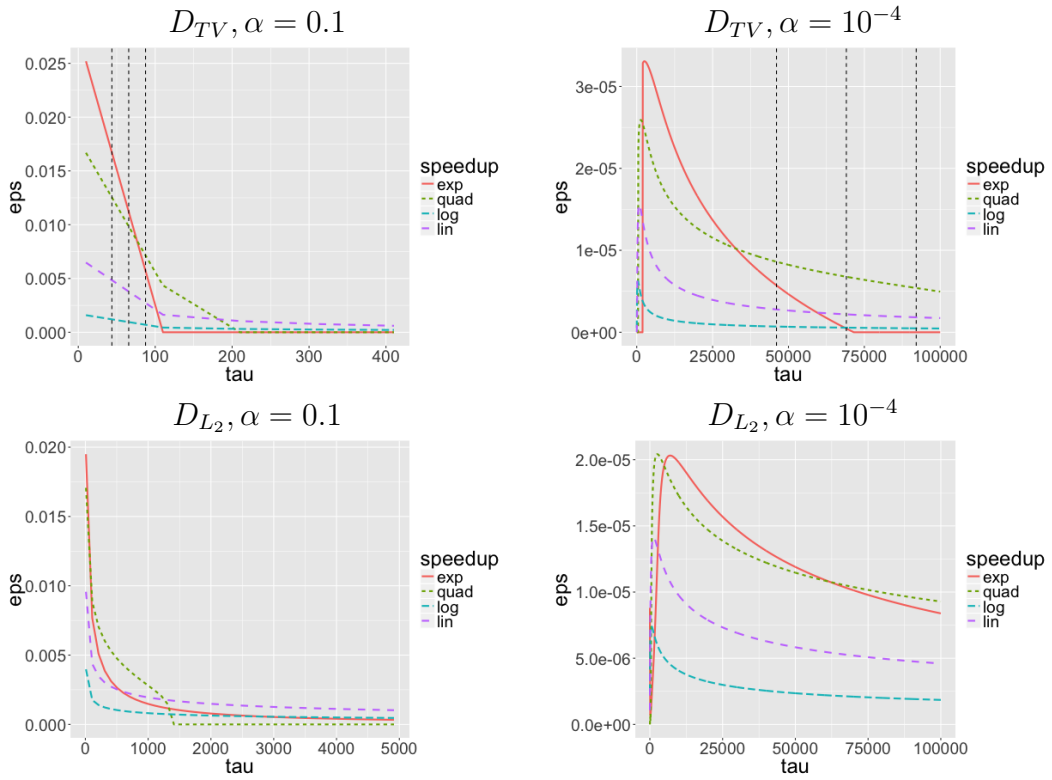


FIGURE 6.2: *Plot of $\epsilon_c(\tau_{\max})$ (vertical axis) for values of $\tau_{\max} \leq 10^5$ (horizontal axis), assuming $\tau_{\mathcal{P}}(t) = t$. Vertical dashed lines in the top two panels shown at the worst-case δ-mixing times for the values of δ shown in Table 6.1. Top two panels show results for D_{TV} and bottom two panels show results for D_{L_2}. Note different horizontal axis scale in the left top and bottom panels – the scales were chosen to make notable features more visible.*

To a first approximation, the results can be understood in terms of the second derivative of the speedup function and the fact that $d(\mathcal{P}_\epsilon)$ can be as small as $\alpha - 2\epsilon$. When the second derivative is positive, the computational benefit of increases in ϵ is largest for large values of ϵ, so for relatively small values of $\tau_{\max}$, the optimal ϵ is large. However, large values of ϵ incur a relatively high cost in terms of worst-case convergence rates and autocorrelations, since $d(\mathcal{P}_\epsilon)$ can be as small as $\alpha - 2\epsilon$. Thus, $\epsilon_c(\tau_{\max})$ goes to zero more rapidly when the second derivative of s_ϵ is positive. In contrast, for negative values, small values of ϵ offer relatively large computational

benefits. Notably, for all forms of $s(\epsilon)$ except the exponential, the optimal value of ϵ is nonzero for values of $\tau_{\max}$ greater than the δ-mixing times for all three values of δ considered. Finally, the observation that the optimal value of ϵ is zero for very small values of $\tau_{\max}$, then increases rapidly to its maximum value, is a result of the difference between the bounds on $d(\mathcal{P})$ and $d(\mathcal{P}_\epsilon)$. For small t, this has a significant effect on the upper bound in (6.9).

The bottom panel in Figure 6.2 shows results for $D = D_{L_2}$. In this case, we assume the chain starts close to its stationary distribution by putting $||\Pi_\epsilon - \nu||_{\mathrm{TV}} = 10^{-4}$ in (6.11). This corresponds to the situation in which a substantial number of burn-in samples are discarded. The choice of D_{L_2} instead of D_{TV} results in approximate values of $\epsilon_c(\tau_{\max})$ that are larger at every value of $\tau_{\max}$. Additionally, values of ϵ significantly larger than zero remain optimal well beyond the maximum value of t considered in each case ($5,000$ when $\alpha = 0.1$ and 10^5 when $\alpha = 10^{-4}$). This reflects the fact that high autocorrelations for worst-case functions make variance of MCMC ergodic averages the dominating factor in the L_2 error bounds even for relatively long sample paths, and these autocorrelations are unaffected by discarding a burn-in. Even when autocorrelations are relatively low, as in the case where $\alpha = 0.1$, nonzero ϵ is optimal for relatively large computational budgets when the speedup function is nonconvex.

6.3 Algorithm case studies

We apply the theoretical results of Section 7.2 to three approximate MCMC algorithms: for mixture models for contingency tables using approximations to Gibbs sampler full conditionals, for logistic regression based on subsets of data, and for Gaussian processes using low-rank approximations. For the first example, we verify both Assumption 2.1 and 2.2. For the other two examples, the focus is on verification of Assumption 6.2.2, which we show holds with high probability. An important

conclusion is that it is usually possible to construct kernels that satisfy Assumption 6.2.2 with high probability, but doing so requires adapting the approximation to the current state of the Markov chain.

6.3.1 Distributional approximations to full conditionals

We consider distributional approximations to full conditionals in Gibbs samplers, where the approximations rely on the complete data. The motivation is that sampling from the approximating distribution may be much faster, either because the sufficient statistics are cheaper to calculate or the sampling algorithm itself scales better in the number of observations.

The specific example we consider is a mixture model for contingency tables. Suppose we have p categorical variables $y = (y_1, \ldots, y_p)$, which for simplicity each take d possible values. We consider a variation on the model of Dunson and Xing (2009), replacing a stick-breaking prior with a Dirichlet:

$$Pr(y_1 = c_1, \ldots, y_p = c_p) = \pi_{c_1,\ldots,c_p} = \sum_{h=1}^{K} \nu_h \prod_{j=1}^{p} \lambda_{hc_j}^{(j)}, \tag{6.14a}$$

$$\lambda_h^{(j)} \sim \text{Dirichlet}(a_h^{(j)}), \quad \nu \sim \text{Dirichlet}(\alpha, \ldots, \alpha). \tag{6.14b}$$

In MCMC algorithms for discrete mixture models, it is common to employ data augmentation. Specifically, re-write the likelihood conditional on a latent class variable z as

$$Pr(y_{i1} = c_1, \ldots, y_{ip} = c_p | z_i = h) = \prod_{j=1}^{p} \lambda_{hc_j}^{(j)}, \quad Pr(z_i = h) = \nu_h.$$

When multiple observations with identical values of $y_1, \ldots, y_p$ exist, the data are more compactly represented as a d^p contingency table, where $n(\boldsymbol{c}) = \sum_{i=1}^{n} \prod_{j=1}^{p} \mathbb{1}_{\{y_{ij}=c_j\}}$ and $\boldsymbol{c} = (c_1, \ldots, c_p)$ is a multi-index identifying the cell of the table. Let $\mathcal{C}^+ = \{\boldsymbol{c} : n(\mathbf{c}) > 0\}$, and for each $\boldsymbol{c} \in \mathcal{C}^+$, let $Z(\boldsymbol{c})$ be a $1 \times K$ vector with entries $Z(\boldsymbol{c})_h = \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{y}_i=\boldsymbol{c}\}} \mathbb{1}_{\{z_i=h\}}$.

A Gibbs sampling algorithm for this model is given by

(1) For each $\boldsymbol{c} \in \mathcal{C}^+$, sample

$$Z(\boldsymbol{c})|\nu, \lambda, Y \sim \text{Multinomial}(n(\boldsymbol{c}), \tilde{\nu}), \quad \tilde{\nu}_h = \frac{\nu_h \prod_{j=1}^p \lambda_{hc_j}^{(j)}}{\sum_{l=1}^K \nu_l \prod_{j=1}^p \lambda_{lc_j}^{(j)}}. \tag{6.15}$$

(2) Sample $\lambda_h^{(j)}$ for $h = 1, \ldots, K$ and $j = 1, \ldots, p$ from

$$\lambda_h^{(j)} \sim \text{Dirichlet}\left(a_{h1}^{(j)} + \sum_{\boldsymbol{c}: c_j = 1} Z(\boldsymbol{c})_h, \ldots, a_{h1}^{(j)} + \sum_{\boldsymbol{c}: c_j = d_j} Z(\boldsymbol{c})_h\right).$$

(3) Sample ν from

$$\nu \sim \text{Dirichlet}\left(\alpha + \sum_{\boldsymbol{c}} Z(\boldsymbol{c})_1, \ldots, \alpha + \sum_{\boldsymbol{c}} Z(\boldsymbol{c})_K\right).$$

The dominating step is sampling of $Z(\boldsymbol{c})$, which has computational complexity increasing linearly in $n(\mathbf{c})$, so that each Gibbs iteration consumes at least order N operations just to sample the $Z(\boldsymbol{c})$. An approximating sampler that facilitates scaling to large N replaces the Multinomial sampling step for $Z(\boldsymbol{c})$ with the following procedure:

(1) Let $H = \{h : n(\boldsymbol{c})_h \tilde{\nu}_h > n_{\min}\}$, $K_H = |H|$, with $n_{\min}$ a pre-specified threshold. For any set $A \subset \{1, \ldots, K\}$ and K-vector v, define $v_A = \{v_h, h \in A\}$.

(2) For entries $h \in H$, sample from the Gaussian approximation to the multinomial,

$$W \sim \text{Normal}\left(n(\boldsymbol{c})\tilde{\nu}_H, n(\boldsymbol{c})[\text{diag}\left(\tilde{\nu}_H\right) - \tilde{\nu}_H \tilde{\nu}_H']\right),$$

and set $Z(\boldsymbol{c})_H$ equal to W with the elements rounded to the nearest integers.

(3) If $K_H < K$, sample $Z(\boldsymbol{c})_{H^c}$ from Multinomial $\left(n(\boldsymbol{c}) - \sum_{h' \in H} Z(\boldsymbol{c})_{h'}, \tilde{\nu}_{H^c}\right)$.

(4) Repeat steps (1)-(3) at every MCMC scan.

The Gaussian approximation can be sampled with computational complexity $\mathcal{O}(|\mathcal{C}^+|K^3)$, resulting in substantial speedup when N is large. The other sampling steps are unchanged. The possible values of $n_{\min}$ define a collection of approximate transition kernels $\mathscr{P}_\epsilon = \{\mathcal{P}_{\epsilon_0} : ||\mathcal{P}(\theta_0, d\theta) - \mathcal{P}_{\epsilon_0}(\theta_0, d\theta)||_{\mathrm{TV}} < \epsilon\}$. We allow rounding to negative integers in step (2) for convenience in proving Lemma 6.3.1, guaranteeing Assumption 6.2.2. In practice, negative integers very rarely occur, and in such cases we set them equal to zero.

Remark 6.3.1 (Uniform error bounds for normal approximations)**.** Consider any approximate MCMC algorithm that replaces some full conditionals in Gibbs steps with the discretized Gaussian approximation to the multinomial described in step (2). For every $\epsilon \in (0, 1)$ there exists $n_{\min}$ such that

$$\sup_{\theta \in \Theta} ||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}} < \epsilon,$$

where $\mathcal{P}_\epsilon(\theta, \cdot)$ corresponds to the algorithm with threshold $n_{\min}$. Moreover, there exist constants $C(K_H - 1)$ depending only on $K_H - 1$ for which

$$n(\boldsymbol{c}) > \frac{C(K_{H(\boldsymbol{c})} - 1)^2}{|\mathcal{C}^+|\epsilon^2} \left(\sum_{h \in H(\boldsymbol{c})} \frac{(1 - \tilde{\nu}_h)(1 - 2\tilde{\nu}_h + 2\tilde{\nu}_h^2)(1 + P_h/\tilde{\nu}_K)}{\sqrt{\tilde{\nu}_h(1 - \tilde{\nu}_h)}} \right)^2, \qquad (6.16)$$

for every $\boldsymbol{c} \in \mathcal{C}^+$ implies $||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}} < \epsilon$, where $P_h = \sum_{h' : \tilde{\nu}_{h'} > \tilde{\nu}_h} \tilde{\nu}_h$ and we have explicitly indicated the dependence of H on $\boldsymbol{c}$ through the notation $H(\boldsymbol{c})$.

Proofs of all results in this section are given in Appendix E. Remark 6.3.1 shows that by adapting the kernel to the current state, it is possible to construct an algorithm that satisfies Assumption 6.2.2. Adaptivity enters because the threshold $n_{\min}$ is applied to $n(\boldsymbol{c})\tilde{\nu}_h$, which depends on the current state. We include the condition in (6.16) to illustrate an interesting connection with condition numbers of covariance matrices of the parameters that will be revisited in later examples. Suppose that

$H = \{1, \ldots, K\}$ and let $\tilde{\nu}_{\max}, \tilde{\nu}_{\min}$ be the largest and smallest entries of $\tilde{\nu}$, respectively. The quantity on the right of (6.16) will be large when $\frac{\tilde{\nu}_{\max}}{\tilde{\nu}_{\min}}$ is large. The Multinomial$(n(\boldsymbol{c}), \tilde{\nu})$ distribution has covariance $n(\boldsymbol{c})(\text{diag}(\tilde{\nu}) - \tilde{\nu}\tilde{\nu}')$. Applying inequalities from Golub (1973), the smallest eigenvalue of this matrix is bounded below by $n(\boldsymbol{c})\tilde{\nu}_{\min}$, and the largest is bounded above by $n(\boldsymbol{c})\left(\tilde{\nu}_{\max} + \sum_{h=1}^{K} \tilde{\nu}_h^2\right)$. Thus the condition number is at least $\frac{\tilde{\nu}_{\max}}{\tilde{\nu}_{\min}}$, so the quantity on the right side of (6.16) will always be large when the condition number of the covariance is large, meaning that we require a larger sample in cell $\boldsymbol{c}$ for an accurate approximation. In fact, one way to think about the adaptive approximation is that by excluding categories h with small $\tilde{\nu}_h$ – thereby resulting in larger $\tilde{\nu}_{\min}$ – the covariance matrix in the Gaussian approximation is better conditioned, ensuring a more accurate approximation. Analogous conclusions are reached for the other two example algorithms in the sequel.

Remark 6.3.1 also allows for analysis of the order of the speedup function $s(\epsilon)$ for this algorithm. (6.16) shows that $\epsilon = \mathcal{O}\left(n^{-1/2}\right)$, so we need to increase the threshold at the rate $\sqrt{n}$ for linear decreases in ϵ. This requires substituting order n computation for order K^3 computation. To a first approximation, this indicates that the speedup function is roughly $s(\epsilon) = \sqrt{\epsilon}$.

Remark 6.3.2 shows that the above exact Gibbs sampler for model (6.14a)-(6.14b) satisfies Assumption 6.2.1. In this example, the state space is compact, and the latent variable Z is discrete, which makes verification of Assumption 6.2.1 fairly straightforward.

Remark 6.3.2 (Mixture model conditions)**.** The Gibbs sampling algorithm described above for the model in (6.14a)-(6.14b) satisfies Assumption 6.2.1.

6.3.2 Approximations based on subsets of data

A variety of aMCMC algorithms that utilize subsets of the data have been proposed. An example is provided in Korattikara et al. (2013), where $V \subset \{1, \ldots, N\}$ is a random subset of indices adaptively chosen to obtain a pre-specified type I error in a Metropolis-Hastings acceptance decision. We instead use a subsampling approximation of a covariance matrix within a Gibbs sampler for logistic regression. We are able to obtain theoretical guarantees on approximation error under weaker conditions on the data than Korattikara et al. (2013).

6.3.3 Model and computational algorithm

Consider a logistic regression model with a Gaussian prior on regression coefficients

$$y_i \sim \text{Bernoulli}\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right), \quad \beta \sim \text{Normal}\left(b, B\right). \tag{6.17}$$

Polson et al. (2013) show that the likelihood in this model satisfies

$$L(y_i \mid \beta) = \left(\frac{e^{y_i x_i \beta}}{1+e^{x_i\beta}}\right) \propto \exp(\kappa_i x_i \beta) \int_0^\infty \exp\left\{-\omega_i (x_i\beta)^2/2\right\} p(\omega_i \mid 1, 0),$$

where $\kappa_i = y_i - 1/2$ and $p(\omega_i \mid 1, 0)$ is the density of a *Pólya-Gamma* random variable with parameters 1 and 0, which we represent as $\text{PG}\left(1, 0\right)$. This results in the Gibbs sampler:

$$\omega_i \mid \beta \sim \text{PG}\left(1, x_i\beta\right) \tag{6.18a}$$

$$\beta \mid y, \omega \sim \text{Normal}\left(m_N, S_N\right), \tag{6.18b}$$

where $S_N = (X'\Omega X + B^{-1})^{-1}$, $m_N = S_N(X'\kappa + B^{-1}b)$, and $\Omega = \text{diag}(\omega_1, \ldots, \omega_N)$.

Approximation

When N is large and p – the dimension of β – is moderate, the main computational bottleneck is calculating $X'\Omega X$. This step has computational complexity $\mathcal{O}(Np^2)$.

An approximating Markov chain that uses the likelihood approximation described above will reduce the computational complexity of each step to $\mathcal{O}(|V|p^2)$, a large computational speedup when $|V| \ll N$. However, assuming a random subset V is chosen at each iteration, the estimated posterior variance of β will be inflated.

We analyze an approximating Gibbs sampler with the update rule

$$V \mid \beta \sim \text{Subset}(|V|, \{1, \ldots, N\}), \tag{6.19a}$$

$$\omega_i \mid \beta, V \sim \text{PG}\left(1, x_i\beta\right), \qquad i \in V, \tag{6.19b}$$

$$\beta \mid y, \omega, V \sim \text{Normal}\left(S_V X'\kappa, S_V\right), \tag{6.19c}$$

where $S_V = \left(\frac{N}{|V|} X_V' \Omega_V X_V + B^{-1}\right)^{-1}$ uses a subsample-based approximation to $X'\Omega X$ and $|V|$ may depend on β. Choi and Hobert (2013) showed that the algorithm in (6.18a)-(6.18b) is uniformly ergodic. Theorem 6.3.1 shows that if $|V|$ is chosen adaptively depending on β, Assumption 6.2.2 is satisfied with high probability at any step of the chain.

Theorem 6.3.1 (Error for random subset approximations)**.** *Suppose the rows of X are iid realizations with a log-concave density that is symmetric about the origin. Let $b = 0$, $B = \eta I_p$ for $\eta > 0$. Let $\mathcal{P}(\theta, \cdot)$ be the transition kernel of Gibbs sampler (6.18a) - (6.18b), and $\mathcal{P}_\epsilon(\theta, \cdot)$ be the transition kernel of sampler (6.19a)-(6.19c). Then, for every $\epsilon > 0$, there exists a kernel $\mathcal{P}_\epsilon(\theta, \cdot)$ that sets $|V| \leq N$ as a function of β, for which*

$$\sup_{\theta \in \Theta} ||\mathcal{P}_\epsilon(\theta, \cdot) - \mathcal{P}(\theta, \cdot)||_{\text{TV}} \leq \epsilon$$

with probability $(1-q)^2$, where q decreases exponentially in $|V|$.

The following remark provides insight into the achievable rates in $|V|$ of the probabilities q and the approximation error ϵ.

Remark 6.3.3 (Rates in Theorem 6.3.1)**.** Let $\omega \sim \mathrm{PG}(1, x\beta)$ and $x = (x_1, \ldots, x_p)$ denote realizations for a random subject, and

$$\Sigma = \mathrm{cov}(\omega^{1/2} x \mid \beta), \quad \Sigma_N = \frac{1}{N} X'\Omega X, \quad \Sigma_V = \frac{1}{|V|} X_V'\Omega_V X_V.$$

Choosing $|V| \geq pCM^4\delta^{-2}\log^2(2M^2\delta^{-2})$, $||\Sigma_N - \Sigma_V|| < \delta\,||\Sigma||$ with probability $(1 - e^{-cM\sqrt{p}})^2$, where C and c are absolute constants. Subsets of size $|V| = \mathcal{O}\big(\sqrt{N}\big)$ result in $M \approx \frac{N^{1/8}}{\log^2(2N^{1/4})}$, achieving q slightly larger than $e^{-N^{1/8}}$. The required value of δ to achieve $||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}} < \epsilon$ is

$$\delta = \frac{2\sqrt{2}\epsilon p^{-1/2}(\lambda_{\min}(\beta)/2)^2}{\lambda_s(\beta)^{3/2}} \wedge \frac{\epsilon^2(\lambda_{\min}(\beta)/2)}{p\lambda_s(\beta)} \wedge \frac{\epsilon p^{-1/2}\lambda_{\min}(\beta)}{\lambda_{\min}(\beta) + \lambda_{\max}(\beta)}, \tag{6.20}$$

where $\lambda_s(\beta) = \lambda_{\max}(\beta) + \lambda_{\min}(\beta)/2$ and $\lambda_{\min}(\beta)$ and $\lambda_{\max}(\beta)$ are the smallest and largest eigenvalues of $\Sigma(\beta)$, respectively.

Remark 6.3.3 shows that the required subset size to achieve approximation error ϵ with probability $(1 - q)^2$ depends on the reciprocal condition number of $\Sigma = \mathrm{cov}(\omega^{1/2} x \mid \beta)$. This is similar to the adaptivity result in Remark 6.3.2, and again indicates that algorithms achieving uniform bounds must be adaptive to the current state. Although the exact rates include unknown constants c and C, the result is still useful in constructing a practical algorithm. The condition number of Σ can be approximated by computing Σ_V, which is required by the subsetting algorithm anyway. An adaptive subsample size can be chosen by starting with a subsample of fixed size, computing the condition number of Σ_V, and increasing the subset size when this value is large relative to, say, its ergodic average.

Remark 6.3.3 also characterizes the speedup function for this algorithm. The computational cost of increasing $|V|$ is $\mathcal{O}(|V|p^2)$, while to a second-order approximation $\delta = \mathcal{O}\big(|V|^{-1/2} + \log V\big)$. Moreover, (6.20) shows that $C_\beta\epsilon^2 < \delta < C_\beta\epsilon$, so $\epsilon = \mathcal{O}\big(|V|^{-1/2} + \log V\big)$ as well. Therefore, the speedup function is $\mathcal{O}(\sqrt{\epsilon})$ up to log factors.

6.3.4 Low-rank approximations to Gaussian processes

Exact MCMC algorithms involving Gaussian processes scale as $\mathcal{O}(n^3)$, leading to numerous proposals for approximations. Prominent examples include the predictive process (Banerjee et al. (2008)) and subset of regressors (Smola and Bartlett (2001)), which both employ low-rank approximations to the Gaussian process covariance matrix.

Model

Consider the nonparametric regression model

$$y_i = f(x_i) + \eta_i, \qquad \eta_i \sim \text{Normal}\left(0, \sigma^2 I_n\right), \qquad i = 1, \ldots, n, \tag{6.21}$$

where y_i are responses, $x_i \in \mathcal{X}$ are $p \times 1$ covariate vectors, and f is an unknown function. A typical Bayesian approach assigns a Gaussian process prior to f, $f \sim \text{GP}(\mu(\beta), c(\gamma))$, with $\mu(\cdot;\beta)$ a mean function with parameter β and $c(\cdot,\cdot;\gamma)$ a covariance function parametrized by γ, so that for $x_1, x_2 \in \mathcal{X}$, $\mathbb{E}[f(x_1)] = \mu(x_1;\beta)$ and $\text{cov} f(x_1) f(x_2) = c(x_1, x_2; \gamma)$. Here we will assume $\mu(x;\beta) = 0$ for all $x \in \mathcal{X}$, so that the model parameters consist of $\theta = (\sigma^2, \gamma)$. Although we focus on model (6.21), our analysis applies to general settings involving Gaussian processes (e.g., for spatial data).

The covariance kernel $c(x_1, x_2; \gamma)$ is positive definite, so that the $n \times n$ covariance matrix S given by $S_{ij} = c(x_i, x_j; \gamma)$ is full rank. However, as noted by Banerjee et al. (2013), in many cases when n is large, the matrix S is poorly conditioned and nearly low-rank. This motivates low-rank approximations of S. As an example, consider the squared exponential kernel $c(x_1, x_2; \gamma) = \tau^2 \exp(-\phi \left|\left|x_1 - x_2\right|\right|^2)$, with $\gamma = (\tau^2, \phi)$ consisting of a decay parameter ϕ and scale τ^2. In this case we write $S = \tau^2 \Sigma$, where $\Sigma_{ij} = \exp(-\phi \left|\left|x_i - x_j\right|\right|^2)$, and we have $\theta = (\sigma^2, \tau^2, \phi)$. We adopt the common prior structure

$$\phi \sim \text{DiscUnif}(\phi_1, \ldots, \phi_d),\ \tau^{-2} \sim \text{Gamma}\left(a_\tau, b_\tau\right),\ \sigma^{-2} \sim \text{Gamma}\left(a_\sigma, b_\sigma\right). \tag{6.22}$$

We consider a marginal MCMC sampler (e.g. Finley et al. (2009)) which iterates

1. Sample $\sigma^{-2}|y, \tau^2, \phi$ using a Metropolis-Hastings step with random walk on $\log(\sigma^2)$ as a proposal.

2. Sample $\tau^{-2}|y, \phi, \sigma^2$ using a Metropolis-Hastings step with random walk on $\log(\tau^2)$ as a proposal.

3. Set $p_l \propto \det\left(\tau^2\Sigma^{(l)} + \sigma^2 I_n\right)^{-1/2} \exp\left(-y'(\tau^2\Sigma^{(l)} + \sigma^2 I_n)^{-1}y\right)$, where $\Sigma_{ij}^{(l)} = \exp(-\phi_l ||x_i - x_j||^2)$, and sample

$$\phi \sim \text{Disc}\left(\{\phi_1, \ldots, \phi_d\}, (p_1, \ldots, p_d)\right).$$

Approximate MCMC for Gaussian processes

We replace Σ with a low-rank approximation $\Sigma_\epsilon \approx \Sigma$ to construct a transition kernel $\mathcal{P}_\epsilon(\theta, \cdot)$. We focus on approximations of the form

$$\Sigma \approx U_\epsilon \Lambda_\epsilon U_\epsilon' = \Sigma_\epsilon, \tag{6.23}$$

where U_ϵ is orthonormal, and Λ_ϵ is nonnegative and diagonal.

All of the steps of the MCMC sampler contain the quadratic form $y'(\tau^2\Sigma + \sigma^2 I)^{-1}y$, and the process f is sampled from

$$p(f \mid y, \theta) \sim N(\Psi y, \Psi), \qquad \Psi = (\tau^2\Sigma + \sigma^2 I)^{-1}$$

to obtain interval estimates. The approximation instead uses

$$p_\epsilon(f \mid y, \theta) \sim N(\Psi_\epsilon y, \Psi_\epsilon), \qquad \Psi_\epsilon = (\tau^2\Sigma_\epsilon + \sigma^2 I)^{-1}.$$

For algorithms in this class, we obtain the result in Theorem 6.3.2.

Theorem 6.3.2 (Gaussian process approximation error bounds)**.** *Suppose data are generated according to (6.21), with* $c(x_1, x_2; \gamma) = \tau^2 \exp(-\phi ||x_1 - x_2||^2)$*, and* $\Sigma_\epsilon =$

$U_\epsilon \Lambda_\epsilon U_\epsilon'$ for U_ϵ a $n \times r$ matrix and Λ_ϵ a $r \times r$ matrix with $r < n$. For every $\epsilon > 0$ there exists $\delta > 0$, which depends on θ, such that if $||\Sigma - \Sigma_\epsilon||_F < \delta$,

$$||p(f \mid y, \theta) - p_\epsilon(f \mid y, \theta)||_{\mathrm{TV}} < \epsilon. \tag{6.24}$$

If additionally Σ_ϵ is a partial rank-r eigendecomposition of Σ and a joint Metropolis-Hastings step is used for (σ^2, τ^2), then for every $\epsilon > 0$, there exists a $\mathcal{P}_\epsilon(\theta, \cdot)$ that replaces Σ with Σ_ϵ achieving $||\Sigma - \Sigma_\epsilon|| < \delta$ with probability $1 - q$, where δ depends on θ, such that

$$\sup_{\theta \in \Theta} ||\mathcal{P}_\epsilon(\theta, \cdot) - \mathcal{P}(\theta, \cdot)||_{\mathrm{TV}} < \epsilon \tag{6.25}$$

with probability $1 - q$.

In practice, although we cannot calculate an exact partial eigendecomposition, Algorithm 2 of Banerjee et al. (2013) provides an accurate approximation, which is equivalent to the *adaptive randomized range finder* (Algorithm 4.2) combined with the *eigenvalue decomposition via Nyström approximation* (algorithm 5.5) in Halko et al. (2011). Algorithm 2 attains approximation error $||\Sigma - \Sigma_\epsilon||_F < \delta$ with probability $1 - 10^{-d}$ where both δ and d can be specified. We provide empirical evidence that the partial eigendecomposition approximation is accurate in Appendix E. Not all low-rank approximations of Σ approximate a partial eigendecomposition, so Theorem 6.3.2 suggests an advantage of Algorithm 2 of Banerjee et al. (2013) over alternatives.

The following remark describes the achievable rates in δ as a function of ϵ and n.

Remark 6.3.4 (Rates for aMCMC for Gaussian process)**.** The value of δ for (6.24) is

$$\delta = \frac{\epsilon^2 \sigma^4}{\tau^2 \sqrt{n(\tau^2 \lambda_{\max}(\Sigma_\epsilon) + \sigma^2)}} \wedge \frac{\epsilon^2 \sigma^2}{\tau^2},$$

where $\lambda_{\max}(\Sigma_\epsilon)$ is the largest eigenvalue of Σ_ϵ. Controlling δ to satisfy Assumption 6.2.2 requires that

$$\left| \exp\left(-\frac{n - r}{2} \left[\frac{\sigma^2 - \sigma_*^2}{\sigma_*^2 \sigma^2} - \log \frac{\sigma_*^2}{\sigma^2} \right] \right) \right.$$

$$- \exp\left(-\frac{n-r}{2}\left[\frac{\tau^2\delta + \sigma^2 - \tau_*^2\delta - \sigma_*^2}{(\tau_*^2\delta + \sigma_*^2)(\tau^2\delta + \sigma^2)} - \log\frac{\tau_*^2\delta + \sigma_*^2}{\tau^2\delta + \sigma^2}\right]\right)\Bigg|$$

be small, where σ_*^2, τ_*^2 are the proposed values of σ^2, τ^2 in the Metropolis-Hastings algorithm. To achieve constant approximation error, δ must decrease with n; if the spectrum of Σ_ϵ decays rapidly, the decrease can be slow. In addition, a smaller value of δ is required when τ^2 is large relative to σ^2, suggesting that a higher signal to noise ratio requires better approximations.

Remark 6.3.4 implies that for the weaker (6.24) to hold, $\epsilon = \mathcal{O}\big(\sqrt{\delta}\big)$; no effective estimate is available from the proof of (6.25). The algorithm scales as $\mathcal{O}(n^2 r)$, where r is the rank of Σ_ϵ, so increasing r to achieve a better approximation has computational cost n^2. However, the relationship between r and δ – and therefore between r and ϵ – depends on the spectrum of Σ. If, for example, $\lambda_r \propto e^{-r}$, then the speedup will be exponential. At the other extreme, the speedup could easily be concave if the spectrum decays too slowly. This is ultimately an empirical question, and is revisited in Appendix E when a specific dataset is analyzed.

6.4 Computational example

The results of Section 6.2 show how to predict the compminimax approximation error ϵ for a discrepancy measure D given α, the speedup function $s(\epsilon)$, and a computational budget τ. Section 6.3 was devoted to showing Assumption 6.2.2, and in one case Assumption 6.2.1. We also obtained the shape of the speedup functions for the three algorithms considered. In this section, we apply the subsetting aMCMC algorithm for logisitic regression to a real dataset and compare the results to the predictions of Section 6.2. Empirical studies of the other two algorithms – including consideration of other discrepancy measures – can be found in Appendix E.

6.4.1 Estimation of convergence rate and approximation error

In theorem 6.3.1, we showed that by controlling the subset size adaptively, the subsetting approximation for logistic regression satisfies Assumption 6.2.2 with high probability, and the exact algorithm is uniformly ergodic Choi and Hobert (2013). However, these results give only bounds on the relevant parameters, and ideally practitioners can apply the theory of aMCMC without showing results for every algorithm and associated approximation. To do this, one must estimate the convergence rate of the chain and the accuracy of the approximation empirically. We suggest one approach to this.

The main interest is in approximating the value of $1 - \alpha$ in Assumption 6.2.1 and the value of ϵ in Assumption 6.2.2. Although some MCMC algorithms may not satisfy Assumption 6.2.1, most will satisy geometric ergodicity.

Definition 6.4.1 (Geometric ergodicity)**.** A Markov chain evolving according to a transition kernel $\mathcal{P}(\theta; \cdot)$ on a state space Θ with invariant measure Π is geometrically ergodic if there exist constants $\rho \in (0, 1)$ and $B < \infty$ and a function $V : \Theta \to [1, \infty)$ such that

$$\left|\left|\mathcal{P}^k(\theta_0; \cdot) - \Pi\right|\right|_{\mathrm{TV}} \leq BV(\theta_0)(1 - \alpha)^k.$$

The Doeblin condition in Assumption 6.2.1 corresponds to the special case of $B = 1$, $V(\theta) = 1$, and uniform ergodicity corresponds to $V(\theta) = 1$. The parameter $(1 - \alpha)$ is the geometric convergence rate of the chain, which determines its mixing and convergence properties. Unlike uniformly mixing chains, the properties of geometrically ergodic chains depend on the initial state in a manner determined by the function V. When the chain is rapidly mixing, so that $1 - \alpha \ll 1$, we expect that aMCMC will outperform the exact chain only for short computation times, while when $1 - \alpha$ is near one aMCMC will dominate for relatively long computation times.

On the other hand, relatively large approximation error ϵ may be tolerable when $1-\alpha$ is small.

We suggest an approach to estimating $1-\alpha$ based on sample path autocorrelations similar to that described in Yang and Dunson (2013); details of the procedure are given in Appendix E. Using this procedure, we obtain an estimate of

$$\widehat{\varphi}_{\max} = \max_{j \leq p} \max_{k \leq k_{\max}} \widehat{\varphi}_{j,k}^{1/k}, \tag{6.26}$$

where $\widehat{\varphi}_{j,k}$ is an estimate of the lag-k autocorrelation for the jth component of θ and $k_{\max} \ll k$. This provides an estimated lower bound on $1-\alpha$. In special cases described in Appendix E, this bound is tight.

Estimates of ϵ are also of interest. If $\mathcal{P}$ satisfies Assumption 6.2.1 and $\mathcal{P}_\epsilon$ satisfies Assumption 6.2.2, then $||\Pi - \Pi_\epsilon||_{\mathrm{TV}} \leq \frac{\epsilon}{\alpha}$, so point estimates of α and $||\Pi - \Pi_\epsilon||_{\mathrm{TV}}$ give a plug-in lower bound on ϵ. If these assumptions are not satisfied, we would still expect $||\Pi - \Pi_\epsilon||_{\mathrm{TV}}$ to provide an indication of the magnitude of ϵ, with smaller values suggesting more accurate approximations. Unfortunately, estimating the total variation distance between two distributions on the basis of samples is difficult in moderate to high dimensions, and the estimates tend to be very noisy. Instead, Minsker et al. (2014) suggest a sample estimate of the Wasserstein-1 distance with respect to a metric kernel K, which we denote $W_{1,d_K}(P,Q)$ for probability measures P, Q. The estimate is robust relative to estimates of the total variation distance in multiple dimensions and easy to compute when $p > 1$. We choose the kernel to give $0 < d_K < 1$ and $W_{1,d_K}(P,Q) = 0$ if and only if $P = Q$. Since $d_K \leq 1$, this provides a *lower bound* on the total variation distance via $2W_{1,d_K}(P,Q) \leq ||P - Q||_{\mathrm{TV}}$. Details of this metric are provided in Appendix E.

6.4.2 Logistic regression using subsets

We applied the sampler in (6.19a)-(6.19c) to the SUSY dataset Baldi et al. (2014). The dataset consists of 4.5 million observations of a binary outcome with nine con-

tinuous covariates. To validate the general applicability of the empirical approach and assess the subset algorithm in a more challenging applied setting, we consider a hierarchical prior structure on β, specifically

$$\beta_j \sim N(0, \kappa^2\lambda_j^2), \quad \kappa \sim C_+(0,1), \quad \lambda_j \sim C_+(0,1),$$

where $C_+(0,1)$ is the Cauchy distribution with location 0 and scale 1 restricted to the real positive half-line. This is referred to as the Horseshoe shrinkage prior **?**. Sampling κ and λ_j adds two slice sampling steps to the update rule (see the online supplement to **?**). This provides an opportunity to demonstrate the applicability of our empirical approach to analyzing the properties of an approximate algorithm that has not been characterized theoretically. As before, $\mathcal{P}_\epsilon$ corresponds to using subsamples of data of size $|V|$ to approximate $X'\Omega X + B^{-1}$, where here $B = \text{diag}(\kappa^2\lambda_j^2)$. We use a fixed subsample size $|V|$ at every iteration. To increase model complexity, we also add two and three way interactions, resulting in $p = 92$.

The approach in Section 6.4.1 was used to approximate $1-\alpha$ and values of ϵ corresponding to different subset sizes $|V|$. Results are shown in Table 6.2. The values of $\widehat{\varphi}_{\max}$ are all about 0.98, giving an approximate value of α of 0.02. This corresponds to relatively rapid mixing: a chain with $\alpha = 0.02$ that satisfies the Doeblin condition has δ-mixing time with $\delta = 0.01$ of only 228 iterations. As expected, $\widehat{W}_{1,d_K}(\Pi_\epsilon, \Pi)$ decreases as $|V|$ increases. The last row of Table 6.2 gives a rough approximation to ϵ from $\widehat{W}_{1,d_K}(\Pi_\epsilon, \Pi)(1-\widehat{\varphi}_{\max}) \approx \alpha\,||\Pi - \Pi_\epsilon||_{\text{TV}}$, where $\widehat{\varphi}_{\max}$ is computed for the exact algorithm. Most values are small relative to $1-\widehat{\varphi}_{\max}$; however, for $|V| = 1,000$ they are approximately equal, so the theory suggests that this subset size may be too small to give useful posterior inference. Based on these estimates, the exact algorithm can be characterized as rapidly mixing, and there is no evidence that mixing degrades as ϵ increases. In Section 6.3, we showed that the speedup function for the subsetting algorithm is concave, given by $s(\epsilon) \propto \epsilon^{1/2}$. For rapidly mixing chains with concave

speedup functions, the theory predicts that moderate values of ϵ will be optimal over very short computational budgets, with small values of ϵ optimal for longer computation time.

Table 6.2: Estimates of $\widehat{\varphi}_{\max}$ and $W_{1,d_K}(\Pi_\epsilon, \Pi)$ for logistic regression aMCMC with different minibatch sizes.

$\lvert V\rvert$	$1,000$	$10,000$	$50,000$	10^4	5×10^5	4.5×10^6
$\widehat{W}_{1,d_K}(\Pi_\epsilon, \Pi)$	0.9843	0.2853	0.0474	0.0326	0.0063	0.0000
$\widehat{\varphi}_{\max}$	0.9781	0.9812	0.9796	0.9765	0.9768	0.9767
$\widehat{W}_{1,d_K}(\Pi_\epsilon, \Pi)(1-\widehat{\varphi}_{\max})$	0.0229	0.0066	0.0011	0.0008	0.0001	0.0000

We now assess this prediction by computing estimates of D_{TV} and D_{L_2} for this algorithm for different computational budgets. The value of D_{TV} is approximated by

$$\widehat{W}_{1,d_K}(\beta, t_0, \epsilon) = \widehat{W}_{1,d_K}\left(\frac{1}{t_0}\sum_{k=1}^{t_0}\delta_{\beta_k^{(\epsilon)}}, \frac{1}{t}\sum_{k=1}^{t}\delta_{\beta_k}\right),$$

for different sample path lengths t_0, where $t = 1,200$ is the full length of the chain for the exact algorithm, which consumed wall clock time of about 12 hours. The values of t_0 are converted into wall clock times. Similarly, D_{L_2} is approximated by

$$\mathrm{RMSE}(\beta, t_0, \epsilon) = \left(\sum_{j=1}^{p}\left(\frac{1}{t_0}\sum_{k=1}^{t_0}\beta_{k,j}^{(\epsilon)} - \frac{1}{t}\sum_{k=1}^{t}\beta_{k,j}\right)^2\right)^{1/2}.$$

The calculations of RMSE use burn-in of 200 iterations.

The left panel of Figure 6.3 shows $\mathrm{RMSE}(\beta)$ as a function of computation time τ in seconds for different $|V|$. Because the calculations use burn-in, for larger sample sizes the graph originates away from $\tau = 0$. For very small computational budgets, the optimal subset size with respect to D_{L_2} among those considered is $|V| = 50,000$ or $|V| = 100,000$, corresponding to relatively large ϵ. For computational budgets between about 15 minutes and two hours, the largest subset size $|V| = 500,000$

is optimal. For larger computational budgets, the exact algorithm is optimal. It is likely that for computational budgets exceeding two hours, some $|V|$ satisfying $500,000 < |V| < 4,500,000$ is optimal. Conversely, the estimates with $|V| = 10,000$ are much less accurate, and the scale of RMSE so different for $|V| = 1,000$ that the result is not shown. Based on table 6.2, $|V| = 1,000$ results in $\epsilon \approx \alpha$. Recalling that the results in Section 6.2 required $\epsilon < \alpha/2$, it is unsurprising that performance is very poor. All of these results are consistent with the predictions of Section 6.2 for fast-mixing chains with concave speedup function.

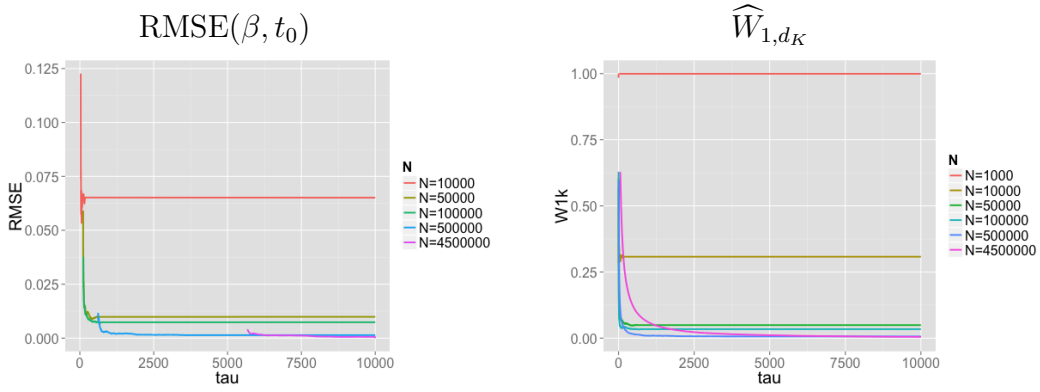


FIGURE 6.3: *Logistic regression RMSE for estimation of β (left) and approximate W_{1,d_K} distance to the exact posterior (right) as a function of computation time τ in seconds.*

In the right panel are lower-bound estimates of D_{TV} using $\widehat{W}_{1,d_K}(\beta, t_0, \epsilon)$. For consistency with Figure 6.2, these estimates do not use burn-in. The smallest mini-batch size $|V| = 1,000$ has $\widehat{W}_{1,d_K} \approx 1$ for any computational budget, which is again consistent with the result in Table 6.2 that $\epsilon \approx \alpha$ for this subset size. The other sample sizes all result in meaningful approximation to the posterior by this metric. The ranges of computation times for which $|V| = 50,000$, $|V| = 100,000$, and $|V| = 500,000$ are optimal are similar to those for D_{L_2}, with $|V| = 500,000$ being optimal for computation times between about 15 minutes and two hours. For budgets greater than two hours, $|V| = 500,000$ still provides a very accurate approximation. These results are again very consistent with the predictions of Section 6.2 based on

the shape of the speedup function and empirical estimates of ϵ and α.

6.5 Discussion

An important and interesting finding of this work is that in many cases, exact MCMC may not be statistically optimal when an approximation that offers computational advantages is available. aMCMC offers longer sample paths at the same computational cost, which, when the approximation is fairly accurate, can easily outweigh the effects of bias and potentially slower convergence of aMCMC. This tradeoff is formalized through the characterization of speedup functions and the compminimax notion of optimal approximation error. It has long been recognized in optimization that noisy gradients are often far superior to exact gradients, but this rationale has only recently entered into the MCMC literature (Korattikara et al. (2013), Ahn et al. (2012)), which has mainly pursued approximations when exact MCMC is considered computationally intractable. Another way to view this possibly surprising result is as a parallel to the well-characterized phenomenon that biased estimators have lower risk than unbiased estimators (see e.g. Stein (1956), James and Stein (1961)). The superiority of aMCMC with respect to D_{L_2} for large computational budgets is conceptually similar – that is, biased Markov chains can often have superior statistical properties to those of unbiased ones – though in this context the optimal level of bias depends on the computation time and the fundamental reason for the improved performance is quite different. That approximate MCMC may offer optimal performance for sample path lengths that exceed those found in typical applications of MCMC in Bayesian statistics suggests there is much room for expanding the use of aMCMC.

The theory of approximate MCMC provides a guide to what can go wrong when approximate kernels are employed, and how to check whether difficulties are likely to occur. This is exemplified by our consideration of low-rank approximations for

Gaussian processes and logistic regression with subsets. Optimal algorithms for a particular model, regardless of the convergence properties of the original chain, can only be approximately determined through numerical approximation of the constants or obtaining theoretically upper and lower bounds. These issues are not conceptually different from the long-standing issue of empirically assessing MCMC convergence, and are important problems for which no definitive solution currently exists. In the interim, the results presented here should provide a level of comfort for practitioners that approximate MCMC algorithms can often result in better performance in statistical estimation with limited computational resources.

7

Large-sample Efficiency of Data Augmentation Gibbs Sampling for Binary Outcomes

Gibbs sampling (Gelfand and Smith (1990), Geman and Geman (1984)), along with various extensions (see Gamerman and Lopes (2006), Robert and Casella (2013)), is arguably the dominant paradigm for computation in Bayesian statistics. Unlike alternatives, these algorithms require little tuning, and Gibbs samplers for a wide range of popular models have been proposed in the literature, making them accessible to practitioners. The simplicity and broad utility of the Gibbs sampler have undoubtedly contributed to the growth of Bayesian statistical methods over the past 25 years.

The efficiency of Gibbs sampling and other Markov chain Monte Carlo (MCMC) algorithms is often studied in terms of three quantities: (1) the computational complexity of taking a single step from the transition kernel of the associated Markov chain, (2) the number of "burn-in" steps required for the law of the chain to be "close" to its stationary measure, and (3) the "effective sample size" of an MCMC sample taken at or near stationarity. (1) can be studied through standard algorithmic

complexity analysis, while (2) and (3) are studied in terms of the mixing properties of the Markov chain and can be described in terms of objects such as the *convergence rate* and the *autocorrelation function* of the chain. The literature on the theory of MCMC convergence for statistical computation has emphasized studying the mixing properties of Markov chains by proving a *geometric ergodicity* condition (see Meyn and Tweedie (1994), Roberts and Rosenthal (1997), Kontoyiannis and Meyn (2003), Meyn and Tweedie (1993)). Such a condition gives bounds on both the required burn-in time and the effective sample size of an MCMC run.

These three quantities contribute to the overall *running time* of the MCMC algorithm as a function of the sample size n. For example, if an MCMC algorithm has computational complexity per step of $\mathcal{O}(n)$, takes $\mathcal{O}(n)$ steps to "burn-in," and has $\mathcal{O}(n^{-1})$ effective sample size so that the number of required samples near stationarity is $\mathcal{O}(n)$, then the running time of the algorithm is $\mathcal{O}(n^2)$, since we require $\mathcal{O}(n)$ samples to obtain a Monte Carlo estimate with fixed error $0 < \epsilon < \infty$ and it takes $\mathcal{O}(n)$ clock time to obtain a single sample. Historically, there has been a focus on finding algorithms that have $\mathcal{O}\big(n^k\big)$, $0 < k < \infty$ (polynomial time) running times, with $\mathcal{O}(e^{cn})\,, c > 0$ (exponential time) considered computationally intractable. It is often suggested that in "big data" contexts, most polynomial time algorithms are too costly, and only $\mathcal{O}\big(n \log^k n\big)$ (quasilinear time) algorithms should be viewed as tractable. Our results are consistent with this latter view.

Studying the computational efficiency of MCMC algorithms as the sample size n increases is an important problem, particularly in the big data setting, where very large sample sizes are often encountered. Previous work on this topic includes Belloni and Chernozhukov (2009), which shows running times that are polynomial in n when the posterior obeys the Bernstein von-Mises theorem, under a warmness condition for the initial state distribution. Yang et al. (2015) shows existence of a MCMC algorithm for computing the posterior for sparse linear models with g-priors with mixing

times that are polynomial in n. Rajaratnam and Sparks (2015) shows that when p grows faster than n, the geometric convergence rate of some MCMC algorithms tends to zero. Other studies have showed mixing times that are exponential in n in a variety of problems (e.g. Mossel and Vigoda (2006)). A commonality among these studies is the extension of computational complexity analysis common in computer science to MCMC by bounding mixing times as a function of n, or showing the order of the geometric convergence rate in n. Implicit in the focus of Belloni and Chernozhukov (2009) and Yang et al. (2015) is the traditional point of view in computer science that polynomial time algorithms are computationally scalable, whereas exponential time algorithms are not.

Here we consider the very popular Gibbs sampling algorithms of Polson et al. (2013) and Albert and Chib (1993), used for posterior computation for generalized linear models with probit and logit links. The algorithm of Polson et al. (2013) is the most recent in a series of ingenious algorithms for posterior computation when logit links are involved, with earlier examples including O'brien and Dunson (2004), Holmes et al. (2006), and Frühwirth-Schnatter and Frühwirth (2010). The algorithms of Albert and Chib (1993) are among the earliest examples of this class, but are still heavily used in applications, particularly for the binary probit. After data augmentation, sampling steps for the parameter of interest are conditionally Gaussian, resulting in a simple Gibbs sampler. The conditional normality of sampling steps for the exponential family parameter facilitates hierarchical modeling. Moreover, the sampler of Polson et al. (2013) is uniformly ergodic (Choi and Hobert (2013)), an atypically strong result for an MCMC algorithm. However, some practical concerns about the performance of the algorithm were raised by Ghosh et al. (2015) in the case where Cauchy priors are used for the coefficients in a logistic regression. They report good mixing with Normal priors, and attribute the poor mixing to the prior choice rather than the data augmentation scheme. A number of papers have

also suggested parameter expansion to improve mixing for the algorithm of Albert and Chib (1993). Roy and Hobert (2007) and Hobert et al. (2008) show that both the parameter expansion and original algorithm are geometrically ergodic, and provide a careful consideration of the superior practical performance of the parameter expanded algorithm in some cases.

