

Incorporating Scalability and Structural Constraints in Bayesian Modeling

by

Shounak Chattopadhyay

Department of Statistical Science
Duke University

Date: _____

Approved:

David B. Dunson, Supervisor

Amy H. Herring

Jason Q. Xu

Anru Zhang

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
2023

ABSTRACT

Incorporating Scalability and Structural Constraints in
Bayesian Modeling

by

Shounak Chattopadhyay

Department of Statistical Science
Duke University

Date: _____

Approved:

David B. Dunson, Supervisor

Amy H. Herring

Jason Q. Xu

Anru Zhang

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
2023

Copyright © 2023 by Shounak Chattopadhyay
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Real-life modeling of probabilistic events often involves incorporating constraints on quantities of interest. Broadly, such constraints can be classified as being either *computational* when facing limitations on computational feasibility or budget, or *structural* when facing limitations in terms of modeling a desirable quantity of interest due to the inherent nature of this quantity. To that end, this work focuses on incorporating computational and structural constraints into modeling real-life data from a Bayesian perspective. In Chapter 2, we focus on the problem of Bayesian nonparametric density estimation. Although well-studied and highly regarded in existing literature due to their flexibility, adaptability, and accuracy along with quantifying uncertainty when estimating probability density functions, Bayesian nonparametric approaches often face major roadblocks in terms of computation via cumbersome Markov chain Monte Carlo (MCMC) algorithms. By leveraging on aspects of nearest neighbor allocation and Bayesian mixture models, we engineer a highly effective hybrid density estimation approach called Nearest Neighbor Dirichlet Mixtures (NN-DM). The NN-DM completely avoids MCMC and is embarrassingly parallel, providing substantial computational gains in comparison to existing approaches, along with providing accurate point estimation and uncertainty quantification both theoretically and empirically. In Chapter 3, we consider the problem of dose response modeling in a public health scenario, where individuals are exposed to toxic chemicals. An overwhelming portion of the current approaches only focus on quantifying the marginal

effects of these exposures on the response, ignoring possible interactions. As an alternative, our focus is on incorporating structural constraints in the form of modeling *synergistic* and *antagonistic* interactions between the chemicals. We developed the Synergistic Antagonistic Interaction Detection (SAID), a novel Bayesian approach shrinking interactions to being synergistic or antagonistic. Instead of focusing only on linear effects, our model is flexible to allow non-linearity and scales well computationally with moderate number of exposures. We apply our approach to an NHANES data set and uncover interactions between heavy metals affecting kidney function. Finally, in Chapter 4, we focus on the problem of Bayesian factor analysis. Bayesian factor models provide an elegant framework to model high-dimensional covariance matrices as the sum of two components, one low rank and another diagonal. Existing approaches utilizing MCMC to obtain posterior draws of the covariance matrix face significant challenges in terms of slow convergence and mode switching due to non-identifiability of the factor model resulting from rotational invariance. As both the sample size and the number of dimensions increase, we focus on a *blessing of dimensionality* phenomenon allowing us to effectively obtain a plug-in estimate of the latent factors. Using this plug-in estimate, our proposed Factor Analysis with BLEssing of dimensionality (FABLE) approach provides a pseudo-posterior for the covariance matrix. FABLE is an embarrassingly parallel technique with immense computational benefits, completely bypassing MCMC and thus its pitfalls. We provide theoretical guarantees on the performance of FABLE, along with evaluating the approach in numerous simulation studies.

Dedication

To Ma and Baba

Contents

Abstract	iv
List of Tables	xi
List of Figures	xii
List of Abbreviations and Symbols	xiii
Acknowledgements	xv
1 Introduction	1
2 Nearest Neighbor Dirichlet Mixtures	6
2.1 Introduction	6
2.2 Methodology	10
2.2.1 Nearest Neighbor Dirichlet Mixture Framework	10
2.2.2 Illustration with Gaussian Kernels	12
2.2.3 Hyperparameter Choice	14
2.3 Theoretical Properties	16
2.3.1 Asymptotic Results	16
2.3.2 Pseudo-Posterior Distribution of Weights	20
2.4 Simulation Experiments	22
2.4.1 Preliminaries	22
2.4.2 Univariate Cases	23
2.4.3 Multivariate Cases	26

2.4.4	Accuracy of Uncertainty Quantification	30
2.4.5	Comparison for High Dimensional Data	31
2.4.6	Runtime Comparison	33
2.4.7	Sensitivity to the Choice of k	36
2.5	Application	37
3	Synergistic Antagonistic Interaction Detection	41
3.1	Introduction	41
3.2	Kidney Function Data Analysis	44
3.2.1	Motivation	44
3.2.2	Data Description	45
3.2.3	Urinary Dilution	46
3.2.4	Issues with Existing Approaches	47
3.3	Structured Interaction Modeling Approach	48
3.3.1	Basic Modeling Structure	48
3.3.2	Modeling Pairwise Interactions	49
3.3.3	Variable Selection	53
3.3.4	Main Effects and Other Parameters	54
3.3.5	Posterior Sampling	55
3.4	Simulation Examples	58
3.4.1	Preliminaries	58
3.4.2	Two Exposures	60
3.4.3	More than Two Exposures	62
3.5	Analysis of Kidney Function Data	64
3.5.1	Preliminaries	64
3.5.2	Model Diagnostics	66

3.5.3	Results	67
4	Factor Analysis with Blessing of Dimensionality	72
4.1	Introduction	72
4.2	Proposed Methodology	75
4.2.1	Initial Approach	75
4.2.2	Coverage Correction	81
4.2.3	Hyperparameter Choice	82
4.2.4	Final Algorithm	84
4.3	Theoretical Support	86
4.4	Simulation Results	92
4.4.1	Preliminaries	92
4.4.2	Estimation Performance	93
4.4.3	Frequentist Coverage	95
5	Conclusion and Future Research	98
A	Further details for Chapter 2	102
A.1	Prerequisites	102
A.2	Proof of Theorem 1	104
A.3	Proof of Theorem 2	108
A.4	Proof of Theorem 4	109
A.5	Proof of Theorem 5	115
A.5.1	A property of the k -nearest neighbor distance	115
A.5.2	Number of effective member points in each neighborhood	117
A.6	Proof of Consistency of KDE	118
A.7	Cross-validation	120
A.7.1	Algorithm for leave-one-out cross-validation	120

A.7.2	Fast Implementation of cross-validation	121
A.8	Algorithm with Gaussian Kernels for Univariate Data	122
A.9	Inverse Wishart Parametrization	123
A.10	Univariate and Multivariate \mathcal{L}_1 Error Tables	124
B	Further details for Chapter 3	126
B.1	B-spline Functions	126
B.2	P-spline Priors	127
B.3	Model Identifiability	127
B.4	Other Approaches	129
B.5	Posterior Sampling	130
B.6	Cutoff in Variable Selection	133
B.7	Further Tables on Simulation Results	134
B.8	Creatinine Data Application	135
C	Further details for Chapter 4	138
C.1	Proofs of Results	138
C.2	Proof of Theorem 8	139
C.2.1	Proof of part (a)	139
C.2.2	Proof of part (b)	142
C.2.3	Proof of part (c)	143
C.2.4	Relevant lemmas and their proofs	144
C.3	Proof of Theorem 10	149
C.4	Proof of Theorem 11	151
C.5	Related Lemmas for Theorems 10 and 11	154
	Bibliography	157
	Biography	167

List of Tables

2.1	Coverage for univariate data	32
2.2	Coverage for bivariate data	32
2.3	Predictive log-likelihood comparison	33
2.4	Runtime comparison for density estimation	36
3.1	Error to estimate whole surface with two exposures.	61
3.2	Error to estimate interaction surface with two exposures.	61
3.3	Simulation results with multiple exposures.	64
4.1	Estimation error results for $\pi_0 = 0$	94
4.2	Estimation error results for $\pi_0 = 0.4$	94
4.3	Estimation error results for $\pi_0 = 0.6$	95
4.4	Coverage comparison between CC-FABLE and FABLE	96
A.1	Univariate \mathcal{L}_1 errors	124
A.2	Multivariate \mathcal{L}_1 errors	125
B.1	Variable selection for different cutoffs	134
B.2	RMSE when estimating whole surface for case QR	134
B.3	RMSE when estimating interaction surface for case QR	134
B.4	RMSE when estimating whole surface for case MIS	135
B.5	RMSE when estimating interaction surface for case MIS	135

List of Figures

2.1	Box plots for univariate density estimation	24
2.2	Smoothness of density estimates for sawtooth density	26
2.3	Smoothness of density estimates for skewed bimodal density	27
2.4	Box plots for multivariate density estimation	29
2.5	Runtime comparison with estimation errors	36
2.6	Choice of k in NN-DM	37
2.7	Comparison of sensitivity and specificity	39
2.8	Comparison of receiver operating characteristic curves	40
2.9	Comparison of predictive log-likelihoods	40
3.1	Heat plot of metal correlations	46
3.2	Effect of penalty on prior specification	53
3.3	Q-Q plot and posterior predictive check in NHANES analysis	67
3.4	Plots of main effects in NHANES analysis	69
3.5	Plots of interaction effects in NHANES analysis	71
4.1	Comparison of pseudo-credible intervals	97
B.1	Plots for main effects of other metals	136
B.2	Interaction between Molybdenum and Tin	137

List of Abbreviations and Symbols

Symbols

\mathbb{R}^p	The set of all p tuples with real entries.
$\mathbb{R}^{m \times n}$	The set of all $m \times n$ matrices with real entries.
Dirichlet(\mathbf{p})	The Dirichlet distribution with parameters $\mathbf{p} = (p_1, \dots, p_k)$.
$N(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2 .
$N_p(\eta, \Sigma)$	p -dimensional multivariate normal distribution with mean η and covariance matrix Σ .
Φ	Standard normal cumulative distribution function.
n	Number of samples.
p	Number of dimensions.
i, j, u, v	Used throughout to denote indices.
IW_p	The Wishart distribution on $p \times p$ matrices.
f_0	A true data generating density.
G, IG	The gamma and inverse-gamma distributions.
\wedge	Minimum.
\vee	Maximum.
$\mathbb{1}$	The indicator function.
0_p	Vector of all zeros in \mathbb{R}^p .
1_p	Vector of all ones in \mathbb{R}^p .
\mathbb{I}_p	The $p \times p$ identity matrix.

$\mathbb{O}_{m \times n}$	The $m \times n$ matrix of zeros.
$\lfloor \cdot \rfloor, \lceil \cdot \rceil$	The floor and ceiling functions.
\mathcal{O}	Computational complexity.
(\cdot)	Combinatorial function.
C^+	Half-Cauchy distribution; refer to Polson and Scott (2010).
\approx	Approximation.
\asymp	Asymptotically equivalent.
\ll	Asymptotically negligible.
\odot	Hadamard product.
$\text{diag}(\mathbf{a})$	A diagonal matrix with diagonal entries $\mathbf{a} = (a_1, \dots, a_k)$.

Abbreviations

CDF	Cumulative distribution function.
PDF	Probability density function.
EP	Embarrassingly parallel.
MCMC	Markov chain Monte Carlo.
UQ	Uncertainty quantification.
NIG, NIW	Normal inverse-gamma and Normal inverse-Wishart.
KNN	k -nearest neighbors.
CV	Cross-validation.
ROC	Receiver operating characteristic.
PIP	Posterior inclusion probability.
HMC	Hamiltonian Monte Carlo.
SVD	Singular value decomposition.
RMSE	Root mean squared error.

Acknowledgements

I am immensely grateful to my advisor, David Dunson, for his mentorship, patience, and encouragement throughout this journey. When confronting doldrums or in doubt, his sustained energy behind research projects was a constant source of motivation. His honesty in our conversations helped me grow, particularly when conducting independent research and developing clear scientific communication. His striking intuition when dealing with roadblocks will continue to inspire me.

In the past few years, I had the immense pleasure of interacting with outstanding researchers and learning from them. I am obliged to the members of my committee, Amy Herring, Jason Xu, and Anru Zhang, for providing insight and guidance. I am thankful to the International Society for Bayesian Analysis (ISBA) for funding my travel to the ISBA World Meeting and the International Conference on Bayesian Nonparametrics. I am grateful to Bani Mallick and Anirban Bhattacharya for introducing me to the beautiful world of Bayesian statistics.

I was fortunate to have the privilege of companionship during my time at Duke, rescuing me from unbearable solitude. Thank you to my friends, old and new, with whom I have shared countless enjoyable moments, memories of which will last a lifetime. To my partner Anandita, thank you for standing by me, especially when I found myself in times of trouble. Finally, I am forever indebted to my parents - Ma and Baba, for their unwavering support and the countless sacrifices they have made for me. I could not have done this without them.

1

Introduction

Statistical modeling of real-life data has an abundance of problems requiring incorporation of constraints. For example, one could face challenges with fitting a complex Bayesian nonparametric model (Ghosal and van der Vaart, 2017) in terms of finite computational budget and lack of scalability. Alternatively, other constraints are woven directly into the quantities being modeled, such as the response of an individual from exposure to harmful substances being monotone in the dose of exposure (Ramsay, 1988). Ignoring such constraints when constructing the model could lead to sub-optimal performance of the model in terms of estimating the desirable quantities of interest, or could lead to infeasible computational runtimes for algorithms due to inefficient exploration of the model space when fitting the model. There is thus a compelling need to develop approaches that incorporate constraints as described above when carrying out statistical inference and prediction.

We are particularly concerned with providing scalable and structured approaches from a Bayesian perspective. In problems involving sampling from a high-dimensional or other complex posterior distributions, one often encounters stiff challenges when implementing Markov chain Monte Carlo (MCMC); for example, the chain can face

slow convergence or can get stuck at modes. There have been a wide variety of attempts to resolve such issues. Analytical approximations to the posterior such as the Bernstein von Mises theorem or Laplace approximations (Ghosal and van der Vaart, 2017) attempt to obtain a simplified distribution which is asymptotically close to the posterior. However, such approximations are often difficult to derive and insufficient in terms of performance under the finite sample regime. Variational Bayesian approaches (Blei and Jordan, 2006) consider a tractable class of distributions and approximate the posterior by the member of the class closest to the posterior. Although exhibiting immense computational gains, such approaches often heavily underestimate uncertainty and are difficult to justify theoretically. Other promising directions involve approximating or modifying the problematic MCMC itself to derive embarrassingly parallelizable approaches (Srivastava et al., 2018) that scale well with the size of the data. In this work, we focus on obtaining scalable and embarrassingly parallel approaches bypassing MCMC altogether for common inference problems, such as density estimation or high-dimensional covariance estimation via factor analysis. Along with scalability constraints, we also focus on incorporating structural constraints in a regression framework; in particular, we consider the problem of dose response modeling as a function of harmful exposures and extracting possible *synergistic* and *antagonistic* interactions.

Algorithmic approaches such as k -nearest neighbor (KNN) density estimation (Mack and Rosenblatt, 1979) provide point estimates that are fast to compute, but often do not provide a valid notion of uncertainty. On the other hand, Bayesian nonparametric approaches such as Dirichlet process mixtures and overfitted mixture models (Ferguson, 1973; Rousseau and Mengersen, 2011) provide an elegant framework for density estimation, along with automatic uncertainty quantification. However, such approaches are accompanied by substantial computational burden, along with difficulty in implementation due to mixing and convergence issues with

the use of MCMC. In Chapter 2, we propose the Nearest Neighbor Dirichlet Mixture (NN-DM) density estimator, an embarrassingly parallel approach borrowing ideas from both the algorithmic viewpoint of KNN allocation and the flexible framework of mixture models. Upon observing data, we rely on the k -nearest neighborhood of each data point and characterize the density within the neighborhood using a parametric model, such as a Gaussian. The local density estimates are combined into a global density estimate using a weight vector. With prior distributions on the neighborhood specific parameters and the weights, we obtain a pseudo-posterior distribution of the density estimator at each input point. The NN-DM completely bypasses MCMC if the prior distributions are chosen to be conjugate to the assumed parametric model in each local neighborhood; for example, the local density could be a Gaussian, with normal-inverse Wishart priors on the neighborhood specific means and covariances. This leads to substantial computational gains along with avoiding the pitfalls of MCMC. The NN-DM pseudo-posterior also has attractive theoretical properties in terms of convergence and uncertainty quantification. To evaluate the proposed method, extensive numerical experiments and an application to the High Time Resolution Universe (HTRU) (Keith et al., 2010) survey data on classifying pulsar stars are carried out. The NN-DM performs very well across a variety of cases along with providing better frequentist coverage than its competitors.

There has been considerable attention in the epidemiological community regarding health effects of mixtures of chemical exposures (Joubert et al., 2022). Most of the existing approaches ignore interactions between these exposures and only focus on characterizing main effects, which can lead to misestimation of the health hazard of chemicals. In Chapter 3, we are particularly motivated by capturing synergistic and antagonistic interactions between heavy metals affecting kidney function, in data collected by NHANES. Synergistic interactions amplify the hazard of an exposure in the presence of another, while antagonistic interactions tend to inhibit each other's

effects. Existing statistical methods to detect such interactions are either parametric and thus too restrictive, or rely on overly flexible nonparametric models leading to wiggly surface estimates which are difficult to interpret. To bypass these issues, we propose the Synergistic Antagonistic Interaction Detection (SAID), which identifies nonlinear synergistic or antagonistic pairwise interactions. SAID carries out Bayesian inference with a carefully structured prior allowing us to capture both the magnitude and direction of interactions, along with an option to relax the prior if the interaction is neither synergistic nor antagonistic. Instead of modeling complicated bivariate surfaces, we use a modification to substantially decrease computational burden and improve scalability. The approach is compared with competitors in simulation studies. SAID succeeds in detecting multiple synergistic and antagonistic interactions in the NHANES data not detected by existing state-of-the-art approaches. The obtained results are of fundamental public health and clinical importance given the epidemic of kidney disease in agricultural workers worldwide, for which heavy metals provide one plausible cause.

Bayesian latent factor models (Bhattacharya and Dunson, 2011; Lopes and West, 2004) have been widely used to reduce dimensionality in characterizing dependence in high-dimensional and complex data. However, such approaches often suffer from inefficient and slow MCMC for inferring the latent factors and factor loadings, particularly struggling with the nonidentifiability arising from orthogonal invariance. Although there are post-processing approaches such as varimax rotations (Rohe and Zeng, 2020) to get rid of such ambiguity, such approaches are often ad-hoc and complicate the inferential process. In Chapter 4, we propose the Factor Analysis with BLEssing of dimensionality (FABLE), an approach which pre-estimates the unknown latent factors in the high-dimensional setup, leveraging a ‘blessing of dimensionality’ phenomenon. Once the latent factors have been pre-estimated, we obtain a pseudo-posterior for the factor loadings in an embarrassingly parallel fashion, pro-

viding a scalable framework for high-dimensional covariance estimation. Compared to existing approaches, FABLE vastly improves computational time, provides accurate estimates, and valid uncertainty quantification. The FABLE pseudo-posterior has desirable theoretical properties showing convergence of the procedure in spectral norm and validity of uncertainty quantification obtained from the pseudo-posterior credible intervals. We also compare FABLE with other state-of-the-art approaches across different simulation cases, highlighting its advantages.

The underlying theme throughout Chapters 2, 3, and 4 is providing frameworks that allow scalable and structured Bayesian inference. A main principle behind developing the scalable algorithms in this work is relying on parallelization. This is obtained by first conditioning on a relevant quantity of interest, pre-estimating this quantity before the sampling procedure, and finally obtaining a simple Monte Carlo sampler that avoids the pitfalls of MCMC. Even when incorporating structural constraints such as searching for synergistic or antagonistic interactions as in Chapter 3, we employ a dimension reduction trick to massively speed up computation. For the convenience of the reader, each Chapter may be read independently.

Nearest Neighbor Dirichlet Mixtures

2.1 Introduction

Bayesian nonparametric methods provide a useful alternative to black box machine learning algorithms, having potential advantages in terms of characterizing uncertainty in inferences and predictions. However, computation can be slow and unwieldy to implement. Hence, it is important to develop simpler and faster Bayesian nonparametric approaches, and *hybrid* methods that borrow the best of both worlds. For example, if one could use the Bayesian machinery for uncertainty quantification and reduction of mean square errors through shrinkage, while incorporating algorithmic aspects of machine learning approaches, one may be able to engineer a highly effective hybrid. The focus of this article is on proposing such an approach for density estimation, motivated by the successes and limitations of nearest neighbor algorithms and Bayesian mixture models.

Nearest neighbor algorithms are popular due to a combination of simplicity and performance. Given a set of n observations $\mathcal{X}^{(n)} = (X_1, \dots, X_n)$ in \mathbb{R}^p , the density at x is estimated as $\hat{f}_{\text{kn}}(x) = k/(nV_p R_k^p)$, where k is the number of neighbors of x in

$\mathcal{X}^{(n)}$, $R_k = R_k(x)$ is the distance of x from its k th nearest neighbor in $\mathcal{X}^{(n)}$, and V_p is the volume of the p -dimensional unit ball (Loftsgaarden and Quesenberry, 1965; Mack and Rosenblatt, 1979). Refer to Biau and Devroye (2015) for an overview of related estimators and corresponding theory.

Nearest neighbor density estimators are a type of locally adaptive kernel density estimators. The literature on such methods identifies two broad classes: balloon estimators and sample smoothing estimators; see Scott (2015); Terrell and Scott (1992) for an overview. Balloon estimators characterize the density at a query point x using a bandwidth function $h(x)$; classical examples include the naive k -nearest neighbor density estimator (Loftsgaarden and Quesenberry, 1965) and its modification in Mack and Rosenblatt (1979). More elaborate balloon estimators face challenges in terms of choice of $h(x)$ and obtaining density estimators that do not integrate to 1. Sample smoothing estimators use n different bandwidths $h(X_i)$, one for each sample point X_i , to estimate the density at a query point x globally. By construction, sample smoothing estimators are *bona fide* density functions integrating to 1. To fit either the balloon or the sample smoothing estimator, one may compute an initial pilot density estimator employing a constant bandwidth and then use this pilot to estimate the bandwidth function (Breiman et al., 1977; Abramson, 1982). Another example of a locally adaptive density estimator is the local likelihood density estimator (Loader, 1996, 2006; Hjort and Jones, 1996), which fits a polynomial model in the neighborhood of a query point x to estimate the density at x , estimating the parameters of the local polynomial by maximizing a penalized local log-likelihood function. The above methods produce a point estimate of the density without uncertainty quantification (UQ).

Alternatively, there is a Bayesian literature on locally adaptive kernel methods,

which express the unknown density as:

$$f(x) = \sum_{h=1}^m \pi_h \mathcal{K}(x; \theta_h), \quad \theta_h \sim P_0, \quad (\pi_h)_{h=1}^m \sim Q_0, \quad (2.1)$$

which is a mixture of m components, with the h th having probability weight π_h and kernel parameters θ_h ; by allowing the location and bandwidth to vary across components, local adaptivity is obtained. A Bayesian specification is completed with prior P_0 for the kernel parameters and Q_0 for the weights. In practice, it is common to rely on an over-fitted mixture model (Rousseau and Mengersen, 2011), which chooses m as a pre-specified finite upper bound on the number of components, and lets

$$\pi = (\pi_1, \dots, \pi_m)^\top \sim \text{Dirichlet}(\alpha, \dots, \alpha). \quad (2.2)$$

Augmenting component indices $c_i \in \{1, \dots, m\}$ for $i = 1, \dots, n$, a simple Gibbs sampler can be used for posterior computation, alternating between sampling

- (i) c_i from a multinomial conditional posterior, for $i = 1, \dots, n$;
- (ii) $\theta_h \mid - \sim P_0(\theta_h) \prod_{i:c_i=h} \mathcal{K}(X_i; \theta_h)$; and
- (iii) $\pi \mid - \sim \text{Dirichlet}(\alpha + n_1, \dots, \alpha + n_m)$, with $n_h = \sum_{i=1}^n \mathbb{1}(c_i = h)$ for $h = 1, \dots, m$.

Relative to frequentist locally adaptive methods, Bayesian approaches are appealing in automatically providing a characterization of uncertainty in estimation, while having excellent practical performance for a broad variety of density shapes and dimensions. However, implementation typically relies on Markov chain Monte Carlo (MCMC), with the Gibbs sampler sketched above providing an example of a common algorithm used in practice. Unfortunately, current MCMC algorithms for posterior sampling in mixture models tend to face issues with *slow mixing*, meaning the sampler can take a very large number of iterations to adequately explore different posterior modes and obtain sufficiently accurate posterior summaries.

MCMC inefficiency has motivated a literature on faster approaches, including sequential approximations (Wang and Dunson, 2011; Zhang et al., 2014) and variational Bayes (Blei and Jordan, 2006). These methods are order dependent, tend to converge to local modes, and/or lack theory support. Newton and Zhang (1999); Newton (2002) instead rely on predictive recursion. Such estimators are fast to compute and have theory support, but are also order dependent and do not provide a characterization of uncertainty. Alternatively, one can use a Polya tree as a conjugate prior (Lavine, 1992, 1994), and there is a rich literature on related multiscale and recursive partitioning approaches, such as the optional Polya tree (Wong and Ma, 2010). However, Polya trees have disadvantages in terms of sensitivity to a base partition and a tendency to favor spiky/erratic densities. These disadvantages are inherited by most of the computationally fast modifications.

This article develops an alternative to current locally adaptive density estimators, obtaining the practical advantages of Bayesian approaches in terms of uncertainty quantification and a tendency to have relatively good performance for a wide variety of true densities, but without the computational disadvantage due to the use of MCMC. This is accomplished with a *Nearest Neighbor-Dirichlet Mixture* (NN-DM) model. The basic idea is to rely on fast nearest neighbor search algorithms to group the data into local neighborhoods, and then use these neighborhoods in defining a Bayesian mixture model-based approach. Section 3.3 outlines the NN-DM approach and describes implementation details for Gaussian kernels. Section 4.3 provides some theory support for NN-DM. Section 4.4 contains simulation experiments comparing NN-DM with a rich variety of competitors in univariate and multivariate examples, including an assessment of UQ performance. Section 2.5 contains a real data application on pulsar data.

2.2 Methodology

2.2.1 Nearest Neighbor Dirichlet Mixture Framework

Let $d(x_1, x_2)$ denote a distance metric between data points $x_1, x_2 \in \mathcal{X}$. For $\mathcal{X} = \mathbb{R}^p$, the Euclidean distance is typically chosen. For each $i \in \{1, 2, \dots, n\}$, let $X_{i[j]}$ denote the j th nearest neighbor to X_i in the data $\mathcal{X}^{(n)} = (X_1, \dots, X_n)$ such that $d(X_i, X_{i[1]}) \leq \dots \leq d(X_i, X_{i[n]})$, with ties broken by increasing order of indices. By convention, we define $X_{i[1]} = X_i$. The indices on the k nearest neighbors to X_i are denoted as $\mathcal{N}_i = \{j : d(X_i, X_j) \leq d(X_i, X_{i[k]})\}$. Denote the set of data points in the i th neighborhood by $\mathcal{S}_i = \{X_j : j \in \mathcal{N}_i\}$. In implementing the proposed method, we typically let the number of neighbors k vary as a function of n . When necessary, we use the notation k_n to express this dependence. However, we routinely drop the n subscript for notational simplicity.

Fixing $x \in \mathcal{X}$, we model the density of the data within the i th neighborhood using

$$f_i(x) = \mathcal{K}(x; \theta_i), \quad \theta_i \sim P_0, \quad (2.3)$$

where θ_i are parameters specific to neighborhood i that are given a global prior distribution P_0 . To combine the $f_i(x)$ s into a single global $f(x)$, similarly to equations (2.1)-(2.2), we let

$$f(x) = \sum_{i=1}^n \pi_i f_i(x), \quad \pi = (\pi_i)_{i=1}^n \sim \text{Dirichlet}(\alpha, \dots, \alpha), \quad \theta_i \sim P_0. \quad (2.4)$$

The key difference relative to standard Bayesian mixture model (2.1) is that in (2.4) we include one component for each data sample and assume that only the data in the k -nearest neighborhood of sample i will inform about θ_i . In contrast, (2.1) lacks any sample dependence, and we infer allocation of samples to mixture components in a posterior inference phase.

Given the restriction that only data in the i th neighborhood \mathcal{S}_i inform about θ_i , the pseudo-posterior density $\tilde{\Pi}_1(\theta_i; \mathcal{S}_i, P_0)$ of θ_i with data \mathcal{S}_i and prior P_0 is

$$\tilde{\Pi}_1(\theta_i; \mathcal{S}_i, P_0) \propto P_0(\theta_i) \prod_{j \in \mathcal{N}_i} \mathcal{K}(X_j; \theta_i), \quad (2.5)$$

where the right-hand side of (2.5) is motivated from Bayes' theorem. This pseudo-posterior is in a simple analytic form if P_0 is conjugate to $\mathcal{K}(x; \theta)$. The prior P_0 can involve unknown parameters and borrows information across neighborhoods; this reduces the large variance problem common to nearest neighbor estimators.

Since the neighborhoods are overlapping, proposing a pseudo-posterior update for π under (2.4) is not straightforward. However, one can define the number of effective members in the i th neighborhood \mathcal{S}_i similar in spirit to the number of points in the h th cluster in mixture models of the form (2.1). By convention, we define the point X_i that generated its neighborhood \mathcal{S}_i to be an effective member of that neighborhood. For any other data point X_j to be a effective member of the neighborhood generated by X_i for $j \neq i$, we require $X_j \in \mathcal{S}_i$ but $X_j \notin \mathcal{S}_u$ for all $u = 1, \dots, n$ such that $u \notin \{i, j\}$. That is, X_j lies in the neighborhood generated by X_i but does not lie in the neighborhood of any other X_u for $u \notin \{i, j\}$. In Section 2.3.2, we show that the number of effective member points defined as above approaches 1 as $n \rightarrow \infty$. This motivates the following Dirichlet pseudo-posterior density $\tilde{\Pi}_2(\pi; \mathcal{X}^{(n)})$ for the neighborhood weights π :

$$\tilde{\Pi}_2(\pi; \mathcal{X}^{(n)}) = \text{Dirichlet}(\pi \mid \alpha + 1, \dots, \alpha + 1), \quad (2.6)$$

where $\text{Dirichlet}(p \mid q_1, \dots, q_d)$ denotes the density of the Dirichlet distribution evaluated at p with parameters (q_1, \dots, q_d) . We provide a justification for the pseudo-posterior update (2.6) in Section 2.3.2. This distribution is inspired from the conditional posterior on the kernel weights in the Dirichlet mixture of equations (2.1)-(2.2), but we use n components and fix the effective number of samples allocated to each

component at one.

Based on equations (2.3)-(2.6), our nearest neighbor-Dirichlet mixture produces a pseudo-posterior distribution for the unknown density $f(x)$ through simple distributions for the parameters characterizing the density within each neighborhood and for the weights. To generate independent Monte Carlo samples from the pseudo-posterior for f , one can simply draw independent samples of $(\theta_i)_{i=1}^n$ and π from (2.5) and (2.6) respectively, and plug these samples into the expression for $f(x)$ in (2.4). The resulting mechanism can be described as

$$\begin{aligned} \theta_i &\stackrel{ind}{\sim} \tilde{\Pi}_1(\cdot; \mathcal{S}_i, P_0) \quad \text{for } i = 1, \dots, n, \\ \pi &\sim \tilde{\Pi}_2(\cdot; \mathcal{X}^{(n)}) \\ f(x) &= \sum_{i=1}^n \pi_i \mathcal{K}(x; \theta_i). \end{aligned} \tag{2.7}$$

In (2.7), we denote the induced pseudo-posterior distribution on f by $f \sim \tilde{\Pi}$. Although this is not exactly a coherent fully Bayesian posterior distribution, we claim that it can be used as a practical alternative to such a posterior in practice. This claim is backed up by theoretical arguments, simulation studies, and a real data application in the sequel.

2.2.2 Illustration with Gaussian Kernels

Suppose we have independent and identically distributed (iid) observations $\mathcal{X}^{(n)}$ from the density f , where $X_i \in \mathbb{R}^p$ for $i = 1, \dots, n$ and f is an unknown density function with respect to the Lebesgue measure on \mathbb{R}^p for $p \geq 1$. Let $\mathbb{R}_+^{p \times p}$ denote the set of all real-valued $p \times p$ positive definite matrices. Fix $x \in \mathbb{R}^p$. We proceed by setting $\mathcal{K}(x; \theta)$ to be the multivariate Gaussian density $\phi_p(x; \eta, \Sigma)$, given by

$$\phi_p(x; \eta, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \{ -(x - \eta)^\top \Sigma^{-1} (x - \eta) / 2 \},$$

where $\theta = (\eta, \Sigma)$, $\eta \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}_+^{p \times p}$. We first compute the neighborhoods \mathcal{N}_i corresponding to X_i as in Section 2.2.1 and place a normal-inverse Wishart (NIW) prior on $\theta_i = (\eta_i, \Sigma_i)$, given by $(\eta_i, \Sigma_i) \sim \text{NIW}_p(\mu_0, \nu_0, \gamma_0, \Psi_0)$ independently for $i = 1, \dots, n$. That is, we let

$$\eta_i \mid \Sigma_i \sim N\left(\mu_0, \frac{\Sigma_i}{\nu_0}\right), \quad \Sigma_i \sim \text{IW}_p(\gamma_0, \Psi_0),$$

with $\mu_0 \in \mathbb{R}^p$, $\nu_0 > 0$, $\gamma_0 > p - 1$ and $\Psi_0 \in \mathbb{R}_+^{p \times p}$; for details about parametrization see Section A.9 of the Appendix.

Monte Carlo samples from the pseudo-posterior of $f(x)$ can be obtained using Algorithm 1. The corresponding steps for the univariate case are provided in Section A.8 of the Appendix. For code, we developed the R package NNDM available at <https://github.com/shounakchattopadhyay/NN-DM>.

Algorithm 1. *Nearest neighbor-Dirichlet mixture algorithm to obtain Monte Carlo samples from the pseudo-posterior of $f(x)$ with Gaussian kernel and normal-inverse Wishart prior.*

- **Step 1:** For $i = 1, \dots, n$, compute the neighborhood \mathcal{N}_i for data point $X_i \in \mathbb{R}^p$ according to distance $d(\cdot, \cdot)$ with $(k - 1)$ nearest neighbors in $\mathcal{X}^{-i} = \mathcal{X}^{(n)} \setminus \{X_i\}$.
- **Step 2:** Update the parameters for neighborhood \mathcal{N}_i to $(\mu_i, \nu_n, \gamma_n, \Psi_i)$, where $\nu_n = \nu_0 + k$, $\gamma_n = \gamma_0 + k$,

$$\mu_i = \frac{1}{\nu_n} (\nu_0 \mu_0 + k \bar{X}_i), \quad \bar{X}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_i} X_j, \quad \text{and}$$

$$\Psi_i = \Psi_0 + \sum_{j \in \mathcal{N}_i} (X_j - \bar{X}_i)(X_j - \bar{X}_i)^\top + \frac{k\nu_0}{\nu_n} (\bar{X}_i - \mu_0)(\bar{X}_i - \mu_0)^\top.$$

- **Step 3:** To compute the t -th Monte Carlo sample $f^{(t)}(x)$ of $f(x)$, sample Dirichlet weights $\pi^{(t)} \sim \text{Dirichlet}(\alpha + 1, \dots, \alpha + 1)$ and neighborhood specific

parameters $(\eta_i^{(t)}, \Sigma_i^{(t)}) \sim NIW_p(\mu_i, \nu_n, \gamma_n, \Psi_i)$ independently for $i = 1, \dots, n$,
and set

$$f^{(t)}(x) = \sum_{i=1}^n \pi_i^{(t)} \phi_p(x; \eta_i^{(t)}, \Sigma_i^{(t)}).$$

Although the pseudo-posterior distribution of $f(x)$ lacks an analytic form, we can obtain a simple form for its pseudo-posterior mean by integrating over the pseudo-posterior distribution of $(\theta_i)_{i=1}^n$ and π . Recall the definitions of μ_i and Ψ_i from Step 2 of Algorithm 1 and define $\Lambda_i = \{\nu_n(\gamma_n - p + 1)\}^{-1}(\nu_n + 1) \Psi_i$. Then the pseudo-posterior mean of $f(x)$ is given by

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n t_{\gamma_n - p + 1}(x; \mu_i, \Lambda_i), \quad (2.8)$$

where $t_\gamma(x; \mu, \Lambda)$ for $x \in \mathbb{R}^p$ is the p -dimensional Student's t -density with degrees of freedom $\gamma > 0$, location $\mu \in \mathbb{R}^p$ and scale matrix $\Lambda \in \mathbb{R}_+^{p \times p}$. We proceed with using Gaussian kernels and NIW conjugate priors when implementing the NN-DM for the remainder of the chapter.

2.2.3 Hyperparameter Choice

The hyperparameters in the prior for the neighborhood-specific parameters need to be chosen carefully – we found results to be sensitive to γ_0 and Ψ_0 . If non-informative values are chosen for these key hyperparameters, we tend to inherit typical problems of nearest neighbor estimators including lack of smoothness and high variance. Suppose $\Sigma \sim IW_p(\gamma_0, \Psi_0)$ and for $i, j = 1, \dots, p$, let Σ_{ij} and $\Psi_{0,ij}$ denote the i, j th entry of Σ and Ψ_0 , respectively. Then $\Sigma_{jj} \sim \text{IG}(\gamma_*/2, \Psi_{0,jj}/2)$ where $\gamma_* = \gamma_0 - p + 1$. For $p = 1$, the $IW_p(\gamma_0, \Psi_0)$ density simplifies to an $\text{IG}(\gamma_0/2, \gamma_0 \delta_0^2/2)$ density with $\delta_0^2 = \Psi_0/\gamma_0$. Thus borrowing from the univariate case, we set $\Psi_{0,jj} = \gamma_* \delta_0^2$ and $\Psi_{0,ij} = 0$ for all $i \neq j$, which implies that $\Psi_0 = (\gamma_* \delta_0^2) \mathbb{I}_p$ and we use leave-one-out

cross-validation to select the optimum δ_0^2 . With p dimensional data, we recommend fixing $\gamma_0 = p$ which implies a multivariate Cauchy prior predictive density. We choose the leave-one-out log-likelihood as the criterion function for cross-validation, which is closely related to minimizing the Kullback-Leibler divergence between the true and estimated density (Hall, 1987; Bowman, 1984). The explicit expression for the pseudo-posterior mean in (2.8) makes cross-validation computationally efficient. The description of a fast implementation is provided in Section A.7 of the Appendix.

The proposed method has substantially faster runtime if one uses a default choice of hyperparameters. In particular, we found the default values $\mu_0 = 0_p, \nu_0 = 0.001, \gamma_0 = p$, and $\Psi_0 = \mathbb{I}_p$ to work well across a number of simulation cases, especially when the true density is smooth. Although using cross-validation to estimate Ψ_0 can lead to improved performance when the underlying density is spiky, cross-validation provides little to no gains for smooth true densities. Furthermore, with low sample size and increasing number of dimensions, we found this improvement to diminish rapidly. In order to obtain desirable uncertainty quantification in simulations and applications, we found small values of α to work well. As a default value, we recommend using $\alpha = 0.001$ for small samples and moderate dimensions.

The other key tuning parameter for NN-DM is the number of nearest neighbors $k = k_n$. The pseudo-posterior mean in (2.8) reduces to a single $t_{\gamma_n - p + 1}$ kernel if $k_n = n$. In contrast, $k_n = 1$ provides a sample smoothing kernel density estimate with a specific bandwidth function (Terrell and Scott, 1992). Therefore, the choice of k can impact the smoothness of the density estimate. To assess the sensitivity of the NN-DM estimate to the choice of k , we investigate how the out-of-sample log-likelihood of a test set changes with respect to k in Section 2.4.7. These simulations suggest that the proposed method is quite robust to the exact choice of k . In practice with finite samples and small dimensions, we recommend a default choice of $k_n = \lfloor n^{1/3} \rfloor + 1$ and $k_n = 10$ for univariate and multivariate cases, respectively. These values led to

good performance across a wide variety of simulation cases as described in Section 4.4.

2.3 Theoretical Properties

2.3.1 Asymptotic Results

There is a rich literature on asymptotic properties of the posterior measure for an unknown density under Bayesian models, providing a frequentist justification for Bayesian density estimation; refer, for example to Ghosal et al. (1999), Ghosal and van der Vaart (2007). Unfortunately, the tools developed in this literature rely critically on the mathematical properties of fully Bayes posteriors, providing theoretical guarantees for a computationally intractable exact posterior distribution under a Bayesian model. Our focus is instead on providing frequentist asymptotic guarantees for our computationally efficient NN-DM approach, with this task made much more complex by the dependence across neighborhoods induced by the use of a nearest neighbor procedure.

We first focus on proving pointwise consistency of the pseudo-posterior of $f(x)$ induced by (2.7) for each $x \in [0, 1]^p$, using Gaussian kernels as in Section 2.2.2. We separately study the mean and variance of the NN-DM pseudo-posterior distribution, first showing that the pseudo-posterior mean in (2.8) is pointwise consistent and then that the pseudo-posterior variance vanishes asymptotically. The key idea behind our proof is to show that the pseudo-posterior mean is asymptotically close to a kernel density estimator with suitably chosen bandwidth for fixed p and $k_n \rightarrow \infty$ at a desired rate. The proof then follows from standard arguments leading to consistency of kernel density estimators. The NN-DM pseudo-posterior mean mimics a kernel density estimator only in the asymptotic regime; in finite sample simulation studies (refer to Section 4.4), NN-DM has much better performance. The detailed proofs of all results in this section are in the Appendix.

Consider independent and identically distributed data $\mathcal{X}^{(n)}$ from a fixed unknown density f_0 with respect to the Lebesgue measure on \mathbb{R}^p equipped with the Euclidean metric, inducing the measure P_{f_0} on $\mathcal{B}(\mathbb{R}^p)$. We use $\tilde{E}\{f(x)\}$, $\tilde{\text{var}}\{f(x)\}$, and $\tilde{\text{pr}}\{f(x) \in B\}$ to denote the mean of $f(x)$, variance of $f(x)$, and probability of the event $\{f(x) \in B\}$ for $B \in \mathcal{B}(\mathbb{R}^p)$, respectively, under the pseudo-posterior distribution of $f(x)$ implied by (2.7). We make the following regularity assumptions on f_0 :

Assumption 1 (Compact support). f_0 is supported on $[0, 1]^p$.

Assumption 2 (Bounded gradient). f_0 is continuous on $[0, 1]^p$ with $\|\nabla f_0(x)\|_2 \leq L$ for all $x \in [0, 1]^p$ and some finite $L > 0$.

Assumption 3 (Bounded sup-norm). $\|\log(f_0)\|_\infty < \infty$.

Our asymptotic analysis relies on analyzing the behavior of the pseudo-posterior updates within each nearest neighborhood. We leverage on key results from Biau and Devroye (2015); Evans et al. (2002) which are based on the assumption that the true density has compact support as in Assumption 1. Assumption 2 ensures that the kernel density estimator has finite expectation. Versions of this assumption are common in the kernel density literature; for example, refer to Tsybakov (2009). Assumptions 1 and 3 imply the existence of $0 < a_1, a_2 < \infty$ such that $0 < a_1 < f_0(x) < a_2 < \infty$ for all $x \in [0, 1]^p$, which is referred to as a positive density condition by Evans (2008); Evans et al. (2002). This is used to establish consistency of the proposed method and justify the choice of the pseudo-posterior distribution of the weights. These assumptions are standard in the literature studying frequentist asymptotic properties of nearest neighbor and Bayesian density estimators.

For $i = 1, \dots, n$, recall the definitions of μ_i and Λ_i from (2.8):

$$\mu_i = \frac{\nu_0}{\nu_n} \mu_0 + \frac{k_n}{\nu_n} \bar{X}_i, \quad \Lambda_i = \frac{\nu_n + 1}{\nu_n(\gamma_n - p + 1)} \Psi_i,$$

where $\nu_n = \nu_0 + k_n$, $\gamma_n = \gamma_0 + k_n$, and \bar{X}_i, Ψ_i are as in Algorithm 1. Define the bandwidth matrix

$$H_n = h_n^2 \mathbb{I}_p, \quad \text{where } h_n^2 = \frac{(\nu_n + 1)(\gamma_0 - p + 1)}{\nu_n(\gamma_n - p + 1)} \delta_0^2. \quad (2.9)$$

We have suppressed the dependence of μ_i and Λ_i on n for notational convenience. It is immediate that $h_n^2 \rightarrow 0$ if $k_n \rightarrow \infty$ as $n \rightarrow \infty$. Fix $x \in [0, 1]^p$. To prove consistency of the pseudo-posterior mean, we first show that $\hat{f}_n(x)$ and $f_K(x) = (1/n) \sum_{i=1}^n t_{\gamma_n - p + 1}(x; X_i, H_n)$ are asymptotically close, that is we show that $E_{P_{f_0}}(|\hat{f}_n(x) - f_K(x)|) \rightarrow 0$ as $n \rightarrow \infty$. To obtain this result, we approximate μ_i by X_i and Λ_i by H_n using successive applications of the mean value theorem. Finally, we exploit the convergence of $f_K(x)$ to the true value $f_0(x)$ to obtain the consistency of $\hat{f}_n(x)$. The proof of convergence of $f_K(x)$ to $f_0(x)$ is provided in Section A.6 of the Appendix. The precise statement regarding the consistency of the pseudo-posterior mean is given in the following theorem. Let $a \wedge b$ denote the minimum of a and b .

Theorem 1. *Fix $x \in [0, 1]^p$. Let $k_n = o(n^{i_0})$ with $i_0 = \{2/(p^2 + p + 2)\} \wedge \{4/(p + 2)^2\}$ such that $k_n \rightarrow \infty$ as $n \rightarrow \infty$, and $\nu_0 = o\{n^{-2/p} k_n^{(2/p)+1}\}$. Then, $\hat{f}_n(x) \rightarrow f_0(x)$ in P_{f_0} -probability as $n \rightarrow \infty$.*

We now look at the pseudo-posterior variance of $f(x)$. We let

$$R_n = \frac{\Gamma\{(\gamma_n - p + 2)/2\}}{\Gamma\{(\gamma_n - p + 1)/2\}} \left[\frac{\nu_n + 2}{4\pi\nu_n(\gamma_n - p + 2)} \right]^{p/2} \quad \text{and} \quad D_n = \frac{(\gamma_n - p + 1)(\nu_n + 2)}{2(\gamma_n - p + 2)(\nu_n + 1)}. \quad (2.10)$$

For $i = 1, \dots, n$, let $B_i = D_n \Lambda_i$ and define

$$\hat{f}_{var}(x) = \frac{1}{n} \sum_{i=1}^n t_{\gamma_n - p + 2}(x; \mu_i, B_i). \quad (2.11)$$

As $n \rightarrow \infty$, we have $D_n \rightarrow 1/2$. Analogous steps to the ones used in the proof of Theorem 1 can be used to imply that $\hat{f}_{var}(x) \rightarrow f_0(x)$ in P_{f_0} -probability. Also,

as $n \rightarrow \infty$, $k_n^{(p-1)/2} R_n = \mathcal{O}(1)$ using Stirling's approximation. We now provide an upper bound on the pseudo-posterior variance of $f(x)$ which shows convergence of the pseudo-posterior variance to 0.

Theorem 2. *Let H_n be the bandwidth matrix defined in (2.9). Let R_n, D_n be as in (2.10) and \hat{f}_{var} be as in (2.11). Under Assumptions 1-3 with x, k_n , and ν_0 as in Theorem 1, we have*

$$\widetilde{\text{var}}\{f(x)\} \leq \frac{R_n D_n^{-p/2} \hat{f}_{var}(x)}{|H_n|^{1/2}} \left\{ \frac{1}{n(\alpha + 1) + 1} + \frac{1}{n} \right\}. \quad (2.12)$$

This implies $\widetilde{\text{var}}\{f(x)\} \rightarrow 0$ in P_{f_0} -probability as $n \rightarrow \infty$.

Refer to Sections A.2 and A.3 in the Appendix for proofs of Theorems 4 and 5, respectively. Pointwise pseudo-posterior consistency follows from Theorems 1 and 2, as shown below.

Theorem 3. *Let f_0 satisfy Assumptions 1-3 with x, k_n and ν_0 as in Theorem 1. Fix $\epsilon > 0$ and define the ϵ -ball around $f_0(x)$ by $U_\epsilon = \{y_* : |y_* - f_0(x)| \leq \epsilon\}$. Let $\tilde{\text{pr}}\{f(x) \in U_\epsilon^c\}$ denote the probability of the set U_ϵ^c under the pseudo-posterior distribution of $f(x)$ as induced by (2.7). Then $\tilde{\text{pr}}\{f(x) \in U_\epsilon^c\} \rightarrow 0$ in P_{f_0} -probability as $n \rightarrow \infty$.*

Proof. Fix $\epsilon > 0$ and consider the ϵ -ball $U_\epsilon = \{y_* : |y_* - f_0(x)| \leq \epsilon\}$. Then by Chebychev's inequality, we have $\tilde{\text{pr}}\{f(x) \in U_\epsilon^c\} \leq [(\hat{f}_n(x) - f_0(x))^2 + \widetilde{\text{var}}\{f(x)\}]/\epsilon^2 \rightarrow 0$ in P_{f_0} -probability as $n \rightarrow \infty$, using Theorems 1 and 2. \square

We next focus on the limiting distribution of $f(x)$ for the univariate case. From Section A.8 of the Appendix, the pseudo-posterior distribution of (η_i, σ_i^2) for $i = 1, \dots, n$ is given by $\text{NIG}(\mu_i, \nu_n, \gamma_n/2, \gamma_n \delta_i^2/2)$, where μ_i, ν_n, γ_n are as before and

$$\gamma_n \delta_i^2 = \gamma_0 \delta_0^2 + \sum_{j \in \mathcal{N}_i} (X_j - \bar{X}_i)^2 + \frac{k_n \nu_0}{\nu_n} (\bar{X}_i - \mu_0)^2.$$

We establish in Theorem 4 that the limiting distribution of $f(x)$ is a Gaussian distribution with appropriate centering and scaling. This allows interpretation of $100(1 - \beta)\%$ pseudo-credible intervals as $100(1 - \beta)\%$ frequentist confidence intervals on average for large n .

Theorem 4. *Fix $x \in [0, 1]$. Suppose f_0 satisfies Assumptions 1-3 and also satisfies $|f_0^{(4)}(x)| \leq C_0$ for all $x \in [0, 1]$ for some finite $C_0 > 0$. Let k_n satisfy $k_n = o(n^{2/7})$ such that $n^{-2/9}k_n \rightarrow \infty$, h_n be as in (2.9) satisfying $h_n \rightarrow 0$, and $\alpha = \alpha_n \rightarrow \infty$, as $n \rightarrow \infty$. For $t \in \mathbb{R}$, define*

$$G_n(t) = \tilde{\text{pr}} \left[(nh_n)^{1/2} \left\{ f(x) - \left(f_0(x) + \frac{h_n^2 f_0^{(2)}(x)}{2} \right) \right\} \leq t \right].$$

Then, we have

$$\lim_{n \rightarrow \infty} E_{P_{f_0}} \{G_n(t)\} = \Phi \left(t; 0, \frac{f_0(x)}{2\pi^{1/2}} \right),$$

where $\Phi(t; 0, \sigma^2)$ denotes the cumulative distribution function of the $N(0, \sigma^2)$ density.

For a proof of Theorem 4, we refer the reader to Section A.4 of the Appendix.

2.3.2 Pseudo-Posterior Distribution of Weights

We investigate the rationale behind the pseudo-posterior update (2.6) of the weight π , which has a symmetric prior distribution $\pi \sim \text{Dirichlet}(\alpha, \dots, \alpha)$ as motivated in Section 2.1. As discussed in Section 2.1, the conditional update for the weights π in a finite Bayesian mixture model with m components given the cluster allocation indices $\{c_1, \dots, c_n\}$ is obtained by $\text{Dirichlet}(\alpha + n_1, \dots, \alpha + n_m)$, where α is the prior concentration parameter and $n_h = \sum_{i=1}^n \mathbb{1}(c_i = h)$ is the number of data points allocated to the h th cluster. This is not true in our case as the k_n -nearest neighborhoods have considerable overlap between them. Instead, we consider the number of effective member data points in each of these neighborhoods.

Define the k_n -nearest neighborhood of X_i to be the set $\mathcal{S}_i = \{X_j : d(X_i, X_j) \leq d(X_i, X_{i[k_n]})\}$ where $X_{i[k_n]}$ is the k_n -th nearest neighbor of X_i in the data $\mathcal{X}^{(n)}$, following the notation in Section 2.2.1. We assume $d(\cdot, \cdot)$ is the Euclidean metric from here on, and let $R_i = d(X_i, X_{i[k_n]}) = \|X_i - X_{i[k_n]}\|_2$ denote the distance of X_i from its k_n -th nearest neighbor in $\mathcal{X}^{(n)}$.

Let N_i denote the number of effective members in \mathcal{S}_i as defined in Section 2.2.1. Then, we can express N_i as

$$N_i = 1 + \sum_{j \neq i} \mathbb{1} \left[X_j \in \mathcal{S}_i, \bigcap_{u \notin \{i, j\}} \{X_j \notin \mathcal{S}_u\} \right], \quad (2.13)$$

where $\mathbb{1}(A)$ is the indicator function of the set A . Under $X_1, \dots, X_n \stackrel{iid}{\sim} f_0$, we have

$$E_{P_{f_0}}(N_1) = 1 + (n-1)P_{f_0} \left[X_2 \in \mathcal{S}_1, \bigcap_{u=3}^n \{X_2 \notin \mathcal{S}_u\} \right], \quad (2.14)$$

by symmetry. Furthermore, N_i are identically distributed for $i = 1, \dots, n$. We now state a result which provides a motivation for our choice of the pseudo-posterior update of π . For two sequences of real numbers (a_n) and (b_n) , we write $a_n \sim b_n$ if $|a_n/b_n| \rightarrow c_0$ as $n \rightarrow \infty$ for some constant $c_0 > 0$.

Theorem 5. *Suppose $X_1, \dots, X_n \stackrel{iid}{\sim} f_0$ with f_0 satisfying Assumptions 1-3. Furthermore, suppose that $k_n \sim n^{i_0 - \epsilon}$ for some $\epsilon \in (0, i_0)$, where i_0 is as defined in Theorem 1. Then,*

$$\lim_{n \rightarrow \infty} n P_{f_0} \left[X_2 \in \mathcal{S}_1, \bigcap_{u=3}^n \{X_2 \notin \mathcal{S}_u\} \right] = 0. \quad (2.15)$$

Proof of Theorem 5 is in Section A.5 of the Appendix. The above theorem suggests we asymptotically have only one effective member per neighborhood \mathcal{S}_i , namely the point X_i that itself generated this neighborhood. This result motivates

our choice of the pseudo-posterior update of the weight vector π . We illustrate uncertainty quantification of the proposed method in finite samples in Section 2.4.4 with this choice of pseudo-posterior update of the weight vector π .

2.4 Simulation Experiments

2.4.1 Preliminaries

In this section, we compare the performance of the proposed density estimator with several other standard density estimators through several numerical experiments. We evaluate estimation performance based on the expected \mathcal{L}_1 distance (Devroye and Györfi, 1985). For the pair (f_0, \hat{f}) , where f_0 is the true data generating density and \hat{f} is an estimator, the expected \mathcal{L}_1 distance is defined as $\mathcal{L}_1(f_0, \hat{f}) = E_{P_{f_0}}\{\int |f_0(x) - \hat{f}(x)| dx\}$. We compute an estimate $\hat{\mathcal{L}}_1(f_0, \hat{f})$ of $\mathcal{L}_1(f_0, \hat{f})$ in two steps. First, we sample n training points $X_1, \dots, X_n \sim f_0$ and obtain \hat{f} based on this sample, and then further sample n_t independent test points $X_{n+1}, \dots, X_{n+n_t} \sim f_0$ and compute

$$\hat{L} = \frac{1}{n_t} \sum_{i=1}^{n_t} \left| \frac{\hat{f}(X_{n+i})}{f_0(X_{n+i})} - 1 \right|.$$

In the second step, to approximate the expectation with respect to P_{f_0} , the first step is repeated R times. Letting \hat{L}_r denote the estimate for the r th replicate, we compute the final estimate as $\hat{\mathcal{L}}_1(f_0, \hat{f}) = (1/R) \sum_{r=1}^R \hat{L}_r$. Then, it follows that $\hat{\mathcal{L}}_1(f_0, \hat{f}) \rightarrow \mathcal{L}_1(f_0, \hat{f})$ as $n_t, R \rightarrow \infty$, by the law of large numbers. In our experiments, we set $n_t = 500$ and $R = 20$. We let 0_p and 1_p denote the vector with all entries equal to 0 and the vector with all entries equal to 1 in \mathbb{R}^p , respectively, for $p \geq 1$.

All simulations were carried out using the R programming language (R Core Team, 2021). For Dirichlet process mixture models, we collect 2,000 Markov chain Monte Carlo (MCMC) samples after discarding a burn-in of 3,000 samples using the `dirichletprocess` package (J. Ross and Markwick, 2019). The default implementa-

tion of the Dirichlet process mixture model in p dimensions in the `dirichletprocess` package uses multivariate Gaussian kernels and has the base measure as $\text{NIW}_p(0_p, p, p, \mathbb{I}_p)$ with the Dirichlet concentration parameter having the $\text{Gamma}(2, 4)$ prior (West, 1992). For the nearest neighbor-Dirichlet mixture, 1,000 Monte Carlo samples are taken. For the kernel density estimator, we select the bandwidth by the default plug-in method `hpi` for univariate cases and `Hpi` for multivariate cases (Sheather and Jones, 1991; Wand and Jones, 1994) using the package `ks` (Duong, 2020). We additionally consider the k -nearest neighbor estimator studied in Mack and Rosenblatt (1979), setting the number of neighbors $k = n^{1/2}$, and the variational Bayes (VB) approximation to Dirichlet process mixture models (Blei and Jordan, 2006). We also compare with the optional Polya tree (OPT) (Wong and Ma, 2010) using the package `PTT`. For univariate cases, we consider the recursive predictive density estimator (RD) from Hahn et al. (2018), Polya tree mixtures (PTM) using the package `DPpackage` (Jara et al., 2011), and the sample smoothing kernel density estimator (A-KDE) using the package `quantreg`. Lastly, we also compare with the local likelihood density estimator (LLDE) using the package `locfit` for both univariate and multivariate cases. Dirichlet process mixture model hyperparameter values are kept the same in both the MCMC and variational Bayes implementations, with the number of components of the variational family set to 10 for all cases. We denote the nearest neighbor-Dirichlet mixture, Dirichlet process mixture (DPM) implemented with MCMC, kernel density estimator, variational Bayes approximation to the DPM, and k -nearest neighbor density estimator by NN-DM, DP-MC, KDE, DP-VB, and KNN, respectively, in tables and figures.

2.4.2 Univariate Cases

We set $n = 200, 500$ with $k_n = \lfloor n^{1/3} \rfloor + 1$. We consider 10 choices of f_0 from the R package `benchden` (Mildenberger and Weinert, 2012); the specific choices are Cauchy

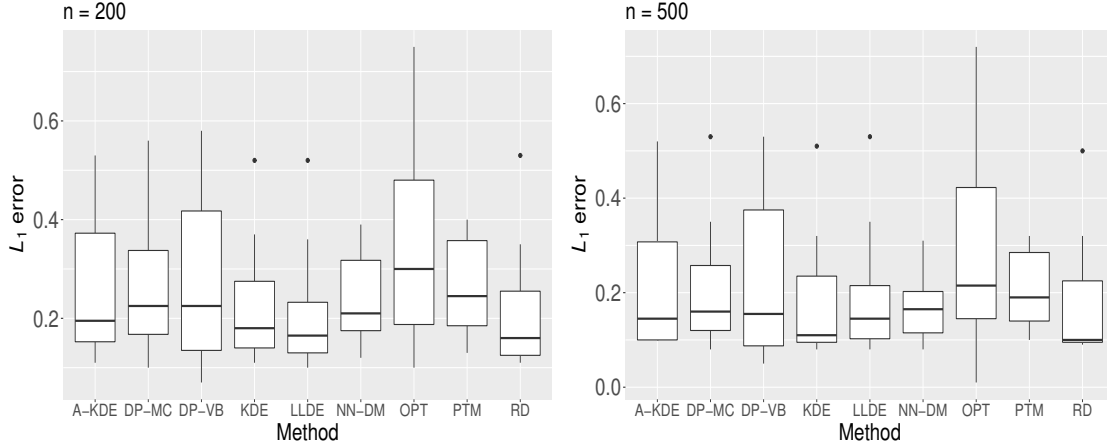


FIGURE 2.1: Box plots of $\hat{\mathcal{L}}_1(f_0, \hat{f})$ for the 10 different choices of the true density f_0 and different estimators \hat{f} for univariate data. The box plots for KDE and RD exclude the heavy-tailed cases CA, IE, and SP.

(CA), claw (CW), double exponential (DE), Gaussian (GS), inverse exponential (IE), lognormal (LN), logistic (LO), skewed bimodal (SB), symmetric Pareto (SP), and sawtooth (ST) with default choices of the corresponding parameters. The prior hyperparameter choices for the proposed method are $\mu_0 = 0, \nu_0 = 0.001, \gamma_0 = 1; \delta_0^2$ is chosen via the cross-validation method of Section 2.2.3. Detailed numerical results are deferred to Table A.1 in the Appendix. Instead, in Figure 2.1, we provide a visual summary of the performance of each method under consideration by forming a box plot of the estimated \mathcal{L}_1 errors of the methods across all the data generating densities. Methods with lower median as indicated by the solid line of the box plot, and smaller overall spread are preferable as they provide higher accuracy and also maintain such accuracy across a collection of true density cases. Results of KNN are omitted in Figure 2.1 due to much higher values compared to other methods. For the KDE and RD estimator, the plot and the table exclude the results for the heavy-tailed densities CA, IE, and SP due to very high \mathcal{L}_1 errors.

Overall, a major advantage of the proposed method is its versatility among the considered methods. The Bayesian nonparametric methods DP-MC, DP-VB, PTM,

OPT, and RD are often close to NN-DM in terms of their performance when the true densities are smooth and do not display locally spiky behavior. However, the NN-DM performs better than other methods in densities where such local behavior is present and performs very close to the best estimator for either the smooth heavy-tailed or thin-tailed densities. The KDE and RD perform well when data are generated from a smooth underlying density. However, there are some cases where the error for KDE and RD is very high. For instance, when $n = 500$ and f_0 is the standard Cauchy (CA) density, the estimated \mathcal{L}_1 error for the KDE is 38501.85 and the algorithm for the RD estimate did not converge. Both the KDE and RD also perform poorly in very spiky multi-modal densities such as the ST. Compared to the LLDE and the A-KDE, the NN-DM displays similar performance in heavy-tailed and smooth densities when $n = 200$, with the NN-DM performing better for the spiky densities. However, when $n = 500$, the NN-DM shows significant improvements over the LLDE and the A-KDE for spiky densities such as the CW and the ST.

In Figure 2.2, we show the performance of the NN-DM estimator \hat{f}_n (with hyperparameters chosen as described earlier) relative to the posterior mean under a DP-MC with default or hand-tuned hyperparameters, when 500 data points are generated from the sawtooth (ST) density. The Dirichlet process mixture with default hyperparameters is unable to detect the multiple spikes, merging adjacent modes to form larger clusters, perhaps due to inadequate mixing of the Markov chain Monte Carlo sampler or to the Gaussian kernels used in the mixture. As a result, we had to hand-tune the hyperparameters for the Dirichlet process mixture to obtain comparable performance with the NN-DM (without hand-tuning). We obtained the best results when changing the hyperparameters of the base measure of the DP-MC to $\text{NIG}(0, 0.01, 1, 1)$ while keeping the prior on α the same as before. This illustrates the deficiency of the DP-MC in estimating densities with spiky local behavior unless we hand-tune the hyperparameters, which requires knowledge of the true density.

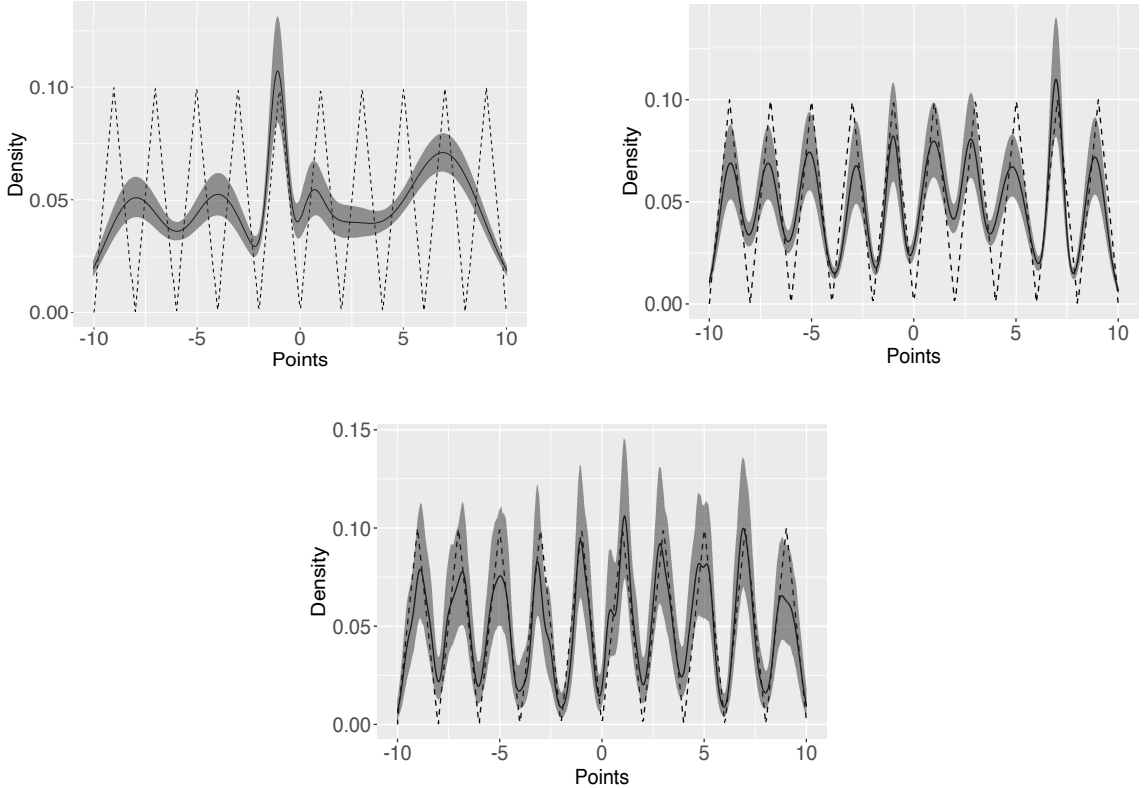


FIGURE 2.2: Plot comparing density estimates for the NN-DM and DP-MC for $n = 500$ samples generated from the sawtooth (ST) density. Shaded regions correspond to 95% (pseudo) posterior credible intervals. The true density is displayed using dotted lines. The top panel shows the performance of DP-MC with default hyperparameters on the left and with hand-tuned hyperparameters on the right. The bottom panel shows the performance of the NN-DM.

We also compare the performance of the two methods with a smoother test density in Figure 2.3, where the data are generated from a skewed bimodal (SB) distribution. Both the estimates are comparable, but the nearest neighbor-Dirichlet mixture provides better uncertainty quantification. Similar results are obtained for $n = 1000$, and hence are omitted.

2.4.3 Multivariate Cases

For the multivariate cases, we consider $n = 200$ and 1000 . The number of neighbors is set to $k = 10$ and the dimension p is chosen from $\{2, 3, 4, 6\}$. Recall the definition

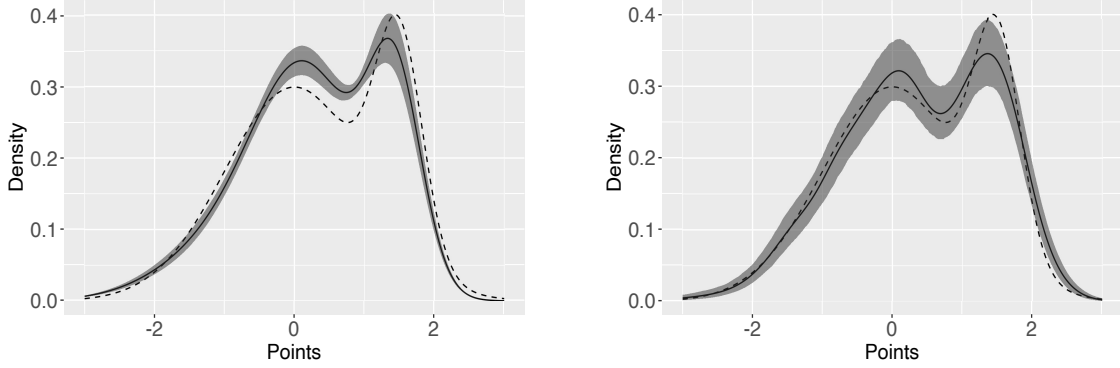


FIGURE 2.3: Similar to Figure 2.2, with data of sample size $n = 500$ generated from the skewed bimodal (SB) density. Left panel shows the DP-MC fit and the right panel shows the NN-DM fit.

of $\phi_p(x; \mu, \Sigma)$ from Section 2.2.2 and let $\Phi(x)$ be the cumulative distribution function of the standard Gaussian density. Let $S_0 = \rho 1_p 1_p^\top + (1 - \rho) \mathbb{I}_p$ with $\rho = 0.8$. Let $x = (x_1, \dots, x_p)^\top$. We consider the following cases.

(1) *Mixture of Gaussians (MG)*: $f_0(x) = 0.4 \phi_p(x; m_1, S_0) + 0.6 \phi_p(x; m_2, S_0)$, where $m_1 = -2 \times 1_p, m_2 = 2 \times 1_p$.

(2) *Skew normal (SN)*: $f_0(x) = 2\phi_p(x; m_0, S_0)\Phi\{s_0^\top W^{-1}(x - m_0)\}$ (Azzalini, 2005), where W is the diagonal matrix with diagonal entries $W_{ii}^2 = S_{0,ii}$ for $i = 1, \dots, p$. We choose $m_0 = 0_p$ and the skewness parameter vector $s_0 = 0.5 \times 1_p$.

(3) *Multivariate t-distribution (T)*: $f_0(x) = t_{d_0}(x; m_*, S_0)$ is the density of the p -dimensional multivariate Student's t-distribution. We set $d_0 = 10$ and $m_* = 1_p$.

(4) *Mixture of multivariate skew t-distributions (MST)*: $f_0(x) = 0.25 t_{d_0}(x; m_1, S_0, s_0) + 0.75 t_{d_0}(x; m_2, S_0, s_0)$. Here, $t_d(\cdot; \mu, S, s)$ is the skew t-density (Azzalini, 2005) with parameters d, μ, S, s , with d_0, s_0 defined as before and m_1, m_2 the same as in the first case.

(5) *Multivariate Cauchy (MVC)*: $f_0(x) \propto \{1 + (x - \mu_*)^\top S_0^{-1}(x - \mu_*)\}^{-1}$ where $\mu_* = 0_p$.

(6) *Multivariate Gamma (MVG)*: $f_0(x) \propto c_\Phi(F_1(x_1), \dots, F_p(x_p) | S_0) \prod_{j=1}^p f_j(x_j; \gamma_{j1}, \gamma_{j2})$ where f_j and F_j denote the density and distribution function of the univariate

gamma distribution with shape parameter γ_{j1} and rate parameter γ_{j2} , respectively, for $j = 1, \dots, p$ and $c_\Phi(\cdot | \Gamma)$ is as described in Song (2000). This is a Gaussian copula based construction of the multivariate gamma distribution. We set $\gamma_{j1} = \gamma_{j2} = 1$ for $j = 1, \dots, p$.

The hyperparameters for the nearest neighbor-Dirichlet mixture are chosen as $\mu_0 = 0_p, \nu_0 = 0.001, \gamma_0 = p$, and $\Psi_0 = \{(\gamma_0 - p + 1)\delta_0^2\}\mathbb{I}_p = \delta_0^2 \mathbb{I}_p$, where the optimal δ_0^2 is chosen via cross-validation as described in Section 2.2.3. Default hyperparameters as described in Section 2.4.1 are chosen for the MCMC and VB implementations of the DPM.

Similar to the univariate case, we defer the numerical results to Table A.2 in the Appendix and in Figure 2.4 display a visual summary consisting of box plot of estimated \mathcal{L}_1 errors over the densities considered. The proposed method is very robust against a wide selection of true distributions, with its \mathcal{L}_1 error scaling nicely with the dimension. The KDE shows a noticeably sharp decline in performance - when the dimension is changed from 2 to 6, the average increase in \mathcal{L}_1 error is by factors of about 5 and 7 for sample sizes 200 and 1000, respectively. This is possibly due to lack of adaptive density estimation in higher dimensions using a single bandwidth matrix, since data in \mathbb{R}^p become increasingly sparse with increasing p . As in the univariate case, we had to exclude the MVC density for the KDE due to the algorithm not converging. The performances of NN-DM, DP-MC, and DP-VB are quite competitive across densities, with NN-DM faring better than the DP-VB when estimating densities such as the MVC and the MVG. Furthermore, the NN-DM is hit the least significantly by the curse of dimensionality out of the three. This is particularly prominent for the DP-MC when the true density is either MG or MST with $n = 200$ and $p = 6$, and for the DP-VB when the true density is MVC. It is also important to keep in mind that the NN-DM provides similar results compared to the DP-MC while being at least an order of magnitude faster,

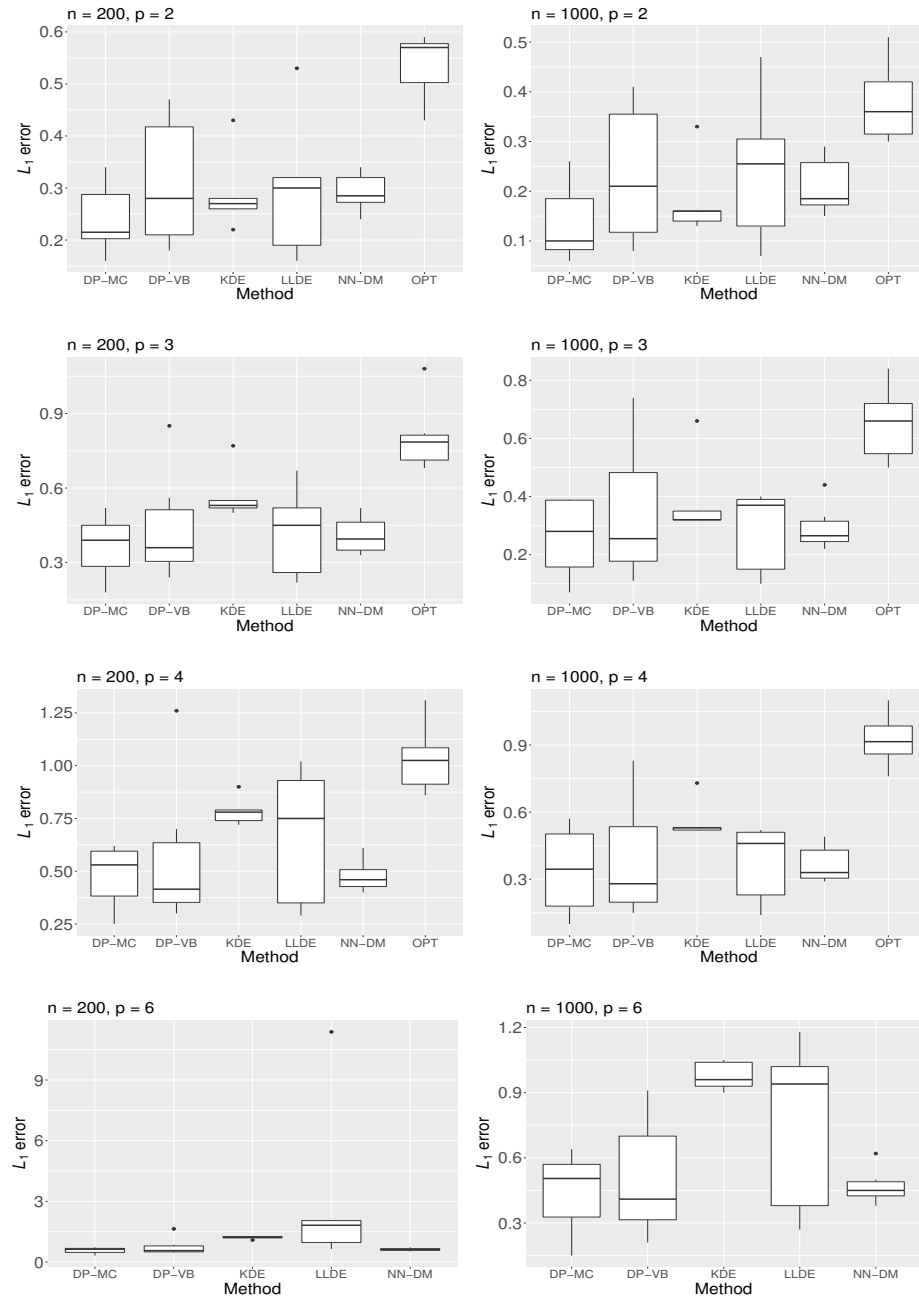


FIGURE 2.4: Box plots of $\hat{\mathcal{L}}_1(f_0, \hat{f})$ for the 6 different choices of the true density f_0 and different estimators \hat{f} for multivariate data. The box plots for KDE and LLDE exclude the MVC density. The box plots for $p = 6$ exclude results from OPT.

as illustrated in Section 2.4.6. The performance of the OPT is hit quite significantly as the number of dimensions increases, along with the algorithm not converging for $p = 6$. The LLDE provides competitive results with the NN-DM in lower dimensions. However, in higher dimensions, the LLDE often does not converge, indicating lack of stability of the algorithm. We reported the average of the replicates for which the algorithm did converge. The results suggest that the performance of the LLDE is also affected quite drastically with increasing dimensions. When compared across all data generating cases considering the variation in densities, dimensions and sample sizes, the proposed method is seen to be more versatile than its competitors.

2.4.4 Accuracy of Uncertainty Quantification

In this section, we assess frequentist coverage of 95% pseudo-posterior credible intervals for the NN-DM and compare with coverage based on the 95% posterior credible intervals obtained from DP-MC and DP-VB. Ghosal and van der Vaart (2017) recommend investigating the frequentist coverage of Bayesian credible intervals. We do not include frequentist coverage for Polya tree mixtures (PTMs) and the optional Polya tree (OPT) due to the lack of available code. We consider the cases $p \in \{1, 2\}$ in our experiments with sample size $n = 500$. For each choice of density f_0 , we fix $n_t = 200$ test points $\mathcal{X}_t = \{X_{t1}, \dots, X_{tn_t}\}$ generated from the density f_0 . With these fixed test points, we generate $n = 500$ data points in our sample for $R_{cov} = 200$ times and check the coverage of posterior/pseudo-posterior credible intervals obtained from the three methods. We implement the DP-MC with base measure $\text{NIW}_p(0_p, 0.01, p, \mathbb{I}_p)$ and a $\text{Gamma}(2, 4)$ prior on the concentration parameter as in West (1992). These choices of hyperparameters were seen to give better frequentist coverage results than using the default values used in Sections 2.4.2 and 2.4.3. Same choices of hyperparameters are maintained for DP-VB. For the NN-DM, we take $k = 8$ in the univariate case and $k = 5$ in the bivariate case, $\alpha = 0.001$, and other hyperparameters chosen as be-

fore. We report the average coverage probability and average length of the (pseudo) credible intervals across all the points in the test data \mathcal{X}_t in Tables 2.1 and 2.2 for the univariate and bivariate cases, respectively.

For univariate densities, both the DP-MC and DP-VB display severe under-coverage. In most of the cases, the DP-VB and NN-DM have similar width of (pseudo) credible intervals but the DP-VB displays dramatically lower coverage than the NN-DM. The under-coverage displayed by the DP-MC may be due to MCMC mixing issues. The NN-DM shows near nominal coverage in the smooth Gaussian (GS) and lognormal (LN) densities, while also attaining near nominal coverage in the skewed bimodal (SB), claw (CW), and sawtooth (ST) densities which are multi-modal. The shortcomings of DP-MC and DP-VB are especially noticeable when dealing with spiky densities such as the claw or sawtooth. For bivariate cases considered in Table 2.2 we see a similar trend; the NN-DM method provides uniformly better uncertainty quantification across all the densities considered. It is clear that in terms of frequentist uncertainty quantification, the NN-DM displays vastly superior coverage to the DP-MC and the DP-VB without inflating the interval width.

2.4.5 Comparison for High Dimensional Data

In addition to the above experiments, we performed a simulation experiment for high-dimensional data. Specifically, we set $n = 1000$, $p = 50$, and consider the same set of true densities in Section 2.4.3. We compared results from the proposed NN-DM method and the DP-VB. Due to severe computational time, we did not consider the DP-MC in this scenario. We also tried optional Polya trees (Wong and Ma, 2010) using the PTT package; however, the current implementation of the

Table 2.1: Comparison of the frequentist coverage of 95% (pseudo) posterior credible intervals of the nearest neighbor-Dirichlet mixture and the MCMC and variational implementations of the Dirichlet process mixture for univariate data. Average length of the intervals are also provided for each case within parentheses. Number of replications and sample size are $R_{cov} = 200$ and $n_{cov} = 500$, respectively.

Method	CA	CW	DE	GS	IE
NN-DM	0.75 (0.05)	0.89 (0.21)	0.75 (0.06)	0.92 (0.08)	0.81 (0.11)
DP-MC	0.48 (0.02)	0.06 (0.01)	0.35 (0.02)	0.37 (0.01)	0.39 (0.04)
DP-VB	0.33 (0.05)	0.18 (0.07)	0.28 (0.07)	0.79 (0.05)	0.14 (0.04)

Method	LN	LO	SB	SP	ST
NN-DM	0.92 (0.17)	0.81 (0.03)	0.88 (0.10)	0.72 (0.01)	0.91 (0.05)
DP-MC	0.31 (0.05)	0.55 (0.03)	0.46 (0.03)	0.46 (0.01)	0.64 (0.03)
DP-VB	0.19 (0.15)	0.10 (0.03)	0.40 (0.10)	0.20 (0.01)	0.07 (0.01)

Table 2.2: Comparison of the frequentist coverage of 95% (pseudo) posterior credible intervals of the nearest neighbor-Dirichlet mixture and the MCMC and variational implementations of the Dirichlet process mixture for bivariate data. Average length of the intervals are also provided for each case within parentheses. Number of replications and sample size are $R_{cov} = 200$ and $n_{cov} = 500$, respectively.

Method	MG	MST	MVC	MVG	SN	T
NN-DM	0.92 (0.04)	0.88 (0.03)	0.69 (0.03)	0.80 (0.31)	0.92 (0.06)	0.88 (0.03)
DP-MC	0.53 (0.01)	0.56 (0.01)	0.47 (0.01)	0.41 (0.16)	0.39 (0.02)	0.55 (0.01)
DP-VB	0.56 (0.03)	0.58 (0.03)	0.18 (0.02)	0.55 (0.26)	0.49 (0.05)	0.57 (0.02)

method breaks down in this high-dimensional setup. Due to numerical instability in estimating the \mathcal{L}_1 error in higher dimensions, we evaluate the methods in terms of their out-of-sample log-likelihood (OOSLL) instead (Gneiting and Raftery, 2007), on a test set of 500 data points. We report the average OOSLL over 30 replications in Table 2.3. The results indicate that both methods perform very similarly in terms of out-of-sample fit to the data, with the NN-DM outperforming the DP-VB when the true density is MVC. We also observed that for this experiment, the NN-DM

Table 2.3: Out-of-sample log-likelihood ($\times 10^4$) of NN-DM and DP-VB on a test set of 500 points for 6 different multivariate densities considered in Section 2.4.3, for $n = 1000$ and $p = 50$. Greater out-of-sample log-likelihood is better.

Method	MG	SN	T	MST	MVC	MVG
NN-DM	-1.75	-1.74	-1.84	-1.84	-1.31	-1.36
DP-VB	-1.75	-1.74	-1.86	-1.84	-1.34	-1.36

methods with default choice of hyperparameters and with cross-validated choice of Ψ_0 have almost identical performance. For the NN-DM, we set $k = 12$ after carrying out a sensitivity analysis on k by considering $k = 5, 7, 10, 15$, and 20 . The best results for the NN-DM were obtained for $k \in \{7, 10, 12\}$ with negligible difference in out-of-sample log-likelihoods between these three choices, with $k = 12$ performing the best.

2.4.6 Runtime Comparison

With n data points in p dimensions, the initial nearest neighbor allocation into n neighborhoods can be carried out in $\mathcal{O}(n \log n)$ steps (Vaidya, 1986; Ma and Li, 2019). Once the neighborhoods are determined with k_n points in each neighborhood, obtaining the neighborhood specific empirical means and covariance matrices has $\mathcal{O}(nk_n p + nk_n p^2) = \mathcal{O}(nk_n p^2)$ complexity. Obtaining the pseudo-posterior mean (2.8) then requires inversion of n such $p \times p$ matrices to evaluate the multivariate t-density, with a runtime of $\mathcal{O}(np^3)$. Therefore, the total runtime to obtain the pseudo-posterior mean is of the order $\mathcal{O}(nk_n p^2 + np^3)$. When we are interested in uncertainty quantification, we require Monte Carlo samples of the NN-DM, which are independently drawn from its pseudo-posterior. This involves sampling the Dirichlet weights, the neighborhood specific unknown mean and covariance matrix parameters of the Gaussian kernel, and evaluating a Gaussian density for each neighborhood, as outlined in Algorithm 1. To obtain M Monte Carlo samples, the combined complex-

ity of this step is thus $\mathcal{O}(Mn + Mnp^3) = \mathcal{O}(Mnp^3)$. Overall the runtime complexity to obtain NN-DM samples is therefore $\mathcal{O}(Mnp^3 + nk_n p^2 + np^3)$. For high dimensional scenarios, this runtime can be greatly improved by using a low rank matrix factorization of both the neighborhood specific empirical covariance matrices and the sampled covariance matrix parameters to make matrix inversion more efficient (Golub and van Loan, 1996). We now provide a detailed simulation study of runtimes of the proposed method, with all the simulations carried out on an M1 MacBook Pro with 16 GB of RAM.

We first focus on some runtime experiments comparing NN-DM and DP-MC. In our experiments, we focus on $p = 1$ and $p = 4$. The runtime for NN-DM consists of the time to estimate δ_0^2 by cross-validation as in Section 2.2.3 and then drawing samples from its pseudo-posterior. For both dimensions, the sample size is varied from $n = 200$ to $n = 1500$ in increments of 100. Data are generated from the standard Gaussian density (GS) for $p = 1$ and from a mixture of skew t-distributions with the parameters as described for the case MST in Section 2.4.3 for $p = 4$. For $p = 1$, we evaluate the two methods at 500 test points, while for $p = 4$ we evaluate the methods at 200 test points. The hyperparameters are kept the same as in Sections 2.4.2 and 2.4.3. We took 1000 Monte Carlo samples for the NN-DM and 2500 MCMC samples for the DP-MC with a burn-in of 1500 samples. We provide a figure summarizing the results in Figure 2.5. In the top panel of Figure 2.5, we plot the average of the logarithm (base 10) of the run times of each approach for 10 independent replications. Corresponding \mathcal{L}_1 errors of the two methods is included in the bottom panel of Figure 2.5.

In Figure 2.5, the NN-DM is at least an order of magnitude faster than DP-MC. The time saved becomes more pronounced in the multivariate case, where for sample size 1500 the NN-DM is ~ 50 times faster. The gain in computing time does not come at the cost of accuracy as can be seen from the right panel; the proposed method

maintains the same order of \mathcal{L}_1 error as the DP-MC in the univariate case and often outperforms the DP-MC in the multivariate case. We did not implement the Monte Carlo sampler for the proposed algorithm in parallel, but such a modification would substantially improve runtime. Bypassing cross-validation and choosing default hyperparameters instead as outlined in Section 2.2.3, NN-DM took 3.3 seconds and 16.4 seconds when $p = 1$ and $p = 4$, respectively, with sample size $n = 1500$. In the same scenario, DP-MC took 99.4 seconds and 1618.1 seconds for $p = 1$ and $p = 4$, respectively. Thus the NN-DM with default hyperparameters is about 30 times faster when $p = 1$ and almost 100 times faster when $p = 4$.

We also compare the runtime of the proposed method with three recent implementations of the DPM, namely the packages `bnpy` (Hughes and Sudderth, 2014), `DPMMSubClusters` (Dinari et al., 2019), and `vdpgm` (Kurihara et al., 2006) available for download at <https://kenichikurihara.com/variational-dirichlet-process-gaussian-mixture-model/>. These three packages implement variational approximations of the DPM posterior with different modifications. We also include the DP-MC and OPT for comparison. All the runtime results are comparable only up to machine and coding language differences. Amongst the competitor package implementations, the `NNDM` and `dirichletprocess` packages are the only ones providing (pseudo) posterior samples of the density estimate at a test point. We consider the average runtime of $R = 10$ replicates to fit a training data set of iid $N(0, 1)$ entries with $n = 1500$ and $p = 4$. For the NN-DM, DP-MC, and OPT, we consider 1000 (pseudo) posterior samples. Table 2.4 provides the runtimes for the different packages considered. Overall, the fastest implementation is observed for the PTT package. The next fastest implementations are the `NNDM` without cross-validation (CV), `DPMMSubClusters`, and `bnpy`. The runtime for `NNDM` with CV closely follows the previous implementations, with both `NNDM` with and without CV providing (pseudo) posterior samples. The major improvement in runtime for NN-DM is mainly due to the fact that neighborhood allocations are fixed here which is not the

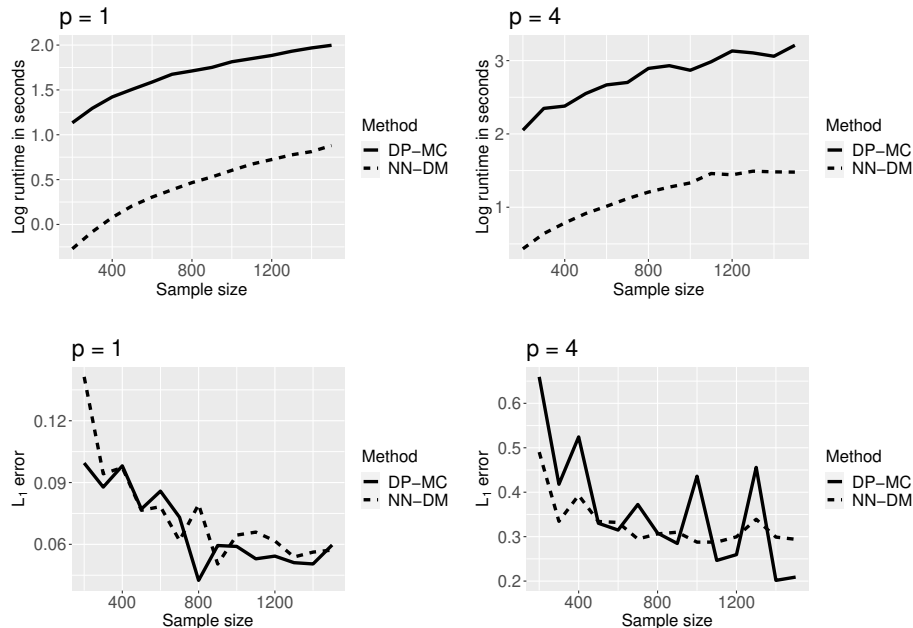


FIGURE 2.5: Runtime comparison of DP-MC and NN-DM in univariate case and for 4-dimensional data. Top panel shows runtimes in \log_{10} scale whereas bottom panel shows corresponding \mathcal{L}_1 error. Sample size n is varied from 200 to 1500 in increments of 100.

case for DP-MC.

Table 2.4: Table comparing the average runtimes of different packages for $n = 1500$, $p = 4$. NN-DM runtimes are provided both with and without cross-validation (CV), using the package NNDM developed by the authors.