Here we undertake a theoretical and empirical study of the performance of the algorithms of Polson et al. (2013) and Albert and Chib (1993) in the case where the posterior for parameters is concentrated near zero or one on the probability scale, corresponding to an observed number of successes y that is small relative to the sample size n. This study was initially motivated by the desire to utilize these samplers for posterior computation in hierarchical models for rare events, a problem that is inspired by a common task in quantitative advertising. By studying the regime where the success probability $p \to 0$ as $n \to \infty$, we derive bounds on the convergence rate. A similar asymptotic regime is studied by Owen (2007), which was also motivated by the modeling of rare events in a non-Bayesian context. We show that best-case computation times will converge to ∞ at least at the rate $n^{1.5}$, up to log factors. Moreover, extensive empirical studies indicate that, when y/n is near zero or one, the algorithm mixes very poorly and is not practical. Alternatives, including a Metropolis-Hastings algorithm and Hamiltonian Monte Carlo, show acceptable and excellent performance on the same data, respectively. This is demonstrated in a number of synthetic data examples as well as in an application to real quantitative advertising data.

Our results suggest that the traditional computational complexity heuristics from computer science – that polynomial time algorithms are loosely speaking "scalable" – is not useful in the context of these Gibbs samplers, and that the more recent focus on quasilinear time algorithms is justified in this context. Practically, our results also suggest that these samplers should be avoided in cases where the posterior places

significant mass on probabilities near zero or one, or, equivalently, when the Markov chain will spend a substantial amount of time in this region on the probability scale. Finally, our results provide an alternative explanation for the observation of Ghosh et al. (2015) that the Pólya-Gamma sampler mixes poorly with Cauchy priors, well with normal priors, and has intermediate performance with student t priors. When the prior has heavy tails on the logit scale, the tail of the posterior will also be heavy, as measured on the scale of typical jump sizes. Our results predict poor mixing in this setting, which is depicted graphically in Figure 7.1 in Section 7.2. Thus, the problem may not be the Cauchy prior per se, but rather that the data augmentation algorithm is exceedingly ill-suited to exploring the posterior in such settings.

7.1 Background

7.1.1 MCMC convergence rates

The fundamental entity in Bayesian statistics is the posterior distribution,

$$\Pi(\theta \mid y) = \frac{L(y \mid \theta)p(\theta)}{\int_\theta L(y \mid \theta)p(\theta)d\theta}, \tag{7.1}$$

where $L(y \mid \theta)$ is the likelihood or sampling model and $p(\theta)$ is the prior.[1] Inference centers on computing expectations with respect to the posterior $\Pi(\theta \mid y)$, as well as other summary statistics of the posterior. In general, the integral in the denominator of (7.1) is not available in closed form, so various computational methods are used to approximate quantities of interest. Arguably the most common approach is MCMC, which constructs an ergodic Markov chain with transition kernel $\mathcal{P}(\theta; \cdot)$ having invariant measure Π, then uses finite-time realizations $\{\theta_t\}_{t=1}^T$ of the chain to estimate expectations via the ergodic average $\Pi(\theta \mid y)(f) \approx \frac{1}{T}\sum_{t=1}^T f(\theta_t)$.

The convergence behavior of MCMC is often studied by initially showing that

[1] To simplify exposition, we abuse notation somewhat by using Π to refer to a distribution function, a density/mass function, or a probability measure, depending on the context

the associated kernel $\mathcal{P}$ obeys a general ergodicity condition, such as the popular *geometric ergodicity* condition:

Definition 7.1.1 (Geometric ergodicity)**.** A Markov chain evolving according to a transition kernel $\mathcal{P}(\theta;\cdot)$ with invariant measure Π is geometrically ergodic if there exists constants $\rho \in (0,1)$ and $B < \infty$ and a function $V : \Theta \to [1,\infty)$ such that

$$\left|\left|\mathcal{P}^t(\theta_0;\cdot) - \Pi\right|\right|_{\mathrm{TV}} \leq BV(\theta_0)\rho^t.$$

The constants ρ and B do not directly lead to reasonable estimates of either the mixing time or the effective sample size discussed in the introduction. Both of these terms can, however, be bounded in terms of the *spectral gap* of the associated chain. The following central limit theorem, quoted from Jones et al. (2004), gives an estimate of the effective sample size of an MCMC estimate for a chain started at stationarity:

Theorem 7.1.2 (Central Limit Theorem for MCMC)**.** *Let $\{\theta_t\}_{t\in\mathbb{N}}$ be a Markov chain evolving according to a transition kernel $\mathcal{P}(\theta;\cdot)$ that satisfies the condition in Definition 7.1.1 for some function V. Let $f : \Theta \to [1,\infty)$ satisfy $f^2 \leq V$, assume θ_1 is distributed according to the unique stationary measure Π, and define*

$$\sigma_f^2 = \mathrm{Var}[f(\theta_1)] + 2\sum_{t=2}^{\infty} \mathrm{Cov}[f(\theta_1), f(\theta_t)]. \tag{7.2}$$

Then $\sigma_f^2 \in [0,\infty)$. Furthermore, if $\sigma_f^2 > 0$, then

$$\lim_{T\to\infty} \sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} f(\theta_t) - \Pi(f)\right) \overset{d}{=} \mathrm{N}(0,\sigma_f^2).$$

This last conclusion holds for any initial distribution on θ_1.

We refer to σ_f^2 as the asymptotic variance. The asymptotic effective sample size of an MCMC sample of size T is usually defined by $\frac{T}{\sigma_f^2}$. The variance σ_f^2 is controlled by the *spectral gap*:

Definition 7.1.3 (Spectral Gap)**.** Let $\mathcal{P}(\theta;\cdot)$ be the transition kernel of a Markov chain with unique stationary distribution Π. The *spectrum* of $\mathcal{P}$ is

$$S = \{\lambda \in \mathbb{C}\backslash\{0\} \,:\, (\lambda \mathbf{I} - \mathcal{P})^{-1} \text{ is not a bounded linear operator on } L^2(\Pi)\}.$$

The *spectral gap* of $\mathcal{P}$ is given by

$$\delta(\mathcal{P}) = 1 - \sup\{|\lambda| :\, \lambda \in S,\, \lambda \neq 1\}$$

when the eigenvalue 1 has multiplicity 1, and $\delta(\mathcal{P}) = 0$ otherwise.

The asymptotic variance σ_f^2 always satisfies

$$\sigma_f^2 \leq \frac{2\mathrm{Var}_\Pi(f)}{\delta(\mathcal{P})}, \tag{7.3}$$

which goes to infinity linearly in $\delta(\mathcal{P})^{-1}$. This bound is sharp for worst case functions when $\mathcal{P}$ has no residual spectrum, which holds, for example, for reversible Markov chains on discrete state spaces. Thus, the spectral gap generally controls the asymptotic effective sample size of an MCMC algorithm via (7.3). Under stronger assumptions that are satisfied by the Markov chains studied in this paper, the spectral gap also controls the burn-in time. To state this carefully, we relate the spectral gap to the *conductance* of a Markov chain:

Definition 7.1.4 (Conductance)**.** Let $\mathcal{P}(\theta;\cdot)$ be the transition kernel of a Markov chain with invariant measure Π. For Π-measurable sets $S \subset \Theta$ with $0 < \Pi(S) < 1$, define

$$\kappa(S) = \frac{\int_{\theta \in S} \mathcal{P}(\theta, S^c)\Pi(ds)}{\Pi(S)(1 - \Pi(S))}$$

and the *Cheeger constant* or *conductance*

$$\kappa = \inf_{0<\Pi(S)<1} \kappa(S).$$

Theorem 2.1 of Lawler and Sokal (1988) relates conductance to the spectral gap:

Theorem 7.1.5. *The spectral gap* $\delta(\mathcal{P}) = 1 - \lambda_1(\mathcal{P})$ *of* $\mathcal{P}$ *satisfies*

$$\frac{\kappa^2}{8} \leq 1 - \lambda_1(\mathcal{P}) \leq \kappa.$$

The conductance (and thus the spectral gap) of a chain may be related to the required burn-in time by the following Theorem from Lovász and Simonovits (1993), which bounds the bias of an MCMC estimate after any number of steps:

Theorem 7.1.6 (Warm Start Bound)**.** *Let* $\mathcal{P}(\theta; \cdot)$ *be the transition kernel of a Markov chain* $\{\theta_t\}_{t \in \mathbb{N}}$ *with invariant measure* Π *and conductance* κ*. Then for all measurable sets* $S \subset \Theta$*,*

$$|\mathbb{P}[\theta_{t+1} \in S] - \Pi(S)| \leq \sqrt{M} \left(1 - \frac{\kappa^2}{2}\right)^t, \tag{7.4}$$

where $M = \sup_{A \subset \Theta} \frac{\mathbb{P}[\theta_0 \in A]}{\Pi(A)}$.

Define the ϵ burn-in time $t_\epsilon = \inf\{t : \sup_S |\mathbb{P}[\theta_{t+1} \in S] - \Pi(S)| < \epsilon\}$. Let κ_{sup} be an upper bound on κ. Then we obtain the following upper bound on t_ϵ:

$$t_\epsilon \leq \left\lceil \frac{\log\left(\frac{\epsilon}{\sqrt{M}}\right)}{\log\left(1 - \frac{\kappa_{\text{sup}}^2}{2}\right)} \right\rceil \asymp -\frac{\kappa_{\text{sup}}^2}{2} \log\left(\frac{\epsilon}{\sqrt{M}}\right). \tag{7.5}$$

As indicated by the $\asymp$ relation in (7.5), this quantity behaves like κ^2 for κ^2 near zero; since our results for the Pólya-Gamma and Albert and Chib samplers show the conductance converging to zero and also describe simple warm starts, this is the relevant asymptotic regime.

In light of these results, it is clear that the spectral gap gives an effective estimate of both the asymptotic effective sample size and required burn-in period for MCMC algorithms. This justifies our emphasis on estimating the spectral gaps of Markov chains throughout the rest of this paper.

7.1.2 Scenario for data generation

The scenario for data generation considered here is motivated by a problem in quantitative advertising, where a focus is performing inference on the probabilities of rare events. In this case, the rare event of interest is sequential visits to pairs of websites. In general, sequential visits to any two pairs of websites are exceedingly rare, but in advertising it is important to find groups of websites where sequential visits happen at a slightly higher rate. Suppose y_{ij} is the count of visits to website j after visiting website i, and n_i is the total number of visits to website i. Then a simple model for the sequential visit probability is

$$y_{ij} \sim \text{Binomial}(n_i, p_{ij}), \quad p_{ij} = g^{-1}(\theta_{ij}), \tag{7.6a}$$

$$\theta_{ij} \sim \text{Normal}\left(0, B\right), \tag{7.6b}$$

and g is the logit or probit link. An important motivation for the logistic or probit link with Gaussian prior is the flexibility to incorporate dependence across site pairs through more complicated hierarchical priors. An example of such a specification is given in Section 7.4.

Polson et al. (2013) introduce a data augmentation Gibbs sampler for posterior computation when g in (7.6a) is the logistic link $\ell(\cdot)$. The sampler has update rule given by

$$\omega \mid \theta \sim \text{PG}(n, \theta) \tag{7.7a}$$

$$\theta \mid \omega \sim \text{Normal}\left((\omega + B^{-1})^{-1}\kappa, (\omega + B^{-1})^{-1}\right), \tag{7.7b}$$

where $\kappa = y - n/2$ and $\text{PG}(a, c)$ is the *Pólya-Gamma* distribution with parameters a and c. The transition kernel $\mathcal{P}_\ell(\theta; \cdot)$ given by this update has θ-marginal invariant measure the posterior $\Pi(\theta \mid y)$ for the model in (7.6a)-(7.6b). Choi and Hobert (2013) show that this sampler is uniformly ergodic.

A similar data augmentation scheme exists for the case where g is the probit function $\Phi(\cdot)$, the distribution function of the standard Normal. Initially proposed

by Albert and Chib (1993), the sampler has update rule

$$Z \mid \theta = \sum_{i=1}^{y} w_i + \sum_{i=1}^{n-y} u_i, \qquad w_i \sim \mathrm{TN}(\theta, 1; 0, \infty), \quad u_i \sim \mathrm{TN}(\theta, 1; -\infty, 0) \tag{7.8}$$

$$\theta \mid Z \sim \mathrm{Normal}\left((n + B^{-1})^{-1} Z, (n + B^{-1})^{-1}\right), \tag{7.9}$$

where $\mathrm{TN}(\mu, \tau^2; a, b)$ is the normal distribution with parameters μ and τ^2 truncated to the interval (a, b). The transition kernel $\mathcal{P}_\Phi(\theta; \cdot)$ for θ defined by this update has θ-marginal invariant distribution $\Pi(\theta \mid y)$ for the model in (7.6a)-(7.6b) when $g = \Phi$. It is clear from (7.8) that the computational complexity per iteration scales linearly in n for this algorithm. Although a recent manuscript proposes some more efficient samplers, the samplers in Polson et al. (2013) for $\mathrm{PG}(n, \theta)$ also scale linearly in n. These observations will factor into our analysis of the overall run-time for these algorithms.

Motivated by inference on probabilities of rare events, in what follows we consider data $y = 1$ for increasing sequences n. Using this simple setup, we obtain a number of results on burn-in time and asymptotic effective sample size of the data augmentation samplers introduced above.

7.2 Main results

In this section we derive some large sample properties of the latent variable samplers of Albert and Chib (1993) and Polson et al. (2013). All of the results provided are for the data sequence (n, y_n) where $y_n = 1$ for every n. A similar asymptotic setup is considered in Owen (2007), where the motivation is also rare events. We provide two types of results: (1) upper bounds on the spectral gap; and (2) results showing that a certain starting distribution is 'warm.' These results are provided in Theorems 7.2.1 and 7.2.2. By Theorems 7.1.2 and 7.1.6, these results give bounds on the asymptotic variance and mixing time for an MCMC algorithm in terms of the problem size n.

Our main conclusions are that the spectral gap converges to zero at least at the rate $\mathcal{O}\left(\frac{\log(n)^c}{\sqrt{n}}\right)$ for some $0 \leq c < \infty$, and that it is straightforward to begin the chain from a warm start. Because the computational complexity per MCMC iteration for these algorithms is $\mathcal{O}(n)$, these results further imply running times of at least $\Omega(n^{1.8})$ ($\Omega(n)$ per iteration and $\Omega(n^{0.8})$ from the scaling of the effective size), where $x = \Omega(f(n))$ indicates $x \geq Cf(n)$ for some $C < \infty$.

7.2.1 Intuition

The intuition for the results that follow can be explained simply using the graphic in Figure 7.1. In either the logit or probit model when $y/n \to 0$, the width of the high probability region of the posterior is $\Omega\left(\frac{1}{\log n}\right)$. This is depicted by the large region in the graphic where the posterior density is non-trivial. However, the steps in the data augmentation samplers depend on the mean of n iid random variables, so the typical move size is $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. A graphic depiction of such a move is provided in Figure 7.1 in the region indicated by $\theta_t \to \theta_{t+1}$. Thus, the algorithm needs at least $\Omega\left(\frac{\log n}{\sqrt{n}}\right)$ moves to traverse the high-probability region of the posterior. This also explains the convergence of the asymptotic effective sample size to zero. These rough computations omit some logarithmic factors that appear in the rigorous bounds, but this is irrelevant to the basic intuition. Note that this reasoning applies to *any data augmentation sampler* with step sizes that are proportional the the variance of the mean of *iid* random variables. If the high probability region of the posterior is contracting at a rate slower than $\frac{1}{\sqrt{n}}$, then the algorithm will mix poorly in large samples.

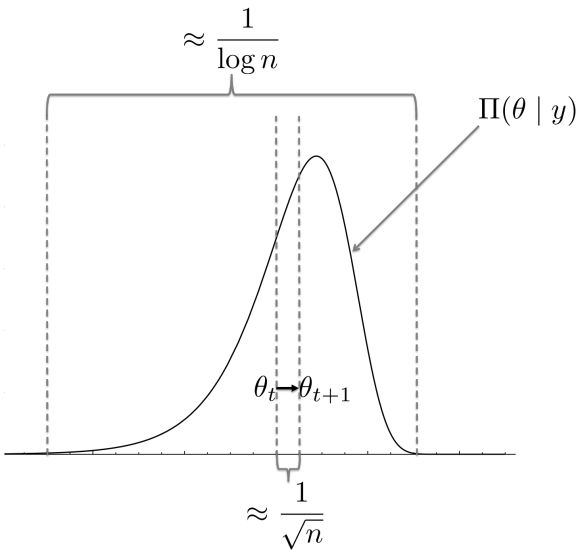


FIGURE 7.1: *Cartoon comparing the posterior mode width and typical move size.*

7.2.2 Convergence rate and spectral gap

We now give results on the Cheeger constant and 'warm start' bounds for the latent variable samplers of Albert and Chib (1993) and Polson et al. (2013). Applying Theorems 7.1.5, 7.1.2 and 7.1.6 will then give bounds on the convergence rates of these samplers. Theorem 7.2.1 gives the result for the Pólya-Gamma sampler. Proofs are found in Appendix F.

Theorem 7.2.1. *Let $\{\theta_t, w_t\}_{t\geq 0}$ be a Polya-Gamma sampler evolving according to (7.7a)-(7.7b) in the case that $y = 1$, and let $\mathcal{P}$ be the associated kernel. Then $\mathcal{P}$ has spectral gap*

$$1 - \lambda_1(\mathcal{P}) = \mathcal{O}\left(\frac{\log(n)^{4.5}}{\sqrt{n}}\right). \tag{7.10}$$

Furthermore, let $\theta_{\max} \equiv \text{argmax}_\theta p(\theta|y = 1)$ be the mode of Π; then the distribution $\mu_n = \text{Unif}([\theta_{\max} - \frac{1}{\log(n)}, \theta_{\max} + \frac{1}{\log(n)}])$ satisfies

$$\sup_{A\subset\mathbb{R}} \frac{\mu_n(A)}{\Pi(A|y)} = O(\log(n)^2) \tag{7.11}$$

and thus provides a 'warm start' distribution for the Polya-Gamma sampler. Note that $\theta_{\max}$ *is the unique solution to* $\frac{\theta_{\max}}{B} + n\frac{e^{\theta_{\max}}}{1+e^{\theta_{\max}}} = 1$.

The result for the Albert and Chib sampler is given in Theorem 7.2.2.

Theorem 7.2.2. *Let* $\{\theta_t, z_t\}_{t\geq 0}$ *be the Albert and Chib sampler evolving according to (7.8)-(7.9) in the case that* $y = 1$*, and let* $\mathcal{P}$ *be the associated kernel. Then* $\mathcal{P}$ *has spectral gap*

$$1 - \lambda_1(\mathcal{P}) = \mathcal{O}\left(\frac{(\log n)^2}{\sqrt{n}}\right). \tag{7.12}$$

Furthermore, letting Φ *be the CDF of the standard normal distribution, the distribution* $\mu_n = \text{Unif}\left[\Phi^{-1}\left(\frac{B+1}{C(Bn+1)}\right), \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right]$ *satisfies*

$$\sup_{A\subset\mathbb{R}} \frac{\mu_n(A)}{\Pi(A|y)} = \mathcal{O}(1) \tag{7.13}$$

and thus provides a 'warm start' distribution for the Polya-Gamma sampler.

Remark 7.2.1 (Asymptotic effective sample sizes)**.** Theorems 7.2.1 and 7.2.2 also allow us to roughly estimate the scaling of the asymptotic effective sample size.

If we assume that the upper bounds on conductance are achieved, then applying Theorem 7.1.5, the spectral gap scales at rate n^{-k} for $1/2 < k < 1$. If we further assume that the bound in (7.3) is achieved, then the asymptotic effective sample size also scales at rate n^{-k} for $1/2 < k < 1$. Therefore, if the bounds are tight, then the asymptotic effective sample size goes to zero at rate between $n^{-1/2}$ and n^{-1}.

Thus, since the asymptotic effective sample size scales at least like $n^{-1/2}$, the overall run time for both algorithms is $\Omega(n^{1.5})$ or worse (neglecting logarithmic factors). In the following section, we provide empirical evidence that the run time empirically scales like $n^{1.85}$, incorporating both the computation time per iteration and the effects on mixing.

7.3 Synthetic Data Examples

In this section we perform a number of example computations to assess the practical performance of the samplers in (7.8)-(7.9) and (7.7a)-(7.7b). We also consider related samplers for Multinomial logit or probit in the situation where one or more of the observed cell counts is 1.

7.3.1 Binomial Logit and Probit

In the first set of computational examples, we consider the model in (7.6a)-(7.6b). We put $y = 1$ and vary n between 10 and $10,000$. We perform computation using either the algorithm of Polson et al. (2013) or Albert and Chib (1993), then estimate autocorrelations and effective sample sizes. For the probit, we use a prior of $B = 49$, whereas for the logit we use a prior of $B = 100$, reflecting the lighter tails of the probit function. For the probit, we initialize the sampler at the MLE and collect no burn-in; for logit, we initialize the sampler at zero and collect burn-in of 5000 iterations. Similar results were achieved using the opposite choice for each algorithm. For probit, we wrote code for the Gibbs sampler in `R`. For logit, we use the package `BayesLogit` for computation. For comparison, we also estimate the model using Hamiltonian Monte Carlo (HMC), implemented with the `Stan` environment and `Rstan` package.

Figure 7.2 shows the autocorrelation function (estimated using the `coda` package for `R`) for the probit and logit at lags 1-100, with computation by the corresponding data augmentation Gibbs sampler, as well as for the logit model estimated using HMC. Clearly, the autocorrelations increase with n for both the data augmentation Gibbs samplers, but are very near zero for HMC after lag ten, and are essentially identical for all values of n. Table 7.1 shows T_{eff}/T (also computed using `coda`) for the same three algorithms with $y = 1$ and increasing values of n. The effective sample size is anemic for the data augmentation Gibbs samplers but about $0.2T$ for

HMC for all values of n.

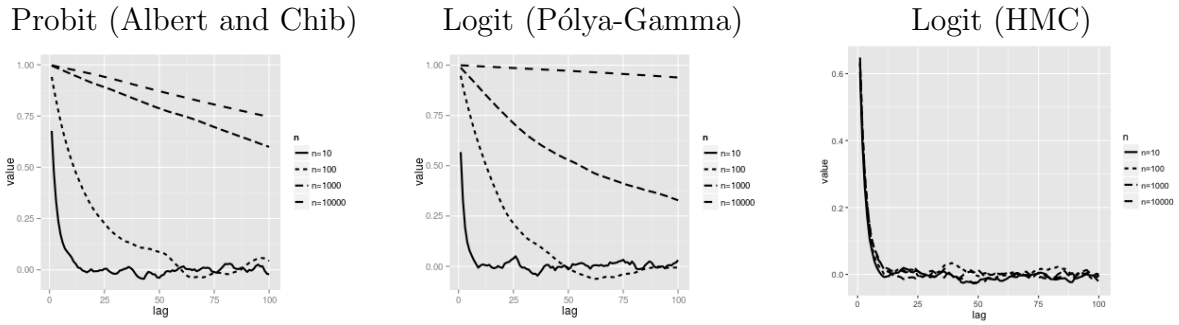


FIGURE 7.2: *Estimated autocorrelations at lags 1-100 for the two data augmentation samplers for binomial probit and logit as well as for logit with computation by HMC. Four different values of n are shown. Note that the HMC plot is on a different scale for readibility; the maximum absolute correlation at any lag for that algorithm is less than 0.03.*

Table 7.1: Estimated value of T_{eff}/T for $T = 5000$ for probit and logit models using sampler of Albert and Chib (1993) (AC) and Polson et al. (2013) (PG), respectively, and for the logit model with computation by HMC. Here $y = 1$ in each case and n varies between 10 *and* $10,000$.

	Probit (AC)	Logit (PG)	Logit (HMC)
n=10	0.1968	0.2755	0.2234
n=50	0.0449	0.0479	0.1867
n=100	0.0322	0.0260	0.2096
n=500	0.0128	0.0073	0.1944
n=1000	0.0064	0.0061	0.1919
n=5000	0.0016	0.0020	0.2063
n=10000	0.0008	0.0003	0.1944

Although the theoretical results in Section 7.2 consider the case where $y = 1$ and n is increasing, empirically we observe poor mixing whenever y/n is small. To demonstrate this, we perform another set of computational examples where y and n both vary in such a way that y/n is constant. Specifically, we consider $n = 10,000$, $n = 50,000$, and $n = 100,000$ with $y = 1, 5, 10$. Computation is performed for the two data augmentation Gibbs samplers as above, and effective sample sizes and autocorrelation functions estimated. The results in Figure 7.3 shows estimated autocorrelations, which are similarly near 1 at lag 1 and decay slowly. Table 7.2 shows values of T_{eff}/T for $T = 5000$ for the two algorithms. Neither measure of

computational efficiency shows a meaningful effect of increasing y when y/n remains constant.

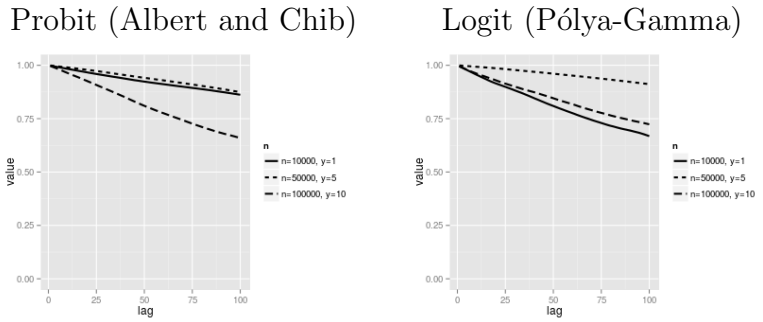


FIGURE 7.3: *Estimated autocorrelation functions for synthetic data examples that vary y and n*

Table 7.2: Values of T_{eff}/T for synthetic data examples that vary y and n.

	AC	PG
n=10000, y=1	0.0008	0.0022
n=50000, y=5	0.0005	0.0003
n=100000, y=10	0.0020	0.0022

7.3.2 Empirical analysis of mixing times

In section 7.2, we provided bounds on the burn-in times for the Albert and Chib and Pólya-Gamma data augmentation Gibbs samplers that indicate these samplers have run times of at least $\Omega(n^{1.5})$, neglecting logarithmic factors. However, this is a lower bound and we do not assess theoretically whether the bound is tight. Here, we provide an empirical evaluation of the mixing times of these algorithms that suggests the reciprocal asymptotic effective sample size scales approximately as $n^{0.85}$, confirming the rough calculation in Remark 7.2.1 is fairly accurate.

To estimate σ_f^2, we use the relationship in 7.2, which, upon rearranging gives

$$\frac{\sigma_f^2}{\mathrm{Var}_\Pi(\sigma)} = \eta_1 + 2\sum_{t=2}^{\infty} \eta_t, \tag{7.14}$$

where η_t is the lag-t autocorrelation of $f(\theta_k)$. We are interested in the rate at which this value grows with n. Let $\sigma_f^2(n)$ be the asymptotic variance when the sample size is n, and assume $\sigma_f^2(n) \approx Cn^k$ for large n, so that $\log(\sigma_f^2(n)) = \log(C) + k\log(n)$. This suggests an approach to estimating k by estimating $\sigma_f^2(n)$ for different values of n, then estimating k by regression of $\log(\sigma_f^2(n))$ on $\log(n)$.

Clearly, we can estimate σ_f^2 based on autocorrelations via (7.14). First, notice that using an estimate of $\rho_f = \frac{\sigma_f^2}{\mathrm{Var}_\Pi(\sigma)}$ instead of an estimate of σ_f^2 will affect only the constant C but not k, so it is enough to compute the quantity on the right in (7.14). A point estimate that can be computed for finite length sample paths is

$$\widehat{\rho_f} = \widehat{\eta_1} + 2\sum_{t=2}^{S} \widehat{\eta_t}, \tag{7.15}$$

where $\widehat{\eta_t}$ is a point estimate of η_t. The choice of S is important. First, S should be small enough relative to the length of the path T that uncertainty in the point estimates $\widehat{\eta_t}$ is small. Second, most of the contribution to the sum in (7.14) occurs by $t \approx \frac{C}{1-\lambda}$, where C is a "large" constant and $1-\lambda$ is the spectral gap of the chain. Thus, we want $S \gg \frac{1}{1-\lambda}$. The lower bounds derived in Section 7.2 have $\frac{1}{1-\lambda} = \Omega n^{1/2}$ up to a log factor, so we use $S = n$ to compute the sum in (7.15).

Because these Markov chains are poorly mixing for larger values of n, estimating η_t can be challenging. To obtain robust estimates of η_t, we obtain ten sample paths of length $T = 1.1 \times 10^6$. For each sample path, the maximum likelihood estimates (MLEs) of η_t are computed, and the final point estimate $\widehat{\eta_t}$ used in (7.15) is calculated by taking the median of the MLEs across sample paths. Figure 7.4 shows plots of $\log(n)$ versus $\log(\widehat{\rho_f})$ for values of N between 10 and 10,000 for the Pólya-Gamma and Albert and Chib sampler. The relationships are apparently linear, and the least squares estimate of the slope is 0.86 for Pólya-Gamma and 0.84 for Albert and Chib, so the associated point estimate of asymptotic variance scales approximately as $n^{0.85}$.

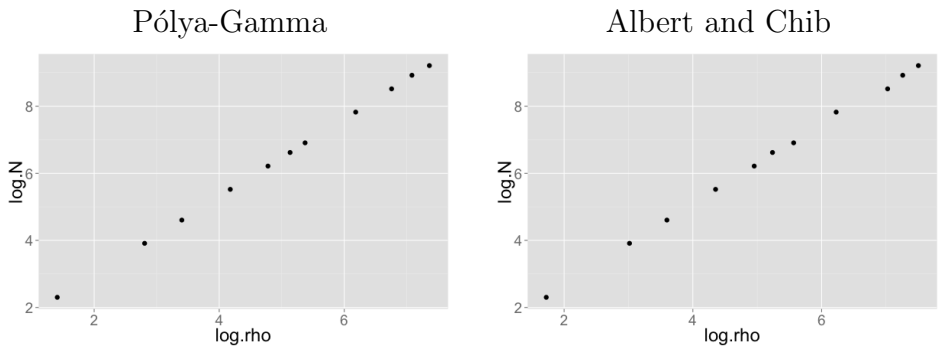


FIGURE 7.4: *Plots of* $\log(n)$ *versus* $\sigma^2(n)$ *for different values of* n*. The estimated values of* k *are* 0.86 *and* 0.84*, respectively, for* $\sigma^2(n) = \mathcal{O}\left(n^k\right)$.

7.3.3 Data augmentation algorithms for Multinomial likelihoods

Thus far we have considered specifically data augmentation algorithms for probit and logit links and binomial likelihoods. Similar algorithms exist for computation in Multinomial logit and probit models, the analogues of these links for multinomial likelihoods. Specifically, consider models of the form

$$\begin{aligned} y &\sim \text{Multinomial}(n, \pi), \quad \pi = g^{-1}(\theta), \\ \theta_j &\sim \text{Normal}\left(0, B_j\right), \end{aligned} \tag{7.16}$$

where here y is a length d vector of nonnegative integers whose sum is n, π resides on the boundary of the d-dimensional simplex $\bar{\mathcal{S}}^d$, $B_j > 0$, and $g^{-1}(\cdot)$ is a map $\mathbb{R}^{d-1} \to \bar{\mathcal{S}}^d$. In general, one category is chosen as the "base," without loss of generality, we assume this is the category labeled 1. Then the Multinomial probit model results by taking $\theta^* = (0, \theta)$, and setting $\pi_j = \mathbb{P}\left[z_j = \max_{j'} z_{j'}\right]$ with $z \sim N(\theta^*, I)$ [2]. The multinomial logit results from putting $g_j(\theta) = \frac{e^{\theta_j}}{1+\sum_{j'} e^{\theta_{j'}}}$ for $j > 1$.

[2] In economics, it is common to make the covariance a parameter to estimate; however, in this simple setting the covariance parameter is not identified

In particular, Polson et al. (2013) describes a Pólya-Gamma data augmentation scheme for the multinomial logit as in (7.16) with update rule

$$\theta_j \mid \Omega_j \sim \text{Normal}\,(m_j, V_j)$$
$$\omega_j \mid \theta_j \sim \text{PG}(n, \eta_j),$$

where

$$V_j^{-1} = (\omega_j + B_j^{-1})^{-1}$$
$$m_j = V_j(y_j - \omega_j c_j),$$

and $c_j = \log \sum_{k \neq j} e^{\theta_k}$. This algorithm is very similar to the data augmentation algorithm for binary logit. A data augmentation Gibbs sampler for the Multinomial probit is given in Imai and van Dyk (2005) and implemented in the `R` package `MNP`. The algorithm and priors are somewhat complicated due to restrictions necessary for identification, so we do not summarize it here. We utilize this package for computation.

We study a synthetic data example where y is a 4×1 count vector with entries adding to n. The first three entries of y are always 1, the final entry is $n - 3$, and a series of values of n between $n = 10$ and $n = 10,000$ are studied. Estimated values of T_{eff}/T for the three entries of θ for both algorithms are shown in Table 7.3. The results are very similar to those for the binomial logit and probit, and are consistent across the different entries of θ. This example is of applied relevance. Although situations in which small numbers of successes are encountered with a large number of trials may be restricted to certain application areas, it is exceedingly common in modeling vectors of counts with large numbers of categoies that many of the entries are small or zero. This suggests that data augmentation for inference with Multinomial likelihoods is of little use when the number of categories is moderate to large.

Table 7.3: Estimated values of T_{eff}/T for the three entries of θ for multinomial logit and probit data augmentation for increasing values of n with data $y = (1, 1, 1, n-3)$. Results are based on 5,000 samples gathered after discarding 5,000 samples as burn-in.

	Multinomial Probit			Multinomial Logit		
	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
$n = 10$	0.1069	0.0926	0.1512	0.2430	0.2221	0.2763
$n = 50$	0.0219	0.0269	0.0396	0.0552	0.0628	0.0643
$n = 100$	0.0122	0.0154	0.0201	0.0306	0.0393	0.0365
$n = 500$	0.0032	0.0046	0.0100	0.0061	0.0083	0.0097
$n = 1000$	0.0024	0.0023	0.0097	0.0068	0.0062	0.0105
$n = 5000$	0.0004	0.0006	0.0018	0.0016	0.0025	0.0018
$n = 10,000$	0.0010	0.0009	0.0052	0.0019	0.0035	0.0019

7.4 Real data example: quantitative advertising

In quantitative advertising, it is important to accurately estimate the probability that users view two websites within a specified time window. In particular, advertisers are interested in the "organic" probability that a user views a client's website during the same browsing session that the user also visits one of thousands of high traffic sites. Here, "organic" means the user views the client's site without clicking on a link in an advertisement. Small differences in these organic transition probabilities often translate to commercially very significant differences in the effectiveness of ads. Ultimately, the goal is to develop a list of potential high-traffic sites to serve ads, ranked in order of the rate at which users organically view the client site.

These transitions are rare events, and in most cases it is necessary to obtain data on tens or hundreds of thousands of users to view even a single organic transition, so the data are noisy, with only a few transitions observed for most of the high traffic sites. Thus, it makes sense to borrow information across the different high-traffic sites to obtain lower risk point estimates of the transition probabilities. Moreover, estimates of uncertainty are useful, since a site with a relatively high transition probability that is estimated very precisely may be a better target than one that has an even higher estimated transition probability with relatively large confidence/credible

bands.

A simple approach is to borrowing information and inducing shrinkage is to take a Bayesian approach to inference and use a hierarchical prior structure. For example

$$y_i \sim \text{Binomial}(n_i, p_i), \quad p_i = \frac{e^{\theta_i}}{1 + e^{\theta_i}}, i = 1, \ldots, N \tag{7.17}$$

$$\theta_i \overset{iid}{\sim} N(\theta_0, \tau^2), \quad p(\tau^2, \theta_0) \sim f(\tau^2, \theta_0) \tag{7.18}$$

Here, i indexes the high-traffic sites, and p_i is the probability of visiting the client site in the same browsing session. Ultimately, it might make sense to add more layers of hierarchy, perhaps grouping the high-traffic sites according to content or topic. However, our interest is in the mixing properties of alternative MCMC algorithms for computation in these models, and if the simple model above fails to mix well, more complex prior structures are unlikely to alleviate the problem.

We use data on transitions from 59,317 high traffic sites to a client site to do computation for the model in (7.17)-(7.18) by MCMC with Pólya-Gamma data augmentation. We use an improper uniform prior on τ and a Normal (θ_{00}, τ_0^2) prior on θ_0, with $\theta_{00} = -12$ and $\tau_0^2 = 49$. The prior on θ_0 is weakly informative on the logit scale and consistent with information solicited from experts in quantitative advertising. As an alternative, we use the following Metropolis-within-Gibbs algorithm:

1. Compute the conditional mode of $\hat{\theta}_i^t = p(\theta_i \mid \theta^t, \boldsymbol{y}, \boldsymbol{n})$ for each $i \in 1, \ldots, N$ using Newton-Raphson. This consists of N independent univariate convex optimization problems for which both the first and second derivatives with respect to θ_i are available. This step is thus quite fast.

2. Propose θ_i^* from a $t_5(\hat{\theta}_i^t, v)$, where $t_\nu(m, v)$ is a t distribution with ν degrees of freedom and scale v centered at m. Tune v to give 30-40 percent acceptance rates.

3. Compute the acceptance probability

$$\log(q^*) = Y \log\left(\frac{p_i^*}{p_i}\right) + (N - Y)\log\left(\frac{1-p_i^*}{1-p_i}\right) - \frac{1}{2}\frac{(\theta_0 - \theta_i^*)^2}{\tau^2} + \frac{1}{2}\frac{(\theta_0 - \theta_i)^2}{\tau^2}$$

where $p_i = e^{\theta_i}/(1+e^{\theta_i})$ and perform a Metropolis step. Note that the proposal did not depend on θ_i so this is not a Metropolis-Hastings step.

4. Sample θ_0 and τ^2 from

$$\theta_0 \mid \theta_i, \tau^2 \sim \text{Normal}\left(sm, s\right), \quad s = (N/\tau^2 + 1/\tau_0^2)^{-1}, m = (\sum_i \theta_i/\tau^2 + \theta_{00}/\tau_0^2)$$

$$\tau^{-2} \mid \theta_i \sim \text{Gamma}\left(\frac{(N-1)}{2}, \frac{\sum_i(\theta_0 - \theta_i)^2}{2}\right).$$

We also performed computation on a subset of 593 of the sites using HMC implemented in the `Stan` language with the `rstan` package (HMC was prohibitively slow on the full data).

Figure 7.5 shows the autocorrelations and estimated values of T_{eff}/T for the Pólya-Gamma data augmentation algorithm, HMC, and the MH algorithm described above. The lag-1 autocorrelation for the data augmentation Gibbs sampler is very near 1, and the autocorrelation function decays vary slowly. In contrast, the HMC algorithm has autocorrelations near zero by lag 5, while the MH algorithm has an autocorrelation function that decays to below 0.25 for the majority of the parameters by lag 20. The values of T_{eff}/T are consistent with what is expected given the autocorrelation structure for each Markov chain. Ultimately, these results demonstrate a practical applied problem in which the data augmentation Gibbs sampler is not useful, while readily available alternatives provide excellent (in the case of HMC) and acceptable (in the case of MH) performance.

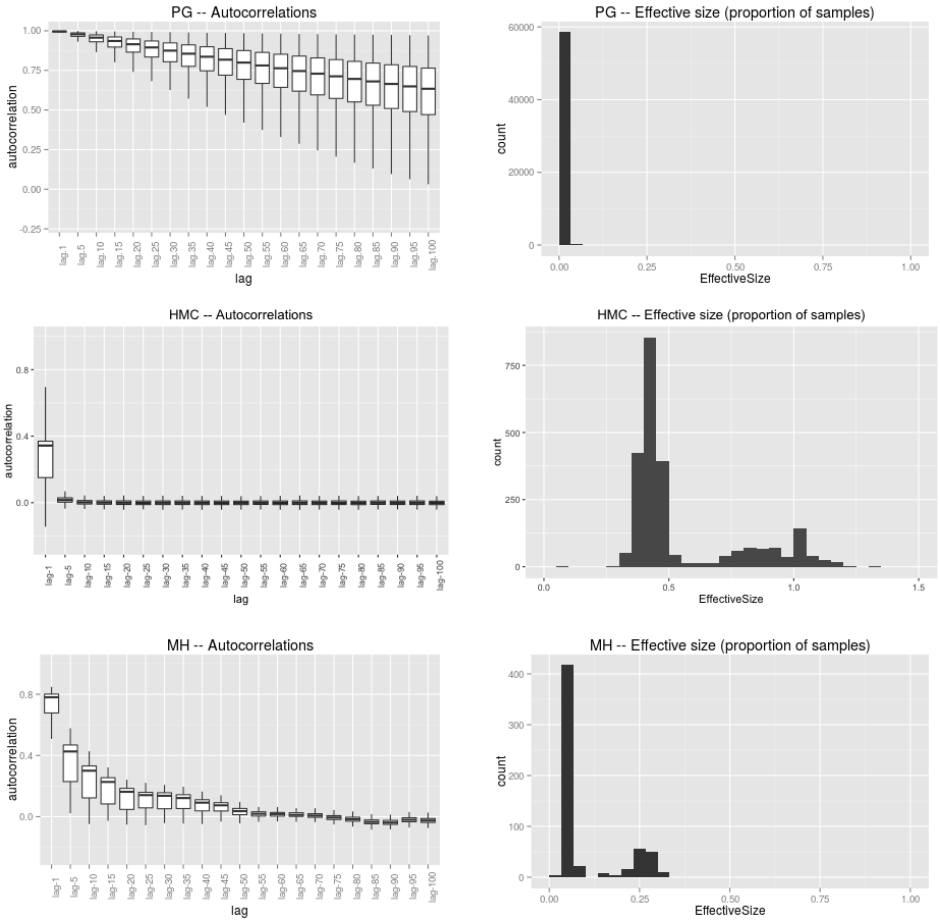


FIGURE 7.5*: Autocorrelations (left column) and Effective sample sizes (right column) for Polya-Gamma (PG), Albert and Chib (AC), and Hamiltonian Monte Carlo (HMC). The PG and HMC examples use the logit link. The boxplot of autocorrelations shows variation in the autocorrelation function across the 59,317 θ_i parameters (593 in the case of HMC), whereas the histograms of effective sample sizes also depict variation across the site-specific parameters.*

7.5 Discussion

Providing easy to implement and reliable algorithms for posterior computation in generalized linear models has been a major focus for decades. Data augmentation Gibbs sampling, particularly for logistic links, has received much of this attention. A series of clever data augmentation schemes, of which Polson et al. (2013) is the most recent, have steadily improved the accessibility of Gibbs sampling for this important class of generalized linear models. This is a specific case of the larger trend in

Bayesian computation over the past several decades of focusing attention on Gibbs samplers, both in algorithm development and theoretical analysis of existing algorithms. This focus reflects the general extensibility of Gibbs samplers, the minimal need for tuning, and the widespread familiarity with the algorithm among applied statisticians. These factors contribute to a short development time for modifications of existing Gibbs samplers to accommodate a wide variety of datasets and prior structures.

That data augmentation Gibbs samplers mix so poorly in the relatively simple cases analyzed here perhaps indicates a re-thinking of priorities in the development of algorithms for Bayesian computation may be in order. Although development time for more exotic algorithms may be considerable, in modern high-dimensional problems one can ill afford the computation time penalties associated with poor mixing in large samples. As we point out, any data augmentation Gibbs sampler in which the step sizes converge to zero faster than the posterior contracts around its mode will inevitably mix poorly in large samples. Therefore, problems such as those elucidated here are probably inevitable in MCMC algorithms that make local moves, unless great care is taken to keep the typical move sizes on the same scale as the width of the high-probability region of the posterior. This also points to the importance of analyzing jointly the properties of the posterior and the properties of MCMC algorithms for approximately sampling from it. Overall, it may be worthwhile to focus more on alternatives to Gibbs sampling, such as Hamiltonian Monte Carlo, if Bayesian computation based on Markov chains is to maintain its dominance in the era of Big Data.

Appendix A

Appendix to Chapter 2

A.1 Proofs and auxiliary results

A.1.1 Auxiliary results

We state and prove lemma A.1.1 which is used to prove Theorem 2.3.1.

Lemma A.1.1. *Let π and ψ be two non-negative d^p tensors. Then, $rnk_P^+(\pi \circ \psi) \leq rnk_P^+(\pi) rnk_P^+(\psi)$, where $\circ$ denotes a Hadamard product, and $rnk_P^+(\pi + \psi) \leq rnk_P^+(\pi) + rnk_P^+(\psi)$.*

Proof. Let $\text{rnk}_P^+(\pi) = m, \text{rnk}_P^+(\psi) = k$ and $\phi = \pi \circ \psi$. For $1 \leq j \leq p$, there exist non-negative vectors $\lambda_h^{(j)} \in \mathbb{R}_+^d, h = 1, \ldots, m$ and $\zeta_l^{(j)} \in \mathbb{R}_+^d, l = 1, \ldots, k$, such that $\pi = \sum_{h=1}^m \lambda_h^{(1)} \otimes \ldots \otimes \lambda_h^{(p)}$ and $\psi = \sum_{l=1}^k \zeta_h^{(1)} \otimes \ldots \otimes \zeta_h^{(p)}$. Then, it is easy to see that

$$\phi = \sum_{h=1}^{m} \sum_{l=1}^{k} \gamma_{hl}^{(1)} \otimes \ldots \otimes \gamma_{hl}^{(p)},$$

where $\gamma_{hl}^{(j)} = \lambda_h^{(j)} \circ \zeta_l^{(j)}$ for $1 \leq j \leq p$. Clearly, for any j, $\gamma_{hl}^{(j)} \in \mathbb{R}_+^d$ for $h = 1, \ldots, m; l = 1, \ldots, k$. Thus, $\text{rnk}_P^+(\phi) \leq mk$.

In particular, if $\text{rnk}_P^+(\psi) = 1$, we have $\text{rnk}_P^+(\phi) \leq m$. This bound cannot be globally improved, or in other words, the upper bound can be achieved. Take for example, $\psi = \zeta^{(1)} \otimes \ldots \otimes \zeta^{(p)}$, with $\zeta^{(j)} = (1, \ldots, 1)^{\mathrm{T}}$ for all j.

Finally, we note that if

$$\pi = \sum_{h=1}^{m_1} \bigotimes_{j=1}^{p} \tilde{\lambda}_h^{(j)} \text{ and } \psi = \sum_{h=1}^{m_2} \bigotimes_{j=1}^{p} \tilde{\zeta}_h^{(j)}$$

then

$$\pi + \psi = \sum_{h=1}^{m_1} \bigotimes_{j=1}^{p} \tilde{\lambda}_h^{(j)} + \sum_{h=1}^{m_2} \bigotimes_{j=1}^{p} \tilde{\zeta}_h^{(j)}$$

so $\text{rnk}_P^+(\pi + \psi) = m_1 + m_2 = \text{rnk}_P^+(\pi) + \text{rnk}_P^+(\psi)$. □

Proof of Theorem 2.3.1

Without loss of generality, we assume σ is the identity permutation and drop the corresponding subscripts. Let $\mathcal{P}^{(1)}$ be the partition of $\mathcal{I}_1$ consisting of the singleton sets $\{c\}$ for $c \in B_1$ and the set $(B_1)^c$. Weak hierachicality ensures that $y_1 \mathbb{1}_{(y_1 \in A)} \perp\!\!\!\perp y_{[-1]}$ for any $A \in \mathcal{P}^{(1)}$. Using the fact that for any two random variables Z_1, Z_2 and any measurable set A, $Z_1 \mathbb{1}_{(Z_1 \in A)} \perp\!\!\!\perp Z_2 \Leftrightarrow Z_1 \perp\!\!\!\perp Z_2 \mid A$, we have $y_1 \perp\!\!\!\perp y_{[-1]} \mid A$ for any $A \in \mathcal{P}^{(1)}$. Enumerating the sets in $\mathcal{P}^{(1)}$ as $A_1, \ldots, A_{m_1}$, with $m_1 = |\mathcal{P}^{(1)}| = |B_1| + 1$, we can write π as

$$\pi_{c_1 \ldots c_p} = \sum_{h=1}^{m_1} \nu_h \lambda_{hc_1} \psi_{hc_2 \ldots c_p}, \tag{A.1}$$

where for each $1 \leq h \leq m_1$, $\nu_h = Pr(A_h)$, $\lambda_h \in \Delta^{(d-1)}$ with $\lambda_{hc} = Pr(y_1 = c \mid A_h)$ and ψ_h is a d^{p-1} non-negative tensor representing the joint probability of $y_{[-1]} | A_h$, i.e.,

$$\psi_{hc_2 \ldots c_p} = Pr(y_2 = c_2, \ldots, y_p = c_p \mid A_h).$$

Define d^p tensors $\{\pi_h^{(1)}\}$ and $\{\pi_h^{(2)}\}$ by

$$\pi_h^{(1)} = \lambda_h \otimes \mathbf{1} \ldots \otimes \mathbf{1}$$

$$(\pi_h^{(2)})_{c_1 \ldots c_p} = \nu_h \psi_{h c_2 \ldots c_p}.$$

The expansion of π in (A.1) can now be written in tensor notation as $\pi = \sum_{h=1}^{m_1} \pi_h^{(1)} \circ \pi_h^{(2)}$. Clearly $\text{rnk}_P^+(\pi_h^{(1)}) = 1$ and it is easily verified that $\text{rnk}_P^+(\pi_h^{(2)}) \leq \text{rnk}_P^+(\psi_h)$ for all h. Therefore, using Lemma A.1.1 we have that $\text{rnk}_P^+(\pi) \leq m_1 r$, where $r = \text{rnk}_P^+(\psi_h)$.

Recursively applying this process for the variables $y_2, \ldots, y_p$, we can show that $r \leq \prod_{j=2}^p m_j = \prod_{j=2}^p (|B_j|+1)$, so that

$$\text{rnk}_P^+(\pi) \leq \prod_{j=1}^p (|B_j|+1).$$

For any permutation σ, we can obtain a result as in the above display by scanning through the variables in the sequence $\sigma(1), \ldots, \sigma(p)$. Taking the minimum over all permutations σ, we obtain the desired result.

Proof of (2.14) in Theorem 2.3.2

Fix $H \in \mathscr{H}$. Let $\bar{H}_j = \mathcal{I}_j \backslash H_j$ and let $\mathcal{P}_{H,j}$ denote the partition of $\mathcal{I}_j$ consisting of the singleton sets $\{i_j\}$ for $i_j \in H_j$ and the set $\bar{H}_j$. Define a partition $\mathcal{P}_H^0$ of $\mathcal{I}_V$ as the Cartesian product of the partitions $\mathcal{P}_{H,j}$ as in (2.11). We show that for any set $A \in \mathcal{P}_H^0$, (2.12) is satisfied, i.e.,

$$Pr(y_1 = i_1, \ldots, y_p = i_p \mid A) = \prod_{j=1}^p Pr(y_j = i_j \mid A), \tag{A.2}$$

for any $\boldsymbol{i} \in \mathcal{I}_V$. Based on the discussion in Section 3.1, the random variable $z = z_H^0$ corresponding to the partition $\mathcal{P}_H^0$ defined via (2.10) will then satisfy (2.9), implying

$$\text{rnk}_P^+(\pi) \leq |\mathcal{P}_H^0| = \prod_{j=1}^p |\mathcal{P}_{H,j}| = \prod_{j=1}^p (|H_j|+1).$$

We now proceed to establish (A.2). Fix $A \in \mathcal{P}_H^0$. By construction,

$$A = \bigtimes_{k \in \bar{J}} \{c_k\} \times \bigtimes_{j \in J} \bar{H}_j \tag{A.3}$$

for some $J \subset V$, $\bar{J} = V \backslash J$ and $c_k \in H_k$ for all $k \in \bar{J}$. Without loss of generality, we assume $J = \{q, \ldots, p\}$ for some integer $q \geq 1$.