Package (Language)	Average Runtime (s)	Samples?
bnpy (Python)	5.79	No
DPMMSubClusters (Julia)	4.33	No
vdpgm (MATLAB)	58.38	No
NNDM (Rcpp and R, with CV)	18.96	Yes
NNDM (Rcpp and R, without CV)	3.52	Yes
dirichletprocess (R)	1068.48	Yes
PTT (Rcpp and R)	0.59	No

2.4.7 Sensitivity to the Choice of k

In this subsection, we investigate the role of $k_n = k$ in finite samples for the proposed method. We consider $n = 200$ samples from the SP density in the univariate case and

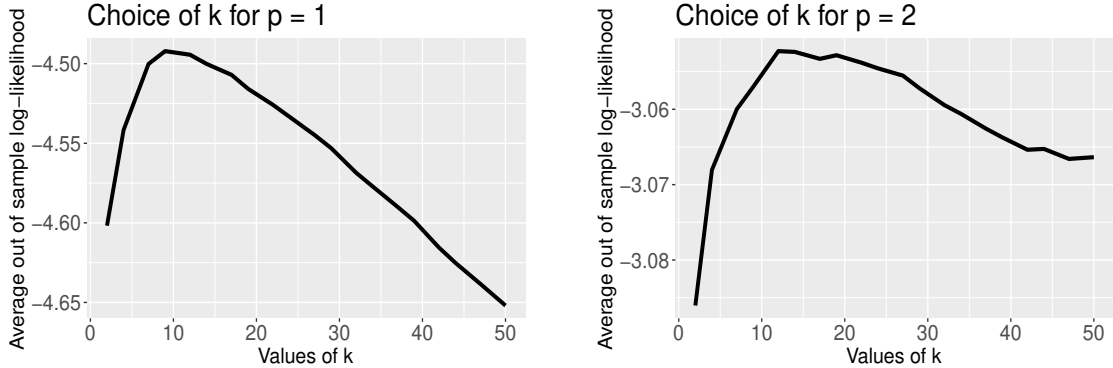


FIGURE 2.6: Average out-of-sample log-likelihood of 500 test points for the NN-DM as a function of k for one-dimensional and two-dimensional data. Number of samples and number of replications are $n = 200$ and $R = 10$, respectively.

the MG density in the bivariate case. In each case, we fix a test set of $n_t = 500$ points, and evaluate the out-of-sample log-likelihood (OOSLL) of the test points for 20 different integer values of k ranging from 2 to 50. Finally, we report results averaged from 10 independent replicates of this setup. We note that for each considered value of k , the parameter δ_0^2 was estimated using leave-one-out cross-validation. Figure 2.6 shows how the OOSLL averaged over replicates changes as a function of k for each density considered. The original OOSLL values of the test data points were scaled by the number of test points $n_t = 500$ for better representability.

For the univariate SP density, the optimal value of k which maximizes the average OOSLL is $\hat{k} = 9$. This is close to the choice of $k = 6$ as taken in Section 2.4.2. For the bivariate MG density, we observe that the choice of k maximizing the OOSLL is $\hat{k} = 12$, which is also close to the choice of $k = 10$ as taken in Section 2.4.3. For both the univariate and the bivariate case, the out-of-sample log-likelihood of the test set shows little variation with changing k . This indicates that the estimates obtained from the proposed method are quite robust to the particular choice of k .

2.5 Application

We apply the proposed density estimator to binary classification. Consider data $\mathcal{D} = \{(X_i, Y_i) : i = 1, \dots, n\}$, where $X_i \in \mathbb{R}^p$ are p -dimensional feature vectors and

$Y_i \in \{0, 1\}$ are binary class labels. To predict the probability that $y_0 = 1$ for a test point x_0 , we use Bayes rule:

$$\text{pr}(y_0 = 1 \mid x_0) = \frac{\tilde{f}_1(x_0) \text{pr}(y_0 = 1)}{\tilde{f}_0(x_0) \text{pr}(y_0 = 0) + \tilde{f}_1(x_0) \text{pr}(y_0 = 1)}, \quad (2.16)$$

where $\tilde{f}_j(x_0)$ is the feature density at x_0 in class j and $\text{pr}(y_0 = j)$ is the marginal probability of class j , for $j = 0, 1$. Based on n_t test data, we let $\hat{\text{pr}}(y_0 = 1) = (1/n_t) \sum_{i=1}^{n_t} Y_i$, with $\hat{\text{pr}}(y_0 = 0) = 1 - \hat{\text{pr}}(y_0 = 1)$. We use either the NN-DM pseudo-posterior mean $\hat{f}_n(\cdot)$, the DP-MC posterior mean $\hat{f}_{\text{DP}}(\cdot)$, or the DP-VB posterior mean $\hat{f}_{\text{VB}}(\cdot)$ for estimating the within class densities, and compare their classification performances in terms of sensitivity, specificity, and probabilistic calibration. We omit the KDE as to the best of our knowledge, no routine R implementation is available for data having more than 6 dimensions.

The high time resolution universe survey data (Keith et al., 2010) contain information on sampled pulsar stars. Pulsar stars are a type of neutron stars and their radio emissions are detectable from the Earth. These stars have gained considerable interest from the scientific community due to their several applications (Lorimer and Kramer, 2012). The data are publicly available from the University of California at Irvine machine learning repository. Stars are classified into pulsar and non-pulsar groups according to 8 attributes (Lyon, 2016). There are a total of 17898 instances of stars, among which 1639 are classified as pulsar stars.

We create a test data set of 200 stars, among which 23 are pulsar stars. The training size is then varied from 300 to 1800 in increments of 300, each time adding 300 training points by randomly sampling from the entire data leaving out the initial test set. In Figure 2.7, we plot the sensitivity and specificity of the three methods in consideration. All the methods exhibit similar sensitivity across various training sizes; the DP-MC has marginally better specificity for training sizes 1200 and 1500, while the NN-DM has better specificity for training sizes 300 and 600. Both the NN-DM and the DP-MC exhibit higher specificity and sensitivity than the DP-VB across all training sample sizes considered.

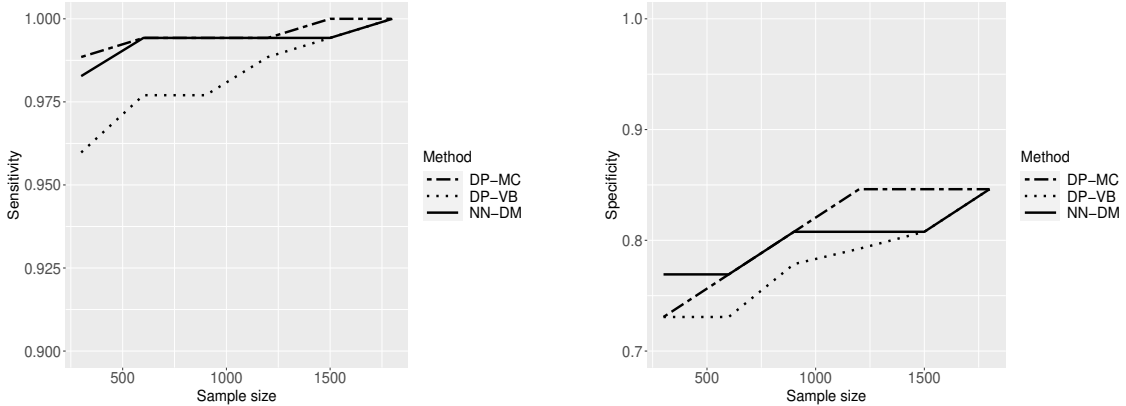


FIGURE 2.7: Sensitivity and specificity of the NN-DM, DP-MC, and DP-VB for the high time resolution universe survey data.

We also compare the methods using the Brier score, a proper scoring rule (Gneiting and Raftery, 2007) for probabilistic classification. Suppose for n_t test points and the i th Monte Carlo sample, p_i denotes the sampled $n_t \times 1$ probability vector for a generic method. We compute the normalized Brier score for the i th sample as $(1/n_t) \|p_i - Y_t\|_2^2$, where Y_t is the vector of class labels in the test set. Then with T samples of p_i , $i = 1, \dots, T$, we compute the mean Brier score for the three methods considered. The mean Brier score for each training size is shown in the right panel of Figure 2.8, which naturally shows a declining trend with increasing training size. There is little to choose between the three classifiers in terms of mean Brier score; the proposed method fairs equally well in terms of calibration of estimated test set probabilities with the MCMC implementation of the Dirichlet process. In the left panel of Figure 2.8, the receiver operating characteristic curve of the methods is shown for 1800 training samples. The area under the curve (AUC) for the NN-DM, the DP-MC and the DP-VB are 0.96, 0.95 and 0.96, respectively. For 1800 training samples, the computation time for the proposed method is about 13 minutes while for the DP-MC it is approximately 5 hours.

Hence, the proposed method is much faster, even without exploiting parallel computation. We also fitted the proposed method using the training set of all 17698 points; DP-MC was too slow in this case. The sensitivity and specificity of the

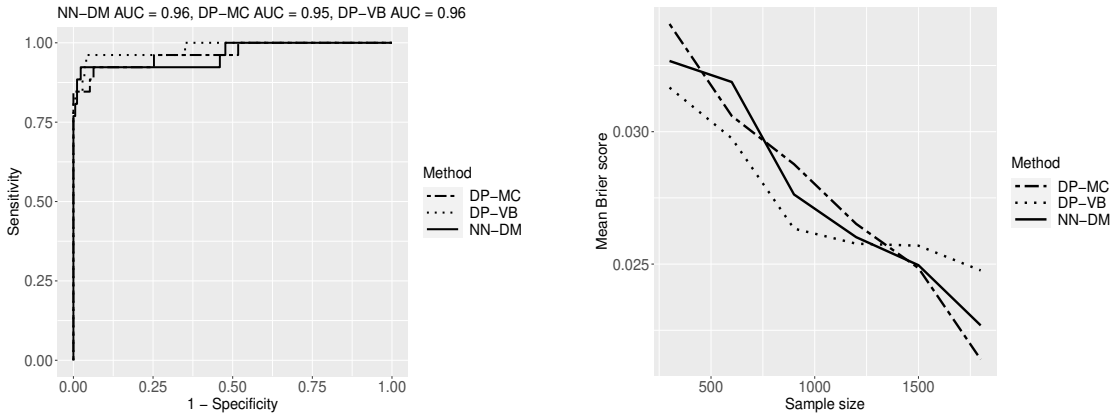


FIGURE 2.8: Left plot shows the receiver operating characteristic curve of the NN-DM, DP-MC, and DP-VB with 1800 training samples. Area under the curve is abbreviated as AUC. Right plot shows normalized Brier scores for the methods with varying training sample size.

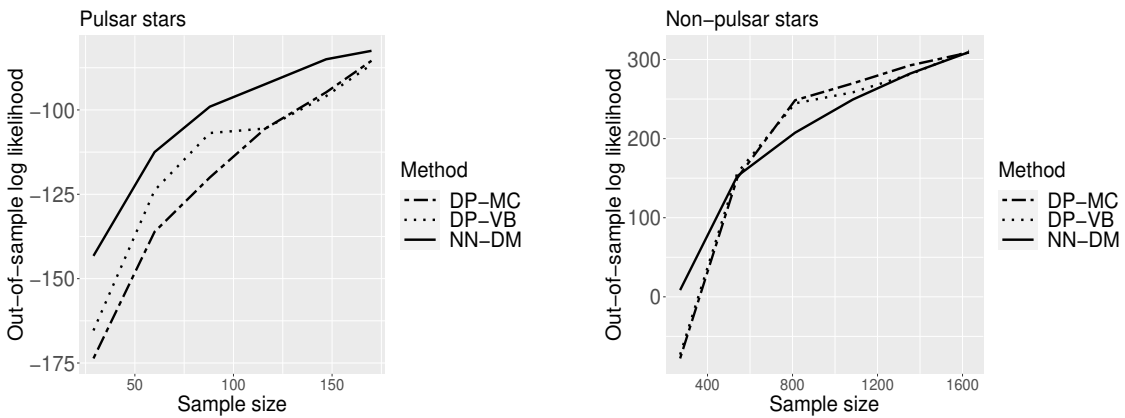


FIGURE 2.9: Left and right plots show the out-of-sample log-likelihoods of NN-DM, DP-MC, and DP-VB for the two different star types.

proposed method increased to 0.99 and 0.91, respectively. We additionally evaluated the methods in terms of the out-of-sample log-likelihood. The results are displayed in Figure 2.9. While the methods perform comparably in terms of their classification performance, NN-DM achieves a better fit overall, especially for the significantly less prevalent pulsar star type.

Synergistic Antagonistic Interaction Detection

3.1 Introduction

There is considerable concern that humans are exposed to a variety of potentially adverse chemicals, and these exposures may have adverse health effects. Classically, such health effects have been assessed through one exposure at a time studies, either collecting *in vitro* or *in vivo* data at different doses of a single chemical or focusing analyses of observational epidemiology studies on single exposures. Then, in order to predict the overall health effect of a mixture of different exposures, one needs to make strong assumptions, such as additivity. Unfortunately, such predictions will misestimate an individual's true adverse health risk if certain chemicals interact. Of particular concern are *synergistic interactions* in which the adverse effect of one chemical is increased due to the presence of another chemical. If regulatory agencies are unaware of such synergistic interactions in setting pollution guidelines, they may inadvertently permit substantial pollution-induced mortality and morbidity. On the other hand, *antagonistic interactions* may lead to additive models over-predicting risk in which case certain chemicals may be over-regulated.

In this chapter, our goal is to analyze the effect of heavy metal exposure on human kidney function. The impact of metal exposures on human renal function has garnered considerable attention from the epidemiological community. A review of the existing literature shows evidence of degraded kidney function following prolonged exposure to heavy metals (Pollack et al., 2015; Luo and Hendryx, 2020). Most of this literature is based on simple one exposure-at-a-time correlation analyses from observational epidemiology data. It is clearly of interest to study joint effects of multiple metals, while adjusting for covariates that may act as potential confounding variables. Mechanistically, it seems unlikely that metals have a simple additive effect on kidney function. For example, healthy kidney function may continue with a single metal exposure at relatively low doses, but as dose increases and/or additional metal exposures are added it is likely that kidney function may worsen rapidly. Such a dose response surface would be reflective of a synergistic interaction. In our analyses, we focus on 2015-16 data from NHANES. These data contain information on heavy metals found in spot urine collections and also on urine creatinine, which can be used as a marker of kidney function when urinary dilution issues are appropriately accounted for.

There is a clear need for new statistical tools for accommodating interactions in assessing the health effects of mixtures of chemical exposures. For a review of recent developments, refer to Joubert et al. (2022). A broad set of strategies have been taken in this literature. The first extends linear regression to include a quadratic term characterizing pairwise interactions; for recent examples, refer to Ferrari and Dunson (2021); Wang et al. (2019). This approach can trivially identify synergistic versus antagonistic interactions through the signs of the quadratic coefficients, but relies on a highly restrictive parametric model. An alternative is to rely on nonparametric regression; for example, using Bayesian Additive Regression Trees (BART) (Chipman et al., 2010) or Gaussian processes (GP) (Williams and Rasmussen, 2006). Bayesian

Kernel Machine Regression (BKMR) (Bobb et al., 2015) implements GP regression with variable selection motivated by the mixtures problem. These approaches do not identify synergistic or antagonistic interactions, and can lead to overly-wiggly and difficult to interpret dose response surfaces. MixSelect (Ferrari and Dunson, 2020) bridges between parametric and nonparametric approaches via an additive expansion. However, it lacks flexibility to characterize nonlinear main effects and interactions due to its reliance on quadratic regression. To simplify dose response modeling, Molitor et al. (2010) propose profile regression based on clustering exposures, while Czarnota et al. (2015) use a single index model based on a weighted sum of the exposures. Neither approach allows for inferences on interactions.

We propose a nonparametric Bayesian approach for identifying synergistic or antagonistic interactions between p exposures. The dose response surface is decomposed additively into p main effects and $\binom{p}{2}$ pairwise interactions. Ruling out higher order interactions substantially reduces dimensionality, while aiding interpretability. Each pairwise interaction is decomposed into the difference of two non-negative functions, facilitating selection of synergistic or antagonistic effects. The proposed *synergistic antagonistic interaction detection* (SAID) approach is shown to improve over unconstrained nonparametric regression, leading to comparatively higher statistical power in simulation studies. To avoid modeling complicated pairwise surfaces, we focus on a special case to dramatically reduce dimensionality and obtain substantial computational gains. The resulting interaction surface retains a high degree of flexibility in capturing nonlinear surfaces. We also outline an approach to carry out variable selection on the pairwise interactions. The SAID approach employs a computationally efficient Markov chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution of the unknown quantities.

The chapter is outlined as follows. We describe the urine creatinine and metal exposure data in Section 3.2 and outline challenges involved with analyzing these data.

Section 3.3 provides details of our SAID approach. Section 4.4 compares SAID and existing competitors in simulation experiments, highlighting benefits of using SAID in terms of estimation accuracy, valid uncertainty quantification, and variable selection. Section 3.5 uses SAID to investigate synergistic and antagonistic interactions of metal exposures in predicting urine creatinine levels based on NHANES 2015-16 data.

3.2 Kidney Function Data Analysis

3.2.1 Motivation

We are interested in assessing the impact of exposure to heavy metals on human renal function. The concentration of creatinine in the blood or urine is an established biomarker of kidney function (Kashani et al., 2020; Barr et al., 2005). Exposure to heavy metals has been linked to changes in renal function and kidney damage (Pollack et al., 2015). In addition to being indicative of renal function, creatinine levels are also related to muscle mass (Forbes and Bruining, 1976; Baxmann et al., 2008). Urine creatinine has been shown to be positively associated with serum creatinine levels in studies excluding individuals with chronic kidney disease (CKD) (Jain, 2016). Existing studies, such as the one in Luo and Hendryx (2020), show a statistical association between biomarkers of chronic kidney disease and blood levels of the heavy metals cobalt (Co), chromium (Cr), mercury (Hg), and lead (Pb) using NHANES data. Kim et al. (2015) find an association between CKD and elevated levels of cadmium (Cd) in blood. These studies primarily focus on marginal associations between individual chemicals and CKD. Our focus is instead on studying how multiple heavy metal exposures relate to kidney function measured through urine creatinine, with an emphasis on identifying synergistic and antagonistic interactions.

3.2.2 Data Description

We analyze data that were collected by NHANES for the year 2015. We consider urine analyte levels of the 13 heavy metals: Antimony (Sb), Barium (Ba), Cadmium (Cd), Cesium (Cs), Cobalt (Co), Lead (Pb), Manganese (Mn), Molybdenum (Mo), Strontium (Sr), Thallium (Tl), Tin (Sn), Tungsten (W), and Uranium (U) as exposure variables and the level of urine creatinine (uCr) as the response variable. The unit of measurement for exposure variables is $\mu\text{g}/\text{mL}$ while the unit for the response variable is mg/dL . We also adjust for age (in years), sex (male or female), ethnicity (Non-Hispanic White, Non-Hispanic Black, Mexican American, Other Hispanic, and Other), and body mass index (BMI).

We start off by only considering the data for 2300 individuals for which values of the response variable (uCr) are not missing. In this data set, we first remove the individuals having serious kidney disease. For this, we used the albumin-to-creatinine ratio (ACR) and removed the subjects who had albuminuria, defined as their ACR satisfying $\text{ACR} \geq 30 \text{ mg/g}$. The resulting data set has 2008 individuals. Furthermore, the response variable for one individual was less than the limit of detection (LOD); this singular data point was removed for purposes of analysis. For the metal exposures, there are both missing data points and data points which are lower than their respective LODs. We remove any individuals with an exposure entry missing. Finally, we removed the entries of the covariate values which were missing. In summary, neither the response variable nor the covariate variables considered in our analysis have any missing data or data less than the LOD; however, the metal exposure data contain points less than the LOD. After removing the individuals missing response, covariate, and exposure measurements, the final sample size is $n = 1979$.

We look at various summary measures of the response and exposure variables.

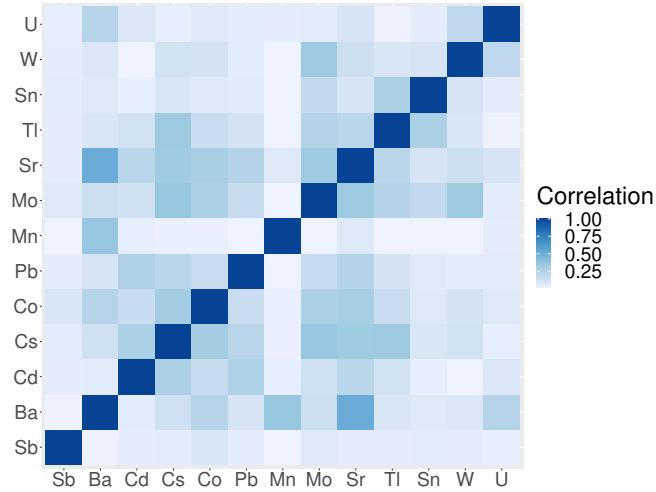


FIGURE 3.1: Heat plot showing correlations between heavy metals in NHANES 2015-16 data.

The uCr levels are right skewed, with mean 119.67 mg/dL, median 104.00 mg/dL, and semi-interquartile range (SIQR) 52.50 mg/dL. The heavy metal exposure variables are mild to moderately correlated, with a heat plot of the correlation matrix shown in Figure 3.1. More specifically, 58 out of the 78 pairwise correlation coefficients are less than 0.20. The median correlation is 0.10 with an SIQR of 0.08, while the maximum correlation is 0.51, between Barium and Strontium. Furthermore, there is wide variation between the range of different exposure variables. For example, the maximum recorded exposure to Cesium in the sample is 110.87 $\mu\text{g}/\text{mL}$, while that of Uranium is only 0.77 $\mu\text{g}/\text{mL}$.

3.2.3 Urinary Dilution

In NHANES, the urine data is collected from spot collections. The urine sample collection procedure is regulated to ensure consistent specimen collection across differing water loadings of the subjects. Regardless, due to the nature of data collection, the concentrations of the heavy metals and creatinine levels can vary substantially across the sampled participants due to different hydration levels. For this reason, we

adjust the raw response (creatinine) and exposure variables (heavy metals) measured in the urine using the urine flow rate (UFR, measured in mL/min) of the subjects. Adjusting for dilution using urine flow rate also helps reduce wide variation in exposure measurements. Adjusting urinary measures for variation in water loading across individuals using urine flow rates is prevalent in existing literature (Jeng et al., 2021; Middleton et al., 2016; Hays et al., 2015). Suppose for the i -th subject, the urine creatinine level is C_i and the urinary concentrations of the heavy metals is the vector $\mathbf{M}_i = (M_{i1}, \dots, M_{ip})^\top$. Provided the urine flow rate for the i -th subject is $\tau_i > 0$, we adjust for urinary dilution by multiplying both the original response and exposure variables by τ_i . That is, the urine flow adjusted response variable and exposure vector are $\tau_i C_i$ and $\tau_i \mathbf{M}_i$, respectively, which we refer to as the dilution-adjusted response and exposure variables, respectively.

3.2.4 *Issues with Existing Approaches*

As two different types of state-of-the-art methods, we apply MixSelect and BKMR on the heavy metals and creatinine data. We (natural) log transform the response and exposures prior to analysis. Both BKMR and MixSelect conduct Bayesian variable selection, providing posterior inclusion probabilities (PIPs) for each exposure. Excluding exposures having PIPs < 0.5 , BKMR excludes Barium (PIP ≈ 0), Lead (PIP ≈ 0), and Tin (PIP = 0.13), while MixSelect excludes Cobalt (PIP = 0.29) and Uranium (PIP = 0.11). With these exposures excluded, BKMR produces a 10 dimensional dose response surface in the remaining metals. Although inferences on main effects and pairwise interactions could rely on examining univariate and bivariate cross-sections of the 10-dimensional surface, such results are exploratory and difficult to interpret. In contrast, MixSelect provides PIPs for both the main effects and pairwise interactions. However, it is not reassuring that MixSelect excludes different exposures than BKMR. In addition, all the interaction PIPs for MixSelect are

< 0.11 , indicating that none of the interactions are selected.

We are motivated by these preliminary results to develop an approach that is flexible enough to characterize nonlinear dose response surfaces, while allowing us to formally detect pairwise interactions that are synergistic or antagonistic. Flexible nonparametric dose response surface methods can characterize synergistic or antagonistic pairwise interactions but they tend to be hidden within a flexible multivariate surface, as highlighted for BKMR above. We describe our proposed Synergistic Antagonistic Interaction Detection (SAID) approach in Section 3.3. SAID will be applied to the motivating metals and creatinine data in Section 3.5.

3.3 Structured Interaction Modeling Approach

3.3.1 Basic Modeling Structure

Denote the health outcome of interest by y_i , for individuals $i = 1, \dots, n$; we assume $y_i \in \mathbb{R}$ but the approach can be easily modified to allow $y_i \in \{0, 1\}$ or ordered categorical or count responses. We denote the exposures as $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top \in [0, 1]^p$, assuming without loss of generality they each fall within $[0, 1]$, and the covariates as $\mathbf{z}_i \in \mathbb{R}^q$. The model we consider is given by

$$y_i = H(\mathbf{x}_i) + \eta^\top \mathbf{z}_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad (3.1)$$

where $H(\mathbf{x}) = H(x_1, \dots, x_p)$ is the dose response function of the exposures, and we follow common practice in adjusting for the covariates linearly. Without loss of generality, we assume that higher values of the health outcome y represent worse health in interpreting exposure effects. As in Wei et al. (2020); Brezger and Lang (2006), we characterize the dose response function of the exposures $H(\mathbf{x})$ via an additive expansion into main effects and pairwise interaction terms:

$$H(x_1, \dots, x_p) = \alpha + \sum_{j=1}^p f_j(x_j) + \sum_{1 \leq u < v \leq p} h_{uv}(x_u, x_v), \quad (3.2)$$

where α is an intercept, $f_j(x_j)$ is the main effect of the j th exposure for $j = 1, \dots, p$, and $h_{uv}(x_u, x_v)$ for $1 \leq u < v \leq p$ is a pairwise interaction. The decomposition in (3.2) allows interpretation of different components in a factorization of the dose response surface H into main effects and pairwise interactions.

The primary innovation in this chapter is our approach for inferences on the pairwise interaction component. Section 3.3.2 introduces the general structure of our model for the h_{uv} s. Section 3.3.3 describes an approach for interaction selection. Section 3.3.4 describes our model for the main effects. Section 3.3.5 contains details on posterior computation.

3.3.2 Modeling Pairwise Interactions

We propose a model for the pairwise interactions h_{uv} s. Assume that h_{uv} is continuous and admits finite partial derivatives up to second order. For $\mathbf{a} = (a_1, \dots, a_d)^\top \in \mathbb{R}^d$, we let $\mathbf{a}^2 = (a_1^2, \dots, a_d^2)^\top \in \mathbb{R}^d$. For a non-negative function $f : [0, 1]^2 \rightarrow [0, \infty)$, we say $f \equiv 0$ if $f(x_1, x_2) = 0$ for all $(x_1, x_2) \in [0, 1]^2$; otherwise, if $f(x_1, x_2) > 0$ for at least one $(x_1, x_2) \in [0, 1]^2$, $f \not\equiv 0$.

In conducting inferences on interactions in environmental epidemiology, it is of primary interest to assess whether exposures work together to magnify their health effects (synergy), tend to block each other's effects (antagonism), or have effectively no interaction (null). Hence, for exposures u and v , we focus on the following classes of interactions.

1. *Synergistic* if $h_{uv}(x_u, x_v) \geq 0$ for all $(x_u, x_v) \in [0, 1]^2$ with at least one strict inequality.
2. *Antagonistic* if $h_{uv}(x_u, x_v) \leq 0$ for all $(x_u, x_v) \in [0, 1]^2$ with at least one strict inequality.
3. *Null* if $h_{uv}(x_u, x_v) = 0$ for all $(x_u, x_v) \in [0, 1]^2$.

Our primary interest is in classifying interactions h_{uv} in terms of Definitions 1-3. Although we will not rule out other types of interaction surfaces *a priori*, we develop inference approaches that are targeted towards this three class hypothesis testing problem. This is accomplished in a Bayesian manner with a carefully-structured model for h_{uv} that shrinks towards the space of synergistic, antagonistic, and null interactions.

If we knew *a priori* that h_{uv} was either synergistic or null, we could let $h_{uv} = P_{uv}$, where $P_{uv} \geq 0$. If $P_{uv}(x_u, x_v) > 0$ for at least one $(x_u, x_v) \in [0, 1]^2$ then h_{uv} is synergistic; otherwise, $P_{uv} \equiv 0$ and h_{uv} is a null interaction. Similarly, if we knew *a priori* that h_{uv} was either antagonistic or null, we could let $h_{uv} = -N_{uv}$, where $N_{uv} \geq 0$. However, the sign of the interaction h_{uv} is usually unknown and pairwise interactions may be only approximately synergistic or antagonistic. To allow for such complexities, we let

$$h_{uv}(x_u, x_v) = P_{uv}(x_u, x_v) - N_{uv}(x_u, x_v), \quad (3.3)$$

for all $(x_u, x_v) \in [0, 1]^2$, where $P_{uv}, N_{uv} : [0, 1]^2 \rightarrow [0, \infty)$ are non-negative functions. If $P_{uv} \not\equiv 0$ and $N_{uv} \equiv 0$, h_{uv} is synergistic; while if $P_{uv} \equiv 0$ and $N_{uv} \not\equiv 0$, h_{uv} is antagonistic. If both $P_{uv} = N_{uv} \equiv 0$, h_{uv} is null. All three cases are contained in the restriction $\int P_{uv} \int N_{uv} = 0$, where $\int f$ is shorthand for $\int_{[0,1]^2} f(x_1, x_2) dx_1 dx_2$ for a bivariate function f . By penalizing $\int P_{uv} \int N_{uv}$, we can favor h_{uv} close to synergistic, antagonistic, or null.

We decompose P_{uv} and N_{uv} as products of non-negative univariate functions:

$$\begin{aligned} h_{uv}(x_u, x_v) &= P_{uv}(x_u, x_v) - N_{uv}(x_u, x_v) \\ &= P_{uv,1}(x_u)P_{uv,2}(x_v) - N_{uv,1}(x_u)N_{uv,2}(x_v). \end{aligned} \quad (3.4)$$

This model is much more flexible than commonly used quadratic regression, which lets $h_{uv}(x_u, x_v) = \gamma_{uv}x_u x_v$. To model $P_{uv,1}, P_{uv,2}, N_{uv,1}, N_{uv,2}$ as flexible one-dimensional

non-negative functions, we rely on squaring B-spline expansions (De Boor, 1978) as follows:

$$\begin{aligned} P_{uv,1}(x_u) &= \{\mathbf{s}_u(x_u)^\top \theta_{uv,1}\}^2, & P_{uv,2}(x_v) &= \{\mathbf{s}_v(x_v)^\top \phi_{uv,1}\}^2 \\ N_{uv,1}(x_u) &= \{\mathbf{s}_u(x_u)^\top \theta_{uv,2}\}^2, & N_{uv,2}(x_v) &= \{\mathbf{s}_v(x_v)^\top \phi_{uv,2}\}^2, \end{aligned} \quad (3.5)$$

where $\mathbf{s}_u(x_u) = (s_{u1}(x_u), \dots, s_{um}(x_u))^\top$ denote B-splines for the u -th variable, chosen so that $s_{uj}(0) = 0$ for $j = 1, \dots, m$. This is obtained by discarding the intercept spline; refer to Section B.1 of the Supplementary Material for further details. Constraining $s_{uj}(0) = 0$ for $j = 1, \dots, m$ and $u = 1, \dots, p$ ensures that the interaction h_{uv} satisfies $h_{uv}(x_u, 0) = 0$ for all $x_u \in [0, 1]$ and $h_{uv}(0, x_v) = 0$ for all $x_v \in [0, 1]$, thereby making h_{uv} identifiable. For further details on identifiability of pairwise interactions, we refer the reader to Sections B.1 and B.3 of the Supplementary Material.

To estimate h_{uv} in a Bayesian manner, we define a prior distribution $\pi(\Psi_{uv})$ on the basis coefficients $\Psi_{uv} = (\theta_{uv,1}^\top, \phi_{uv,1}^\top, \theta_{uv,2}^\top, \phi_{uv,2}^\top)^\top$. In order to formulate $\pi(\Psi_{uv})$, we first define a prior distribution $\pi_0(\Psi_{uv})$ and then augment it with a penalty term based on $\mathcal{Q}(P_{uv}, N_{uv}) = \int P_{uv} \int N_{uv}$. The initial prior $\pi_0(\Psi_{uv})$ is defined as

$$\theta_{uv,1}, \phi_{uv,1} \sim N(0, \nu^2 \tau_{uv,1}^2 \Sigma_0), \quad \theta_{uv,2}, \phi_{uv,2} \sim N(0, \nu^2 \tau_{uv,2}^2 \Sigma_0), \quad (3.6)$$

where Σ_0 is a P-spline covariance matrix as in Lang and Brezger (2004); for more details, refer to Section B.2 of the Supplementary Material. Let $w_{uv} = (\tau_{uv,1}, \tau_{uv,2})$ denote the uv -th interaction specific variance parameters. Then, the augmented prior is

$$\pi(\Psi_{uv} \mid \kappa_{uv}, w_{uv}, \nu) \propto \pi_0(\Psi_{uv} \mid w_{uv}, \nu) \exp\{-\kappa_{uv} \mathcal{Q}(P_{uv}, N_{uv})\}.$$

By letting $A_u = \int \mathbf{s}_u(x_u) \mathbf{s}_u(x_u)^\top dx_u$ for $u = 1, \dots, p$, $\mathcal{Q}(P_{uv}, N_{uv})$ can be expressed as

$$\mathcal{Q}(P_{uv}, N_{uv}) = \tilde{\mathcal{Q}}(\Psi_{uv}) = (\theta_{uv,1}^\top A_u \theta_{uv,1}) (\phi_{uv,1}^\top A_v \phi_{uv,1}) (\theta_{uv,2}^\top A_u \theta_{uv,2}) (\phi_{uv,2}^\top A_v \phi_{uv,2}).$$

We can therefore rewrite $\pi(\cdot \mid \kappa_{uv}, w_{uv}, \nu)$ as

$$\pi(\Psi_{uv} \mid \kappa_{uv}, w_{uv}, \nu) \propto \pi_0(\Psi_{uv} \mid w_{uv}, \nu) \exp\{-\kappa_{uv} \tilde{\mathcal{Q}}(\Psi_{uv})\}. \quad (3.7)$$

The prior is completed with hyperpriors for the variance and penalty parameters:

$$\tau_{uv,1}, \tau_{uv,2} \sim C^+(0, 1), \quad \log(\kappa_{uv}) \sim N(0, 1), \quad \nu \sim C^+(0, 1), \quad (3.8)$$

where $C^+(0, 1)$ denotes a half-Cauchy prior.

Prior (3.7)-(3.8) is chosen to have a global-local shrinkage form motivated by the horseshoe prior (Carvalho et al., 2009). Small values of the global parameter ν favor most of the interactions $h_{uv} \approx 0$, while the heavy-tailed prior on local parameters $\tau_{uv,1}$ and $\tau_{uv,2}$ allow certain interactions to have P_{uv} and N_{uv} components arbitrarily far from zero. The role of the penalty is to favor h_{uv} s that are close to synergistic, antagonistic, or null. However, the prior on κ_{uv} allows the penalty to be greatly relaxed for certain pairs of exposures.

To investigate the effect of the penalty parameter κ_{uv} , we let $p = 2$, corresponding to a single interaction h_{12} , and simulate draws from the prior π in (3.7). We fix $\tau_{12,1} = \tau_{12,2} = \nu = 1$ and vary $\kappa_{12} \in \{0, 0.1, 1, 10, 100, 1000\}$. Given a value of κ_{12} , we draw $\Psi_{12} \sim \pi$ and compute $\mathcal{W} = \{\int h_{12}^+(x_u, x_v) dx_u dx_v\} \{\int h_{12}^-(x_u, x_v) dx_u dx_v\}$, where $h_{12}^+(x_u, x_v) = \max\{h_{12}(x_u, x_v), 0\}$ and $h_{12}^-(x_u, x_v) = \max\{-h_{12}(x_u, x_v), 0\}$. It is clear that $\mathcal{W} = 0$ if and only if h_{12} is either synergistic, antagonistic, or null. In Figure 3.2, we plot the proportion of prior draws of \mathcal{W} less than 0.001 for each value of κ_{12} . The proportion of prior samples of \mathcal{W} less than 0.001 increases from 0.13 to 0.87 as we increase the penalty κ_{12} from 0 to 1000. We also looked at the maximum of \mathcal{W} over prior draws, which decreased from 939.03 to 0.05 as κ_{12} increased from 0 to 1000. This shows that the proposed prior distribution is flexible enough to capture arbitrary interactions for small values of κ_{12} , while favoring interactions that are very close to being synergistic, antagonistic, or null for large values of κ_{12} .

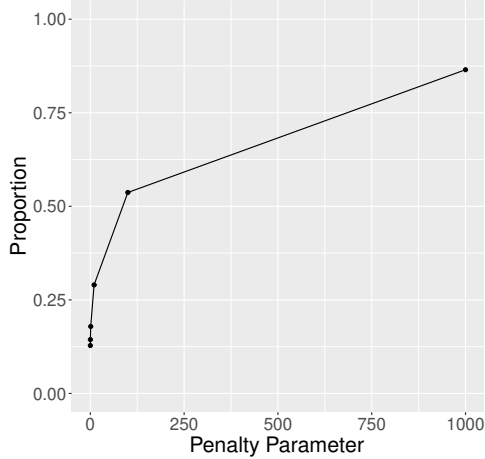


FIGURE 3.2: Plot showing proportion of prior draws of \mathcal{W} less than 0.001 as κ_{12} increases. \mathcal{W} is a measure of deviation of the interaction h_{12} from being synergistic, antagonistic, or null.

We also tried other approaches to model h_{uv} . Instead of squaring unconstrained functions, one could potentially use a Bayesian monotone spline formulation based on restricting the sign of the basis coefficients. However, we found squaring an unconstrained function to provide better estimates, particularly when the non-negative function being modeled is near 0. We also tried using a tensor product spline model for the bivariate functions, but using products of univariate functions had better power. We describe these alternative approaches in more detail in Section B.4 of the Supplementary Material.

3.3.3 Variable Selection

To select the non-null interactions h_{uv} s, we first decompose

$$h_{uv}(x_u, x_v) = h_{uv}^+(x_u, x_v) - h_{uv}^-(x_u, x_v), \quad (3.9)$$

where $h_{uv}^+(x_u, x_v) = \max\{h_{uv}(x_u, x_v), 0\} \geq 0$ and $h_{uv}^-(x_u, x_v) = \max\{-h_{uv}(x_u, x_v), 0\} \geq 0$, for any $(x_u, x_v) \in [0, 1]^2$. The interaction h_{uv} is synergistic if and only if $h_{uv}^+ \not\equiv 0$ and $h_{uv}^- \equiv 0$, antagonistic if and only if $h_{uv}^+ \equiv 0$ and $h_{uv}^- \not\equiv 0$, and null if and only if $h_{uv}^+ = h_{uv}^- \equiv 0$. These conditions may be rewritten in terms of their integrals since

for a continuous non-negative function P , $P \equiv 0$ if and only if $\int P = 0$ and $P \neq 0$ if and only if $\int P > 0$.

Because we are using a continuous shrinkage prior that places zero probability on h_{uv}^+ or h_{uv}^- being *exactly* zero, we treat the integrals $\int h_{uv}^+$ and $\int h_{uv}^-$ as effectively zero if they are below a cutoff $c_0 > 0$. We call c_0 the *integral cutoff*. Let $C_{uv} = \{\int h_{uv}^+ \leq c_0\}$ and $D_{uv} = \{\int h_{uv}^- \leq c_0\}$ denote the events that the sizes of h_{uv}^+ and h_{uv}^- are less than the integral cutoff c_0 , respectively. In practice, we compute Monte Carlo estimates of the posterior probabilities $\mathbf{P}(C_{uv} \cap D_{uv})$, $\mathbf{P}(C_{uv}^c \cap D_{uv})$, and $\mathbf{P}(C_{uv} \cap D_{uv}^c)$; these are then used as the estimates of the posterior probabilities that h_{uv} is null, synergistic, and antagonistic, respectively. In particular, the posterior inclusion probability (PIP) of the interaction h_{uv} is given by

$$\text{PIP}(h_{uv}) = 1 - \mathbf{P}(C_{uv} \cap D_{uv}) = \mathbf{P}(C_{uv}^c \cup D_{uv}^c) \quad (3.10)$$

In practice, we tried with different integral cutoff values $c_0 = 0.005, 0.01, 0.05, 0.10$. We provide further details on sensitivity analysis to c_0 in Section B.6 of the Supplementary Material. Based on the sensitivity analysis, we recommend using an integral cutoff of $c_0 = 0.01$, after standardizing y to have variance 1 prior to fitting SAID. As a rough rule of thumb, loosely based on Kass and Raftery (1995), one can view interactions having $\text{PIP}(h_{uv}) \in (0.5, 0.75)$ as barely worth a mention in terms of weight of evidence against the null, $\text{PIP}(h_{uv}) \in [0.75, 0.95)$ as weakly to moderately suggestive, $\text{PIP}(h_{uv}) \in [0.95, 0.99)$ as strong evidence, and $\text{PIP}(h_{uv}) \geq 0.99$ as very strong evidence.

3.3.4 Main Effects and Other Parameters

For reasons of identifiability, we assume the main effects f_1, \dots, f_p either satisfy $\int_0^1 f_j(x_j) dx_j = 0$ for each $j = 1, \dots, p$ or $f_j(0) = 0$ for $j = 1, \dots, p$. We call the former set of conditions as *integral constraints* and the latter as *origin constraints*. Suppose

that for exposure j , $\{b_{j,1}(\cdot), \dots, b_{j,d}(\cdot)\}$ represent one-dimensional basis functions, such as B-splines. To enforce the identifiability conditions, we let the functions $b_{j,1}, \dots, b_{j,d}$ either satisfy $\int_0^1 b_{j,u}(x_j) dx_j = 0$ for $u = 1, \dots, d$, $j = 1, \dots, p$ if the main effects satisfy the *integral constraint*, or $b_{j,u}(0) = 0$ for $u = 1, \dots, d$, $j = 1, \dots, p$ if the main effects satisfy the *origin constraint*. We provide further details in Sections B.1 and B.3 of the Supplementary Material. Let $\mathbf{b}_j(x_j) = (b_{j,1}(x_j), \dots, b_{j,d}(x_j))^\top$. We now model f_j as

$$f_j(x_j) = \sum_{u=1}^d b_{j,u}(x_j) \gamma_{j,u} = \mathbf{b}_j(x_j)^\top \boldsymbol{\Gamma}_j, \quad (3.11)$$

where $\boldsymbol{\Gamma}_j = (\gamma_{j,1}, \dots, \gamma_{j,d})^\top$. To estimate the coefficient vector $\boldsymbol{\Gamma}_j$, we assume a univariate P-spline prior distribution (Lang and Brezger, 2004) on $\boldsymbol{\Gamma}_j$. With an appropriate positive-definite matrix Σ_M , we can write this prior as

$$\boldsymbol{\Gamma}_j \mid \lambda_j \sim N\left(0, \frac{\Sigma_M}{\lambda_j}\right), \quad \lambda_j \sim G(a, a), \quad (3.12)$$

where $\lambda_j > 0$ is a scale parameter corresponding to the j -th main effect. The sensitivity to the choice of hyperparameter a for the distribution of λ_j has been investigated in Lang and Brezger (2004); we found $a = 0.5$ to work the best across simulations and applications.

To complete prior specification for the parameters in (3.1)-(3.2), we put vague prior distributions on the intercept α and covariate effects η , and a non-informative prior on the measurement error variance σ^2 :

$$\alpha \sim N(0, 10^4), \quad \eta \sim N(0, 10^4 \mathbb{I}_q), \quad \sigma^2 \sim \sigma^{-2}. \quad (3.13)$$

3.3.5 Posterior Sampling

We rely on a Hamiltonian Monte Carlo (HMC) (Neal et al., 2011; Betancourt and Girolami, 2015; Hoffman et al., 2014)-within-Gibbs algorithm, with HMC used to

sample the interaction parameters Ψ_{uv} for $1 \leq u < v \leq p$, and other parameters updated in Gibbs steps. Although this approach is highly effective in our experiments, it is important to carefully choose the step size e_0 and step length L_0 in HMC. In practice, we found $e_0 \approx 0.01$ and $L_0 \approx 10$ to work well.

The augmented prior distribution π introduced in Section 3.3.2 creates difficulty when sampling from the posterior using MCMC. This is because the normalizing constant

$$Z(\kappa, \tau_1, \tau_2, \nu) = \int \exp\{-\kappa \tilde{\mathcal{Q}}(\Psi)\} \pi_0(\Psi \mid \tau_1, \tau_2, \nu) d\Psi$$

is not easy to evaluate numerically. To bypass this difficulty, we proceed as in Rao et al. (2016), by augmenting rejected proposals originating from a rejection sampling mechanism. Introducing the rejected proposals eliminates the normalizing constant Z and makes sampling from the posterior feasible using standard MCMC algorithms.

Let $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ and $\mathbf{Z} = [\mathbf{z}_1 \mid \dots \mid \mathbf{z}_n]^\top \in \mathbb{R}^{n \times q}$. Following Sections 3.3.2 and 3.3.4, let $B_1, \dots, B_p \in \mathbb{R}^{n \times d}$ and $S_1, \dots, S_p \in \mathbb{R}^{n \times m}$ be such that the i th row of B_j and S_j is given by $\mathbf{b}_j^\top(x_{ij})$ and $\mathbf{s}_j^\top(x_{ij})$, respectively. Let \mathcal{M} be the block diagonal matrix given by $\mathcal{M} = \text{block-diag}(10^4, 10^4 \mathbb{I}_q, \Sigma_M/\lambda_1, \dots, \Sigma_M/\lambda_p)$. We also let $\tilde{\mathbf{B}} = [1_n \mid \mathbf{Z} \mid B_1 \mid \dots \mid B_p] \in \mathbb{R}^{n \times (1+q+pd)}$, $\mathcal{G} = (\alpha, \eta^\top, \Gamma_1^\top, \dots, \Gamma_p^\top)^\top$, and for $1 \leq u < v \leq p$, we let $\mathbf{h}_{uv} = (S_u \theta_{uv,1})^2 (S_v \phi_{uv,1})^2 - (S_u \theta_{uv,2})^2 (S_v \phi_{uv,2})^2$ denote the vector of the uv th interaction evaluated at the data points, and $\Theta_{uv} = (\Psi_{uv}^\top, \tau_{uv,1}, \tau_{uv,2}, \kappa_{uv})^\top$ be the set of parameters for the uv -th interaction. For a matrix M_0 and a finite set S_0 , denote their trace and cardinality by $\text{tr}(M_0)$ and $|S_0|$, respectively. Posterior sampling then proceeds as follows.