Let $\tilde{\mathcal{I}}_V$ denote the subset of $\mathcal{I}_V$ consisting of cells $\boldsymbol{i}$ such that $i_k = c_k$ for all $k \in \bar{J}$ and $i_j \in \bar{H}_j$ for all $j \in J$. It is easy to see that for any $\boldsymbol{i} \notin \tilde{\mathcal{I}}_V$, (A.2) is satisfied trivially since both sides are reduced to zero or one simultaneously. Hence, it suffices to show that (A.2) holds for any $\boldsymbol{i} \in \tilde{\mathcal{I}}_V$.

Fix $\boldsymbol{i} \in \tilde{\mathcal{I}}_V$. Let $A_{\boldsymbol{i}}$ denote the subset of $\mathcal{I}_V$ corresponding to the event $\{y_j = i_j, j \in V\}$ in $\mathcal{Y}$, so that

$$A_{\boldsymbol{i}} = \bigtimes_{j \in V} \{i_j\}, \quad Pr(A_{\boldsymbol{i}}) = \pi_{\boldsymbol{i}}.$$

Clearly, $A_{\boldsymbol{i}} \subset A$, which implies $Pr(A_{\boldsymbol{i}} \mid A) = \pi_{\boldsymbol{i}}/Pr(A)$. Further, $Pr(y_k = i_k \mid A) = 1$ for any $k \in \bar{J}$, since $i_k = c_k$ for $k \in \bar{J}$. Therefore, (A.2) reduces to showing

$$\frac{\pi_{\boldsymbol{i}}}{Pr(A)} = \prod_{l \in J} Pr(y_l = i_l \mid A). \tag{A.4}$$

For $E \subset V$, we introduce the notation

$$\bar{\boldsymbol{H}}_E = \prod_{j \in E} \bar{H}_j.$$

We shall use $\boldsymbol{\alpha}$ to generically denote an element of $\bar{\boldsymbol{H}}_J$, i.e., $\boldsymbol{\alpha}$ is a $|J|$-vector of indices with α_j the entry in $\boldsymbol{\alpha}$ corresponding to variable $j \in J$. For $l \in J$, $J^{(-l)}$ shall denote the set $J \backslash \{l\}$. We use $\boldsymbol{\alpha}^{(l)}$ to generically denote an element of $\bar{\boldsymbol{H}}_{J^{(-l)}}$, with $\alpha_j^{(l)}$ the entry in $\boldsymbol{\alpha}^{(l)}$ corresponding to variable $j \in J^{(-l)}$.

Finally, for a partition of V into J_1, J_2, J_3, denote [1]

$$\pi^{(J_1,J_2,J_3)}_{f_j g_k h_l} := Pr\Big[\bigtimes_{j\in J_1} \{f_j\} \times \bigtimes_{k\in J_2} \{g_k\} \times \bigtimes_{l\in J_3} \{h_l\} \Big]. \tag{A.5}$$

For any $l \in J$,

$$\begin{aligned} Pr(y_l = i_l \mid A) &= \frac{Pr\Big[\times_{k\in\bar{J}} \{c_k\} \times \{i_l\} \times \times_{j\in J^{(-l)}} \bar{H}_j \Big]}{Pr(A)} \\ &= \frac{\pi_{\boldsymbol{i}}}{Pr(A)} \sum_{\boldsymbol{\alpha}^{(l)} \in \bar{\boldsymbol{H}}_{J^{(-l)}}} \frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha^{(l)}_j}}{\pi_{\boldsymbol{i}}}. \end{aligned} \tag{A.6}$$

In the above display, we adopt the notation in (A.5), with V partitioned into $(\bar{J}, \{l\}, J^{(-l)})$ and

$$\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha^{(l)}_j} = Pr\Big[\bigtimes_{k\in\bar{J}} \{c_k\} \times \{i_l\} \times \bigtimes_{j\in J^{(-l)}} \{\alpha^{(l)}_j\} \Big].$$

From (A.6), we have

$$\prod_{l\in J} Pr(y_l = i_l \mid A) = \left[\frac{\pi_{\boldsymbol{i}}}{Pr(A)} \right]^{|J|} \sum_{\boldsymbol{\alpha}^{(q)} \in \bar{\boldsymbol{H}}_{J^{(-q)}}} \cdots \sum_{\boldsymbol{\alpha}^{(p)} \in \bar{\boldsymbol{H}}_{J^{(-p)}}} \prod_{l\in J} \frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha^{(l)}_j}}{\pi_{\boldsymbol{i}}}.$$

Substituting this in (A.4), we have (A.4) is equivalent to showing

$$\left[\frac{Pr(A)}{\pi_{\boldsymbol{i}}} \right]^{|J|-1} = \sum_{\boldsymbol{\alpha}^{(q)} \in \bar{\boldsymbol{H}}_{J^{(-q)}}} \cdots \sum_{\boldsymbol{\alpha}^{(p)} \in \bar{\boldsymbol{H}}_{J^{(-p)}}} \prod_{l\in J} \frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha^{(l)}_j}}{\pi_{\boldsymbol{i}}}. \tag{A.7}$$

Recalling the set A from (A.3), we have

$$\frac{Pr(A)}{\pi_{\boldsymbol{i}}} = \sum_{\boldsymbol{\alpha} \in \bar{\boldsymbol{H}}_J} \frac{\pi^{(\bar{J},J)}_{c_k \alpha_j}}{\pi_{\boldsymbol{i}}},$$

[1] Recall our convention, noted in Section 2.3.1, of identifying the event $\{y_1 \in B_1, \ldots, y_p \in B_p\}$ with the event $\times_{j=1}^{p} B_j$ in the discrete σ-algebra generated by $\mathcal{I}_V$.

implying

$$\left[\frac{Pr(A)}{\pi_{\boldsymbol{i}}}\right]^{|J|-1} = \sum_{\boldsymbol{\alpha}_q \in \bar{\boldsymbol{H}}_J} \cdots \sum_{\boldsymbol{\alpha}_{p-1} \in \bar{\boldsymbol{H}}_J} \prod_{l \in J^{(-p)}} \frac{\pi^{(\bar{J},J)}_{c_k \alpha_{lj}}}{\pi_{\boldsymbol{i}}}, \tag{A.8}$$

where $\boldsymbol{\alpha}_q, \ldots, \boldsymbol{\alpha}_{p-1}$ denote $|J|-1$ independent copies of the running index $\boldsymbol{\alpha}$, and α_{lj} is the entry in $\boldsymbol{\alpha}_l$ corresponding to variable j.

It now amounts to show that the expressions in the right hand side of (A.7) and (A.8) are the same. We first argue that both expressions contain the same number of terms. To see this, let $|\bar{H}_j| = m_j$. The expression of $Pr(y_l = i_l \mid A)$ in (A.6) is a sum over $\prod_{j \neq l} m_j$ terms, and so $\prod_{l \in J} Pr(y_l = i_l \mid A)$ has $\prod_{l \in J} \prod_{j \neq l} m_j = \prod_{l \in J} m_l^{(|J|-1)}$ terms. Accordingly, the right hand side in (A.7) has $\prod_{l \in J} m_l^{(|J|-1)}$ many terms. On the other hand, $Pr(A)/\pi_{\boldsymbol{i}}$ is a sum over $\prod_{j \in J} m_j$ terms, and hence $\{Pr(A)/\pi_{\boldsymbol{i}}\}^{(|J|-1)}$ in (A.8) also has $\prod_{j \in J} m_j^{(|J|-1)}$ terms.

Therefore, it now amounts to show that each term inside the summation in the right hand side of (A.7) has a one-to-one correspondence with a term in the right hand side of (A.8). We establish this by showing

$$\prod_{l \in J} \frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha^{(l)}_j}}{\pi_{\boldsymbol{i}}} = \prod_{l \in J^{(-p)}} \frac{\pi^{(\bar{J},J)}_{c_k \alpha_{lj}}}{\pi_{\boldsymbol{i}}}, \tag{A.9}$$

when for each l, $\alpha^{(l)}_j = \alpha_{lj}$ for all $j \neq l$. Introducing additional notation, let $\mathcal{E} = \{E = E_1 \cup \{j\} : E_1 \subset \bar{J}, j \in J_2\}, \mathcal{E}^{(-l)} = \{E = E_1 \cup \{j\} : E_1 \subset \bar{J}, j \in J^{(-l)}\}$ and $\mathcal{E}^{(l)} = \{E = E_1 \cup \{l\} : E_1 \subset \bar{J}\}$. For any l, clearly $\mathcal{E}$ is a disjoint union of $\mathcal{E}^{(-l)}$ and $\mathcal{E}^{(l)}$. Let $\boldsymbol{i}^{(l)}$ denote the cell such that $i^{(l)}_k = c_k$ for $k \in \bar{J}$ and $i^{(l)}_j = \alpha_{lj}$ for $j \in J$.

First, consider the expression in the right hand side of (A.9). We have

$$\frac{\pi^{(\bar{J},J)}_{c_k \alpha_{lj}}}{\pi_{\boldsymbol{i}}} = \exp\left[\sum_{E \subset V} \left\{\theta_E(\boldsymbol{i}^{(l)}_E) - \theta_E(\boldsymbol{i}_E)\right\}\right] = \exp\left[\sum_{E \subset \mathcal{E}} \left\{\theta_E(\boldsymbol{i}^{(l)}_E) - \theta_E(\boldsymbol{i}_E)\right\}\right]$$

$$= \exp\Big[\sum_{E \subset \mathcal{E}^{(-l)}} \Big\{ \theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E) \Big\} \Big] \exp\Big[\sum_{E \subset \mathcal{E}^{(l)}} \Big\{ \theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E) \Big\} \Big] \tag{A.10}$$

The second equality in the above display simply follows from the expression of the cell probabilities for log-linear models in (2.2). The third equality is the key one which uses (i) since $i_k^{(l)} = i_k = c_k$ for all $k \in \bar{J}$, all interaction terms corresponding to $E \subset \bar{J}$ cancel out; and (ii) any $E \subset V$ such that $|E \cap J| \geq 2$, $\theta_E(\boldsymbol{i}_E^{(l)}) = \theta_E(\boldsymbol{i}_E) = 0$, given weak hierarchically and the condition $C_\theta = T_{C_\theta,H}$. To see this, suppose that there exists $E \subset V$ with $|E \cap J| \geq 2$ such that $\theta_E(\boldsymbol{i}_E) \neq 0$ for some $\boldsymbol{i} \in A$. By weak hierarchicality, there must be $j, j^* \in J$ such that $\theta_{\{j,j^*\}}(\alpha_j, \alpha_{j^*}) \neq 0$ for some $(\alpha_j, \alpha_{j^*}) \in \bar{\boldsymbol{H}}_j \times \bar{\boldsymbol{H}}_{j^*}$. Then $\theta_{\{j,j^*\}}(\alpha_j, \alpha_{j^*}) \notin T_{C_\theta,H}$, contradicting $C_\theta = T_{C_\theta,H}$.

Using the same argument and additionally the fact that $\alpha_j^{(l)} = \alpha_{lj}$ for all $j \neq l$, we can simplify the expression in left hand side of (A.9) as

$$\frac{\pi_{c_k i_l \alpha_j^{(l)}}^{(\bar{J},\{l\},J^{(-l)})}}{\pi_{\boldsymbol{i}}} = \exp\Big[\sum_{E \subset \mathcal{E}^{(-l)}} \Big\{ \theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E) \Big\} \Big]. \tag{A.11}$$

Therefore,

$$\prod_{l \in J^{(-p)}} \frac{\pi_{c_k \alpha_{lj}}^{(\bar{J},J)}}{\pi_{\boldsymbol{i}}}$$

$$= \prod_{l \in J^{(-p)}} \exp\Big[\sum_{E \subset \mathcal{E}^{(-l)}} \Big\{ \theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E) \Big\} \Big] \prod_{l \in J^{(-p)}} \exp\Big[\sum_{E \subset \mathcal{E}^{(l)}} \Big\{ \theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E) \Big\} \Big]$$

$$= \prod_{l \in J} \exp\Big[\sum_{E \subset \mathcal{E}^{(-l)}} \Big\{ \theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E) \Big\} \Big] = \prod_{l \in J} \frac{\pi_{c_k i_l \alpha_j^{(l)}}^{(\bar{J},\{l\},J^{(-l)})}}{\pi_{\boldsymbol{i}}},$$

establishing (A.9). The second inequality in the above display used

$$\prod_{l \in J^{(-p)}} \exp\Big[\sum_{E \subset \mathcal{E}^{(l)}} \Big\{ \theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E) \Big\} \Big] = \exp\Big[\sum_{E \subset \mathcal{E}^{(-p)}} \Big\{ \theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E) \Big\} \Big],$$

since $\mathcal{E}^{(-p)} = \bigcup_{l \neq p} \mathcal{E}^{(l)}$ is a disjoint union.

Proof of (2.15) in Theorem 2.3.2

The main idea in this part of the proof is that we can merge certain sets in $\mathcal{P}_H^0$ to create a coarser partition without sacrificing the conditional independence.

For a set $A = \times_{j \in V} A_j$ in $\mathcal{P}_H^0$ and $J \subset V$, let $\Pi_J(A)$ denote

$$\Pi_J(A) = \prod_{j \in J} A_j.$$

With a slight abuse of notation, we shall use $\Pi_l(A)$ to denote the lth coordinate projection, i.e., $\Pi_l(A) = A_l$.

Fix $l \in V$ and let $V^{(-l)} = V \setminus \{l\}$. In this proof, we shall use $\boldsymbol{\alpha}$ to denote a $V^{(-l)}$-cell suppressing the dependence on l. Given $\boldsymbol{\alpha}$, let

$$\mathcal{P}_{H,l}^{\boldsymbol{\alpha}} = \{A \in \mathcal{P}_H^0 : \Pi_{V^{(-l)}}(A) = \bigtimes_{j \neq l} \{\alpha_j\}\}. \tag{A.12}$$

Let $\mathcal{A}$ denote the collection of all $V^{(-l)}$-cells $\boldsymbol{\alpha}$ such that $\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}$ is non-empty. For $\boldsymbol{\alpha} \in \mathcal{A}$, let

$$B^{\boldsymbol{\alpha}} = \bigcup_{A \in \mathcal{P}_{H,l}^{\boldsymbol{\alpha}}} A. \tag{A.13}$$

Note that for any $\boldsymbol{\alpha} \in \mathcal{A}$, $\mid \mathcal{P}_{H,l}^{\boldsymbol{\alpha}} \mid = |H_l|+1$, since $\Pi_l(A)$ ranges over the elements of $\mathcal{P}_{H,l}$, i.e., $\{i_l\}$ for $i_l \in H_l$ and $\bar{H}_l$. It is also evident that $B^{\boldsymbol{\alpha}} = \times_{j \neq l} \{\alpha_j\} \times \mathcal{I}_l$.

We now create a coarser partition $\mathcal{P}_H^{(l)}$ out of $\mathcal{P}_H^0$ by replacing the collection of sets $\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}$ by the single set $B^{\boldsymbol{\alpha}}$ for every $\boldsymbol{\alpha} \in \mathcal{A}$, so that

$$\mathcal{P}_{H,l} = \bigcup_{\boldsymbol{\alpha} \in \mathcal{A}} \Big[(\mathcal{P}_H^0 \setminus \mathcal{P}_{H,l}^{\boldsymbol{\alpha}}) \cup \{B^{\boldsymbol{\alpha}}\} \Big]. \tag{A.14}$$

The main idea is that if $(|V|-1)$ coordinate projections $\Pi_j(A)$ are singletons $\{\alpha_j\}$, we can simply set the lth coordinate projection of A to be $\mathcal{I}_l$ and achieve conditional independence (A.2). This follows immediately from the expression in the display

after (2.11). However, our construction of $\mathcal{P}_H^0$ clearly contains sets of the form $\times_{j\neq l}\{\alpha_j\}\times\{i_l\}$ for $i_l \in H_l$ and $\times_{j\neq l}\{\alpha_j\}\times\bar{H}_l$ which are redundant. To avoid this redundancy, we merge these sets in $\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}$ to form $B^{\boldsymbol{\alpha}} = \times_{j\neq l}\{\alpha_j\}\times\mathcal{I}_l$ for every $\boldsymbol{\alpha}\in\mathcal{A}$.

It only remains to calculate the cardinality of $\mathcal{P}_{H,l}$ now. As pointed out in the previous paragraph, $|\ \mathcal{P}_{H,l}^{\boldsymbol{\alpha}}\ | = |H_l|+1$ for all $\boldsymbol{\alpha}\in\mathcal{A}$, and hence the net reduction in the number of elements from $\mathcal{P}_H^0$ to $\mathcal{P}_{H,l}$ is

$$\mathcal{P}_H^0 - \mathcal{P}_{H,l} = |\ \mathcal{A}\ |\ |\ H_l\ |\ .$$

It thus remains to calculate $|\mathcal{A}|$. We need to count the number of distinct $\boldsymbol{\alpha}$ such that (A.12) is satisfied. Recall that for any $A\in\mathcal{P}_H^0$ and any $j\in V$, $\Pi_j(A)$ ranges over the elements of the partition $\mathcal{P}_{H,j}$. The number of singleton sets in $\mathcal{P}_{H,j}$ is $|H_j|$ as long as $|H_j| < (d-1)$ (the sets $\{i_j\}$ for $i_j\in H_j$). However, when $|H_j| = (d-1)$, $\bar{H}_j$ is also a singleton set and hence the number of singleton sets in $\mathcal{P}_{H,j}$ in that case becomes $|H_j|+1$. Therefore, we conclude,

$$\mathcal{A} = \Bigg[\prod_{j\neq l:|H_j|=d-1}(|\ H_j\ |+1)\Bigg]\Bigg[\prod_{j\neq l:|H_j|<d-1}|\ H_j\ |\ \Bigg].$$

The proof is completed by noting $|\mathcal{A}||H_l| = \prod_{j\in W_l}(|H_j|+1)\prod_{j\in\bar{W}_l}|H_j|$ and taking minimum over $l\in V$ and $H\in\mathscr{H}$.

Proof of Theorem 2.4.2

The condition that $\bigvee_{F\in\mathscr{F}}|F| = o(\log_2(p))$ gives that for each clique F the number of terms in the PARAFAC expansion corresponding to that clique is linear in p. This follows because the maximum PARAFAC rank corresponding to the joint distribution of the variables in each clique is bounded by $2^{\lceil\log_2(p)-1\rceil} = o(p)$. So the joint distribution can be represented by the Hadamard product of k probability tensors $\pi^{(1)},\ldots,\pi^{(k)}$,

with $\mathrm{rnk}_P^+(\pi^{(l)}) = \mathcal{O}(p)$ for every $l = 1, \ldots, k$. Thus, $\sum_{s=1}^k \mathrm{rnk}_P^+(\pi^{(s)}) = \mathcal{O}(kp)$. Note that in the special case where $\log_2(p)$ is an integer and all cliques have identical size, this will give $\sum_{s=1}^k \mathrm{rnk}_P^+(\pi^{(s)}) = p^2/\log_2(p)$.

Dependence graphs associated with PARAFAC and c-Tucker

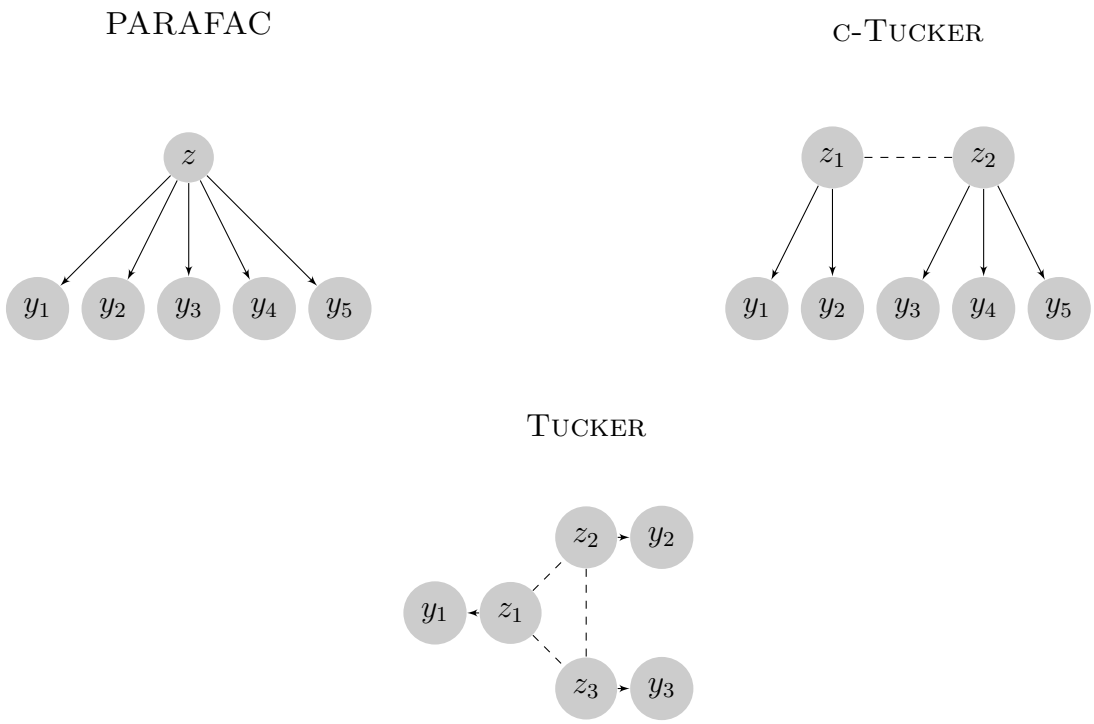


FIGURE A.1*: Graphical representations of hierarchical models inducing PARAFAC, c-Tucker, and Tucker decompositions of π. Dashed edges indicate that there may or may not be an edge between nodes.*

A.2 Supplemental Results

A.2.1 Proof of Remark 3.4

Let

$$\mathscr{H}_g = \{H \in \mathscr{H} : \text{ there exists no } H^* \in \mathscr{H} \text{ for which } H_j^* \subset H_j \text{ for every } j \in U\}.$$

We claim that if $H \in \mathscr{H}_g$, the only partitions satisfying conditional independence with fewer elements than $\mathcal{P}_H^0$ are those in which events in one or more collections $\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}$ are replaced with sets $\mathcal{B}^{\boldsymbol{\alpha}}$ as in the proof of (14). We demonstrate that replacement of any other collections of events by their union gives an event conditional upon which $y_{[1:p]}$ are not independent.

Choose $H \in \mathscr{H}_g$. Let $B_1, B_2 \in \mathcal{P}_H^0$ for which there exists no $l \in V, \boldsymbol{\alpha} \in \mathcal{A}$ such that B_1 and B_2 belong to the same collection $\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}$. Set $A = B_1 \cup B_2$, $J_1 = \{j : H_j^c \cap A = A\}$, $J_2 = \{j : C_j \cap A = A, |C_j| > 1, C_j \neq H_j^c\}$, $J = J_1 \cup J_2$, $J^{(-l)} = J \setminus \{l\}$, and $\bar{J} = V \setminus J$. Then by construction

$$A = \bigtimes_{k \in \bar{J}} \{c_k\} \times \bigtimes_{j \in J_1} \bar{H}_j \times \bigtimes_{j \in J_2} C_j.$$

The conditions imply that J_2 must be nonempty. Otherwise, there would exist $c_l \in H_l$ for some $l \in V$ such that if we set $H_l = H_l \setminus \{c_l\}$ and $H^* = \{\{H_j : j \neq l\}, H_l\}$ then $T_{C_\theta, H^*} = C_\theta$ and $\sum_j \mid H_j^* \mid < \sum_j \mid H_j \mid$, contradicting $H \in \mathscr{H}_g$. Without loss of generality, we suppose $\max_j\{j \in \bar{J}\} < \min_j\{j \in J_1\} = q$ and $\max_j\{j \in J_1\} < \min_j\{j \in J_2\} = r$. Define

$$\bar{\boldsymbol{C}}_E = \bigtimes_{j \in (E \cap J_1)} H_j^c \times \bigtimes_{j \in (E \cap J_2)} C_j.$$

From the proof of (13) in Theorem 3.2, we need to show that with this choice of A, there exists a $\boldsymbol{i} \in \tilde{\mathcal{I}}_V$ such that

$$\sum_{\boldsymbol{\alpha}^{(q)} \in \bar{\boldsymbol{C}}_{J^{(-q)}}} \cdots \sum_{\boldsymbol{\alpha}^{(p)} \in \bar{\boldsymbol{C}}_{J^{(-p)}}} \prod_{l \in J} \frac{\pi_{c_k i_l \alpha_j^{(l)}}^{(\bar{J}, \{l\}, J^{(-l)})}}{\pi_{\boldsymbol{i}}} \neq \sum_{\boldsymbol{\alpha}_q \in \bar{\boldsymbol{C}}_J} \cdots \sum_{\boldsymbol{\alpha}_{p-1} \in \bar{\boldsymbol{C}}_J} \prod_{l \in J^{(-p)}} \frac{\pi_{c_k \alpha_{lj}}^{(\bar{J}, J)}}{\pi_{\boldsymbol{i}}}.$$

Introducing additional notation, let

$$\mathcal{E}_1 = \{E : E \subset (\bar{J} \cup \{j\}), j \in J, E \cap J \neq \emptyset\},$$

$$\mathcal{E}_2 = \{E : E \cap J_2 \neq \emptyset, |E \cap J| \geq 2\},$$

$$\mathcal{E}_1^{(-l)} = \{E : E \subset (\bar{J} \cup \{j\}), j \in J^{(-l)}, E \cap J \neq \emptyset\}, \text{ and}$$

$$\mathcal{E}_2^{(-l)} = \{E : l \notin E, E \cap J_2 \neq \emptyset, |E \cap J| \geq 2\}.$$

Let $\boldsymbol{i}^{(l)}$ be defined as in the proof of (13), and make the additional definition that $\tilde{\boldsymbol{i}}^{(l)}$ is the cell such that $\tilde{i}_k^{(l)} = i_k^{(l)}$ if $k \neq l$ and $\tilde{i}_k^{(l)} = c_k$ if $k = l$. Clearly

$$\begin{aligned}\frac{\pi^{(\bar{J},\{l\},J^{(-l)})}_{c_k i_l \alpha_j^{(l)}}}{\pi_{\boldsymbol{i}}} &= \exp\left[\sum_{E \in (\mathcal{E}_1^{(-l)} \cup \mathcal{E}_2)} \left\{\theta_E(\tilde{\boldsymbol{i}}_E^{(l)}) - \theta_E(\boldsymbol{i}_E)\right\}\right] \\ &= \exp\left[\sum_{E \in \mathcal{E}_1^{(-l)}} \left\{\theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E)\right\}\right] \exp\left[\sum_{E \in \mathcal{E}_2} \left\{\theta_E(\tilde{\boldsymbol{i}}_E^{(l)}) - \theta_E(\boldsymbol{i}_E)\right\}\right],\end{aligned} \tag{A.15}$$

since the l coordinate of the cell $\tilde{\boldsymbol{i}}^{(l)}$ is irrelevant when summing over $E \in \mathcal{E}^{(-l)}$, and

$$\begin{aligned}\frac{\pi^{(\bar{J},J)}_{c_k \alpha_{lj}}}{\pi_{\boldsymbol{i}}} &= \exp\left[\sum_{E \in (\mathcal{E}_1 \cup \mathcal{E}_2)} \left\{\theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E)\right\}\right] \\ &= \exp\left[\sum_{E \in \mathcal{E}_1} \left\{\theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E)\right\}\right] \exp\left[\sum_{E \in \mathcal{E}_2} \left\{\theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E)\right\}\right].\end{aligned} \tag{A.16}$$

For any $E \in \mathcal{E}_1$, we can proceed exactly as in the proof of (13), and therefore we consider the second terms in (A.15) and (A.16). Since $p \in J_2$, there must exist $E^* \subset J$ with $p \in E^*$ such that for some $\boldsymbol{i}^* \in \tilde{\mathcal{I}}_V$, $\theta_E(\boldsymbol{i}_E^*) \neq 0$. Without loss of generality, suppose $E^* = \{p, q\}$ where as above $q = p - |J|$. Let $\tilde{I}_V^{(1)}$ be the set of cells corresponding to the event B_1 and $\tilde{I}_V^{(2)}$ those corresponding to B_2. One of these sets must have $i_q = i_q^*$ and $i_p = i_p^*$ for every element, since $\theta_E(\boldsymbol{i}_E^*) \neq 0$. Take this to be $\tilde{\mathcal{I}}_V^{(1)}$, in which case no $\boldsymbol{i} \in \tilde{\mathcal{I}}_V^{(2)}$ has both $i_q = i_q^*$ and $i_p = i_p^*$. Otherwise, there exists $l \in V$ such that B_1 and B_2 belong to the same collection $\mathcal{P}_{H,l}^{\boldsymbol{\alpha}}$, which is false by choice of B_1 and B_2.

Without loss of generality, suppose that $\boldsymbol{i} \in \tilde{\mathcal{I}}_V^{(2)}$ implies $i_q \neq i_q^*$ and take $c_q \neq i_q^*$. Then $\theta_{E^*}(\tilde{\boldsymbol{i}}_{E^*}^{(l)}) \neq \theta_{E^*}(\boldsymbol{i}_{E^*}^{(l)})$ whenever $\alpha_{lq} = i_q^*$, in which case we have

$$\prod_{l \in J} \exp\left[\sum_{E \in \mathcal{E}_2} \left\{\theta_E(\tilde{\boldsymbol{i}}_E^{(l)}) - \theta_E(\boldsymbol{i}_E)\right\}\right] \neq \prod_{l \in J^{(-p)}} \exp\left[\sum_{E \in \mathcal{E}_2} \left\{\theta_E(\boldsymbol{i}_E^{(l)}) - \theta_E(\boldsymbol{i}_E)\right\}\right].$$

Therefore

$$\sum_{\boldsymbol{\alpha}^{(q)} \in \bar{\boldsymbol{C}}_{J^{(-q)}}} \cdots \sum_{\boldsymbol{\alpha}^{(p)} \in \bar{\boldsymbol{C}}_{J^{(-p)}}} \prod_{l \in J} \frac{\pi_{c_k i_l \alpha_j^{(l)}}^{(\bar{J}, \{l\}, J^{(-l)})}}{\pi_{\boldsymbol{i}}}$$

never has $\exp(\theta_{E^*}(\boldsymbol{i}_{E^*}^*))$ as a factor of any terms of the summand, instead replacing these with $\exp(\theta_{E^*}(\tilde{\boldsymbol{i}}_{E^*}^{(l)}))$ (which may be one), whereas

$$\sum_{\boldsymbol{\alpha}_q \in \bar{\boldsymbol{C}}_J} \cdots \sum_{\boldsymbol{\alpha}_{p-1} \in \bar{\boldsymbol{C}}_J} \prod_{l \in J^{(-p)}} \frac{\pi_{c_k \alpha_{lj}}^{(\bar{J}, J)}}{\pi_{\boldsymbol{i}}}$$

includes $\exp(m(\boldsymbol{\alpha}^{(q)}, \ldots, \boldsymbol{\alpha}^{(p)})\theta_{E^*}(\boldsymbol{i}_{E^*}^*))$ as a factor of every summand, with $m(\boldsymbol{\alpha}^{(q)}, \ldots, \boldsymbol{\alpha}^{(p)}) = |\{l : \alpha_{lq} = i_q^*\}|$. Therefore there exists $\boldsymbol{i} \in \tilde{\mathcal{I}}_V$ for which $Pr(\boldsymbol{y} = \boldsymbol{i}|A) \neq \prod_j Pr(y_j = i_j|A)$, completing the proof.

It follows that for any $H \in \mathscr{H}$ meeting the conditions of Remark 3.4, the only smaller partitions satisfying conditional independence that can be formed from elements of $\mathcal{P}_H^0$ are obtained by replacing events in the collections $\mathcal{P}_{H,A}^{\boldsymbol{\alpha}}$ with events $\mathcal{B}^{\boldsymbol{\alpha}}$. Therefore, noting that every partition of the sample space can be formed from unions of events in the partition consisting solely of Cartesian products of singletons, it follows that every partition satisfying conditional indepndence has at least the number of events given in (14), so long as the conditions of Remark 3.4 are satisfied.

A.2.2 Constructive nonnegative matrix rank result

The following proposition shows that in the two-dimensional case, the nonnegative rank can be bounded by one plus the minimum number of rows and columns that

contain all of the cells that differ from a rank one nonnegative matrix. Figure A.2 shows several examples of the essential principle the proof, which is constructive. Although in the case of probability tensors corresponding to log-linear models, this result is a corollary of Theorem 3.2, the constructive approach is very instructive and provided intuition for the general result.

Proposition A.2.1. *Suppose M is a $d \times d$ nonnegative matrix. Let $\lambda^{(1)}, \lambda^{(2)}$ be nonnegative vectors and set $\tilde{M} = \lambda^{(1)} \otimes \lambda^{(2)}$ with*

$$C_M = \{(c_1, c_2) : M_{c_1 c_2} - \tilde{M}_{c_1 c_2} \neq 0\}, \qquad C_M^{(1)} = \{c_1 : (c_1, c_2) \in C_M\}$$

$$C_M^{(2)} = \{c_2 : (c_1, c_2) \in C_M\},$$

and $\mathscr{H} = \{H : T_{(C_M, H)} = C_M\}$. Define $|H| = |H_1| + |H_2|$. Then $rnk_P^+(M) \leq 1 + \bigwedge_{H \in \mathscr{H}} |H|$.

Proof. Let $H = (H_1, H_2)$ be any element of $\mathscr{H}$. Set

$$\lambda_{0c_1}^{(1)} = \lambda^{(1)} \mathbb{1}(c_1 \notin H_1), \text{ and}$$

$$\lambda_{0c_2}^{(2)} = \lambda^{(2)} \mathbb{1}(c_2 \notin H_2),$$

and put $M^{(0)} = \lambda_0^{(1)} \otimes \lambda_0^{(2)}$. Then for $1 \leq h \leq \mid H_1 \mid$ set

$$\lambda_{hc_1}^{(1)} = \mathbb{1}(c_1 = H_{1h}), \text{ and}$$

$$\lambda_{hc_2}^{(2)} = M_{H_{1h} c_2},$$

where H_{1h} is the hth element of (any ordering of) H_1. Then set $M^{(1)} = \sum_{h=1}^{|H_1|} \lambda_h^{(1)} \otimes \lambda_h^{(2)}$. Finally for $1 \leq h \leq |H_2|$ set

$$\lambda_{hc_1}^{(1)} = M_{c_1 H_{2h}} \mathbb{1}(c_1 \notin H_1) \text{ and,}$$

$$\lambda_{hc_2}^{(2)} = \mathbb{1}(c_2 = H_{2h}),$$

and put $M^{(2)} = \sum_{h=1}^{|H_1|} \lambda_h^{(1)} \otimes \lambda_h^{(2)}$. Then $M^{(0)} + M^{(1)} + M^{(2)} = M$ and therefore M has a $1 + |H| = 1 + \bigwedge_{H' \in \mathscr{H}} (|H'|)$-term nonnegative PARAFAC expansion. □

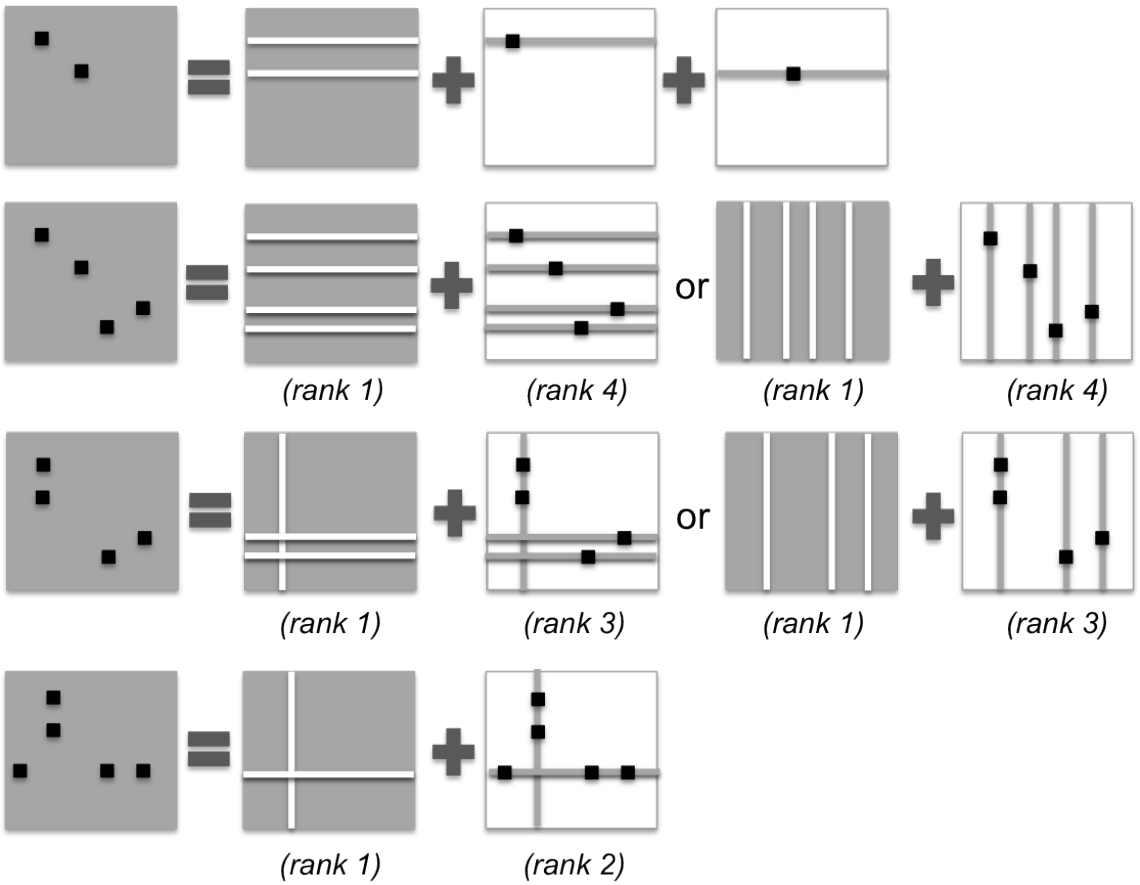


FIGURE A.2*: Examples of nonnegative PARAFAC expansions for matrices. Black indicates cells containing interaction terms, gray indicates cells that do not contain interaction terms, and white indicates cells containing zeros.*

A.3 Posterior computation for c-Tucker models

The conditional posteriors for all the parameters can be derived in closed form using standard algebra and the sampler cycles through the following steps,

Step 1. For $j : s_j = s$ and $h = 1, \ldots, m$, update $\boldsymbol{\lambda}_h^{(j)}$ from the following Dirichlet full conditional posterior distribution,

$$\pi(\boldsymbol{\lambda}_h^{(j)} \mid -) \sim \text{Diri}\left(a_{j1} + \sum_{i:z_{is}=h} 1(y_{ij} = 1), \ldots, a_{jd_j} + \sum_{i:z_{is}=h} 1(y_{ij} = d_j)\right).$$

Step 2. Sample the latent class indicators z_{is} for $s \in \{1, \ldots, k\}$ from the following

full conditional distribution,

$$\text{pr}(z_{is} = h_s \mid -) \propto \left(\prod_{j:s_j=s} \lambda^{(j)}_{h_s y_{ij}} \right) \psi^{(s)}_{w_i h_s},\ h_s = 1, \ldots, m.$$

Step 3. Sample w_i from the following full-conditional distribution,

$$\text{pr}(w_i = l \mid -) \propto \nu_l \prod_{s=1}^{k} \psi^{(s)}_{l z_{is}},\ l = 1, \ldots, k.$$

Step 4. Sample ν^*_l from the following full-conditional distribution,

$$\pi(\nu^*_l \mid -) \sim \text{beta}(1 + m_l, \beta + m_{l+})\, l = 1, \ldots, k,$$

where $m_l = \sum_{i=1}^{n} 1(w_i = l)$ and $m_{l+} = \sum_{i=1}^{n} 1(w_i > l)$.

Step 5. To update $\phi^{(s)}_{lh}$ for $s \in \{1, \ldots, k\}$ define $n^{(s)}_{lh} = \sum_{i:w_i=l} 1(z_{is} = h)$ and $n^{(s)}_{lh+} = \sum_{i:w_i=l} 1(z_{is} > h)$. Then, the full conditional posterior of $\phi^{(s)}_{lh}$ is

$$\pi(\phi^{(s)}_{lh} \mid -) \sim \text{Beta}\left(1 + n^{(s)}_{lh}, \delta_1 + n^{(s)}_{lh+}\right).$$

Step 6. Assuming a gamma(a_β, b_β) prior for β, the full conditional posterior is

$$\pi(\beta \mid -) \sim \text{gamma}\left(a_\beta + k, b_\beta - \sum_{l=1}^{k} \log(1 - \nu^*_l)\right).$$

Step 7. Assuming a gamma$(a^{(s)}_\delta, b^{(s)}_\delta)$ prior for δ_s for each $s \in \{1, \ldots, k\}$, the full conditional posterior is

$$\pi(\delta_1 \mid -) \sim \text{gamma}\left(a^{(s)}_\delta + mk, b^{(s)}_\delta - \sum_{l=1}^{k} \sum_{h=1}^{m} \log(1 - \phi^{(s)}_{lh})\right).$$

Step 8. The groups s_j are updated sequentially. Set $L_{ijl} = Pr(y_j = y_{ij} | s_j = l) = \lambda^{(j)}_{z_l y_j}$, and $L_{jl} = \prod_i L_{ijl}$. Then sample s_j as

$$\pi(s_j = l \mid -) = \frac{\xi_l L_{jl}}{\sum_{l=1}^{k} \xi_l L_{jl}}.$$

Step 9. Let $n_l = \sum_j \mathbb{1}_{s_j=l}$, and sample ξ from

$$p(\xi \mid -) \sim \text{Dirichlet}(n_1 + 1/k, \ldots, n_k + 1/k).$$

A.4 Supplemental figures and tables for section 5

Table A.1: Estimated Cramér's V associations (elements under the main diagonal) and posterior probabilities $Pr(H_{1,\rho}|y^{(1:n)})$ (elements above the main diagonal) in the NLTCS data estimated using the c-Tucker model.

	ADL						**IADL**									
	1	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ADL																
1	-	1.00	1.00	0.00	0.00	0.37	1.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00
2	0.21	-	1.00	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.99	0.99	1.00	0.01	1.00	1.00
3	0.26	0.25	-	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.25	0.17	1.00	0.00	0.68	1.00
4	0.06	0.09	0.12	-	1.00	1.00	1.00	1.00	1.00	1.00	0.05	1.00	0.06	1.00	0.99	0.00
5	0.03	0.06	0.05	0.20	-	1.00	0.99	0.99	0.97	1.00	0.98	1.00	0.36	1.00	1.00	0.00
6	0.10	0.14	0.18	0.38	0.21	-	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.00
IADL																
1	0.20	0.27	0.25	0.16	0.11	0.27	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.14	0.20	0.19	0.20	0.15	0.32	0.42	-	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
3	0.18	0.24	0.20	0.13	0.11	0.22	0.48	0.37	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.04	0.08	0.08	0.19	0.19	0.25	0.14	0.21	0.12	-	0.04	1.00	0.00	1.00	0.99	0.00
5	0.08	0.14	0.10	0.09	0.13	0.14	0.20	0.17	0.22	0.09	-	1.00	1.00	1.00	1.00	0.99
6	0.07	0.12	0.10	0.13	0.18	0.19	0.19	0.21	0.19	0.15	0.22	-	1.00	1.00	1.00	1.00
7	0.15	0.22	0.15	0.09	0.10	0.15	0.30	0.22	0.32	0.09	0.31	0.21	-	1.00	1.00	1.00
8	0.05	0.09	0.07	0.13	0.37	0.17	0.16	0.18	0.18	0.14	0.26	0.25	0.21	-	1.00	0.99
9	0.09	0.14	0.10	0.11	0.19	0.16	0.22	0.21	0.24	0.11	0.30	0.24	0.31	0.41	-	1.00
10	0.17	0.19	0.15	0.06	0.05	0.09	0.22	0.15	0.24	0.05	0.21	0.13	0.33	0.12	0.21	-

Table A.2: Estimated Cramér's V associations (elements under the main diagonal) and posterior probabilities $Pr(H_{1,\rho}|y^{(1:n)})$ (elements above the main diagonal) in the NLTCS data estimated using copula Gaussian graphical model in Dobra and Lenkoski (2011).

	ADL						**IADL**									
	1	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ADL																
1	-	1.00	1.00	0.00	0.00	0.61	1.00	1.00	1.00	0.00	0.00	0.00	0.99	0.00	0.00	1.00
2	0.21	-	1.00	0.43	0.00	1.00	1.00	1.00	1.00	0.00	0.99	1.00	1.00	0.08	1.00	1.00
3	0.26	0.25	-	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.05	0.45	1.00	0.00	0.14	0.99
4	0.07	0.10	0.15	-	1.00	1.00	1.00	1.00	1.00	1.00	0.38	1.00	0.32	1.00	0.98	0.00
5	0.03	0.06	0.05	0.21	-	1.00	0.99	1.00	0.98	1.00	1.00	1.00	0.32	1.00	1.00	0.00
6	0.10	0.14	0.20	0.38	0.21	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.03
IADL																
1	0.21	0.28	0.28	0.18	0.11	0.28	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.14	0.19	0.19	0.21	0.16	0.34	0.43	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	0.16	0.22	0.18	0.13	0.11	0.22	0.48	0.40	-	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.04	0.08	0.08	0.19	0.18	0.25	0.15	0.23	0.13	-	0.33	1.00	0.13	1.00	1.00	0.00
5	0.06	0.12	0.09	0.10	0.14	0.14	0.18	0.17	0.19	0.10	-	1.00	1.00	1.00	1.00	1.00
6	0.06	0.12	0.10	0.14	0.19	0.20	0.19	0.21	0.18	0.17	0.27	-	1.00	1.00	1.00	1.00
7	0.13	0.21	0.14	0.10	0.10	0.14	0.27	0.21	0.28	0.09	0.30	0.23	-	1.00	1.00	1.00
8	0.05	0.09	0.07	0.14	0.39	0.19	0.16	0.19	0.17	0.15	0.26	0.26	0.20	-	1.00	0.95
9	0.07	0.13	0.09	0.11	0.20	0.16	0.20	0.21	0.23	0.12	0.31	0.25	0.30	0.42	-	1.00
10	0.14	0.16	0.13	0.06	0.05	0.09	0.20	0.13	0.20	0.05	0.19	0.12	0.32	0.11	0.20	-

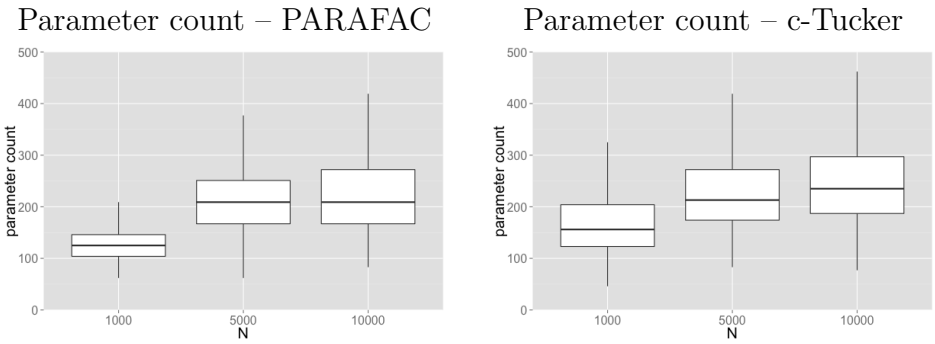


FIGURE A.3*: Left figure: box plot of parameter count for PARAFAC model for the simulation study based on example 4.3. Right figure: box plot of parameter count for c-Tucker model for the same simulation study.*

Appendix B

Appendix to Chapter 3

B.1 Proof of Remark 3.2.1

To see why this is true, note that V is defined by the system of equations

$$\prod_{j:i_j>0} \lambda_{i_j}^{(j)} \prod_{j:i_j=0} \left(1 - \sum_{c_j>0} \lambda_{c_j}^{(j)}\right), = \pi_{i_1,\ldots,i_p} \tag{B.1a}$$

$$\prod_{j:i_j>0} \lambda_{i_j}^{(j)} \prod_{j:i_j=0} \left(1 - \sum_{c_j>0} \lambda_{c_j}^{(j)}\right) \geq 0, \tag{B.1b}$$

$$\lambda_{i_j}^{(j)} \geq 0 \tag{B.1c}$$

for $(i_1, \ldots, i_p) \in \{1, \ldots, d\}^p$ and $j = 1, \ldots, p$, which is a set of polynomial constraints of the form in definition 3.2.3.

B.2 Proof of Proposition 3.2.5

The mixture is clearly contained in the secant variety. We now show that if ν is sampled from a distribution that is absolutely continuous with respect to Lebesgue measure on the simplex, then the corresponding point in $\text{Mixt}^s(V)$ is a non-singular

point in $\text{Sec}^s(\bar{V})$. Recall that

$$\text{Mixt}^s(V) = \left\{ \sum_{h=1}^{s} \nu_h v_h, v_h \in V, \nu \in \Delta^{(s-1)} \right\}.$$

If V is semi-algebraic, then it is defined by finite unions of sets of the form (3.7), and its Zariski closure $\bar{V}$ is obtained by removing the inequalities $h_j(x) > 0$ from the set of constraints. The singularities of $\text{Sec}^s(\bar{V})$ are those points where the Jacobian matrix defined by the partial derivatives $\partial f_i / \partial x_j$ – where the f_i are as in (3.7) and x_j are the elements of V – has lower rank than its maximum rank on $\text{Sec}^s(\bar{V})$.

We now need the following results from Geiger et al. (2001)

Lemma B.2.1 (Geiger et al. 2001, Lemma 9)**.** *Let $g : \mathbb{R}^m \to \mathbb{R}^n$ be a polynomial mapping. Let $J(x) = \partial g / \partial dx$ be the Jacobian matrix at x. Then the rank of $J(x)$ equals the maximal rank almost everywhere.*

We also recall Theorem 3.2.4

Theorem B.2.1 (Geiger et al. 2001, Theorem 10)**.** *Let $g : A \subseteq \mathbb{R}^m \to \mathbb{R}^n$ be a polynomial mapping where A is a semialgebraic open set. Let $J(x) = \partial g / \partial x$ be the Jacobian matrix at x. Then the maximal rank of $J(x)$ is equal to the dimension of $g(A)$.*

Thus, the singularities of $\text{Sec}^s(\bar{V})$ are a set of Lebesgue measure zero, and we can determine the dimension of the image of a semialgebraic set under a polynomial map by computing the maximum rank of the Jacobian (which will be its rank at a random point with probability one). Clearly, the mixture $\text{Mixt}^s(V)$ is not a set of (Lebesgue) measure zero in $\text{Sec}^s(V)$. To see this, note that the parameter space for the mixture is the product of $\Delta^{s-1} \times \left(\times_{j=1}^{p} \mathbb{R}_+^d \right)$, whereas the parameter space for the secant variety is the product $\mathbb{R}^{s-1} \times \left(\times_{j=1}^{p} \mathbb{R}^d \right)$. The former has positive measure

with respect to Lebesgue measure defined over the latter, so a point in $\text{Mixt}^s(V)$ is a singularity of $\text{Sec}^s(V)$ with probability zero. The result follows.