1. Sample $\mathcal{G} \mid - \sim N\left(\mathcal{A}^{-1} \frac{\tilde{\mathbf{B}}^\top \xi}{\sigma^2}, \mathcal{A}^{-1}\right)$ using Rue (2001), where

$$\mathcal{A} = \frac{\tilde{\mathbf{B}}^\top \tilde{\mathbf{B}}}{\sigma^2} + \mathcal{M}^{-1}, \quad \xi = \mathbf{y} - \sum_{1 \leq u < v \leq p} \mathbf{h}_{uv}.$$

2. For each $j = 1, \dots, p$, sample $\lambda_j \mid - \sim G\left(a + \frac{d}{2}, a + \frac{\mathbf{\Gamma}_j^\top \Sigma_M^{-1} \mathbf{\Gamma}_j}{2}\right)$.
3. For each u, v with $1 \leq u < v \leq p$, let $\mathcal{I}_{uv} = \{(i_1, i_2) : 1 \leq i_1 < i_2 \leq p, (i_1, i_2) \neq (u, v)\}$.

(a) Define $\mathbf{\Delta}_{uv} = \mathbf{y} - \left(\tilde{\mathbf{B}}\mathbf{g} + \sum_{(u', v') \in \mathcal{I}_{uv}} \mathbf{h}_{u'v'} \right)$.

- (b) Given $\kappa_{uv}, \tau_{uv,1}, \tau_{uv,2}, \nu$, repeatedly sample independent $\mathcal{Y}_{uv,j} \sim \pi_0(\cdot \mid \tau_{uv,1}, \tau_{uv,2}, \nu)$ for $j = 1, 2, \dots$ until a sample \mathcal{Y}_{uv,j_0} is accepted; accept $\mathcal{Y}_{uv,j}$ with probability $\exp\{-\kappa_{uv} \tilde{\mathcal{Q}}(\mathcal{Y}_{uv,j})\}$ for $j = 1, 2, \dots$. Discard \mathcal{Y}_{uv,j_0} and form the set of rejected proposals $\mathcal{R}_{uv} = \{\mathcal{Y}_{uv,1}, \dots, \mathcal{Y}_{uv,|\mathcal{R}_{uv}|}\} = \{\mathcal{Y}_{uv,1}, \dots, \mathcal{Y}_{uv,j_0-1}\}$.

- (c) Given $\mathbf{\Delta}_{uv}, \mathcal{R}_{uv}$, and ν , use HMC to draw one sample of $\mathbf{\Theta}_{uv}$, targeting

$$\begin{aligned} \Pi(\mathbf{\Theta}_{uv} \mid \mathbf{\Delta}_{uv}, \mathcal{R}_{uv}, \nu) &\propto \pi_0(\mathbf{\Psi}_{uv} \mid \tau_{uv,1}, \tau_{uv,2}, \nu) \exp\{-\kappa_{uv} \tilde{\mathcal{Q}}(\mathbf{\Psi}_{uv})\} \\ &\times \prod_{j=1}^{|\mathcal{R}_{uv}|} \pi_0(\mathcal{Y}_{uv,j} \mid \tau_{uv,1}, \tau_{uv,2}, \nu) \left[1 - \exp\{-\kappa_{uv} \tilde{\mathcal{Q}}(\mathcal{Y}_{uv,j})\}\right] \\ &\times \pi_\tau(\tau_{uv,1}) \pi_\tau(\tau_{uv,2}) \pi_\kappa(\kappa_{uv}) N(\mathbf{\Delta}_{uv} \mid \mathbf{h}_{uv}, \sigma^2 \mathbb{I}_n), \end{aligned}$$

where π_τ is the $C^+(0, 1)$ density, and π_κ is the standard lognormal density.

- (d) Update the uv -th interaction $\mathbf{h}_{uv} = (S_u \theta_{uv,1})^2 (S_v \phi_{uv,1})^2 - (S_u \theta_{uv,2})^2 (S_v \phi_{uv,2})^2$.

4. Following Makalic and Schmidt (2015), we introduce W such that $\nu^2 \mid W \sim \text{IG}(1/2, 1/W)$ and $W \sim \text{IG}(1/2, 1)$, which leads to $\nu \sim C^+(0, 1)$. For each $1 \leq u < v \leq p$, we decompose each rejected proposal $\mathcal{Y}_{uv,j}$ as $\mathcal{Y}_{uv,j} = (\tilde{\mathcal{Y}}_{uv,j1}^\top, \tilde{\mathcal{Y}}_{uv,j2}^\top, \tilde{\mathcal{Y}}_{uv,j3}^\top, \tilde{\mathcal{Y}}_{uv,j4}^\top)^\top$, for $j = 1, \dots, |\mathcal{R}_{uv}|$. Form $\mathbf{R}_{uv,l} \in \mathbb{R}^{|\mathcal{R}_{uv}| \times m}$ with its j -th row given by $\tilde{\mathcal{Y}}_{uv,jl}^\top$, for $j = 1, \dots, |\mathcal{R}_{uv}|$ and $l = 1, 2, 3, 4$. Let $n_R = \sum_{1 \leq u < v \leq p} |\mathcal{R}_{uv}|$. The full conditional updates for ν^2 and W are given by:

(a) $\nu^2 \mid - \sim \text{IG} \left(\frac{1}{2} + 2m \left\{ \binom{p}{2} + n_R \right\}, \frac{1}{W} + \frac{1}{2} \sum_{1 \leq u < v \leq p} (r_{uv} + t_{uv}) \right)$, where

$$r_{uv} = \frac{\theta_{uv,1}^\top \Sigma_0^{-1} \theta_{uv,1} + \phi_{uv,1}^\top \Sigma_0^{-1} \phi_{uv,1}}{\tau_{uv,1}^2} + \frac{\theta_{uv,2}^\top \Sigma_0^{-1} \theta_{uv,2} + \phi_{uv,2}^\top \Sigma_0^{-1} \phi_{uv,2}}{\tau_{uv,2}^2},$$

$$t_{uv} = \frac{\text{tr} \left\{ \Sigma_0^{-1} (\mathbf{R}_{uv,1}^\top \mathbf{R}_{uv,1} + \mathbf{R}_{uv,2}^\top \mathbf{R}_{uv,2}) \right\}}{\tau_{uv,1}^2} + \frac{\text{tr} \left\{ \Sigma_0^{-1} (\mathbf{R}_{uv,3}^\top \mathbf{R}_{uv,3} + \mathbf{R}_{uv,4}^\top \mathbf{R}_{uv,4}) \right\}}{\tau_{uv,2}^2}.$$

(b) $W \mid - \sim \text{IG}(1, 1 + \nu^{-2})$.

5. Sample $\sigma^2 \mid - \sim \text{IG} \left(\frac{n}{2}, \frac{(\mathbf{y} - \mu)^\top (\mathbf{y} - \mu)}{2} \right)$, where $\mu = \tilde{\mathbf{B}}\mathcal{G} + \sum_{1 \leq u < v \leq p} \mathbf{h}_{uv}$.

We repeat Steps 1-5 for a large number of iterations, and base inference on the resulting samples after discarding a burn-in. We provide further details in Section B.5 of the Supplementary Material. For code, we developed the R package SAID available for download at <https://github.com/shounakchattopadhyay/SAID>, which was used in the application and numerical experiments.

3.4 Simulation Examples

3.4.1 Preliminaries

We carry out simulation studies comparing SAID with the competitors BKMR (Bobb et al., 2015), MixSelect (Ferrari and Dunson, 2020), HierNet (Bien et al., 2013), Family (Haris et al., 2016), PIE (Wang et al., 2019), and RAMP (Hao et al., 2018). We generate data according to:

$$y_i = H(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_0^2), \quad (3.14)$$

where H is the exposure dose response surface decomposed as in (3.2), and σ_0^2 is the true error variance. As before, we assume that the exposures satisfy $\mathbf{x}_i \in [0, 1]^p$. We consider two dimensions: $p = 2$ and $p = 10$. For $p = 2$, we compare the methods in terms of their estimation performance for varying interaction signal strength and error variance, across three different types of interactions. For $p = 10$, we consider accuracy in estimation and variable selection for pairwise interactions.

HierNet, Family, PIE, and RAMP estimate the dose response surface H using quadratic regression, which also provides estimates of pairwise interaction functions using bilinear surfaces of the form $f(x, x') = \gamma xx'$. MixSelect combines quadratic regression with an additional nonlinear deviation term in the model, assumed to be orthogonal to the quadratic regression. This approach provides estimates of pairwise interactions in the spirit of quadratic regression, while improving flexibility in capturing H . BKMR estimates H using unconstrained Gaussian processes incorporating variable selection. Thus, we do not consider BKMR as a competitor when estimating pairwise interactions, as BKMR does not provide estimates of these interactions.

Given training points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we first estimate H and the pairwise interactions h_{uv} using the relevant method and denote the estimates by \hat{H} and \hat{h}_{uv} , respectively. For the Bayesian methods, we use the posterior mean as the estimator. We evaluate the true surface H and true interactions h_{uv} at test exposure points $\{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{n_t}\}$, denoted by

$$\tilde{\mathbf{H}} = (H(\tilde{\mathbf{x}}_1), \dots, H(\tilde{\mathbf{x}}_{n_t}))^\top$$

and

$$\tilde{\mathbf{h}}_{uv} = (h_{uv}(\tilde{x}_{1u}, \tilde{x}_{1v}), \dots, h_{uv}(\tilde{x}_{n_t,u}, \tilde{x}_{n_t,v}))^\top,$$

respectively. Finally, we estimate $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{h}}_{uv}$ for $1 \leq u < v \leq p$ by replacing H and h_{uv} with \hat{H} and \hat{h}_{uv} , respectively. The estimates are evaluated using root mean squared error (RMSE).

We also consider the *variable selection accuracy* of competing methods. For each method, we estimate the case 1 and case 2 errors from classifying an interaction as synergistic, antagonistic, or null. For any category, the case 1 error probability is given by the probability of misclassifying an interaction as not belonging to that category, when in truth the interaction is in that category. The case 2 error probability is given by the probability of misclassifying an interaction as belonging to that category when in truth, it does not belong to that category.

3.4.2 Two Exposures

We first carry out simulation experiments with $p = 2$ exposures. We consider three different scenarios and evaluate the competitors on test set RMSE in estimating H and the pairwise interaction h_{12} . In all three scenarios, we train the methods on a sample of size $n = 500$ and evaluate on $n_t = 500$ test points. The pairwise interaction h_{12} is taken to be of the form $h_{12}(x_1, x_2) = \gamma_0 f(x_1, x_2)$, where γ_0 is the interaction strength and f is a baseline interaction function which we vary across the three different scenarios. We vary $\gamma_0 \in \{1, 2\}$ and $\sigma_0^2 \in \{0.1, 0.5\}$, thereby considering low-to-moderate interaction strength and error variance. For each method in a particular scenario and a pair (γ_0, σ_0^2) , we replicate the experiment $R = 10$ times and report the average error across replicates. When fitting SAID, we assume that the main effects satisfy the origin constraint.

We first consider a scenario where the true interaction is synergistic and nonlinear in nature and denote this case by SN. To elaborate, the dose response surface $H(\cdot)$ is given by

$$H(x_1, x_2) = 0.5 + x_1^2 + x_2^2 + \gamma_0 x_1^2 x_2^2.$$

The pairwise interaction $h_{12}(x_1, x_2)$ in this case is $h_{12}(x_1, x_2) = \gamma_0 x_1^2 x_2^2$. The RMSEs of each method for varying signal and noise components are provided in Tables 3.1 and 3.2. Although methods such as BKMR and RAMP have similar out-of-sample

Table 3.1: RMSE of competitors when estimating H in case SN. Here $n = 500$ and $p = 2$.

Signal and Noise	SAID	BKMR	MixSelect	HierNet	Family	PIE	RAMP
$\gamma_0 = 1, \sigma_0^2 = 0.1$	0.05	0.04	0.15	0.22	0.40	1.41	0.05
$\gamma_0 = 1, \sigma_0^2 = 0.5$	0.10	0.10	0.16	0.22	0.41	1.29	0.09
$\gamma_0 = 2, \sigma_0^2 = 0.1$	0.05	0.05	0.13	0.24	0.51	2.18	0.07
$\gamma_0 = 2, \sigma_0^2 = 0.5$	0.10	0.10	0.19	0.24	0.51	2.16	0.10

Table 3.2: RMSE of competitors when estimating h_{12} in case SN. Here $n = 500$ and $p = 2$.

Signal and Noise	SAID	MixSelect	HierNet	Family	PIE	RAMP
$\gamma_0 = 1, \sigma_0^2 = 0.1$	0.06	0.16	0.19	0.18	0.18	0.18
$\gamma_0 = 1, \sigma_0^2 = 0.5$	0.11	0.19	0.16	0.18	0.24	0.21
$\gamma_0 = 2, \sigma_0^2 = 0.1$	0.08	0.24	0.17	0.37	0.34	0.34
$\gamma_0 = 2, \sigma_0^2 = 0.5$	0.13	0.28	0.18	0.37	0.37	0.36

RMSEs with respect to SAID, the proposed approach shows superior performance in estimating the interaction term. In particular, the gains of using SAID are the most prominent when the signal strength of the interaction is the highest, indicating lack of flexibility of the other methods.

Secondly, we assume the data generating process has linear main effects and a synergistic linear interaction, namely a quadratic regression (QR) setup. We let H be

$$H(x_1, x_2) = 0.5 + x_1 + x_2 + \gamma_0 x_1 x_2.$$

The interaction is given by $h_{12}(x_1, x_2) = \gamma_0 x_1 x_2$. Lastly, we consider a scenario where the interaction is nonlinear and is neither synergistic, antagonistic, or null, while keeping the main effects as before. We call this the mis-specified interaction (MIS) case and let H be

$$H(x_1, x_2) = 0.5 + x_1 + x_2 + \gamma_0(x_1 x_2 - 2x_1^2 x_2^2).$$

In this case, the interaction is given by $h_{12}(x_1, x_2) = \gamma_0(x_1 x_2 - 2x_1^2 x_2^2)$. The interaction

h_{12} is non-negative for $x_1x_2 \leq 1/2$ and non-positive for $x_1x_2 \geq 1/2$, and thus is neither synergistic, antagonistic, or null. To ease exposition, we defer the results obtained from scenarios QR and MIS to Section B.7 of the Supplementary Material. In the scenario QR, quadratic regression approaches such as RAMP and MixSelect perform better as the data is also generated from a quadratic regression setup. However, the performance of SAID is similar to its competitors for estimating both the dose response surface and the interaction surface. In the scenario MIS, BKMR and SAID perform the best in terms of estimating H . When estimating the interaction, SAID outperforms its competitors due to the lack of flexibility for other pairwise interaction approaches.

3.4.3 More than Two Exposures

In this subsection, we consider a moderate dimensional example with $p = 10$ and thus $\binom{10}{2} = 45$ pairwise interactions. We assume that the true data generating model has 5 synergistic, 5 antagonistic, and 35 null pairwise interactions. We let $\sigma_0^2 = 0.2$ with the number of training and test points taken to be $n = 1000$ and $n_t = 500$, respectively. The experiment is replicated $R = 20$ times. We fit all the quadratic regression methods assuming weak heredity of the pairwise interactions. Family is not considered due to unstable estimates; furthermore, except when estimating H , we do not consider BKMR.

We assume (3.14) and decompose $H(\mathbf{x})$ in the following way:

$$H(\mathbf{x}) = \alpha_0 + M_0(\mathbf{x}) + S_0(\mathbf{x}) + A_0(\mathbf{x}),$$

where

$$\begin{aligned}\alpha_0 &= -5/6 \\ M_0(\mathbf{x}) &= (x_1 + x_1^2) + \frac{x_2}{2} + x_7^3, \\ S_0(\mathbf{x}) &= 4(x_1 - x_1^2)x_2 + x_1x_9 + x_2^2x_3^2 + x_3x_8 + \frac{(e^{x_5} - 1)x_{10}}{e - 1}, \\ A_0(\mathbf{x}) &= - \left[x_1x_3 + x_2^2x_5 + \frac{27}{4}x_4^2(1 - x_4)x_9 + x_7x_{10} + x_8x_9^2 \right].\end{aligned}$$

Here, M_0 , S_0 , A_0 denote the true main effects, synergistic interaction effects, and the antagonistic interaction effects, respectively. The form of the interactions include pairwise linear, nonlinear polynomial, and nonlinear interactions which are not of polynomial form. Each pairwise interaction has absolute maximum value equal to 1. As before, we assume the main effects start from the origin when fitting SAID. The results comparing the methods are given in Table B.1. For comparison, the RMSE when estimating H using BKMR is 0.17.

In terms of estimation accuracy, SAID performs uniformly better than its competitors in estimating both the overall surface H and the interaction surface I , given by $I(\mathbf{x}) = S_0(\mathbf{x}) + A_0(\mathbf{x})$. We believe that this is due to the SAID prior on the pairwise interactions being flexible enough to capture nonlinear interactions while also efficiently extracting interaction signal in the presence of noise. Furthermore, the SAID framework does not require any heredity assumptions, helping detection of pairwise interactions even in the absence of main effects of one or both of the corresponding exposures. In terms of variable selection, the case 1 error probabilities for synergistic and antagonistic detections are similar across the methods. However, SAID has considerably less case 1 error when detecting null interactions, indicating that SAID more accurately classifies null interactions to be null compared with its competitors. SAID also shows superior performance in terms of case 2 error when classifying both synergistic and antagonistic interactions. This indicates that SAID

Table 3.3: Comparison of the methods in terms of RMSE and variable selection accuracy for $p = 10$.

Criterion	SAID	MixSelect	HierNet	PIE	RAMP
Overall Surface RMSE	0.13	0.29	0.30	0.75	0.18
Interaction Surface RMSE	0.21	0.52	0.48	0.42	0.43
Synergistic Case 1 Probability	0.008	0.040	0.001	0.008	0.015
Synergistic Case2 Probability	0.001	0.220	0.450	0.210	0.230
Antagonistic Case1 Probability	0.005	0.013	0.002	0.006	0.008
Antagonistic Case2 Probability	0.001	0.050	0.240	0.040	0.050
Null Case1 Probability	0.001	0.135	0.345	0.125	0.140
Null Case2 Probability	0.014	0.056	0.002	0.016	0.026

has a lower probability of misclassifying an interaction as synergistic or antagonistic when it is not so, compared with its competitors. Out of the methods considered, SAID is the only one with case 1 and case 2 classification errors less than 0.05 for each case. In terms of uncertainty quantification of interactions evaluated at the test points, the 95% posterior credible intervals obtained from SAID have 94.6% coverage, averaged over $R = 20$ replicates and all 45 interactions.

3.5 Analysis of Kidney Function Data

3.5.1 Preliminaries

In this Section, we apply the proposed Synergistic Antagonistic Interaction Detection (SAID) approach to the NHANES 2015-16 data. We are interested in detecting synergistic and antagonistic interactions between heavy metals affecting kidney function. Following Section 3.2 and the results in Section 4.4, it is evident that the current methods either do not provide inferences on interactions or lack flexibility in characterizing interactions. We now illustrate how SAID detects synergistic, antagonistic, and null interactions.

As discussed in Section 3.2.1, we assess kidney function of individuals through their urine creatinine (uCr) levels, measured in mg/dL. Heavy metal concentrations

are also measured in urine. We consider 13 heavy metals, namely Antimony (Sb), Barium (Ba), Cadmium (Cd), Cesium (Cs), Cobalt (Co), Lead (Pb), Manganese (Mn), Molybdenum (Mo), Strontium (Sr), Thallium (Tl), Tin (Sn), Tungsten (W), and Uranium (U), all measured in $\mu\text{g}/\text{mL}$. Following Section 3.2.2, we remove individuals with albuminuria and missing entries from the original data set. The sample size after removal of such entries is $n = 1979$. As described in Section 3.2.3, we multiply both the uCr and the heavy metal concentration levels by the individual-specific urine flow rate to obtain their dilution-adjusted versions. Furthermore, we also consider age (in years), sex (0 for males and 1 for females), ethnicity (Non-Hispanic White, Non-Hispanic Black, Mexican American, Other Hispanic, and Other), and body mass index (BMI) of the subject as covariates.

Before beginning our analysis, we first (natural) log-transformed the dilution-adjusted urine creatinine levels. We use a marginal cumulative distribution function (CDF) transformation for standardizing exposure variables. Suppose the dilution-adjusted exposure variables are denoted by E_1, \dots, E_p , with the marginal CDF of variable j denoted by F_j for $j = 1, \dots, p$. To estimate F_j , we use kernel density estimation to first estimate the marginal density of E_j and then estimate the induced CDF \hat{F}_j from the density estimate. The transformed exposures are defined as $x_j = \hat{F}_j(E_j) \in [0, 1]$. Standardizing exposures in this manner facilitates statistical inferences by reducing the tendency for the exposure data to be unevenly distributed, with very sparse observations for certain ranges of exposure. Transforming the exposures does not complicate interpretation, as we can convert dose response surfaces back to the original units. In addition, the transformed exposures are directly interpretable as quantiles of the exposure distribution in the sample; for example, $x_j = 0.5$ reflects a median value for the j th exposure.

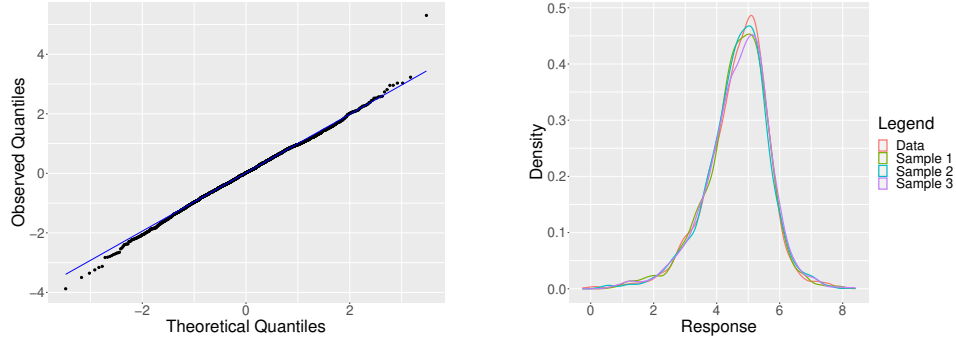
For the i -th individual in the data set, let $y_i \in \mathbb{R}$ be the log of the dilution-

adjusted urinary creatinine level, $\mathbf{x}_i \in [0, 1]^{13}$ be the 13 dilution-adjusted urinary metal concentrations after CDF transformations, and $\mathbf{z}_i \in \mathbb{R}^7$ be the covariate vector, with dummy variables for the different ethnic groups. We use “Non-Hispanic White” as the baseline category in defining indicators. The dimensions of exposures and covariates are $p = 13$ and $q = 7$, respectively. We also standardized the covariates age and BMI to have variance 1 before fitting the model.

Following Section 3.3.5, we employed a Hamiltonian Monte Carlo (HMC)-within-Gibbs sampler to obtain posterior samples of the model parameters. We ran the sampler for a total of 15000 iterations and discarded the first 5000 iterations as burn-in. We standardized the y to have variance 1 before fitting the model and then used an integral cutoff of 0.01 as described in Section 3.3.3 in order to compute PIPs of pairwise interactions. After computing the PIPs, we present our inferences in the original scale, that is, where y is not standardized. This is achieved simply by multiplying the estimates of the intercept, main effects, interaction effects, covariate effects, and measurement error standard deviation obtained from the fitted model by the standard deviation of y .

3.5.2 Model Diagnostics

We first assess MCMC convergence. To improve mixing, we perturb the step-size e_0 by a small factor every 500 iterations. As a measure of mixing of the chain, we look at the MCMC samples of the error variance σ^2 . The effective sample size of σ^2 is 33.7% of the total number of MCMC samples, after discarding the burn-in. We found this proportion to remain fairly robust across longer chains. On a MacBook Pro with M1 CPU and 16 GB of RAM, it took ~ 40 minutes to fit SAID on this dataset; for comparison, MixSelect took ~ 6 hours with the same number of MCMC iterates. Computation time can be improved by taking shorter chains, as 15000 was conservative based on our assessments.



(a) Q-Q Plot of standardized residuals. (b) Posterior predictive check.

FIGURE 3.3: Figure showing Q-Q plot of standardized residuals and marginal density plots of posterior predictive samples of urine creatinine levels, obtained from the NHANES 2015-16 data. Deviation from the reference blue line in Q-Q plot is deviation from normality.

Next, we assess goodness-of-fit by inspecting standardized residuals and carrying out posterior predictive checks (Gelman et al., 1996). We generate posterior predictive samples corresponding to each training point. To validate the assumption of normality of the measurement error, we look at a Q-Q plot of standardized residuals, defined to be the difference of the observed and predicted responses and then standardized. We also compare the marginal density of the observed response variable with the marginal densities of 3 randomly chosen MCMC samples of predicted responses. We provide both the Q-Q plot of the obtained standardized residuals and the comparison of marginal density plots in Figure 3.3. Figure 3.3(a) indicates that the normality assumption is justified; furthermore, the densities of the randomly chosen predicted response samples in Figure 3.3(b) very closely resemble the marginal density of the observed response. Lastly, the coverage of 95% posterior predictive intervals of the responses, averaged over all observed responses, is 96%.

3.5.3 Results

In this subsection, we discuss the main results of our analyses of the data described in Section 3.5.1. For each exposure, we constrain the main effects to start from 0 at

the minimum observed value of that exposure. The 95% posterior credible interval of the error variance σ^2 is [0.080, 0.091] with a posterior mean of 0.086. The exposures and covariates together explain 89% of the variation in the response.

We observed nonlinear main effects for Antimony, Cadmium, Cobalt, Cesium, Molybdenum, Tin, and Uranium. The general trend for the main effects of these heavy metals is an increase in log dilution-adjusted urine creatinine as exposure to metal concentrations increases. Since urine concentrations of creatinine are directly correlated with serum concentrations of creatinine, this might indicate higher kidney stress at higher levels of metal exposure. The main effects have either a monotonic or hill-shaped pattern, which is as expected in studying health effects of potentially toxic exposures. We found exposure to high doses of Cesium to increase log dilution-adjusted urine creatinine the most, followed by Cadmium. For illustration, we plot the main effects of Cadmium, Cesium, Molybdenum, and Uranium in Figure 3.4, with both the response and the exposures shown in log scale. We provide further plots for the main effects of the other exposures in Section B.8 of the Supplementary Material. When the exposures are CDF transformed, we can evaluate the main effect of an exposure at a desired quantile. As an illustration, the metals Antimony, Cadmium, Cobalt, Cesium, Molybdenum, Tin, and Uranium at their median exposure levels have a main effect of 0.23, 0.69, 0.52, 0.88, 0.39, 0.22, and 0.22, respectively, on log dilution-adjusted Creatinine. Similar main effects of heavy metal exposures on kidney function have been detected in earlier literature. For example, Ferraro et al. (2010) found Cadmium exposure to be associated with increased risk of chronic kidney disease, based on an analysis of NHANES data from 1999-2006. In a recent study, Rahman et al. (2022) found Cesium, Cadmium, and Antimony to be associated with kidney damage.

The proposed approach detects multiple synergistic and antagonistic interactions between exposures. Excluding the interactions with $PIP < 0.5$, the detected inter-

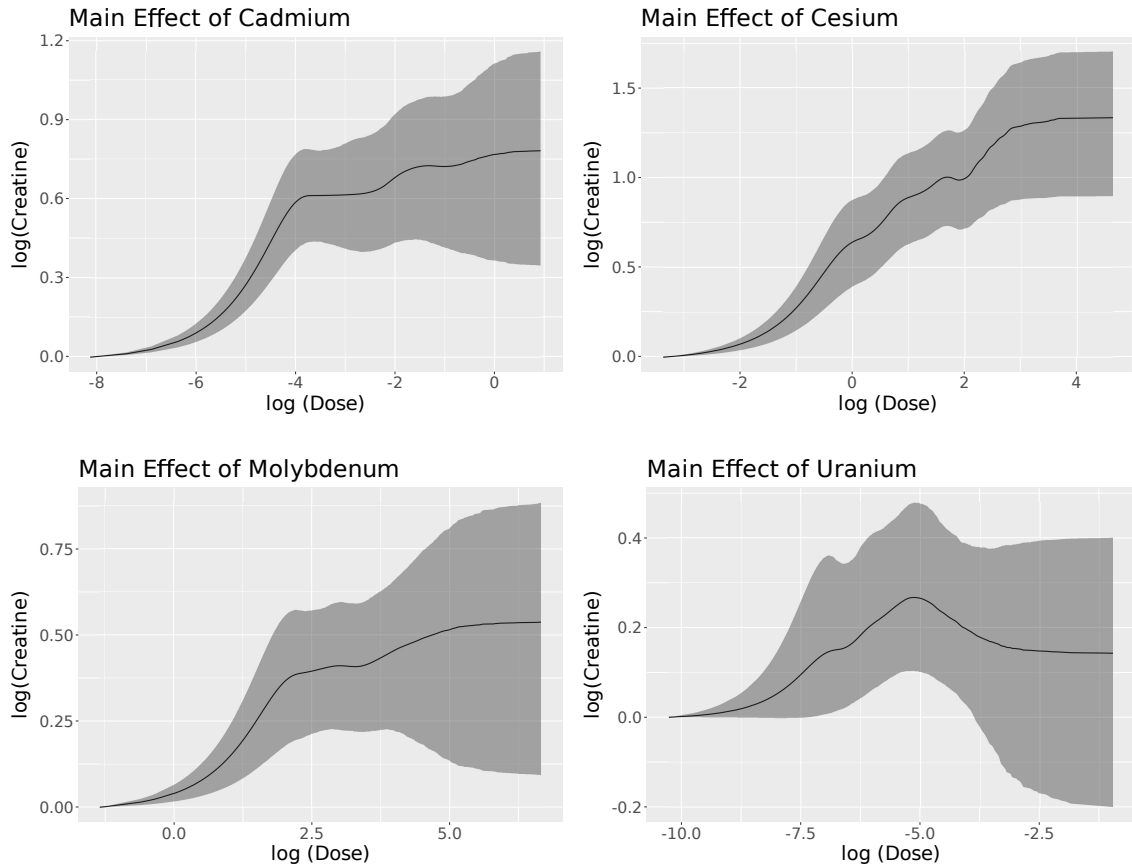


FIGURE 3.4: Plots showing the main effects of dilution-adjusted Cadmium, Cesium, Molybdenum, and Uranium on log dilution-adjusted Creatinine. Exposure levels are in log scale. Black line denotes posterior mean and shaded regions denote pointwise 95% posterior credible intervals.

actions were Cadmium and Tin (PIP > 0.99), Molybdenum and Tin (PIP > 0.99), Cadmium and Manganese (PIP = 0.99), and Cobalt and Manganese (PIP = 0.98). For these interactions, we also compute the posterior synergistic probability (PSP) and posterior antagonistic probability (PAP). The interactions Cadmium and Tin (PSP > 0.99) and Cobalt and Manganese (PSP = 0.98) have high posterior probability of being synergistic. On the other hand, the interactions Molybdenum and Tin (PAP > 0.99) and Cadmium and Manganese (PAP = 0.99) have high posterior probabilities of being antagonistic. The interactions typically demonstrate flat behavior for most of the exposure domain and synergy/antagonism for the rest of the

domain. Nonlinear surfaces of this kind cannot be captured by quadratic regression approaches. We believe this to be a possible reason behind MixSelect being unable to detect these interactions. In Figure 3.5, we plot the interaction surfaces of Cadmium and Tin, Cobalt and Manganese, and Cadmium and Manganese. The plots for the interaction between Molybdenum and Tin can be found in Section B.8 of the Supplementary Material. Each row in Figure 3.5, from left to right, shows the pointwise 2.5% quantile, mean, and pointwise 97.5% quantile of the posterior samples of the pairwise interaction, evaluated on a 30×30 regular grid of points across the exposure values.

Regarding the covariate effects, we detected a negative association with age with an estimated coefficient -0.10 and 95% CI given by $[-0.12, -0.08]$. As expected, females had significantly lower urine creatinine levels than males, with the estimated coefficient of sex being -0.25 with 95% CI given by $[-0.28, -0.22]$. In addition, average creatinine levels increased with BMI, with the coefficient on BMI estimated as 0.11 with 95% CI of $[0.10, 0.13]$. The ethnicity category “Non-Hispanic Black” was found to have higher log dilution-adjusted urine creatinine concentrations than that of the baseline category “Non-Hispanic White”, with an estimated coefficient of 0.17 and a 95% CI $[0.13, 0.21]$. The ethnicity categories “Mexican American”, “Other Hispanic”, and “Other” were found to have lower log dilution-adjusted urine creatinine concentrations than that of “Non-Hispanic White”; their estimated coefficients are -0.12 with 95% CI given by $[-0.16, -0.08]$, -0.05 with 95% CI given by $[-0.10, -0.01]$, and -0.11 with a 95% CI given by $[-0.16, -0.07]$, respectively. Similar observations have been made previously in literature; refer to James et al. (1988).

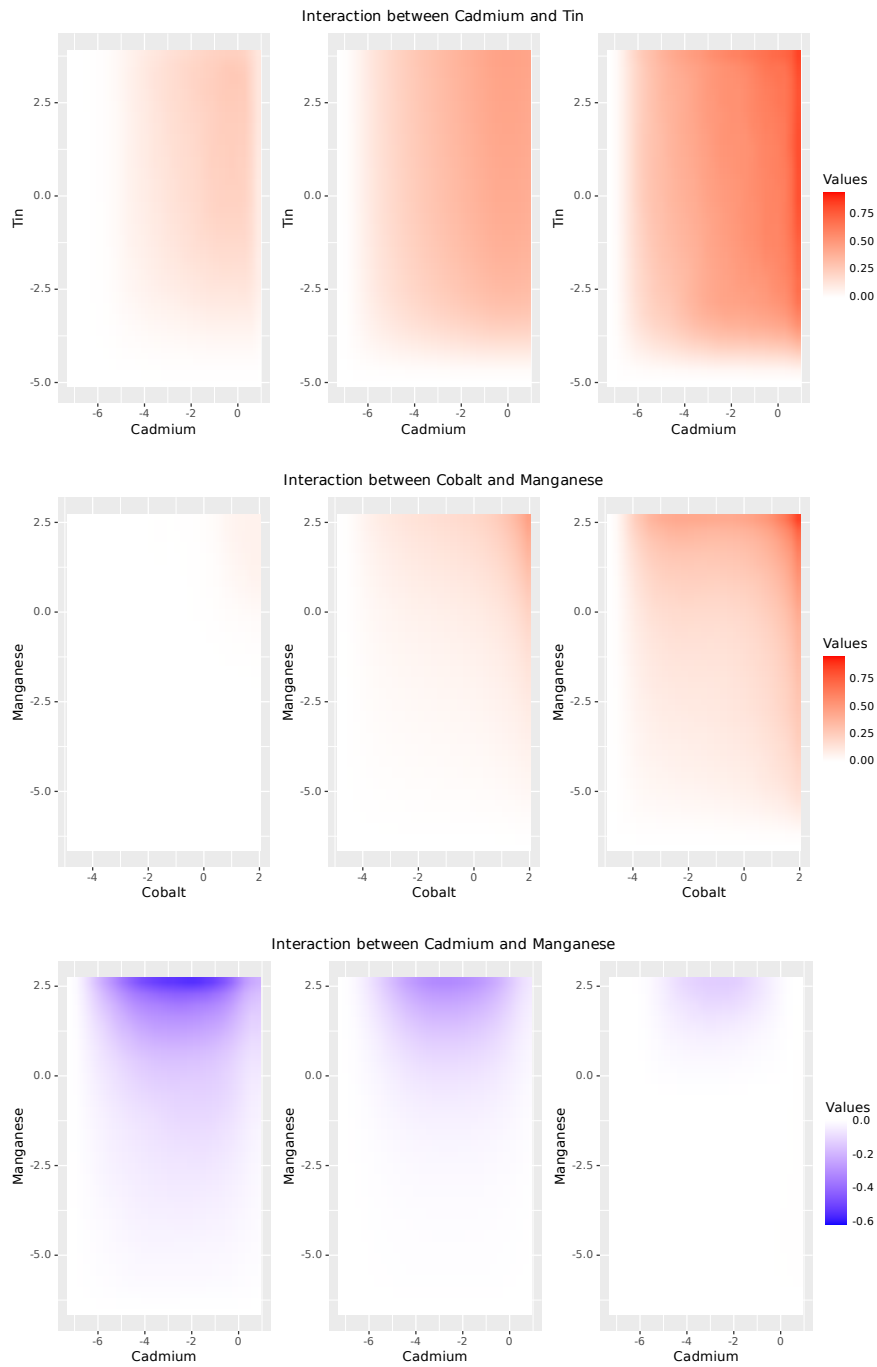


FIGURE 3.5: Plots showing the interaction effects of dilution-adjusted Cadmium and Tin, Cobalt and Manganese, and Cadmium and Manganese on dilution-adjusted log Creatinine. Exposure levels are in log scale. Each plot shows the pointwise 2.5% posterior credible surface, the posterior mean, and the 97.5% posterior credible surface from left to right.

Factor Analysis with Blessing of Dimensionality

4.1 Introduction

Inference on covariance in high-dimensional data is a key focus in many application areas, motivating a rich literature on associated statistical methods. One thread of this literature avoids modeling of the data and instead focuses on high-dimensional covariance matrix estimators under various assumptions on the inherent low-dimensional structure in the data, including (but not limited to) banded covariance (Bickel and Levina, 2008), low rank structure (Shikhaliyev et al., 2019), low rank with sparsity (Richard et al., 2012), sparse covariance (Bien and Tibshirani, 2011), and sparse inverse precision matrix estimation (Zhang and Zou, 2014). Our interest is instead in model-based Bayesian approaches, which have advantages in terms of ability to naturally accommodate complexities in the data and uncertainty quantification, while having disadvantages in terms of computational efficiency.

While there is a rich Bayesian literature on inference for high-dimensional covariance matrices, one of the most popular and routinely applied approaches is Bayesian factor analysis (Bhattacharya and Dunson, 2011; Lopes and West, 2004).

There continue to be regular developments improving upon and expanding the scope of Bayesian factor analysis methods (Schiavon et al., 2022; De Vito et al., 2021; Frühwirth-Schnatter, 2023; Roy et al., 2021; Ma and Liu, 2022; Bolfarine et al., 2022; Xie et al., 2022). Even with increasingly rich classes of priors and data types, the canonical approach for posterior computation remains Gibbs samplers that iterate between updating latent factors, factor loadings, residual variances, hyperparameters controlling the hierarchical prior, and other model parameters. This approach has the advantage of being simple to implement in broad model classes, but nonetheless commonly faces problems with slow mixing, particular as data dimensionality and complexity increase. Furthermore, when carrying out posterior inference with Gibbs samplers, there is a need to rotate the obtained samples to tackle the rotational ambiguity inherently woven into factor models and induce sparsity. Such ambiguity resulting in the non-identifiability of parameters can manifest as Markov chain Monte Carlo (MCMC) algorithms cycling between visiting multiple posterior modes and/or getting stuck at a mode, reducing the effective sample size (ESS) and thus the quality of posterior inference obtained from the algorithm. Although approaches such as varimax rotations in Rohe and Zeng (2020) or automatic rotations to induce sparsity as in Ročková and George (2016) resolve such ambiguity, the need for post-processing the obtained posterior samples introduces unnecessary complexity when inferring the covariance matrix and makes the overall process more cumbersome.

There are various strategies that have been developed to improve mixing of Gibbs samplers, including blocking, marginalization, and parameter expansion. However, computational hurdles remain, particularly as the number of dimensions increase. This has motivated a rich literature on scalable approaches for Bayesian inference of the covariance matrix in factor models, relying on variational approximations (Attias, 2013), maximum a posteriori estimation under sparsity priors (Srivastava et al., 2017), empirical Bayes approaches (Wang and Stephens, 2021), scalable spar-

sity inducing priors (Zhao et al., 2016), point estimation using the expectation-maximization (EM) algorithm (Avalos-Pacheco et al., 2022), and efficient posterior sampling to tackle non-identifiability (Man and Culpepper, 2022). Although greatly improving computational efficiency over the traditional Gibbs sampling algorithms, such approaches often provide little to no quantification of uncertainty in the covariance matrix, or provide further challenges to resolve non-identifiability of the latent factors and factor loadings. Furthermore, one may run into the same issues as getting stuck at local modes when optimizing the criterion of interest to obtain point estimates of the covariance matrix.

In general, fast algorithms for Bayesian factor analysis that are capable of scaling efficiently to high-dimensional data sacrifice the ability to provide an accurate characterization of uncertainty. The focus of this article is on proposing a simple approach for overcoming this limitation, providing a fast algorithm for high-dimensional Bayesian covariance matrix inference in factor analysis. With increasing dimensions, we obtain an increasing number of observations sharing a common latent factor. As a consequence, we obtain a blessing of dimensionality phenomenon allowing us to first pre-estimate the latent factors and then use this point estimate to obtain pseudo-posterior samples of the factor loadings and the residual variances by employing conjugate priors for the unknown parameters. This in turn induces a pseudo-posterior distribution on the covariance matrix and allows us to draw samples from this pseudo-posterior in a computationally efficient manner. The proposed Factor Analysis with BLEssing of dimensionality (FABLE) approach completely bypasses Markov chain Monte Carlo, instead providing an embarrassingly parallel framework to obtain pseudo-posterior samples of the high dimensional covariance matrix. The work of Fan et al. (2023) also considers pre-estimating the latent factors using principal components analysis (PCA), obtained from the observed high-dimensional covariates in a regression setting. In their work, high-dimensional regression is carried

out using latent factors to alleviate issues of multicollinearity. Our framework for estimating the latent factors is more general as it accounts for arbitrary rotation ambiguity, yielding the estimate of Fan et al. (2023) as a special case. The primary innovation of the proposed FABLE methodology is leveraging on the blessing of dimensionality phenomenon to obtain fast pseudo-posterior samples of the covariance matrix, with guarantees on both estimation accuracy and uncertainty quantification through the lens of asymptotic theory and simulation experiments. In contrast, Fan et al. (2023) considers estimating the latent factors to aid high-dimensional regression and relates joint latent factor models with sparse regression models.

In Section 4.2, we describe the FABLE methodology in detail, along with a discussion on choosing key hyperparameter values. In Section 4.3, we provide theoretical results demonstrating the blessing of dimensionality phenomenon, along with pseudo-posterior contraction rates and guarantees on uncertainty quantification when estimating the underlying data generating high-dimensional covariance matrix. In Section 4.4, we validate our approach from the viewpoint of estimation error and frequentist coverage on a wide variety of numerical experiments by comparisons with existing state-of-the-art approaches.

4.2 Proposed Methodology

4.2.1 Initial Approach

The observed data consists of $\mathbf{Y} = [y_1, \dots, y_n]^\top \in \mathbb{R}^{n \times p}$, where $y_i \in \mathbb{R}^p$ for $i = 1, \dots, n$. We consider the following latent factor model:

$$y_i = \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N_p(0, \Sigma), \quad (4.1)$$

where we have omitted the intercept term, assuming the data have been centered prior to analysis. Here, $\Lambda \in \mathbb{R}^{p \times k}$ denotes an unknown matrix of factor loadings, $\eta_i \stackrel{iid}{\sim} N_k(0, \mathbb{I}_k)$ denotes latent factors, $k \ll p$, and ϵ_i is a zero-mean error having

diagonal covariance $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$. Integrating out the latent factors provides the following marginal distribution of y_i for $i = 1, \dots, n$:

$$y_i \mid \Lambda, \Sigma \stackrel{iid}{\sim} N_p(0, \Lambda\Lambda^\top + \Sigma).$$

Therefore, the covariance $\Psi = \Lambda\Lambda^\top + \Sigma$ is decomposed as a sum of two parts; one low rank and the other diagonal. In this paper, our goal is to estimate Σ , $L = \Lambda\Lambda^\top$, and Ψ . We first illustrate our methodology assuming k is known. Later, we discuss an approach to estimate k in Section 4.2.3.

Let $\mathbf{M} = [\eta_1, \dots, \eta_n]^\top \in \mathbb{R}^{n \times k}$ and $\Lambda = [\lambda_1, \dots, \lambda_p]^\top \in \mathbb{R}^{p \times k}$, with λ_j^\top denoting the j th row of Λ . We also denote the j th column of \mathbf{Y} as $y^{(j)}$, so that $\mathbf{Y} = [y^{(1)}, \dots, y^{(p)}]$. The latent factor model (4.1) may be expressed in the following alternate way:

$$y^{(j)} = \mathbf{M}\lambda_j + \epsilon^{(j)}, \quad (4.2)$$

where $\epsilon^{(j)}$ is the j th column of the matrix $\mathbf{E} = [\epsilon_1, \dots, \epsilon_n]^\top = [\epsilon^{(1)}, \dots, \epsilon^{(p)}]$, where $\epsilon^{(j)} \stackrel{iid}{\sim} N_n(0, \sigma_j^2 \mathbb{I}_n)$. Writing in matrix form, we obtain

$$\mathbf{Y} = \mathbf{M}\Lambda^\top + \mathbf{E}. \quad (4.3)$$

In our approach, we obtain a proxy $\widehat{\mathbf{M}}$ for \mathbf{M} and substitute the proxy in (4.2) to effectively reduce the latent factor model to a parallel regression problem with $\widehat{\mathbf{M}}$ as the design matrix and λ_j the regression coefficient for the j th regression, $j = 1, \dots, p$.

To propose our choice for $\widehat{\mathbf{M}}$, we first begin with the singular value decomposition of \mathbf{Y} , given by

$$\mathbf{Y} = UDV^\top + U_\perp D_\perp V_\perp^\top, \quad (4.4)$$

where $U \in \mathbb{R}^{n \times k}$, $D \in \mathbb{R}^{k \times k}$, $V \in \mathbb{R}^{p \times k}$, $U_\perp \in \mathbb{R}^{n \times (r-k)}$, $D_\perp \in \mathbb{R}^{(r-k) \times (r-k)}$, and $V_\perp \in \mathbb{R}^{p \times (r-k)}$, with $r = p \wedge n$. The columns of U, U_\perp, V, V_\perp consist of orthonormal vectors, with $U^\top U_\perp = V^\top V_\perp = \mathbb{O}_{k \times (r-k)}$. D, D_\perp are diagonal matrices, with all the diagonal

entries of D strictly positive. Let us define

$$\mathbf{A} = \frac{\mathbf{YV}}{\sqrt{p}} = \frac{UD}{\sqrt{p}}.$$

Let $\widehat{C} \in \mathbb{R}^{k \times k}$ be an invertible matrix satisfying

$$\widehat{C}\widehat{C}^\top = \frac{1}{n}\mathbf{A}^\top\mathbf{A} = \frac{D^2}{np}. \quad (4.5)$$

There does exist at least one such \widehat{C} ; for example, $\widehat{C} = D/\sqrt{np}$ satisfies (4.5) since D is invertible. For any choice of \widehat{C} satisfying (4.5) such that $(\widehat{C})^{-1}$ exists, we define the estimator $\widehat{\mathbf{M}}$ of \mathbf{M} to be

$$\widehat{\mathbf{M}} = \mathbf{A}(\widehat{C}^\top)^{-1}. \quad (4.6)$$

Given a particular choice of $\widehat{\mathbf{M}}$ satisfying (4.6), we now consider the following surrogate regression model:

$$y^{(j)} = \widehat{\mathbf{M}}\tilde{\lambda}_j + \tilde{\epsilon}^{(j)}, \quad \tilde{\epsilon}^{(j)} \stackrel{ind}{\sim} N_n(0, \tilde{\sigma}_j^2 \mathbb{I}_n) \quad (4.7)$$

for $j = 1, \dots, p$. The model (4.7) could be interpreted as a version of (4.2) with $\widehat{\mathbf{M}}$ substituted for the original matrix of latent factors \mathbf{M} and new parameters $\tilde{\lambda}_j \in \mathbb{R}^k$, $\tilde{\sigma}_j^2 > 0$ for $j = 1, \dots, p$. Next, we endow $(\tilde{\lambda}_j, \tilde{\sigma}_j^2)_{j=1}^p$ with normal-inverse gamma (NIG) priors $(\tilde{\lambda}_j, \tilde{\sigma}_j^2) \stackrel{iid}{\sim} \text{NIG}(0_k, \tau^2 \mathbb{I}_k, \gamma_0/2, \gamma_0 \delta_0^2/2)$. That is, we let

$$\tilde{\lambda}_j \mid \tilde{\sigma}_j^2 \sim N_k(0, \tilde{\sigma}_j^2 \tau^2 \mathbb{I}_k), \quad \tilde{\sigma}_j^2 \sim \text{IG}\left(\frac{\gamma_0}{2}, \frac{\gamma_0 \delta_0^2}{2}\right). \quad (4.8)$$

The global shrinkage parameter τ^2 allows us to *a priori* shrink the factor loadings towards zero, regularizing $\tilde{\Lambda} = [\tilde{\lambda}_1, \dots, \tilde{\lambda}_p]^\top$ and favoring sparsity, which is commonly assumed in existing literature. We discuss a strategy to obtain a data-driven choice for τ^2 in Section 4.2.3.

The surrogate model (4.7) and the prior specification (4.8) motivate the pseudo-posterior densities $\tilde{\Pi}_j$ for $(\tilde{\lambda}_j, \tilde{\sigma}_j^2)_{j=1}^p$, given by

$$\begin{aligned}\tilde{\Pi}_j(\tilde{\lambda}_j, \tilde{\sigma}_j^2) &= \text{NIG}(\tilde{\lambda}_j, \tilde{\sigma}_j^2 \mid \mu_j, \mathbf{K}, \gamma_n/2, \gamma_n\delta_j^2/2) \\ &= N_k(\tilde{\lambda}_j \mid \mu_j, \tilde{\sigma}_j^2 \mathbf{K}) \text{IG}\left(\tilde{\sigma}_j^2 \mid \frac{\gamma_n}{2}, \frac{\gamma_n\delta_j^2}{2}\right),\end{aligned}\tag{4.9}$$

with the updated hyperparameters given by

$$\begin{aligned}\mu_j &= \left(\widehat{\mathbf{M}}^\top \widehat{\mathbf{M}} + \frac{\mathbb{I}_k}{\tau^2}\right)^{-1} \widehat{\mathbf{M}}^\top \mathbf{y}^{(j)}, \\ \mathbf{K} &= \left(\widehat{\mathbf{M}}^\top \widehat{\mathbf{M}} + \frac{\mathbb{I}_k}{\tau^2}\right)^{-1}, \\ \gamma_n &= \gamma_0 + n,\end{aligned}\tag{4.10}$$

$$\gamma_n\delta_j^2 = \gamma_0\delta_0^2 + (\mathbf{y}^{(j)\top} \mathbf{y}^{(j)} - \mu_j^\top \mathbf{K}^{-1} \mu_j) = \gamma_0\delta_0^2 + \mathbf{y}^{(j)\top} \left(\mathbb{I}_n - \frac{\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top}{n + \tau^{-2}}\right) \mathbf{y}^{(j)}.$$

The pseudo-posterior for $(\tilde{\lambda}_j, \tilde{\sigma}_j^2)$ for $j = 1, \dots, p$ is motivated by applying Bayes' rule on the j th regression 4.7 and conditioning on a fixed $\widehat{\mathbf{M}}$. To obtain pseudo-posterior samples of $(\tilde{\lambda}_j, \tilde{\sigma}_j^2)$, we simply draw independent samples $(\tilde{\lambda}_j, \tilde{\sigma}_j^2) \stackrel{\text{ind}}{\sim} \tilde{\Pi}_j$ for $j = 1, \dots, p$, and let $\tilde{\Lambda} = [\tilde{\lambda}_1, \dots, \tilde{\lambda}_p]^\top$, $\tilde{L} = \tilde{\Lambda}\tilde{\Lambda}^\top$, $\tilde{\Sigma} = \text{diag}(\tilde{\sigma}_1^2, \dots, \tilde{\sigma}_p^2)$, and $\tilde{\Psi} = \tilde{L} + \tilde{\Sigma}$ denote the pseudo-posterior samples of Λ , L , Σ , and Ψ , respectively. Our proposed Factor Analysis with BLEssing of dimensionality (FABLE) approach completely bypasses Markov chain Monte Carlo (MCMC) by obtaining independent samples from the pseudo-posterior of the relevant quantities in an embarrassingly parallel fashion.

Although we observed good performance of $\tilde{\Psi}$ when estimating Ψ both in terms of simulations and posterior contraction rates, the entry-wise pseudo-posterior credible intervals of $\tilde{\Psi}$ underestimated the uncertainty when estimating the entry-wise

elements of Ψ , in terms of frequentist coverage. In Section 4.2.2, we build on the initial FABLE approach developed in this subsection and provide a Coverage-Corrected version of FABLE (CC-FABLE) that provides good uncertainty quantification along with good estimation when estimating the entrywise components of Ψ .

We now consider the effect of choosing $\widehat{\mathbf{M}}$ on the FABLE procedure. A natural concern is whether choosing a different $\widehat{\mathbf{M}}$ would affect the pseudo-posterior distribution of $\tilde{\Psi}$. The answer is no, as obtained from

Proposition 6. (i) For any $\widehat{\mathbf{M}}$ satisfying (4.6), we have $\widehat{\mathbf{M}}^\top \widehat{\mathbf{M}} = n\mathbb{I}_k$ and $\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top = nUU^\top$. (ii) The pseudo-posterior distribution of \tilde{L} , $\tilde{\Sigma}$, and $\tilde{\Psi}$ obtained from the FABLE approach only depends on $\widehat{\mathbf{M}}$ through the two quantities $\widehat{\mathbf{M}}^\top \widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top$.

Proof: For (i), we simply observe that

$$\widehat{\mathbf{M}}^\top \widehat{\mathbf{M}} = \widehat{C}^{-1} \mathbf{A}^\top \mathbf{A} (\widehat{C}^\top)^{-1} = n\widehat{C}^{-1} \widehat{C} \widehat{C}^\top (\widehat{C}^\top)^{-1} = n\mathbb{I}_k.$$

Similarly,

$$\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top = \mathbf{A} (\widehat{C} \widehat{C}^\top)^{-1} \mathbf{A}^\top = (np/p) U D D^{-2} D U^\top = nUU^\top.$$

Next, (ii) is immediate from observing that the distribution of the entries of \tilde{L} and $\tilde{\Sigma}$ depend on $\mu_j^\top \mu_{j'}$ and δ_l^2 , for $1 \leq j, j' \leq p$ and $1 \leq l \leq p$. Since both of these quantities only depend on $\widehat{\mathbf{M}}$ through $\widehat{\mathbf{M}}^\top \widehat{\mathbf{M}}$ and $\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top$, we have proven the claim.

Thus, the particular choice of $\widehat{\mathbf{M}}$ does not affect the pseudo-posterior samples and therefore our inference procedure. We now describe the heuristic behind the choice of $\widehat{\mathbf{M}}$ in (4.6). Starting from the SVD of \mathbf{Y} , we have

$$\mathbf{Y} = U D V^\top + U_\perp D_\perp V_\perp^\top.$$

From (4.3), it is immediate that the matrix $\mathbf{A} = \mathbf{Y}V/\sqrt{p}$ satisfies

$$\mathbf{A} = \mathbf{M} \frac{\Lambda^\top V}{\sqrt{p}} + \frac{\mathbf{E}V}{\sqrt{p}}.$$

Let $C = V^\top \Lambda / \sqrt{p} \in \mathbb{R}^{k \times k}$. Based on the consistency of spectral estimates, we expect \mathbf{E} to be approximately independent of V as both n, p grow. As a result, we expect $\mathbf{E}V/\sqrt{p} \approx 0$ for increasing p . This leads us to

$$\mathbf{A} \approx \mathbf{M}C^\top \tag{4.11}$$

or equivalently, $a_i \approx C\eta_i$ for $i = 1, \dots, n$. Since $\eta_i \sim N_k(0, \mathbb{I}_k)$, the marginal density of a_i for $i = 1, \dots, n$ is approximately $a_i \stackrel{iid}{\sim} N_k(0, CC^\top)$. This motivates the following estimator $\hat{C}\hat{C}^\top$ of CC^\top :

$$\hat{C}\hat{C}^\top = \frac{1}{n}A^\top A = \frac{D^2}{np}.$$

Given any \hat{C} satisfying (4.5), we use (4.11) to propose $\hat{\mathbf{M}} = \mathbf{A}(\hat{C}^\top)^{-1}$ as a surrogate for \mathbf{M} . An immediately available choice is given by letting $\hat{C} = D/\sqrt{np}$, implying $\hat{\mathbf{M}} = \sqrt{n}U$. We shall refer to this as the canonical choice of $\hat{\mathbf{M}}$, also seen to be the spectral estimate of \mathbf{M} . As described in Fan et al. (2023), the canonical estimator $\sqrt{n}U$ is seen to be the same as the estimator obtained by carrying out principal components analysis (PCA) on the matrix YY^\top . To obtain the PCA estimate of \mathbf{M} , one solves the optimization problem

$$\arg \min_{F, B} \|Y - FB^\top\|_{\mathbf{F}}^2,$$

subject to $F^\top F/n = \mathbb{I}_k$ and $B^\top B$ is diagonal, where $\|Q\|_{\mathbf{F}} = \sqrt{\text{tr}(Q^\top Q)}$ denotes the Frobenius norm of Q . The resulting estimate \hat{F} is such that the columns of \hat{F}/\sqrt{n} are the eigenvectors corresponding to the k -largest eigenvalues of YY^\top , or equivalently, the left singular vectors corresponding to the k -largest singular values of Y , leading us to $\hat{F} = \sqrt{n}U$. However, the estimate of the latent factors $\hat{\mathbf{M}}$ as in (4.6) also allows choices other than the one obtained from PCA, providing a general framework for obtaining estimates of the latent factors under different rotations. In Section 4.3, we

demonstrate the validity of approximating \mathbf{M} using $\widehat{\mathbf{M}}$ as both $n, p \rightarrow \infty$, providing a blessing of dimensions when estimating the high-dimensional covariance matrix Ψ .

4.2.2 Coverage Correction

The pseudo-posterior updates as in (4.9) provide a natural approach to obtain pseudo-posterior samples $\tilde{\Psi}$ of the high-dimensional covariance matrix Ψ of y_1, \dots, y_n . One could then consider measures such as the entry-wise posterior mean or the entry-wise quantiles of the sampled $\tilde{\Psi}_{ij}$ to provide summaries for the (i, j) entry Ψ_{ij} of Ψ . Although the pseudo-posterior mean of $\tilde{\Psi}_{ij}$ obtained from (4.9) was observed to be a good estimator of Ψ_{ij} across numerous simulation studies that we carried out, we observed under-coverage of the entry-wise pseudo-posterior credible intervals of $\tilde{\Psi}_{ij}$ in terms of containing Ψ_{ij} , particularly when the signal-to-noise ratios $R_j = \|\lambda_j\|_2^2/\sigma_j^2$ for $j = 1, \dots, p$ are larger in magnitude. To counter underestimation of uncertainty of the pseudo-credible intervals, we now provide a coverage-corrected version of FABLE (CC-FABLE), improving the frequentist coverage of the entrywise pseudo-credible intervals of $\tilde{\Psi}_{ij}$ obtained from the FABLE methodology introduced in Section 4.2.1.

Let G be the matrix with uv th entry given by $G_{uv} = \mu_u^\top \mu_v$ for $1 \leq u, v \leq p$ and let $s_j^2 = \|(\mathbb{I}_n - UU^\top)y^{(j)}\|_2^2/n$ for $j = 1, \dots, p$. Suppose we have obtained pseudo-posterior samples $(\tilde{L}, \tilde{\Sigma})$ from the initial approach in Section 4.2.1. We then define a coverage-corrected pseudo-posterior sample of (L, Σ) to be (L_C, Σ_C) , where

$$L_C = G + B \odot (\tilde{L} - G), \quad (4.12)$$

$$\Sigma_C = \tilde{\Sigma}, \quad (4.13)$$

where $B = (b_{uv})_{1 \leq u, v \leq p}$ with b_{uv} as

$$\begin{aligned} b_{uv} &= \left(1 + \frac{\|\mu_u\|_2^2 \|\mu_v\|_2^2 + (\mu_u^\top \mu_v)^2}{s_u^2 \|\mu_v\|_2^2 + s_v^2 \|\mu_u\|_2^2} \right)^{1/2}, \quad \text{if } u \neq v, \\ &= \left(1 + \frac{\|\mu_u\|_2^2}{2s_u^2} \right)^{1/2}, \quad \text{if } u = v, \\ &= 1, \quad \text{otherwise,} \end{aligned} \tag{4.14}$$

and $A_1 \odot A_2$ denoting the Hadamard product of two matrices A_1 and A_2 . Finally, we let a coverage-corrected pseudo-posterior sample of Ψ be Ψ_C , where

$$\Psi_C = L_C + \Sigma_C. \tag{4.15}$$

Re-scaling of $(\tilde{L} - G)$ by the scaling matrix B provides the correct entry-wise asymptotic pseudo-posterior variance of the coverage-corrected factor loading matrix L_C . As we see in Section 4.3, this leads to a Bernstein-von Mises (BvM) type result implying that the entry-wise credible intervals obtained from the coverage-corrected samples of L_C will have the nominal asymptotic frequentist coverage. In our simulations, we observed this scaling to work well for finite data as well, with the proposed CC-FABLE greatly improving entrywise under-coverage of the FABLE approach.

4.2.3 Hyperparameter Choice

Before generating the pseudo-posterior samples of the covariance matrix, one needs to carefully choose two key hyperparameters, namely, the number of factors k and the global variance of the factor loadings τ^2 . We now describe a data-driven approach to choose k and τ^2 . To choose the number of factors k , we start out by looking at the singular values of the matrix $\mathbf{Y} = [y_1, \dots, y_n]^\top$, denoted by $s_1 \geq \dots \geq s_{n \wedge p}$. Provided the signal-to-noise ratio is sufficiently large, we expect a sharp decrease in the magnitude of the singular values after the first k from a spectral perspective, since

$s_{k+1}, \dots, s_{n \wedge p}$ are associated with noise. This guides us to our heuristic estimate

$$\hat{k} = \operatorname{argmax}_{1 \leq j \leq n \wedge p - 1} (s_j - s_{j+1}). \quad (4.16)$$

Estimating k using \hat{k} performs reasonably well in an extensive number of simulations that we carried out. However, this approach is often insufficient when the signal-to-noise ratio in the model is too low, which typically signifies the spectrum of \mathbf{Y} being very close together. As a rough guideline to the practitioner, we recommend first investigating the quantity

$$\Delta = \frac{1}{s_1} \left[\max_{1 \leq j \leq n \wedge p - 1} (s_j - s_{j+1}) \right] \in [0, 1],$$

and identifying a scenario as having low signal-to-noise ratio if $\Delta \leq 0.1$. In such scenarios, we recommend proceeding with an overestimated value of k ; for example, one could proceed with $k^* = \hat{k} + 10$, where \hat{k} is the estimate obtained from (4.16). In our simulations, using an overestimated value of k performed quite well, with the performance decaying the further away our guess of k is from the true value of k .

Once we have chosen a value for the number of factors k , we proceed to choose a value for the common variance parameter τ^2 , which can also be viewed as a global shrinkage factor, shrinking the factor loadings towards 0. To do this, we employ an approximate empirical Bayesian approach. Since

$$\tilde{\lambda}_j \mid \tilde{\sigma}_j^2 \sim N_k(0, \tau^2 \tilde{\sigma}_j^2 \mathbb{I}_k),$$

we obtain an estimate of τ^2 by conditioning on $(\tilde{\lambda}_j, \tilde{\sigma}_j^2)$ and then maximizing the conditional likelihood, with the optimal value given by

$$\hat{\tau}^2 = \frac{1}{kp} \sum_{j=1}^p \frac{\|\tilde{\lambda}_j\|_2^2}{\tilde{\sigma}_j^2}.$$

We now use plug-in replacements of $\|\tilde{\lambda}_j\|_2^2$ and $\tilde{\sigma}_j^2$ as follows. Following from the pseudo-posterior updates (4.9) and (4.10), we simply replace $\|\tilde{\lambda}_j\|_2^2$ by \mathcal{L}_j^2 and $\tilde{\sigma}_j^2$ by s_j^2 , where

$$\begin{aligned}\mathcal{L}_j^2 &= \frac{1}{n} \|U^\top y^{(j)}\|_2^2, \\ s_j^2 &= \frac{1}{n} \|(\mathbb{I}_n - UU^\top)y^{(j)}\|_2^2.\end{aligned}$$

This leads us to the plug-in estimate of τ^2 given by

$$\hat{\tau}^2 = \frac{1}{kp} \sum_{j=1}^p \frac{\mathcal{L}_j^2}{s_j^2}. \quad (4.17)$$

We found this estimate of τ^2 to perform very well across a number of simulations that were investigated. Once the number of factors k and the common variance τ^2 have been estimated, we now proceed to obtain pseudo-posterior samples.

4.2.4 Final Algorithm

We now provide an algorithm for implementing FABLE or CC-FABLE on a given data set to obtain pseudo-posterior samples of the high-dimensional covariance matrix modeled using factor analysis. We first tune the relevant hyperparameters k and τ^2 as in Section 4.2.3, obtain initial pseudo-posterior samples of the covariance matrix as in Section 4.2.1, and correct for under-coverage as in Section 4.2.2 when obtaining entrywise interval estimates of the underlying covariance matrix. The relevant steps are described in Algorithm 2.

Algorithm 2. *Steps to obtain pseudo-posterior samples of the covariance matrix using CC-FABLE.*

Input: The data matrix $\mathbf{Y} \in \mathbf{R}^{n \times p}$, number of Monte Carlo (MC) samples N , and the error variance hyperparameters (γ_0, σ_0^2) . Let $r = n \wedge p$.

Step 1: Carry out the singular value decomposition of $\mathbf{Y} = U^* D^* V^{*\top}$ with $U \in \mathbf{R}^{n \times r}$, $D \in \mathbf{R}^{r \times r}$, $V \in \mathbf{R}^{p \times r}$; suppose the singular values are $s_1 \geq \dots \geq s_r \geq 0$.

Step 2: Estimate the number of factors k using

$$\hat{k} = \arg \max_{1 \leq j \leq r} (s_j - s_{j+1}).$$

Step 3: Let U consist of the columns of U^* corresponding to the \hat{k} largest singular values. For $1 \leq j \leq p$, let $y^{(j)}$ denote the j th column of \mathbf{Y} , and obtain

$$\begin{aligned} \mathcal{L}_j^2 &= \frac{1}{n} \|U^\top y^{(j)}\|_2^2 \\ s_j^2 &= \frac{1}{n} \|(\mathbb{I}_n - UU^\top)Y^{(j)}\|_2^2. \end{aligned}$$

Step 4: Estimate τ^2 by

$$\hat{\tau}^2 = \frac{1}{kp} \sum_{j=1}^p \frac{\mathcal{L}_j^2}{s_j^2}.$$

Step 5: For $1 \leq j \leq p$, let

$$\mu_j = \frac{\sqrt{n}}{n + \hat{\tau}^{-2}} U^\top y^{(j)}.$$

Let $B = (b_{uv})_{1 \leq u \leq v \leq p}$, with

$$\begin{aligned} b_{uv} &= \left(1 + \frac{\|\mu_u\|_2^2 \|\mu_v\|_2^2 + (\mu_u^\top \mu_v)^2}{s_u^2 \|\mu_v\|_2^2 + s_v^2 \|\mu_u\|_2^2} \right)^{1/2}, \quad \text{if } u \neq v, \\ &= \left(1 + \frac{\|\mu_u\|_2^2}{2s_u^2} \right)^{1/2}, \quad \text{if } u = v, \\ &= 1, \quad \text{otherwise,} \end{aligned}$$

Step 6: Let $\gamma_n = \gamma_0 + n$ and for $j = 1, \dots, p$, evaluate

$$\gamma_n \delta_j^2 = \gamma_0 \delta_0^2 + y^{(j)\top} \left(\mathbb{I}_n - \frac{nUU^\top}{n + \hat{\tau}^{-2}} \right) y^{(j)}.$$

Step 7: For each $t = 1, \dots, N$, independently sample $(\tilde{\lambda}_{j,t}, \tilde{\sigma}_{j,t}^2)$ across $j = 1, \dots, p$, such that

$$\begin{aligned} \tilde{\sigma}_{j,t}^2 &\sim IG\left(\frac{\gamma_n}{2}, \frac{\gamma_n \delta_j^2}{2}\right), \\ \tilde{\lambda}_{j,t} \mid \tilde{\sigma}_{j,t}^2 &\sim N_k\left(\mu_j, \frac{\tilde{\sigma}_{j,t}^2}{n + \hat{\tau}^{-2}} \mathbb{I}_k\right). \end{aligned}$$

Form $\tilde{\Lambda}_t = [\tilde{\lambda}_{1,t}, \dots, \tilde{\lambda}_{p,t}]^\top$ and $\tilde{\Sigma}_t = \text{diag}(\tilde{\sigma}_{1,t}^2, \dots, \tilde{\sigma}_{p,t}^2)$, which denote the t -th Monte Carlo (MC) samples of Λ and Σ , respectively.

Step 8: Let $G = [\mu_1, \dots, \mu_p]^\top$. For each $t = 1, \dots, N$, compute the coverage-corrected samples as

$$\begin{aligned} L_{C,t} &= G + B \odot (\tilde{\Lambda}_t \tilde{\Lambda}_t^\top - G), \\ \Sigma_{C,t} &= \tilde{\Sigma}_t, \\ \Psi_{C,t} &= L_{C,t} + \Sigma_{C,t}. \end{aligned}$$

Output: The N MC samples of the covariance matrix $\Psi_{C,1}, \dots, \Psi_{C,N}$.

4.3 Theoretical Support

In this Section, we turn our attention to providing theoretical guarantees of the proposed FABLE procedure in estimating the high-dimensional covariance matrix. Most of the existing literature (Bhattacharya and Dunson, 2011; Pati et al., 2014) on obtaining provable guarantees for Bayesian factor models utilize the machinery of

Bayes' rule in deriving posterior contraction rates. In contrast, the proposed method provides pseudo-posterior samples of the underlying covariance matrix, and as a result, we are not directly able to utilize the Bayesian framework to obtain theoretical guarantees. However, we overcome this challenge by leveraging on a blessing of dimensionality phenomenon which requires both n and p to grow, providing results on both the pseudo-posterior contraction and uncertainty quantification of our approach. The blessing of dimensionality phenomenon allows accurate estimation of the latent factor subspace up to rotational ambiguity, which serves as an important component behind the theoretical results. The proofs of all the results can be found in Appendix C.

We assume that the data are generated from the following data generating model:

$$y_i = \Lambda_0 \eta_{0i} + \epsilon_i, \quad (4.18)$$

where $\epsilon_i \stackrel{iid}{\sim} N_p(0, \Sigma_0)$ for $i = 1, \dots, n$ with $\Sigma_0 = \text{diag}(\sigma_{01}^2, \dots, \sigma_{0p}^2)$, $\eta_{0i} \stackrel{iid}{\sim} N_k(0, \mathbb{I}_k)$ for $i = 1, \dots, n$, and Λ_0 is the true matrix of factor loadings. η_{0i} are the true latent factors; integrating them out provides us $y_i \stackrel{iid}{\sim} N_p(0, \Lambda_0 \Lambda_0^\top + \Sigma_0)$ for $i = 1, \dots, n$ as the marginal distribution of the data. Let $\mathbf{Y} = [y_1, \dots, y_n]^\top \in \mathbb{R}^{n \times p}$ and $M_0 = [\eta_{01}, \dots, \eta_{0n}]^\top \in \mathbb{R}^{n \times k}$ be the data matrix and the true matrix of latent factors, respectively, so that the true data generating model may be written as

$$\mathbf{Y} = M_0 \Lambda_0^\top + \mathbf{E}, \quad (4.19)$$

where $\mathbf{E} = [\epsilon_1, \dots, \epsilon_n]^\top$. Our primary goal is the estimation of the covariance matrix $\Psi_0 = \Lambda_0 \Lambda_0^\top + \Sigma_0$.

For a matrix $A \in \mathbb{R}^{n_1 \times n_2}$, suppose the singular values of A are given by $s_1(A) \geq \dots \geq s_{n_1 \wedge n_2}(A)$; let $\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2 = s_1(A)$ denote the operator norm of A .

For two sequences a_m, b_m , we say $a_m \asymp b_m$ if $|(a_m/b_m) - 1| \rightarrow 0$ as $m \rightarrow \infty$. For

our theoretical requirements, we will assume that k and τ^2 are known and fixed. Furthermore, we assume the following conditions on the true data generating model:

Assumption 4. Λ_0 satisfies $s_k(\Lambda_0) \asymp \|\Lambda_0\| \asymp \sqrt{p}$ and $\|\Lambda_0\|_\infty < \infty$.

Assumption 5. The true residual variances satisfy

$$\max_{1 \leq j \leq p} \sigma_{0j}^2 = O(1), \quad \min_{1 \leq j \leq p} \sigma_{0j}^2 > 0.$$

Assumption 6. The hyperparameters $k, \tau^2, \gamma_0, \delta_0^2$ are fixed constants.

Assumption 7. The number of dimensions p increases satisfies $\log p = o(na_n^2)$ for any a_n such that $a_n = o(1), na_n^2 \rightarrow \infty$ as $n \rightarrow \infty$.

Such assumptions are standard in existing literature on asymptotic theoretical properties of latent factor models. Assumption 4 ensures that the true loadings matrix Λ_0 is well-conditioned and the low-rank portion $\Lambda_0 \Lambda_0^\top$ can be identified from noise in the asymptotic regime. Assumptions 5 and 6 simply assume the scalar error variances and model hyperparameters are finite. Assumption 7 allows the number of dimensions p to increase at any polynomial rate for any polynomially decaying choice of a_n . In particular, $a_n = \log n / \sqrt{n}$ satisfies Assumption 7. Asymptotically larger choices of a_n allow p to grow faster, at the cost of providing slower rates of convergence when estimating Σ_0 .

Let $M_0 \Lambda_0^\top = U_0 D_0 V_0^\top$ be the singular value decomposition of the signal, with $U_0 \in \mathbb{R}^{n \times k}, V_0 \in \mathbb{R}^{p \times k}$ having orthonormal columns and D_0 a diagonal matrix of positive singular values. Suppose the singular value decomposition of \mathbf{Y} is given by

$$\mathbf{Y} = U D V^\top + U_\perp D_\perp V_\perp^\top,$$

where $U \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{p \times k}$ have orthonormal columns, and $D \in \mathbb{R}^{k \times k}$ contain the k largest singular values of \mathbf{Y} . We first provide a result showcasing the blessing of

dimensionality when estimating U_0 by U , which forms a key part of the results that follow. Let us denote the induced pseudo-posterior measure, the true data generating measure, and the expectation under the true data generating measure by $\tilde{\Pi}$, P_0 , and E_0 , respectively.

Proposition 7. *Suppose Assumptions 4 - 7 hold. Then, there exists a constant $G_1 > 0$ such that*

$$\lim_{n,p \rightarrow \infty} P_0 \left\{ \|UU^\top - U_0U_0^\top\| > G_1 \left(\frac{1}{n} + \frac{1}{p} \right) \right\} = 0.$$

Recall the definition of $\tilde{\Lambda}$, $\tilde{\Sigma}$, and $\tilde{\Psi}$ from Section 4.2.1. First, in Theorem 8, we provide pseudo-posterior contraction rates when estimating $\Lambda_0\Lambda_0^\top$, Σ_0 , and Ψ_0 using $\Lambda_C\Lambda_C^\top$, Σ_C , and $\Psi_C = \Lambda_C\Lambda_C^\top + \Sigma_C$, respectively, obtained from the CC-FABLE methodology in Section 4.2.2. Later, using Theorems 10 and 11, we provide a justification on why quantifying uncertainty of the entrywise elements of Ψ_0 via pseudo-posterior credible intervals obtained CC-FABLE is valid, in the sense that $100(1 - \alpha)\%$ pseudo-posterior credible intervals are also $100(1 - \alpha)\%$ confidence intervals asymptotically.

Theorem 8. *Suppose Assumptions 4 - 7 hold. Then, as $n, p \rightarrow \infty$,*

(a) *There exists a constant $C_1 > 0$ such that*

$$E_0 \left[\tilde{\Pi} \left\{ \frac{\|\Lambda_C\Lambda_C^\top - \Lambda_0\Lambda_0^\top\|}{\|\Lambda_0\Lambda_0^\top\|} > C_1 \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}} \right) \right\} \right] \rightarrow 0.$$

(b) *There exists a constant $C_2 > 0$ such that*

$$E_0 \left[\tilde{\Pi} \left\{ \|\Sigma_C - \Sigma_0\| > C_2 \left(a_n + \frac{1}{p} \right) \right\} \right] \rightarrow 0.$$

(c) *There exists a constant $C > 0$ such that*

$$E_0 \left[\tilde{\Pi} \left\{ \frac{\|\Psi_C - \Psi_0\|}{\|\Psi_0\|} > C \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}} \right) \right\} \right] \rightarrow 0.$$

Theorem 8 shows pseudo-posterior concentration of the relevant quantities of CC-FABLE procedure. The relative error when estimating Ψ_0 converges to 0 at the rate $n^{-1/2} + p^{-1/2}$. This showcases the blessing of dimensionality, with pseudo-posterior concentration taking place when both the number of samples n and the number of dimensions p increase.

We now consider the problem of quantifying the uncertainty when estimating the entrywise elements of the covariance matrix. For $1 \leq u \leq v \leq p$, let $\Psi_{C,uv}$ and $\Psi_{0,uv}$ denote the uv th element of Ψ_C and Ψ_0 , respectively, where Ψ_C denotes the coverage-corrected pseudo-posterior sample obtained from the CC-FABLE discussed in Section 4.2.2. It is immediate that $\Psi_{0,uv} = \lambda_{0u}^\top \lambda_{0v} + \sigma_{0u}^2 \mathbb{1}(u = v)$ by $\tilde{\Psi}_{uv}$ for $1 \leq u \leq v \leq p$. To exercise finer control over bounding entrywise terms, we require a stronger version of the blessing of dimensionality result in Lemma 7 when estimating U_0 by U . This is obtained by providing an upper bound to the max norm of $UU^\top - U_0U_0^\top$. For any matrix A , the max norm of A is defined as $\|A\|_\infty = \max_{ij} |A_{ij}|$, where A_{ij} is the ij th entry of A .

Proposition 9. *Suppose Assumptions 4 – 7 hold. Then, there exists a constant $G_2 > 0$ such that*

$$\lim_{n,p \rightarrow \infty} P_0 \left\{ \|UU^\top - U_0U_0^\top\|_\infty > G_2 \left(\frac{1}{n} + \sqrt{\frac{\log(n+p)}{np}} \right) \right\} = 0.$$

Let $T_{uv} = \mu_u^\top \mu_v + \delta_u^2 \mathbb{1}(u = v)$ be an estimator of $\Psi_{0,uv}$. Then, we have the following result approximating the pseudo-posterior distribution of $\Psi_{C,uv}$ by a suitable Gaussian distribution as both n, p increase.

Theorem 10. *Suppose Assumptions 4-7 hold. Furthermore, assume that $\log n = o(p)$. For $1 \leq u \leq v \leq p$, let*

$$\begin{aligned} \mathcal{S}_{0,uv}^2 &= \sigma_{0v}^2 \|\lambda_{0u}\|_2^2 + \sigma_{0u}^2 \|\lambda_{0v}\|_2^2 + \|\lambda_{0u}\|_2^2 \|\lambda_{0v}\|_2^2 + (\lambda_{0u}^\top \lambda_{0v})^2, \quad \text{for } u \neq v, \\ &= 2(\|\lambda_{0u}\|_2^2 + \sigma_{0u}^2)^2, \quad \text{for } u = v. \end{aligned}$$

Then, as $n, p \rightarrow \infty$,

$$\sup_{x \in \mathbb{R}} \left| \tilde{\Pi} \left\{ \frac{\sqrt{n}(\Psi_{C,uv} - T_{uv})}{\mathcal{S}_{0,uv}} \leq x \right\} - \Phi(x) \right| \xrightarrow{P_0} 0.$$

Theorem 10 allows us to approximate the pseudo-posterior distribution of each element of the covariance matrix using a Gaussian with suitably chosen mean T_{uv} and variance $\mathcal{S}_{0,uv}^2$ asymptotically. In what follows, we first state a result regarding the asymptotic law of $\sqrt{n}(T_{uv} - \Psi_{0,uv})$ and then demonstrate how this result can be used to conclude that the entrywise $100(1 - \alpha)\%$ pseudo-posterior credible intervals also serve as valid $100(1 - \alpha)\%$ confidence intervals for any $0 < \alpha < 1$, allowing us to obtain nominal frequentist coverage of the pseudo-posterior intervals. For general random variables X_n and X , we denote X_n converging in distribution to X by $X_n \implies X$.

Theorem 11. *Suppose the Assumptions of Theorem 10 hold. Then, as $n, p \rightarrow \infty$, one has*

$$\sqrt{n}(T_{uv} - \Psi_{0,uv}) \implies N(0, \mathcal{S}_{0,uv}^2).$$

Theorems 10 and 11 together imply that the pseudo-posterior variance of the entrywise elements of the coverage-corrected covariance matrix and the variance of the estimator T_{uv} under repeated sampling agree in the asymptotic limit. Under Theorem 10, the $100(1 - \alpha)\%$ pseudo-posterior credible interval of $\Psi_{C,uv}$ may be asymptotically approximated by the interval

$$\mathcal{C}_{uv} = \left[T_{uv} \mp z_\alpha \frac{\mathcal{S}_{0,uv}}{\sqrt{n}} \right],$$

where $z_\alpha = \Phi^{-1}\{1 - (\alpha/2)\}$. To see that this pseudo-posterior credible interval indeed has the correct asymptotic frequentist coverage for $\Psi_{0,uv}$, we consider the probability of coverage under repeated sampling

$$\begin{aligned} P_0[\Psi_{0,uv} \in \mathcal{C}_{uv}] &= P_0\left[\sqrt{n}\frac{|T_{uv} - \Psi_{0,uv}|}{\mathcal{S}_{0,uv}} \leq z_\alpha\right] \\ &\rightarrow 2\Phi(z_\alpha) - 1 = 1 - \alpha, \end{aligned}$$

as both $n, p \rightarrow \infty$, using Theorem 11. The form of the asymptotic variance $\mathcal{S}_{0,uv}^2$ is seen to be greater for larger values of the signal-to-noise ratio in the data.

4.4 Simulation Results

4.4.1 Preliminaries

In this Section, we compare the proposed CC-FABLE approach with competitors when judging their performance from the viewpoint of estimation error and uncertainty quantification. For fixed data generating matrices Λ_0 and Σ_0 , we let the common high-dimensional covariance matrix of the observed data be $\Psi_0 = \Lambda_0\Lambda_0^\top + \Sigma_0$. Given an estimator $\hat{\Psi}$ of Ψ_0 , we judge the estimator using the relative spectral error, defined as

$$\mathcal{L}(\Psi_0, \hat{\Psi}) = \frac{\|\Psi_0 - \hat{\Psi}\|}{\|\Psi_0\|}.$$

Normalization of the spectral error by $\|\Psi_0\|$ ensures that the relative spectral errors are comparable on a same scale, across cases with different Ψ_0 with larger or smaller values. For a given Ψ_0 , we obtain the average of the relative spectral error when using $\hat{\Psi}$ as an estimate with $R = 20$ replications of the data. When using a Bayesian method, we consider $\hat{\Psi}$ to be the posterior mean of Ψ .

We also evaluate the frequentist coverage of the (pseudo) posterior credible intervals of the entrywise elements of Ψ used to estimate Ψ_0 for both CC-FABLE and FABLE, illustrating the role of coverage-correction. For sake of convenience, we only

considered the coverage for the first 100×100 submatrix of Ψ_0 with $100 \times (100+1)/2 = 5050$ distinct entries. For evaluation, we consider the 95% (pseudo) posterior credible intervals of the considered entrywise elements and compare their frequentist coverage with the nominal value of 0.95, across $R_C = 100$ replications of the data.

As competitors to the proposed CC-FABLE approach, we consider the approaches in Bhattacharya and Dunson (2011) and Ročková and George (2016) and denote them by MGSP and ROTATE, respectively, across all the simulations. The approach of Bhattacharya and Dunson (2011) carries out posterior sampling with the multiplicative gamma shrinkage prior (MGSP), yielding posterior samples of the covariance matrix. In contrast, the approach of Ročková and George (2016) only yields point estimates of the covariance matrix, using an expectation-maximization (EM) approach to approximate the posterior mean with spike and slab prior distributions to learn the entries. As a result, we do not consider ROTATE for our uncertainty quantification experiments. For CC-FABLE and MGSP, we used 1000 Monte Carlo iterates and 4000 MCMC iterates, respectively, along with discarding the first 2000 as burn-in in the case of MGSP. As illustrated earlier, CC-FABLE does not require MCMC for inference, instead providing direct Monte Carlo samples. We used the `infinitefactor` package (Poworoznek et al., 2021) for implementing the MGSP, while code to implement ROTATE was obtained from http://veronikarock.com/FACTOR_ANALYSIS.zip. All the simulations were carried out in the R programming language (R Core Team, 2021).

4.4.2 Estimation Performance

We first consider estimation performance. For this criteria, we consider four different combinations of n and p , namely

$$(n, p) \in \{(500, 1000), (1000, 1000), (500, 5000), (1000, 5000)\}.$$

Table 4.1: Estimation error results for $\pi_0 = 0$.

Cases	CC-FABLE	MGSP	ROTATE
$n = 500, p = 1000$	0.49	0.69	0.65
$n = 1000, p = 1000$	0.40	0.44	0.48
$n = 500, p = 5000$	0.53	-	0.72
$n = 1000, p = 5000$	0.40	-	0.54

Table 4.2: Estimation error results for $\pi_0 = 0.4$.

Cases	CC-FABLE	MGSP	ROTATE
$n = 500, p = 1000$	0.61	0.79	0.65
$n = 1000, p = 1000$	0.49	0.51	0.49
$n = 500, p = 5000$	0.59	-	0.71
$n = 1000, p = 5000$	0.46	-	0.55

For all choices of (n, p) , we let the true number of factors be $k = 10$. For a given choice of p , we consider three different regimes for the true factor loadings matrix, differing in the amount of sparsity present. We assume the ij th element $\Lambda_{0,ij}$ of Λ_0 for $1 \leq i \leq j \leq p$ is generated independently from

$$\Lambda_{0,ij} \sim \pi_0 \mathcal{P}_0 + (1 - \pi_0)N(0, 0.01),$$

where \mathcal{P}_0 represents a point mass at 0 and $\pi_0 \in \{0, 0.4, 0.6\}$ represents the proportion of exact zeroes present in the matrix, with three different values varying between no sparsity, moderate sparsity, and high sparsity. Although it is common to assume sparsity while modeling the factor loadings matrix, often in practice the factor loadings may not be sparse, but only small in magnitude, reflecting $\pi_0 = 0$. The true error variances σ_{0j}^2 for $1 \leq j \leq p$ for all the 4 different cases of (n, p) are generated independently from $\mathcal{U}(0.5, 2)$. The obtained results are in Tables 4.1, 4.2, and 4.3. For $p = 5000$, we do not report the results for MGSP, whose implementation ran into memory overflow problems.

In all the three scenarios across different combinations of (n, p) , the performance

Table 4.3: Estimation error results for $\pi_0 = 0.6$.

	CC-FABLE	MGSP	ROTATE
$n = 500, p = 1000$	0.76	0.94	0.67
$n = 1000, p = 1000$	0.60	0.62	0.49
$n = 500, p = 5000$	0.64	-	0.71
$n = 1000, p = 5000$	0.51	-	0.55

of CC-FABLE is either the best or close to being the best. CC-FABLE performs better than the ROTATE approach for low to moderate sparsity, while performing slightly worse when the factor loadings are highly sparse, which aids the high sparsity inducing prior specifications of both ROTATE and MGSP. However, a remarkable observation is that for a fixed sample size, an increase in dimensions leads to a decrease of relative error for CC-FABLE. This further highlights the blessing of dimensionality when carrying out estimation using CC-FABLE. For both ROTATE and MGSP, an increase in dimensions with the same sample size leads to a worse relative error than before. As a result, CC-FABLE is able to perform better than ROTATE even in the high sparsity case when the number of dimensions are much larger than the sample size. MGSP seems to suffer particularly in the case of small n and large p . We think this is due to well documented mixing problems leading to suboptimal performance of the MCMC for posterior sampling.

4.4.3 Frequentist Coverage

We now compare the CC-FABLE and FABLE in terms of frequentist coverage of the entrywise elements of Ψ_0 by (pseudo) credible intervals. We carry out a comparison of CC-FABLE and FABLE to investigate the effect of the coverage-correction on the proposed procedure. We let $n = 500, p = 5000, \pi_0 = 0.5$ and assume that

$$\Lambda_{0,ij} \stackrel{ind}{\sim} \pi_0 \mathcal{P}_0 + (1 - \pi_0)N(0, \nu^2),$$

Table 4.4: Comparison of frequentist coverage of entrywise pseudo-credible intervals obtained from CC-FABLE and FABLE.

Case	CC-FABLE	FABLE
$\nu = 0.5$	0.951	0.893
$\nu = 1$	0.950	0.768
$\nu = 2$	0.949	0.540

where we vary $\nu \in \{0.5, 1, 2\}$. We generate the error variances $\sigma_{0j}^2 \stackrel{iid}{\sim} \mathcal{U}(0.1, 2)$ as before. We provide values for the coverage of 95% pseudo-posterior credible intervals for the uncorrected FABLE and the modified CC-FABLE in Table 4.4.

In Table 4.4, it is clear that the CC-FABLE approach provides vastly superior coverage to the FABLE approach, particularly as the loadings grow larger in magnitude with increasing values of ν . The primary reason for this phenomenon is that asymptotically, the pseudo-posterior variance of the covariance matrix entries when using the uncorrected FABLE is smaller than the variance of its pseudo-posterior mean under repeated sampling of data. The coverage-correction uses the factors b_{uv} as introduced in Section 4.2.2 to adjust for this disagreement in variances, improving the coverage. In Figure 4.1, we compare the pseudo-posterior credible intervals for CC-FABLE and FABLE for an arbitrarily chosen replicate of the data for $\nu = 2$. For FABLE, the intervals are too narrow and often miss the $y = x$ line, implying that the pseudo-posterior credible intervals do not capture the respective true entry of Ψ_0 . On the other hand, the intervals for CC-FABLE appear to be much better calibrated, with the $y = x$ line almost completely captured by the plotted intervals. Thus, we conclude that the coverage-correction mechanism in CC-FABLE provides substantially improved frequentist coverage over the unadjusted FABLE variant.

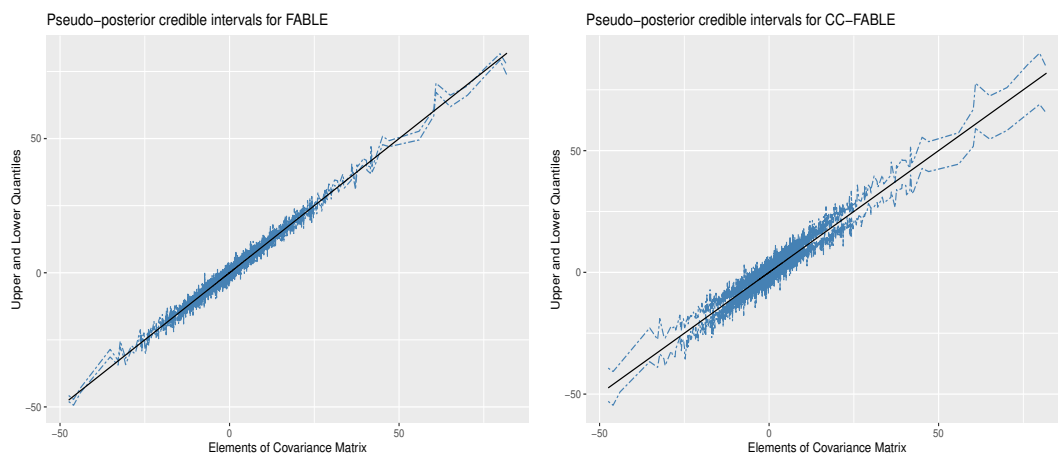


FIGURE 4.1: Comparison of pseudo-posterior credible intervals for CC-FABLE and FABLE for $\nu = 2$. On the x-axis, we plot the true entries of the first 100×100 submatrix of Ψ_0 . Dotted blue lines represent the 2.5% and 97.5% quantiles, with the solid black line representing the $y = x$ line. Greater coverage corresponds to greater part of the solid black line being captured by the dotted blue lines.