B.3 Proof of Theorem 3.3

We will make use of the following result from Drton et al. (2009)

Proposition B.3.1. *Suppose that* $\text{Sec}^s(V) \neq \text{Aff}(V)$*, the affine hull of* V*. Then*

$$\text{Sec}^{s-1}(V) \subseteq \text{Sing}(\text{Sec}^s(V)),$$

where $\text{Sing}(A)$ *is the set of singularities of* A*.*

Proof. We showed previously that the set V is a semi-algebraic set, and so if ν is sampled from a distribution on Δ_{K-1} that has a density with respect the Lebesgue measure, the corresponding point in $\text{Mixt}^K(V)$ is a non-singular point in $\text{Sec}^K(\bar{V})$, where $\bar{V}$ is the Zariski closure of V. Note that the set of nonnegative matrices of nonnegative PARAFAC rank one coincide with those of ordinary PARAFAC rank one (see my paper with David and Anirban). Now, since $K + (K+1)\sum_j (d_j - 1) < \prod_j d_j - 1$ and $K < \min_l \prod_{j \neq l} d_j$, $\text{Sec}^K(\bar{V}) \neq \text{aff}(V)$, the affine hull of $\bar{V}$. This follows since the rank is less than the maximal possible PARAFAC rank ($K < \min_l \prod_{j \neq l} d_j$) and the number of parameters in $\text{Mixt}^{K+1}(V)$ is less than the number of possible free parameters (choosing $\text{Mixt}^{K+1}(V)$ was done to avoid the case where the number of parameters in $\text{Mixt}^K(V)$ is less than the maximum but the number of parameters in $\text{Mixt}^{K+1}(V)$ is greater than the maximum, which would mean $\text{aff}(V) = \text{Mixt}^K(V)$).

It follows that $\text{Sec}^{K-1}(\bar{V}) \subseteq \text{Sing}(\text{Sec}^K(\bar{V}))$. Since the points of $\text{Mixt}^K(V)$ are not singularities of $\text{Sec}^K(\bar{V})$ almost surely, and since $\text{Sec}^{K-1}(\bar{V}) \supset \text{Mixt}^{K-1}(V)$, it follows that

$$\mu(\text{Mixt}^{K-1}(V)) = 0,$$

since $\text{Mixt}^{K-1}(V) \subset \text{Sing}(\text{Sec}^K(\bar{V}))$. That the singularities are a set of measure zero follows from noting that $\text{Sec}^K(\bar{V})$ is defined by a set of polynomial equations and applying Lemma B.2.1 from Geiger et al. (2001). □

Appendix C

Appendix to Chapter 4

C.1 Log-linear model details

The discussion here largely follows Massam et al. (2009) and Lauritzen (1996) in its presentation. Let V be the set of variables that will be collected into a contingency table. Let $\mathcal{I}_\gamma, \gamma \in V$ denote the set of possible levels of values of γ. Without loss of generality, we can take this set to be a finite collection of sequential nonnegative integers. Let $\mathcal{I} = \times_{\gamma \in V} \mathcal{I}_\gamma$ be the set of all possible combinations of levels of the variables in V. Every cell i of the contingency table corresponds to an element of V; thus $|\mathcal{I}| = d + 1$, where d is defined as in the main text.

Following Lauritzen (1996), define a cell of the contingency table as $i = (i_\gamma, \gamma \in V)$, and let $\pi(i) = \text{pr}[y_1 = i_1, \ldots, y_p = i_p]$. For any $E \subset V$, let $i_E = (i_\gamma, \gamma \in E)$ be the cell of the E-marginal table corresponding to the values in i of the variables in E. Finally, designate the "base" cell $i^* = (0, 0, \ldots, 0)$. Thus, every i can be written as $i = (i_E, i^*_{E^c})$, where E is the subset of V on which $i \neq 0$. Then, the log-linear

model in the corner parametrization is given by

$$\log \frac{\pi(i_E, i^*_{E^c})}{\pi(i^*)} = \sum_{F \subseteq_\emptyset E} \theta_F(i_F),$$

where for any $F \subset V$, $\theta_F(i_F)$ is a parameter corresponding the the variables in F taking the values in i_F, and the notation $\subseteq_\emptyset$ means all subsets excluding the empty set. Refer to Proposition 2.1 in Letac and Massam (2012) for a result showing how the model can be expressed in the form in (4.5).

C.2 Proof of Proposition 4.2.2

This is readily seen by the change of variable theorem; one only needs some work to calculate the Jacobian term for the change of variable. The matrix of partial derivatives $J = (\partial\theta_j/\partial\pi_r)_{jr}$ is given by

$$\frac{\partial\theta_j}{\partial\pi_j} = \frac{1 - \sum_{l\neq j}\pi_l}{\pi_j(1-\sum_{l=1}^d \pi_l)}, \quad \frac{\partial\theta_j}{\partial\pi_r} = -\frac{1}{1-\sum_{l=1}^d \pi_l}, \quad (1 \leq j \neq r \leq d).$$

Write $J = U + uu^{\mathrm{T}}$, where $u = (1-\sum_{l=1}^d \pi_l)^{-1/2}(1, -1, \ldots, -1)^{\mathrm{T}}$ and $U = \mathrm{Diag}(1/\pi_1, \ldots, 1/\pi_d)$. We then have $|J| = |U|(1 + u^{\mathrm{T}}U^{-1}u)$ and therefore,

$$|J|^{-1} = \pi_1 \ldots \pi_d \left(1 - \sum_{l=1}^d \pi_l\right) = \frac{e^{\sum_{l=1}^d \theta_l}}{(1+\sum_{l=1}^d e^{\theta_l})^{d+1}}.$$

The proof is concluded by noting that $p(\theta;\alpha) = q(\ell(\theta);\alpha)\,|J|^{-1}$. □

C.3 Proof of main results

We first state some preparatory results that are used to prove the main results.

C.3.1 Preliminaries

The following identity for the Gamma function is well known (see, e.g., Abramowitz and Stegun (1964)). For $z > 0$,

$$\Gamma(z) = \frac{\log(2\pi)}{2} + \left(z - \frac{1}{2}\right)\log z - z + R(z), \tag{C.1}$$

where $0 < R(z) < 1/(12z)$.

The digamma function $\psi(z) = \frac{d}{dz}\log\Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}$ satisfies $\psi(z+1) = \psi(z) + 1/z$ for any $z > 0$. We use the following bound for the digamma function from Lemma 1 of Chen and Qi (2003). For any $z > 0$,

$$\frac{1}{2z} - \frac{1}{12z^2} < \psi(z+1) - \log z < \frac{1}{2z}. \tag{C.2}$$

The trigamma function $\psi'(z) = \frac{d}{dz}\psi(z)$ is the derivative of the digamma function. We derive a simple bound for the trigamma function that is used in the sequel.

Lemma C.3.1. *For any $z > 1/3$,*

$$\frac{1}{z} < \psi'(z) < \frac{1}{z} + \frac{1}{z^2}. \tag{C.3}$$

The condition $z > 1/3$ is only required for the upper bound.

Proof. From Chen and Qi (2003), the trigamma function admits a series expansion

$$\psi'(z) = \sum_{j=0}^{\infty} \frac{1}{(z+j)^2}$$

valid for any $z > 0$. The function $t \mapsto t^{-2}$ is monotonically decreasing on $(0, \infty)$ and hence $x^{-2} > \int_x^{x+1} t^{-2}dt$ for any $x > 0$. Therefore, for any $z > 0$, $\psi'(z) > \sum_{j=0}^{\infty}\int_{z+j}^{z+j+1} t^{-2}dt = \int_z^{\infty} t^{-2}dt = z^{-1}$. For the upper bound, we use Lemma 1 of Chen and Qi (2003) which states that $1/z - \psi'(z+1) > 1/(2z^2) - 1/(6z^3)$ for

any $z > 0$. Since $\psi(z+1) = \psi(z) + 1/z$, $\psi'(z+1) = \psi'(z) - 1/z^2$, which yields $\psi'(z) - 1/z < 1/z^2 - 1/(2z^2) + 1/(6z^3) = 1/(2z^2) + 1/(6z^3)$ for any $z > 0$. The conclusion follows since $1/(6z^3) < 1/(2z^2)$ for any $z > 1/3$. □

Finally, we state a useful result in Lemma C.3.2.

Lemma C.3.2. *Let $X \in \mathbb{R}^d$ be a random vector with $EX = \mu_X$ and $var(X) = \Sigma_X$. For $\mu \in \mathbb{R}^d$ and $d \times d$ positive definite matrix Σ, the mapping*

$$(\mu, \Sigma) \mapsto g(\mu, \Sigma) = \log|\Sigma| + E(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu) \tag{C.4}$$

attains its minima when $\mu = \mu_X$ and $\Sigma = \Sigma_X$. The minimum value of the objective function $g(\mu_X, \Sigma_X) = \log|\Sigma_X| + d$.

Proof. To start with,

$$\mathbb{E}\left[(X-\mu_X)^{\mathrm{T}}\Sigma_X^{-1}(X-\mu_X)\right] = \mathrm{tr}\left(\mathbb{E}\left[(X-\mu_X)(X-\mu_X)^{\mathrm{T}}\Sigma_X^{-1}\right]\right) = \mathrm{tr}\left(\mathrm{I}_d\right) = d,$$

and hence $g(\mu_X, \Sigma_X) = \log|\Sigma_X| + d$. Fix $\mu \in \mathbb{R}^d$ and Σ positive definite. We can write

$$\begin{aligned}
\mathbb{E}\left[(X-\mu)\Sigma^{-1}(X-\mu)\right] &= \mathrm{tr}\left(\mathbb{E}\left[(X-\mu)(X-\mu)^{\mathrm{T}}\Sigma^{-1}\right]\right) \\
&= \mathrm{tr}\left(\mathbb{E}\left[(X-\mu_X)(X-\mu_X)^{\mathrm{T}}\Sigma^{-1}\right] + (\mu_X-\mu)\Sigma^{-1}(\mu_X-\mu)\right) \\
&= \mathrm{tr}\left(\Sigma_X\Sigma^{-1}\right) + (\mu_X-\mu)^{\mathrm{T}}\Sigma^{-1}(\mu_X-\mu).
\end{aligned}$$

Therefore,

$$g(\mu, \Sigma) - g(\mu_X, \Sigma_X) = \mathrm{tr}\left(\Sigma_X\Sigma^{-1}\right) + (\mu_X-\mu)^{\mathrm{T}}\Sigma^{-1}(\mu_X-\mu) - d - \log|\Sigma_X\Sigma^{-1}|.$$

The above quantity is non-negative since it equals $2D\{N(\mu_X, \Sigma_X) \,||\, N(\mu, \Sigma)\}$, i.e., twice the Kullback–Leibler divergence between $N(\mu_X, \Sigma_X)$ and $N(\mu, \Sigma)$. Since μ and Σ were arbitrary, the first part is proved. The second part has been already proved at the beginning. □

C.3.2 Proof of Theorem 4.3.1 and Corollary 4.3.2

We can now give a proof of Theorem 4.3.1. Recall the Dirichlet density q from (4.6) and the logistic normal density $\widetilde{q}$ from (4.11). We shall write $q(\pi)$ and $\widetilde{q}(\pi)$ in place of $q(\pi \mid \beta)$ and $\widetilde{q}(\pi \mid \mu, \Sigma)$ henceforth for brevity. From (4.6) and (4.11),

$$\log \frac{q(\pi)}{\widetilde{q}(\pi)} = \log B_\beta + \frac{d \log(2\pi)}{2} + \sum_{j=0}^{d} \beta_j \log \pi_j + + \frac{\log|\Sigma|}{2}$$

$$+ \frac{1}{2}\{ \log(\pi/\pi_0) - \mu\}^{\mathrm{T}} \Sigma^{-1} \{ \log(\pi/\pi_0) - \mu\}.$$

Observe that μ and Σ appear only in the last two terms in the right hand side of the above display. Invoking Lemma C.3.2, it is therefore evident that $D(q \mid\mid \widetilde{q}) = E_q \log(q/\widetilde{q})$ is minimized when $\mu^* = E_q \log(\pi/\pi_0)$ and $\Sigma^* = \mathrm{var}_q\{\log(\pi/\pi_0)\}$, and the minimum vaue of the Kullback–Leibler divergence is

$$\log B_\beta + \sum_{j=0}^{d} \beta_j E_q \log \pi_j + \frac{d}{2}\{1 + \log(2\pi)\} + \frac{\log|\Sigma^*|}{2}. \tag{C.5}$$

Using standard properties of the Dirichlet distribution or Exponential family differential identities, with $\beta = \sum_{j=0}^{d} \beta_j$,

$$E_q \log \pi_j = \psi(\beta_j) - \psi(\beta), \quad j = 0, 1, \ldots d, \tag{C.6}$$

$$\mathrm{cov}_q(\log \pi_j, \log \pi_l) = \psi'(\beta_j)\delta_{jl} - \psi'(\beta), \quad j, l = 0, 1, \ldots, d. \tag{C.7}$$

Therefore, $\mu_j^* = E_q \log \pi_j - E_q \log \pi_0 = \psi(\beta_j) - \psi(\beta_0)$ for $j = 1, \ldots d$. Next, $\sigma_{jj'}^* = \mathrm{cov}_q(\log \pi_j - \log \pi_0, \log \pi_{j'} - \log \pi_0) = \delta_{jj'}\psi'(\beta_j) + \psi'(\beta_0)$ for $j, j' = 1, \ldots, d$. The expressions for μ^* and Σ^* are identical to (4.8), proving the first part of the theorem. Note this also establishes Proposition 4.2.3.

We now proceed to bound each term in the expression for the minimum Kullback–Leibler divergence in (C.5); refer to them by T_1, T_2, T_3 and T_4 respectively. First, we have,

$$T_1 := \log B_\beta = \log \Gamma(\beta) - \sum_{j=0}^{d} \Gamma(\beta_j)$$

$$< -\frac{d\log(2\pi)}{2} + \left(\beta\log\beta - \sum_{j=0}^{d}\beta_j\log\beta_j\right) - \frac{1}{2}\left(\log\beta - \sum_{j=0}^{d}\log\beta_j\right) + \frac{1}{12\beta}. \tag{C.8}$$

In the above display, we used (C.1) to bound $\log\Gamma(\beta)$ from above and $\log\Gamma(\beta_j)$s from below. The $(-\beta)$ term in upper bound to $\log\Gamma(\beta)$ cancels out the $(-\sum_{j=0}^{d}\beta_j)$ contribution from the lower bounds to the $\log\Gamma(\beta_j)$s. Next,

$$\begin{aligned} T_2 &:= \sum_{j=0}^{d}\beta_j E_q\pi_j = \sum_{j=0}^{d}\beta_j\{\psi(\beta_j) - \psi(\beta)\} \\ &= \sum_{j=0}^{d}\beta_j\{\psi(\beta_{j+1}) - \psi(\beta+1)\} - \sum_{j=0}^{d}\beta_j\left(\frac{1}{\beta_j} - \frac{1}{\beta}\right) \\ &= \left\{\sum_{j=0}^{d}\beta_j\psi(\beta_{j+1}) - \beta\psi(\beta)\right\} - d \\ &< \left(\sum_{j=0}^{d}\beta_j\log\beta_j - \beta\log\beta\right) - \frac{d}{2} + \frac{1}{12\beta}. \end{aligned} \tag{C.9}$$

In the first line of the above display, we used (C.6). From the first to the second line, we used the identity $\psi(z+1) = \psi(z) + 1/z$. From the second to the third line, we only use $\sum_{j=0}^{d}\beta_j = \beta$. From the third to the fourth line, we made use of the bound (C.2) for the digamma function ψ. From the upper bound in (C.2), $\beta_j\psi(\beta_{j+1}) < \beta_j\log\beta_j + 1/2$ and hence $\sum_{j=0}^{d}\beta_j\psi(\beta_{j+1}) < \sum_{j=0}^{d}\beta_j\log\beta_j + (d+1)/2$. From the lower bound in (C.2), $\beta\psi(\beta) > \beta\log\beta + 1/2 - 1/(12\beta)$.

Finally, from (C.7), we can write $\Sigma^* = D + \psi'(\beta_0)\mathbf{1}\mathbf{1}^{\mathrm{T}}$, with $D = \mathrm{diag}(\psi'(\beta_1), \ldots, \psi'(\beta_d))$. Using the fact $|X + uv^{\mathrm{T}}| = |X|(1 + v^{\mathrm{T}}X^{-1}u)$, we obtain

$$|\Sigma^*| = \left\{1 + \sum_{j=1}^{d}\psi'(\beta_0)/\psi'(\beta_j)\right\}\left\{\prod_{j=1}^{d}\psi'(\beta_j)\right\} = \left\{\sum_{j=0}^{d}\frac{\psi'(\beta_0)}{\psi'(\beta_j)}\right\}\left\{\prod_{j=1}^{d}\psi'(\beta_j)\right\}.$$

From Lemma C.3.1, $\psi'(\beta_j) > 1/\beta_j$, implying

$$T_4 := \frac{\log|\Sigma^*|}{2} = \frac{1}{2}\Bigg[\log\Bigg\{ \sum_{j=0}^{d} \frac{\psi'(\beta_0)}{\psi'(\beta_j)} \Bigg\} + \sum_{j=1}^{d} \log \psi'(\beta_j) \Bigg]$$

$$< \frac{1}{2}\Bigg\{ \log \beta + \sum_{j=0}^{d} \log \psi'(\beta_j) \Bigg\}. \tag{C.10}$$

Recalling $T_3 = d\{1 + \log(2\pi)\}/2$ and substituting the bounds for T_1, T_2 and T_4 from (C.8), (C.9) and (C.10) in (C.5), we obtain, after plenty of cancellations,

$$\sum_{j=1}^{4} T_j < \frac{1}{2} \sum_{j=0}^{d} \log\{\beta_j \psi'(\beta_j)\} + \frac{1}{6\beta}$$

$$< \frac{1}{2} \sum_{j=0}^{d} \frac{1}{\beta_j} + \frac{1}{6\beta}.$$

From the first to the second line, we invokeed Lemma C.3.1 to bound $\beta_j \psi'(\beta_j) < 1 + 1/\beta_j$ and used $\log(1+x) < x$ for $x > 0$. We have obtained the desired bound, concluding the proof.

Now, to show Corollary 4.3.2, just note that by the invariance of D under one-to-one transformations, we have that for any full rank matrix X,

$$D\Big\{ \mathcal{LD}(\beta) \;||\; \mathcal{N}(\mu, \Sigma) \Big\} = D\Big\{ \mathcal{P}_X(\cdot; \beta) \;||\; \mathcal{N}(X\mu, X^{\mathrm{T}}\Sigma X) \Big\}. \tag{C.11}$$

So

$$\inf_{\mu,\Sigma} \Big\{ \mathcal{LD}(\beta) \;||\; \mathcal{N}(\mu, \Sigma) \Big\} = \inf_{\widetilde{\mu},\widetilde{\Sigma}} D\Big\{ \mathcal{P}_X(\cdot; \beta) \;||\; \mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma}) \Big\}. \tag{C.12}$$

Since the infimum on the left side in (C.12) is attained by μ^*, Σ^*, we have by (C.11) that

$$D\left(\mathcal{P}_X(\cdot; \beta) \;||\; \mathcal{N}(\cdot; X\mu^*, X^T\Sigma^* X)\right) = \inf_{\mu,\Sigma} D\left(\mathcal{P}_X(\cdot; \beta) \;||\; \mathcal{N}(\cdot; \mu, \Sigma)\right),$$

which gives Corollary 4.3.2.

Appendix D

Appendix to Chapter 5

Throughout, $\mathcal{N}(\cdot) = \mathcal{N}(du\, d\xi\, ds\, d\tau\, dv\, d\Lambda)$ will refer to a Poisson random measure on $\Omega = [0,\infty) \times \mathbb{R}^d \times \mathbb{R} \times [0,\infty) \times \mathbb{R}^d \times \mathcal{L}$. We use ω or ω_i to represent an element of Ω. The intensity measure for the process is given by $u^{-2}\, du\, d\xi\, ds\, \delta e^{-\delta\tau}\, \pi(dv d\Lambda)$. Here, s represents birth times, τ durations, v velocities, and $\Lambda \in \mathcal{L}$ parameters controling the shape of the kernels $k(x,y) = k(x,y;\Lambda)$. The parameters Λ may differ depending on k so we represent the space abstractly as $\mathcal{L}$. Parts of some results will take k to be the isotropic Gaussian kernel, in which case Λ is a positive-definite $d \times d$ matrix. The results all refer to the process

$$Y(x,t) = \sup_j u_j k(x, \xi_j + v_j(t - s_j)) \mathbb{1}_{\{t-\tau_j < s_j < t\}},$$

with other restrictions or assumptions specified in theorem statements and in the course of giving the arguments.

D.1 Proof of Theorem 5.2.2

First, we show that the number of points $\mathcal{N}$ places in an arbitrary connected set B at time t is a spatial Poisson process with time-invariant intensity.

Let A be an arbitrary connected set in $\mathbb{R}^d$ with boundary ∂A, and consider the time interval $(t, t+\eta]$ for small η. Focus on a point σ on the boundary, or an interval of length ϵ around σ. To cross ∂A at an interval of length ϵ around σ in time η, a particle with velocity v needs to be within distance $\eta|v \cdot \nu|$ where ν is the outward pointing unit normal vector, so $(v \cdot \nu)$ is the velocity component perpendicular to the segment. If $(v \cdot \nu)$ is positive then the point is leaving A; if $(v \cdot \nu)$ is negative it is entering A. So the points going into A through the interval of boundary around σ will be those in a rectangle that is ϵ long and $\eta|v \cdot \nu|$ wide. Taking expectations for the random velocity v, we get

$$\mathbb{E}[\#\text{going into } A \text{ through this boundary element in time } \eta]$$
$$= \epsilon\eta \int (-v \cdot \nu) \mathbb{1}_{\{v \cdot \nu < 0\}} dv = \epsilon\eta(-V \cdot \nu) \mathbb{1}_{\{v \cdot \nu < 0\}} dv$$

where V is the mean velocity vector. To get the $\mathbb{E}[\#$ entering] in a macroscopic time interval, integrate $(-V \cdot \nu)\mathbb{1}_{\{v \cdot \nu < 0\}} dt$ over the time interval; So $\mathbb{E}[\#$ entering] over the whole boundary is $(-V \cdot \nu)\mathbb{1}_{\{v \cdot \nu < 0\}} dt d\sigma$ over the boundary (ν depends on the boundary point). To get the net flux across the boundary, remove the indicator and integrate $(-V \cdot \nu) dt d\sigma$ over the boundary— and get zero, by the divergence theorem, i.e.

$$\oiint_{\partial A} (-V \cdot \nu) d\sigma = \int \int_A -\nabla \cdot V dX.$$

Since $G(x, y) = (x, y) \cdot V$ is the potential function and $V = \nabla G(x, y)$ its gradient, we have that

$$\nabla \cdot V = \nabla \cdot \nabla G(x, y) = \Delta G(x, y)$$

is the Laplacian of G, which is zero (second derivative of a linear function). So the entire expression $\iint_A -\nabla \cdot V dX$ is zero, confirming that the distribution for number of particles in A is time invariant, so long as $\mathbb{E}[v] < \infty$. The result for arbitrary sets

in $\mathbb{R}^d$ follows by considering countable unions of connected sets. It follows that when $\nu(d\xi) = \beta d\xi$, the number of points $\mathcal{N}$ places in any set B is homogeneous in both space and time and does not depend on the distribution of v.

Now, fix a time t, and consider the distribution of $Y(x,t)$. Since the number of points in any set $A \subset \mathbb{R}^k$ at time t follows a homogeneous spatial Poisson process with time-invariant intensity for any $\pi(dv)$ with finite expectation, this holds for the specific case of $\pi(dv) = \delta_0$. So the spatial intensity measure for $Y(x,t)$ is given by

$$\int_{\tau\in[0,\infty)} \int_{t-\tau}^{t} \beta\delta e^{-\delta\tau} ds d\tau = \frac{\beta}{\delta}.$$

Therefore for fixed t, $Y(x,t) = \tilde{Y}(x) = \sup_j u_j k(x,\xi_j)$ for (ξ_j, u_j) a spatial Poisson process with intensity measure $\frac{\beta}{\delta} d\xi u^{-2} du$. Then by Theorem 5.2.1, $Y(x,t)$ for fixed t is max-stable.

Finally, consider $Y*(x,t) := \sup_{0<t^*<t} Y(x,t^*)$. When $v = 0$, we have

$$Y^*(x,t) = \sup_j u_j k(x,\xi_j) \mathbb{1}_{\{-\tau<s<t\}},$$

so $Y*$ is max-stable with intensity measure $u^{-2} du \nu(d\xi)$ where

$$\nu(d\xi) = \int_{\tau\in[0,\infty)} \int_{-\tau<s<t} \beta\delta e^{-\delta\tau} ds d\tau d\xi = \left[t + \frac{1}{\delta}\right] d\xi.$$

Lemma D.1.1. *If k is isotropic, the function $k^*(\omega,x,t) = \sup_{0<t^*<t} k(x, \xi + v(t^*-s))$ satisfies $\int_{\omega\in\Omega} k^*(\omega,t)\pi(d\omega) \le t \int_v (1+|v|)\pi(dv) = t(1+\mathbb{E}[|v|])$.*

Proof. Consider any point $x \in \mathbb{R}^d$. The function $k^*(\omega,t)$ is defined by the point along the line segment $[\xi - vs, \xi + v(t-s)]$ that is nearest to x in the distance defined by the kernel $k(x,y) = k(0, x-y)$. Let $A \subset \mathbb{R}^d$ be any set. We can bound $\int k^*(\omega,t) d\xi$ in the following way. Fix $v > 0$. Then the rate of flow of volume passing through A is given by $|v|dt$. Because net flux is zero, the total volume of space containing points that could possibly reside in A at any time $0 < t^* < t$ is given by $|A| + t|A||v|$.

Now, since the birth locations ξ are uniformly distributed in $\mathbb{R}^d$, for any fixed v, the coordinates of the point that achieves $\sup_{0<t^*<t} k(x, \xi + v(t^* - s))$ will be uniformly distributed in the orthogonal complement of the line of travel. Denote the space containing the line of travel by V, its orthogonal complement $V^\perp$, and note that $V^\perp$ is a $d-1$ dimensional Euclidean space. Denote a point in $V^\perp$ by ξ^*, and the projection of x onto $V^\perp$ by x^*. Let $\Lambda^\perp$ be the parameter of the kernel shape in $V^\perp$. Therefore if k is isotropic, we have

$$\begin{aligned}\int_{v,\Lambda}\int_{\tau\in\mathbb{R}_+}\int_{-\tau<s<t}\int_{\xi\in\mathbb{R}^d} k^*(\omega,t)\pi(d\omega) &< \int_{v,\Lambda}\int_{\tau\in\mathbb{R}_+}\int_{-\tau<s<t}\int_{\xi^*\in\mathbb{R}^{d-1}} \beta k(||x^*-\xi^*||)(1+|v|)t\pi(d\omega)\\ &< \int_v\int_{\tau\in\mathbb{R}^d}\int_{-\tau<s<t} \beta\delta e^{-\delta\tau}\frac{c_d(\Lambda)}{c_{d-1}(\Lambda^\perp)}(1+|v|)td\omega\\ &< \int_v \frac{c_d(\Lambda)}{c_{d-1}(\Lambda^\perp)}\frac{\beta t}{\delta}(1+|v|)dv.\end{aligned}$$

where $c_d(\Lambda)$ is the normalizing constant of the kernel in dimension d with parameter Λ and $c_{d-1}(\Lambda^\perp)$ is the normalizing constant of the kernel in dimension $d-1$ with parameters Λ^*. The last line is finite so long as $\mathbb{E}_{\pi(dv)}[|v|] < \infty$. If Λ is random, then the integrability condition is that

$$\mathbb{E}_{\pi(dv,d\Lambda)}\left[\frac{c_d(\Lambda)}{c_{d-1}(\Lambda^\perp)}|v|\right] \tag{D.1}$$

is finite. □

D.2 Results on Marginal and Joint distributions of Max-stable velocity process

D.2.1 Proof of Theorem 5.2.3

The CDF at time t is also $\mathbb{P}[\mathcal{N}(A)] = 0$ for the set

$$A = \{\omega : \sup_j u_j k(x, \xi_j + v_j(t - s_j))\mathbb{1}_{\{t-\tau_j<s_j<t\}} > y\},$$

which is available by direct integration. The condition

$$u_j k(x, \xi_j + v_j(t - s_j)) > y$$

corresponds to $u_j > y/k(x, \xi_j + v_j(t - s_j))$, giving the integral expression

$$|A| = \int_{(v,\Lambda)\in\mathbb{R}^k\times\mathcal{L}} \int_{\tau\in\mathbb{R}_+} \int_{t-\tau<s<t} \int_{\xi\in\mathbb{R}^k} \int_{u>\frac{y}{k(x,\xi+v(t-s))}} \beta u^{-2}\delta e^{-\delta\tau}\pi(dv, d\Lambda)du d\xi ds d\tau dv d\Lambda$$

$$\int_{(v,\Lambda)\in\mathbb{R}^k\times\mathcal{L}} \int_{\tau\in\mathbb{R}_+} \int_{t-\tau<s<t} \int_{\xi\in\mathbb{R}^k} \frac{\beta k(x, \xi + v(t - s))}{y}\delta e^{-\delta\tau}\pi(dv, d\Lambda)d\xi ds d\tau dv d\Lambda.$$

Change variables to $z = x - (\xi + v(t - s))$ to obtain

$$|A| = \int_{(v,\Lambda)\in\mathbb{R}^k\times\mathcal{L}} \int_{\tau\in\mathbb{R}_+} \int_{t-\tau<s<t} \int_{z\in\mathbb{R}^k} \frac{\beta k(0, z)}{y}\delta e^{-\delta\tau}\pi(dv, d\Lambda)dz ds d\tau dv d\Lambda.$$

Since $k(0, z)$ satisfies $\int_{z\in\mathbb{R}^d} k(0, z)dz = 1$, then if $\nu(d\xi) = \beta d\xi$, we have

$$\begin{aligned}
|A| = & \int_{(v,\Lambda)\in\mathbb{R}^k\times\mathcal{L}} \int_{\tau\in\mathbb{R}_+} \int_{t-\tau<s<t} \frac{\beta}{y}\delta e^{-\delta\tau}\pi(dv, d\Lambda)ds d\tau dv d\Lambda \\
& \frac{\beta}{y} \int_{(v,\Lambda)\in\mathbb{R}^k\times\mathcal{L}} \int_{\tau\in\mathbb{R}_+} \tau\delta e^{-\delta\tau}\pi(dv, d\Lambda)d\tau dv d\Lambda \\
& \frac{\beta}{\delta y} \int_{(v,\Lambda)\in\mathbb{R}^k\times\mathcal{L}} \pi(dv, d\Lambda)dv d\Lambda \\
& \frac{\beta}{\delta y},
\end{aligned}$$

so $\mathbb{P}[Y(x, t) < y] = e^{-\beta/(\delta y)}$, a Fréchet distribution with unit shape.

When $v = 0$, the survival function of the waiting time is given by $\mathbb{P}[\mathcal{N}(B)] = 0$ for the set

$$B = \{(\xi, u, a) : \sup_j u_j k(x, \xi_j)\mathbb{1}_{\{-\tau_j<s_j<t\}} > y\}, \tag{D.2}$$

so the integral will be nearly identical, with the result $|B| = \frac{\beta(t+\delta^{-1})}{y}$, giving $\mathbb{P}[\kappa(y) > t] = e^{-\beta(t+\delta^{-1})/y}$.

When $v \neq 0$, the waiting time distribution can be derived from the measure of three sets

1. A_0, the set of points that causes an exceedance at $t = 0$
2. A_1, the set of points that causes an exceedance before time t but after time 0, where the exceedance occurs at birth
3. A_2, the set of points that causes an exceedance before time t but after time 0 where the exceedance does not occur at birth if the storm is born after time 0.

The measure of $A_0 \cup A_1$ is exactly the measure of the set B in (D.2); so $|A_0| = \frac{\beta}{\delta y}$ and $|A_1| = \frac{\beta t}{y}$. Now, define

$$k^*(x,\omega,t) = \sup_{0<t^*<t} k(x, \xi + v(t^* - s))\mathbb{1}_{\{s+\tau>t^*\}}.$$

Then

$$\begin{aligned}
|A_2| = &\int_{v,\Lambda}\int_{\tau\in(0,\infty)}\int_{-\tau<s<0}\int_{\xi\in\mathbb{R}^d}\int_{u=\frac{y}{k^*(x,\omega,t)}}^{u=\frac{y}{k(x,\xi-vs)}} \beta u^{-2}\mathbb{1}_{\{\frac{k^*(x,\omega,t)}{k(x,\xi-vs)}>1\}}du\pi(d\omega) \\
&+\int_{v,\Lambda}\int_{\tau\in(0,\infty)}\int_{0<s<t}\int_{\xi\in\mathbb{R}^d}\int_{u=\frac{y}{k^*(x,\omega,t)}}^{u=\frac{y}{k(x,\xi)}} \beta u^{-2}\mathbb{1}_{\{\frac{k^*(x,\omega,t)}{k(x,\xi)}>1\}}du\pi(d\omega) \\
&=\int_{v,\Lambda}\int_{\tau\in(0,\infty)}\int_{-\tau<s<0}\int_{\xi\in\mathbb{R}^d}\frac{\beta(k^*(x,\omega,t)-k(x,\xi-vs))}{y}\mathbb{1}_{\{\frac{k^*(x,\omega,t)}{k(x,\xi-vs)}>1\}}\pi(d\omega) \\
&+\int_{v,\Lambda}\int_{\tau\in(0,\infty)}\int_{0<s<t}\int_{\xi\in\mathbb{R}^d}\frac{\beta(k^*(x,\omega,t)-k(x,\xi))}{y}\delta e^{-\delta\tau}\mathbb{1}_{\{\frac{k^*(x,\omega,t)}{k(x,\xi)}>1\}}\pi(d\omega) \\
&=\frac{\beta f(t)}{y},
\end{aligned}$$

for some positive, monotone increasing function f. Then, since the sets A_0, A_1, A_2 are disjoint, the survival function for the waiting time until first exceedance is given by

$$\mathbb{P}[\kappa_1 > t] = \exp(-(|A_0|+|A_1|+|A_2|)) = \exp\left(-\frac{\beta(\delta^{-1}+t+f(t))}{y}\right).$$

When k is isotropic, f is bounded by $\mathbb{E}_{\pi(dv,d\Lambda)}\left[\frac{c_d(\Lambda)}{c_{d-1}(\Lambda^\perp)}|v|\right]$ by Lemma D.1.1.

D.2.2 Proof of Theorem 5.2.4

Here we derive the joint distribution of the max-stable velocity process at two points x_1, x_2.

Fix $x_1, x_2 \in \mathbb{R}^d$ and $-\infty < t_1 \le t_2 < \infty$. Then

$$\mathbb{P}[Y(x_1,t_1) \le y_1, \quad Y(x_2,t_2) \le y_2] = \mathbb{P}[\mathcal{N}(B) = \emptyset] = \exp(-|B|),$$

where

$$B = \{\omega : \ u\,k(x_j, \xi + (t_j - s)v)\mathbb{1}_{\{s<t_j\le s+\tau\}} > y_j \text{ for } j = 1 \text{ or } j = 2\}$$

$$|B| = \mathbb{E}\mathcal{N}(B) = \int_B \beta\delta e^{-\delta\tau}u^{-2}\,du\,d\xi\,ds\,d\tau\pi(dv\,da)$$

$$= |B_1|+|B_2|$$

for the disjoint sets B_1, B_2 given by

$$B_1 = \{\omega : u\,k(x_1, \xi + (t_1 - s)v)\mathbb{1}_{\{s<t_1\le s+\tau\}} > y_1\}$$

$$\cap\,\{\omega : y_1\,k(x_2, \xi + (t_2 - s)v)\mathbb{1}_{\{s<t_2\le s+\tau\}} \le y_2\,k(x_1, \xi + (t_1 - s)v)\mathbb{1}_{\{s<t_1\le s+\tau\}}\}$$

$$B_2 = \{\omega : u\,k(x_2, \xi + (t_2 - s)v)\mathbb{1}_{\{s<t_2\le s+\tau\}} > y_2\}$$

$$\cap\,\{\omega : y_2\,k(x_1, \xi + (t_1 - s)v)\mathbb{1}_{\{s<t_1\le s+\tau\}} \le y_1\,k(x_2, \xi + (t_2 - s)v)\mathbb{1}_{\{s<t_2\le s+\tau\}}\}.$$

To calculate $|B_1|$ first integrate wrt $d\tau\,du$, separately over the sets $t_1 < s+\tau \le t_2$ and $t_2 < s+\tau$:

$$|B_1| = \int_{B_1\cap[t_1<s+\tau\le t_2]} \beta\delta e^{-\delta\tau}u^{-2}\,du\,d\xi\,ds\,d\tau\,\pi(dv,da)$$

$$+\int_{B_1\cap[t_2<s+\tau<\infty]}\beta\delta e^{-\delta\tau}u^{-2}\,du\,d\xi\,ds\,d\tau\,\pi(dv,da)$$

$$=\int[e^{-\delta(t_1-s)}-e^{-\delta(t_2-s)}]\frac{\beta}{y_1}\mathbb{1}_{\{s\le t_1\}}k(x_1,\xi+(t_1-s)v)\,d\xi\,ds\,\pi(dv,da)$$

$$+\int e^{-\delta(t_2-s)}\frac{\beta}{y_1}\mathbb{1}_{\{s\le t_1\}}\mathbb{1}_{\{\frac{k(x_1,\xi+(t_1-s)v)}{k(x_2,\xi+(t_2-s)v)}\ge\frac{y_1}{y_2}\}}k(x_1,\xi+(t_1-s)v)\,d\xi\,ds\,\pi(dv,da)$$

Changing variables from ξ to $z:=[\xi+(t_1-s)v-x_1]\in\mathbb{R}^d$, setting

$$\Delta(x_1,x_2,t_1,t_2,v)=\Delta(v):=x_2-x_1-(t_2-t_1)v; \tag{D.3}$$

we will often suppress dependence of Δ on the non-stochastic arguments x_1,x_2,t_1,t_2. So $[\xi+(t_2-s)v-x_2]=z-\Delta(v)$, and using $k(x,y)=k(0,x-y)$ we have

$$|B_1|=\int[e^{-\delta(t_1-s)}-e^{-\delta(t_2-s)}]\frac{\beta}{y_1}\mathbb{1}_{\{s\le t_1\}}k(0,z)\,dz\,ds\,\pi(dv,d\Lambda)$$

$$+\int e^{-\delta(t_2-s)}\frac{\beta}{y_1}\mathbb{1}_{\{s\le t_1\}}\mathbb{1}_{\{\frac{k(0,z)}{k(0,z-\Delta(v))}\ge\frac{y_1}{y_2}\}}k(0,z)\,dz\,ds\,\pi(dv,d\Lambda)$$

$$=\int[1-e^{-\delta(t_2-t_1)}]\frac{\beta/\delta}{y_1}k(0,z)\,dz\,\pi(dv,d\Lambda)$$

$$+\int e^{-\delta(t_2-t_1)}\frac{\beta/\delta}{y_1}\mathbb{1}_{\{\frac{k(0,z)}{k(0,z-\Delta(v))}\ge\frac{y_1}{y_2}\}}k(0,z)\,dz\,\pi(dv,d\Lambda)$$

$$=\int\frac{\beta/\delta}{y_1}k(0,z)\,dz\,\pi(dv,d\Lambda)-\int e^{-\delta(t_2-t_1)}\frac{\beta/\delta}{y_1}\mathbb{1}_{\{\frac{k(0,z)}{k(0,z-\Delta(v))}<\frac{y_1}{y_2}\}}k(0,z)\,dz\,\pi(dv,d\Lambda)$$

Using $\int_{z\in\mathbb{R}^d}k(0,z)dz=1$, we have

$$|B_1|=\int\frac{\beta/\delta}{y_1}\,\pi(dv,d\Lambda)-\int e^{-\delta(t_2-t_1)}\frac{\beta/\delta}{y_1}\mathbb{1}_{\{\frac{k(0,z)}{k(0,z-\Delta(v))}<\frac{y_1}{y_2}\}}k(0,z)\,dz\,\pi(dv,d\Lambda)$$

$$=\frac{\beta/\delta}{y_1}-\frac{\beta/\delta}{y_1}\int e^{-\delta(t_2-t_1)}F(\Lambda,\Delta(v);y_1,y_2)\,\pi(dv,d\Lambda)$$

$$=\frac{\beta/\delta}{y_1}\left[1-e^{-\delta(t_2-t_1)}\int F(\Lambda,\Delta(v);y_1,y_2)\,\pi(dv,d\Lambda)\right] \tag{D.4}$$

for some function F satisfying $0\le F\le 1$. Since $\pi(dvd\Lambda)$ is a probability measure, this is enough to guarantee $|B_1|<\infty$, giving the general result. We now take k to be

the isotropic Gaussian kernel – so that $k^*(z) = k(0,z)$ – and obtain a specific result in this case.

$$
\begin{aligned}
|B_1| &= \frac{\beta/\delta}{y_1} \int \left\{1 - e^{-\delta(t_2-t_1)} + e^{-\delta(t_2-t_1)} \mathbb{1}_{\{\Delta'\Lambda z < 1/2\Delta'\Lambda\Delta - \log(y_1/y_2)\}}\right\} k(0,z)\, dz\, \pi(dv, d\Lambda) \\
&= \frac{\beta/\delta}{y_1} \int \left\{1 - e^{-\delta(t_2-t_1)} \mathbb{1}_{\{\Delta'\Lambda z \geq 1/2\Delta'\Lambda\Delta - \log(y_1/y_2)\}}\right\} k(0,z)\, dz\, \pi(dv, d\Lambda) \\
&= \frac{\beta/\delta}{y_1} \left\{1 - e^{-\delta(t_2-t_1)} \int \mathbb{P}[\Delta'\Lambda Z \geq 1/2\Delta'\Lambda\Delta - \log(y_1/y_2)] \pi(dv, d\Lambda)\right\}.
\end{aligned}
$$

where $Z \sim k(0,z)dz$. But $\Delta'\Lambda Z \sim N(0, \sigma^2)$ for $\sigma^2 := \Delta'\Lambda\Delta$, so

$$
|B_1| = \frac{\beta/\delta}{y_1} \left\{1 - e^{-\delta(t_2-t_1)} \int \Phi\left(-\frac{\sigma}{2} + \frac{1}{\sigma} \log \frac{y_1}{y_2}\right) \pi(dv, d\Lambda)\right\}.
$$

Therefore, F for the isotropic Gaussian kernel is given by

$$
F(\Lambda, \Delta(v); y_1, y_2) = \Phi\left(-\frac{\sigma(\Delta(v), \Lambda)}{2} + \frac{1}{\sigma(\Delta(v), \Lambda)} \log \frac{y_1}{y_2}\right). \tag{D.5}
$$

Now, by an obvious symmetry argument, we obtain

$$
|B_2| = \frac{\beta/\delta}{y_2} \left[1 - e^{-\delta(t_2-t_1)} \int F(\Lambda, \bar{\Delta}(v); y_2, y_1)\, \pi(dv, d\Lambda)\right]
$$

where

$$
\bar{\Delta}(v) := \Delta(x_2, x_1, t_2, t_1, v) = x_1 - x_2 - (t_1 - t_2)v \tag{D.6}
$$

Finally, $\mathbb{P}[Y(x_1,t_1) \leq y_1,\ Y(x_2,t_2) \leq y_2] = \exp(-|B|)$ with

$$
\begin{aligned}
|B| = &\frac{\beta/\delta}{y_1} \left[1 - e^{-\delta|t_2-t_1|} \int F(\Lambda, \Delta(v); y_1, y_2)\, \pi(dv, d\Lambda)\right] \\
&+ \frac{\beta/\delta}{y_2} \left[1 - e^{-\delta|t_2-t_1|} \int F(\Lambda, \bar{\Delta}(v); y_2, y_1)\, \pi(dv, d\Lambda)\right].
\end{aligned}
$$

In the Gaussian model this is

$$
|B| = \frac{\beta/\delta}{y_1} \left\{1 - e^{-\delta|t_2-t_1|} \Phi\Big(-\frac{\sigma(\Delta(v), \Lambda)}{2} + \frac{1}{\sigma(\Delta(v), \Lambda)} \log \frac{y_1}{y_2}\Big)\right\}
$$

$$+\frac{\beta/\delta}{y_2}\left\{1-e^{-\delta|t_2-t_1|}\Phi\Big(-\frac{\sigma(\bar{\Delta}(v),\Lambda)}{2}+\frac{1}{\sigma(\bar{\Delta}(v),\Lambda)}\log\frac{y_2}{y_1}\Big)\right\}$$

where

$$\sigma^2(\Delta(v),\Lambda)=\Delta(v)'\Lambda\Delta(v).$$

The equation above reduces to the unnumbered displayed equation on page 9 in Smith (1990), with σ^2 equal to a^2 in Equation (3.2) of Smith (1990), when $t_1=t_2$ and $\beta=\delta$.

D.2.3 Proof of Theorem 5.2.5

Recall that the survival function for the marginal tail waiting time is $\mathbb{P}[\kappa(y)>t]=e^{-\beta(\delta^{-1}+t+f(t))/y}$. So $\kappa(y)=0$ with probability $1-e^{-\beta/(y\delta)}$ and otherwise has survival function $e^{-\beta(t+f(t))/y}$, which is stochastically dominated by an exponential distribution with rate β/y. Under independence, we can calculate the distribution of the waiting time between exceedances $\kappa_{i_1i_2}(y)=|\kappa_{i_1}(y)-\kappa_{i_2}(y)|$, specifically

$$\begin{aligned}
\mathbb{P}[|\kappa_1(y)-\kappa_2(y)|=0] =&\mathbb{P}[\kappa_1=0,\kappa_2=0]\\
\mathbb{P}[|\kappa_1(y)-\kappa_2(y)|>t] =&\mathbb{P}[\kappa_2>t\mid\kappa_1=0,\kappa_2>0]\mathbb{P}[\kappa_1=0,\kappa_2>0]\\
&+\mathbb{P}[\kappa_1>t\mid\kappa_2=0,\kappa_1>0]\mathbb{P}[\kappa_2=0,\kappa_1>0]\\
&+\mathbb{P}[|\kappa_1-\kappa_2|>t\mid\kappa_1>0,\kappa_2>0]\mathbb{P}[\kappa_1>0,\kappa_2>0]
\end{aligned}$$

First we perform the calculation for $v=0$

$$\begin{aligned}
\mathbb{P}[\kappa_1=0,\kappa_2=0] &= (1-e^{-\beta/(y_1\delta)})(1-e^{-\beta/(y_2\delta)})\\
\mathbb{P}[\kappa_2>t\mid\kappa_1=0,\kappa_2>0]\mathbb{P}[\kappa_1=0,\kappa_2>0] &= e^{-\beta t/y_2}(1-e^{-\beta/(y_1\delta)})e^{-\beta/(y_2\delta)}\\
\mathbb{P}[\kappa_2>t\mid\kappa_1>0,\kappa_2=0]\mathbb{P}[\kappa_1>0,\kappa_2=0] &= e^{-\beta t/y_1}(1-e^{-\beta/(y_2\delta)})e^{-\beta/(y_1\delta)}\\
\mathbb{P}[|\kappa_1-\kappa_2|>t\mid\kappa_1>0,\kappa_2>0]\mathbb{P}[\kappa_1>0,\kappa_2>0] &= \left[e^{-\beta t/y_1}\frac{y_2}{y_1+y_2}+e^{-\beta t/y_2}\frac{y_1}{y_1+y_2}\right]\\
&\quad\times e^{-\beta/(y_1\delta)}e^{-\beta/(y_2\delta)}.
\end{aligned}$$

Set $\lambda_1 = e^{-\beta/(y_1\delta)}$, $\lambda_2 = e^{-\beta/(y_2\delta)}$, and get

$$\mathbb{P}[\kappa_1 = 0, \kappa_2 = 0] = (1-\lambda_1)(1-\lambda_2)$$

$$\mathbb{P}[|\kappa_{i_1}(y) - \kappa_{i_2}(y)| > t] = e^{-\beta t/y_2}\left(\lambda_2(1-\lambda_1) + \frac{y_1\lambda_1\lambda_2}{y_1+y_2}\right) + e^{-\beta t/y_1}\left(\lambda_1(1-\lambda_2) + \frac{y_2\lambda_1\lambda_2}{y_1+y_2}\right),$$

a mixture of an atom at zero with mass $(1-\lambda_1)(1-\lambda_2)$ and exponential distributions with rates β/y_1 and β/y_2.

Now, when $\pi(dv) \neq \delta_0$, we get

$$\begin{aligned}
\mathbb{P}[\kappa_1 = 0, \kappa_2 = 0] &= (1 - e^{-\beta/(y_1\delta)})(1 - e^{-\beta/(y_2\delta)}) \\
\mathbb{P}[\kappa_2 > t, \kappa_1 = 0, \kappa_2 > 0] &= e^{-\beta(t+f_2(t))/y_2}(1 - e^{-\beta/(y_1\delta)})e^{-\beta/(y_2\delta)} \\
\mathbb{P}[\kappa_2 > t, \kappa_1 > 0, \kappa_2 = 0] &= e^{-\beta(t+f_1(t))/y_1}(1 - e^{-\beta/(y_2\delta)})e^{-\beta/(y_1\delta)} \\
\mathbb{P}[|\kappa_1 - \kappa_2| > t, \kappa_1 > 0, \kappa_2 > 0] &= \Bigg[1 - \Bigg(\int_{\kappa_2=t}^{\infty} (e^{-\beta(\kappa_2 - t + f_1(\kappa_2 - t))/y_1} \\
&\quad - e^{-\beta(\kappa_2 + t + f_1(\kappa_2 + t))/y_1})\mu_2(d\kappa_2) \\
&\quad + \int_{\kappa_2=0}^{t} (1 - e^{-\beta(\kappa_2 + t + f_1(\kappa_2 + t))/y_1})\mu_2(d\kappa_2)\Bigg)\Bigg] \\
&\quad \times e^{-\beta/(y_1\delta)}e^{-\beta/(y_2\delta)}.
\end{aligned}$$

This is stochastically bounded above by the distribution when $v = 0$ and, by Lemma D.1.1, below by a mixture of exponentials with the same weights and rates $\frac{\beta}{y_j}\left(1 + \mathbb{E}_{\pi(dv,d\Lambda)}\left[\frac{c_1(\Lambda)}{c_2(\Lambda)}(1+|v|)\right]\right)$.

D.2.4 Proof of Theorem 5.2.6

We first calculate the measures of three sets:

1. A_0, the set where one or more support points is alive between time 0 and time t that causes an exceedance of y_1 at x_1 and y_2 at x_2. This exceedance must occur at birth time. In this case, the waiting time between exceedances is zero.