Conclusion and Future Research

The topics covered in this work only scratch the tip of the iceberg when it comes to incorporating scalability and structured constraints in Bayesian modeling. Our work may be extended to multiple exciting future research directions, some of which are outlined below.

In Chapter 2, the proposed NN-DM approach provides a useful alternative to Bayesian density estimation based on Dirichlet mixtures with much faster computational speed and stability in avoiding MCMC, along with providing provably accurate guarantees on estimation and uncertainty quantification for a wide variety of data generating densities. MCMC can have very poor performance in mixture models and other multimodal cases, due to difficulty in mixing, and hence can lead to posterior inferences that are unreliable. The main conceptual disadvantage of the proposed approach is the lack of a coherent Bayesian posterior updating rule. However, it is important to keep in mind that Bayesian kernel mixtures have key disadvantages that are difficult to remove within a fully coherent Bayesian modeling framework. These include a strong sensitivity to the choice of kernel and prior on the weights on these kernels; refer, for example to Miller and Dunson (2019). There are several

important next steps. The first is to develop fast and robust algorithms for using the NN-DM not just for density estimation but also as a component of more complex hierarchical models. For example, one may want to model the residual density in regression nonparametrically or treat a random effects distribution as unknown. In such settings, one can potentially update other parameters within a Bayesian model using MCMC, while using algorithms related to those proposed in this article to update the nonparametric part conditionally on these other parameters.

In Chapter 3, we proposed SAID, a Bayesian framework for inference on main effects and pairwise synergistic, antagonistic, or null interaction effects, motivated by the *mixtures problem* in environmental epidemiology. Instead of sharply imposing constraints, we use a shrinkage prior that penalizes deviations from synergistic, antagonistic, or null interactions. In our NHANES analysis using SAID, we found a variety of significant interactions among metal exposures impacting kidney function. To our knowledge, these interactions are currently unknown to the epidemiology and public health communities. It will be very interesting to validate these results in other cohorts, to study the mechanisms by which synergistic and antagonistic interactions occur in these particular metals, and also assess the regulatory implications. Our approach for assessing interactions in observational epidemiology data can be used for identifying specific pairs of chemicals to study in more detail in *in vivo* and *in vitro* assays. Although we assumed the response to be continuous, it is straightforward to incorporate other types of responses in our framework, such as binary or count variables. In particular, we can model the response variables as falling in an exponential family with the linear predictor having exactly the form described in Chapter 3. For binary outcomes and assuming a logistic link function, our proposed computational algorithm can be easily modified by relying on Polya-Gamma data augmentation (Polson et al., 2013), with a related approach which may be used for counts. It is additionally straightforward to include further applied complications,

such as spatially or temporally dependent data, by incorporating appropriate terms in the linear predictor. The exposures in the NHANES 2015-16 data considered in this chapter are mild-to-moderately correlated. In scenarios where the exposures are more heavily correlated, individually interpreting main effects and interaction effects is often infeasible. In this scenario, one could think of dividing the exposure variables into groups based on chemical class, clustering algorithms, or latent factor models. It is then of interest to investigate presence of interactions between chemical groups.

In Chapter 4, we developed an algorithm providing pseudo-posterior samples of a high-dimensional covariance matrix modeled using ideas from factor analysis, providing accurate estimation, uncertainty quantification, and immense computational benefits. The proposed FABLE approach, like the NN-DM in Chapter 2, completely bypasses MCMC and the pitfalls associated with it; the key piece is a blessing of dimensionality phenomenon allowing us to pre-estimate the unknown latent factors. However, similar to the NN-DM, the lack of a coherent Bayesian updating rule in FABLE makes it harder to incorporate it as a part of a more complex hierarchical model where the end goal is not factor analysis or high-dimensional covariance estimation. Instead of linear latent factor models, one often encounters binary and/or count data, particularly in ecological settings where such data are available in abundance. Such ecological data could be very large in scale, with both the number of samples and the number of dimensions potentially exceeding tens or hundreds of thousands. In such settings, modeling the data using a logistic/Poisson model with latent factors and factor loadings is commonplace to introduce dependence among the observations, and it is of interest to obtain a computationally efficient estimate of the factor loadings matrix after obtaining a pre-estimate of the factors. Perhaps, this could be achieved by the blessing of dimensionality observed for linear factor models extended to the generalized linear model case.

Broadly, our efforts in this work primarily focused on improving scalability by

developing embarrassingly parallel algorithms to bypass MCMC. Although providing immense computational gains when compared to traditional Bayesian approaches employing MCMC for posterior sampling, one loses the coherence obtained from a Bayesian updating mechanism. This makes the process of incorporating the developed approaches for density estimation / factor analysis as a part of more complicated hierarchical specifications not straightforward. For example, the approach of Ferrari and Dunson (2021) focuses on regression of health outcomes based on high-dimensional exposures, introducing a joint factor model to alleviate multicollinearity between the exposures when fitting the regression. A natural question for future research combining all the three directions in Chapters 2, 3, and 4 is: could we obtain a blessing of dimensionality phenomenon in such a scenario for high-dimensional factor regression, with modeling the measurement error density in a nonparametric fashion to reduce misspecification? It is also of independent interest to develop theoretical properties of such pseudo-Bayesian procedures. In particular, one could investigate how the contraction rates between that of a coherent Bayesian posterior and a proposed pseudo-posterior compare. From the viewpoint of uncertainty quantification, it is compelling to investigate whether Bernstein-von Mises results hold for such pseudo-posteriors, which would allow us to interpret pseudo-posterior credible intervals as frequentist confidence intervals asymptotically.

Appendix A

Further details for Chapter 2

A.1 Prerequisites

We first introduce some notation with accompanying technical details which will be used hereafter. We denote the Frobenius norm and determinant of $A \in \mathbb{R}^{p \times p}$ by $\|A\|_F = \{\text{tr}(A^\top A)\}^{1/2}$ and $|A|$, respectively. For $v \in \mathbb{R}^p$, one has $\|vv^\top\|_F = \|v\|_2^2$ where $\|a\|_2 = (a^\top a)^{1/2}$ is the Euclidean norm of a . For two symmetric matrices $A, B \in \mathbb{R}^{p \times p}$, we say that $A \geq B$ if $A - B$ is positive semi-definite, that is $x^\top(A - B)x \geq 0$ for all $x \in \mathbb{R}^p, x \neq 0_p$ where $0_p = (0, \dots, 0)^\top$. For a real symmetric matrix A_* , let the eigenvalues of A_* be $e_1(A_*), \dots, e_p(A_*)$, arranged such that $e_1(A_*) \geq \dots \geq e_p(A_*)$. If $A \geq B$, then it follows by the min-max theorem (Teschl, 2009) that for each $j = 1, \dots, p$, we have $e_j(A) \geq e_j(B)$. In particular, we have $|A| \geq |B|$ and $\|A\|_F \geq \|B\|_F$.

Now consider a true data generating density $X_1, \dots, X_n \stackrel{iid}{\sim} f_0$ satisfying Assumptions 1-3 as in Section 2.3.1. Let $\mathcal{X}^{(n)} = (X_1, \dots, X_n)$ and suppose f_0 induces the measure P_{f_0} on the Borel σ -field on \mathbb{R}^p , denoted by $\mathcal{B}(\mathbb{R}^p)$. We form the k -nearest neighborhood of X_i using the Euclidean norm for $i = 1, \dots, n$. We also let k depend

on n and express this dependence as k_n when required. However, we routinely drop this dependence for notational simplicity. For X_i , let Q_i be its k th nearest neighbor in $\mathcal{X}^{(n)}$ (for $k = 1$, $Q_i = X_i$) and let R_i be the distance between X_i and Q_i , given by $R_i = \|X_i - Q_i\|_2$. Define the ball

$$B_i = \{y \in [0, 1]^p : 0 < \|y - X_i\|_2 < R_i\}$$

and the probability

$$G(X_i, R_i) = \int_{B_i} f_0(u) du$$

of the ball B_i under P_{f_0} . Let $Y_1^{(i)} = X_i$ and $Y_2^{(i)}, \dots, Y_{k-1}^{(i)}$ denote the rest of the interior points in B_i . Let the mean \bar{X}_i and covariance matrix S_i of the i th neighborhood be

$$\bar{X}_i = \frac{1}{k_n} \left\{ \sum_{j=1}^{k-1} Y_j^{(i)} + Q_i \right\},$$

$$S_i = \frac{1}{k_n} \left\{ \sum_{j=1}^{k-1} (Y_j^{(i)} - \bar{X}_i)(Y_j^{(i)} - \bar{X}_i)^\top + (Q_i - \bar{X}_i)(Q_i - \bar{X}_i)^\top \right\}.$$

We observe that $(Y_2^{(i)}, \dots, Y_{k-1}^{(i)}, Q_i)$ is identically distributed for $i = 1, \dots, n$ since X_1, \dots, X_n are independent and identically distributed. Thus we only consider the case $i = 1$ from here on. For sake of brevity, denote $Y_u^{(1)}$ by Y_u for $u = 2, \dots, k-1$ and Q_1 by Q .

Conditional on $X_1 = x_1 \in [0, 1]^p$ and $R_1 = r_1 > 0$, following Mack and Rosenblatt (1979), the conditional joint density of Y_2, \dots, Y_{k-1} and Q is

$$p(y_2, \dots, y_{k-1}, q \mid x_1, r_1) = \left\{ \prod_{j=2}^{k-1} \frac{f_0(y_j)}{G(x_1, r_1)} \mathbb{1}(y_j \in B_1) \right\} \frac{f_0(q)}{G'(x_1, r_1)} \mathbb{1}(\|q - x_1\| = r_1),$$

where $G'(x_1, r_1) = \partial G(x_1, r_1) / \partial r_1$ and $\mathbb{1}(A)$ denotes the indicator function of the event $A \in \mathcal{B}(\mathbb{R}^p)$. Thus conditional on X_1 and R_1 , the random variables Y_2, \dots, Y_{k-1} are independent and identically distributed, and independent of Q .

Let the function $\rho(x_1, r_1) = r_1^{\kappa_1}$ where κ_1 is a non-negative integer. This function can be identified with $\phi(\cdot)$ in equation (11) of Mack and Rosenblatt (1979). In the results that follow, we will require the expectation of $\rho(x_1, r_1)$ under P_{f_0} for different choices of κ_1 . To that end, we shall repeatedly make use of the equation (12) from Mack and Rosenblatt (1979) adapted to our setting:

$$E_{P_{f_0}}\{R_1^{\kappa_1} \mid X_1 = x_1\} = \frac{(n-1)!}{(k-2)!(n-k)!} \int_0^1 \left\{ \left(\frac{t}{C_p f_0(x_1)} \right)^{\kappa_1/p} + o(t^{\kappa_1/p}) \right\} t^{k-2} (1-t)^{n-k} dt. \quad (\text{A.1})$$

Finally, we let \tilde{E} and $\tilde{\text{var}}$ denote the expectation and variance, respectively, of the NN-DM estimator $f(x)$ under the pseudo-posterior density $\tilde{\Pi}$, described in (2.7). Conditioning notation under $\tilde{\Pi}$ is as usual; for example, the conditional expectation

$$\tilde{E}\{f(x) \mid \pi_1, \dots, \pi_n\} = \sum_{i=1}^n \pi_i \tilde{E}\{\phi_p(x; \eta_i, \Sigma_i)\},$$

where the expectation $\tilde{E}\{\phi_p(x; \eta_i, \Sigma_i)\}$ is with respect to the pseudo-posterior density of (η_i, Σ_i) as described in Section 2.2.2.

A.2 Proof of Theorem 1

Suppose X_1, \dots, X_n are independent and identically distributed random variables generated from the density f_0 supported on $[0, 1]^p$ satisfying Assumptions 1-3. For $i = 1, \dots, n$, recall the definitions of μ_i and Λ_i from (2.8):

$$\mu_i = \frac{\nu_0}{\nu_n} \mu_0 + \frac{k}{\nu_n} \bar{X}_i, \quad \Lambda_i = \frac{\nu_n + 1}{\nu_n (\gamma_n - p + 1)} \Psi_i.$$

We want to show that $\hat{f}_n(x) = (1/n) \sum_{i=1}^n t_{\gamma_n - p + 1}(x; \mu_i, \Lambda_i) \rightarrow f_0(x)$ in P_{f_0} -probability as $n \rightarrow \infty$ for any $x \in [0, 1]^p$, where $\hat{f}_n(x)$ is as described in (2.8). We first prove

two propositions involving successive mean value theorem type approximations to $\hat{f}_n(x)$, which will imply the final result. We now state the two propositions, with accompanying proofs, before stating the final theorem.

Proposition 12. *Fix $x \in [0, 1]^p$. Let $f_A(x) = (1/n) \sum_{i=1}^n t_{\gamma_n-p+1}(x; X_i, \Lambda_i)$. Also, let $k = o(n^{i_1})$ with $i_1 = 2/(p^2 + p + 2)$ and $\nu_0 = o(n^{-1/p} k^{(1/p)+1})$. Then, we have $\mathbf{E}(|\hat{f}_n(x) - f_A(x)|) \rightarrow 0$ as $n \rightarrow \infty$.*

Proof. Since the $(\Lambda_i)_{i=1}^n$ are identically distributed and $(\mu_i)_{i=1}^n$ are identically distributed, we have $\mathbf{E}(|\hat{f}_n(x) - f_A(x)|) \leq \mathbf{E}\{|t_{\gamma_n-p+1}(x; \mu_1, \Lambda_1) - t_{\gamma_n-p+1}(x; X_1, \Lambda_1)|\}$.

The multivariate mean value theorem now implies that

$$\mathbf{E}(|\hat{f}_n(x) - f_A(x)|) \leq \mathbf{E}\left\{|\Lambda_1|^{-1/2} \|\nabla t_{\gamma_n-p+1}(\xi; 0_p, I_p)\|_2 \|\Lambda_1^{-1/2}(X_1 - \mu_1)\|_2\right\}, \quad (\text{A.2})$$

where $\nabla t_{\gamma_n-p+1}(\xi; 0_p, I_p) = [\partial t_{\gamma_n-p+1}(x; 0_p, \mathbb{I}_p)/\partial x]_\xi$ for some ξ in the convex hull of $\Lambda_1^{-1/2}(x - X_1)$ and $\Lambda_1^{-1/2}(x - \mu_1)$.

Using standard results and the min-max theorem, we have

$$\|\Lambda_1^{-1/2}(X_1 - \mu_1)\|_2 \leq \|\Lambda_1^{-1/2}\|_F \|X_1 - \mu_1\|_2.$$

If we let $H_n = H = \{\nu_n(\gamma_n - p + 1)\}^{-1}(\nu_n + 1)\Psi_0 = h^2 I_p$ where $h^2 = h_n^2 = \{\nu_n(\gamma_n - p + 1)\}^{-1}\{(\nu_n + 1)(\gamma_0 - p + 1)\} \delta_0^2$ following the choice of Ψ_0 from Section 2.2.3, then it is clear that $\Lambda_1 \geq H$. Therefore, we have $\|\Lambda_1^{-1/2}(X_1 - \mu_1)\|_2 \leq \|H^{-1/2}\|_F \|X_1 - \mu_1\|_2$. Straightforward calculations show that $\|H^{-1/2}\|_F = h^{-1} p^{1/2}$ and $\|X_1 - \mu_1\|_2 \leq R_1 + \{\nu_n^{-1}(1 + \|\mu_0\|_2)\nu_0\}$ where $R_1 = \|X_1 - X_{1[k]}\|_2$. Using Theorem 2.4 from Biau and Devroye (2015) for $p \geq 2$ and (A.1) for $p = 1$, one gets

$$E_{P_{f_0}}(R_1^2) \leq d_p^2 \left(\frac{k}{n}\right)^{2/p}, \quad (\text{A.3})$$

for an appropriate constant $d_p > 0$. Thus, we have $\mathbf{E}(R_1) \leq \{\mathbf{E}(R_1^2)\}^{1/2} \leq d_p(k/n)^{1/p}$

for sufficiently large n . This implies that

$$E(\|X_1 - \mu_1\|_2) \leq d_p \left(\frac{k}{n}\right)^{1/p} + o\left(\frac{k}{n}\right)^{1/p}. \quad (\text{A.4})$$

We also have $|\Lambda_1|^{-1/2} \leq |H|^{-1/2} = h^{-p}$. Finally, simple calculations yield that

$$\|\nabla t_{\gamma_n-p+1}(\xi; 0_p, I_p)\|_2 \leq L_{1,n,p}$$

where $L_{1,n,p} > 0$ satisfies $L_{1,n,p} \rightarrow (2\pi)^{-p/2} e^{-1/2}$ as $n \rightarrow \infty$. Plugging all these back in (A.2), we obtain a finite constant $L_{2,n,p} > 0$ such that

$$\mathbf{E}(|\hat{f}_n(x) - f_A(x)|) \leq L_{2,n,p} (n^{-i_1} k)^{(p^2+p+2)/(2p)} + o\{(n^{-i_1} k)^{(p^2+p+2)/(2p)}\}, \quad (\text{A.5})$$

which goes to 0 as $n \rightarrow \infty$, completing the proof. \square

We now provide the second mean value theorem type approximation which approximates the random bandwidth matrix Λ_i in $f_A(x)$ by $H = H_n$ for each $i = 1, \dots, n$.

Proposition 13. *Fix $x \in [0, 1]^p$. Let $f_K(x) = (1/n) \sum_{i=1}^n t_{\gamma_n-p+1}(x; X_i, H)$. Also, let $k = o(n^{i_2})$ with $i_2 = 4/(p+2)^2$ and $\nu_0 = o\{n^{-2/p} k^{(2/p)+1}\}$. Then, we have $\mathbf{E}(|f_A(x) - f_K(x)|) \rightarrow 0$ as $n \rightarrow \infty$.*

Proof. Using the identically distributed properties of $(\Lambda_i)_{i=1}^n$ and $(X_i)_{i=1}^n$, we obtain $\mathbf{E}(|f_A(x) - f_K(x)|) \leq \mathbf{E}(|t_{\gamma_n-p+1}(x; X_1, \Lambda_1) - t_{\gamma_n-p+1}(x; X_1, H)|)$. Using the multivariate mean value theorem, we obtain that

$$\mathbf{E}(|t_{\gamma_n-p+1}(x; X_1, \Lambda_1) - t_{\gamma_n-p+1}(x; X_1, H)|) \leq \mathbf{E}(\|M_1\|_F \|\Lambda_1 - H\|_F), \quad (\text{A.6})$$

where $M_1 = [\partial\{t_{\gamma_n-p+1}(x; X_1, \Sigma)\}/\partial\Sigma]_{\Sigma_0}$ for some Σ_0 , with Σ_0 in the convex hull of Λ_1 and H . Since $\Lambda_1 \geq H$, we immediately have $\Sigma_0 \geq H$ as well. Using the definitions of Λ_1 and H , we have

$$\|\Lambda_1 - H\|_F \leq \frac{(\nu_n + 1)}{\nu_n(\gamma_n - p + 1)} \left\{ \left\| \sum_{j \in \mathcal{N}_1} (X_j - \bar{X}_1)(X_j - \bar{X}_1)^\top \right\|_F + \frac{k\nu_0}{\nu_n} \|\bar{X}_1 \bar{X}_1^\top\|_F \right\}.$$

Since $\|\sum_{j \in \mathcal{N}_1} (X_j - \bar{X}_1)(X_j - \bar{X}_1)^\top\|_F \leq \sum_{j \in \mathcal{N}_1} \|(X_j - \bar{X}_1)(X_j - \bar{X}_1)^\top\|_F = \sum_{j \in \mathcal{N}_1} \|X_j - \bar{X}_1\|_2^2 \leq \sum_{j \in \mathcal{N}_1} R_1^2 = kR_1^2$, we get for sufficiently large n the following:

$$\mathbf{E}(\|\Lambda_1 - H\|_F) \leq \mathbf{E}(R_1^2) + o\left(\frac{k}{n}\right)^{2/p}, \quad (\text{A.7})$$

$$\leq d_p^2 \left(\frac{k}{n}\right)^{2/p} + o\left(\frac{k}{n}\right)^{2/p}, \quad (\text{A.8})$$

using (A.3) and $\nu_0 = o\{n^{-2/p}k^{(2/p)+1}\}$. Taking partial derivatives of the logarithm of the t density $\log\{t_{\gamma_n-p+1}(x; X_1, \Sigma)\}$ with respect to Σ evaluated at Σ_0 and taking Frobenius norm of both sides, we obtain

$$\|t_{\gamma_n-p+1}^{-1}(x; X_1, \Sigma_0) M_1\|_F \leq h^{-2}(\gamma_n + 1)$$

for sufficiently large n . We now observe that

$$t_{\gamma_n-p+1}(x; X_1, \Sigma_0) \leq c_{p, \gamma_n-p+1} |\Sigma_0|^{-1/2} \leq c_{p, \gamma_n-p+1} |H|^{-1/2} = h^{-p} c_{p, \gamma_n-p+1},$$

where $c_{p, \beta} = (\pi\beta)^{-p/2} \{\Gamma(\beta/2)\}^{-1} \Gamma\{(\beta+p)/2\}$ for $p \geq 1, \beta > 0$. Note that $c_{p, \beta} \rightarrow (2\pi)^{-p/2}$ as $\beta \rightarrow \infty$ for any $p \geq 1$. This immediately implies that $\|M_1\|_F \leq h^{-(p+2)} c_{p, \gamma_n-p+1} (\gamma_n + 1)$ for sufficiently large n . Plugging all these back in equation (A.6), we obtain for sufficiently large n , a finite $L_{3,n,p} > 0$ such that

$$\mathbf{E}(|f_A(x) - f_K(x)|) \leq L_{3,n,p} (n^{-i_2} k)^{(p+2)^2/(2p)} + o\{(n^{-i_2} k)^{(p+2)^2/(2p)}\}, \quad (\text{A.9})$$

which goes to 0 as $n \rightarrow \infty$, proving the proposition. \square

We now prove Theorem 1.

Theorem 4. $\mathbf{E}(|\hat{f}_n(x) - f_K(x)|) \leq \mathbf{E}(|\hat{f}_n(x) - f_A(x)|) + \mathbf{E}(|f_A(x) - f_K(x)|)$ by the triangle inequality. Using Propositions 12 and 13, we obtain that $E_{P_{f_0}}(|\hat{f}_n(x) - f_K(x)|) \rightarrow 0$ as $n \rightarrow \infty$. From Section A.6 of the Appendix, we obtain $f_K(x) \rightarrow f_0(x)$ in P_{f_0} -probability. This immediately implies that given the conditions on k, ν_0 , and for any $x \in [0, 1]^p$, we have $\hat{f}_n(x) \rightarrow f_0(x)$ in P_{f_0} -probability. \square

A.3 Proof of Theorem 2

Proof. Fix $x \in [0, 1]^p$. For $i = 1, \dots, n$, let $z_i = \phi_p(x; \eta_i, \Sigma_i)$ and suppose $z^{(n)} = (z_1, \dots, z_n)^\top$. Then, we have $f(x) = \sum_{i=1}^n \pi_i z_i = z^{(n)\top} \pi^{(n)}$ where $\pi^{(n)} = (\pi_1, \dots, \pi_n)^\top$.

We begin with the identity

$$\widetilde{\text{var}}\{f(x)\} = \widetilde{\text{var}}[\tilde{E}\{f(x) \mid z^{(n)}\}] + \tilde{E}[\widetilde{\text{var}}\{f(x) \mid z^{(n)}\}]. \quad (\text{A.10})$$

We start with the first term on the right hand side of (A.10). Observe that z_1, \dots, z_n are independent under $\tilde{\Pi}$ and $\tilde{E}(\pi_i) = 1/n$ for $i = 1, \dots, n$. Thus, we have

$$\begin{aligned} \widetilde{\text{var}}[\tilde{E}\{f(x) \mid z^{(n)}\}] &= \widetilde{\text{var}}\left(\frac{1}{n} \sum_{i=1}^n z_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \widetilde{\text{var}}(z_i) \\ &\leq \frac{1}{n^2} \sum_{i=1}^n \tilde{E}(z_i^2) \\ &= \frac{1}{n^2} \sum_{i=1}^n R_n |B_i|^{-1/2} t_{\gamma_n - p + 2}(x; \mu_i, B_i), \end{aligned}$$

since for $i = 1, \dots, n$, we have

$$\tilde{E}(z_i^2) = R_n |B_i|^{-1/2} t_{\gamma_n - p + 2}(x; \mu_i, B_i), \quad (\text{A.11})$$

where

$$R_n = \frac{\Gamma\{(\gamma_n - p + 2)/2\}}{\Gamma\{(\gamma_n - p + 1)/2\}} \left[\frac{\nu_n + 2}{4\pi\nu_n(\gamma_n - p + 2)} \right]^{p/2}, \quad B_i = D_n \Lambda_i,$$

and $D_n = \{2(\gamma_n - p + 2)(\nu_n + 1)\}^{-1}(\gamma_n - p + 1)(\nu_n + 2)$. To obtain (A.11), we integrate over the pseudo-posterior distribution of $(\eta_i, \Sigma_i)_{i=1}^n$, namely $\text{NIW}(\mu_i, \nu_n, \gamma_n, \Psi_i)$.

For $i = 1, \dots, n$, since $|\Lambda_i| \geq |H_n|$, we have $|B_i| \geq D_n^p |H_n|$. Letting $\hat{f}_{\text{var}}(x) = (1/n) \sum_{i=1}^n t_{\gamma_n - p + 2}(x; \mu_i, B_i)$, we have

$$\widetilde{\text{var}}[\tilde{E}\{f(x) \mid z^{(n)}\}] \leq \frac{R_n D_n^{-p/2} \hat{f}_{\text{var}}(x)}{n |H_n|^{1/2}}. \quad (\text{A.12})$$

We now analyze the second term on the right hand side of (A.10). Recall that $\pi^{(n)}$ is independent of $z^{(n)}$ under $\tilde{\Pi}$. Let Σ_π denote the pseudo-posterior covariance matrix of $\pi^{(n)}$. Standard results yield $\Sigma_\pi = V_n\{(1 - C_n)\mathbb{I}_n + C_n\mathbf{1}_n\mathbf{1}_n^\top\}$, where $V_n = (n - 1)/[n^2\{n(\alpha + 1) + 1\}]$, and $C_n = -1/(n - 1)$. Then, we have

$$\tilde{E}[\widetilde{\text{var}}\{f(x) \mid z^{(n)}\}] = \tilde{E}[z^{(n)\top} \Sigma_\pi z^{(n)}]. \quad (\text{A.13})$$

Using the expression for Σ_π along with (A.13), we obtain,

$$\tilde{E}[\widetilde{\text{var}}\{f(x) \mid z^{(n)}\}] = \frac{1}{n(\alpha + 1) + 1} \tilde{E} \left\{ \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 \right\}, \quad (\text{A.14})$$

where $\bar{z} = (1/n) \sum_{i=1}^n z_i$. We now have

$$\begin{aligned} \tilde{E}[\widetilde{\text{var}}\{f(x) \mid z^{(n)}\}] &= \frac{1}{n\{n(\alpha + 1) + 1\}} \left\{ \sum_{i=1}^n \tilde{E}(z_i^2) - n\tilde{E}(\bar{z}^2) \right\} \\ &\leq \frac{1}{n\{n(\alpha + 1) + 1\}} \sum_{i=1}^n \tilde{E}(z_i^2) \\ &= \frac{1}{n\{n(\alpha + 1) + 1\}} \sum_{i=1}^n R_n |B_i|^{-1/2} t_{\gamma_n - p + 2}(x; \mu_i, B_i), \end{aligned}$$

using (A.11). Using $|B_i| \geq D_n^p |H|$ for $i = 1, \dots, n$ as before, we have

$$\tilde{E}[\widetilde{\text{var}}\{f(x) \mid z^{(n)}\}] \leq \frac{R_n D_n^{-p/2} \hat{f}_{\text{var}}(x)}{\{n(\alpha + 1) + 1\} |H_n|^{1/2}}. \quad (\text{A.15})$$

Combining (A.12) and (A.15) and putting the results back in (A.10), we have the desired result. If we let $n \rightarrow \infty$, we immediately obtain that $\widetilde{\text{var}}\{f(x)\} \rightarrow 0$ in P_{f_0} -probability. \square

A.4 Proof of Theorem 4

Proof. We have iid data $\mathcal{X}^{(n)} = (X_1, \dots, X_n)$ such that $X_1, \dots, X_n \stackrel{iid}{\sim} f_0$, with f_0 satisfying Assumptions 1-3 for $p = 1$. Given the NN-DM estimator $f(x) =$

$\sum_{i=1}^n \pi_i \phi(x; \eta_i, \sigma_i^2)$, we define the simplified NN-DM density estimator to be

$$g(x) = \frac{1}{n} \sum_{i=1}^n \phi(x; \eta_i, \sigma_i^2),$$

The simplified estimator $g(x)$ can be interpreted as a version of $f(x)$ with the Dirichlet weights being replaced by their pseudo-posterior mean. That is, $g(x) = \tilde{E}\{f(x) \mid (\eta_1, \sigma_1^2), \dots, (\eta_n, \sigma_n^2)\}$. The pseudo-posterior distribution of $g(x)$ is induced through the pseudo-posterior distributions of $\{(\eta_i, \sigma_i^2)\}_{i=1}^n$. The pseudo-posterior mean is of the form

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i} t_{\gamma_n} \left(\frac{x - \mu_i}{\lambda_i} \right),$$

where $\lambda_i = \{(\nu_n + 1)/\nu_n\}^{1/2} \delta_i$. Let $h_n = (\nu_n \gamma_n)^{-1/2} (\nu_n + 1)^{1/2} (\gamma_0 \delta_0^2)^{1/2}$. Then

$$(nh_n)^{1/2} E_{P_{f_0}} |\hat{f}_n(x) - f_K(x)| \rightarrow 0, \quad (\text{A.16})$$

for $k_n = o(n^{2/7})$ and $k_n \rightarrow \infty$ as $n \rightarrow \infty$ from Section A.2 of the Appendix, where

$$f_K(x) = \frac{1}{nh_n} \sum_{i=1}^n t_{\gamma_n} \left(\frac{x - X_i}{h_n} \right).$$

We want to investigate the asymptotic distribution of $f(x)$ as $n \rightarrow \infty$. For that, we first investigate the asymptotic distribution of the simplified NN-DM estimator $g(x)$, and then show that $f(x)$ and $g(x)$ are asymptotically close in P_{f_0} -probability.

To derive the asymptotic distribution of $g(x)$, we begin with the asymptotic distribution of $f_K(x)$, which can be expressed as $f_K(x) = n^{-1} \sum_{i=1}^n u_{in}$, where $u_{in} = h_n^{-1} t_{\gamma_n} \{(x - X_i)/h_n\}$. Using Lyapunov's central limit theorem and denoting convergence in distribution under f_0 by d_0 , we have

$$\frac{f_K(x) - E_{P_{f_0}} \{f_K(x)\}}{[\text{var}_{P_{f_0}} \{f_K(x)\}]^{1/2}} \xrightarrow{d_0} N(0, 1)$$

if

$$\frac{(\sum_{i=1}^n \rho_{in})^{1/r}}{(\sum_{i=1}^n \tau_{in}^2)^{1/2}} \rightarrow 0, \text{ as } n \rightarrow \infty, \quad (\text{A.17})$$

for some $r > 2$, where $\rho_{in} = E|u_{in} - E(u_{in})|^r$ and $\tau_{in}^2 = E\{u_{in} - E(u_{in})\}^2$ for $i = 1, \dots, n$. By standard calculations, we have

$$\tau_{in}^2 = \frac{f_0(x)}{h_n} \int t_{\gamma_n}^2(u) du + o\left(\frac{1}{h_n}\right).$$

For $r = 3$,

$$\rho_{in} \leq \frac{8f_0(x)}{h_n^2} \int t_{\gamma_n}^3(u) du + o\left(\frac{1}{h_n^2}\right).$$

It is straightforward to see that $\int t_{\gamma_n}^r(u) du / \int t_{\gamma_n}(u) du = \mathcal{O}(1)$ for any $r \geq 1$. So, Lyapunov's condition is satisfied as the ratio in this case satisfies $\mathcal{O}\{(nh_n)^{-1/6}\}$ and $nh_n \rightarrow \infty$. Additionally, $|\tau_{in}^2 - \{f_0(x)/h_n\} \int \phi^2(u) du| \rightarrow 0$. So by a combination of Lyapunov's central limit theorem and Slutsky's theorem, we have

$$(nh_n)^{1/2} \left[f_K(x) - E_{P_{f_0}} \{f_K(x)\} \right] \xrightarrow{d_0} \text{N} \left(0, \frac{f_0(x)}{2\pi^{1/2}} \right), \quad (\text{A.18})$$

since $\int \phi^2(u) du = (2\pi^{1/2})^{-1}$. From the calculations in Section A.6 of the Appendix, we can expand the Taylor series to two more terms to obtain

$$E_{P_{f_0}} \left\{ f_K(x) - f_0(x) - \frac{h_n^2 f_0''(x)}{2} \right\} = \mathcal{O}(h_n^4),$$

since $|f_0^{(4)}(x)| \leq C_0$ for all $x \in [0, 1]$. Thus,

$$(nh_n)^{1/2} \left[f_K(x) - \left\{ f_0(x) + \frac{h_n^2 f_0^{(2)}(x)}{2} \right\} \right] \xrightarrow{d_0} \text{N} \left(0, \frac{f_0(x)}{2\pi^{1/2}} \right), \quad (\text{A.19})$$

provided $n^{-2/9} k_n \rightarrow \infty$ as $n \rightarrow \infty$, implying $(nh_n)^{1/2} h_n^4 \rightarrow 0$.

We now argue that $(nh_n)^{1/2}|g(x) - \hat{f}_n(x)| \rightarrow 0$ in P_{f_0} -probability. For this, we first look at

$$\begin{aligned} E_{P_{f_0}} \left[nh_n \{g(x) - \hat{f}_n(x)\}^2 \right] &= nh_n E_{P_{f_0}} \left[\tilde{E} \left\{ (g(x) - \hat{f}_n(x))^2 \right\} \right] \\ &= nh_n E_{P_{f_0}} \left[\widetilde{\text{var}}\{g(x)\} \right], \end{aligned}$$

since $\tilde{E}\{g(x)\} = \hat{f}_n(x)$. The pseudo-posterior variance of $g(x)$ is given by

$$\widetilde{\text{var}}\{g(x)\} = \frac{1}{n^2} \sum_{i=1}^n \widetilde{\text{var}}\{Z_i(x)\},$$

where $Z_i(x) = \phi(x; \eta_i, \sigma_i^2)$ for $i = 1, \dots, n$. It is straightforward to show that

$$\widetilde{\text{var}}\{Z_i(x)\} \sim |\Delta_n| \frac{1}{\tilde{\lambda}_i^2} t_{2\gamma_n+1} \left(\frac{x - \mu_i}{\tilde{\lambda}_i} \right), \quad (\text{A.20})$$

as $n \rightarrow \infty$, where $\tilde{\lambda}_i^2 = \lambda_i^2/2$ for $i = 1, \dots, n$ and

$$\Delta_n = \frac{u_{\gamma_n}^2}{u_{2\gamma_n+1}} - \frac{1}{(2\pi)^{1/2}},$$

with $u_d = \Gamma\{(d+1)/2\}/\{(d\pi)^{1/2}\Gamma(d/2)\}$ being the normalizing constant of the Student's t-density with degrees of freedom $d > 0$. Using Stirling's approximation, $\Delta_n \rightarrow 0$ as $n \rightarrow \infty$. This immediately implies

$$nh_n \widetilde{\text{var}}\{g(x)\} \leq |\Delta_n| v_g(x),$$

where

$$v_g(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\tilde{\lambda}_i} t_{2\gamma_n+1} \left(\frac{x - \mu_i}{\tilde{\lambda}_i} \right).$$

Using the techniques of Section A.2 of the Appendix, it can be shown that $E_{P_{f_0}}\{v_g(x)\} \rightarrow$

$f_0(x)$ as $n \rightarrow \infty$. Therefore, we have

$$\begin{aligned}
E_{P_{f_0}} \left[nh_n \{g(x) - \hat{f}_n(x)\}^2 \right] &= nh_n E_{P_{f_0}} \left[\tilde{E} \left\{ (g(x) - \hat{f}_n(x))^2 \right\} \right] \\
&= nh_n E_{P_{f_0}} \left[\widetilde{\text{var}} \{g(x)\} \right] \\
&\leq E_{P_{f_0}} \{ |\Delta_n| v_g(x) \} \\
&\rightarrow 0,
\end{aligned}$$

as $n \rightarrow \infty$. A simple application of Chebychev's inequality implies $(nh_n)^{1/2} |g(x) - \hat{f}_n(x)| \rightarrow 0$ in P_{f_0} -probability as $n \rightarrow \infty$. Combining this with (A.16) and (A.19) and using Slutsky's theorem, we obtain the desired result for $g(x)$.

We now demonstrate that $f(x)$ and $g(x)$ are asymptotically close to derive the same result for $f(x)$. We start out with

$$\begin{aligned}
\text{var}_{P_{f_0}} \left[(nh_n)^{1/2} \{f(x) - g(x)\} \right] &= nh_n E_{P_{f_0}} \left[\widetilde{\text{var}} \{f(x) - g(x)\} \right] \\
&= nh_n E_{P_{f_0}} \left[\widetilde{\text{var}} \left\{ \sum_{i=1}^n \left(\pi_i - \frac{1}{n} \right) Z_i(x) \right\} \right] \tag{A.21}
\end{aligned}$$

We now focus on the term inside $E_{P_{f_0}}$ in (A.21) and further decompose it as

$$\begin{aligned}
\widetilde{\text{var}} \left\{ \sum_{i=1}^n \left(\pi_i - \frac{1}{n} \right) Z_i(x) \right\} &= \tilde{E} \left[\widetilde{\text{var}} \left\{ \sum_{i=1}^n \left(\pi_i - \frac{1}{n} \right) Z_i(x) \mid \pi^{(n)} \right\} \right] \\
&\quad + \widetilde{\text{var}} \left[\tilde{E} \left\{ \sum_{i=1}^n \left(\pi_i - \frac{1}{n} \right) Z_i(x) \mid \pi^{(n)} \right\} \right], \tag{A.22}
\end{aligned}$$

where $\pi^{(n)} = (\pi_1, \dots, \pi_n)^\top$. In (A.22), let Ξ_{1n} and Ξ_{2n} be as follows:

$$\begin{aligned}
\Xi_{1n} &= \tilde{E} \left[\widetilde{\text{var}} \left\{ \sum_{i=1}^n \left(\pi_i - \frac{1}{n} \right) Z_i(x) \mid \pi^{(n)} \right\} \right], \\
\Xi_{2n} &= \widetilde{\text{var}} \left[\tilde{E} \left\{ \sum_{i=1}^n \left(\pi_i - \frac{1}{n} \right) Z_i(x) \mid \pi^{(n)} \right\} \right].
\end{aligned}$$

Thus, we can write (A.21) as

$$\text{var}_{P_{f_0}} \left[(nh_n)^{1/2} \{f(x) - g(x)\} \right] = nh_n E_{P_{f_0}}(\Xi_{1n}) + nh_n E_{P_{f_0}}(\Xi_{2n}). \quad (\text{A.23})$$

It is straightforward to see that $\Xi_{1n} = \sum_{i=1}^n \widetilde{\text{var}}(\pi_i) \widetilde{\text{var}}\{Z_i(x)\}$. As $n \rightarrow \infty$, we use the fact that $\pi_i \sim \text{Beta}(\alpha_n + 1, (n-1)(\alpha_n + 1))$ under $\tilde{\Pi}$ and (A.20), to get

$$nh_n E_{P_{f_0}}(\Xi_{1n}) \sim \frac{n \Delta_n}{n(\alpha_n + 1) + 1} \tilde{\Xi}_{1n}, \quad (\text{A.24})$$

for some $\tilde{\Xi}_{1n}$ satisfying $\tilde{\Xi}_{1n} \rightarrow f_0(x)$ and $\Delta_n \rightarrow 0$ as $n \rightarrow \infty$. Therefore, $nh_n E_{P_{f_0}}(\Xi_{1n}) \rightarrow 0$ as $n \rightarrow \infty$. For the second part, let $d_i(x) = (1/\lambda_i) t_{\gamma_n} \{(x - \mu_i)/\lambda_i\}$ so that the pseudo-posterior mean $\hat{f}_n(x) = (1/n) \sum_{i=1}^n d_i(x)$. We first observe that

$$\begin{aligned} \Xi_{2n} &= \widetilde{\text{var}} \left[\sum_{i=1}^n \left(\pi_i - \frac{1}{n} \right) \frac{1}{\lambda_i} t_{\gamma_n} \left(\frac{x - \mu_i}{\lambda_i} \right) \right] \\ &= \widetilde{\text{var}} \left[\sum_{i=1}^n \pi_i d_i(x) \right] \\ &= \frac{1}{n(\alpha_n + 1) + 1} \left[\frac{1}{n} \sum_{i=1}^n \{d_i(x) - \hat{f}_n(x)\}^2 \right] \\ &\leq \frac{1}{n(\alpha_n + 1) + 1} \left[\frac{1}{n} \sum_{i=1}^n d_i^2(x) \right]. \end{aligned}$$

It now follows from some algebra that

$$\frac{1}{n} \sum_{i=1}^n d_i^2(x) \leq \frac{d_0(n) \gamma_n}{2\gamma_n + 1} \left(\frac{2\gamma_n + 1}{\gamma_n} \right)^{1/2} \frac{1}{h_n} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{\tilde{\lambda}_i} t_{2\gamma_n+1} \left(\frac{x - \mu_i}{\tilde{\lambda}_i} \right) \right],$$

where $d_0(n) \rightarrow (2\pi)^{-1/2}$ as $n \rightarrow \infty$. Therefore, we have

$$nh_n E_{P_{f_0}}(\Xi_{2n}) \leq \mathcal{O} \left\{ \frac{1}{2\pi^{1/2}} \frac{n}{n(\alpha_n + 1) + 1} \tilde{\Xi}_{2n} \right\}, \quad (\text{A.25})$$

for some $\tilde{\Xi}_{2n}$ satisfying $\tilde{\Xi}_{2n} \rightarrow f_0(x)$ as $n \rightarrow \infty$. By the conditions of the theorem, we have $nh_n E_{P_{f_0}}(\Xi_{2n}) \rightarrow 0$ as $n \rightarrow \infty$. This, along with (A.24) substituted in (A.23) provides $(nh_n)^{1/2} |f(x) - g(x)| \rightarrow 0$ in P_{f_0} -probability. This implies the desired result for $f(x)$ using Slutsky's theorem. As a result, we can interpret pseudo-credible intervals to be frequentist confidence intervals, on average, asymptotically. \square

A.5 Proof of Theorem 5

A.5.1 A property of the k -nearest neighbor distance

Suppose $X_1, \dots, X_n \stackrel{iid}{\sim} f_0$ with f_0 a density on \mathbb{R}^p satisfying Assumptions 1-3. We denote the induced probability measure P_{f_0} by P_0 in this section for the sake of convenience. We define the smoothed k -nearest neighborhood of X_i as $\mathcal{B}_i = \{y \in \mathbb{R}^p : \|X_i - y\|_2 \leq R_i\}$, where $R_i = \|X_i - X_{i[k_n]}\|_2$ is the Euclidean distance between X_i and the k_n -nearest neighbor of X_i for $i = 1, \dots, n$. By symmetry, R_1, \dots, R_n are identically distributed. Suppose $r_n = (k_n/n)^{1/p}$ and define the quasi-neighborhood $\tilde{\mathcal{B}}_i(r) = \{y \in \mathbb{R}^p : \|X_i - y\|_2 \leq r\}$, where the random variables R_i have been replaced by $r \geq 0$. Let

$$\omega_x(r) = \int_{\{y: \|y-x\|_2 \leq r\}} f_0(y) dy.$$

The positive density condition on f_0 obtained from Assumptions 1 and 3 (Evans et al., 2002; Evans, 2008) ensures the existence of $A > 1$ and $\rho > 0$ such that for all $0 \leq r \leq \rho$ and for all $x \in [0, 1]^p$,

$$\frac{r^p}{A} \leq \omega_x(r) \leq Ar^p. \tag{A.26}$$

We first state a Lemma proving some important properties of R_1 . We next use this Lemma to prove Theorem 5. Recall that two non-negative sequences (a_n) and (b_n) are said to be asymptotically equivalent if $|a_n/b_n| \rightarrow c_0$ for some $c_0 > 0$, denoted by $a_n \sim b_n$.