2. A_1, the set where one or more support points is alive between time 0 and time t that causes an exceedance of y_1 at x_1 but no support points are alive during this time that cause an exceedance of y_2 at x_2.

3. A_2, the set where one or more support points is alive between time 0 and time t that causes an exceedance of y_2 at x_2 but no support points are alive during this time that cause an exceedance of y_1 at x_1.

These calculations will then be used to derive the waiting time distribution with zero velocity. The symbol $\omega = (u, \xi, a)$ will be used to denote a generic support point.

First calculate $|A_0|$, which is

$$A_0 = \{\exists\ \omega_j : u_j k(x_1, \xi_j) > y_1, u_j k(x_2, \xi_j) > y_2, -\tau \leq s_j \leq t\}.$$

The condition on u_j is $u_j > \frac{y_1}{k(x_1,\xi_j)} \vee \frac{y_2}{k(x_2,\xi_j)}$. If $\frac{y_1}{k(x_1,\xi_j)} > \frac{y_2}{k(x_2,\xi_j)}$ then $\frac{k(x_1,\xi_j)}{k(x_2,\xi_j)} < \frac{y_1}{y_2}$, and interchange the indices for the other case. So we can calculate $|A_0|$ as $|B_1|+|B_2|$ with

$$A_{01} = \left\{\exists\ \omega_j : u_j k(x_1, \xi_j) \mathbb{1}_{\{k(x_1,\xi_j)/k(x_2,\xi_j) < y_1/y_2\}} \mathbb{1}_{\{-\tau \leq s_j \leq t\}} > y_1\right\}$$
$$A_{02} = \left\{\exists\ \omega_j : u_j k(x_2, \xi_j) \mathbb{1}_{\{k(x_2,\xi_j)/k(x_1,\xi_j) < y_2/y_1\}} \mathbb{1}_{\{-\tau \leq s_j \leq t\}} > y_2\right\}.$$

Calculate $|A_{01}|$ first as

$$\begin{aligned}
|A_{01}| &= \int_{\Lambda\in\mathcal{L}} \int_{\xi\in\mathbb{R}^d} \int_{u>\frac{y_1}{k(x_1,\xi)}} \int_{\tau=0}^{\infty} \int_{s=-\tau}^{t} \beta\delta e^{-\delta\tau} \mathbb{1}_{\{\frac{k(x_1,\xi_j)}{k(x_2,\xi_j)} < \frac{y_1}{y_2}\}} \beta u^{-2} ds\, d\tau\, du\, d\xi\, \pi(d\Lambda) \\
&= \int_{\Lambda} \int_{\xi\in\mathbb{R}^d} \int_{u>\frac{y_1}{k(x_1,\xi)}} \left[\frac{1}{\delta} + t\right] \beta u^{-2} \mathbb{1}_{\{\frac{k(x_1,\xi_j)}{k(x_2,\xi_j)} < \frac{y_1}{y_2}\}} du\, d\xi\, \pi(d\Lambda) \\
&= \int_{\Lambda} \int_{\xi\in\mathbb{R}^d} \left[\frac{1}{\delta} + t\right] \beta \frac{k(x_1,\xi)}{y_1} \mathbb{1}_{\{\frac{k(x_1,\xi_j)}{k(x_2,\xi_j)} < \frac{y_1}{y_2}\}} d\xi\, \pi(d\Lambda).
\end{aligned}$$

Now changing variables to $z = \xi - x_1$, so $\xi - x_2 = z - \Delta(0)$ for $\Delta(0)$ as in (D.3) we obtain

$$|A_{01}| = \int_\Lambda \int_{z\in\mathbb{R}^d} \left[\frac{1}{\delta} + t\right] \beta \frac{k(0,z)}{y_1} \mathbb{1}_{\{\frac{k(0,z)}{k(0,z-\Delta(0))} < \frac{y_1}{y_2}\}} dz\, \pi(d\Lambda)$$

$$= \left[\frac{1}{\delta} + t\right] \frac{\beta}{y_1} \int_\Lambda F(\Lambda, \Delta(0); y_1, y_2)\, \pi(d\Lambda)$$

for F as in (D.4).

In the isotropic Gaussian case, we obtain

$$|A_{01}| = \left[\frac{1}{\delta} + t\right] \frac{\beta}{y_1} \int_{\Lambda\in\mathcal{L}} \Phi\left(-\frac{\sigma(\Delta(0),\Lambda)}{2} + \frac{1}{\sigma(\Delta(0),\Lambda)} \log\frac{y_1}{y_2}\right) \pi(d\Lambda).$$

A symmetry argument gives

$$|A_{02}| = \left[\frac{1}{\delta} + t\right] \frac{\beta}{y_2} \int_\Lambda F(\Lambda, \bar{\Delta}(0); y_2, y_1)\, \pi(d\Lambda)$$

with $\bar{\Delta}$ as in (D.6). In the specific case of the isotropic Gaussian kernel

$$|A_{02}| = \left[\frac{1}{\delta} + t\right] \frac{\beta}{y_2} \int_{\Lambda\in\mathcal{L}} \Phi\left(-\frac{\sigma(\bar{\Delta}(0),\Lambda)}{2} + \frac{1}{\sigma(\bar{\Delta}(0),\Lambda)} \log\frac{y_2}{y_1}\right) \pi(d\Lambda).$$

Then we obtain the probability of a contemporaneous exceedance in the interval $[0, t]$ as $\exp(-|A_0|)$ with

$$|A_0| = \beta \left[\frac{1}{\delta} + t\right] \left[\frac{1}{y_1} \int_\Lambda F(\Lambda, \Delta(0); y_1, y_2)\, \pi(d\Lambda) + \frac{1}{y_2} \int_\Lambda F(\Lambda, \bar{\Delta}(0); y_2, y_1)\, \pi(d\Lambda)\right].$$

Now, the set A_2, given by

$$A_2 = \{\omega : \max_j u_j k(x_1, \xi_j) < y_1, \max_j u_j k(x_2, \xi_j) > y_2, -\tau \le s_j \le t\}$$

$$= \left\{\omega : \frac{y_2}{k(x_2,\xi)} < u < \frac{y_1}{k(x_1,\xi)}, \frac{k(x_1,\xi)}{k(x_2,\xi)} < \frac{y_1}{y_2}, \tau \le s \le t\right\}$$

is the set where one or more exceedances at x_2 occur by time t without any exceedances at x_1.

$$|A_2| = \int_{\Lambda\in\mathcal{L}} \int_{\mathbb{R}^d} \int_{u=\frac{y_2}{k(x_2,\xi)}}^{\frac{y_1}{k(x_1,\xi)}} \int_{\tau=0}^{\infty} \int_{s=-\tau}^{t} \beta u^{-2} \delta e^{-\delta\tau} \mathbb{1}_{\{\frac{k(x_1,\xi)}{k(x_2,\xi)} < \frac{y_1}{y_2}\}} ds\, d\tau\, du\, d\xi\, \pi(d\Lambda)$$

$$= \left[\frac{1}{\delta} + t\right] \beta \int_{\Lambda \in \mathcal{L}} \int_{\mathbb{R}^d} \left(\frac{k(x_2, \xi)}{y_2} - \frac{k(x_1, \xi)}{y_1} \right) \mathbb{1}_{\{\frac{k(x_1,\xi)}{k(x_2,\xi)} < \frac{y_1}{y_2}\}} d\xi \pi(d\Lambda)$$

$$= \left[\frac{1}{\delta} + t\right] \beta \int_{\Lambda \in \mathcal{L}} \left[\int_{\mathbb{R}^d} \frac{k(x_2, \xi)}{y_2} \mathbb{1}_{\{\frac{k(x_1,\xi)}{k(x_2,\xi)} < \frac{y_1}{y_2}\}} d\xi - \int_{\mathbb{R}^d} \frac{k(x_1, \xi)}{y_1} \mathbb{1}_{\{\frac{k(x_1,\xi)}{k(x_2,\xi)} < \frac{y_1}{y_2}\}} d\xi \right] \pi(d\Lambda)$$

$$= \left[\frac{1}{\delta} + t\right] \beta \int_{\Lambda \in \mathcal{L}} \left[\int_{\mathbb{R}^d} \frac{k(x_2, \xi)}{y_2} \mathbb{1}_{\{\frac{k(x_1,\xi)}{k(x_2,\xi)} < \frac{y_1}{y_2}\}} d\xi - \frac{1}{y_1} F(\Lambda, \Delta(0); y_1, y_2) \right] \pi(d\Lambda).$$

where in the last step we used (D.4). Now, focus on the inner integral with respect to ξ and make the substitution $z = \xi - x_2$ so that $\xi - x_1 = z - \bar{\Delta}$. Then we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{k(x_2, \xi)}{y_2} \mathbb{1}_{\{\frac{k(x_1,\xi)}{k(x_2,\xi)} < \frac{y_1}{y_2}\}} d\xi &= \frac{1}{y_2} \int_{\mathbb{R}^d} k(0, z) \mathbb{1}_{\{\frac{k(0,z-\bar{\Delta})}{k(0,z)} < \frac{y_1}{y_2}\}} dz \\
&= \frac{1}{y_2} \int_{\mathbb{R}^d} k(0, z) \left(1 - \mathbb{1}_{\{\frac{k(0,z)}{k(0,z-\Delta)} < \frac{y_2}{y_1}\}} \right) dz \\
&= \frac{1}{y_2} [1 - F(\Lambda, \bar{\Delta}(0); y_2, y_1)].
\end{aligned}$$

So this gives

$$|A_2| = \left[\frac{1}{\delta} + t\right] \beta \int_{\Lambda \in \mathcal{L}} \left[\frac{1}{y_2} [1 - F(\Lambda, \bar{\Delta}(0); y_2, y_1)] - \frac{1}{y_1} F(\Lambda, \Delta(0); y_1, y_2) \right] \pi(d\Lambda),$$

with $\mathbb{P}[\mathcal{N}(A_2) = 0] = \exp(-|A_2|)$. For the isotropic Gaussian kernel we get

$$\begin{aligned}
|A_2| = \left[\frac{1}{\delta} + t\right] \beta \Bigg[&\frac{1}{y_2} \Phi \left(\frac{\sigma(\bar{\Delta}(0), \Lambda)}{2} + \frac{1}{\sigma(\bar{\Delta}(0), \Lambda)} \log \frac{y_1}{y_2} \right) \\
&- \frac{1}{y_1} \Phi \left(-\frac{\sigma(\Delta(0), \Lambda)}{2} + \frac{1}{\sigma(\Delta(0), \Lambda)} \log \frac{y_1}{y_2} \right) \Bigg].
\end{aligned}$$

Note that in the Gaussian case, $\sigma(\bar{\Delta}(0), \Lambda) = \sigma(\Delta(0), \Lambda)$, since $(x_1 - x_2)' \Lambda (x_1 - x_2) = (x_2 - x_1)' \Lambda (x_2 - x_1)$.

A symmetry argument shows that

$$|A_1| = \left[\frac{1}{\delta} + t\right] \beta \int_{\Lambda \in \mathcal{L}} \left[\frac{1}{y_1} [1 - F(\Lambda, \Delta(0); y_1, y_2)] - \frac{1}{y_2} F(\Lambda, \bar{\Delta}(0); y_2, y_1) \right] \pi(d\Lambda),$$

In the case where $y_1 = y_2 = y$, all of the expressions for the Gaussian case simplify considerably, resulting in

$$|A_0^*| = \left[\frac{1}{\delta} + t\right] \frac{2\beta}{y} \int_\Lambda \Phi\left(-\frac{\sigma(\Delta(0), \Lambda)}{2}\right) \pi(d\Lambda)$$

$$|A_1^*| = |A_2^*| = \left[\frac{1}{\delta} + t\right] \frac{2\beta}{y} \int_\Lambda \left[\Phi\left(\frac{\sigma(\Delta(0), \Lambda)}{2}\right) - \frac{1}{2}\right] \pi(d\Lambda)$$

Now, set

$\kappa_1^* = \kappa_{x_1}^*(y_1) =$ time until first exceedance of y_1 at x_1

$\kappa_2^* = \kappa_{x_2}^*(y_2) =$ time until first exceedance of y_2 at x_2

$\kappa_1^* =$ time until first exceedance of y_1 at x_1 without exceeding y_2 at x_2

$\kappa_2^* =$ time until first exceedance of y_2 at x_2 without exceeding y_1 at x_1

$\kappa_0^* =$ time until first common exceedance.

and also define λ_j by

$$|A_j| = \left[t + \frac{1}{\delta}\right] \beta\lambda_j,$$

and

$$m^* \equiv \bigwedge_{j=0}^{2} \kappa_j^*.$$

Notice that

$$\lambda_0 + \lambda_1 = \frac{1}{y_1}$$

$$\lambda_0 + \lambda_2 = \frac{1}{y_2}$$

Now calculate the waiting times. There are eight cases relating to which of $\kappa_0^*, \kappa_1^*, \kappa_2^*$ are zero, summarized – along with the resulting conditional distribution of $\kappa_{(1,2)}(y_1, y_2)$ – in the table below.

case	$\kappa_0^* = 0$	$\kappa_1^* = 0$	$\kappa_2^* = 0$	Dist'n of $\kappa_{(1,2)}(y_1, y_2)$
1	T	T	T	δ_0
2	T	T	F	δ_0
3	T	F	T	δ_0
4	T	F	F	δ_0
5	F	T	T	δ_0
6	F	F	T	$\kappa_0^* \wedge \kappa_1^*$
7	F	T	F	$\kappa_0^* \wedge \kappa_2^*$
8	F	F	F	see below

In case eight, we have three further subcases (all conditional on case 8):

1. Case 8a: If case 8 holds and $\kappa_0^* = \bigwedge_{j=0}^2 \kappa_j^* = m^*$, then $\kappa_{(1,2)}(y_1, y_2) = 0$

2. Case 8b: If case 8 holds and $\kappa_1^* = m^*$, then $\kappa_{(1,2)}(y_1, y_2)$ has the distribution of $\kappa_0^* \wedge \kappa_2^*$.

3. Case 8c: If case 8 holds and $\kappa_2^* = m^*$, then $\kappa_{(1,2)}(y_1, y_2)$ has the distribution of $\kappa_0^* \wedge \kappa_1^*$.

First, the zero probability. The first four cases in the table have probability equal to the marginal probability that $\kappa_0^* = 0$, given by

$$\mathbb{P}\left[\kappa_0^* = 0\right] = 1 - e^{-\beta\lambda_0/\delta}.$$

The fifth case has probability

$$\mathbb{P}\left[\kappa_0^* > 0, \kappa_1^* = \kappa_2^* = 0\right] = e^{-\beta\lambda_0/\delta}(1 - e^{-\beta\lambda_1/\delta})(1 - e^{-\beta\lambda_2/\delta}).$$

And case 8a has probability

$$\mathbb{P}[\kappa_0^* = m^*, m^* > 0] = \frac{\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2} e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta},$$

which together give the zero probability.

Case 6 gives

$$\begin{aligned}\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t, \kappa_0^* > 0, \kappa_1^* > 0, \kappa_2^* = 0] &= \mathbb{P}[\kappa_0^* \wedge \kappa_1^* > t] e^{-\beta(\lambda_1+\lambda_0)/\delta}(1 - e^{-\beta\lambda_2/\delta}) \\ &= e^{-t\beta(\lambda_0+\lambda_1)} e^{-\beta(\lambda_1+\lambda_0)/\delta}(1 - e^{-\beta\lambda_2/\delta}).\end{aligned}$$

Case 7 gives, by symmetry

$$\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t, \kappa_0^* > 0, \kappa_1^* > 0, \kappa_2^* = 0] = e^{-t\beta(\lambda_0+\lambda_2)} e^{-\beta(\lambda_2+\lambda_0)/\delta}(1 - e^{-\beta\lambda_1/\delta}).$$

Finally, the last two subcases of case 8 give

$$\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t, \kappa_1^* = m^*, \kappa_0^* > 0, \kappa_1^* > 0, \kappa_2^* > 0] = e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta} \frac{\lambda_1 e^{-t\beta(\lambda_0+\lambda_2)}}{\lambda_0 + \lambda_1 + \lambda_2},$$

and

$$\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t, \kappa_2^* = m^*, \kappa_0^* > 0, \kappa_1^* > 0, \kappa_2^* > 0] = e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta} \frac{\lambda_2 e^{-t\beta(\lambda_0+\lambda_1)}}{\lambda_0 + \lambda_1 + \lambda_2}.$$

So the final expression is

$$\begin{aligned}
\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) = 0] =& 1 - e^{-\beta\lambda_0/\delta} + e^{-\beta\lambda_0/\delta}(1 - e^{-\beta\lambda_1/\delta})(1 - e^{-\beta\lambda_2/\delta}) + \frac{\lambda_0 e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta}}{\lambda_0 + \lambda_1 + \lambda_2} \\
\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t] =& e^{-t\beta(\lambda_0+\lambda_1)} \left(e^{-\beta(\lambda_0+\lambda_1)/\delta}(1 - e^{-\beta\lambda_2/\delta}) + \frac{\lambda_2 e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta}}{\lambda_0 + \lambda_1 + \lambda_2} \right) \\
& + e^{-t\beta(\lambda_0+\lambda_2)} \left(e^{-\beta(\lambda_2+\lambda_0)/\delta}(1 - e^{-\beta\lambda_1/\delta}) + \frac{\lambda_1 e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta}}{\lambda_0 + \lambda_1 + \lambda_2} \right).
\end{aligned} \tag{D.7}$$

Now, since $\lambda_j = \mathcal{O}(1/y_1 + 1/y_2)$, $\exp(-c\lambda_j) \to 1$ as $y_1, y_2 \to \infty$ faster than λ_j, which converge to zero at the rate $1/y_j$. So as $y_1, y_2 \to \infty$ we have

$$\begin{aligned}
\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) = 0] \asymp& \frac{\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2} \\
\mathbb{P}[\kappa_{(1,2)}(y_1, y_2) > t] \asymp& e^{-t\beta(\lambda_0+\lambda_1)} \frac{\lambda_2}{\lambda_0 + \lambda_1 + \lambda_2} + e^{-t\beta(\lambda_0+\lambda_2)} \frac{\lambda_1}{\lambda_0 + \lambda_1 + \lambda_2}.
\end{aligned}$$

D.3 Proof of Theorem 5.2.7

We construct the waiting time distribution in the general case. As in the proof of Theorem 5.2.6, define

$$\kappa_1^* = \text{ time until first exceedance of } y_1 \text{ at } x_1 \text{ without exceeding } y_2 \text{ at } x_2$$

$\kappa_2^* =$ time until first exceedance of y_2 at x_2 without exceeding y_1 at x_1

$\kappa_0^* =$ time until first simultaneous exceedance.

Since the sets of support points leading to each of these three events are disjoint, κ_j^* are independent, though their distributions differ from the zero velocity case.

We will make use of two features of $Y(x,t)$ that allow us to use some of the calculations for Theorem 5.2.6. First, by Theorem 5.2.2, the distribution of $Y(x,0)$ does not depend on velocity, so $\mathbb{P}[\kappa_0^* = 0]$, $\mathbb{P}[\kappa_1^* = 0]$, and $\mathbb{P}[\kappa_2^* = 0]$ are the same as in the zero velocity case. Second, let $E_1(\mathcal{F}_T)$ be any event measurable with respect to $\mathcal{F}_T$, the filtration at time T, and let $E_2(\mathcal{F}_{t+T})$ be any event measurable with respect to the filtration at time $t+T$. In general we can write $E_2(\mathcal{F}_{t+T}) = E_{21}(\mathcal{F}_{t+T}) \cup E_{22}(\mathcal{F}_{t+T} \setminus \mathcal{F}_T)$ with $E_{21} \cap E_{22} = \emptyset$, i.e. we can partition the event into events that are measurable with respect to $\mathcal{F}_{t+T}$ and events that are measurable with respect to $\mathcal{F}_{t+T}$ but about which $\mathcal{F}_T$ contains no information. In the max-stable velocity process, events of the form $E_{22}(\mathcal{F}_{t+T} \setminus \mathcal{F}_T)$ relate to support points that have $s_j > T$. Moreover, this is true for any T, t, that is, events caused by support points born after time T are always in $E_{22}(\mathcal{F}_{t+T} \setminus \mathcal{F}_T)$ Since

$$\begin{aligned}
\mathbb{P}[E_2(\mathcal{F}_{t+T}) \mid E_1(\mathcal{F}_T)] &= \mathbb{P}[E_2(\mathcal{F}_{t+T}) \mid E_1(\mathcal{F}_T)] + \mathbb{P}[E_2(\mathcal{F}_{t+T} \setminus \mathcal{F}_T) \mid E_1(\mathcal{F}_T)] \\
&= \mathbb{P}[E_2(\mathcal{F}_{t+T}) \mid E_1(\mathcal{F}_T)] + \mathbb{P}[E_2(\mathcal{F}_{T+t} \setminus \mathcal{F}_T)], \\
&\geq \mathbb{P}[E_2(\mathcal{F}_{t+T} \setminus \mathcal{F}_T)],
\end{aligned}$$

we can bound the probability of events of the form $\mathbb{P}[E_2(\mathcal{F}_{t+T}) \mid E_1(\mathcal{F}_T)]$ by the marginal probability of an event in $\mathcal{F}_{t+T} \setminus \mathcal{F}_T$. The event that an exceedance is caused by a kernel born after time T for any T is always contained in $\mathcal{F}_{t+T} \setminus \mathcal{F}_T$, and in particular, the event that an exceedance is caused at birth time by a kernel born after time T is contained in $\mathcal{F}_{T+t} \setminus \mathcal{F}_T$. This is quite useful, since the time t until first exceedance of y at x at kernel birth given that first exceedance does not occur at time T has exactly the same distribution as in the zero velocity case, with survival

function $e^{-\beta t/y}$, which does not depend on T at all. Exponential distributions of this form will dominate all of the nonatomic distributions in the mixture that we derive here.

Now define some additional random variables

1. $\kappa_{12}^{(0)}$, the time until first exceedance of y_2 at x_2 given that exceedance of y_1 at x_1 but not exceedance of y_2 at x_2 occurs at time 0

2. $\kappa_{21}^{(0)}$, the time until first exceedance of y_1 at x_1 given that exceedance of y_2 at x_2 but not exceedance of y_1 at x_1 occurs at time 0

3. $\kappa_{12}^{(+)}$, the time until first exceedance of y_2 at x_2 given that $\kappa_1^* = \min_j \kappa_j^*$ and $\kappa_1^* \neq 0$

4. $\kappa_{21}^{(+)}$, the time until first exceedance of y_1 at x_1 given that $\kappa_2^* = \min_j \kappa_j^*$ and $\kappa_2^* \neq 0$.

Using the decomposition into events measurable with respect to $\mathcal{F}_{t+T}$ and events measurable with respect to $\mathcal{F}_{t+T} \setminus \mathcal{F}_T$, for any of the above four random variables we have

$$\mathbb{P}[\kappa < t] \geq \mathbb{P}[\kappa(\mathcal{F}_{t+T} \setminus \mathcal{F}_T) < t],$$

where $\kappa(\mathcal{F}_{t+T} \setminus \mathcal{F}_T)$ is the waiting time until exceedance by a support point born after time T. This will be similar in every case, and it is given by $1 - \exp(-\beta(t + g(t))/y)$ where

$$\frac{\beta g(t)}{y} = \int_{v,\Lambda} \int_{\tau \in (0,\infty)} \int_{0<s<t} \int_{\xi \in \mathbb{R}^d} \frac{\beta(k^*(x,\omega,t) - k(x,\xi))}{y} \delta e^{-\delta\tau} \mathbb{1}_{\{\frac{k^*(x,\omega,t)}{k(x,\xi)} > 1\}} \pi(d\omega),$$

a positive, monotone nondecreasing function (see the proof of Theorem 5.2.3 and Lemma D.1.1). To obtain the final expression for $\mathbb{P}[\kappa < t]$, use the fact that any waiting time in a Poisson process is $\mathbb{P}[\mathcal{N}(A) = 0] = e^{-|A|}$ for a set A, and the

calculation above lower bounds $|A|$. So we have $\mathbb{P}[\kappa > t] = e^{-\beta([t+g(t)]/y+h(t))}$. We do not derive an expression for h, but several properties are clear. h must be a nonnegative function. $h(t)$ relates to the probability of exceedance by time t by a support point that was alive when exceedance occurred at another location, so it must be monotone. However, because the support points have exponential lifetimes, $\lim_{t\to\infty} h(t) \neq \infty$, since any support points alive when the initial event occurred will die almost surely, and the probability of exceedance at x_2 in the future given exceedance at x_1 is strictly less than one unless $x_1 = x_2$. Using the bound derived in Lemma D.1.1, we have $\mathbb{P}[\kappa > t] \leq e^{-\beta([t+g(t)]/y+c)}$ for a monotone nondecreasing function g that is bounded by (D.1) for isotropic k and a constant c depending on x_1, x_2, y_1, y_2.

Now, consider the possible values of κ_j^* and the resulting distributions of $\kappa_{(1,2)}(y_1, y_2)$:

case	$\kappa_0^* = 0$	$\kappa_1^* = 0$	$\kappa_2^* = 0$	Dist'n of $\kappa_{(1,2)}(y_1, y_2)$
1	T	T/F	T/F	δ_0
2	F	T	T	δ_0
3	F	F	T	$\kappa_{21}^{(0)}$
4	F	T	F	$\kappa_{12}^{(0)}$
5	F	F	F	mixture of δ_0, $\kappa_{12}^{(+)}$, $\kappa_{21}^{(+)}$

Define λ_0, λ_1, λ_2 as in the proof of Theorem 5.2.6, and define

$$p_l = \mathbb{P}\left[\kappa_l^* = \bigwedge_j \kappa_j^* \;\middle|\; \bigwedge_j \kappa_j^* \neq 0\right],$$

and let $\mu_l(dT)$ be the distribution of k_l^* conditional on the event

$$\left\{\kappa_l^* = \bigwedge_j \kappa_j^*, \quad \bigwedge_j \kappa_j^* \neq 0\right\}.$$

Then

$$\mathbb{P}[\kappa_{(1,2)} = 0] = 1 - e^{-\beta\lambda_0/\delta} + e^{-\beta\lambda_0/\delta}(1 - e^{-\beta\lambda_1/\delta})(1 - e^{-\beta\lambda_1/\delta}) + e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta}p_0,$$

where we have used the fact that $\mathbb{P}[\kappa_j^* = 0]$ is the same as in the zero velocity case. Then we have

$$\begin{aligned}\mathbb{P}[\kappa_{(1,2)} > t] &= e^{-\beta(\lambda_0+\lambda_2)/\delta} e^{-\beta(t/y_2 + g_{12}(t,y_1,y_2) + h_{12}^{(0)}(t,y_1,y_2))} \\ &\quad + e^{-\beta(\lambda_0+\lambda_1)/\delta} e^{-\beta(t/y_1 + g_{21}(t,y_1,y_2) + h_{21}^{(0)}(t,y_1,y_2))} \\ &\quad + e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta} p_1 e^{-\beta(t/y_1 + g_{12}(t,y_1,y_2))} \left(\int_{T=0}^{\infty} e^{-h_{12}^{(+)}(t,y_1,y_2,T)} \mu_1(dT) \right) \\ &\quad + e^{-\beta(\lambda_0+\lambda_1+\lambda_2)/\delta} p_2 e^{-\beta(t/y_2 + g_{21}(t,y_1,y_2))} \left(\int_{T=0}^{\infty} e^{-h_{21}^{(+)}(t,y_1,y_2,T)} \mu_2(dT) \right)\end{aligned}$$

where $h_{jj'}^{(+)}(t, y_1, y_2, T)$ are positive and monotone nondecreasing in t for every T.

D.4 Algorithms and Computation

D.4.1 Gibbs sampler for mixture models

Consider the Bayesian model

$$\kappa \sim q_0 \delta_0 + \sum_{j=1}^{K-1} q_j \text{Exponential}(\lambda_j) \tag{D.8}$$

$$q \sim \text{Dirichlet}(\boldsymbol{\alpha}), \qquad \lambda_j \overset{iid}{\sim} \text{Gamma}\left((, a), b\right). \tag{D.9}$$

For observed waiting times κ_i for $i \in \{1, \ldots, N\}$, a Gibbs sampler for this model, including imputation of continuous waiting times from censored waiting times, cycles through the following steps

1. For each observation $i \in \{1, \ldots, N\}$, sample a latent variable $z_i \sim \text{Categorical}(\tilde{q})$, where
$$\tilde{q}_j = \begin{cases} \frac{q_j \lambda_j e^{-\lambda \tilde{\kappa}_i}}{q_0 \mathbb{1}_{\{\tilde{\kappa}_i = 0\}} + \sum_{j=1}^{K} \lambda_j e^{-\lambda_j \tilde{\kappa}_i}} & j > 0 \\ \frac{q_0 \mathbb{1}_{\{\tilde{\kappa}_i = 0\}}}{q_0 \mathbb{1}_{\{\tilde{\kappa}_i = 0\}} + \sum_{j=1}^{K} \lambda_j e^{-\lambda_j \tilde{\kappa}_i}} & j = 0 \end{cases}$$

2. Sample λ_j from
$$\lambda_j \sim \text{Gamma}\left(a + \sum_i \mathbb{1}_{\{z_i = j\}}, b + \sum_i \tilde{\kappa}_i \mathbb{1}_{\{z_i = j\}} \right).$$

Table D.1: Key indicating identity of currencies corresponding to each column of the colormap images in Figure 5.6.

column/row number	symbol	name
1	AUD	Australian Dollar
2	BEF	Belgian Franc
3	CAD	Canadian Dollar
4	FRF	French Franc
5	DEM	German Deutschmark
6	JPY	Japanese Yen
7	NLG	Dutch Guilder
8	NZD	New Zealand Dollar
9	ESP	Spanish Peseta
10	SEK	Swedish Kroner
11	CHF	Swiss Franc
12	GBP	British Pound

3. Sample q from

$$q \sim \text{Dirichlet}\left(\alpha + \sum_i \mathbb{1}_{\{z_i=1\}}, \ldots, \alpha + \sum_i \mathbb{1}_{\{z_i=K\}}\right)$$

4. Sample $w_i \sim \text{Categorical}(\tilde{q})$ where

$$\tilde{q}_j = \begin{cases} \frac{q_j(F_{\lambda_j}(\kappa_i+1)-F_{\lambda_j}(\kappa_i))}{q_0\mathbb{1}_{\{\kappa_i=0\}}+\sum_j q_j(F_{\lambda_j}(\kappa_i+1)-F_{\lambda_j}(\kappa_i))} & j>0 \\ \frac{q_0\mathbb{1}_{\{\kappa_i=0\}}}{q_0\mathbb{1}_{\{\kappa_i=0\}}+\sum_j q_j(F_{\lambda_j}(\kappa_i+1)-F_{\lambda_j}(\kappa_i))} & j=0 \end{cases}$$

then sample

$$\tilde{\kappa}_i \sim \begin{cases} \delta_0 & w_i = 1 \\ \text{Exponential}_{[\kappa_i,\kappa_i+1]}(\lambda_j) & w_i = j+1 \end{cases}$$

where $\text{Exponential}_{[l,u]}$ is an exponential distribution truncated to the interval $[l,u]$.

D.5 Supplemental Figures

Table D.2: Key indicating identity of stocks corresponding to each column of the colormap images in Figure 5.5

column/row number	symbol	name
1	axp	American Express
2	ba	Boeing
3	cat	Caterpillar
4	csco	Cisco Systems
5	cvx	Chevron
6	dd	DuPont
7	dis	Disney
8	ge	General Electric
9	gs	Goldman Sachs
10	hd	Home Depot
11	ibm	IBM
12	intc	Intel
13	jnj	Johnson & Johnson
14	jpm	J.P. Morgan Chase
15	ko	Coca-Cola
16	mcd	McDonald's
17	mmm	3M
18	mrk	Merck
19	msft	Microsoft
20	nke	Nike
21	pfe	Pfizer
22	pg	Proctor & Gamble
23	t	AT&T
24	trv	Travelers
25	unh	United Healthcare
26	utx	United Technologies
27	v	Visa
28	vz	Verizon
29	wmt	Wal-Mart
30	xom	Exxon-Mobil

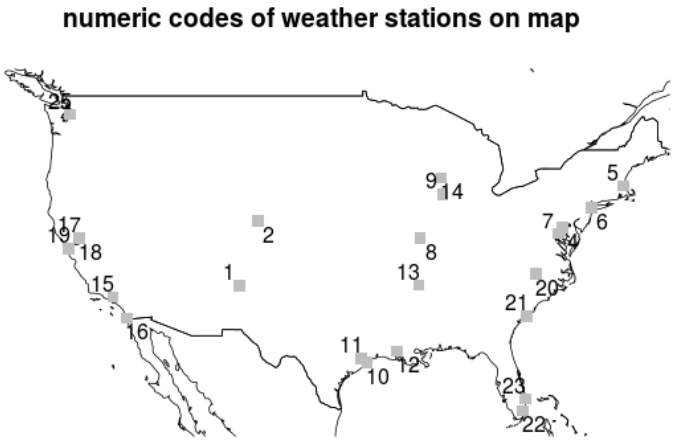


FIGURE D.1*: Map with numbers labeling locations of weather stations; the numbers correspond to the order in which the stations appear in the colormap images in Figure 5.4.*

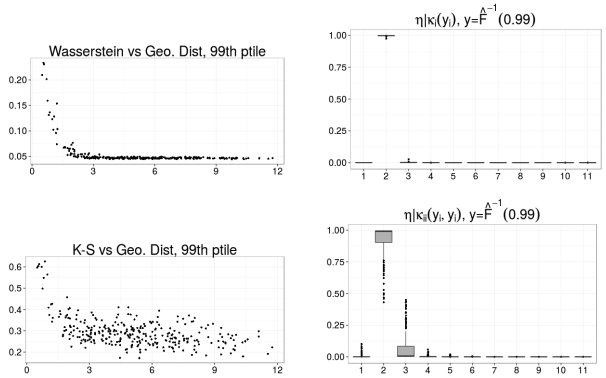


FIGURE D.2*: Left: plots of $\hat{\gamma}_d(y_1, y_2)$ for $d = W_{1,\varphi}$ and $d = TV$ for threshold $y_i = \hat{F}_i^{-1}(0.99)$ versus Euclidean distance between points. Right: Posterior estimates of $\eta_i \mid \kappa_i(y_i)$ and $\eta_{(i,i')} \mid \kappa_{(i,i')}(y_i, y_{i'})$ for $y = F^{-1}(0.99)$.*

Appendix E

Appendix to Chapter 6

E.1 Proof of Theorem 6.2.4

E.1.1 Preparatory results

The following is a standard result of the Doeblin condition in Assumption 6.2.1.

Theorem E.1.1 (Convergence under Doeblin condition)**.** *Under assumption 6.2.1, there exists a unique stationary measure* Π *for* $\mathcal{P}$*. Furthermore for any initial probability measures* ν_1, ν_2*, one has*

$$\left|\left|\nu_1 \mathcal{P}^t - \nu_2 \mathcal{P}^t\right|\right|_{\mathrm{TV}} \leq (1-\alpha)^t \left|\left|\nu_1 - \nu_2\right|\right|_{\mathrm{TV}}.$$

In particular, taking $\nu_1 = \Pi$*, we have*

$$\left|\left|\Pi - \nu_2 \mathcal{P}^t\right|\right|_{\mathrm{TV}} \leq (1-\alpha)^t \left|\left|\Pi - \nu_2\right|\right|_{\mathrm{TV}} \leq (1-\alpha)^t.$$

Proposition E.1.2. *Under Assumptions 6.2.1 and 6.2.2, any stationary measure* Π_ϵ *of* $\mathcal{P}_\epsilon$ *satisfies*

$$\left|\left|\Pi - \Pi_\epsilon\right|\right|_{\mathrm{TV}} \leq \frac{\epsilon}{\alpha}.$$

Proof.

$$\left|\left|\Pi - \Pi_\epsilon\right|\right|_{\mathrm{TV}} \leq \left|\left|\Pi\mathcal{P} - \Pi_\epsilon\mathcal{P}\right|\right|_{\mathrm{TV}} + \left|\left|\Pi_\epsilon\mathcal{P} - \Pi_\epsilon\mathcal{P}_\epsilon\right|\right|_{\mathrm{TV}} \leq (1-\alpha)\left|\left|\Pi - \Pi_\epsilon\right|\right|_{\mathrm{TV}} + \epsilon$$

The first inequality follows from the triangle inequality the second used Assumption 6.2.1 for the first term and Assumption 6.2.2 for the second term. Rearranging the resulting inequality produces the result. □

Proposition E.1.3. *Let Assumption 6.2.1 and 6.2.2 hold. For any $\epsilon \in (0, \alpha/2)$ Assumption 6.2.1 holds for the Markov operator $\mathcal{P}_\epsilon$ with the constant "α" equal to $\alpha - 2\epsilon$, which is less than 1 by construction. Hence for such ϵ the chain has a unique stationary distribution Π_ϵ to which it converges exponentially.*

Proof. We have

$$\begin{aligned}
||\mathcal{P}_\epsilon(\theta, \cdot) - \mathcal{P}_\epsilon(\theta', \cdot)||_{\mathrm{TV}} &= ||\mathcal{P}_\epsilon(\theta, \cdot) - \mathcal{P}(\theta, \cdot) + \mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta', \cdot)||_{\mathrm{TV}} \\
&\leq ||\mathcal{P}_\epsilon(\theta, \cdot) - \mathcal{P}(\theta, \cdot)||_{\mathrm{TV}} + ||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta', \cdot)||_{\mathrm{TV}} \\
&\leq \epsilon + ||\mathcal{P}(\theta, \cdot) - \mathcal{P}(\theta', \cdot) + \mathcal{P}(\theta', \cdot) - \mathcal{P}_\epsilon(\theta', \cdot)||_{\mathrm{TV}} \\
&\leq \epsilon + ||\mathcal{P}(\theta, \cdot) - \mathcal{P}(\theta', \cdot)||_{\mathrm{TV}} + ||\mathcal{P}(\theta', \cdot) - \mathcal{P}_\epsilon(\theta', \cdot)||_{\mathrm{TV}} \\
&\leq \epsilon + (1 - \alpha) + \epsilon = 1 - (\alpha - 2\epsilon)
\end{aligned}$$

□

Corollary E.1.4. *If $\mathcal{P}_\epsilon$ satisfies assumption 6.2.2 with $\epsilon < \alpha/2$, and $\mathcal{P}$ satisfies Assumption 6.2.1, then for any initial state measure ν*

$$\left|\left|\nu\mathcal{P}_\epsilon^t - \Pi\right|\right|_{\mathrm{TV}} \leq (1 - (\alpha - 2\epsilon))^t \, ||\nu - \Pi_\epsilon||_{\mathrm{TV}} + \frac{\epsilon}{\alpha}.$$

Proof. This follows by applying the triangle inequality and the results of Propositions E.1.2 and E.1.3. □

Corollary E.1.5 (Upper bounds on covariances)**.** *Suppose $\mathcal{P}$ satisfies assumption 6.2.1. Let f and g be bounded functions. Then*

$$\mathrm{cov} f(\theta_t) g(\theta_s) \leq (1 - \alpha)^{|t-s|} ||f||_* ||g||_*,$$

where $||f||_* = \inf_{c \in \mathbb{R}} ||f - c||_\infty$.

Proof. Our strategy follows some of the discussion in Yang and Dunson (2013). Suppose f satisfies $\Pi f = 0$, and $f \in L_2(\Pi)$. Define the forward operator

$$Ff(\theta) := \int f(\theta')\mathcal{P}(\theta, \theta')d\theta' = \mathbb{E}\left[f(\theta_1) \mid \theta_0 = \theta'\right].$$

From Lemma 12.6.4 in **?**,

$$\sup_{f,g \in L_2(\Pi)} \text{corr}(f(\theta_0), g(\theta_t)) = \sup_{||f||=1, ||g||=1} \langle F^t f, g\rangle = ||F^t||, \tag{E.1}$$

where $||F^t||$ is the operator norm of F^t. Since $F^t f(\theta') = \mathbb{E}\left[f(\theta_t) \mid \theta_0 = \theta'\right]$, we have that

$$F^t f - \Pi f = \mathbb{E}\left[f(\theta_t) \mid \theta_0 = \theta'\right] - 0 \leq ||f||_\infty (1-\alpha)^t$$

by Theorem E.1.1, so $\langle F^t f, g\rangle < ||f||_\infty ||g||_\infty (1-\alpha)^t$, giving $||F^t|| \leq (1-\alpha)^t$. Now, since

$$\text{corr}(f(\theta_0), g(\theta_t)) = \text{corr}(f(\theta_0) - c, g(\theta_t) - c')$$

for any $c, c' \in \mathbb{R}$, the bound in (E.1) also holds for functions with nonzero expectation with respect to Π. Therefore

$$\sup_{f,g \in L_2(\Pi)} \text{cov}(f(\theta_0), g(\theta_t)) \leq ||f||_* ||g||_* (1-\alpha)^t.$$

Finally, since the above holds for any starting measure $\theta_0 \sim \nu$, and $\text{cov}(\theta_t, \theta_0) = \text{cov}(\theta_0, \theta_t)$, we obtain

$$\text{cov} f(\theta_t) g(\theta_s) \leq (1-\alpha)^{|t-s|} ||f||_* ||g||_*.$$

□

E.1.2 Error bounds for exact chain

We want to show upper bounds on

$$\mathbb{E}\left[\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k)\right)^2\right] \quad \text{and} \quad \left|\left|\Pi - \frac{1}{t}\sum_{k=0}^{t-1} \nu\mathcal{P}^k\right|\right|_{\text{TV}}.$$

A simple way to obtain a bound on the expected square is to proceed analogously to a bias-variance decomposition

$$\mathbb{E}\left[\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k)\right)^2\right] =$$

$$\mathbb{E}\left[\left(\Pi f + \frac{1}{t}\sum_{k=1}^{t-1}(\nu\mathcal{P}^k f - \nu\mathcal{P}^k f) + \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k)\right)^2\right] =$$

$$\mathbb{E}\left[\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}^k f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k) - \nu\mathcal{P}^k f\right)^2\right] =$$

$$\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}^k f\right)^2 + \mathbb{E}\left[\left(\frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k) - \nu\mathcal{P}^k f\right)^2\right] =$$

$$\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}^k f\right)^2 + \frac{1}{t^2}\sum_{k=0}^{t-1}\sum_{j=0}^{t-1}\mathrm{cov} f(\theta_k) f(\theta_j).$$

Now applying Corollary E.1.5 and Theorem E.1.1,

$$\begin{aligned}\mathbb{E}\left[\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k)\right)^2\right] &\leq 4||f||_*^2\left(\left\|\Pi - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}^k\right\|_{\mathrm{TV}}\right)^2 \\ &\quad + \frac{||f||_*^2}{t^2}\sum_{k=0}^{t-1}\sum_{j=0}^{t-1}(1-\alpha)^{|j-k|} \\ &\leq 4||f||_*^2\left(\frac{\left(1-(1-\alpha)^t\right)||\Pi - \nu||_{\mathrm{TV}}}{\alpha t}\right)^2 \\ &\quad + \frac{||f||_*^2}{t^2}\sum_{k=0}^{t-1}\sum_{j=0}^{t-1}(1-\alpha)^{|j-k|}.\end{aligned}$$

Concentrating on the second term, we have

$$\frac{||f||_*^2}{t^2}\sum_{k=0}^{t-1}\sum_{j=0}^{t-1}(1-\alpha)^{|j-k|} = \frac{||f||_*^2}{t^2}\sum_{k=0}^{t-1}\left(\sum_{j=0}^{k}(1-\alpha)^{k-j} + \sum_{j=k+1}^{t-1}(1-\alpha)^{j-k}\right)$$

$$= \frac{||f||_*^2}{\alpha t^2} \sum_{k=0}^{t-1} \left(1 - (1-\alpha)^{k+1} - (1-\alpha)^{t-k}\right)$$

$$= \frac{||f||_*^2}{t^2} \left(\frac{2t+2}{\alpha} + \frac{2(1-\alpha)^{t+1}}{\alpha^2} - t - \frac{2}{\alpha^2}\right)$$

$$= ||f||_*^2 \left(\frac{2}{\alpha t} + \frac{2}{\alpha t^2} + \frac{2(1-\alpha)^{t+1}}{\alpha^2 t^2} - \frac{1}{t} - \frac{2}{\alpha^2 t^2}\right),$$

which gives the result.

To get a total variation bound, just apply Theorem E.1.1

$$\left|\left|\Pi - \frac{1}{t}\sum_{k=0}^{t-1} \nu\mathcal{P}^k\right|\right|_{\text{TV}} = \frac{1}{t}\sum_{k=0}^{t-1} \left|\left|\Pi - \nu\mathcal{P}^k\right|\right|_{\text{TV}}$$

$$\leq \frac{1}{t}\sum_{k=0}^{t-1} (1-\alpha)^k \left|\left|\Pi - \nu\right|\right|_{\text{TV}} = \frac{\left(1-(1-\alpha)^t\right)\left|\left|\Pi - \nu\right|\right|_{\text{TV}}}{\alpha t}.$$

E.1.3 Basic closeness properties of $\mathcal{P}_\epsilon$

Here we follow a similar approach as with the exact chain, except an additional asymptotic bias term will appear. First the total variation result

$$\mathbb{E}\left[\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k^\epsilon)\right] = \mathbb{E}\Bigg[(\Pi - \Pi_\epsilon)f + \frac{1}{t}\sum_{k=0}^{t-1}(\Pi_\epsilon - \nu\mathcal{P}_\epsilon^k)f$$

$$- \frac{1}{t}\sum_{k=0}^{t-1}(f(\theta_k^\epsilon) - \nu\mathcal{P}_\epsilon^k f)\Bigg]$$

$$\leq \frac{2||f||_*\epsilon}{\alpha} + \frac{2||f||_*\left(1-(1-\alpha_\epsilon)^t\right)\left|\left|\Pi_\epsilon - \nu\right|\right|_{\text{TV}}}{t\alpha_\epsilon},$$

so

$$\left|\left|\Pi - \frac{1}{t}\sum_{k=0}^{t-1} \nu\mathcal{P}_\epsilon^k\right|\right|_{\text{TV}} = \sup_{f:||f||_\infty \leq 1} \frac{1}{2}\mathbb{E}\left[\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k^\epsilon)\right]$$

$$\leq \frac{\epsilon}{\alpha} + \frac{\left(1-(1-\alpha_\epsilon)^t\right)\left|\left|\Pi_\epsilon - \nu\right|\right|_{\text{TV}}}{t\alpha_\epsilon},$$

since $||f||_\infty \leq 1$ implies $||f||_* \leq 1$.

Now we use it to get the L_2 bound

$$\begin{aligned}
\mathbb{E}\left[\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k^\epsilon)\right)^2\right] &= \mathbb{E}\left[\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} \nu\mathcal{P}_\epsilon^k f + \frac{1}{t}\sum_{k=0}^{t-1} \nu\mathcal{P}_\epsilon^k f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k^\epsilon)\right)^2\right] \\
&= \left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} \nu\mathcal{P}_\epsilon^k f\right)^2 + \mathbb{E}\left[\left(\frac{1}{t}\sum_{k=0}^{t-1}(f(\theta_k^\epsilon) - \nu\mathcal{P}_\epsilon^k f)\right)^2\right] \\
&\leq 4||f||_*^2\left(\frac{\epsilon}{\alpha} + \frac{(1-(1-\alpha_\epsilon)^t)\,||\Pi_\epsilon - \nu||_{\mathrm{TV}}}{t\alpha_\epsilon}\right)^2 \\
&+ \frac{\sum_{k=0}^{t-1}\sum_{j=0}^{t-1}\mathrm{cov} f(\theta_k^\epsilon) f(\theta_j^\epsilon)}{t^2} \\
&\leq 4||f||_*^2\left(\frac{\epsilon}{\alpha} + \frac{(1-(1-\alpha_\epsilon)^t)\,||\Pi_\epsilon - \nu||_{\mathrm{TV}}}{t\alpha_\epsilon}\right)^2 \\
&+ ||f||_*^2\left(\frac{2}{\alpha_\epsilon t} + \frac{2}{\alpha_\epsilon t^2} + \frac{2(1-\alpha_\epsilon)^{t+1}}{\alpha_\epsilon^2 t^2} - \frac{1}{t} - \frac{2}{\alpha_\epsilon^2 t^2}\right)
\end{aligned}$$

where $\alpha_\epsilon = \alpha - 2\epsilon$.

E.2 Proof of Remark 6.2.1

Now we show that the total variation bound for the exact chain is tight by exhibiting a Markov chain satisfying the assumptions that achieves the bound. Let

$$\mathcal{P} = \begin{pmatrix} 1-a & a \\ a & 1-a \end{pmatrix}$$

for $a \leq 1/2$. It is easy to verify by direct calculation that the invariant measure is $\Pi = (1/2, 1/2)$ and P satisfies the Doeblin condition with $\alpha = 2a$. $\mathcal{P}$ has eigenvectors

$$\phi_1 = (1/2, 1/2), \quad \phi_2 = (-1/2, 1/2)$$

with eigenvalues 1 and $1 - 2a$, respectively. Any possible starting measure ν can be expressed as $\nu_\gamma = (\gamma, 1-\gamma)$ for some $\gamma \leq 1/2$ (if $\gamma > 1/2$, just switch the definitions

of the two states). Then $||\nu_\gamma - \Pi||_{\mathrm{TV}} = \frac{1}{2}\left(|1/2-\gamma|+|1/2-(1-\gamma)|\right) = \frac{1}{2}-\gamma$ when $\gamma < 1/2$. This can be expressed in terms of the eigenvectors as

$$(\gamma, 1-\gamma) = (1/2, 1/2) + (1-2\gamma)(-1/2, 1/2).$$

So then

$$\begin{aligned}\nu_\gamma \mathcal{P}^k &= (1/2,1/2) + (1-2\gamma)(1-2a)^k(-1/2,1/2) \\ &= (1/2,1/2) + ||\nu_\gamma - \Pi||_{\mathrm{TV}}\,(1-\alpha)^k(-1/2,1/2)\end{aligned}$$

and therefore

$$\frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}^k = (1/2,1/2) + (1-2\gamma)\frac{1-(1-2a)^t}{2at}(-1/2,1/2).$$

It follows that

$$\begin{aligned}\left\|\Pi - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}^k\right\|_{\mathrm{TV}} &= \frac{1}{2}\left(\left|\frac{1}{2} - \left(\frac{1}{2} - \frac{1}{2}2\,||\nu_\gamma - \Pi||_{\mathrm{TV}}\,\frac{1-(1-\alpha)^t}{\alpha t}\right)\right|\right. \\ &\qquad \left. + \left|\frac{1}{2} - \left(\frac{1}{2} + \frac{1}{2}2\,||\nu_\gamma - \Pi||_{\mathrm{TV}}\,\frac{1-(1-\alpha)^t}{\alpha t}\right)\right|\right) \\ &= \frac{(1-(1-\alpha)^t)\,||\nu_\gamma - \Pi||_{\mathrm{TV}}}{\alpha t},\end{aligned}$$

as required.