Lemma 14. Define $i_0 = \{2/(p^2 + p + 2)\} \wedge \{4/(p + 2)^2\}$ as in Theorem 1. Assume $k_n \sim n^{i_0 - \epsilon}$ for some $\epsilon \in (0, i_0)$. Suppose $\delta > 0$ satisfies

$$\delta < (1 - i_0 + \epsilon)^{-1} - 1.$$

Define

$$r_n = \left(\frac{k_n}{n}\right)^{1/p}, \quad c_n = \frac{1}{(Ae)^{1/p}} r_n^{1+\delta}, \quad \text{and} \quad n_0 = \left\lceil \left\{ \frac{1 - \frac{i_0 - \epsilon}{2}}{\delta(1 - i_0 + \epsilon)} + 1 \right\}^{\frac{1}{i_0 - \epsilon}} \right\rceil + 1.$$

Then, the following results hold:

$$(i) \quad p_n = P_0(R_1 \leq c_n) = \mathcal{O} \left[k_n^{-1/2} \left(\frac{k_n}{n}\right)^{(k_n-1)\delta} \right] \text{ as } n \rightarrow \infty.$$

$$(ii) \quad \sum_{n=n_0+1}^{\infty} p_n < \infty. \text{ That is, } \{p_n\}_{n=1}^{\infty} \text{ is summable.}$$

$$(iii) \quad P \left[\limsup_{n \rightarrow \infty} \{R_1 \leq c_n\} \right] = 0.$$

$$(iv) \quad nc_n^p \rightarrow \infty \text{ as } n \rightarrow \infty.$$

Proof. (i) Note that $c_n \leq \rho$ for sufficiently large n . From Lemma 4.1 of Evans et al. (2002) we have

$$\begin{aligned} P_0(R_1 \leq c_n \mid X_i = x) &= \int_0^{\omega_x(c_n)} (k_n - 1) \binom{n-1}{k_n-1} y^{k_n-2} (1-y)^{n-k_n} dy \\ &\leq \binom{n-1}{k_n-1} \omega_x(c_n)^{k_n-1} \\ &\sim k_n^{-1/2} \left(\frac{k_n}{n}\right)^{(k_n-1)\delta}, \end{aligned}$$

for any $x \in [0, 1]^p$, using (A.26). This immediately implies

$$P_0(R_1 \leq c_n) = \int_{[0,1]^p} P_0(R_i \leq c_n \mid X_i = x) f_0(x) dx \leq \mathcal{O} \left[k_n^{-1/2} \left(\frac{k_n}{n}\right)^{(k_n-1)\delta} \right].$$

- (ii) For $n > n_0$, we have $p_n = \mathcal{O}\{n^{-(1+\Theta_n)}\}$ for a sequence $\Theta_n \rightarrow \infty$, $\Theta_n > 0$. This ensures that $\sum_{n=n_0+1}^{\infty} p_n < \infty$.
- (iii) Since $\sum_{n=1}^{\infty} p_n < \infty$, a direct application of the first Borel-Cantelli lemma proves the statement.
- (iv) We have, using the condition on δ ,

$$\begin{aligned} nc_n^p &= \frac{n}{Ae} \left(\frac{k_n}{n} \right)^{1+\delta} \\ &= \frac{1}{Ae} \frac{k_n^{1+\delta}}{n^\delta} \\ &\rightarrow \infty, \quad \text{as } n \rightarrow \infty. \end{aligned}$$

□

We now use the above Lemma to prove Theorem 5. The key idea is to leverage the fact that $R_i > c_n$ for all $i = 1, \dots, n$ with probability 1 for all but finite n .

A.5.2 Number of effective member points in each neighborhood

We now prove Theorem 5.

Proof. Using (iii) from Lemma 14, for $i = 1, \dots, n$, we have an integer \tilde{N}_i such that for all $n \geq \tilde{N}_i$, $P_0(R_i > c_n) = 1$. However, since R_1, \dots, R_n are identically distributed, $\tilde{N}_1 = \dots = \tilde{N}_n = \tilde{N}$, say. Thus, for all $i = 1, \dots, n$, we have $P_0(R_i > c_n) = 1$ for all $n \geq \tilde{N}$. This immediately implies that $P_0[\bigcap_{i=1}^n \{R_i > c_n\}] = 1 - P_0[\bigcup_{i=1}^n \{R_i \leq c_n\}] \geq 1 - \sum_{i=1}^n P_0[R_i \leq c_n] = 1$ for all $n \geq \tilde{N}$, which shows

$P_0 [\bigcap_{i=1}^n \{R_i > c_n\}] = 1$. Therefore, we have

$$\begin{aligned}
nQ_n &= nP_0 \left[X_2 \in \mathcal{B}_1, \bigcap_{i=3}^n \{X_2 \notin \mathcal{B}_i\}, \bigcap_{i=1}^n \{R_i > c_n\} \right] \\
&\leq nP_0 \left[\bigcap_{i=3}^n \{X_2 \notin \tilde{\mathcal{B}}_i(c_n)\} \right] \\
&= nP_0 \left[\bigcap_{i=3}^n \{\|X_2 - X_i\| > c_n\} \right] \\
&= n \int \theta_n^{n-2}(x) f_0(x) dx,
\end{aligned}$$

where $\theta_n(x) = 1 - \omega_x(c_n)$, since $X_1, \dots, X_n \stackrel{iid}{\sim} f_0$. Using (A.26), we have $\theta_n(x) \leq 1 - (c_n^p/A)$ for all $x \in [0, 1]^p$. Given the conditions on k , it follows that as $n \rightarrow \infty$,

$$\log n + n \log \left(1 - \frac{c_n^p}{A} \right) \sim \log n - \frac{n^\xi}{A^2 e} \rightarrow -\infty,$$

for all $x \in [0, 1]^p$, where $\xi = 1 - (1 + \epsilon - i_0)(1 + \delta) > 0$. Therefore, we have

$$\begin{aligned}
nQ_n &= n \int \theta_n^{n-2}(x) f_0(x) dx \\
&\leq n \left[1 - \frac{c_n^p}{A} \right]^{n-2} \int f_0(x) dx \\
&= \mathcal{O} \left[\exp \left\{ \log n + n \log \left(1 - \frac{c_n^p}{A} \right) \right\} \right] \\
&\rightarrow 0,
\end{aligned}$$

as $n \rightarrow \infty$. This proves the result. \square

A.6 Proof of Consistency of KDE

Define the standard multivariate t-density with $d > 0$ degrees of freedom to be $g_d(x) = t_d(x; 0_p, \mathbb{I}_p)$. Since $H = H_n = h_n^2 \mathbb{I}_p$ as defined in Section 2.3.1 is diagonal, it

immediately follows that $t_{\gamma_n-p+1}(x; \mu, H) = h_n^{-p} g_{\gamma_n-p+1}\{h_n^{-1}(x - \mu)\}$. The following lemma proves the consistency of any such generic kernel density estimator with kernel depending on n , say

$$f_K(x) = \frac{1}{nw^p} \sum_{i=1}^n g_{\gamma_n-p+1} \left(\frac{x - X_i}{w} \right),$$

where the bandwidth $w = w_n$ satisfies $w_n \rightarrow 0$ and $nw_n^p \rightarrow \infty$ as $n \rightarrow \infty$, with independent and identically distributed data $X_1, \dots, X_n \sim f_0$ satisfying Assumptions 1-3.

Lemma 15. *Suppose w_n is a sequence satisfying $w_n \rightarrow 0$ and $nw_n^p \rightarrow \infty$ as $n \rightarrow \infty$. Let $f_K(x) = (nw_n^p)^{-1} \sum_{i=1}^n g_{\gamma_n-p+1}\{w_n^{-1}(x - X_i)\}$. Then $f_K(x) \rightarrow f_0(x)$ in P_{f_0} -probability for each $x \in [0, 1]^p$.*

Proof. It is enough to show that $\mathbf{E}\{f_K(x)\} \rightarrow f_0(x)$ and $\text{var}_{P_{f_0}}\{f_K(x)\} \rightarrow 0$ as $n \rightarrow \infty$. Let us start first with $\mathbf{E}\{f_K(x)\}$. We have

$$\begin{aligned} \mathbf{E}\{f_K(x)\} &= \mathbf{E} \left\{ \frac{1}{w_n^p} g_{\gamma_n-p+1} \left(\frac{x - X_1}{w_n} \right) \right\} \\ &= \int_{[0,1]^p} \frac{1}{w_n^p} g_{\gamma_n-p+1} \left(\frac{y - x}{w_n} \right) f_0(y) dy \\ &= \int_{\left[-\frac{x}{w_n}, \frac{1-x}{w_n}\right]^p} g_{\gamma_n-p+1}(u) f_0(x + w_n u) du, \\ &= \int_{\left[-\frac{x}{w_n}, \frac{1-x}{w_n}\right]^p} g_{\gamma_n-p+1}(u) \{f_0(x) + w_n u^\top \nabla f_0(\xi)\} du \\ &= f_0(x) \int_{\left[-\frac{x}{w_n}, \frac{1-x}{w_n}\right]^p} g_{\gamma_n-p+1}(u) du + w_n \int_{\left[-\frac{x}{w_n}, \frac{1-x}{w_n}\right]^p} g_{\gamma_n-p+1}(u) u^\top \nabla f_0(\xi) du, \\ &= f_0(x) \{1 - o_n(1)\} + w_n \mathcal{O}_n(1), \end{aligned}$$

using the mean value theorem and Polya's theorem (Pólya, 1920) along with Assumption 2 to bound $\nabla f_0(\cdot)$. As $n \rightarrow \infty$, this implies that $E_{P_{f_0}}\{f_K(x)\} \rightarrow f_0(x)$ since $w_n \rightarrow 0$ as $n \rightarrow \infty$.

The variance may be dealt with in a similar manner. Following the same steps as before we get

$$\begin{aligned}
\text{var}_{P_{f_0}}\{f_K(x)\} &= \frac{1}{n} \text{var}_{P_{f_0}} \left\{ \frac{1}{w_n^p} g_{\gamma_n-p+1} \left(\frac{x - X_1}{w_n} \right) \right\} \leq \frac{1}{n} \mathbf{E} \left\{ \frac{1}{w_n^{2p}} g_{\gamma_n-p+1}^2 \left(\frac{x - X_1}{w_n} \right) \right\} \\
&\leq \frac{1}{nw_n^{2p}} \int_{[0,1]^p} g_{\gamma_n-p+1}^2 \left(\frac{y-x}{w_n} \right) f_0(y) dy \\
&\leq \frac{1}{nw_n^p} \int_{[-\frac{x}{w_n}, \frac{1-x}{w_n}]^p} g_{\gamma_n-p+1}^2(u) \{f_0(x) + w_n u^\top \nabla f_0(\xi)\} du, \\
&\leq \frac{f_0(x) \mathcal{O}_n(1)}{nw_n^p},
\end{aligned}$$

which shows that the variance goes to 0 as $n \rightarrow \infty$, since $nw_n^p \rightarrow \infty$ as $n \rightarrow \infty$. \square

For the NN-DM, recall

$$f_K(x) = \frac{1}{n} \sum_{i=1}^n t_{\gamma_n-p+1}(x; X_i, H_n)$$

from Section 2.3.1 of the main document, where $H_n = h_n^2 \mathbb{I}_p$ and $h_n^2 = \{\nu_n(\gamma_n - p + 1)\}^{-1} \{(\nu_n + 1)(\gamma_0 - p + 1)\} \delta_0^2$. Here, the bandwidth h_n satisfies $h_n \rightarrow 0$ and $nh_n^p \rightarrow \infty$ as $n \rightarrow \infty$. Lemma 15 then shows that $f_K(x)$ converges to $f_0(x)$ in P_{f_0} -probability as $n \rightarrow \infty$.

A.7 Cross-validation

A.7.1 Algorithm for leave-one-out cross-validation

Consider independent and identically distributed data $X_1, \dots, X_n \in \mathbb{R}^p \sim f$ with f having the NN-DM formulation. The prior of the neighborhood parameters (η_i, Σ_i) following Sections 2.2.2 and 2.2.3 is $(\eta_i, \Sigma_i) \sim \text{NIW}_p(\mu_0, \nu_0, \gamma_0, \Psi_0)$ where $\Psi_0 = (\gamma_* \delta_0^2) \mathbb{I}_p$ with $\gamma_* = \gamma_0 - p + 1$. We use the pseudo-posterior mean in (2.8) to compute leave-one-out log-likelihoods $\mathbf{L}(\delta_0^2)$ for different choices of the hyperparameter

δ_0^2 , choosing $\delta_{0,CV}^2 = \arg \sup_{\delta_0^2} \mathbf{L}(\delta_0^2)$ to maximize this criteria. The details of the computation of $\mathbf{L}(\delta_0^2)$ for a fixed δ_0^2 are provided in Algorithm 3.

Algorithm 3. *Leave-one-out cross-validation for choosing the hyperparameter δ_0^2 in nearest neighbor-Dirichlet mixture method.*

- Consider data $\mathcal{X}^{(n)} = (X_1, \dots, X_n)$ where $X_i \in \mathbb{R}^p$, $p \geq 1$.

Fix the number of neighbors k and other hyperparameters μ_0, ν_0, γ_0 .

- For $i \in \{1, \dots, n\}$, consider the data set leaving out the i th data point, given by $\mathcal{X}^{-i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. Compute the pseudo-posterior mean density estimate at X_i , namely $\hat{f}_{-i}(X_i)$, using \mathcal{X}^{-i} and (2.8); let $\mathbf{L}_i(\delta_0^2) = \hat{f}_{-i}(X_i)$. Finally, compute the leave-one-out log-likelihood given by

$$\mathbf{L}(\delta_0^2) = \frac{1}{n} \sum_{i=1}^n \log\{\mathbf{L}_i(\delta_0^2)\}.$$

- For $\delta_0^2 > 0$, obtain $\delta_{0,CV}^2 = \arg \sup_{\delta_0^2} \mathbf{L}(\delta_0^2)$.

A.7.2 Fast Implementation of cross-validation

In Algorithm 3, the nearest neighborhood specification for each \mathcal{X}^{-i} is different for $i = 1, \dots, n$. However, we bypass this computation by initially forming a neighborhood of size $(k + 1)$ for each data point using the entire data and storing the respective neighborhood means and covariance matrices. Suppose for X_i , the indices of the $(k + 1)$ -nearest neighbors are given by $\tilde{\mathcal{N}}_i = \{j \in \{1, \dots, n\} : \|X_i - X_j\|_2 \leq \|X_i - X_{i[k+1]}\|_2\}$, arranged in increasing order according to their distance from X_i with $X_{i[1]} = X_i$. Define the neighborhood mean $m_i = \{1/(k + 1)\} \sum_{j \in \tilde{\mathcal{N}}_i} X_j$ and the neighborhood covariance matrix $S_i = (k + 1)^{-1} \{\sum_{j \in \tilde{\mathcal{N}}_i} (X_j - m_i)(X_j - m_i)^\top\}$. Then, to form a k -nearest neighborhood for the new data \mathcal{X}^{-i} , a single pass over the initial

neighborhoods $\tilde{\mathcal{N}}_i$ is sufficient to update the new neighborhood means and covariance matrices. Below, we describe the update for the neighborhood means $m_j^{(-i)}$ and covariance matrices $S_j^{(-i)}$ for $j = 1, \dots, n$ and $j \neq i$, considering the data \mathcal{X}^{-i} . For $j = 1, \dots, n$ and $j \neq i$, we have,

$$m_j^{(-i)} = \begin{cases} (1/k)\{(k+1)m_j - X_{j[k+1]}\} & \text{if } i \notin \tilde{\mathcal{N}}_j, \\ (1/k)\{(k+1)m_j - X_i\} & \text{if } i \in \tilde{\mathcal{N}}_j. \end{cases}$$

$$S_j^{(-i)} = \begin{cases} S_j - \{(k+1)/k\}(m_j - X_{j[k+1]})(m_j - X_{j[k+1]})^\top & \text{if } i \notin \tilde{\mathcal{N}}_j, \\ S_j - \{(k+1)/k\}(m_j - X_i)(m_j - X_i)^\top & \text{if } i \in \tilde{\mathcal{N}}_j. \end{cases} \quad (\text{A.27})$$

A.8 Algorithm with Gaussian Kernels for Univariate Data

For $p = 1$, we have a univariate Gaussian density $\phi(x; \eta_i, \sigma_i^2)$ in neighborhood i and normal-inverse gamma priors $(\eta_i, \sigma_i^2) \sim \text{NIG}(\mu_0, \nu_0, \gamma_0/2, \gamma_0\delta_0^2/2)$ independently for $i = 1, \dots, n$, with $\mu_0 \in \mathbb{R}$ and $\nu_0, \gamma_0, \delta_0^2 > 0$. That is,

$$\eta_i \mid \sigma_i^2 \sim \text{N}\left(\mu_0, \frac{\sigma_i^2}{\nu_0}\right), \quad \sigma_i^2 \sim \text{IG}\left(\frac{\gamma_0}{2}, \frac{\gamma_0\delta_0^2}{2}\right).$$

Monte Carlo samples from the pseudo-posterior of the unknown density f at any point x can be generated following the steps of Algorithm 4.

Algorithm 4. *Nearest neighbor-Dirichlet mixture algorithm to obtain Monte Carlo samples from the pseudo-posterior of $f(x)$ with Gaussian kernel and normal-inverse gamma prior.*

- **Step 1:** Compute the k -nearest neighborhood \mathcal{N}_i for data point X_i with $X_{i[1]} = X_i$, using the distance $d(\cdot, \cdot)$.
- **Step 2:** Update the parameters for neighborhood \mathcal{N}_i to $(\mu_i, \nu_n, \gamma_n/2, \gamma_n\delta_i^2/2)$ where $\nu_n = \nu_0 + k$, $\gamma_n = \gamma_0 + k$,

$$\mu_i = \frac{\nu_0\mu_0 + k\bar{X}_i}{\nu_n}, \quad \bar{X}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_i} X_j,$$

and $\gamma_n \delta_i^2 = \gamma_0 \delta_0^2 + \sum_{j \in \mathcal{N}_i} (X_j - \bar{X}_i)^2 + k \nu_0 \nu_n^{-1} (\mu_0 - \bar{X}_i)^2$.

- **Step 3:** To compute the t -th Monte Carlo sample of $f(x)$, sample Dirichlet weights $\pi^{(t)} \sim \text{Dirichlet}(\alpha + 1, \dots, \alpha + 1)$ and neighborhood-specific parameters $(\eta_i^{(t)}, \sigma_i^{(t)2}) \sim \text{NIG}(\mu_i, \nu_n, \gamma_n/2, \gamma_n \delta_i^2/2)$ independently for $i = 1, \dots, n$, and set

$$f^{(t)}(x) = \sum_{i=1}^n \pi_i^{(t)} \phi(x; \eta_i^{(t)}, \sigma_i^{(t)2}). \quad (\text{A.28})$$

A.9 Inverse Wishart Parametrization

The parametrization of the inverse Wishart density defined on the set of all $p \times p$ matrices with real entries used in Chapter 2 is given as follows. Suppose $\gamma > p - 1$ and Ψ is a $p \times p$ positive definite matrix. If $\Sigma \sim \text{IW}_p(\gamma, \Psi)$, then Σ has the following density function:

$$g(\Sigma) = \begin{cases} \frac{|\Psi|^{\gamma/2}}{2^{\gamma p/2} \Gamma_p\left(\frac{\gamma}{2}\right)} |\Sigma|^{-(\gamma+p+1)/2} \text{etr}\left(-\frac{1}{2} \Psi \Sigma^{-1}\right) & \text{if } \Sigma \text{ is positive definite,} \\ 0 & \text{otherwise,} \end{cases}$$

where $\Gamma_p(\cdot)$ is the multivariate gamma function given by

$$\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma\left(a + \frac{1-j}{2}\right),$$

for $a \geq (p-1)/2$ and the function $\text{etr}(A) = \exp\{\text{tr}(A)\}$ for a square matrix A . When $p = 1$, the $\text{IW}_p(\gamma, \Psi)$ density is the same as the $\text{IG}(\gamma/2, \gamma \delta^2/2)$ density, where $\delta^2 = \Psi/\gamma$. The $\text{IW}_p(\gamma, \Psi)$ distribution has mean $\Psi/(\nu - p - 1)$ for $\nu > p + 1$ and mode $\Psi/(\nu + p + 1)$.

A.10 Univariate and Multivariate \mathcal{L}_1 Error Tables

Table A.1: Comparison of the methods in terms of \mathcal{L}_1 error in the univariate case. Number of test points and replications considered are $n_t = 500$ and $R = 20$, respectively.

Sample size	Estimator	CA	CW	DE	GS	IE	LN	LO	SB	SP	ST
200	NN-DM	0.20	0.31	0.19	0.12	0.36	0.20	0.13	0.16	0.30	0.31
	NN-DM (default)	0.21	0.37	0.17	0.12	0.34	0.20	0.14	0.17	0.31	0.32
	DP-MC	0.17	0.37	0.14	0.10	0.36	0.22	0.13	0.23	0.27	0.55
	KDE	-	0.37	0.16	0.12	-	0.18	0.11	0.18	-	0.52
	KNN	5.99	0.58	0.59	0.28	3.46	0.54	0.48	0.39	6.02	0.46
	DP-VB	0.20	0.35	0.15	0.08	0.53	0.25	0.11	0.11	0.44	0.57
	RD	-	0.35	0.13	0.12	-	0.16	0.11	0.16	-	0.53
	PTM	0.29	0.27	0.18	0.13	0.38	0.22	0.13	0.20	0.40	0.39
	LLDE	0.19	0.36	0.14	0.10	-	0.15	0.10	0.18	-	0.55
	OPT	0.32	0.36	0.28	0.17	0.55	0.21	0.02	0.18	0.75	0.52
	A-KDE	0.22	0.35	0.15	0.14	0.46	0.16	0.11	0.17	0.38	0.53
500	NN-DM	0.16	0.17	0.13	0.08	0.30	0.16	0.10	0.10	0.24	0.20
	NN-DM (default)	0.16	0.36	0.12	0.09	0.30	0.17	0.10	0.12	0.25	0.22
	DP-MC	0.11	0.35	0.10	0.08	0.27	0.18	0.09	0.13	0.22	0.53
	KDE	-	0.32	0.11	0.08	-	0.15	0.08	0.11	-	0.51
	KNN	3.62	0.47	0.48	0.27	3.39	0.40	0.30	0.31	5.64	0.35
	DP-VB	0.14	0.33	0.11	0.05	0.48	0.19	0.08	0.08	0.45	0.55
	RD	-	0.32	0.10	0.09	-	0.13	0.09	0.10	-	0.50
	PTM	0.24	0.19	0.14	0.10	0.32	0.19	0.11	0.14	0.32	0.30
	LLDE	0.17	0.35	0.11	0.08	-	0.15	0.08	0.14	-	0.53
	OPT	0.27	0.31	0.16	0.12	0.51	0.16	0.01	0.14	0.72	0.46
	A-KDE	0.16	0.32	0.10	0.10	0.40	0.13	0.10	0.10	0.27	0.52

Table A.2: Comparison of the methods in terms of \mathcal{L}_1 error in the multivariate case. Number of test points and replications considered are $n_t = 500$ and $R = 20$, respectively.

Density		MG					MST					MVC					MVG					SN					T				
Sample size	Dimension	2	3	4	6	2	3	4	6	2	3	4	6	2	3	4	6	2	3	4	6	2	3	4	6	2	3	4	6		
200	NN-DM	0.29	0.39	0.46	0.63	0.27	0.37	0.43	0.59	0.33	0.48	0.52	0.62	0.33	0.50	0.61	0.74	0.23	0.32	0.40	0.56	0.26	0.33	0.38	0.52						
	NN-DM (default)	0.29	0.40	0.47	0.65	0.28	0.38	0.45	0.61	0.37	0.49	0.61	0.80	0.38	0.50	0.62	0.75	0.24	0.34	0.42	0.58	0.26	0.33	0.40	0.53						
	DP-MC	0.20	0.36	0.62	0.67	0.22	0.42	0.55	0.68	0.31	0.46	0.51	0.60	0.34	0.52	0.61	0.73	0.16	0.18	0.25	0.32	0.21	0.26	0.34	0.44						
	KDE	0.28	0.52	0.79	1.29	0.27	0.55	0.72	1.21	-	-	-	-	0.43	0.77	0.90	1.09	0.22	0.50	0.78	1.26	0.26	0.53	0.74	1.21						
	KNN	1.93	3.82	4.80	18.83	7.28	8.22	9.25	11.05	4.92	7.31	15.24	20.5	3.30	3.65	4.75	5.54	2.21	5.16	8.37	10.04	2.97	6.37	10.22	17.54						
	DP-VB	0.29	0.38	0.41	0.50	0.24	0.36	0.44	0.59	0.48	0.85	1.28	1.69	0.45	0.58	0.71	0.86	0.17	0.23	0.31	0.52	0.19	0.29	0.34	0.46						
1000	LLDE	0.32	0.52	0.93	1.82	0.28	0.67	1.02	11.38	0.53	-	-	-	0.32	0.45	0.75	0.97	0.16	0.22	0.29	0.65	0.17	0.26	0.35	2.05						
	OPT	0.58	0.79	1.04	-	0.57	1.08	1.31	-	0.59	0.82	1.01	-	0.43	0.69	0.86	-	0.48	0.68	0.88	-	0.57	0.78	1.10	-						
	NN-DM	0.18	0.26	0.34	0.46	0.19	0.27	0.32	0.44	0.28	0.33	0.46	0.50	0.29	0.44	0.49	0.62	0.15	0.22	0.29	0.42	0.17	0.24	0.30	0.38						
	NN-DM (default)	0.22	0.30	0.37	0.48	0.21	0.29	0.33	0.44	0.31	0.41	0.52	0.66	0.36	0.48	0.55	0.68	0.22	0.29	0.36	0.45	0.22	0.28	0.33	0.39						
	DP-MC	0.08	0.39	0.57	0.58	0.11	0.18	0.21	0.47	0.26	0.39	0.48	0.54	0.21	0.38	0.51	0.64	0.06	0.07	0.10	0.15	0.09	0.15	0.17	0.28						
	KDE	0.16	0.32	0.52	0.96	0.16	0.35	0.53	1.04	-	-	-	-	0.33	0.66	0.73	0.93	0.13	0.32	0.53	1.05	0.14	0.32	0.52	0.90						
1000	KNN	0.92	2.62	4.01	15.28	5.96	6.48	7.04	9.63	4.68	6.29	13.7	17.04	2.01	2.59	3.88	4.09	1.89	4.39	6.84	8.15	2.37	5.30	9.66	13.28						
	DP-VB	0.25	0.29	0.33	0.36	0.15	0.24	0.25	0.45	0.42	0.74	0.82	0.91	0.38	0.54	0.61	0.77	0.10	0.12	0.15	0.20	0.10	0.15	0.18	0.31						
	LLDE	0.31	0.37	0.52	1.02	0.22	0.40	0.46	1.18	0.47	-	-	-	0.29	0.39	0.51	0.94	0.07	0.10	0.14	0.27	0.10	0.15	0.23	0.38						
	OPT	0.39	0.72	1.00	-	0.43	0.84	1.10	-	0.51	0.72	0.89	-	0.30	0.60	0.85	-	0.31	0.50	0.76	-	0.33	0.53	0.94	-						

Appendix B

Further details for Chapter 3

B.1 B-spline Functions

For an extensive overview of B-spline basis functions, we refer the reader to De Boor (1978). As a default, we choose cubic splines for modeling both the main and interaction effects throughout the paper; however, one could vary the degree of the B-splines as required. We now describe the choice of B-splines for modeling the interaction effects and main effects in Sections 3.3.2 and 3.3.4, respectively.

For the interaction function h_{uv} , we let h_{uv} satisfy $h_{uv}(x_u, 0) = 0$ for all $x_u \in [0, 1]$ and $h_{uv}(0, x_v) = 0$ for all $x_v \in [0, 1]$. To achieve this, we first construct the set of B-spline functions $s_{u0}(x_u), \dots, s_{um}(x_u)$ along the u -th dimension for $u = 1, \dots, p$, where s_{u0} is the intercept spline. The intercept spline s_{u0} the only B-spline basis function satisfying $s_{u0}(0) \neq 0$. Thus, to model the interaction effects, we simply ignore the intercept spline s_{u0} and only choose the B-spline functions $s_{uj}(\cdot)$ that satisfy $s_{uj}(0) = 0$.

In order to model the main effects, we proceed as follows. Suppose we have B-spline basis functions t_{u0}, \dots, t_{ud} , with t_{u0} being the intercept spline. If the main

effects satisfy the origin constraint, we simply ignore t_{u0} and let the basis functions be $b_{u,j}(x_u) = t_{uj}(x_u)$ for $u = 1, \dots, p$ and $j = 1, \dots, d$. If the main effects satisfy the integral constraint, we let $b_{u,j}(x_u) = t_{uj}(x_u) - \int_0^1 t_{uj}(x) dx$ for $u = 1, \dots, p$ and $j = 0, \dots, d$.

B.2 P-spline Priors

In order to ensure a function $g(x) = \sum_{j=1}^m s_j(x)\beta_j$ expressed as a linear combination of the B-splines s_1, \dots, s_m is sufficiently smooth, we follow the P-spline approach described in Lang and Brezger (2004). P-spline priors aim to promote smoothness by shrinking coefficients of adjacent splines towards each other.

We use a P-spline prior of order 1 along with a slight modification to make it a proper prior distribution. The order 1 prior described in Lang and Brezger (2004) is obtained by considering the hierarchical model $\beta_j | \beta_{j-1} \sim N(\beta_{j-1}, \tau^2)$ for $j = 2, \dots, m$, and then letting β_1 have an improper prior, given by

$$\pi(\beta_1) \propto 1.$$

However, we avoid the use of improper priors, and let $\beta_1 \sim N(0, \tau^2)$ as well. Under this modified prior specification, the joint distribution of $\beta = (\beta_1, \dots, \beta_m)^\top$ is given by

$$\pi(\beta) \propto \exp \left[-\frac{1}{2\tau^2} \left\{ \beta_1^2 + \sum_{j=2}^m (\beta_j - \beta_{j-1})^2 \right\} \right].$$

This can be rewritten as $\beta \sim N(0, \tau^2 \Sigma_0)$ for a positive-definite matrix Σ_0 whose entries are known.

B.3 Model Identifiability

We first state a result which ensures identifiability of the main and interaction effects in the assumed model.

Proposition 16. *Suppose the dose response function is given by*

$$H(\mathbf{x}) = \alpha + \sum_{j=1}^p f_j(x_j) + \sum_{1 \leq u < v \leq p} h_{uv}(x_u, x_v),$$

for $\mathbf{x} = (x_1, \dots, x_p)^\top \in [0, 1]^p$, $p \geq 2$. Assume the following conditions hold:

1. f_1, \dots, f_p together satisfy either (a) or (b), where

(a) **Origin Constraint:** $f_j(0) = 0$ for $j = 1, \dots, p$, or

(b) **Integral Constraint:** $\int_0^1 f_j(x_j) dx_j = 0$ for $j = 1, \dots, p$.

2. For all $1 \leq u < v \leq p$, $h_{uv}(x_u, 0) = 0$ and $h_{uv}(0, x_v) = 0$ for all $x_u, x_v \in [0, 1]$.

If two tuples $(\alpha, \{f_j\}_{j=1}^p, \{h_{uv}\}_{1 \leq u < v \leq p})$ and $(\tilde{\alpha}, \{\tilde{f}_j\}_{j=1}^p, \{\tilde{h}_{uv}\}_{1 \leq u < v \leq p})$ satisfy

$$\alpha + \sum_{j=1}^p f_j(x_j) + \sum_{1 \leq u < v \leq p} h_{uv}(x_u, x_v) = \tilde{\alpha} + \sum_{j=1}^p \tilde{f}_j(x_j) + \sum_{1 \leq u < v \leq p} \tilde{h}_{uv}(x_u, x_v),$$

for all $\mathbf{x} \in [0, 1]^p$, then $(\alpha, \{f_j\}_{j=1}^p, \{h_{uv}\}_{1 \leq u < v \leq p}) = (\tilde{\alpha}, \{\tilde{f}_j\}_{j=1}^p, \{\tilde{h}_{uv}\}_{1 \leq u < v \leq p})$.

Proof. We first prove the result for $p = 2$ and then generalize to higher dimensions, under both sets of constraints on the main effects. Suppose the following holds for all $x_1, x_2 \in [0, 1]$:

$$\alpha + f_1(x_1) + f_2(x_2) + h_{12}(x_1, x_2) = \tilde{\alpha} + \tilde{f}_1(x_1) + \tilde{f}_2(x_2) + \tilde{h}_{12}(x_1, x_2). \quad (\text{B.1})$$

Proof for origin constraint: We first let $x_1 = x_2 = 0$. Due to the constraints on the main effects and interactions, we obtain $\alpha = \tilde{\alpha}$. Next, we let $x_2 = 0$ for any x_1 . We then obtain $\alpha + f_1(x_1) = \tilde{\alpha} + \tilde{f}_1(x_1)$ for all x_1 , which leads us to $f_1 = \tilde{f}_1$.

Similarly, by letting $x_1 = 0$ for any x_2 , we obtain $f_2 = \tilde{f}_2$. Substituting these results into (B.1), we find $h_{12} = \tilde{h}_{12}$.

For $p \geq 3$, we fix $1 \leq u < v \leq p$. We first let $x_j = 0$ for all $j \notin \{u, v\}$, which implies

$$\alpha + f_u(x_u) + f_v(x_v) + h_{uv}(x_u, x_v) = \tilde{\alpha} + \tilde{f}_u(x_u) + \tilde{f}_v(x_v) + \tilde{h}_{uv}(x_u, x_v).$$

The result for $p = 2$ implies $f_u = \tilde{f}_u$, $f_v = \tilde{f}_v$, and $h_{uv} = \tilde{h}_{uv}$ for any $1 \leq u < v \leq p$.

Proof for integral constraint: We first let $x_2 = 0$, which implies $\alpha + f_1(x_1) + f_2(0) = \tilde{\alpha} + \tilde{f}_1(x_1) + \tilde{f}_2(0)$ for all $x_1 \in [0, 1]$. Integrating both sides with respect to x_1 and using the fact that $\int_0^1 f_1(x_1) dx_1 = \int_0^1 \tilde{f}_1(x_1) dx_1 = 0$, we obtain $\alpha + f_2(0) = \tilde{\alpha} + \tilde{f}_2(0)$, which when substituted into the previous equation, implies that $f_1 = \tilde{f}_1$. Similarly, letting $x_1 = 0$ leads us to $f_2 = \tilde{f}_2$, which implies $\alpha = \tilde{\alpha}$ since $\alpha + f_2(0) = \tilde{\alpha} + \tilde{f}_2(0)$. Substituting these results into (B.1) implies $h_{12} = \tilde{h}_{12}$.

For $p \geq 3$, we fix $1 \leq u < v \leq p$. We first let $x_j = 0$ for all $j \notin \{u, v\}$, which implies

$$\rho + f_u(x_u) + f_v(x_v) + h_{uv}(x_u, x_v) = \tilde{\rho} + \tilde{f}_u(x_u) + \tilde{f}_v(x_v) + \tilde{h}_{uv}(x_u, x_v),$$

where $\rho = \alpha + \sum_{j \notin \{u, v\}} f_j(0)$ and $\tilde{\rho} = \tilde{\alpha} + \sum_{j \notin \{u, v\}} \tilde{f}_j(0)$. Using the result for $p = 2$, this immediately implies that $f_u = \tilde{f}_u$ for all $u = 1, \dots, p$, $h_{uv} = \tilde{h}_{uv}$ for all $1 \leq u < v \leq p$, and $\rho = \tilde{\rho}$, which implies $\alpha = \tilde{\alpha}$.

□

B.4 Other Approaches

We also considered alternative Bayesian monotone spline formulations based on restricting the sign of the coefficients (Ramsay, 1988; Shively et al., 2009) instead of squaring unconstrained functions. However, squaring unconstrained functions led to superior estimation performance in simulations, particularly when the non-negative

function being estimated was close to 0. To see why, suppose $g(x) = \{\sum_{j=1}^m s_j(x)c_j\}^2$, where $\mathbf{s}(x) = (s_j(x))_{j=1}^m$ denotes m B-spline basis functions and $\mathbf{c} = (c_1, \dots, c_m)^\top$ denotes the vector of basis coefficients. If we assume $\mathbf{c} \sim N(0, C)$ for some covariance matrix C , then for each x , $g(x)/[\mathbf{s}(x)^\top C \mathbf{s}(x)] \sim \chi_1^2$. Since the χ_1^2 density function has an infinite spike at 0, the squared model provides superior shrinkage of the non-negative function towards 0 as compared to the model $g(x) = \mathbf{s}(x)^\top \mathbf{c}$, where we assume the truncated Gaussian prior on $\mathbf{c} \sim N(0, C)\mathbb{1}[c_1 \geq 0, \dots, c_m \geq 0]$.

As an alternative to decomposing P_{uv} and N_{uv} as products of univariate non-negative functions, we tried modeling P_{uv} and N_{uv} as squares of linear combinations of tensor products of B-splines (De Boor, 1978). That is, we let $P_{uv}(x_u, x_v) = \left\{ \sum_{j=1}^m \sum_{j'=1}^m s_{uj}(x_u) s_{vj'}(x_v) \beta_{uv,jj'} \right\}^2$ and $N_{uv}(x_u, x_v) = \left\{ \sum_{j=1}^m \sum_{j'=1}^m s_{uj}(x_u) s_{vj'}(x_v) \delta_{uv,jj'} \right\}^2$. Assuming P_{uv} and N_{uv} to be products of univariate non-negative functions is a special case of the above model, obtained by letting $\beta_{uv,jj'} = \theta_{uv,1j} \phi_{uv,1j'}$ and $\delta_{uv,jj'} = \theta_{uv,2j} \phi_{uv,2j'}$. This is equivalent to assuming the matrices $\mathcal{B}_{uv} = (\beta_{uv,jj'})_{1 \leq j, j' \leq m}$ and $\mathcal{D}_{uv} = (\delta_{uv,jj'})_{1 \leq j, j' \leq m}$ are rank 1. Due to the limited sample sizes and signal-to-noise ratios typical of environmental epidemiology, we found restricting the rank to 1 to be much more effective at identifying synergistic, antagonistic, or null interactions. Furthermore, the rank 1 assumption reduces the number of spline coefficients for the uv th interaction from $2m^2$ to $4m$, leading to substantial computational gains.

B.5 Posterior Sampling

Following Section 3.3.5, we now provide further details for sampling from the posterior under the proposed SAID approach. Let $w_{uv} = (\tau_{uv,1}, \tau_{uv,2})$. The conditional prior for the spline coefficients Ψ_{uv} given the penalty parameter κ_{uv} and variance

parameters w_{uv}, ν is given by

$$\pi(\Psi_{uv} \mid \kappa_{uv}, w_{uv}, \nu) \propto \pi_0(\Psi_{uv} \mid w_{uv}, \nu) \exp \left\{ -\kappa_{uv} \tilde{Q}(\Psi_{uv}) \right\},$$

where the unpenalized prior $\pi_0(\cdot \mid w_{uv}, \nu)$ is the product density given by:

$$\pi_0(\Psi_{uv} \mid w_{uv}, \nu) = \prod_{l=1}^2 \left\{ N(\theta_{uv,l} \mid 0, \nu^2 \tau_{uv,l}^2 \Sigma_0) N(\phi_{uv,l} \mid 0, \nu^2 \tau_{uv,l}^2 \Sigma_0) \right\}.$$

Let $g(\Psi_{uv}, \kappa_{uv}, w_{uv}, \nu) = \pi_0(\Psi_{uv} \mid w_{uv}, \nu) \exp \{ -\kappa_{uv} \tilde{Q}(\Psi_{uv}) \}$. Removing the proportionality, we get:

$$\pi(\Psi_{uv} \mid \kappa_{uv}, w_{uv}, \nu) = \frac{g(\Psi_{uv}, \kappa_{uv}, w_{uv}, \nu)}{Z(\kappa_{uv}, w_{uv}, \nu)} = \frac{\pi_0(\Psi_{uv} \mid w_{uv}, \nu) \exp \left\{ -\kappa_{uv} \tilde{Q}(\Psi_{uv}) \right\}}{Z(\kappa_{uv}, w_{uv}, \nu)},$$

where the normalizing constant is given by

$$Z(\kappa_{uv}, w_{uv}, \nu) = \int g(\Psi, \kappa_{uv}, w_{uv}, \nu) d\Psi.$$

With π as the prior for Ψ_{uv} , it is difficult to sample from the posterior of Ψ_{uv} using standard MCMC algorithms, since $Z(\kappa_{uv}, w_{uv}, \nu)$ is an intractable integral. To make MCMC sampling easier, we follow the variable augmentation technique described in Rao et al. (2016). Although Rao et al. (2016) develops their algorithm in the case of intractable likelihoods, we use the same principle to facilitate joint sampling of $\Theta_{uv} = (\Psi_{uv}^\top, \kappa_{uv}, w_{uv})^\top$ and then conditionally sampling ν given all the parameters $\{\Theta_{uv}\}_{1 \leq u < v \leq p}$, by augmenting new variables generated using rejection sampling.

We first remark that $g(\Psi_{uv}, \kappa_{uv}, w_{uv}, \nu) \leq M \pi_0(\Psi_{uv} \mid w_{uv}, \nu)$ with the constant $M = 1$. Following Rao et al. (2016), we introduce auxiliary variables $\mathcal{R}_{uv} = \{\mathcal{Y}_{uv,1}, \dots, \mathcal{Y}_{uv,|\mathcal{R}_{uv}|}\}$, which are considered as the rejected samples before Ψ_{uv} is accepted. The rejected proposals \mathcal{R}_{uv} are sampled using the following rejection sampling algorithm: given Θ_{uv} and ν , we independently sample $\mathcal{Y}_{uv,j} \sim \pi_0(\cdot \mid w_{uv}, \nu)$ until the first acceptance; accepting $\mathcal{Y}_{uv,j}$ with probability $g(\mathcal{Y}_{uv,j}, \kappa_{uv}, w_{uv}, \nu) / [M \pi_0(\mathcal{Y}_{uv,j} \mid$

$w_{uv}, \nu]$ = $\exp\{-\kappa_{uv}\tilde{\mathcal{Q}}(\mathcal{Y}_{uv,j})\}$. The conditional joint distribution of Ψ_{uv} and the rejected proposals \mathcal{R}_{uv} , given κ_{uv}, w_{uv} , and ν is

$$\begin{aligned} & \pi_1(\Psi_{uv}, \mathcal{R}_{uv} \mid \kappa_{uv}, w_{uv}, \nu) \\ &= g(\Psi_{uv}, \kappa_{uv}, w_{uv}, \nu) \prod_{j=1}^{|\mathcal{R}_{uv}|} \{\pi_0(\mathcal{Y}_{uv,j} \mid w_{uv}, \nu) - g(\mathcal{Y}_{uv,j}, \kappa_{uv}, w_{uv}, \nu)\} \\ &= g(\Psi_{uv}, \kappa_{uv}, w_{uv}, \nu) \prod_{j=1}^{|\mathcal{R}_{uv}|} \pi_0(\mathcal{Y}_{uv,j} \mid w_{uv}, \nu) \left[1 - \exp\{-\kappa_{uv}\tilde{\mathcal{Q}}(\mathcal{Y}_{uv,j})\}\right]. \end{aligned}$$

The above joint distribution allows us to compute the prior density of the uv -th interaction specific parameters Θ_{uv} , conditional on ν and the rejected proposals \mathcal{R}_{uv} :

$$p(\Psi_{uv}, \kappa_{uv}, w_{uv} \mid \mathcal{R}_{uv}, \nu) \propto \pi_1(\Psi_{uv}, \mathcal{R}_{uv} \mid \kappa_{uv}, w_{uv}, \nu) \pi_\kappa(\kappa_{uv}) \pi_w(w_{uv}). \quad (\text{B.2})$$

Here, π_κ is the standard lognormal density; that is, the density of $\log(W)$ when $W \sim N(0, 1)$, and $\pi_w(w_1, w_2) = C^+(w_1 \mid 0, 1) C^+(w_2 \mid 0, 1)$. From Section 3.3.5, the conditional likelihood for Θ_{uv} when sampling the uv -th interaction is given by $\mathcal{L}(\Theta_{uv}) = N(\Delta_{uv} \mid \mathbf{h}_{uv}, \sigma^2 \mathbb{I}_n)$. To sample all the parameters using MCMC, we adopt the following HMC-within-Gibbs strategy:

1. Sample the intercept, main effect, and covariate coefficient parameters as in Section 3.3.5.
2. For $1 \leq u < v \leq p$,
 - (a) Construct Δ_{uv} after estimating the intercept, main effect, covariate coefficients, and the rest of the interactions $\{h_{u'v'} : (u', v') \neq (u, v)\}$.
 - (b) **Sample \mathcal{R}_{uv} given Θ_{uv}, ν :** Given Θ_{uv} and ν , we independently sample $\mathcal{Y}_{uv,j} \sim \pi_0(\cdot \mid w_{uv}, \nu)$ for $j = 1, 2, \dots$ until a sample is accepted; we accept a sample $\mathcal{Y}_{uv,j}$ with probability $\exp\{-\kappa_{uv}\tilde{\mathcal{Q}}(\mathcal{Y}_{uv,j})\}$. Once we accept the

first sample, we discard the accepted sample and form the set of rejected samples $\mathcal{R}_{uv} = \{\mathcal{Y}_{uv,1}, \dots, \mathcal{Y}_{uv,|\mathcal{Y}_{uv}|}\}$.

(c) **Sample Θ_{uv} given \mathcal{R}_{uv} and ν :** Given $\mathcal{R}_{uv}, \nu, \Delta_{uv}$ and (B.2), we form

$$\Pi(\Psi_{uv}, \kappa_{uv}, w_{uv} \mid \Delta_{uv}, \mathcal{R}_{uv}, \nu) \propto N(\Delta_{uv} \mid \mathbf{h}_{uv}, \sigma^2 \mathbb{I}_n) p(\Psi_{uv}, \kappa_{uv}, w_{uv} \mid \mathcal{R}_{uv}, \nu),$$

and target sampling from Π using HMC.

3. **Sample ν given $\{\Theta_{uv}\}_{1 \leq u < v \leq p}$ and $\{\mathcal{R}_{uv}\}_{1 \leq u < v \leq p}$:** Sample ν from

$$p(\nu \mid \{\Theta_{uv}\}_{1 \leq u < v \leq p}, \{\mathcal{R}_{uv}\}_{1 \leq u < v \leq p}) \\ \propto \pi_\nu(\nu) \prod_{1 \leq u < v \leq p} \left\{ \pi_0(\Psi_{uv} \mid w_{uv}, \nu) \prod_{j=1}^{|\mathcal{R}_{uv}|} \pi_0(\mathcal{Y}_{uv,j} \mid w_{uv}, \nu) \right\}.$$

4. Sample σ^2 as in Section 3.3.5.

Since $\nu \sim C^+(0, 1)$, Step 2 above can be easily accomplished by introducing an auxiliary parameter W such that $\nu^2 \mid W \sim \text{IG}(1/2, 1/W)$ and $W \sim \text{IG}(1/2, 1)$ (Makalic and Schmidt, 2015) and then obtaining conditionally conjugate inverse-gamma updates for both ν^2 and W . We now repeat the above steps for a large number of iterations and discard a burn-in to obtain posterior samples of the unknown parameters.

B.6 Cutoff in Variable Selection

We provide case 1 and case 2 probabilities for classification using SAID in Section 3.4.3, by varying the integral cutoff in $\{0.005, 0.01, 0.05, 0.10\}$.

Table B.1: Comparison of different cutoffs when fitting SAID to assess variable selection accuracy for $p = 10$.