Now, the perturbation

$$\mathcal{P}_\epsilon = \begin{pmatrix} 1-(a-\epsilon) & a-\epsilon \\ a+\epsilon & 1-(a+\epsilon)\end{pmatrix},$$

satisfies $\sup_{\theta\in\Theta}||\mathcal{P}_\epsilon(\theta,\cdot) - \mathcal{P}(\theta,\cdot)||_{\mathrm{TV}} = \epsilon$. For $\epsilon < \alpha/2$, $\mathcal{P}_\epsilon$ satisfies the Doeblin condition with $\alpha_\epsilon = 2a = \alpha$, and has invariant measure $\Pi_\epsilon = \left(\frac{a+\epsilon}{2a}, \frac{a-\epsilon}{2a}\right)$. Therefore, we have

$$||\Pi - \Pi_\epsilon||_{\mathrm{TV}} = \frac{1}{2}\left(\left|\frac{1}{2} - \frac{a+\epsilon}{2a}\right| + \left|\frac{1}{2} - \frac{a-\epsilon}{2a}\right|\right)$$

$$= \frac{1}{2}\left(\left|\frac{2a - 2(a+\epsilon)}{2(2a)}\right| + \left|\frac{2a - 2(a-\epsilon)}{2(2a)}\right|\right)$$
$$= \frac{1}{2}\left(\frac{\epsilon}{\alpha} + \frac{\epsilon}{\alpha}\right) = \frac{\epsilon}{\alpha}$$

for this chain, showing that for every $\alpha < 1/2$ and $\epsilon < \alpha/2$, there exists a Markov chain satisfying both the Doeblin condition and uniform approximation error conditions for which $||\Pi - \Pi_\epsilon||_{\mathrm{TV}} = \frac{\epsilon}{\alpha}$.

A similar perturbation

$$\mathcal{P}_\epsilon = \begin{pmatrix} 1-(a-\epsilon) & a-\epsilon \\ a-\epsilon & 1-(a-\epsilon) \end{pmatrix},$$

can be represented as

$$\mathcal{P}_\epsilon = \begin{pmatrix} 1-a_\epsilon & a_\epsilon \\ a_\epsilon & 1-a_\epsilon \end{pmatrix},$$

for $a_\epsilon = a - \epsilon$, and has $\alpha_\epsilon = 2a - 2\epsilon = \alpha - 2\epsilon$, and invariant measure $(1/2, 1/2)$. So applying the result proved for $\mathcal{P}$, $\mathcal{P}_\epsilon$ achieves

$$\left|\left|\Pi - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}_\epsilon^k\right|\right|_{\mathrm{TV}} = \left|\left|\Pi_\epsilon - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}_\epsilon^k\right|\right|_{\mathrm{TV}} = \frac{\left(1-(1-\alpha_\epsilon)^t\right)||\nu_\gamma - \Pi||_{\mathrm{TV}}}{\alpha_\epsilon t}$$
$$= \frac{\left(1-(1-(\alpha-2\epsilon))^t\right)||\nu_\gamma - \Pi||_{\mathrm{TV}}}{(\alpha-2\epsilon)t}.$$

So there exist perturbations that achieve both of the components of the bound for $\left|\left|\Pi - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}_\epsilon^k\right|\right|_{\mathrm{TV}}$, but the perturbations exhibited differ.

Now, recall that

$$\mathbb{E}\left[\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k)\right)^2\right] = \left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1}\nu\mathcal{P}^k f\right)^2 + \frac{1}{t^2}\sum_{k=0}^{t-1}\sum_{j=0}^{t-1}\mathrm{cov} f(\theta_k) f(\theta_j).$$

For a discussion of tightness of the covariance bound $\mathrm{cov} f(\theta_0) f(\theta_t) \leq ||F^t||$ when the forward operator is compact and self-adjoint, see Yang and Dunson (2013). Now,

note that

$$\mathcal{P} = \begin{pmatrix} 1-a & a \\ a & 1-a \end{pmatrix}$$

is the transition matrix of a reversible Markov chain on a finite state space, so F is compact and self-adjoint. We showed that

$$\frac{1}{t}\sum_{k=0}^{t-1} \nu\mathcal{P}^k = (1/2, 1/2) + (1-2\gamma)\frac{1-(1-2a)^t}{2at}(-1/2, 1/2).$$

The only non-trivial functions on this state space have different values in the two states. To make $|f| \leq 1$, put $f(0) = -1$ and $f(1) = 1$. Then $\Pi f = 0$ and

$$\begin{aligned}\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} \nu\mathcal{P}^k f =& -1\left(\frac{1}{2} - \frac{1}{2}(1-2\gamma)\frac{1-(1-2a)^t}{2at}\right) \\ &+ \left(\frac{1}{2} + \frac{1}{2}(1-2\gamma)\frac{1-(1-2a)^t}{2at}\right) \\ =&(1-2\gamma)\frac{1-(1-2a)^t}{2at} = 2\,||\Pi - \nu||_{\mathrm{TV}}\,\frac{1-(1-\alpha)^t}{\alpha t}\end{aligned}$$

so

$$\left(\Pi f - \frac{1}{t}\sum_{k=0}^{t-1} \nu\mathcal{P}^k\right)^2 = 4||f||_*^2||\Pi - \nu||_{\mathrm{TV}}^2\left(\frac{1-(1-\alpha)^t}{\alpha t}\right)^2$$

E.3 Additional Computational Examples

This section applies the aMCMC algorithms considered in Section 6.3 of Chapter 6 to real and simulated data to assess their performance. For the mixture model with Gaussian approximations to full conditionals, a similar analysis to that for the logistic regression with data subsets is performed. That is, we estimate ϵ and α from sample paths and use these estimates to compare the behavior of the algorithm to the predictions in Section 6.4 of Chapter 6. Such an analysis is not possible for the Gaussian process example because the exact algorithm is numerically and

computationally infeasible due to the need to invert a very large matrix. For aMCMC for logisitic regression and Gaussian processes, we also measure performance with respect to a larger variety of discrepancy measures or loss functions as a function of computation time.

E.3.1 Details of procedures for approximating $1-\alpha$ *and* ϵ

We first provide additional details on estimation of $1-\alpha$ and ϵ. To estimate the convergence rate $1-\alpha$, we use the following approach based on sample path autocorrelations. When $\mathcal{P}$ is reversible, there exist B and V in Definition 7.1.1 in Chapter 6 such that $(1-\alpha)^k = ||F^k||$ for the forward operator F defined in (6.12) of Chapter 6. When F is compact, we also have

$$\sup_{f \in L_2(\Pi)} \text{Corr}\left(f(\theta_0), f(\theta_k)\right) = ||F^k||$$

where $L_2(\Pi)$ is the space of (Π) square-integrable functions (for a more detailed discussion of this equivalence, see section 3 of Yang and Dunson (2013)). Although not all MCMC algorithms are reversible – and in particular, Gibbs samplers often are not – in practice, a useful lower bound estimate of $1-\alpha$ can be obtained from the autocorrelations. Specifically, if $\text{Corr}\left(f(\theta_0), f(\theta_k)\right) \approx (1-\alpha)^k$, then we can estimate $1-\alpha$ using the estimator

$$\widehat{\varphi}_{\max} = \max_{j \leq p} \max_{k \leq k_{\max}} \widehat{\varphi}_{j,k}^{1/k}, \tag{E.2}$$

This estimator can be unreliable when $\widehat{\varphi}_{j,k}$ is near zero, particularly for large k. Thus, when using the maximum likelihood estimator of the sample autocorrelations to compute (E.2), we consider only the values of j for which $\widehat{\varphi}_{j,k}$ exceeds $\frac{\Phi^{-1}(0.95^{1/k_{\max}})}{\sqrt{t-k_{\max}}}$, which corresponds to a union bound multiplicity correction at the 0.95 level based on the asymptotic distribution of the MLE.

Estimates of the Wasserstein-1 distance with respect to a metric kernel K are obtained as follows. Let $K : \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ be a reproducing kernel and $\mathbb{H}$ a Hilbert

space, and define

$$d_K(\theta_1, \theta_2) := ||K(\cdot, \theta_1) - K(\cdot, \theta_2)||_{\mathbb{H}},$$

where $||f - g||_{\mathbb{H}} = \sqrt{\langle f, g \rangle}$ is the usual Hilbert space norm. For probability measures P, Q satisfying quite general conditions, Minsker et al. (2014) shows the Wasserstein-1 distance defined with respect to this metric is

$$W_{1,d_K}(P, Q) := \left|\left| \int_\Theta k(\theta, \cdot) d(P - Q)(\theta) \right|\right|_{\mathbb{H}},$$

for which there exists a simple sample estimator (see equation 2.12 in Minsker et al. (2014)). Choosing the (unnormalized) isotropic Gaussian kernel

$$K(\theta_1, \theta_2) = \frac{1}{\sigma} \exp(-\phi(\theta_1 - \theta_2)'(\theta_1 - \theta_2)),$$

results in $d_K \leq \frac{1}{\sigma}$. As default choices, we put $\phi, \sigma = 1$, giving $0 \leq d_K \leq 1$. Moreover, we have $W_{1,d_K}(P, Q) = 0$ if and only if $P = Q$. Since $d_K \leq 1$, this provides a *lower bound* on the total variation distance via $2W_{1,d_K}(P, Q) \leq ||P - Q||_{\mathrm{TV}}$.

E.3.2 Distributional approximations – Mixture model

We simulated a $1000 \times 1000 = d \times d$ contingency table with cell count $N = 10^9$ according to model (6.14a)-(6.14b), with $a_h^{(j)} = 1/d$, $\alpha = 1$, and $k = 7$, and implemented either aMCMC or exact MCMC for 10,000 iterations after a burn-in of 10,000 iterations. During the burn-in, data were gradually added to prevent the chain from becoming trapped in a local mode. We focus comparisons on the samples after burn-in.

We first analyze this example for consistency with the theory in Section 7.2 of Chapter 6. Table E.1 shows estimates of $\widehat{\varphi}_{\max}$, W_{1,d_K}, $s(\epsilon)$, and an approximation to ϵ from $\widehat{W}_{1,d_K}(\Pi_\epsilon, \Pi)(1 - \widehat{\varphi}_{\max}) \approx \alpha ||\Pi - \Pi_\epsilon||_{\mathrm{TV}}$, where $\widehat{\varphi}_{\max}$ is computed for the exact algorithm. The estimates are based on the same sample path quantities

as those in Table 6.2 in Chapter 6. Memory requirements for storing sample paths for all 10^6 entries of π were prohibitive, so all results are based on the 100 largest entries of π. The largest value of $\widehat{\varphi}_{\max}$ is 0.93, though the tendency of this algorithm to become trapped in local modes suggests that this is a substantial underestimate. The approximate values of ϵ are all very small, indicating that the approximation is highly accurate. However, it offers limited speedup, with the speedup for the least accurate approximation only 11.5. There is no evidence that the mixing and convergence properties of the exact chain are superior to those of the approximate chain. A similar result was obtained for the logistic regression algorithm in Chapter 6, suggesting that the results in Section 7.2 of Chapter 6, which incorporate the effect of possibly slower mixing and convergence of the approximation, may understate the performance of some real aMCMC algorithms.

Table E.1: Estimates of $\widehat{\varphi}_{\max}$ and $W_{1,d_K}(\Pi_\epsilon, \Pi)$ for mixture model aMCMC with different values of $n_{\min}$.

	$n_{\min} = \infty$	$n_{\min} = 1000$	$n_{\min} = 200$	$n_{\min} = 100$
$\widehat{W}_{1,d_K}(\Pi_\epsilon, \Pi)$	0	3.01×10^{-6}	3.31×10^{-6}	4.35×10^{-6}
$\widehat{\varphi}_{\max}$	0.93	0.90	0.93	0.89
$\widehat{W}_{1,d_K}(\Pi_\epsilon, \Pi)(1 - \widehat{\varphi}_{\max})$	0	1.99×10^{-7}	2.19×10^{-7}	2.88×10^{-7}
$s(\epsilon)$	1.00	10.08	11.45	11.51

Figure E.1 shows $\text{RMSE}(\pi, t_0, \epsilon)$ as defined in Chapter 6 based on different length sample paths t_0 corresponding to the computation times shown on the horizontal axis. Burn-in time is incorporated into the total computation times shown on the horizontal axis. For computation times up to around 25,000 seconds (about seven hours), the approximate algorithm with $n_{\min} = 1000$ performs better with respect to D_{L_2} than the exact algorithm. In both cases, the estimated values of $\text{RMSE}(\pi, t_0, \epsilon)$ are very small. Most of the computational benefit of aMCMC is a result of the longer computation time for the burn-in period for the exact algorithm. After burn-in, the exact algorithm becomes optimal in less than an hour of additional computation time.

However, the data in the example consist of a few cells with very large cell counts, so it is likely that much larger values of $n_{\min}$ would offer comparable speedup to the values considered, and that the corresponding algorithm would be compminimax optimal for longer computation times. The observed results are consistent with the concave speedup function and relatively small speedup provided by the approximation.

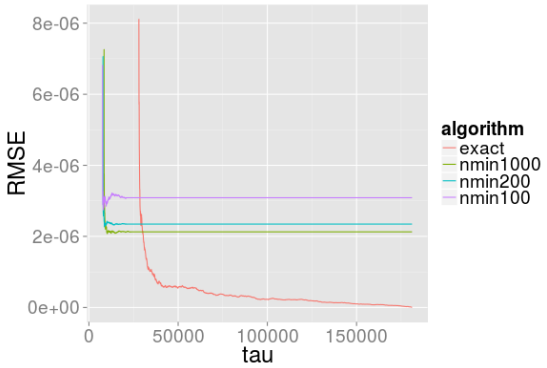


FIGURE E.1: RMSE(π, t_0, ϵ) *between ergodic average of* π *for the entire sample path from the exact algorithm and the ergodic average of* π *computed from sample paths from the exact and approximate transition kernels for different computation times* τ *with* $N = 10^9$.

E.3.3 Performance of aMCMC for different discrepancy measures

We now consider performance of the logistic regression and Gaussian process aMCMC algorithms with respect to a variety of discrepancy measures, some of which are not analyzed theoretically. The results support two general conclusions: (1) overall, the insights from theoretical analysis of compminimax optimality with respect to D_{TV} and D_{L_2} are transferrable to other discrepancy measures; and (2) there can be substantial differences in the performance of approximations for any given value of ϵ depending on the object of inference.

E.3.4 Logistic regression using subsets – additional results

We now present some additional results for logistic regression using subsets, in this case using exactly the sampler in (6.19a)-(6.19c) applied to the SUSY dataset (Baldi

et al. (2014)). The dataset consists of 5 million observations of a binary outcome with 18 continuous covariates. The data are divided into a training set consisting of 4.5 million observations and a test set of 0.5 million observations. Computation was performed for a range of seven subset sizes between $|V| = 1,000$ and $|V| = 4,500,000$. In each case, the following functionals were estimated based on the sample path of θ_ϵ^t. We do not adapt the subset sizes based on the state of the chain for this example, so it is not expected to achieve the uniform error bound.

1. The mean of the regression coefficients β, based on $\frac{1}{t}\sum_{k=0}^{t-1}\beta^k$. Root mean square error (RMSE) was used as the discrepancy measure.

2. The median of the regression coefficients, given by $m = \text{argmax}_{m^*} : \left(\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{1}_{\{\beta^k < m^*\}}\right) < 0.5$. Mean absolute error was chosen for the discrepancy measure.

3. The endpoints of 95 percent posterior credible intervals,

$$m_q = \text{argmax}_{m^*} : \left(\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{1}_{\{\beta^k < m^*\}}\right) < q$$

   for $q = 0.025, 0.975$. The discrepancy measure is RMSE. In addition, each 95 percent credible interval was classified according to whether it included zero, and the L_0 loss for this classification was calculated.

4. Prediction of the outcome y on the test set. Predictive accuracy was measured with the area under curve metric.

5. The L_1 and L_2 norms of the regression coefficients, $\sum_j |\beta_j|$ and $\sum_j \beta_j^2$, respectively.

We focus on the accuracy of estimates based on samples from the (putative) stationary distributions of approximate samplers with subsample sizes $|V| =$ 1,000,

Table E.2: Posterior discrepancy for estimation of various functionals at different values of $|V|$ for logistic regression example on SUSY data.

$\|V\|\rightarrow$	1,000	10,000	50,000	500,000	4,500,000
RMSE	0.12	0.02	0.01	0.00	0.00
RMSE (c.i.)	1.36	0.44	0.17	0.03	0.00
AUC	0.86	0.86	0.86	0.86	0.86
L1 norm beta	19.71	18.33	17.86	17.76	17.76
L2 norm beta	52.27	44.13	42.27	42.05	42.09
MAE	0.08	0.01	0.00	0.00	0.00
Mean L0 Error (c.i. cross zero)	0.53	0.16	0.11	0.05	0.00
Clock time (seconds)	24.64	78.16	333.24	3207.83	29876.50
Effective sample size	413.99	405.28	424.36	372.06	369.70

5,000, 10,000, 50,000, 100,000, 250,000, 500,000 and 4,500,000. The first 1,000 samples were discarded and the subsequent 1,000 samples used to compute ergodic averages. All discrepancy measures used an estimate based on 1,000 samples from the exact Markov chain after a 1,000 sample burn-in as the "truth." Because the exact sampler mixes rapidly and has low autocorrelation (see Chapter 6), the error in these "true" posterior estimates is expected to be small.

Table E.2 shows posterior discrepancy for the parameters described above. As expected, the discrepancy invariably decreases as $|V|$ grows, which corresponds to smaller values of ϵ. However, there are substantial differences in the rate at which the discrepancy converges to zero as ϵ decreases. For example, $|V| = 1,000$ is sufficient to obtain the best possible out of sample predictive performance measured by AUC, while even with $|V| = 500,000$, one of the 18 regression coefficients is improperly classified as having a posterior credible interval that includes zero. Similarly, RMSE for estimation of β decreases more slowly with ϵ than MAE for estimation of β.

E.3.5 Low-rank Gaussian process

Computation was performed for the low-rank Gaussian process approximations as described in section 6.3.4. Computation for the exact transition kernel is infeasible due to the need to invert a large matrix, so we focus solely on performance of the approximate algorithm in prediction for different levels of approximation error. Six

values of δ – corresponding to approximation error for Σ in the Frobenius norm of $\delta = 0.001, 0.01, 0.02, 0.03, 0.04,$ and 0.05 – were chosen to assess the computation time-approximation accuracy tradeoff. Smaller values of δ correspond to smaller values of ϵ, but because the exact algorithm is infeasible we cannot estimate the value of ϵ corresponding to each value of δ. We do not adapt δ to the state of the chain in this example. The model in (6.21) with prior in (6.22) was estimated on Sarcos robot arm data (see Vijayakumar et al. (2005)). A grid of ϕ values corresponding to correlations between 0.99 and 0.01 at the maximum pairwise distance in X was used for the prior on ϕ, and Gamma $(1, 1)$ priors chosen on τ^{-2} and σ^{-2}. The data consist of 48,933 observations on 21 continuous covariates and one continuous outcome. Of these, 4,449 observations are commonly designated the test set. We divided the dataset into ten subsets of approximately equal size and performed computation independently on each subset. The results provided here are combined over the ten independent datasets.

Table E.3 shows discrepancy measures for estimation of various functionals of y_{test}, the vector of response values in the test set. As in previous examples, this table is based on estimates obtained from the chains at putative stationarity. In particular, $t = 1,000$ samples were gathered after discarding $B = 1,000$ samples as burn-in. In summary, the discrepancy measures and estimators are:

1. RMSE (mean of y_{test}): RMSE for out of sample prediction of y calculated using $\frac{1}{t}\sum_{k=B+1}^{B+t} y_{\text{test}}^k$ as the point estimate.

2. MAE (median of y_{test}): MAE for out of sample prediction of y calculated using $m = \text{argmax}_{m^*} : \left(\frac{1}{t}\sum_{k=B+1}^{B+t} \mathbb{1}_{\{\beta^k < m^*\}}\right) < 0.5$ as the point estimate.

3. MAE |c.i. coverage - 0.95|: MAE for coverage of credible intervals (difference between empirical coverage and 0.95) for out of sample predictive intervals for

Table E.3: Summaries of results for the aMCMC algorithm in the Gaussian process regression example for varying levels of approximation accuracy of the covariance.

δ_ϵ	0.001	0.01	0.02	0.03	0.04	0.05
RMSE (mean of y_{test})	3.57	3.60	3.66	3.72	3.80	3.85
MAE (median of y_{test})	2.43	2.47	2.52	2.53	2.61	2.64
MAE |c.i. coverage - 0.95|	0.01	0.00	0.01	0.01	0.00	0.00
Effective Size per sample	1.01	1.01	1.00	1.00	1.00	1.00
Geweke test proportion	0.06	0.06	0.06	0.07	0.06	0.06
Seconds per sample	0.17	0.12	0.11	0.11	0.11	0.11

y:

$$m_q = \text{argmax}_{m^*} : \left(\frac{1}{t} \sum_{k=0}^{t-1} \mathbb{1}_{\{\beta^k < m^*\}} \right) < q$$

for $q = 0.025, 0.975$ as the point estimate for the credible intervals

Also shown are the effective sample size per iteration, the proportion of the Geweke convergence z-scores that are greater than 1.96 in magnitude, and the computational intensity in seconds per sample.

That the seconds per sample in Table E.3 increases by less than a factor of two over the entire range of δ values considered reflects our empirical finding that the spectrum of the covariance matrix Σ decays slowly. As such, increasing δ does not result in a large decrease in r. The analysis of the approximate algorithm for the Gaussian process indicated that when the spectrum decays very slowly, the speedup function is likely concave, which explains why smaller values of δ appear to give noticeable performance improvements for small computational cost.

Figure E.2 shows discrepancy measures as a function of computation time for the six different δ values. In this case, the threshold time at which the most accurate approximation is preferred is very low. The most accurate approximate chain, with $\delta = 0.001$, is optimal with respect to RMSE among the six values tested for out of sample prediction with a computational budget of 10 seconds or greater, and is optimal with respect to MAE with a budget of 15 seconds or greater. This provides

further empirical support for the hypothesis that the speedup function for this algorithm on this dataset is concave, and that the achieved speedups are small for the range of ϵ values corresponding to the different values of δ considered here.

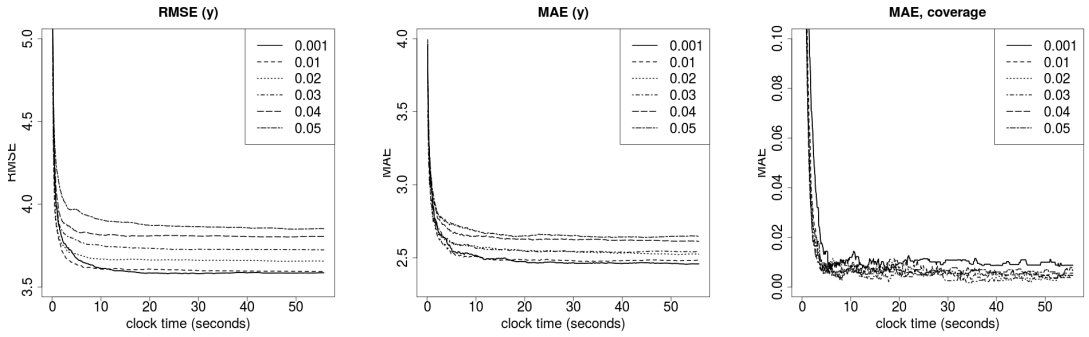


FIGURE E.2*: RMSE for estimation of y_{test} by its ergodic average (left), MAE for estimation of y_{test} by its sample path median (center), and L_1 loss for empirical coverage of 95 percent posterior credible intervals based on the empirical quantiles of the sample path for low-rank GP approximate MCMC algorithms run on Sarcos robot arm data.*

E.4 Alternative to Assumption 6.2.1

We give an alternative set of conditions that are together equivalent to Assumption 6.2.1 but are easier to verify. A classic idea in Markov chain theory is that a minorization condition on the state space, $\inf_{\theta \in \Theta} \mathcal{P}(\theta, \cdot) \geq \gamma m(\cdot)$ where $m(\cdot)$ is a probability measure, implies the Doeblin condition stated in Assumption 6.2.1 (Nummelin, 1978; Athreya and Ney, 1978; Meyn and Tweedie, 2009; Rosenthal, 1994). Here we use a slight variation on the standard minorization condition. Specifically, we divide the state space into a good set Θ_0 and a bad set Θ_0^c, with standard minorization conditions holding on the good set and a lower bound on the probability of transitioning from the bad set to the good set; these conditions are stated in Assumption E.4.1. One can show that two steps of any Markov chain satisfying Assumption E.4.1 will satisfy the standard minorization condition. This implies that satisfying Assumption E.4.1 is equivalent to satisfying Assumption 6.2.1.

Assumption E.4.1 (Minorization and return condition for exact chain)**.** *Let $\mathcal{P}$ be a Markov transition kernel with state space Θ. There exist $\Theta_0 \subset \Theta$, constants $\gamma, \beta > 0$, and a probability measure $m(\cdot)$ supported on Θ such that*

$$\inf_{\theta \in \Theta_0} \mathcal{P}(\theta, \cdot) \geq \gamma m(\cdot), \tag{E.3a}$$

$$\inf_{\theta \in \Theta_0^c} \mathcal{P}(\theta, \Theta_0) \geq \beta, \text{ and}$$

$$m(\Theta_0) > 0.$$

E.5 Proof of Remark 6.3.2

Here we show that there exists a probability measure $m(\cdot)$, a set $\Theta_0 \subset \Theta$, and constants $\gamma, \beta > 0$ such that the Gibbs sampling algorithm in Section 6.3.1 for the mixture model in (6.14a)-(6.14b) satisfies Assumption E.4.1 and hence Assumption 6.2.1.

The state space for this Gibbs sampler is given by $\Theta = \Lambda \times \mathcal{N} \times \mathcal{Z}$, where

$$\Lambda = \bigtimes_{j=1}^{p} \bigtimes_{h=1}^{K} \Delta^{(d_j - 1)}, \quad \mathcal{N} = \Delta^{(K-1)}, \quad \mathcal{Z} = \bigtimes_{\boldsymbol{c} \in \mathcal{C}^+} \mathcal{Z}_{\boldsymbol{c}}$$

$$\mathcal{Z}_{\boldsymbol{c}} = \left\{ Z(\boldsymbol{c}) \in \mathbb{N}^K : \sum_{h=1}^{K} Z(\boldsymbol{c})_h = n(\boldsymbol{c}) \right\},$$

$\mathbb{N}$ are the nonnegative integers, $n(\boldsymbol{c})$ is the observed count in cell $\boldsymbol{c}$ and $\boldsymbol{c} \in \times_{j=1}^{p} \{1, \ldots, d_j\}$. Fix $0 < \delta < 1$ and define $\Theta_0 = \Lambda_0 \times \mathcal{N}_0 \times \mathcal{Z}$, where

$$\Lambda_0 = \bigtimes_{j=1}^{p} \bigtimes_{h=1}^{K} \Delta_0^{(d_j - 1)}, \quad \Delta_0^{(d_j - 1)} = \{\lambda \in \Delta^{(d_j - 1)} : \delta < \lambda_c < 1 - \delta \ \forall \ c\},$$

$$\mathcal{N}_0 = \Delta_0^{(K-1)},$$

Δ^K is the K-dimensional unit simplex, and $\times$ represents a Cartesian product.

E.5.1 Minorization condition

First we construct a measure $m(\cdot)$ such that

$$\inf_{\theta \in \Theta_0} \mathcal{P}(\theta, \cdot) \geq \gamma m(\cdot).$$

For a function of two variables $f(x, y)$, let $f_{\inf(y)}(x) = \inf_y f(x, y)$ be the function defined by the pointwise infimum over y. Let $p(\nu \mid Z)$ and $p(\lambda \mid Z)$ be the conditional densities of ν, λ given Z in the Gibbs sampling algorithm.

It is enough to show that (1) every configuration of Z has positive probability for $\nu, \lambda \in \mathcal{N}_0 \times \Lambda_0$ and (2) the functions $p_{\inf(Z)}(\lambda)$, $p_{\inf(Z)}(\nu)$ satisfy $\int_{\nu \in \mathcal{N}} p_{\inf(Z)}(\nu) d\nu > 0$, $\int_{\lambda \in \Lambda} p_{\inf(Z)}(\lambda) > 0$.

The conditional distribution for Z given λ, ν is

$$Z(\boldsymbol{c}) \mid \nu, \lambda, Y \sim \text{Multinomial}\left(n(\boldsymbol{c}), \tilde{\nu}\right), \qquad \tilde{\nu}_h = \frac{\nu_h \prod_{j=1}^{p} \lambda_{hc_j}^{(j)}}{\sum_{l=1}^{K} \nu_l \prod_{j=1}^{p} \lambda_{lc_j}^{(j)}}$$

so that for any $\theta \in \Theta_0$, $\tilde{\nu}_h > \frac{1}{K}\left(\frac{\delta}{(1-\delta)}\right)^{p+1}$ for every $h \in \{1, \ldots, K\}$. This immediately implies that $\inf_{\lambda \in \Lambda_0, \nu \in \mathcal{N}_0} p(Z \mid \lambda, \nu) > 0$.

To show (2), note that $p(\nu \mid Z)$ and $p(\lambda_h^{(j)} \mid Z)$ are both Dirichlet densities (since $\lambda_h^{(j)}$ are conditionally independent given Z, it is enough to show (2) for an arbitrary $\lambda_h^{(j)}$). The parameter of $p(\nu \mid Z)$ is $\alpha(Z) = \alpha + \sum_{\boldsymbol{c} \in \mathcal{C}^+} Z(\boldsymbol{c})$, with density

$$p(\nu \mid Z) = \frac{1}{B(\alpha(Z))} \prod_{h=1}^{K} \nu_h^{\alpha(Z)_h - 1},$$

where $B(\alpha(Z)) = \frac{\prod_h \Gamma(\alpha(Z)_h)}{\Gamma(N + K\alpha)}$. Consider any compact subset of $\mathcal{N}$ with nonzero Lebesgue measure that has empty intersection with the boundaries of the simplex. For simplicity, we can take $\mathcal{N}_0$. Because $\mathcal{Z}$ is a finite set, and for any $Z \in \mathcal{Z}$, $\alpha(Z)_h > 1$ for all h so long as $\alpha > 0$, $\inf_{\nu \in \mathcal{N}_0} p_{\inf Z}(\nu) = \gamma^* > 0$. This is enough to give $\int_{\nu \in \mathcal{N}} p_{\inf(Z)}(\nu) d\nu > \gamma^* \text{Vol}(\mathcal{N}_0) > 0$, where $\text{Vol}(\mathcal{N}_0)$ is the Lebesgue meaure of the set $\mathcal{N}_0$. A result for $\lambda_h^{(j)}$ follows by a similar argument.

E.5.2 Return condition

Now we show that for $\theta \in \Theta_0^c$, $\inf_{\theta \in \Theta_0^c} P(\Theta|\theta) > 0$. Since $\lambda \perp\!\!\!\perp \nu \mid Z$ and $\nu \perp\!\!\!\perp \lambda \mid Z$, the return probability does not depend on (λ, ν) but only on Z. Conditional on any value of Z, $P(\Theta_0|Z)$ is strictly positive so long as the prior hyperparameters $a_h^{(j)}$ and α have strictly positive entries. Since $\mathcal{Z}$ is finite, minimize over all elements of $\mathcal{Z}$ to obtain $\beta = \bigwedge_{Z \in \mathcal{Z}} P(\Theta_0|Z) > 0$, a lower bound for $P(\Theta_0|\theta)$ that holds for any $\theta \in \Theta$, so in particular it holds for any $\theta \in \Theta_0$.

E.5.3 Nonnegativity condition

Now we just want $m(\Theta_0) > 0$, but this is easy since we showed that $p_{\inf(Z)}(\nu), p_{\inf(Z)}(\lambda)$ are bounded below on Θ_0.

E.6 Proof of Remark 6.3.1

We rely on the Berry-Esséen result in Weiss (1978). The result is given for a Multinomial (n, ν) distribution with number of classes K that may be increasing in n, but in our setting n is fixed so we state the result in this special case.

E.6.1 Result from Weiss (1978)

Suppose there exists $\delta > 0$ such that $\min_{1 \leq h \leq K}(1 - \nu_h) > \delta$. Let $W(n)$ be a random variable having distribution given by the usual normal approximation to the Multinomial, so that

$$W \sim \text{Normal}\left(n\nu, n[\text{diag}\left(\nu\right) - \nu\nu']\right),$$

and for $h = 1, \ldots, K-1$, define the random variable $\overline{W}_h$ as the closest value to W_h (in the L_1 sense) which makes $n\nu_h + \sqrt{n\nu_h}\overline{W}_h$ an integer. $\overline{W}_K(n)$ is given by the identity

$$\sum_{h=1}^{K} \sqrt{\nu_h}\overline{W}_h = 0.$$

Note that this is equivalent to rounding the entries $1, \ldots, K-1$ to the nearest integer and defining the final entry to ensure that the full vector $\overline{W}$ sums to n.

Let $\mu_{\overline{W}}(\cdot)$ be the measure on $\mathbb{Z}^K$ induced by the definition of $\overline{W}$ and let $\mu_Y(\cdot)$ be the Multinomial (n, ν) measure. The result in Weiss (1978) is

$$||\mu_{\overline{W}} - \mu_Y||_{\mathrm{TV}} \leq \frac{C(K-1)}{\sqrt{n}} \sum_{h=1}^{K-1} \frac{(1-\nu_h)(1+P_h/\nu_K)(1-2\nu_h+2\nu_h^2))}{\sqrt{\nu_h(1-\nu_h)}},$$

where $P_h = \sum_{h' \leq h} \nu_{h'}$ and $C(K-1)$ is a constant depending on $K-1$.

By constructing a result of this sort from first principles, it should be possible to obtain a bound on the magnitude of $C(K)$, as has been shown for Berry-Essén results in other settings. However, as our goal is only to show that any approximation error can be obtained with sufficiently large n, we do not pursue this here.

E.6.2 Construction of $\mathcal{P}_\epsilon$

To construct $\mathcal{P}_\epsilon$ satisfying assumption 6.2.2, let $p_\epsilon(Z(\boldsymbol{c}) \mid \nu, \lambda)$ be the pmf of a random variable corresponding to the measure $\mu_{\overline{W}}(\cdot)$.

Use the independence of $Z(\boldsymbol{c})$ conditional on λ, ν to obtain

$$||p_\epsilon(Z \mid \lambda, \nu) - p(Z \mid \lambda, \nu)||_{\mathrm{TV}} \leq \sum_{\boldsymbol{c} \in \mathcal{C}^+} ||p_\epsilon(Z(\boldsymbol{c}) \mid \lambda, \nu) - p(Z(\boldsymbol{c}) \mid \lambda, \nu)||_{\mathrm{TV}}.$$

This implies that $||p_\epsilon(Z(\boldsymbol{c}) \mid \lambda, \nu) - p(Z(\boldsymbol{c}) \mid \lambda, \nu)||_{\mathrm{TV}} < \epsilon/N_z$, where $N_z = |\mathcal{C}^+|$, is sufficient for $||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}} < \epsilon$. For any $\boldsymbol{c}$ and any set $H \subset \{1, \ldots, K\}$, recall that $Z(\boldsymbol{c})_H = Z(\boldsymbol{c})_h, h \in H$. Define $n^*_H(\boldsymbol{c}) = n(\boldsymbol{c}) - Z(\boldsymbol{c})_H$, and $\tilde{\nu}^*_H = 1 - \sum_{h \in H} \tilde{\nu}_h$. Let $Z^*(\boldsymbol{c})_H = (Z(\boldsymbol{c})_H, n^*_H(\boldsymbol{c}))$, which is distributed Multinomial$(n(\boldsymbol{c}), (\tilde{\nu}_H, \tilde{\nu}^*_H))$. Put

$$\mathcal{H}_{\boldsymbol{c}} = \{H \subset \{1, \ldots, K\} : ||p_\epsilon(Z^*(\boldsymbol{c})_H \mid \lambda, \nu) - p(Z^*(\boldsymbol{c})_H \mid \lambda, \nu)||_{\mathrm{TV}} < \epsilon/N_z\},$$

and for each $\boldsymbol{c}$ define the subset H by

$$H = \left\{ H \in \mathcal{H}_{\boldsymbol{c}} : \sum_{h \in H} \tilde{\nu}_h = \bigvee_{H \in \mathcal{H}_{\boldsymbol{c}}} \sum_{h \in H} \tilde{\nu}_h \right\},$$

where $\bigvee$ is the max function. Define $\mathcal{P}_\epsilon$ by the update rule:

1. For every $\boldsymbol{c} \in \mathcal{C}^+$, sample $Z^*(\boldsymbol{c})_H$ from the normal approximation $\overline{W}$ defined above.

2. Conditional on $n^*_H(\boldsymbol{c})$, sample $Z(\boldsymbol{c})_{H^c} \sim \text{Multinomial}(n^*(\boldsymbol{c})_H, \tilde{\nu}_{H^c})$ from its exact multinomial distribution.

3. Sample ν, λ from their exact full conditionals.

This chain satisfies assumption 6.2.2.

E.7 Proof of Theorem 6.3.1

First we show a lemma that is used in the proof of the main result.

Lemma E.7.1. *The* $\text{PG}(1, \alpha)$ *distribution is a log-concave probability law.*

Proof. If $\omega \sim \text{PG}(1, \alpha)$, then it is equal in distribution to the infinite sum of Exponentials

$$\omega \sim \sum_{k=0}^{\infty} \varphi_k, \quad \varphi_k = \frac{g_k}{\pi^2(k - 1/2)^2 + \alpha^2/2},$$

where $g_k \sim \text{Exp}(1)$, $\varphi_k \sim \text{Exp}(\pi^2(k - 1/2)^2 + \alpha^2/2)$, and φ_k has a log-concave probability distribution since $\text{Exp}(\lambda)$ is log-concave for all finite λ (see e.g. Bagnoli and Bergstrom (2005)). Consider the sequence of random variables

$$\omega_n \sim \sum_{k=0}^{n} \frac{g_k}{\pi^2(k - 1/2)^2 + \alpha^2/2} = \sum_{k=0}^{\infty} \varphi_k$$

for $n = 0, \ldots, \infty$. For any finite n, ω_n has a log-concave distribution since the sum of independent random variables having log-concave distributions is log-concave (see Proposition 3.5 in Saumard et al. (2014)). As $\omega_n \xrightarrow{D} \omega$ (indicating convergence in distribution), ω is log concave from Proposition 3.6 in Saumard et al. (2014). $\square$

E.7.1 Proof of main result

We want to show $\sup_{\theta\in\Theta} ||\mathcal{P}(\theta,\cdot) - \mathcal{P}_\epsilon(\theta,\cdot)||_{\mathrm{TV}} < \epsilon$ with high probability. Here, $\mathcal{P}(\theta,\cdot)$ is the transition kernel based on the full sample of N observations for the Gibbs sampler in (6.18a)-(6.18b), and $\mathcal{P}_\epsilon(\theta,\cdot)$ uses subsets of data of size $|V| \leq N$ to approximate $X'\Omega X$ by $\frac{N}{|V|}X_V'\Omega_V X_V$, in accordance with the update rule in (6.19a)-(6.19c).

We begin by showing how to construct a transition kernel $\mathcal{P}_\epsilon(\theta,\cdot)$ that achieves $||\mathcal{P}(\theta,\cdot) - \mathcal{P}_\epsilon(\theta,\cdot)||_{\mathrm{TV}} < \epsilon$ *conditional* on the current state $\theta = (\beta,\omega)$, then we show that we can control the supremum. First, notice that the Gibbs sampling update rule in (6.19a)-(6.19c) depends on θ only through β, so we need only condition on β. Define

$$\Sigma(\beta) = \mathrm{cov}(\omega^{1/2}x \mid \beta), \qquad \Sigma_N(\beta) = \frac{1}{N}X'\Omega X, \qquad \Sigma_V(\beta) = \frac{1}{|V|}X_V'\Omega_V X_V$$

$$S_N(\beta) = \frac{1}{N}(\Sigma_N(\beta) + B^{-1}/N)^{-1}, \qquad S_V(\beta) = \frac{1}{N}(\Sigma_V(\beta) + B^{-1}/N)^{-1};$$

we will sometimes suppress dependence on β for notational convenience. Recall that the distribution of β_{t+1} given ω_{t+1} is $\mathrm{Normal}(S_N X'\kappa, S_N)$, with $\kappa = y - 1/2$. Let $\mathcal{N}(\cdot; m, M)$ be the measure induced by a normal random variable with mean m and covariance M.

We first show that for every δ and every $0 < q < 1$ there exists a $|V|$ for which

$$||\Sigma_N - \Sigma_V|| \leq \delta||\Sigma||$$

with probability $1-q$ whenever $N > |V|$. In practice, the achievable q with $|V| < N$ will depend on N and δ. We then apply this to bound the Kullback-Leibler divergence

$$\begin{aligned}&\mathrm{KL}\left(\mathcal{N}(\cdot; S_V X'\kappa, S_V) \;||\; \mathcal{N}(\cdot; S_N X'\kappa, S_N)\right)\\ &= \frac{1}{2}\left(\mathrm{tr}\left(S_N^{-1}S_V\right) - p + \log\left(\frac{|S_N|}{|S_V|}\right) + Q\right),\end{aligned}$$

with $Q = (S_N X'\kappa - S_V X'\kappa)' S_N^{-1} (S_N X'\kappa - S_V X'\kappa)$. We then use Pinsker's inequality to obtain a total variation bound. We will choose δ as a function of ϵ and quantities depending on β to obtain $||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}} < \epsilon$; thus, the supremum is controlled by adaptive choice of δ. When this requires $|V| > N$, put $V = \{1, \ldots, N\}$ and obtain the exact kernel.

We proceed in four steps:

1. Showing we can control $||\Sigma_V - \Sigma_N||$ with high probability;
2. Obtaining bounds on the eigenvalues of Σ_V and Σ_N when $||\Sigma_V - \Sigma_N|| < \delta||\Sigma||$;
3. Using (a) and (b) to control the KL; and
4. Showing how to choose δ as a function of β to achieve uniform control of $||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}}$.

Part (a): Control of $||\Sigma_V - \Sigma_N||$.

The sample covariance matrix of $z_i = \omega_i^{1/2} x_i$ is $X'\Omega X$. The z_i are iid given β, since

$$f_z(z_1, \ldots, z_N \mid \beta) = \int \left(\prod_{i=1}^N f_{z_i}(z_i \mid x_i, \beta) g_{x_i}(x_i \mid \beta) \right) dx_1 \ldots dx_N,$$

$$= \prod_{i=1}^N \left(\int f_{z_i}(z_i \mid x_i, \beta) g_{x_i}(x_i \mid \beta) dx_i \right) = \prod_{i=1}^N f_{z_i}(z_i \mid \beta),$$

where the first line used independence of z_i given x_i and β and the second line used Fubini. Now we show that $\mathbb{E}\left[\omega^{1/2} x\right] = 0$. Since $x \sim f_x(x; \alpha)$,

$$\mathbb{E}\left[\omega^{1/2} x\right] = \int_{x \in \mathbb{R}^p} \int_{\omega \in \mathbb{R}_+} \omega^{1/2} x e^{-(x\beta)^2 \omega/2} f_\omega(\omega) \cosh\left(\frac{x\beta}{2}\right) f_x(x; \alpha) d\omega dx, \qquad \text{(E.5)}$$

where $f_\omega(\omega)$ is the PG$(1, 0)$ density and $f_x(x; \alpha)$ is symmetric about the origin by assumption. All of the terms in the integrand involving x are symmetric about 0, so the expectation is zero. Since $f_{\sqrt{\omega}}(y) = 2y f_\omega(y^2)$ is the density of $\sqrt{\omega}$, and

$f_\omega(y)$ is log-concave by Lemma E.7.1, $f_{\sqrt{\omega}}(y)$ is log-concave. Since the product of log-concave functions is log-concave, and $f_x(x;\alpha)$ is log-concave by assumption, the distribution of $\omega^{1/2}x$ is log-concave. This allows us to apply the following Theorem from Adamczak et al. (2010).

Theorem E.7.1 (Adamczak 2010, Theorem 4.1)**.** *Let $Z_1, \ldots, Z_N$ be i.i.d. random vectors distributed according to an isotropic, log-concave probability measure on $\mathbb{R}^p$. For every $\delta \in (0,1)$ and $M > 1$ there exists $C(\delta, M) > 0$ such that if $C(\delta, M)p \leq N$, then with probability at least $1 - e^{-cM\sqrt{p}}$,*

$$||\Sigma_N - I_p|| \leq \delta,$$

where $c > 0$ is an absolute constant and Σ_N is the sample covariance matrix based on N samples. Moreover, one can take $C(\delta, M) = CM^4\delta^{-2}\log^2(2M^2\delta^{-2})$, where C is an absolute constant.

Here, $||\Sigma||$ is the spectral norm of Σ, i.e.

$$\sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{||\Sigma x||_2}{||x||_2}.$$

Adamczak et al. (2010) notes that when the distribution of Z is not isotropic but does have zero mean, we instead have

$$||\Sigma_N - \Sigma|| < \delta||\Sigma||$$

with the same probability.

Note that at best we can achieve probabilities on the order of $1 - e^{-N^{1/4}}$, since $C(\delta, M)$ grows like M^4 up to a log factor. Now, fix a q and suppose that for $|V| < N$ we can achieve $||\Sigma_V - \Sigma|| < \frac{\delta}{2}||\Sigma||$ with probability at least $1-q$, for some value of δ to be determined subsequently. Then with probability $(1-q)^2$ we have $||\Sigma_N - \Sigma|| < \frac{\delta}{2}||\Sigma||$, and by the triangle inequality, with the same probability we have $||\Sigma_N - \Sigma_V|| < \delta||\Sigma||$. We now show that for sufficiently small δ, this allows us to bound the eigenvalues of Σ_N and Σ_V.

Part (b) : Control of eigenvalues of Σ_N and Σ_V

If $||\Sigma_N - \Sigma|| < \frac{\delta}{2}||\Sigma||$ then

$$||\Sigma_N + B^{-1}/N - (\Sigma + B^{-1}/N)|| = ||\Sigma_N - \Sigma|| \leq \frac{\delta}{2}||\Sigma||.$$

Now, use $||\Sigma_N - \Sigma||_F^2 \leq p\,||\Sigma_N - \Sigma||^2$, and that Σ, Σ_N are Hermitian, and apply the Hoffman-Weilandt inequality (Bhatia (2013), Hoffman et al. (1953), Tao (2015)), which ensures existence of a permutation ρ of the eigenvalues of Σ_N such that

$$\sum_{j=1}^{p}(\lambda_{\rho(j)}(\Sigma_N) - \lambda_j(\Sigma))^2 < ||\Sigma_N - \Sigma||_F^2 \leq (\sqrt{p}||\Sigma_N - \Sigma||)^2 \leq p\frac{\delta^2}{4}||\Sigma||^2,$$

where $\lambda_j(\Sigma)$ is the jth eigenvalue of the matrix Σ. So there exists a j such that

$$(\lambda_{\max}(\Sigma_N) - \lambda_j(\Sigma))^2 < p\frac{\delta^2}{4}||\Sigma||^2, \tag{E.6}$$

where $\lambda_{\max}(\Sigma_N)$ is the largest eigenvalue of Σ_N. This implies that

$$\lambda_{\max}(\Sigma_N) < \lambda_{\max}(\Sigma) + \sqrt{p}\frac{\delta}{2}||\Sigma||. \tag{E.7}$$

This is immediate if $j = 1$ in (E.6). If $j > 1$ in (E.6), then we must have (E.7), since otherwise

$$(\lambda_{\max}(\Sigma_N) - \lambda_j(\Sigma))^2 \geq (\lambda_{\max}(\Sigma) + \sqrt{p}\frac{\delta}{2}||\Sigma|| - \lambda_j(\Sigma))^2 \geq p\frac{\delta^2}{4}||\Sigma||^2.$$

Furthermore, there exists a j' for which

$$(\lambda_{\min}(\Sigma_N) - \lambda_{j'}(\Sigma))^2 < p\frac{\delta^2}{4}||\Sigma||^2,$$

with $\lambda_{\min}(\Sigma_N)$ the smallest eigenvalue of Σ_N, implying

$$\lambda_{\min}(\Sigma_N) > \lambda_{\min}(\Sigma) - \sqrt{p}\frac{\delta}{2}||\Sigma||$$

by analogous argument. So if

$$\delta < p^{-1/2}\frac{\lambda_{\min}(\Sigma)}{(\lambda_{\max}(\Sigma) + \lambda_{\min}(\Sigma))},$$

we have $\lambda_{\min}(\Sigma_N) > \lambda_{\min}(\Sigma)/2$, ensuring the smallest eigenvalue of Σ_N is bounded away from zero, and $\lambda_{\max}(\Sigma_N) < \lambda_{\max}(\Sigma)+\lambda_{\min}(\Sigma)/2$. Now, put $\ell_{\max}(\beta) = \lambda_{\max}(\Sigma(\beta))+\lambda_{\min}(\Sigma(\beta))/2$ and $\ell_{\min}(\beta) = \lambda_{\min}(\Sigma(\beta))/2$. With $B = \eta I_p$,

$$\lambda_{\min}((\Sigma_N + B^{-1}/N)^{-1}) \geq \frac{1}{\ell_{\max}(\beta) + (N\eta)^{-1}},$$

$$\lambda_{\max}((\Sigma_N + B^{-1}/N)^{-1}) \leq \frac{1}{\ell_{\min}(\beta) + (N\eta)^{-1}},$$

$$\lambda_{\min}((\Sigma_V + B^{-1}/N)^{-1}) \geq \frac{1}{\ell_{\max}(\beta) + (N\eta)^{-1}},$$

$$\lambda_{\max}((\Sigma_V + B^{-1}/N)^{-1}) \leq \frac{1}{\ell_{\min}(\beta) + (N\eta)^{-1}},$$

where the result for Σ_V follows because we also have $||\Sigma_V - \Sigma|| < \frac{\delta}{2}||\Sigma||$.

Part (c): Control of KL Divergence

Now we show control of Q, assuming that $||\Sigma - \Sigma_N|| \leq \delta||\Sigma||$.

$$\begin{aligned}
Q =& \left(\left(\frac{N}{|V|}X_V'\Omega_V X_V + B^{-1}\right)^{-1} X'\kappa - (X'\Omega X + B^{-1})^{-1}X'\kappa\right)' (X'\Omega X + B^{-1}) \\
& \left(\left(\frac{N}{|V|}X_V'\Omega_V X_V + B^{-1}\right)^{-1} X'\kappa - (X'\Omega X + B^{-1})^{-1}X'\kappa\right) \\
=& \left(\frac{1}{N}\left(\frac{1}{|V|}X_V'\Omega_V X_V + \frac{B^{-1}}{N}\right)^{-1} X'\kappa - \frac{1}{N}\left(\frac{1}{N}X'\Omega X + \frac{B^{-1}}{N}\right)^{-1} X'\kappa\right)' \\
& N\left(\frac{1}{N}X'\Omega X + \frac{B^{-1}}{N}\right) \\
& \left(\frac{1}{N}\left(\frac{1}{|V|}X_V'\Omega_V X_V + \frac{B^{-1}}{N}\right)^{-1} X'\kappa - \frac{1}{N}\left(\frac{1}{N}X'\Omega X + \frac{B^{-1}}{N}\right)^{-1} X'\kappa\right) \\
=& \frac{1}{N}\left(\left(\Sigma_V + \frac{B^{-1}}{N}\right)^{-1} X'\kappa - \left(\Sigma_N + \frac{B^{-1}}{N}\right)^{-1} X'\kappa\right)' \left(\Sigma_N + \frac{B^{-1}}{N}\right)
\end{aligned}$$

$$
\begin{aligned}
&\left(\left(\Sigma_V + \frac{B^{-1}}{N}\right)^{-1} X'\kappa - \left(\Sigma_N + \frac{B^{-1}}{N}\right)^{-1} X'\kappa\right) \\
&\leq \frac{(\ell_{\max}(\beta) + (N\eta)^{-1})}{N} \left\|\left(\Sigma_V + \frac{B^{-1}}{N}\right)^{-1} X'\kappa - \left(\Sigma_N + \frac{B^{-1}}{N}\right)^{-1} X'\kappa\right\|^2 \\
&= \frac{(\ell_{\max}(\beta) + (N\eta)^{-1})}{N} \\
&\times \left\|\left(\Sigma_V + \frac{B^{-1}}{N}\right)^{-1} \left[X'\kappa - \left(\Sigma_V + \frac{B^{-1}}{N}\right)\left(\Sigma_N + \frac{B^{-1}}{N}\right)^{-1} X'\kappa\right]\right\|^2 \\
&\leq \frac{(\ell_{\max}(\beta) + (N\eta)^{-1})}{N(\ell_{\min}(\beta) + (N\eta)^{-1})^2} \left\|\left[I - \left(\Sigma_V + \frac{B^{-1}}{N}\right)\left(\Sigma_N + \frac{B^{-1}}{N}\right)^{-1}\right] X'\kappa\right\|^2 \\
&\leq \frac{(\ell_{\max}(\beta) + (N\eta)^{-1})}{N(\ell_{\min}(\beta) + (N\eta)^{-1})^2} \left\|I - \left(\Sigma_V + \frac{B^{-1}}{N}\right)\left(\Sigma_N + \frac{B^{-1}}{N}\right)^{-1}\right\|^2 ||X'\kappa||^2 \\
&\leq \frac{(\ell_{\max}(\beta) + (N\eta)^{-1})}{(\ell_{\min}(\beta) + (N\eta)^{-1})^2} \left(\frac{\delta^2 ||X'\kappa||^2 \ell_{\max}(\beta)^2}{(\ell_{\min}(\beta) + (N\eta)^{-1})^2}\right) \\
&\leq \frac{(\ell_{\max}(\beta) + (N\eta)^{-1})}{(\ell_{\min}(\beta) + (N\eta)^{-1})^2} \left(\frac{\delta^2 p \ell_{\max}(\beta)^2}{4(\ell_{\min}(\beta) + (N\eta)^{-1})^2}\right)
\end{aligned}
$$

where various steps used Cauchy-Schwartz, assume X is standardized to unit variance, $\kappa_i \in \{-1/2, 1/2\}$, $||\Sigma_V - \Sigma_N|| < \delta \ell_{\max}(\beta)$, and

$$
\begin{aligned}
&\left\|I - \left(\Sigma_V + \frac{B^{-1}}{N}\right)\left(\Sigma_N + \frac{B^{-1}}{N}\right)^{-1}\right\| \\
&\leq \left\|\left(\Sigma_N + \frac{B^{-1}}{N}\right) - \left(\Sigma_V + \frac{B^{-1}}{N}\right)\right\| \left\|\left(\Sigma_N + \frac{B^{-1}}{N}\right)^{-1}\right\|.
\end{aligned}
$$

To bound the other terms in the KL, first note that

$$
\begin{aligned}
&\operatorname{tr}\left((X'\Omega X + B^{-1})\left(\frac{N}{|V|} X_V'\Omega_V X_V + B^{-1}\right)^{-1}\right) - p = \\
&= \operatorname{tr}\left(N\left(\Sigma_N + \frac{B^{-1}}{N}\right)\frac{1}{N}\left(\Sigma_V + \frac{B^{-1}}{N}\right)^{-1} - I\right)
\end{aligned}
$$

$$= \text{tr}\left((\Sigma_N - \Sigma_V)\left(\Sigma_V + \frac{B^{-1}}{N}\right)^{-1}\right)$$

$$\leq \lambda_{\max}(\Sigma_N - \Sigma_V)\text{tr}\left((\Sigma_V + B^{-1}/N)^{-1}\right)$$

$$\leq \frac{p\delta\ell_{\max}(\beta)}{\ell_{\min}(\beta) + (N\eta)^{-1}}.$$

Further, from Lemma B.2 in Pati et al. (2014), since S_V and S_N are both positive definite for $|V| > p$,

$$\log|S_N S_V^{-1}| < \text{tr}\left(S_N^{-1}S_V - I\right).$$

So putting all of the bounds together,

$$\text{KL}\left(\mathcal{N}(\cdot; S_V X'\kappa, S_V) \;||\; \mathcal{N}(\cdot; S_N X'\kappa, S_N)\right)$$

$$\leq \frac{1}{2}\frac{(\ell_{\max}(\beta) + (N\eta)^{-1})}{(\ell_{\min}(\beta) + (N\eta)^{-1})^2}\left(\frac{\delta^2 p\ell_{\max}(\beta)^2}{4(\ell_{\min}(\beta) + (N\eta)^{-1})^2}\right) + \frac{p\delta\ell_{\max}(\beta)}{\ell_{\min}(\beta) + (N\eta)^{-1}}$$

$$\leq \frac{\delta^2 p\ell_{\max}(\beta)^3}{8\ell_{\min}(\beta)^4} + \frac{p\delta\ell_{\max}(\beta)}{\ell_{\min}(\beta)}$$

Part (d): Uniform control of $||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\text{TV}}$

Notice that

$$\frac{\lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma) + \lambda_{\min}(\Sigma)} = \frac{2\ell_{\min}(\beta)}{\ell_{\min}(\beta) + \ell_{\max}(\beta)} > \frac{\ell_{\min}(\beta)}{\ell_{\max}(\beta)}.$$

So, put

$$\delta = \frac{2\sqrt{2}\epsilon p^{-1/2}\ell_{\min}(\beta)^2}{\ell_{\max}(\beta)^{3/2}} \wedge \frac{\epsilon^2\ell_{\min}(\beta)}{p\ell_{\max}(\beta)} \wedge \frac{\epsilon p^{-1/2}\lambda_{\min}(\beta)}{\lambda_{\min}(\beta) + \lambda_{\max}(\beta)},$$

with $0 < \epsilon < 1$, where now we explicitly indicate the dependence of $\lambda_{\min}(\Sigma)$, $\lambda_{\max}(\Sigma)$ on β through the notation $\lambda_{\min}(\beta), \lambda_{\max}(\beta)$, and the final term ensures we satisfy the earlier condition on δ. Thus we obtain

$$\text{KL}\left(\mathcal{N}(\cdot; S_V X'\kappa, S_V) \;||\; \mathcal{N}(\cdot; S_N X'\kappa, S_N)\right) \leq 2\epsilon^2.$$

Thus by adaptively choosing δ as a function of β, we obtain approximation error that does not depend on β. Now, apply Pinsker's inequality, so that

$$||\mathcal{N}(\cdot; S_V X'\kappa, S_V) - \mathcal{N}(\cdot; S_N X'\kappa, S_N)||_{\mathrm{TV}} \leq \epsilon$$

with probability at least

$$(1 - e^{-cM\sqrt{p}})^2$$

whenever $|V| \geq pCM^4\delta^{-2}\log^2(2M^2\delta^{-2})$.

E.8 Proof of Theorem 6.3.2

The results in this section concern the model in (6.21) with priors in (6.22). The transition kernel $\mathcal{P}$ is induced by the marginal sampler defined in section 6.3.4, and the approximating kernel $\mathcal{P}_\epsilon$ substitutes $\Sigma_\epsilon = U_\epsilon \Lambda_\epsilon U_\epsilon'$ for Σ where U_ϵ is $n \times r$, Λ_ϵ is $r \times r$, and $r \leq n$.

E.8.1 Result for predictive $p(f \mid \theta)$

First we show that for every $\epsilon \in (0, 1)$ there exists a δ depending on the state $\theta = (\sigma^2, \tau^2, \phi)$ such that $||\Sigma - \Sigma_\epsilon|| < \delta$ implies

$$||p(f \mid \theta) - p_\epsilon(f \mid \theta)||_{\mathrm{TV}} < \epsilon,$$

where f is the latent Gaussian process in (6.21), $p(f \mid \theta)$ is its full conditional in the exact MCMC algorithm (we repress the dependence on y for notational brevity), and $p_\epsilon(f \mid \theta)$ is its full conditional in the approximate sampler. The strategy is to show a bound on

$$\begin{aligned}\mathrm{KL}\left(p(f \mid \theta) \;||\; p_\epsilon(f \mid \theta)\right) = \frac{1}{2}(&\mathrm{tr}\left((\Psi_\epsilon)^{-1}\Psi\right) - n + \log\left(|\Psi|^{-1}|\Psi_\epsilon|\right) \\ &+ y'(\Psi_\epsilon - \Psi)'(\Psi_\epsilon)^{-1}(\Psi_\epsilon - \Psi)y)\end{aligned}$$

where $\Psi = (\tau^2\Sigma + \sigma^2 I)^{-1}$ and $\Psi_\epsilon = (\tau^2\Sigma_\epsilon + \sigma^2 I)^{-1}$, then use Pinsker's inequality. We now bound each term separately following the proof of Theorem 6.3.1.

The eigenvalues of Ψ and Ψ_ϵ satisfy

$$\lambda_{\min}(\Psi) > \frac{1}{\tau^2\lambda_{\max}(\Sigma) + \sigma^2}, \quad \lambda_{\max}(\Psi) < \frac{1}{\sigma^2}$$

$$\lambda_{\min}(\Psi_\epsilon) > \frac{1}{\tau^2\lambda_{\max}(\Sigma_\epsilon) + \sigma^2}, \quad \lambda_{\max}(\Psi_\epsilon) < \frac{1}{\sigma^2}.$$

We assume that the approximation achieves $||\Sigma_\epsilon - \Sigma||_F < \delta$ with probability $1 - 10^{-d}$. So then using the strategy in the proof of Theorem 6.3.1,

$$\begin{aligned}
Q &= (\Psi_\epsilon y - \Psi y)'(\Psi_\epsilon)^{-1}(\Psi_\epsilon y - \Psi y) \\
&\leq (\tau^2\lambda_{\max}(\Sigma_\epsilon) + \sigma^2)||\Psi_\epsilon y - \Psi y||^2 \\
&\leq (\tau^2\lambda_{\max}(\Sigma_\epsilon) + \sigma^2)||\Psi||^2||\Psi^{-1}\Psi_\epsilon - I||^2 \ ||y||^2 \\
&\leq \frac{(\tau^2\lambda_{\max}(\Sigma_\epsilon) + \sigma^2)}{\sigma^4}||\Psi^{-1}\Psi_\epsilon - I||^2 \ ||y||^2 \\
&\leq \frac{(\tau^2\lambda_{\max}(\Sigma_\epsilon) + \sigma^2)}{\sigma^4}||\Sigma - \Sigma_\epsilon||^2 \ ||\Psi_\epsilon||^2 \ ||y||^2 \\
&\leq \frac{\tau^4\delta^2||y||^2(\tau^2\lambda_{\max}(\Sigma_\epsilon) + \sigma^2)}{\sigma^8} \\
&\leq \frac{\tau^4\delta^2 n(\tau^2\lambda_{\max}(\Sigma_\epsilon) + \sigma^2)}{\sigma^8},
\end{aligned}$$

where we used that y is standardized to unit variance. Now since

$$\begin{aligned}
\text{tr}\left(\Psi_\epsilon^{-1}\Psi\right) - n = \text{tr}\left((\Psi_\epsilon^{-1} - \Psi^{-1})\Psi\right) &\leq ||\Psi||\text{tr}\left((\Psi_\epsilon^{-1} - \Psi^{-1})\right) \\
&\leq \frac{n\tau^2\delta}{\sigma^2},
\end{aligned}$$

applying Lemma B.2 in Pati et al. (2014), we obtain the KL bound

$$\text{KL}\left(p(f \mid \theta) \ || \ p_\epsilon(f \mid \theta)\right) \leq \frac{\tau^4\delta^2 n(\tau^2\lambda_{\max}(\Sigma_\epsilon) + \sigma^2)}{2\sigma^8} + \frac{n\delta\tau^2}{\sigma^2}.$$

Apply Pinsker's inequality and get

$$||p(f \mid \theta) - p_\epsilon(f \mid \theta)||_{\text{TV}} \leq \sqrt{n\left(\frac{\tau^4\delta^2(\tau^2\lambda_{\max}(\Sigma_\epsilon) + \sigma^2)}{4\sigma^8} + \frac{\delta\tau^2}{2\sigma^2}\right)}.$$

So choose

$$\delta = \frac{\epsilon^2\sigma^4}{\tau^2\sqrt{n(\tau^2\lambda_{\max}(\Sigma_\epsilon)+\sigma^2)}} \wedge \frac{\epsilon^2\sigma^2}{n\tau^2}$$

for $0 < \epsilon < 1$ to achieve TV error of ϵ. By adapting the required accuracy δ to the state, and noting that one can always achieve $\delta = 0$ by utilizing the exact Σ so that no value of δ is unachievable, this is sufficient to show that the total variation error can be controlled uniformly. Note we did not need the assumption that Σ_ϵ is a partial eigendecomposition; this will be used below.

E.8.2 Result for $\mathcal{P}_\epsilon$

We first prove a lemma that will be used to obtain the main result. We will in general use θ_* to represent the proposal value in Metropolis-Hastings algorithms and θ to represent the current state.

Lemma E.8.1. *Consider transition kernels $\mathcal{P}(\theta,\cdot), \mathcal{P}_\epsilon(\theta,\cdot)$ constructed by Metropolis-Hastings algorithms with identical proposal distributions and acceptance probabilities $p(\theta \to \theta_*)$, $p_\epsilon(\theta \to \theta_*)$ for any $\theta, \theta_* \in \Theta$. If*

$$p_{\sup} = \sup_{\theta_* \in \Theta} \sup_{\theta \in \Theta} |p(\theta \to \theta_*) - p_\epsilon(\theta \to \theta_*)| < \frac{\epsilon}{2},$$

then

$$\sup_\theta ||\mathcal{P}(\theta,\cdot) - \mathcal{P}_\epsilon(\theta,\cdot)||_{\mathrm{TV}} < \epsilon.$$

Proof. Let $Q(\theta; d\theta_*)$ denote the proposal distribution, which may depend on the current state θ. Then

$$\mathcal{P}(\theta,\theta_*) = \int p(\theta \to \theta_*) Q(\theta; d\theta_*) + \delta_\theta(\theta_*) \int (1 - p(\theta \to \theta_*)) Q(\theta; d\theta_*)$$

$$\mathcal{P}_\epsilon(\theta,\theta_*) = \int p_\epsilon(\theta \to \theta_*) Q(\theta; d\theta_*) + \delta_\theta(\theta_*) \int (1 - p_\epsilon(\theta \to \theta_*)) Q(\theta; d\theta_*)$$

Hence, we have

$$
\begin{aligned}
&\sup_{\theta} ||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}} \\
&= \sup_{\theta \in \Theta} \sup_{A \subset \Theta} \left| \int_A p(\theta \to \theta_*) Q(\theta; d\theta_*) + \mathbb{1}_{\{\theta \in A\}} \int_A (1 - p(\theta \to \theta_*)) Q(\theta; d\theta_*) \right. \\
&\left. - \int_A p_\epsilon(\theta \to \theta_*) Q(\theta; d\theta_*) - \mathbb{1}_{\{\theta \in A\}} \int_A (1 - p_\epsilon(\theta \to \theta_*)) Q(\theta; d\theta_*) \right| \\
&\leq \sup_{\theta \in \Theta} \left| \int_\Theta p(\theta \to \theta_*) Q(\theta; d\theta_*) + \int_\Theta (1 - p(\theta \to \theta_*)) Q(\theta; d\theta_*) \right. \\
&\left. - \int_\Theta p_\epsilon(\theta \to \theta_*) Q(\theta; d\theta_*) - \int_\Theta (1 - p_\epsilon(\theta \to \theta_*)) Q(\theta; d\theta_*) \right| \\
&\leq \sup_{\theta \in \Theta} \left(\left| \int_\Theta [p(\theta \to \theta_*) - p_\epsilon(\theta \to \theta_*)] Q(\theta; d\theta_*) \right| \right. \\
&\left. + \left| \int_\Theta [p_\epsilon(\theta \to \theta_*) - p(\theta \to \theta_*)] Q(\theta; d\theta_*) \right| \right) \\
&\leq \sup_{\theta \in \Theta} \int_\Theta (|p(\theta \to \theta_*) - p_\epsilon(\theta \to \theta_*)| + |p_\epsilon(\theta \to \theta_*) - p(\theta \to \theta_*)|) Q(\theta; d\theta_*) \\
&\leq (\sup_{\theta \in \Theta} \sup_{\theta_* \in \Theta} [|p(\theta \to \theta_*) - p_\epsilon(\theta \to \theta_*)| + |p_\epsilon(\theta \to \theta_*) - p(\theta \to \theta_*)|]) \int_\Theta Q(\theta; d\theta_*) \\
&\leq \epsilon \int_\Theta Q(\theta; d\theta_*) \leq \epsilon
\end{aligned}
$$

□

Main result: approximation error for GP MH steps

We now show that for every $\epsilon > 0$, the kernel $\mathcal{P}_\epsilon$ that replaces Σ with Σ_ϵ, achieving $||\Sigma - \Sigma_\epsilon|| < \delta$ with probability $1-q$, satisfies Assumption 6.2.2 (also with probability $1 - q$). This result uses the additional assumption that Σ_ϵ is a partial eigendecomposition. To simplify the exposition and reduce length, the result is obtained for a joint Metropolis-Hastings step for (σ^2, τ^2). A similar result could be obtained for the

sequential Metropolis-Hastings steps in the marginal sampler described in Section 6.3.4 by appropriately re-defining the acceptance probability.

Applying lemma E.8.1, we need only control p_{sup}. The absolute difference in MH acceptance probabilities for the marginal sampler is

$$\begin{aligned} D_\epsilon(\theta, \theta_*) &= |p_\epsilon(\theta \to \theta_*) - p(\theta \to \theta_*)| \\ &= \left| \left(\frac{L_\epsilon(y \mid \theta_*) p(\theta_*) q(\theta \mid \theta_*)}{L_\epsilon(y \mid \theta) p(\theta) q(\theta_* \mid \theta)} \wedge 1 \right) - \left(\frac{L(y \mid \theta_*) p(\theta_*) q(\theta \mid \theta_*)}{L(y \mid \theta) p(\theta) q(\theta_* \mid \theta)} \wedge 1 \right) \right| \\ &= \left| (r_\epsilon(\theta \to \theta_*) \wedge 1) - (r(\theta \to \theta_*) \wedge 1) \right|. \end{aligned}$$

Initially focus on the case where both $r_\epsilon(\theta \to \theta_*)$ and $r(\theta \to \theta_*)$ are less than one, and set $M(\theta, \theta_*) = \frac{p(\theta_*) q(\theta | \theta_*)}{p(\theta) q(\theta_* | \theta)}$. Then

$$\begin{aligned} D_\epsilon(\theta, \theta_*) &= M(\theta, \theta_*) \left| \frac{L_\epsilon(y \mid \theta_*)}{L_\epsilon(y \mid \theta)} - \frac{L(y \mid \theta_*)}{L(y \mid \theta)} \right| \\ &= M(\theta, \theta_*) \left| \frac{|2\pi(\tau_*^2 \Sigma_\epsilon + \sigma_*^2 I)|^{-1/2} \exp(-y'(\tau_*^2 \Sigma_\epsilon + \sigma_*^2 I)^{-1} y/2)}{|2\pi(\tau^2 \Sigma_\epsilon + \sigma^2 I)|^{-1/2} \exp(-y'(\tau^2 \Sigma_\epsilon + \sigma^2 I)^{-1} y/2)} \right. \\ &\quad \left. - \frac{|2\pi(\tau_*^2 \Sigma + \sigma_*^2 I)|^{-1/2} \exp(-y'(\tau_*^2 \Sigma + \sigma_*^2 I)^{-1} y/2)}{|2\pi(\tau^2 \Sigma + \sigma^2 I)|^{-1/2} \exp(-y'(\tau^2 \Sigma + \sigma^2 I)^{-1} y/2)} \right|. \\ &= M(\theta, \theta_*) \left| \frac{\left(\prod_{i=1}^n \tau_*^2 \lambda_i^\epsilon + \sigma_*^2\right)^{-1/2} \exp(-y'(\tau_*^2 \Sigma_\epsilon + \sigma_*^2 I)^{-1} y/2)}{\left(\prod_{i=1}^n \tau^2 \lambda_i^\epsilon + \sigma^2\right)^{-1/2} \exp(-y'(\tau^2 \Sigma_\epsilon + \sigma^2 I)^{-1} y/2)} \right. \\ &\quad \left. - \frac{\left(\prod_{i=1}^n \tau_*^2 \lambda_i + \sigma_*^2\right)^{-1/2} \exp(-y'(\tau_*^2 \Sigma + \sigma_*^2 I)^{-1} y/2)}{\left(\prod_{i=1}^n \tau^2 \lambda_i + \sigma^2\right)^{-1/2} \exp(-y'(\tau^2 \Sigma + \sigma^2 I)^{-1} y/2)} \right|. \end{aligned}$$

Now use that Σ_ϵ is a rank r partial eigendecomposition of Σ satisfying $||\Sigma_\epsilon - \Sigma||_F < \delta$, implying the following

$$\tau^2 \Sigma_\epsilon + \sigma^2 I = U(\tau^2 \Lambda_\epsilon + \sigma^2 I) U', \qquad \tau^2 \Sigma + \sigma^2 I = U(\tau^2 \Lambda + \sigma^2 I) U'$$

$$\lambda_i^\epsilon = \lambda_i, i \leq r, \qquad \lambda_i^\epsilon = 0, i > r, \qquad \lambda_i < \delta, i > r,$$

where $\Lambda_\epsilon = \text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$, and λ_i is the ith eigenvalue of Σ. Now put

$y_U = y'U$, with ith entry $y_{U,i}$, and obtain $D_\epsilon(\theta, \theta_*) =$

$$M(\theta,\theta_*)\Bigg| \exp\left(-\frac{1}{2}y_U \mathrm{diag}\left[\frac{1}{\tau_*^2\lambda_i^\epsilon+\sigma_*^2}-\frac{1}{\tau^2\lambda_i^\epsilon+\sigma^2}\right]y_U' - \frac{1}{2}\sum_{i=1}^n \log\frac{\tau_*^2\lambda_i^\epsilon+\sigma_*^2}{\tau^2\lambda_i^\epsilon+\sigma^2}\right)$$

$$- \exp\left(-\frac{1}{2}y_U \mathrm{diag}\left[\frac{1}{\tau_*^2\lambda_i+\sigma_*^2}-\frac{1}{\tau^2\lambda_i+\sigma^2}\right]y_U' - \frac{1}{2}\sum_{i=1}^n \log\frac{\tau_*^2\lambda_i+\sigma_*^2}{\tau^2\lambda_i+\sigma^2}\right)\Bigg|$$

$$M(\theta,\theta_*)\exp\left(-\frac{1}{2}\sum_{i=1}^r y_{U,i}^2\left[\frac{1}{\tau_*^2\lambda_i+\sigma_*^2}-\frac{1}{\tau^2\lambda_i+\sigma^2}\right] - \frac{1}{2}\sum_{i=1}^r \log\frac{\tau_*^2\lambda_i+\sigma_*^2}{\tau^2\lambda_i+\sigma^2}\right)$$

$$\times\Bigg|\exp\left(-\frac{1}{2}\sum_{i=r+1}^n y_{U,i}^2\left[\frac{1}{\tau_*^2\lambda_i^\epsilon+\sigma_*^2}-\frac{1}{\tau^2\lambda_i^\epsilon+\sigma^2}\right] - \frac{1}{2}\sum_{i=r+1}^n \log\frac{\tau_*^2\lambda_i^\epsilon+\sigma_*^2}{\tau^2\lambda_i^\epsilon+\sigma^2}\right)$$

$$- \exp\left(-\frac{1}{2}\sum_{i=r+1}^n y_{U,i}^2\left[\frac{1}{\tau_*^2\lambda_i+\sigma_*^2}-\frac{1}{\tau^2\lambda_i+\sigma^2}\right] - \frac{1}{2}\sum_{i=r+1}^n \log\frac{\tau_*^2\lambda_i+\sigma_*^2}{\tau^2\lambda_i+\sigma^2}\right)\Bigg|$$

Put

$$M_1(\theta,\theta_*) = M(\theta,\theta_*)\exp\Bigg(-\frac{1}{2}\sum_{i=1}^r y_{U,i}^2\left[\frac{1}{\tau_*^2\lambda_i+\sigma_*^2}-\frac{1}{\tau^2\lambda_i+\sigma^2}\right]$$

$$-\frac{1}{2}\sum_{i=1}^r \log\frac{\tau_*^2\lambda_i+\sigma_*^2}{\tau^2\lambda_i+\sigma^2}\Bigg)$$

giving

$$D_\epsilon(\theta,\theta_*) = M_1(\theta,\theta_*)\Bigg|\exp\left(-\frac{1}{2}\sum_{i=r+1}^n y_{U,i}^2\left[\frac{1}{\sigma_*^2}-\frac{1}{\sigma^2}\right] - \frac{1}{2}\sum_{i=r+1}^n \log\frac{\sigma_*^2}{\sigma^2}\right)$$

$$- \exp\Bigg(-\frac{1}{2}\sum_{i=r+1}^n y_{U,i}^2\left[\frac{1}{\tau_*^2\lambda_i+\sigma_*^2}-\frac{1}{\tau^2\lambda_i+\sigma^2}\right]$$

$$-\frac{1}{2}\sum_{i=r+1}^n \log\frac{\tau_*^2\lambda_i+\sigma_*^2}{\tau^2\lambda_i+\sigma^2}\Bigg)\Bigg|$$

Call the term inside the absolute value $\Delta(\delta, \theta, \theta_*)$, and simplify to obtain

$$\Delta(\delta,\theta,\theta_*) = \Bigg|\exp\left(-\frac{1}{2}\sum_{i=r+1}^n y_{U,i}^2\left[\frac{1}{\sigma_*^2}-\frac{1}{\sigma^2}\right] - \frac{1}{2}\sum_{i=r+1}^n \log\frac{\sigma_*^2}{\sigma^2}\right)$$

$$
\begin{aligned}
&- \exp\left(-\frac{1}{2}\sum_{i=r+1}^{n} y_{U,i}^2 \left[\frac{1}{\tau_*^2\lambda_i + \sigma_*^2} - \frac{1}{\tau^2\lambda_i + \sigma^2}\right]\right. \\
&\left. - \frac{1}{2}\sum_{i=r+1}^{n} \log\frac{\tau_*^2\lambda_i + \sigma_*^2}{\tau^2\lambda_i + \sigma^2}\right)\Bigg| \\
&\leq \left|\exp\left(-\frac{n-r}{2}\left[\frac{\sigma^2 - \sigma_*^2}{\sigma_*^2\sigma^2} - \log\frac{\sigma_*^2}{\sigma^2}\right]\right)\right. \\
&\left. - \exp\left(-\frac{n-r}{2}\left[\frac{\tau_*^2\delta + \sigma^2 - \tau_*^2\delta - \sigma_*^2}{(\tau_*^2\delta + \sigma_*^2)(\tau^2\delta + \sigma^2)} - \log\frac{\tau_*^2\delta + \sigma_*^2}{\tau^2\delta + \sigma^2}\right]\right)\right|.
\end{aligned}
$$

Taking $\delta \to 0$, $\Delta(\delta, \theta, \theta_*)$ can be made arbitrarily small.

Finally, because the prior on ϕ is finitely supported and depends only on likelihood ratios of the same form as those considered above, control of the approximation error for sampling of ϕ follows easily. Thus, the following algorithm achieves $\sup_{\theta\in\Theta} ||\mathcal{P}(\theta, \cdot) - \mathcal{P}_\epsilon(\theta, \cdot)||_{\mathrm{TV}} < \epsilon$: (a) Take a draw from $q(\theta_* \mid \theta)$; (b) Choose δ such that $\Delta(\delta, \theta, \theta_*) < \frac{\epsilon}{2M_1(\theta,\theta_*)}$; (c) Compute $r_\epsilon(\theta \to \theta_*)$; (d) Use this quantity in the MH acceptance decision; and (e) Sample ϕ from its discrete full conditional distribution.

Although the only case considered above was that where $r_\epsilon(\theta \to \theta_*) < 1$, note that if $|r_\epsilon(\theta \to \theta_*) - r(\theta \to \theta_*)| < \frac{\epsilon}{2}$, then

$$
|(1 \wedge r_\epsilon(\theta \to \theta_*)) - (1 \wedge r_\epsilon(\theta \to \theta_*))| < \frac{\epsilon}{2}.
$$

Noting that $\delta = 0$ is always achievable by taking $\Sigma_\epsilon = \Sigma$, this is sufficient to control the approximation error everywhere in the state space.

E.9 Simulation study : accuracy of approximate eigendecompositions

In each simulation, 1000 points are generated in $\mathbb{R}^k$ and pairwise distances computed. A grid of 100 values of ϕ is constructed, corresponding to evenly spaced values such that the minimum value of ϕ corresponds to a correlation of 0.99 at the maximum

observed distance, and the maximum value of ϕ corresponds to a correlation of 0.01. In every simulation, τ^2 is set to one. Four approaches to generating pairwise distances were considered: (1) the points x were evenly distributed on the interval $[0.001, 1]$ (Grid case); (2) the points x were sampled uniformly on the unit interval; (3) the points x were sampled from Gamma $(1, 1)$; and, (4) the points are vectors in $\mathbb{R}^5$ with independent standard normal entries. Naturally, the first three cases correspond to approximately low-rank Σ, while (4) corresponds to a Σ with a much more slowly decaying spectrum.

To assess the accuracy of the approximate partial eigendecomposition, both a complete eigendecomposition and an approximate partial eigendecomposition were computed, producing $\Sigma = U\Lambda U'$ and $\Sigma_\epsilon = U_\epsilon \Lambda_\epsilon U_\epsilon'$, where U_ϵ is $n \times m$ and Λ_ϵ $m \times m$ with $m < n$. The approximate partial eigendecomposition was computed using Algorithms 4.2 and 5.5 of Halko et al. (2011) with $\delta_\epsilon = 0.001$. Let Λ^* and U^* be the diagonal matrix consisting of the largest m eigenvalues of Σ and the corresponding m eigenvectors, respectively. We then compute

$$R(\Lambda^*, \Lambda_\epsilon) = \sqrt{\sum_{i=1}^{m} (\lambda_i^* - \lambda_{\epsilon,i})^2}, \quad F(U^*, U_\epsilon) = ||I - U_\epsilon' U^*||_F / \sqrt{n}, \quad \text{and}$$

$$C(U^*, U_\epsilon) = \text{Corr}\left(y, U_\epsilon (U_\epsilon' U_\epsilon)^{-1} U_\epsilon' y\right),$$

where $y = U^*\beta$ for $\beta_j \sim N(0, 1)$ a random $m \times 1$ vector with independent standard normal entries. Essentially, R measures the quality of the approximation to the eigenvalues and F and C measure the quality of approximation to the column space of U^*. Table E.4 shows results. For the Grid, Uniform, and Gamma cases, the approximation is extremely accurate; the approximate eigendecomposition is almost identical to the partial eigendecomposition. For the Normal case, the approximation to the eigenvalues is still very accurate, but there is noticeable error in the column space approximation. It should be noted that for the first three cases, typical values

Table E.4: Results of simulation study for approximation error using approximate eigendecomposition. The median, maximum, and minimum values of C, R, and F are shown across the 100 values of ϕ specified in the text.

	$R(\Lambda^*, \Lambda_\epsilon)$			$F(U^*, U_\epsilon)$		
	median	max	min	median	max	min
Grid	2.55e-12	3.562e-10	8.21e-16	1.54e-07	7.42e-15	0.00e+00
Uniform	1.70e-12	4.41e-10	1.04e-17	1.36e-07	1.19e-15	0.00e+00
Gamma	2.56e-12	1.41e-09	1.14e-16	2.60e-07	2.54e-15	0.00e+00
Normal	9.36e-09	1.24e-08	1.14e-09	3.87e-01	3.65e-02	0.00e+00

	$C(U^*, U_\epsilon)$		
	median	max	min
Grid	1.00	1.00	1.00
Uniform	1.00	1.00	1.00
Gamma	1.00	1.00	1.00
Normal	0.99	1.00	0.90

of m ranged from 10 to 50, whereas in the Normal case, m is nearly 500 for most values of ϕ. In general, we expect the approximate eigendecomposition to be less accurate in cases where the spectrum decays very slowly, so the results in Table E.4 are not surprising.

Appendix F

Appendix to Chapter 7

F.1 Proofs of Spectral Gap Results

F.1.1 Proof of Corollary F.1.1

The following Corollary to Theorem 7.1.5 is useful in our setting.

Corollary F.1.1. *Let (θ_t, Z_t) be a data-augmentation Markov chain on state space $\Omega_1 \times \Omega_2 \subset \mathbb{R} \times \mathbb{R}^n$. Denote by $\mathcal{P} = \mathcal{P}_1\mathcal{P}_2$ the transition kernel of this chain, where $\mathcal{P}_1((\theta, Z), \Omega_1 \times \{Z\}) = \mathcal{P}_2((\theta, Z), \{\theta\} \times \Omega_2) = 1$ for all $(\theta, Z) \in \Omega_1 \times \Omega_2$. Denote by Π the stationary measure of $\mathcal{P}$, and denote by Π_1 and Π_2 the marginals of this stationary measure on Ω_1 and Ω_2; denote by μ, μ_1 and μ_2 their densities. Assume that there exists an interval $I = (a, b) \subset \Omega_1$ that satisfies*

$$\pi_1(I) \geq 1 - \epsilon \tag{F.1}$$

$$c \leq \inf_{\theta \in I} \mu_1(\theta) \leq \sup_{\theta \in I} \mu_1(\theta) \leq C$$

$$\sup_{\theta \in I, z \in \Omega_2} \mathbb{P}\left[(\theta_{s+1} - \theta_s)^2 > r\delta \mid (\theta_s, Z_s) = (\theta, z)\right] \leq r^{-2} + \gamma$$

for some $\epsilon, \delta > 0$, some $0 \leq \gamma < \infty$ and some $0 < c < C < \infty$, and for all

$0 < r < \infty$. *Assume that* $\delta \leq \frac{1-\epsilon}{4C}$. *Then*

$$1 - \lambda_1(\mathcal{P}) \leq \frac{16C\delta}{(1-\epsilon)^2} + \frac{2C\gamma}{c(1-\epsilon)}.$$

Proof. Let $m = \inf\{x > a : \int_a^x \mu_1(y)dy \geq \frac{\pi_1(I)}{2}\} \geq a + \frac{1-\epsilon}{2C}$ be the median of the restriction of Π_1 to I and let $S = (a, m] \times \Omega_2$. Note that, by the second part of Inequality (F.1),

$$\frac{1-\epsilon}{2c} \geq m - a \geq \frac{1-\epsilon}{2C}.$$

We then calculate

$$\begin{aligned}
\kappa(S) &= \frac{\int_{(x,y)\in S} \mathcal{P}((x,y), S^c)\mu(x,y)dxdy}{\Pi(S)(1-\Pi(S))} \\
&\leq \frac{4}{(1-\epsilon)^2} \int_{(x,y)\in S} \mathcal{P}((x,y), S^c)\mu(x,y)dxdy \\
&\leq \frac{4}{(1-\epsilon)^2} \int_a^m C(\min(1, \frac{\delta^2}{\min(x-a, m-x)^2}) + \gamma)dx \\
&= \frac{8C}{(1-\epsilon)^2}(\int_0^\delta (1+\gamma)dx + \int_\delta^{\frac{m-a}{2}} (\frac{\delta^2}{x^2} + \gamma)dx) \\
&= \frac{8C}{(1-\epsilon)^2}(\delta + \delta^2(\delta^{-1} - \frac{2}{m-a}) + \gamma\frac{m-a}{2}) \\
&\leq \frac{16C\delta}{(1-\epsilon)^2} + \frac{8C\gamma}{(1-\epsilon)^2}\frac{1-\epsilon}{4c} \\
&= \frac{16C\delta}{(1-\epsilon)^2} + \frac{2C\gamma}{c(1-\epsilon)}.
\end{aligned}$$

The result now follows immediately from an application of Theorem 7.1.5. □

F.1.2 Proof of Theorem 7.2.1

We begin by proving inequality (7.10) with an application of Corollary F.1.1. The proof consists of verifying the three conditions given by inequality (F.1), beginning with the third.

This requires bounding from above the typical move size $|\theta_{t+1} - \theta_t|^2$. Note that our chain is uniformly ergodic Choi and Hobert (2013), and so moves can be very large if θ_t is far from its typical value of approximately $-\log(n)$. For that reason, we will fix a constant $0 < C < 1$ and will bound the move size $|\theta_{t+1} - \theta_t|^2$ only when θ_t satisfies $\theta_t = -\log(n)(1 + a_t)$ for some $|a_t| \leq C$. In this regime, we have

$$\begin{aligned}
\mathbb{E}[w_{t+1}] &= \frac{n}{2\theta_t}\tanh(\frac{\theta_t}{2}) \qquad \text{(F.2)} \\
&= \frac{n}{-2\log(n)(1+a_t)}\frac{1 - e^{\log(n)(1+a_t)}}{1 + e^{\log(n)(1+a_t)}} \\
&= \frac{n}{-2\log(n)(1+a_t)}\frac{1 - n^{1+a_t}}{1 + n^{1+a_t}} \\
&= \frac{n}{2\log(n)(1+a_t)}(1 - 2n^{-1-a_t}(1 - o(1)))
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[w_{t+1}] &= \frac{n}{4\theta_t^3}(\sinh(\theta_t) - \theta_t)\mathrm{sech}^2(\frac{\theta_t}{2}) \qquad \text{(F.3)} \\
&= \frac{-n}{4(1+a_t)^3\log(n)^3}(\frac{1 - e^{2(1+a_t)\log(n)}}{2e^{(1+a_t)\log(n)}} + (1+a_t)\log(n))(\frac{2e^{\frac{1}{2}(1+a_t)\log(n)}}{1 + e^{(1+a_t)\log(n)}})^2 \\
&= \frac{n}{4(1+a_t)^3\log(n)^3}(\frac{1}{2}n^{1+a_t}(1 + o(1)))(\frac{4}{n^{1+a_t}}(1 + o(1))) \\
&= \frac{n}{2(1+a_t)^3\log(n)^3}(1 + o(1)).
\end{aligned}$$

Combining inequalities (F.2) and (F.3), we have by Chebyshev's inequality that

$$\mathbb{P}\left[|w_{t+1} - \frac{n}{2\log(n)(1+a_t)}| > r\frac{\sqrt{n}}{\log(n)^{1.5}}\right] = O(r^{-2}) \qquad \text{(F.4)}$$

for any $r > 0$. Next, we estimate θ_{t+1}. Recall

$$\theta_{t+1}|w_{t+1} \sim \text{Normal}\left(\sigma_{w_{t+1}}^{-1}(y - n/2), \sigma_{w_{t+1}}^{-1}\right), \quad \sigma_{w_{t+1}}^{-1} = (w_{t+1} + B^{-1})^{-1}.$$

Define r_t by $w_{t+1} = \frac{n}{2\log(n)(1+a_t)} + r_t\frac{\sqrt{n}}{\log(n)^{1.5}}$. Conditional on $r_t\sqrt{\frac{4\log(n)}{n}} \leq \frac{1}{8}$ and $\frac{4B^{-1}\log(n)}{n} \leq \frac{1}{8}$, we have

$$\begin{aligned}
\sigma_{w_{t+1}}^{-1} &= (\frac{n}{2\log(n)(1+a_t)} + r_t\frac{\sqrt{n}}{\log(n)^{1.5}} + B^{-1})^{-1} \\
&= \frac{2\log(n)(1+a_t)}{n}(1 + B^{-1}\frac{2\log(n)(1+a_t)}{n} + r_t\frac{2(1+a_t)}{\sqrt{n\log(n)}})^{-1} \\
&= \frac{2\log(n)(1+a_t)}{n}(1 - O(\frac{r_t+1}{\sqrt{n\log(n)}})).
\end{aligned}$$

Thus, still conditional on $r_t\sqrt{\frac{4\log(n)}{n}} \leq \frac{1}{8}$ and $\frac{4B^{-1}\log(n)}{n} \leq \frac{1}{8}$,

$$\begin{aligned}
\theta_{t+1}|w_{t+1} &\sim \text{No}\left(\sigma_{w_{t+1}}^{-1}(y - n/2), \sigma_{w_{t+1}}^{-1}\right) \\
&= \text{No}\Bigg((2-n)\frac{\log(n)(1+a_t)}{n}(1 + O(\frac{r_t+1}{\sqrt{n\log(n)}})), \\
&\quad \frac{\log(n)(1+a_t)}{n}(1 + O(\frac{r_t+1}{\sqrt{n\log(n)}}))\Bigg) \\
&= \text{No}\Bigg(-\log(n)(1+a_t)(1 + O(\frac{r_t+1}{\sqrt{n\log(n)}})), \\
&\quad \frac{\log(n)(1+a_t)}{n}(1 + O(\frac{r_t+1}{\sqrt{n\log(n)}}))\Bigg) \\
&= \text{No}\left(\theta_t(1 + O(\frac{r_t+1}{\sqrt{n\log(n)}})), \frac{\log(n)(1+a_t)}{n}(1 + O(\frac{r_t+1}{\sqrt{n\log(n)}}))\right).
\end{aligned}$$

Combining this bound with inequality (F.4), we conclude that for fixed $r > 0$

$$\begin{aligned}
\mathbb{P}\left[|\theta_{t+1} - \theta_t| > 2r\sqrt{\frac{\log(n)}{n}}\right] &\leq \mathbb{P}\left[|w_{t+1} - \frac{n}{2\log(n)(1+a_t)}| > r\frac{\sqrt{n}}{\log(n)^{1.5}}\right] \quad \text{(F.5)} \\
&\quad + \mathbb{P}\Bigg[|\theta_{t+1} - \theta_t| > 2r\sqrt{\frac{\log(n)}{n}}\ \Bigg| \\
&\quad |w_{t+1} - \frac{n}{2\log(n)(1+a_t)}| \leq r\frac{\sqrt{n}}{\log(n)^{1.5}}\Bigg]
\end{aligned}$$

$$= O(r^{-2}) + O(\sqrt{\frac{\log(n)}{n}}).$$

Thus, the third part of inequality (F.1) is satisfied for two sequences of constants $\delta = \delta(n)$ and $\gamma = \gamma(n)$ that satisfy

$$\delta(n) = O(\sqrt{\frac{\log(n)}{n}}),\ \gamma(n) = O(\sqrt{\frac{\log(n)}{n}}) \tag{F.6}$$

on any sequence of sets $I = I(n)$ satisfying $I(n) \subset (-\log(n)(1+\zeta), -\log(n)(1-\zeta))$ and fixed $0 < \zeta < 1$.

Next, we must provide bounds for the first two parts of inequality (F.1). Recall that the posterior density of θ is

$$p(\theta|y=1) = \frac{n}{\sqrt{2\pi}B}(1+e^{\theta})^{-n}e^{\theta}e^{-\frac{\theta^2}{2B}}.$$

We will show that $p(\theta|y=1)$ is roughly constant on a region of size roughly $\frac{1}{\log(n)}$ and that it is negligible outside of a region of size roughly $\log(n)$.

We begin by showing that $p(\theta|y=1)$ is near-constant on a small region around the mode $\theta_{\max} \equiv \text{argmax}_{\theta} p(\theta|y=1)$. By straightforward calculus, $\theta_{\max}$ satisfies

$$\frac{\theta_{\max}}{B} + n\frac{e^{\theta_{\max}}}{1+e^{\theta_{\max}}} = 1,$$

and so $\theta_{\max} = -\log(n) + O(\log(\log(n)))$.

Fix θ_1, θ_2 that satisfy $|\theta_1 - \theta_2| \leq \frac{1}{\log(n)}$ and also $|\theta_1 + \log(n)|, |\theta_2 + \log(n)| \leq A\log(\log(n))$ for some $0 < A < \infty$. Define δ_1, δ_2 by $\theta_1 = -\log(n) + \delta_1$, $\theta_2 = -\log(n) + \delta_2$. Then we calculate

$$\begin{aligned}
\frac{p(\theta_1|y=1)}{p(\theta_2|y=1)} &= e^{\theta_1-\theta_2}\left(\frac{1+e^{\theta_2}}{1+e^{\theta_1}}\right)^n e^{\frac{1}{2B}(\theta_2^2-\theta_1^2)} \\
&= e^{\delta_1-\delta_2}\left(\frac{1+\frac{1}{n}e^{\delta_2}}{1+\frac{1}{n}e^{\delta_1}}\right)^n e^{\frac{1}{2B}(\delta_1-\delta_2)(2\log(n)-\delta_1-\delta_2)} \\
&\geq (e^{-2})(2e)^{-2A}(e^{\frac{-2}{B}}).
\end{aligned} \tag{F.7}$$

Next, we show that $p(\theta|y=1)$ is negligible outside of the interval $(-5\log(n), 3\log(n))$. If $\theta = -\log(n) + C\log(n)$ for some $C \geq 4$,

$$\begin{aligned} p(\theta|y=1) &\leq \frac{n}{\sqrt{2\pi B}} n^{C-1}(1+n^{C-1})^{-n} e^{-\frac{(C-1)^2\log(n)^2}{2B}} \\ &\leq \frac{1}{\sqrt{2\pi B}} n^{C-n(C-1)-\frac{(C-1)^2}{2B}\log(n)}. \end{aligned}$$

Thus,

$$\int_{3\log(n)}^{\infty} p(\theta|y=1)d\theta \leq \sum_{C=4}^{\infty} \log(n) \times \frac{1}{\sqrt{2\pi B}} n^{C-n(C-1)-\frac{(C-1)^2}{2B}\log(n)} = o(1). \quad \text{(F.8)}$$

If $\theta = -\log(n) - C\log(n)$ for some $C \geq 4$, then

$$\begin{aligned} p(\theta|y=1) &\leq \frac{n}{\sqrt{2\pi B}} n^{-C-1}(1+n^{-C-1})^{-n} e^{-\frac{(C+1)^2\log(n)^2}{2B}} \\ &\leq \frac{2}{\sqrt{2\pi B}} n^{-C-\frac{(C+1)^2}{2B}\log(n)}. \end{aligned}$$

Thus,

$$\int_{3\log(n)}^{\infty} p(\theta|y=1)d\theta \leq \sum_{C=4}^{\infty} \log(n) \times \frac{2}{\sqrt{2\pi B}} n^{-C-\frac{(C+1)^2}{2B}\log(n)} = o(1). \quad \text{(F.9)}$$

Combining inequalities (F.8) and (F.9) gives

$$\int_{(-5\log(n),3\log(n))^c} p(\theta|y=1)d\theta = o(1). \quad \text{(F.10)}$$

By inequalities (F.7) and (F.10), the first two parts of inequality (F.1) are satisfied with $\epsilon = \epsilon(n)$, $c = c(n)$ and $C = C(n)$ satisfying

$$(1-\epsilon(n))^{-1} = O(\log(n)^2),\ c(n) = \Theta(1),\ C(n) = \Theta(1) \quad \text{(F.11)}$$

and a set $I(n) \subset (-\log(n) - \frac{1}{\log(n)} - \eta_n, -\log(n) + \frac{1}{\log(n)} - \eta_n)$ for some $\eta_n = O(\log(\log(n)))$. Combining this with (F.6) and Corollary F.1.1 completes the proof of Equation (7.10).

Finally, Equality (7.11) follows immediately from inequalities (F.7) and (F.10). This completes the proof of the Theorem.

F.1.3 Proof of Theorem 7.2.2

First we give a lemma that is used in the main proof to bound $\Phi^{-1}(x)$ and $(\Phi^{-1}(x))^2$.

Lemma F.1.1. *Let $\Phi(\cdot)$ be the standard normal distribution function and fix $x > 0$. Then, taking the asymptotic as $n \to \infty$,*

$$|\Phi^{-1}\left(\frac{x}{n}\right) + \sqrt{2\log\left(\frac{n}{x}\right)} - \frac{\log\left(\log\left(\frac{n}{x}\right)\right)}{2\sqrt{2\log\left(\frac{n}{x}\right)}}| = \mathcal{O}\left(\frac{1}{(\log(n/x))^{1.5}}\right). \tag{F.12}$$

Furthermore,

$$\left(\Phi^{-1}\left(\frac{x}{n}\right)\right)^2 = 2\log\left(\frac{n}{x}\right) - \log\left(2\log\left(\frac{n}{x}\right)\right) - \log(2\pi) + \mathcal{O}\left(\frac{1}{(\log(n/x))}\right) \tag{F.13}$$

Proof. From equation 7.1.13 of Olver (2010) we have for $x > 0$

$$\frac{1}{x + \sqrt{x^2+4}} \leq \sqrt{2\pi}e^{\frac{x^2}{2}}(1 - \Phi(x)) \leq \frac{1}{x + \sqrt{x^2 + \frac{8}{\pi}}}. \tag{F.14}$$

Thus, we can write

$$\sqrt{2\pi}e^{x^2/2}(1 - \Phi(x)) = \frac{1}{x + \sqrt{x^2 + h(x)}}$$

for some function $h(x)$ that satisfies $\frac{8}{\pi} \leq h(x) \leq 4$ for all $x > 0$. Now

$$(1 - \Phi(x)) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{x + \sqrt{x^2 + h(x)}}$$

$$\Phi(x) = 1 - \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{x + \sqrt{x^2 + h(x)}}.$$

Writing $y = \Phi(x)$ and inverting gives

$$1 - y = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}(x + \sqrt{x^2 + h(x)})^{-1} \tag{F.15}$$

$$\log(1-y) = -\frac{1}{2}\log(2\pi) - \frac{x^2}{2} - \log(x + \sqrt{x^2 + h(x)})$$

$$x^2 = -2\log(1-y) - \log(2\pi) - 2\log(x + \sqrt{x^2 + h(x)}).$$

We now claim that for any fixed $\epsilon > 0$ and any sufficiently large $x > X(\epsilon)$, we have $\sqrt{-(2-\epsilon)\log(1-y)}x < \sqrt{-(2+\epsilon)\log(1-y)}$. To see this, recall that Inequality (F.14) clearly implies that, for any fixed $\epsilon > 0$,

$$\frac{1}{\sqrt{2\pi}} e^{-(1+\epsilon)x^2/2} \leq 1 - \Phi(x) \leq \frac{1}{\sqrt{2\pi}} e^{-(1-\epsilon)x^2/2}$$

for all sufficiently large x. Substituting this bound into (F.15) we obtain

$$x^2 = -2\log(1-y) - \log(2\pi) - \log(-2\log(1-y)) + \mathcal{O}\left(\frac{1}{-\log(1-y)}\right),$$

which gives

$$x = \sqrt{-2\log(1-y) - \log(2\pi) - \log(-2\log(1-y)) + \mathcal{O}\left(\frac{1}{-\log(1-y)}\right)}$$

$$= \sqrt{-2\log(1-y)}\left(1 - \frac{\log(-2\log(1-y))}{\sqrt{-2\log(1-y)}} + \mathcal{O}\left(\frac{1}{(-\log(1-y))^{1.5}}\right)\right),$$

and therefore

$$\left|x - \sqrt{-2\log(1-y)}\right| = -\frac{\log(-2\log(1-y))}{\sqrt{-2\log(1-y)}} + \mathcal{O}\left(\frac{1}{(-\log(1-y))^{1.5}}\right).$$

Putting $1 - y = x/n$ for $x/n < 1/2$ we have

$$(\Phi^{-1}(x/n))^2 = 2\log(n/x) - \log(2\pi) - \log(2\log(n/x)) + o(1)$$

$$\Phi^{-1}(x/n) = \sqrt{2\log(n/x)} - \frac{\log(2\log(n/x))}{\sqrt{2\log(n/x)}} + \mathcal{O}\left(\frac{1}{(\log(n/x))^{1.5}}\right),$$

completing the proof. □

Proof of main result

The structure of this proof is quite similar to the proof of Theorem 7.2.1. We start by proving Equation (7.12). The expectation and variance of Z_{t+1} given β_t are

$$\begin{aligned}\mathbb{E}[Z_{t+1} \mid \beta_t] &= (n-1)\left(\beta_t - \frac{\phi(\beta_t)}{1-\Phi(\beta_t)}\right) + \left(\beta_t + \frac{\phi(\beta_t)}{\Phi(\beta_t)}\right) \\ &= n\beta_t - (n-1)\frac{\phi(\beta_t)}{1-\Phi(\beta_t)} + \frac{\phi(\beta_t)}{\Phi(\beta_t)}\end{aligned}$$

$$\mathrm{Var}[Z_{t+1} \mid \beta_t] = v_t = n + (n-1)\left(\beta_t \frac{\phi(\beta_t)}{1-\Phi(\beta_t)} - \left(\frac{\phi(\beta_t)}{1-\Phi(\beta_t)}\right)^2\right) - \beta_t \frac{\phi(\beta_t)}{\Phi(\beta_t)} - \frac{\phi(\beta_t)^2}{\Phi(\beta_t)^2}.$$

Now, for n large enough, using (F.13)

$$\begin{aligned}\phi_B\left(\Phi^{-1}\left(\frac{x}{n}\right)\right) &= \frac{1}{\sqrt{2\pi B}} \exp\left(-\frac{2\log(n/x)}{2B} + \frac{1}{2B}\log\left(2\log\left(\frac{n}{x}\right)\right) + \frac{1}{2B}\log(2\pi) + o(1)\right) \\ &= \frac{1}{\sqrt{2\pi B}}\left(\frac{x}{n}\right)^{1/B}\left(\sqrt{2\log(n/x)}\right)^{1/B}(\sqrt{2\pi})^{1/B}\exp(o(1)) \qquad \text{(F.16)}\end{aligned}$$

where ϕ_B is the density of $N(0, B)$.

We now compute the posterior mode $\hat{\beta}$. We begin by reparameterizing our problem by the one-to-one transformation $\beta = \Phi^{-1}(x/n)$. We will compute $\hat{x}$, the posterior mode under this transformation, and then use this to compute the true posterior mode $\hat{\beta}$ by the equation $\hat{\beta} = \Phi^{-1}(\hat{x}/n)$.

The posterior density when $y = 1$ is proportional to

$$p(\beta|n, y) \propto n(\Phi(\beta))(1-\Phi(\beta))^{n-1}\phi_B(\beta).$$

Under our reparameterization,

$$p(x|n, y) \propto x\left(1-\frac{x}{n}\right)^{n-1}\frac{1}{\sqrt{2\pi B}}\left(\frac{x}{n}\right)^{1/B}\sqrt{\log(n/x)}\exp(o(1)).$$

Also, since

$$\log p(x \mid n, y) \propto \log x + (n-1)\log\left(1-\frac{x}{n}\right) - \frac{(\Phi^{-1}(x/n))^2}{2B}$$

we have

$$\frac{\partial}{\partial x}\log p(x \mid n, y) \propto \frac{1}{x} - \frac{n-1}{n-x} - \frac{\sqrt{2\pi}}{Bn}\exp((\Phi^{-1}(x/n))^2/2)\Phi^{-1}(x/n).$$

Combining this with (F.16) and (F.12), we find

$$\begin{aligned}\frac{\partial}{\partial x}\log p(x \mid n, y) &= \frac{1}{x} - \frac{n-1}{n-x} - \frac{\sqrt{2\pi}}{Bn}\left(\frac{n}{x}\right)\frac{1}{\sqrt{2\log(n/x)}}\frac{1}{\sqrt{2\pi}} \\ &\quad \times \exp(o(1))\left(-\sqrt{2\log\left(\frac{n}{x}\right)} + o(1)\right) \\ &= \frac{1}{x} - \frac{n-1}{n-x} + \frac{1}{Bx}\left(1 + o(1)\right),\end{aligned}$$

so in the limit as $n \to \infty$ the posterior mode is

$$\frac{\widehat{x}}{n} = \frac{B + 1 + o(1)}{Bn + 1}. \tag{F.17}$$

In particular, for large enough n the mode is less than $2/n$.

A region outside of which the posterior is negligible Now we show an interval outside of which the posterior is negligible. Fix $C > 2$ and consider the interval $[\Phi^{-1}(C/n^2), \Phi^{-1}(C/\sqrt{n})]$. Assume $B \geq 1$ and set $\theta = \Phi^{-1}(C/\sqrt{n})$; applying Lemma F.1.1, we have for $n > N(C)$ sufficiently large that

$$\begin{aligned}p(\theta \mid y = 1) &= n\frac{C}{\sqrt{n}}\left(1 - \frac{C}{\sqrt{n}}\right)^{n-1}\phi_B(\Phi^{-1}(x/n)) \\ &= \frac{n\sqrt{n}}{\sqrt{n} - C}\frac{C}{\sqrt{n}}\left(1 - \frac{C}{\sqrt{n}}\right)^{n}\frac{1}{\sqrt{2\pi B}}\left(\frac{C}{\sqrt{n}}\right)^{1/B} \\ &\quad \times \left(\sqrt{2\log(\sqrt{n}/C)}\right)^{1/B}(\sqrt{2\pi})^{1/B}\exp(o(1)) \\ &\leq \frac{nC}{\sqrt{n} - C}\exp(-C\sqrt{n})\left(\frac{C}{\sqrt{n}}\right)^{1/B}\left(\sqrt{2\log(\sqrt{n}/C)}\right)^{1/B}\exp(o(1)) \\ &\leq \sqrt{n}C^{2+1/B}\exp(-C\sqrt{n})\left(\frac{C}{\sqrt{n}}\right)^{1/B}\left(\sqrt{2\log(\sqrt{n}/C)}\right)^{1/B},\end{aligned}$$

where in the last step we chose n large enough that $\exp(o(1)) < C$.

Since

$$\lim_{n\to\infty} \frac{-\sqrt{-2\log C/\sqrt{n}} + \sqrt{2\log C/n}}{\sqrt{\log n}} = \sqrt{2} - 1,$$

we have

$$\int_{\Phi^{-1}(2/\sqrt{n})}^{\lfloor\sqrt{n}\rfloor} p(\theta \mid y = 1) \leq \frac{1}{2}\sqrt{2\log(\sqrt{n})}\sqrt{\log n}\sqrt{n} \sum_{C=3}^{\lfloor\sqrt{n}\rfloor} C^{2+1/B} \exp(-C\sqrt{n}) = o(1).$$

Similarly, for $\theta = \Phi^{-1}(1/(Cn^2))$

$$\begin{aligned}
p(\theta \mid y = 1) &= n\frac{1}{Cn^2}\left(1 - \frac{1}{Cn^2}\right)^{n-1} \phi_B(\Phi^{-1}(1/(Cn^2))) \\
&= n\frac{Cn^2}{Cn^2-1}\frac{1}{Cn^2}\left(1 - \frac{1}{Cn^2}\right)^{n} \frac{1}{\sqrt{2\pi B}}\left(\frac{1}{Cn^2}\right)^{1/B} \\
&\times \left(\sqrt{2\log(Cn^2)}\right)^{1/B} (\sqrt{2\pi})^{1/B} \exp(o(1)) \\
&\leq \frac{n}{Cn^2-1}\left(\frac{1}{Cn^2}\right)^{1/B} e^{-1/(Cn)} \frac{(2\pi)^{1/B}}{\sqrt{2\pi B}}\left(\sqrt{2\log(Cn^2)}\right)^{1/B} \exp(o(1)) \\
&\leq \frac{1}{Cn-1/n}\left(\frac{1}{Cn^2}\right)^{1/B} \frac{(2\pi)^{1/B}}{\sqrt{2\pi B}}\left(\sqrt{2\log(Cn^2)}\right)^{1/B} \exp(o(1)) \\
&\leq \frac{2}{Cn-1/n}\left(\frac{1}{Cn^2}\right)^{1/B} \frac{(2\pi)^{1/B}}{\sqrt{2\pi B}}\left(\sqrt{2\log(Cn^2)}\right)^{1/B}
\end{aligned}$$

when n is large enough that $e^{o(1)} < 2$ and $B \geq 1$. And since

$$\lim_{n\to\infty} \frac{\Phi^{-1}(C/n) - \Phi^{-1}(1/(Cn^2))}{\sqrt{\log n}} = 2 - \sqrt{2}$$

we have for large enough n that $\int_{-\infty}^{\Phi^{-1}(1/(2n^2))} p(\theta \mid y = 1)$

$$\leq \frac{\sqrt{\log n}}{n^{2/B}} \sum_{C=3}^{\infty} \frac{2}{Cn-1/n}\left(\frac{1}{C}\right)^{1/B} \frac{(2\pi)^{1/B}}{\sqrt{2\pi B}}\left(\log(C^2) + \log(n^4)\right)^{1/(2B)},$$

$$\leq \frac{\sqrt{\log n}}{n^{2/B}} \sum_{C=3}^{\infty} \frac{2}{Cn - 1/n} \left(\frac{1}{C}\right)^{1/B} \frac{(2\pi)^{1/B}}{\sqrt{2\pi B}} \left((\log(C^2))^{1/(2B)} + (\log(n^4))^{1/(2B)}\right)$$

$$\leq \frac{\sqrt{\log n}}{n^{1+2/B}} \frac{(2\pi)^{1/B}}{\sqrt{2\pi B}} \sum_{C=3}^{\infty} \frac{2}{C-1} \left(\frac{1}{C}\right)^{1/B} \left((\log(C^2))^{1/(2B)} + (\log(n^4))^{1/(2B)}\right).$$

Now, there exists a constant C_1 such that $(\log(C^2))^{1/(2B)} < C_1 C^{1/(2B)}$, giving

$$\int_{-\infty}^{\Phi^{-1}(1/(2n^2))} p(\theta \mid y = 1)$$

$$\leq \frac{\sqrt{\log n}}{n^{1+2/B}} \frac{C_1 (2\pi)^{1/B}}{\sqrt{2\pi B}} \sum_{C=3}^{\infty} \frac{2}{C-1} \left(\frac{1}{C}\right)^{1/(2B)} \left(1 + (4\log n)^{1/(2B)}\right) = o(1)$$

We conclude

$$\int_{[\Phi^{-1}(1/(Cn^2)), \Phi^{-1}(C/\sqrt{n})]^c} p(\theta \mid y = 1) d\theta = o(1). \tag{F.18}$$

for $C \geq 2$.

An interval on which the posterior is almost constant

We now fix a (new) constant $2 < C < \frac{Bn+1}{B+1}$ and show that the posterior is almost constant on the interval $\left[\Phi^{-1}\left(\frac{B+1}{C(Bn+1)}\right), \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right]$. As shown in Equality (F.17), this interval includes the posterior mode for all large enough n. This interval has width $\mathcal{O}\left(\frac{1}{\sqrt{\log(n)}}\right)$ since

$$\lim_{n\to\infty} \sqrt{\log(n)} \left| -\sqrt{-2\log\left(\frac{C(B+1)}{Bn+1}\right)} + \sqrt{-2\log\left(\frac{B+1}{Bn+1}\right)} \right|$$

$$= \frac{\log(C(B+1)/B) - \log((B+1)/B)}{\sqrt{2}}. \tag{F.19}$$

Repeatedly applying Lemma F.1.1, we estimate the likelihood ratio

$$\frac{p\left(y = 1 \mid \theta = \Phi^{-1}\left(\frac{B+1}{Bn+1}\right)\right)}{p\left(y = 1 \mid \theta = \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right)} = \frac{n\left(\frac{B+1}{Bn+1}\right)\left(1 - \left(\frac{B+1}{Bn+1}\right)\right)^{n-1} \phi_B\left(\frac{B+1}{Bn+1}\right)}{n\left(\frac{C(B+1)}{Bn+1}\right)\left(1 - \left(\frac{C(B+1)}{Bn+1}\right)\right)^{n-1} \phi_B\left(\frac{C(B+1)}{Bn+1}\right)}$$

$$
\begin{aligned}
&= \frac{\left(\frac{B+1}{Bn+1}\right)\left(\frac{Bn+1}{Bn-B}\right)\left(1-\frac{B+1}{Bn+1}\right)^n \phi_B\left(\frac{B+1}{Bn+1}\right)}{\left(\frac{C(B+1)}{Bn+1}\right)\left(\frac{Bn+1}{Bn-BC}\right)\left(1-\frac{C(B+1)}{Bn+1}\right)^n \phi_B\left(\frac{C(B+1)}{Bn+1}\right)} \\
&= \frac{(n-C)\left(1-\frac{B+1}{Bn+1}\right)^n \phi_B\left(\frac{B+1}{Bn+1}\right)}{C(n-1)\left(1-\frac{C(B+1)}{Bn+1}\right)^n \phi_B\left(\frac{C(B+1)}{Bn+1}\right)} \\
&= \left(\frac{1}{C}+o(1)\right)\left(e^{(B+1)(C-1)/B}+o(1)\right)\frac{\phi_B\left(\frac{B+1}{Bn+1}\right)}{\phi_B\left(\frac{C(B+1)}{Bn+1}\right)} \\
&= \left(\frac{1}{C}+o(1)\right)\left(e^{(B+1)(C-1)/B}+o(1)\right) \\
&\times \frac{\left(\frac{B+1}{Bn+1}\right)^{1/B}\left(\sqrt{2\log\left(\frac{Bn+1}{B+1}\right)}\right)^{1/B} e^{o(1)}}{\left(\frac{C(B+1)}{Bn+1}\right)^{1/B}\left(\sqrt{2\log\left(\frac{Bn+1}{C(B+1)}\right)}\right)^{1/B} e^{o(1)}} \\
&= \left(\frac{1}{C}+o(1)\right)\left(e^{(B+1)(C-1)/B}+o(1)\right)\left(\frac{1}{C}\right)^{1/B} \\
&\times \left(\frac{\log\left(\frac{Bn+1}{B+1}\right)}{\log\left(\frac{Bn+1}{C(B+1)}\right)}\right)^{1/(2B)} e^{o(1)} \\
&= \left(\frac{1}{C}+o(1)\right)\left(e^{(B+1)(C-1)/B}+o(1)\right)\left(\frac{1}{C}\right)^{1/B} \\
&\times \left(\frac{\log(Bn+1)}{\log(Bn+1)-\log(C(B+1))}-o(1)\right)^{1/(2B)} e^{o(1)}.
\end{aligned}
$$

This means that, for all n sufficiently large,

$$
\frac{e^{(B+1)(C-1)/B}}{2C^{1+1/B}} \leq \frac{p\left(y=1 \mid \theta=\Phi^{-1}\left(\frac{B+1}{Bn+1}\right)\right)}{p\left(y=1 \mid \theta=\Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right)} \leq \frac{2e^{(B+1)(C-1)/B}}{C^{1+1/B}}. \qquad \text{(F.20)}
$$

Noting that we would obtain the same result replacing C with its reciprocal on the interval $\left[\Phi^{-1}\left(\frac{B+1}{C(Bn+1)}\right), \Phi^{-1}\left(\frac{B+1}{Bn+1}\right)\right]$, and combining with (F.19), this implies the posterior is almost constant on an interval of width $\mathcal{O}\left(\frac{1}{\sqrt{\log(n)}}\right)$.

Bounding typical move sizes.

Following the proof of Theorem 7.2.1, write $\beta_t = \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)$, with $C \in [\frac{1}{C^*}, C^*]$ for $2 < C^* < \frac{Bn+1}{B+1}$, in other words, we have β_t inside the interval $\left[\Phi^{-1}\left(\frac{B+1}{C(Bn+1)}\right), \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right]$. We do this so that we can write $\beta_t = \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)$ and make clear that this includes values of β_t on either side of the mode for large enough n. The term $\phi\left(\Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right)$ will appear often. We have that

$$\phi\left(\Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right) = \mathcal{O}\left(\frac{\sqrt{2\log(Bn+1)}}{Bn+1}\right)$$

by (F.16).

So for the definition of β_t above we have

$$\begin{aligned}
\mathbb{E}\left[Z_{t+1}/n \mid \beta_t\right] &= \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right) - \phi\left(\Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right) \\
&\quad \times \left[\frac{(n-1)}{n}\frac{1}{1-\Phi\left(\Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right)} - \frac{1}{n\Phi\left(\Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right)}\right] \\
&= \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right) - \phi\left(\Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right) \\
&\quad \times \left[\frac{(n-1)}{n}\frac{Bn+1}{Bn+1-C(B+1)} - \frac{Bn+1}{nC(B+1)}\right] \\
&= \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right) - \mathcal{O}\left(\frac{\sqrt{2\log(Bn+1)}}{Bn+1}\right)\mathcal{O}(1) \\
&= \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right) + o(1)
\end{aligned}$$

Now

$$\mathrm{Var}\left[\frac{Z_{t+1}}{n} \;\middle|\; \beta_t\right] = \frac{1}{n} + \frac{n-1}{n^2}\left(\beta_t\frac{\phi(\beta_t)}{1-\Phi(\beta_t)} - \left(\frac{\phi(\beta_t)}{1-\Phi(\beta_t)}\right)^2\right) - \frac{\beta_t}{n^2}\frac{\phi(\beta_t)}{\Phi(\beta_t)} - \frac{\phi(\beta_t)^2}{n^2\Phi(\beta_t)^2}$$

$$
\begin{aligned}
&= \frac{1}{n} + \beta_t \phi(\beta_t) \left(\frac{n-1}{n^2} \frac{1}{1-\Phi(\beta_t)} - \frac{1}{n^2} \frac{1}{\Phi(\beta_t)} \right) \\
&- \phi(\beta_t)^2 \left(\frac{n-1}{n^2(1-\Phi(\beta_t))^2} - \frac{1}{n^2 \Phi(\beta_t)^2} \right) \\
&= \frac{1}{n} + \beta_t \phi(\beta_t) \left(\frac{n-1}{n^2} \frac{Bn+1}{Bn+1-C(B+1)} - \frac{1}{n^2} \frac{Bn+1}{C(B+1)} \right) \\
&- \phi(\beta_t)^2 \left(\frac{n-1(Bn+1)^2}{n^2(Bn+1-C(B+1))^2} - \frac{(Bn+1)^2}{n^2(C(B+1))^2} \right) \\
&= \frac{1}{n} + \beta_t \phi(\beta_t) \, \mathcal{O}(1) - \phi(\beta_t)^2 \, \mathcal{O}(1) \\
&= \frac{1}{n} + \Phi^{-1} \left(\frac{C(B+1)}{Bn+1} \right) \mathcal{O}\left(\frac{\sqrt{2\log(Bn+1)}}{Bn+1} \right) + \mathcal{O}\left(\frac{2\log(Bn+1)}{(Bn+1)^2} \right) \\
&= \frac{1}{n} + \left(\sqrt{2 \log \left(\frac{Bn+1}{C(B+1)} \right)} + \mathcal{O}\left(\frac{\log\left(2\log\left(Bn+1\right)\right)}{\sqrt{2\log\left(Bn+1\right)}} \right) \right) \\
&+ \mathcal{O}\left(\frac{\sqrt{2\log(Bn+1)}}{Bn+1} \right) + \mathcal{O}\left(\frac{2\log(Bn+1)}{(Bn+1)^2} \right) \\
&= \frac{1}{n} + \mathcal{O}\left(\frac{2\log(Bn+1)}{Bn+1} \right) \\
&+ \mathcal{O}\left(\frac{\log 2\log(Bn+1)}{Bn+1} \right) + \mathcal{O}\left(\frac{2\log(Bn+1)}{(Bn+1)^2} \right) \\
&= \mathcal{O}\left(\frac{\log n}{n} \right)
\end{aligned}
$$

Now, we have that $\beta_t = \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)$ and want to show an upper bound on $\mathbb{P}\left[|\beta_t - \beta_{t+1}| > r\delta\right]$. Our strategy is to show a lower bound on $\mathbb{P}\left[|\beta_t - \beta_{t+1}| < r\delta\right]$ for fixed $r, \delta > 0$. By the triangle inequality,

$$
|\beta_t - \beta_{t+1}| < \left| \beta_t - \frac{Z_{t+1}}{n} \right| + \left| \frac{Z_{t+1}}{n} - \beta_{t+1} \right|.
$$

It follows that, for all $r, \delta > 0$,

$$\begin{aligned}
\mathbb{P}\left[|\beta_t - \beta_{t+1}| < r\delta\right] &\geq \mathbb{P}\left[\left|\beta_t - \frac{Z_{t+1}}{n}\right| < \frac{r\delta}{2}, \left|\frac{Z_{t+1}}{n} - \beta_{t+1}\right| < \frac{r\delta}{2}\right] \\
&\geq \mathbb{P}\left[\left|\beta_t - \frac{Z_{t+1}}{n}\right| < \frac{r\delta}{2}\right] \\
&\quad \times \mathbb{P}\left[|Z_{t+1}/n - \beta_{t+1}| < \frac{r\delta}{2} \quad \middle| \quad |\beta_t - Z_{t+1}/n| < \frac{r\delta}{2}\right]
\end{aligned}$$

Since $\beta_t = \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)$, the first term on the right side is

$$\mathbb{P}\left[\left|\Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right) - \frac{Z_{t+1}}{n}\right| < \frac{r\delta}{2}\right].$$

Putting $\delta = \frac{\sqrt{\log n}}{\sqrt{n}}$ and recognizing that the distribution of $Z_{t+1} \mid \beta_t$ is sub-Gaussian, we have

$$\mathbb{P}\left[\left|\frac{Z_{t+1}}{n} - \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right| > \frac{r\sqrt{\log n}}{2\sqrt{n}}\right] \leq e^{-r^2(1+o(1))/8}. \tag{F.21}$$

Now for the second term, recall

$$\begin{aligned}
\theta_{t+1} \mid Z_{t+1} &\sim N((n + B^{-1})^{-1} Z_{t+1}, (n + B^{-1})^{-1}) \\
&\sim N\left(\frac{n}{(n + B^{-1})}\frac{Z_{t+1}}{n}, (n + B^{-1})^{-1}\right).
\end{aligned}$$

So then the following holds uniformly for *any* Z_{t+1},

$$\begin{aligned}
\mathbb{P}\left[|Z_{t+1}/n - \beta_{t+1}| > \frac{r\sqrt{\log n}}{2\sqrt{n}} + \left|\frac{Z_{t+1}}{n} - \frac{Z_{t+1}}{n + B^{-1}}\right|\right] &\leq e^{-(r^2 \log n/8)(1-o(1))} \\
\mathbb{P}\left[|Z_{t+1}/n - \beta_{t+1}| > \frac{r\sqrt{\log n}}{2\sqrt{n}}(1 + o(1))\right] &\leq e^{-(r^2 \log n/8)(1-o(1))}, \qquad \text{(F.22)}
\end{aligned}$$

so in particular this holds conditional on $|\beta_t - Z_{t+1}/n| < \frac{r\sqrt{\log n}}{2\sqrt{n}}$. So then putting together (F.21) and (F.22) we have

$$\mathbb{P}\left[|\beta_t - \beta_{t+1}| < r\frac{\sqrt{\log n}}{\sqrt{n}}\right] \geq (1 - e^{-(r^2 \log n/8)(1-o(1))})(1 - e^{-r^2/8(1-o(1))})$$

so

$$\mathbb{P}\left[|\beta_{t+1} - \beta_t| > \frac{r\sqrt{\log n}}{\sqrt{n}}\right] = \mathcal{O}\Big(e^{-r^2/8}\Big),$$

So then for the interval $\left[\Phi^{-1}\left(\frac{B+1}{C(Bn+1)}\right), \Phi^{-1}\left(\frac{C(B+1)}{Bn+1}\right)\right]$, we have $1-\epsilon = \mathcal{O}\big(1/\sqrt{\log n}\big)$, and $c = C = \mathcal{O}(1)$. So we have $1 - \lambda_1(K) = \frac{(\log n)^2}{\sqrt{n}}$.

Finally, we prove Inequality (7.13). Combining inequalities (F.17) and (F.20) with Lemma F.1.1, we have shown that the mode is contained within an interval of length $\Theta(\sqrt{\log(n)})$ for which the density is $\Theta(1)$. Combining inequality (F.18) with Lemma F.1.1, we have shown that the posterior distribution is negligible outside of an interval of length $\Theta(\sqrt{\log(n)})$. Inequality (7.13) follows immediately.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964), *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, no. 55, Courier Corporation.

Adamczak, R., Litvak, A., Pajor, A., and Tomczak-Jaegermann, N. (2010), "Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles," *Journal of the American Mathematical Society*, 23, 535–561.

Agresti, A. (2002), *Categorical data analysis*, vol. 359, John Wiley & Sons.

Ahn, S., Korattikara, A., and Welling, M. (2012), "Bayesian posterior sampling via stochastic gradient Fisher scoring," *arXiv preprint arXiv:1206.6380*.

Albert, J. H. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American statistical Association*, 88, 669–679.

Alquier, P., Friel, N., Everitt, R., and Boland, A. (2014), "Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels," *Statistics and Computing*, 25, 1–19.

Anderson, T. W. (1954), "On estimation of parameters in latent structure analysis," *Psychometrika*, 19, 1–10.

Athreya, K. and Ney, P. (1978), "A new approach to the limit theory of recurrent Markov chains," *Trans. Amer. Math. Soc*, 245, 493–501.

Attias, H. (1999), "Inferring parameters and structure of latent variable models by variational Bayes," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 21–30, Morgan Kaufmann Publishers Inc.

Bagnoli, M. and Bergstrom, T. (2005), "Log-concave probability and its applications," *Economic theory*, 26, 445–469.

Baldi, P., Sadowski, P., and Whiteson, D. (2014), "Searching for exotic particles in high-energy physics with deep learning," *Nature communications*, 5, 1–9.

Balkema, A. and Embrechts, P. (2007), *High risk scenarios and extremes: a geometric approach*, European Mathematical Society.

Ballani, F. and Schlather, M. (2011), "A construction principle for multivariate extreme value distributions," *Biometrika*, 98, 633–645.

Banerjee, A., Dunson, D. B., and Tokdar, S. T. (2013), "Efficient Gaussian process regression for large datasets," *Biometrika*, 100, 75–89.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 825–848.

Beguería, S. and Vicente-Serrano, S. M. (2006), "Mapping the hazard of extreme rainfall by peaks over threshold extreme value analysis and spatial regression techniques," *Journal of applied meteorology and climatology*, 45, 108–124.

Beguería, S., Angulo-Martínez, M., Vicente-Serrano, S. M., López-Moreno, J. I., and El-Kenawy, A. (2011), "Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis: a case study in northeast Spain from 1930 to 2006," *International Journal of Climatology*, 31, 2102–2114.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2006), *Statistics of extremes: theory and applications*, John Wiley & Sons.

Belloni, A. and Chernozhukov, V. (2009), "On the computational complexity of MCMC-based estimators in large samples," *The Annals of Statistics*, pp. 2011–2055.

Bhatia, R. (2013), *Matrix analysis*, vol. 169, Springer Science & Business Media.

Bhattacharya, A. and Dunson, D. B. (2010), "Nonparametric Bayesian density estimation on manifolds with applications to planar shapes," *Biometrika*, 102, 851–865.

Bhattacharya, A. and Dunson, D. B. (2012), "Simplex factor models for multivariate unordered categorical data," *Journal of the American Statistical Association*, 107, 362–377.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007), *Discrete multivariate analysis: theory and practice*, Springer.

Bondell, H. D. and Reich, B. J. (2012), "Consistent high-dimensional Bayesian variable selection via penalized credible regions," *Journal of the American Statistical Association*, 107, 1610–1624.

Buishand, T., de Haan, L., and Zhou, C. (2008), "On spatial extremes: with application to a rainfall problem," *The Annals of Applied Statistics*, pp. 624–642.

Chen, C.-P. and Qi, F. (2003), "The best lower and upper bounds of harmonic sequence," *RGMIA research report collection*, 6.

Choi, H. M. and Hobert, J. P. (2013), "The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic," *Electronic Journal of Statistics*, 7, 2054–2064.

Cohen, J. E. and Rothblum, U. G. (1993), "Nonnegative ranks, decompositions, and factorizations of nonnegative matrices," *Linear Algebra and its Applications*, 190, 149–168.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001), *An introduction to statistical modeling of extreme values*, vol. 208, Springer.

Coles, S. G. and Tawn, J. A. (1991), "Modelling extreme multivariate events," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 377–392.

Dahinden, C., Kalisch, M., and Bühlmann, P. (2010), "Decomposition and model selection for large contingency tables," *Biometrical Journal*, 52, 233–252.

Darroch, J. N., Lauritzen, S. L., and Speed, T. (1980), "Markov fields and log-linear interaction models for contingency tables," *The Annals of Statistics*, pp. 522–539.

Das, B., Resnick, S. I., et al. (2011), "Conditioning on an extreme component: Model consistency with regular variation on cones," *Bernoulli*, 17, 226–252.

Davison, A. C. and Smith, R. L. (1990), "Models for exceedances over high thresholds," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 393–442.

Dawid, A. P. and Lauritzen, S. L. (1993), "Hyper Markov laws in the statistical analysis of decomposable graphical models," *The Annals of Statistics*, pp. 1272–1317.

De Haan, L. (1984), "A spectral representation for max-stable processes," *The Annals of Probability*, pp. 1194–1204.

De Haan, L. and Ferreira, A. (2007), *Extreme value theory: an introduction*, Springer Science & Business Media.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000a), "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, 21, 1253–1278.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000b), "On the best rank-1 and rank-(R 1, R 2,..., Rn) approximation of higher-order tensors," *SIAM Journal on Matrix Analysis and Applications*, 21, 1324–1342.

Dellaportas, P. and Forster, J. J. (1999), "Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models," *Biometrika*, 86, 615–633.

Diaconis, P. and Ylvisaker, D. (1979), "Conjugate priors for exponential families," *The Annals of statistics*, 7, 269–281.

Dobra, A. and Lenkoski, A. (2011), "Copula Gaussian graphical models and their application to modeling functional disability data," *The Annals of Applied Statistics*, 5, 969–993.

Dobra, A. and Massam, H. (2010), "The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors," *Statistical Methodology*, 7, 240–253.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004), "Sparse graphical models for exploring gene expression data," *Journal of Multivariate Analysis*, 90, 196–212.

Drton, M., Sturmfels, B., and Sullivant, S. (2009), *Lectures on algebraic statistics*, Springer Science & Business Media.

Dunson, D. B. and Xing, C. (2009), "Nonparametric Bayes modeling of multivariate categorical data," *Journal of the American Statistical Association*, 104.

Dzirasa, K., Phillips, H. W., Sotnikova, T. D., Salahpour, A., Kumar, S., Gainetdinov, R. R., Caron, M. G., and Nicolelis, M. A. (2010), "Noradrenergic control of cortico-striato-thalamic and mesolimbic cross-structural synchrony," *The Journal of Neuroscience*, 30, 6387–6397.

Engelke, S., Malinowski, A., Kabluchko, Z., and Schlather, M. (2014), "Estimation of Hüsler–Reiss distributions and Brown–Resnick processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Evanno, G., Regnaut, S., and Goudet, J. (2005), "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study," *Molecular ecology*, 14, 2611–2620.

Fearnhead, P. (2004), "Particle filters for mixture models with an unknown number of components," *Statistics and Computing*, 14, 11–21.

Ferré, D., Hervé, L., Ledoux, J., et al. (2013), "Regular perturbation of V-geometrically ergodic Markov chains," *Journal of applied probability*, 50, 184–194.

Fienberg, S. E. and Rinaldo, A. (2007), "Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation," *Journal of Statistical Planning and Inference*, 137, 3430–3445.

Fienberg, S. E., Hersh, P., Rinaldo, A., and Zhou, Y. (2007), "Maximum likelihood estimation in latent class models for contingency table data," *arXiv preprint arXiv:0709.3535*.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009), "Improving the performance of predictive process modeling for large datasets," *Computational statistics & data analysis*, 53, 2873–2884.

Frühwirth-Schnatter, S. and Frühwirth, R. (2010), "Data augmentation and MCMC for binary and multinomial logit models," in *Statistical modelling and regression structures*, pp. 111–132, Springer.

Gamerman, D. and Lopes, H. F. (2006), *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, CRC Press, 2 edn.

Garcia, L. D., Stillman, M., and Sturmfels, B. (2005), "Algebraic geometry of Bayesian networks," *Journal of Symbolic Computation*, 39, 331–355.

Geiger, D., Heckerman, D., King, H., and Meek, C. (2001), "Stratified exponential families: graphical models and model selection," *The Annals of statistics*, 29, 505–529.

Gelfand, A. E. and Smith, A. F. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, 85, 398–409.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pp. 721–741.

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000), "Convergence rates of posterior distributions," *Annals of Statistics*, 28, 500–531.

Ghosh, J., Li, Y., and Mitra, R. (2015), "On the use of Cauchy prior distributions for Bayesian logistic regression," *arXiv preprint arXiv:1507.07170*.

Gibson, W. A. (1955), "An extension of Anderson's solution for the latent structure equations," *Psychometrika*, 20, 69–73.

Golub, G. H. (1973), "Some modified matrix eigenvalue problems," *Siam Review*, 15, 318–334.

Goodman, L. A. (1974), "Exploratory latent structure analysis using both identifiable and unidentifiable models," *Biometrika*, 61, 215–231.

Green, P. J. and Richardson, S. (2001), "Modelling heterogeneity with and without the Dirichlet process," *Scandinavian journal of statistics*, pp. 355–375.

Gregory, D. A. and Pullman, N. J. (1983), "Semiring rank: Boolean rank and nonnegative rank factorizations," *J. Combin. Inform. System Sci*, 8, 223–233.

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2014), "Bayesian Conditional Density Filtering," *arXiv preprint arXiv:1401.3632*.

Haberman, S. J. (1974), "Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations," *The Annals of Statistics*, pp. 911–924.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011), "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, 53, 217–288.

Harrison, J. and West, M. (1999), *Bayesian Forecasting & Dynamic Models*, Springer.

Harshman, R. A. (1970), "Foundations of the PARAFAC procedure: models and conditions for an" explanatory" multimodal factor analysis," .

Heffernan, J. E. and Resnick, S. I. (2007), "Limit laws for random vectors with an extreme component," *The Annals of Applied Probability*, pp. 537–571.

Heffernan, J. E. and Tawn, J. A. (2004), "A conditional approach for multivariate extreme values (with discussion)," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 497–546.

Heffernan, J. E., Tawn, J. A., and Zhang, Z. (2007), "Asymptotically (in) dependent multivariate maxima of moving maxima processes," *Extremes*, 10, 57–82.

Hobert, J. P., Marchev, D., et al. (2008), "A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms," *The Annals of Statistics*, 36, 532–554.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1998), "Bayesian model averaging," in *In Proceedings of the AAAI Workshop on Integrating Multiple Learned Models*, pp. 77–83, Citeseer.

Hoffman, A. J., Wielandt, H. W., et al. (1953), "The variation of the spectrum of a normal matrix," *Duke Math. J*, 20, 37–39.

Hoffman, H. E., Blair, E. R., Johndrow, J. E., and Bishop, A. C. (2005), "Allele-specific inhibitors of protein tyrosine phosphatases," *Journal of the American Chemical Society*, 127, 2824–2825.

Holmes, C. C., Held, L., et al. (2006), "Bayesian auxiliary variable models for binary and multinomial regression," *Bayesian Analysis*, 1, 145–168.

Hu, J., Joshi, A., and Johnson, V. E. (2009), "Log-linear models for gene association," *Journal of the American Statistical Association*, 104, 597–607.

Huelsenbeck, J. P. and Andolfatto, P. (2007), "Inference of population structure under a Dirichlet process model," *Genetics*, 175, 1787–1802.

Hughes, J. and Haran, M. (2013), "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 139–159.

Humphreys, K. and Titterington, D. (2003), "Variational approximations for categorical causal modeling with latent variables," *Psychometrika*, 68, 391–412.

Imai, K. and van Dyk, D. A. (2005), "A Bayesian analysis of the multinomial probit model using marginal data augmentation," *Journal of econometrics*, 124, 311–334.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96.

James, W. and Stein, C. (1961), "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 361–379.

Johndrow, J., Dunson, D., and Lum, K. (2013), "Diagonal orthant multinomial probit models," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pp. 29–38.

Johndrow, J., Lum, K., and Ball, P. (2015a), *dga: Capture-Recapture Estimation using Bayesian Model Averaging*, R package version 1.2.

Johndrow, J. E. and Bhattacharya, A. (2015), "Optimal Gaussian approximations to the posterior for log-linear models with Diaconis-Ylvisaker priors," *arXiv preprint arXiv:1511.00764*.

Johndrow, J. E. and Wolpert, R. L. (2015), "Tail waiting times and the extremes of stochastic processes," *arXiv preprint arXiv:1512.07848*.

Johndrow, J. E., Magie, C. R., and Parkhurst, S. M. (2004), "Rho GTPase function in flies: insights from a developmental and organismal perspective," *Biochemistry and cell biology*, 82, 643–657.

Johndrow, J. E., Battacharya, A., and Dunson, D. B. (2014a), "Supplement to: Tensor decompositions and sparse log-linear models," *arXiv preprint arXiv:1404.0396*.

Johndrow, J. E., Battacharya, A., and Dunson, D. B. (2014b), "Tensor decompositions and sparse log-linear models," *arXiv preprint arXiv:1404.0396*.

Johndrow, J. E., Mattingly, J. C., Mukherjee, S., and Dunson, D. (2015b), "Approximations of Markov Chains and High-Dimensional Bayesian Inference," *arXiv preprint arXiv:1508.03387*.

Jones, G. L. et al. (2004), "On the Markov chain central limit theorem," *Probability surveys*, 1, 5–1.

Joulin, A., Ollivier, Y., et al. (2010), "Curvature, concentration and error estimates for Markov chain Monte Carlo," *The Annals of Probability*, 38, 2418–2442.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002), "Statistics of extremes in hydrology," *Advances in water resources*, 25, 1287–1304.

Kolda, T. G. and Bader, B. W. (2009), "Tensor decompositions and applications," *SIAM review*, 51, 455–500.

Kontoyiannis, I. and Meyn, S. P. (2003), "Spectral theory and limit theorems for geometrically ergodic Markov processes," *Annals of Applied Probability*, pp. 304–362.

Korattikara, A., Chen, Y., and Welling, M. (2013), "Austerity in MCMC land: Cutting the Metropolis-Hastings budget," *arXiv preprint arXiv:1304.5299*.

Kunihama, T. and Dunson, D. B. (2012), "Bayesian modeling of temporal dependence in large sparse contingency tables," *arXiv preprint arXiv:1205.2816*.

Lauritzen, S. L. (1996), *Graphical models*, Oxford University Press.

Lawler, G. F. and Sokal, A. D. (1988) *Transactions of the American mathematical society*, 309, 557–580.

Lazarsfeld, P. F. and Henry, N. W. (1968), *Latent structure analysis*, Houghton, Mifflin.

Letac, G. and Massam, H. (2012), "Bayes factors and the geometry of discrete hierarchical loglinear models," *The Annals of Statistics*, 40, 861–890.

Lim, L.-H. and Comon, P. (2009), "Nonnegative approximations of nonnegative tensors," *Journal of Chemometrics*, 23, 432–441.

Liu, I. A., Johndrow, J. E., Abe, J., Lüpold, S., Yasukawa, K., Westneat, D. F., and Nowicki, S. (2015), "Genetic diversity does not explain variation in extra-pair paternity in multiple populations of a songbird," *Journal of evolutionary biology*, 28, 1156–1169.

Liu, R., Woolner, S., Johndrow, J. E., Metzger, D., Flores, A., and Parkhurst, S. M. (2008), "Sisyphus, the Drosophila myosin XV homolog, traffics within filopodia transporting key sensory and adhesion cargos," *Development*, 135, 53–63.

Lovász, L. and Simonovits, M. (1993), "Random walks in a convex body and an improved volume algorithm," *Random structures and algorithms*.

Machado, F. S., Johndrow, J. E., Esper, L., Dias, A., Bafica, A., Serhan, C. N., and Aliberti, J. (2006), "Anti-inflammatory actions of lipoxin A4 and aspirin-triggered lipoxin are SOCS-2 dependent," *Nature medicine*, 12, 330–334.

Madansky, A. (1960), "Determinantal methods in latent class analysis," *Psychometrika*, 25, 183–198.

Massam, H., Liu, J., and Dobra, A. (2009), "A conjugate prior for discrete hierarchical log-linear models," *The Annals of Statistics*, 37, 3431–3467.

Medvedovic, M. and Sivaganesan, S. (2002), "Bayesian infinite mixture model based clustering of gene expression profiles," *Bioinformatics*, 18, 1194–1206.

Meinguet, T. (2012), "Maxima of moving maxima of continuous functions," *Extremes*, 15, 267–297.

Méndez, F. J., Menéndez, M., Luceño, A., and Losada, I. J. (2006), "Estimation of the long-term variability of extreme significant wave height using a time-dependent Peak Over Threshold (POT) model," *Journal of Geophysical Research: Oceans (1978–2012)*, 111.

Meyn, S. and Tweedie, R. L. (2009), *Markov chains and stochastic stability*, Cambridge University Press, Cambridge, second edn., With a prologue by Peter W. Glynn.

Meyn, S. P. and Tweedie, R. L. (1993), "Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes," *Advances in Applied Probability*, pp. 518–548.

Meyn, S. P. and Tweedie, R. L. (1994), "Computable bounds for geometric convergence rates of Markov chains," *The Annals of Applied Probability*, pp. 981–1011.

Miller, J. (2014), "Nonparametric and Variable-Dimension Bayesian Mixture Models: Analysis, Comparison, and New Methods," Ph.D. thesis, Brown University.

Miller, J. W. and Harrison, M. T. (2014), "Inconsistency of Pitman-Yor process mixtures for the number of components," *The Journal of Machine Learning Research*, 15, 3333–3370.

Miller, J. W. and Harrison, M. T. (2015), "Mixture models with a prior on the number of components," *arXiv preprint arXiv:1502.06241*.

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. B. (2014), "Robust and scalable Bayes via a median of subset posterior measures," *arXiv preprint arXiv:1403.2660*.

Mitrophanov, A. Y. (2005), "Sensitivity and convergence of uniformly ergodic Markov chains," *Journal of Applied Probability*, pp. 1003–1014.

Mossel, E. and Vigoda, E. (2006), "Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny," *The Annals of Applied Probability*, pp. 2215–2234.

Nardi, Y. and Rinaldo, A. (2012), "The log-linear group-lasso estimator and its asymptotic properties," *Bernoulli*, 18, 945–974.

Nobile, A. (1994), "Bayesian Analysis of Finite Mixture Distributions," Ph.D. thesis, Carnegie Mellon University.

Nobile, A. and Fearnside, A. T. (2007), "Bayesian finite mixtures with an unknown number of components: The allocation sampler," *Statistics and Computing*, 17, 147–162.

Nummelin, E. (1978), "A splitting technique for Harris recurrent Markov chains," *Z. Wahrsch. Verw. Gebiete.*, 43, 309–318.

O'brien, S. M. and Dunson, D. B. (2004), "Bayesian multivariate logistic regression," *Biometrics*, 60, 739–746.

Olver, F. W. (2010), *NIST handbook of mathematical functions*, Cambridge University Press.

Onogi, A., Nurimoto, M., and Morita, M. (2011), "Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods," *BMC bioinformatics*, 12, 263.

Otranto, E. and Gallo, G. M. (2002), "A nonparametric Bayesian approach to detect the number of regimes in Markov switching models," *Econometric Reviews*, 21, 477–496.

Owen, A. B. (2007), "Infinitely imbalanced logistic regression," *The Journal of Machine Learning Research*, 8, 761–773.

Park, M. Y. and Hastie, T. (2007), "L1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 659–677.

Pati, D., Bhattacharya, A., Pillai, N. S., Dunson, D., et al. (2014), "Posterior contraction in sparse Bayesian factor models for massive covariance matrices," *The Annals of Statistics*, 42, 1102–1130.

Pillai, N. S. and Smith, A. (2014), "Ergodicity of Approximate MCMC Chains with Applications to Large Data Sets," *arXiv preprint arXiv:1405.0182*.

Polson, N. G., Scott, J. G., and Windle, J. (2013), "Bayesian inference for logistic models using Pólya–Gamma latent variables," *Journal of the American Statistical Association*, 108, 1339–1349.

Prado, R. and West, M. (2010), *Time series: modeling, computation, and inference*, CRC Press.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000), "Inference of population structure using multilocus genotype data," *Genetics*, 155, 945–959.

Rajaratnam, B. and Sparks, D. (2015), "MCMC-Based Inference in the Era of Big Data: A Fundamental Analysis of the Convergence Complexity of High-Dimensional Chains," *arXiv preprint arXiv:1508.00947*.

Richardson, S. and Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 731–792.

Ritter, C. and Tanner, M. A. (1992), "Facilitating the Gibbs sampler: The Gibbs stopper and griddy-Gibbs sampler," *Journal of the American Statistical Association*, 87, 861–868.

Robert, C. and Casella, G. (2013), *Monte Carlo statistical methods*, Springer Science & Business Media.

Robert, C. P. and Casella, G. (2004), *Monte Carlo statistical methods*, Springer, 2 edn.

Roberts, G. O. and Rosenthal, J. S. (1997), "Geometric ergodicity and hybrid Markov chains," *Electron. Comm. Probab*, 2, 13–25.

Rootzén, H. and Tajvidi, N. (2006), "Multivariate generalized Pareto distributions," *Bernoulli*, pp. 917–930.

Rosales-Nieves, A. E., Johndrow, J. E., Keller, L. C., Magie, C. R., Pinto-Santini, D. M., and Parkhurst, S. M. (2006), "Coordination of microtubule and microfilament dynamics by Drosophila Rho1, Spire and Cappuccino," *Nature cell biology*, 8, 367–376.

Rosenthal, J. S. (1994), "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo," *Journal of the American Statistical Association*, 90, 558–566.

Roth, V. and Fischer, B. (2008), "The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms," in *Proceedings of the 25th international conference on Machine learning*, pp. 848–855, ACM.

Rousseau, J. and Mengersen, K. (2011), "Asymptotic behaviour of the posterior distribution in overfitted mixture models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 689–710.

Roy, V. and Hobert, J. P. (2007), "Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 607–623.

Rudolf, D. and Schweizer, N. (2015), "Perturbation theory for Markov chains via Wasserstein distance," *arXiv preprint arXiv:1503.04123*.

Rusakov, D. and Geiger, D. (2002), "Asymptotic model selection for naive Bayesian networks," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 438–455, Morgan Kaufmann Publishers Inc.

Saumard, A., Wellner, J. A., et al. (2014), "Log-concavity and strong log-concavity: a review," *Statistics Surveys*, 8, 45–114.

Schlather, M. (2002), "Models for stationary max-stable random fields," *Extremes*, 5, 33–44.

Settimi, R. and Smith, J. Q. (1998), "On the geometry of Bayesian graphical models with hidden variables," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pp. 472–479, Morgan Kaufmann Publishers Inc.

Shun, Z. and McCullagh, P. (1995), "Laplace approximation of high dimensional integrals," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 749–760.

Smith, J. Q. and Croft, J. (2003), "Bayesian networks for discrete multivariate data: an algebraic approach to inference," *Journal of Multivariate Analysis*, 84, 387–402.

Smith, R. and Weissman, I. (1996), "Characterization and Estimation of the Multivariate Extremal Index," .

Smith, R. L. (1984), "Threshold methods for sample extremes," in *Statistical extremes and applications*, pp. 621–638, Springer.

Smith, R. L. (1989), "Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone," *Statistical Science*, 4, 367–377.

Smith, R. L. (1990), "Max-stable processes and spatial extremes," *Unpublished manuscript, Univer.*

Smola, A. J. and Bartlett, P. (2001), "Sparse greedy Gaussian process regression," in *Advances in Neural Information Processing Systems 13*.

Stanley, S. A., Johndrow, J. E., Manzanillo, P., and Cox, J. S. (2007), "The Type I IFN response to infection with Mycobacterium tuberculosis requires ESX-1-mediated secretion and contributes to pathogenesis," *The Journal of Immunology*, 178, 3143–3152.

Stein, C. (1956), "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 197–206.

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., and Clausen, J. A. (1950), "Measurement and prediction." .

Tao, T. (2015), "254A, Notes 3a: Eigenvalues and sums of Hermitian matrices," `https://terrytao.wordpress.com/2010/01/12/254a-notes-3a-eigenvalues-and-sums-of-hermitian-matrices/#more-3341`, Accessed: 2015-11-02.

Tawn, J. A. (1988), "Bivariate extreme value theory: models and estimation," *Biometrika*, 75, 397–415.

Tawn, J. A. (1990), "Modelling multivariate extreme value distributions," *Biometrika*, 77, 245–253.

Tierney, L. and Kadane, J. B. (1986), "Accurate approximations for posterior moments and marginal densities," *Journal of the american statistical association*, 81, 82–86.

Tucker, L. R. (1966), "Some mathematical notes on three-mode factor analysis," *Psychometrika*, 31, 279–311.

Vijayakumar, S., D'souza, A., and Schaal, S. (2005), "Incremental online learning in high dimensions," *Neural computation*, 17, 2602–2634.

Wang, B. and Titterington, D. (2004), "Lack of consistency of mean field and variational Bayes approximations for state space models," *Neural Processing Letters*, 20, 151–170.

Wang, B. and Titterington, D. (2005), "Inadequacy of interval estimates corresponding to variational Bayesian approximations," *AISTATS 2005*, p. 373.

Weiss, L. (1978), "The error in the normal approximation to the multinomial with an increasing number of classes," *Naval Research Logistics Quarterly*, 25, 257–261.

West, M. and Escobar, M. D. (1993), *Hierarchical priors and mixture models, with application in regression and density estimation*, Institute of Statistics and Decision Sciences, Duke University.

Wilson, A., Bondell, H. D., and Reich, B. J. (2015), "BayesPen: Bayesian Penalized Credible Regions. R package version 1.2," .

Xing, E. P., Sohn, K.-A., Jordan, M. I., and Teh, Y.-W. (2006), "Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture," in *Proceedings of the 23rd international conference on Machine learning*, pp. 1049–1056, ACM.

Yang, Y. and Dunson, D. B. (2013), "Sequential Markov Chain Monte Carlo," *arXiv preprint arXiv:1308.3861*.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2015), "On the Computational Complexity of High-Dimensional Bayesian Variable Selection," *arXiv preprint arXiv:1505.07925*.

Zhang, Z. and Smith, R. L. (2010), "On the estimation and application of max-stable processes," *Journal of Statistical Planning and Inference*, 140, 1135–1153.

Zhou, J., Bhattacharya, A., Herring, A., and Dunson, D. (2013), "Bayesian factorizations of big sparse tensors," *arXiv preprint arXiv:1306.1598*.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

Biography

James Edward Johndrow was born in Pittsfield, Massachusetts on July 4, 1981. He holds a BA in Chemistry from Amherst College, *Summa cum laude*, and an MS and PhD in Statistical Science from Duke University. He was a James B. Duke and University Scholar graduate fellow at Duke University. He is the author of twelve papers and preprints (Machado et al. (2006), Stanley et al. (2007), Rosales-Nieves et al. (2006), Liu et al. (2015), Liu et al. (2008), Hoffman et al. (2005), Johndrow et al. (2004), Johndrow et al. (2013), Johndrow et al. (2014b), Johndrow and Bhattacharya (2015), Johndrow et al. (2015b), Johndrow and Wolpert (2015)) and one piece of statistical software (Johndrow et al. (2015a)).