Cutoff	= 0.005	= 0.01	= 0.05	= 0.10
Synergistic Case 1 Probability	0.014	0.008	0.003	0.002
Synergistic Case2 Probability	0.001	0.001	0.010	0.120
Antagonistic Case1 Probability	0.008	0.005	0.002	0.001
Antagonistic Case2 Probability	0.010	0.001	0.010	0.200
Null Case1 Probability	0.001	0.001	0.010	0.160
Null Case2 Probability	0.024	0.014	0.005	0.001

B.7 Further Tables on Simulation Results

We provide further tables for the results described in Section 3.4.2. Tables B.2 and B.3 provide RMSE values for estimating the dose response surface H and the interaction h_{12} , respectively, under the QR setup. Tables B.4 and B.5 provide RMSE values for estimating the dose response surface H and the interaction h_{12} , respectively, under the MIS setup. Details on the QR and MIS setups are provided in Section 3.4.2. Let us define Cases 1, 2, 3, and 4 to be (γ_0, σ_0^2) equalling $(1, 0.1)$, $(1, 0.5)$, $(2, 0.1)$, and $(2, 0.5)$, respectively. Throughout, $n = 500$ and $p = 2$.

Table B.2: RMSE of competing methods in estimating H for scenario QR.

Case	SAID	BKMR	MixSelect	HierNet	FAMILY	PIE	RAMP
1	0.04	0.03	0.03	0.17	0.43	0.21	0.03
2	0.10	0.07	0.06	0.17	0.44	0.24	0.08
3	0.05	0.04	0.04	0.17	0.57	0.62	0.03
4	0.10	0.08	0.06	0.18	0.56	0.62	0.07

Table B.3: RMSE of competing methods in estimating h_{12} for scenario QR.

Case	SAID	MixSelect	HierNet	FAMILY	PIE	RAMP
1	0.08	0.08	0.33	0.31	0.05	0.04
2	0.12	0.11	0.29	0.31	0.10	0.17
3	0.12	0.08	0.37	0.62	0.05	0.04
4	0.17	0.09	0.35	0.62	0.11	0.10

Table B.4: RMSE of competing methods in estimating H for scenario MIS.

Case	SAID	BKMR	MixSelect	HierNet	PIE	RAMP
1	0.06	0.05	0.10	0.19	1.44	0.08
2	0.12	0.12	0.12	0.20	1.10	0.13
3	0.01	0.02	0.04	0.05	0.52	0.03
4	0.08	0.08	0.14	0.19	1.59	0.11

Table B.5: RMSE of competing methods in estimating h_{12} for scenario MIS.

Case	SAID	MixSelect	HierNet	PIE	RAMP
1	0.11	0.38	0.14	0.30	0.30
2	0.12	0.22	0.14	0.28	0.24
3	0.04	0.14	0.06	0.13	0.13
4	0.14	0.39	0.21	0.42	0.42

B.8 Creatinine Data Application

We provide plots for rest of the main effects and interaction effects not presented in Section 3.5.3. In Figure B.1, we plot the estimated main effects of the metals Antimony, Barium, Cobalt, Lead, Manganese, Strontium, Thallium, Tin, and Tungsten. In Figure B.2, we plot the estimated interaction surface between Molybdenum and Tin.

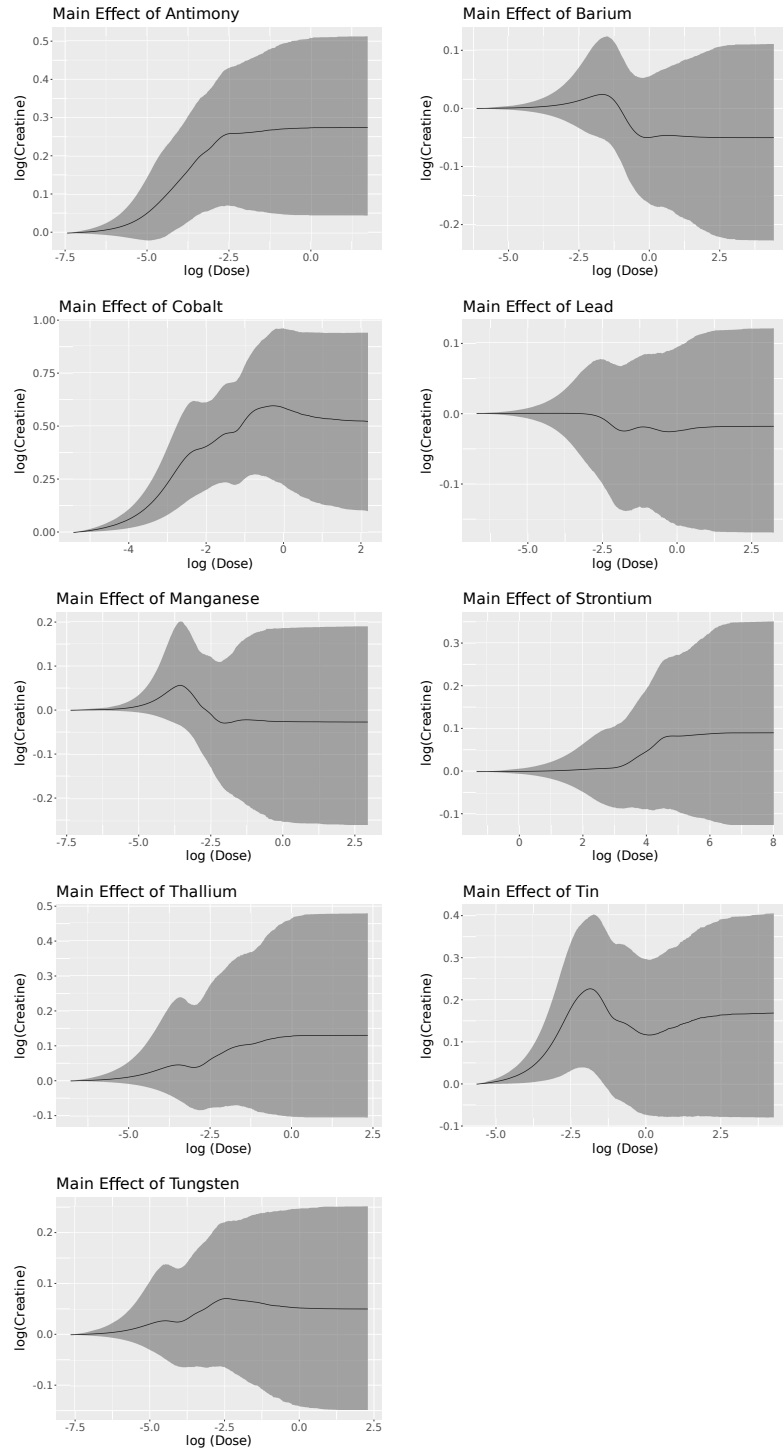


FIGURE B.1: Plots showing the main effects of dilution-adjusted Antimony, Barium, Cobalt, Lead, Manganese, Strontium, Thallium, Tin, and Tungsten on log dilution-adjusted Creatinine. Exposure levels are in log scale. Black line denotes posterior mean and shaded regions denote pointwise 95% posterior credible intervals.

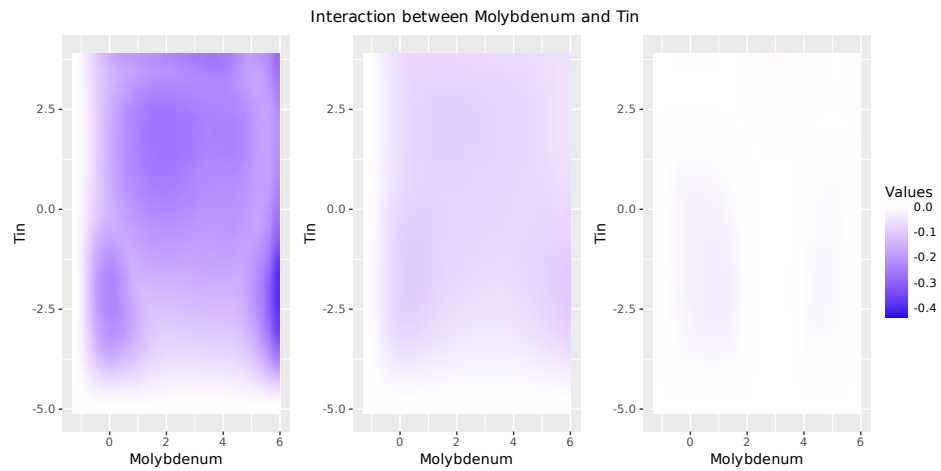


FIGURE B.2: Plot showing the interaction effect of dilution-adjusted Molybdenum and Tin on dilution-adjusted log Creatinine. Exposure levels are in log scale. Plot shows the pointwise 2.5% posterior credible surface, the posterior mean, and the 97.5% posterior credible surface from left to right.

Appendix C

Further details for Chapter 4

C.1 Proofs of Results

We assume the true data generating setup as discussed in Section 4.3. Suppose the singular value decomposition of \mathbf{Y} is given by

$$\mathbf{Y} = UDV^\top + U_\perp D_\perp V_\perp^\top$$

with $D \in \mathbb{R}^{k \times k}$ and let $A = YV/\sqrt{p} = UD/\sqrt{p}$. Let us rewrite \mathbf{Y} as $\mathbf{Y} = [y^{(1)}, \dots, y^{(p)}]$, where $y^{(j)}$ is the j th column of Y , $j = 1, \dots, p$. We let the singular value decomposition of $M_0\Lambda_0^\top = U_0D_0V_0^\top$.

Let us denote the expectation under the true data-generating distribution as E_0 , the induced pseudo-posterior distribution in (4.9) by $\tilde{\Pi}$, and let $\|A\|$ denote the operator norm of a matrix A . For two sequences a_n and b_n , we say $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $|a_n/b_n| \leq C$ for all sufficiently large n .

C.2 Proof of Theorem 8

C.2.1 Proof of part (a)

We first start with the pseudo-posterior contraction of L_C to L_0 . Let

$$G = \frac{n}{(n + \tau^{-2})^2} Y^\top U U^\top Y.$$

Recall that

$$\begin{aligned} L_C - L_0 &= G + B \odot (\tilde{\Lambda} \tilde{\Lambda}^\top - G) - \Lambda_0 \Lambda_0^\top \\ &= B \odot \left[\frac{Y^\top \widehat{M} \tilde{E}^\top + \tilde{E} \widehat{M}^\top Y}{n + \tau^{-2}} + \tilde{E} \tilde{E}^\top \right] + \frac{n}{(n + \tau^{-2})^2} Y^\top U U^\top Y - \Lambda_0 \Lambda_0^\top. \end{aligned}$$

In addition,

$$\begin{aligned} \frac{n}{(n + \tau^{-2})^2} Y^\top U U^\top Y - \Lambda_0 \Lambda_0^\top &= \frac{n}{(n + \tau^{-2})^2} (Y^\top Y - Y^\top U_\perp U_\perp^\top Y) - \Lambda_0 \Lambda_0^\top \\ &= \frac{-n}{(n + \tau^{-2})^2} Y^\top U_\perp U_\perp^\top Y + \frac{n}{(n + \tau^{-2})^2} (M_0 \Lambda_0^\top + E)^\top (M_0 \Lambda_0^\top + E) - \Lambda_0 \Lambda_0^\top \\ &= \frac{-n}{(n + \tau^{-2})^2} Y^\top U_\perp U_\perp^\top Y + \frac{n}{(n + \tau^{-2})^2} \Lambda_0 (M_0^\top M_0 - nI) \Lambda_0^\top + \left(\frac{n^2}{(n + \tau^{-2})^2} - 1 \right) \Lambda_0 \Lambda_0^\top \\ &\quad + \frac{n}{(n + \tau^{-2})^2} (\Lambda_0 M_0^\top E + E^\top M_0 \Lambda_0^\top + E^\top E). \end{aligned}$$

To develop an upper bound of $\|L_C - L_0\|$, we aim to develop an upper bound for the spectral norm of each term in the inequality above. We enumerate them as follows.

- (i) First, we develop a probabilistic upper bound for $\tilde{\sigma}_j^2$. Recall that $\|y^{(j)}\|_2^2 = \sum_{i=1}^n (y_i^{(j)})^2 \sim \chi^2(n) \cdot \text{Var}(y_i^{(j)})$. We apply the tail inequality bound of χ^2 distribution (c.f., Lemma 1 in Laurent and Massart (2000)),

$$\mathbb{P} \left(\max_{1 \leq j \leq p} \|y^{(j)}\|_2^2 \leq Cn \right) \geq 1 - o(1). \quad (\text{C.1})$$

Then for some constant $C > 0$,

$$\mathbb{P} \left(\frac{\gamma_0 \delta_0^2 + \|y^{(j)}\|_2^2 - \mu_j^\top \mu_j \cdot (n + \tau^{-2})}{2} \leq Cn \right) \geq 1 - o(1).$$

As

$$\tilde{\sigma}_j^2 \sim \text{IG} \left(\frac{\gamma_0 + n}{2}, \frac{\gamma_0 \delta_0^2 + \|y^{(j)}\|_2^2 - \mu_j^\top \mu_j \cdot (n + \tau^{-2})}{2} \right),$$

the tail bound of Gamma distribution based on Zhang and Zhou (2020) shows that there exists a constant $C' > 0$ such that

$$\tilde{\Pi} \left(\max_{1 \leq j \leq p} \tilde{\sigma}_j^2 \leq C' \right) = 1 - o_{P_0}(1). \quad (\text{C.2})$$

(ii) Since $\|\Lambda_0\| = \sqrt{p}$, we have

$$\left\| \left\{ \frac{n^2}{(n + \tau^{-2})^2} - 1 \right\} \Lambda_0 \Lambda_0^\top \right\| = \left| \frac{n^2 - (n + \tau^{-2})^2}{(n + \tau^{-2})^2} \right| \|\Lambda_0\|^2 \lesssim \frac{p}{n}. \quad (\text{C.3})$$

(iii) Recall M_0 is an n -by- k matrix with i.i.d. Gaussian entries. By Lemma 17,

$$\begin{aligned} \left\| \frac{n}{(n + \tau^{-2})^2} \Lambda_0 (M_0^\top M_0 - nI) \Lambda_0^\top \right\| &\lesssim \frac{1}{n} \|\Lambda_0\|^2 \|M_0^\top M_0 - nI\| \\ &\lesssim \frac{1}{n} \|\Lambda_0\|^2 \|M_0^\top M_0 - nI\| \\ &\lesssim \frac{p(k + \sqrt{nk})}{n} \end{aligned}$$

with probability at least $1 - o(1)$.

(iv) Next, we look at

$$\left\| \frac{n}{(n + \tau^{-2})^2} (\Lambda_0^\top M_0 E + E^\top M_0 \Lambda_0^\top + E^\top E) \right\| \lesssim \frac{1}{n} (2 \|\Lambda_0^\top M_0 E\| + \|E^\top E\|).$$

By Lemma 18, $\|E^\top E\| = \|E\|^2 \lesssim n + p$ with probability at least $1 - o(1)$. By Lemma 17, $\|\Lambda_0 M_0^\top E\| \lesssim \|\Lambda_0\| \cdot \|M_0^\top E\| \lesssim \sqrt{p}\sqrt{n} \cdot (\sqrt{n} + \sqrt{p})$ with probability at least $1 - o(1)$. Thus, with probability at least $1 - o(1)$,

$$\left\| \frac{n}{(n + \tau^{-2})^2} (\Lambda_0^\top M_0 E + E^\top M_0 \Lambda_0^\top + E^\top E) \right\| \lesssim \frac{n + p + p\sqrt{n} + n\sqrt{p} + \sqrt{np}}{n}.$$

(v) Note that U corresponds to the top k left singular vectors of Y . Therefore, $\|Y^\top U_\perp U_\perp^\top Y\| = \|Y^\top U_\perp\|^2 = \lambda_{k+1}^2(Y)$, where $\lambda_{k+1}^2(Y)$ is the $(k + 1)$ -st singular value of Y . By Eckart–Young Theorem (Lemma 20), we have

$$\|Y^\top U_\perp U_\perp^\top Y\| = \lambda_{k+1}^2(Y) = \min_{B: \text{rank}(B) \leq k} \|Y - B\|^2 \leq \|Y - M_0 \Lambda_0^\top\|^2 = \|E\|^2 \lesssim n + p.$$

Then,

$$\left\| \frac{-n}{(n + \tau^{-2})^2} Y^\top U_\perp U_\perp^\top Y \right\| \lesssim \frac{n + p}{n}.$$

(vi) Finally, we consider

$$\left\| \frac{Y^\top \widehat{M} \widetilde{E}^\top + \widetilde{E} \widehat{M}^\top Y}{n + \tau^{-2}} + \widetilde{E} \widetilde{E}^\top \right\| \lesssim \frac{2}{n} \|Y^\top \widehat{M} \widetilde{E}^\top\| + \|\widetilde{E} \widetilde{E}^\top\|.$$

By Lemma 18, we have

$$\begin{aligned} \|\widetilde{E} \widetilde{E}^\top\| &= \|\widetilde{E}\|^2 \lesssim \frac{p}{n} \max_{1 \leq j \leq p} \tilde{\sigma}_j^2 \lesssim \frac{p}{n} \\ \frac{1}{n} \|Y^\top \widehat{M} \widetilde{E}^\top\| &\lesssim \frac{1}{n} \|Y^\top \sqrt{n} U\| \cdot \|\widetilde{E}\| \lesssim \frac{1}{\sqrt{n}} \|Y\| \cdot \sqrt{\frac{p}{n} \max_{1 \leq j \leq p} \tilde{\sigma}_j} \\ &\lesssim \frac{\sqrt{np}}{\sqrt{n}} \cdot \sqrt{\frac{p}{n}} = \frac{p}{\sqrt{n}} \end{aligned}$$

with probability at least $1 - o_{P_0}(1)$. Lemma 21 implies that

$$\left\| B \odot \left[\frac{1}{(n + \tau^{-2})} \left[Y^\top \widehat{M} \widetilde{E}^\top + \widetilde{E} \widehat{M}^\top Y \right] + \widetilde{E} \widetilde{E}^\top \right] \right\| \lesssim \frac{p}{\sqrt{n}} \|B\|_\infty \sqrt{2k} \lesssim \frac{p}{\sqrt{n}},$$

with probability at least $1 - o_{P_0}(1)$, since the rank of $\tilde{\Lambda}\tilde{\Lambda}^\top - G$ is at most $2k$, and from Lemma 23, we obtain $\|B\|_\infty = O_{P_0}(1)$.

By combining the previous steps (i)-(vi), and $\|\Lambda_0\| \asymp \sqrt{p}$, we have proved the desired result.

C.2.2 Proof of part (b)

We now show the contraction result for the pseudo-posterior of $\Sigma_C = \tilde{\Sigma}$. We start out by observing that $\|\tilde{\Sigma} - \Sigma_0\| = \max_{1 \leq j \leq p} |D_j|$, where $D_j = \tilde{\sigma}_j^2 - \sigma_{0j}^2$ for $j = 1, \dots, p$.

Let us denote by $\kappa_j = \tilde{\sigma}_j^{-2}$, such that $\kappa_j | Y \stackrel{ind}{\sim} G(\gamma_n/2, \gamma_n \delta_j^2/2)$. Let

$$U_j = \frac{\gamma_n \delta_j^2}{2} \kappa_j - \frac{\gamma_n}{2}.$$

We can now express D_j as

$$D_j = \frac{(\delta_j^2 - \sigma_{0j}^2) - \frac{2U_j}{\gamma_n} \sigma_{0j}^2}{1 + \frac{2U_j}{\gamma_n}}.$$

Consider any a_n such that $a_n = o(1)$ and $\sqrt{na_n} \rightarrow \infty$ as $n \rightarrow \infty$. Using Lemma 6, we have $\max_{1 \leq j \leq p} |U_j|/\gamma_n \lesssim a_n$ with probability at least $1 - O(p \exp(-na_n^2)) = 1 - o(1)$. This immediately implies that $\min_{1 \leq j \leq p} |1 + (2U_j/\gamma_n)| \gtrsim 1/2$ with probability at least $1 - o(1)$. Therefore,

$$\max_{1 \leq j \leq p} |D_j| \lesssim 2 \max_{1 \leq j \leq p} |\delta_j^2 - \sigma_{0j}^2| + 4\sigma_{0j}^2 \max_{1 \leq j \leq p} \frac{|U_j|}{\gamma_n},$$

with probability at least $1 - o(1)$.

Let $Q_j = \delta_j^2 - \sigma_{0j}^2$. Using Lemma 24, we can represent

$$Q_j = \frac{\sigma_{0j}^2}{n} \left[\frac{Z_j}{\sigma_{0j}^2} - (n - k) \right] - \frac{k\sigma_{0j}^2}{n} + F_j,$$

where $Z_j/\sigma_{0j}^2 \sim \chi_{n-k}^2$ and

$$\max_{1 \leq j \leq p} |F_j| \lesssim \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}}$$

with probability at least $1 - o(1)$. Using Zhang and Zhou (2020), we obtain

$$\max_{1 \leq j \leq p} \left| \frac{Z_j}{\sigma_{0j}^2} - (n - k) \right| \lesssim na_n$$

with probability at least $1 - O(p \exp(-na_n^2/4)) = 1 - o(1)$, since $\log p = o(na_n^2)$.

Thus, with probability at least $1 - o(1)$, we get

$$\max_{1 \leq j \leq p} |Q_j| \lesssim a_n + \frac{1}{\sqrt{p}}.$$

As $\max_{1 \leq j \leq p} |U_j|/\gamma_n \lesssim a_n$ with probability at least $1 - o(1)$, we obtain

$$\max_{1 \leq j \leq p} |D_j| \lesssim a_n + \frac{1}{\sqrt{p}}$$

with probability at least $1 - o(1)$, for all $j = 1, \dots, p$ or equivalently,

$$\|\tilde{\Sigma} - \Sigma_0\| \lesssim a_n + \frac{1}{\sqrt{p}}$$

with probability at least $1 - o_{P_0}(1)$. This proves the result.

C.2.3 Proof of part (c)

Under the conditions of Theorem 1, we have $\|\Psi_0\| \asymp \|L_0\| \asymp \sqrt{p}$. By the triangle inequality, $\|\Psi_C - \Psi_0\| \leq \|L_C - L_0\| + \|\Sigma_C - \Sigma_0\|$. Thus, we have

$$\frac{\|\Psi_C - \Psi_0\|}{\|\Psi_0\|} \lesssim \frac{1}{\sqrt{p}} + \frac{1}{\sqrt{n}} + o\left(\frac{1}{p}\right) \lesssim \frac{1}{\sqrt{p}} + \frac{1}{\sqrt{n}}$$

with probability at least $1 - o_{P_0}(1)$, proving the result.

C.2.4 Relevant lemmas and their proofs

Lemma 17 (Spectral norm bound of a random matrix; Corollary 5.35 in Vershynin (2010)). *Let E be a p -by- n matrix with i.i.d. standard Gaussian entries. Then for every $t > 0$, one has*

$$\mathbb{P}(\|E\| \leq \sqrt{p} + \sqrt{n} + t) \geq 1 - 2 \exp(-t^2/2).$$

$$\mathbb{E}\|EE^\top - nI\| \leq (p^{1/2} + n^{1/2} + t)^2 - n.$$

Lemma 18 (Spectral norm of random matrix with heteroskedastic entries). *Suppose $E \in \mathbb{R}^{p \times n}$ with independent entries such that $E_{ij} = g_{ij}b_{ij}$ for $g_{ij} \stackrel{iid}{\sim} N(0, 1)$ and $\{b_{ij} : i \leq j\}$ fixed scalars. Let $\sigma_1 = \max_i \sqrt{\sum_j b_{ij}^2}$, $\sigma_2 = \max_j \sqrt{\sum_i b_{ij}^2}$, and $\sigma_* = \max_{i,j} |b_{ij}|$. Then for every $\epsilon \in (0, 1/2)$ there exists a $c'_\epsilon > 0$ such that for all $t \geq 0$,*

$$\mathbb{P}(\|E\| \geq (1 + \epsilon)(\sigma_1 + \sigma_2) + t) \leq (p \wedge n) \exp\{-t^2/(c'_\epsilon \sigma_*^2)\}.$$

Proof of Lemma 18. The proof follows from Corollary 3.9 in Bandeira and Van Handel (2016).

Lemma 19 (Spectral norm bound for product matrices; Theorem 1.1 in Vershynin (2011)). *Let E be a p -by- n matrix with i.i.d. standard Gaussian entries; $W \in \mathbb{R}^{n \times q}$ is a fixed matrix such that $\|W\| \leq 1$. Then,*

$$\mathbb{E}\|EW\| \leq C(\sqrt{p} + \sqrt{q}).$$

Lemma 20 (Eckart–Young Theorem). *For any matrix Y ,*

$$\lambda_{k+1}(Y) = \min_{\text{rank}(M) \leq k} \|Y - M\|.$$

Lemma 21. *For any two matrices C_1, C_2 , let $C_1 \odot C_2$ denote the Hadamard product of C_1 and C_2 . If C_2 has rank r , then*

$$\|C_1 \odot C_2\| \leq \|C_1\|_\infty \|C_2\| \sqrt{r}.$$

Proof: For a matrix A , let $\|A\|_F$ denote the Frobenius norm of A , given by

$$\|A\|_F = \sqrt{\sum_i s_i^2(A)},$$

where $\{s_i(A)\}_i$ denote the singular values of A . It is immediate that $\|A\|_F \leq \sqrt{\text{rank}(A)} \max_i s_i(A) = \sqrt{\text{rank}(A)} \|A\|$. The result follows by noting the following chain of inequalities:

$$\|C_1 \odot C_2\| \leq \|C_1 \odot C_2\|_F \leq \|C_1\|_\infty \|C_2\|_F \leq \|C_1\|_\infty \|C_2\| \sqrt{r}.$$

Lemma 22. *For all $1 \leq u \leq v \leq p$, we have as $n, p \rightarrow \infty$,*

$$\frac{1}{n} y^{(u)\top} U U^\top y^{(v)} \xrightarrow{P_0} \lambda_{0u}^\top \lambda_{0v}$$

and

$$\frac{1}{n} \|(\mathbb{I}_n - U U^\top) y^{(u)}\|_2^2 \xrightarrow{P_0} \sigma_{0u}^2.$$

Proof: We first note the identity

$$\begin{aligned} & \frac{1}{n} y^{(u)\top} U U^\top y^{(v)} - \lambda_{0u}^\top \lambda_{0v} \\ &= \lambda_{0u}^\top \left(\frac{M_0^\top M_0}{n} - \mathbb{I}_k \right) \lambda_{0v} + \frac{1}{n} (\lambda_{0u}^\top M_0^\top \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top \epsilon^{(u)}) \\ &+ \frac{1}{n} \epsilon^{(u)\top} U_0 U_0^\top \epsilon^{(v)} + \frac{1}{n} y^{(u)\top} (U U^\top - U_0 U_0^\top) y^{(v)}. \end{aligned}$$

The first two terms tend to 0 in probability by the weak law of large numbers, the third term is seen to be $O_{P_0}(1/n)$, and the fourth term is seen to be $O_{P_0}(n^{-1} + p^{-1})$ since $\|y^{(u)}\|_2 \lesssim \sqrt{n}$ with probability at least $1 - o(1)$ and by Proposition 7. Therefore, one has the desired result for any $1 \leq u \leq v \leq p$.

To see the second result, we first note that $\|y^{(u)}\|_2^2 \sim (\|\lambda_{0u}\|_2 + \sigma_{0u}^2) \chi_n^2$. An application of the above result along with the weak law of large numbers provides that this quantity converges in probability to $\|\lambda_{0u}\|_2 + \sigma_{0u}^2 - \|\lambda_{0u}\|_2^2 = \sigma_{0u}^2$ for any $1 \leq u \leq p$.

Lemma 23. *If B represents the matrix of coverage-correction coefficients, then $\|B\|_\infty = O_{P_0}(1)$ under the assumptions made on the data generating model.*

Proof: Consider u, v such that $u \neq v$ and $\|\lambda_{0u}\|_2 \|\lambda_{0v}\|_2 > 0$. Following from Assumption 4, there exists $W_1 > 0$ such that $\Lambda_{0,uv} > W_1$ for all u, v such that $\|\lambda_{0u}\|_2 \|\lambda_{0v}\|_2 > 0$. Let $W_2 = \min_u \sigma_{0u}^2 > 0$ from Assumption 5. Following from Lemma 22, we have for sufficiently large n : (1) $\|\mu_u\|_2^2 > W_1/2$ and $s_u^2 > W_2/2$, (2) $|\lambda_{0u}^\top \lambda_{0v}| \leq \|\lambda_{0u}\|_2 \|\lambda_{0v}\|_2 \leq k \|\Lambda_0\|_\infty^2$. This implies that

$$b_{uv}^2 \leq 1 + \frac{4k^2 \|\Lambda_0\|_\infty^4}{W_1 W_2} < \infty.$$

One can analogously show a similar bound on b_{uu} with the bound independent of u . This immediately implies $\|B\|_\infty < \infty$.

Lemma 24. *Suppose $\log p = o(n)$. For each $j = 1, \dots, p$, we have*

$$\delta_j^2 = \frac{Z_j}{n} + F_j,$$

where we have $Z_j/\sigma_{0j}^2 \sim \chi_{n-k}^2$ and

$$\max_{1 \leq j \leq p} |F_j| \lesssim \frac{1}{n} + \frac{1}{p}$$

with probability at least $1 - o(1)$.

Proof: Suppose the SVD of $X_0 = M_0 \Lambda_0^\top = U_0 D_0 V_0^\top$. Let $P_{U_0} = U_0 (U_0^\top U_0)^{-1} U_0^\top = U_0 U_0^\top$ denote the projection matrix onto the column space of U_0 . We can express δ_j^2 as

$$\delta_j^2 = \frac{Z_j}{n} + F_j,$$

where we have

$$Z_j = y^{(j)\top} (\mathbb{I}_n - P_{U_0}) y^{(j)},$$

$$F_j = \frac{\gamma_0 \delta_0^2}{\gamma_n} + \frac{1}{n} y^{(j)\top} \left(P_{U_0} - \frac{\widehat{M} \widehat{M}^\top}{n + \tau^{-2}} \right) y^{(j)} - \frac{\gamma_0}{n \gamma_n} y^{(j)\top} \left(\mathbb{I}_n - \frac{\widehat{M} \widehat{M}^\top}{n + \tau^{-2}} \right) y^{(j)}.$$

Letting $Y = M_0\Lambda_0^\top + E$ where $E = [\epsilon^{(1)} \mid \dots \mid \epsilon^{(p)}]$, we have $(\mathbb{I}_n - P_{U_0})Y = (\mathbb{I}_n - P_{U_0})E$, implying $(\mathbb{I}_n - P_{U_0})y^{(j)} = (\mathbb{I}_n - P_{U_0})\epsilon^{(j)} \sim N_n(0, \sigma_{0j}^2(\mathbb{I}_n - P_{U_0}))$. Since $\mathbb{I}_n - P_{U_0}$ is idempotent, we have

$$\frac{Z_j}{\sigma_{0j}^2} = \frac{\|(\mathbb{I}_n - P_{U_0})y^{(j)}\|^2}{\sigma_{0j}^2} \sim \chi_{\text{tr}(\mathbb{I}_n - P_{U_0})}^2 \equiv \chi_{n-k}^2.$$

We now obtain the stated probabilistic upper bound on $|F_j|$. We observe

$$\max_{1 \leq j \leq p} |F_j| \leq \frac{\gamma_0 \delta_0^2}{\gamma_n} + \frac{1}{n} \left\| P_{U_0} - \frac{\widehat{M}\widehat{M}^\top}{n + \tau^{-2}} \right\| \max_{1 \leq j \leq p} \|y^{(j)}\|_2^2 + \frac{\gamma_0}{n\gamma_n} \left\| \mathbb{I}_n - \frac{\widehat{M}\widehat{M}^\top}{n + \tau^{-2}} \right\| \max_{1 \leq j \leq p} \|y^{(j)}\|_2^2.$$

We first observe that $\max_{1 \leq j \leq p} \|y^{(j)}\|_2^2 \lesssim n$ with probability at least $1 - o(1)$, since $\log p = o(n)$. We also remark that $\widehat{M}\widehat{M}^\top = nUU^\top$, where $Y = UDV^\top + U_\perp D_\perp V_\perp^\top$ is the SVD of Y . We start with

$$\begin{aligned} G_1 &:= \left\| P_{U_0} - \frac{\widehat{M}\widehat{M}^\top}{n + \tau^{-2}} \right\| \\ &\leq \|U_0U_0^\top - UU^\top\| + \left\| UU^\top - \frac{nUU^\top}{n + \tau^{-2}} \right\| \\ &\lesssim \frac{1}{n} + \frac{1}{p} + \frac{1}{n} \lesssim \frac{1}{n} + \frac{1}{p}, \end{aligned}$$

with probability at least $1 - o(1)$, using Proposition 7. Next, we work with

$$\begin{aligned} G_2 &:= \left\| \mathbb{I}_n - \frac{\widehat{M}\widehat{M}^\top}{n + \tau^{-2}} \right\| \\ &\leq \left\| \mathbb{I}_n - \frac{\widehat{M}\widehat{M}^\top}{n} \right\| + \frac{\tau^{-2}}{n^2} \|\widehat{M}\widehat{M}^\top\| \\ &= \|\mathbb{I}_n - UU^\top\| + \frac{\tau^{-2}}{n} \|UU^\top\| \\ &\lesssim 1. \end{aligned}$$

Combining, we have with probability at least $1 - o(1)$,

$$\max_{1 \leq j \leq p} |F_j| \lesssim \frac{1}{n} + \frac{1}{p}.$$

Lemma 25. *Let $V_n \sim G(\gamma_n/2, 1)$ such that $\gamma_n \asymp n$ and let $U_n = V_n - (\gamma_n/2)$. For any a_n satisfying $a_n \rightarrow 0$ and $\sqrt{n}a_n \rightarrow \infty$,*

$$P(|U_n| \geq \gamma_n a_n) \lesssim \exp(-na_n^2).$$

Proof: Immediate from Theorem 5 in Zhang and Zhou (2020).

Lemma 26. *Let $E \in \mathbb{R}^{n \times k}$ be a matrix of iid $N(0, 1)$ entries with $n > k$ and let s_{max} and $s_{min}(M_0)$ be the smallest and largest singular values of M_0 , respectively. Then,*

$$\begin{aligned} s_{max}(M_0) &\equiv \|M_0\| \lesssim \sqrt{n} + \sqrt{k} \\ s_{min}(M_0) &\gtrsim \sqrt{n} - \sqrt{k-1}, \end{aligned}$$

with probability at least $1 - o(1)$.

In particular, if $k = o(n)$, we have $\|M_0\| \asymp \sqrt{n}$ with probability at least $1 - o(1)$.

Proof: Refer to Sections 1.1 and 1.3 in Vershynin (2011).

Proof of Proposition 7:

We start with $\|X_0^\top UU^\top X_0 - X_0^\top U_0 U_0^\top X_0\|$. Theorem 2 in Luo et al. (2021) implies

$$\|X_0^\top UU^\top X_0 - X_0^\top U_0 U_0^\top X_0\| = \|X_0^\top UU^\top X_0 - X_0^\top X_0\| = \|(\mathbb{I}_n - UU^\top)X_0\|^2 \leq 4\|E\|^2.$$

Corollary 3.9 in Bandeira and Van Handel (2016) implies $\|E\| \lesssim (\sigma_{sum} + \sqrt{n}\sigma_{max})$ with probability at least $1 - o(1)$. The result is obtained upon observing that

$$\|X_0^\top UU^\top X_0 - X_0^\top U_0 U_0^\top X_0\| \geq s_{min}(X_0)^2 \|UU^\top - U_0 U_0^\top\|.$$

Using Lemma 7, we have $s_{\min}(M_0) \asymp \|M_0\| \asymp \sqrt{n}$ with probability at least $1 - o(1)$, since $k = o(n)$. Under the assumption $s_{\min}(\Lambda_0) \asymp \|\Lambda_0\| \asymp \sqrt{p}$, then $s_{\min}(X_0) \asymp \|X_0\| \asymp \sqrt{np}$, as for any two matrices A, B , we have $s_{\min}(AB) \geq s_{\min}(A)s_{\min}(B)$. Thus, with probability at least $1 - o(1)$,

$$\|UU^\top - U_0U_0^\top\| \lesssim \frac{(\sqrt{n} + \sqrt{p})^2}{np} \lesssim \frac{1}{n} + \frac{1}{p}.$$

C.3 Proof of Theorem 10

We first assume $u \neq v$, so that $\Psi_{C,uv} = L_{C,uv}$, where $L_{C,uv}$ is the uv th element of L_C . Let $\mu_j = (n + \tau^{-2})^{-1} \widehat{M}^\top y^{(j)}$ for $j = 1, \dots, p$. According to the pseudo-posterior generation mechanism after adjusting for coverage,

$$L_{C,uv} = \mu_u^\top \mu_v + b_{uv} \left[\frac{\tilde{\sigma}_v \mu_u^\top \tilde{e}_v + \tilde{\sigma}_u \mu_v^\top \tilde{e}_u}{\sqrt{n + \tau^{-2}}} + \frac{\tilde{\sigma}_u \tilde{\sigma}_v \tilde{e}_u^\top \tilde{e}_v}{n + \tau^{-2}} \right],$$

for $\tilde{e}_u, \tilde{e}_v \stackrel{\text{ind}}{\sim} N_k(0, \mathbb{I}_k)$. We first observe that with high probability,

$$\sqrt{n} \left(\frac{\tilde{\sigma}_u \tilde{\sigma}_v \tilde{e}_u^\top \tilde{e}_v}{n + \tau^{-2}} \right) \lesssim \frac{1}{\sqrt{n}}.$$

Furthermore, for $R_{uv} \sim N(0, 1)$ such that R_{uv} independent of Y ,

$$\begin{aligned} \tilde{\sigma}_v \mu_u^\top \tilde{e}_v + \tilde{\sigma}_u \mu_v^\top \tilde{e}_u &= (\tilde{\sigma}_v^2 \|\mu_u\|_2^2 + \tilde{\sigma}_u^2 \|\mu_v\|_2^2)^{1/2} R_{uv} \\ &= l_{0,uv} R_{uv} + [(\tilde{\sigma}_v^2 \|\mu_u\|_2^2 + \tilde{\sigma}_u^2 \|\mu_v\|_2^2)^{1/2} - l_{0,uv}] R_{uv}, \end{aligned}$$

where $l_{0,uv}^2 = \sigma_{0v}^2 \|\lambda_{0u}\|_2^2 + \sigma_{0u}^2 \|\lambda_{0v}\|_2^2$. Let $d_{uv} = [(\tilde{\sigma}_v^2 \|\mu_u\|_2^2 + \tilde{\sigma}_u^2 \|\mu_v\|_2^2)^{1/2} - l_{0,uv}]$. We can now express

$$\sqrt{n}(L_{C,uv} - \mu_u^\top \mu_v) = \frac{\sqrt{n}\beta_{uv}}{\sqrt{n + \tau^{-2}}} l_{0,uv} R_{uv} + T_{uv},$$

where

$$T_{uv} = \sqrt{\frac{n}{n + \tau^{-2}}} l_{0,uv} R_{uv} (b_{uv} - \beta_{uv}) + \sqrt{\frac{n}{n + \tau^{-2}}} d_{uv} b_{uv} R_{uv} + \sqrt{n} \left(\frac{\tilde{\sigma}_u \tilde{\sigma}_v \tilde{e}_u^\top \tilde{e}_v}{n + \tau^{-2}} \right) b_{uv},$$

where the coefficients β_{uv} are given by

$$\begin{aligned}\beta_{uv} &= \sqrt{1 + \frac{(\lambda_{0u}^\top \lambda_{0v})^2 + \|\lambda_{0u}\|_2^2 \|\lambda_{0v}\|_2^2}{\sigma_{0u}^2 \|\lambda_{0v}\|_2^2 + \sigma_{0v}^2 \|\lambda_{0u}\|_2^2}}, \quad \text{if } u \neq v \text{ and } \|\lambda_{0u}\|_2 \|\lambda_{0v}\|_2 > 0, \\ &= \sqrt{1 + \frac{\|\lambda_{0u}\|_2^2}{2\sigma_{0u}^2}}, \quad \text{if } u = v \text{ and } \|\lambda_{0u}\|_2 \|\lambda_{0v}\|_2 > 0, \\ &= 1, \quad \text{otherwise.}\end{aligned}$$

From Lemma 22 and the continuous mapping theorem, one has $b_{uv} \xrightarrow{P_0} \beta_{uv}$ as $n, p \rightarrow \infty$. Due to pseudo-posterior concentration of $\tilde{\sigma}_j^2$ to σ_{0j}^2 and convergence in probability of $\|\mu_j\|_2^2$ to $\|\lambda_{0j}\|_2^2$, combined with $\hat{b}_{uv} \xrightarrow{P_0} b_{uv}$, we have pseudo-posterior concentration of T_{uv} around 0 as $n, p \rightarrow \infty$. Since $\beta_{uv} l_{0,uv} = \mathcal{S}_{0,uv}$, we have from Lemma 29

$$\sup_x \left| \tilde{\Pi} \{ \sqrt{n}(L_{C,uv} - \mu_u^\top \mu_v) \leq \mathcal{S}_{0,uv} x \} - \Phi(x) \right| \xrightarrow{P_0} 0.$$

The above result is a Bernstein-von Mises theorem guaranteeing correct coverage of the pseudo-credible intervals asymptotically, since $\sqrt{n}(\mu_u^\top \mu_v - \lambda_{0u}^\top \lambda_{0v}) \implies N(0, \mathcal{S}_{0,uv}^2)$ as $n, p \rightarrow \infty$.

For $u = v$, we first observe that $\sqrt{n}(\Psi_{C,uv} - \Psi_{0,uv}) = \sqrt{n}(L_{C,uv} - \lambda_{0u}^\top \lambda_{0v}) + \sqrt{n}(\tilde{\sigma}_j^2 - \sigma_{0j}^2)$. Using the decomposition in Lemma 24, the bound on $\max_j |F_j|$ can be improved using Lemma 9 to yield

$$\max_{1 \leq j \leq p} |F_j| \lesssim \frac{1}{n} + \sqrt{\frac{\log(n+p)}{np}}.$$

Finally, approximating the Gamma distribution using a Gaussian distribution with the central limit theorem and using Lemma 28 yields the desired result.

C.4 Proof of Theorem 11

We first assume $u \neq v$. Let

$$S_{uv} = \frac{n}{(n + \tau^{-2})^2} y^{(u)\top} U U^\top y^{(v)} - \lambda_{0u}^\top \lambda_{0v}.$$

Let $b_n = n/(n + \tau^{-2})^2$. Since $y^{(u)} = M_0 \lambda_{0u} + \epsilon^{(u)}$, we can decompose S_{uv} as

$$\begin{aligned} S_{uv} &= \lambda_{0u}^\top (b_n M_0^\top U U^\top M_0 - \mathbb{I}_k) \lambda_{0v} \\ &\quad + b_n (\lambda_{0u}^\top M_0^\top U U^\top \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top U U^\top \epsilon^{(u)} + \epsilon^{(u)\top} U U^\top \epsilon^{(v)}). \end{aligned}$$

Let $b_n = (1/n) + \Delta_n$, where $|\Delta_n| \asymp 1/n^2$. We first break up S_{uv} into two parts $S_{uv} = D_{uv} + R_{uv}$, where

$$\begin{aligned} D_{uv} &= \lambda_{0u}^\top \left(\frac{1}{n} M_0^\top U_0 U_0^\top M_0 - \mathbb{I}_k \right) \lambda_{0v} + \frac{1}{n} (\lambda_{0u}^\top M_0^\top U_0 U_0^\top \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top U_0 U_0^\top \epsilon^{(u)}) \\ R_{uv} &= \Delta_n \lambda_{0u}^\top M_0^\top U_0 U_0^\top M_0 \lambda_{0v} + b_n \lambda_{0u}^\top M_0^\top (U U^\top - U_0 U_0^\top) M_0 \lambda_{0v} \\ &\quad + \Delta_n (\lambda_{0u}^\top M_0^\top U_0 U_0^\top \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top U_0 U_0^\top \epsilon^{(u)}) \\ &\quad + b_n (\lambda_{0u}^\top M_0^\top (U U^\top - U_0 U_0^\top) \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top (U U^\top - U_0 U_0^\top) \epsilon^{(u)}) \\ &\quad + b_n \epsilon^{(u)\top} U U^\top \epsilon^{(v)}. \end{aligned}$$

We first deal with R_{uv} and show that $\sqrt{n} R_{uv} = o_{P_0}(1)$. We use the results in Lemma 27 as required. First consider bounding $\sqrt{n} b_n \epsilon^{(u)\top} U U^\top \epsilon^{(v)}$ as follows:

$$\begin{aligned} | \sqrt{n} b_n \epsilon^{(u)\top} U U^\top \epsilon^{(v)} | &\leq \sqrt{n} b_n \| E^\top U U^\top E \|_\infty \\ &\leq \sqrt{n} b_n \| E^\top U_0 U_0^\top E \|_\infty + \sqrt{n} b_n \| E^\top (U U^\top - U_0 U_0^\top) E \|_\infty. \end{aligned}$$

The first term is $\sqrt{n} b_n \max_{u,v} | \epsilon^{(u)\top} U_0 U_0^\top \epsilon^{(v)} | = O_{P_0}(\sqrt{n} b_n) = O_{P_0}(1/\sqrt{n})$, since $U_0^\top \epsilon^{(u)} \sim N(0, \sigma_{0u}^2 \mathbb{I}_k)$ and k is finite. The second term is handled by observing that

$\|\epsilon^{(u)}\|_2 \lesssim \sqrt{n}$ with probability at least $1 - o(1)$, so that

$$\begin{aligned} \sqrt{nb_n} \max_{u,v} |\epsilon^{(u)\top} (UU^\top - U_0U_0^\top) \epsilon^{(v)}| &\leq \sqrt{nb_n} \|UU^\top - U_0U_0^\top\|_\infty \|\epsilon^{(u)}\|_2 \|\epsilon^{(v)}\|_2 \\ &\lesssim n \left(\frac{1}{n^2} + \frac{\log(n+p)}{np} \right) \sqrt{nb_n} \\ &= o_{P_0}(1). \end{aligned}$$

We now consider the rest of the terms.

(i) We first observe that

$$\begin{aligned} \sqrt{n}\Delta_n |\lambda_{0u}^\top M_0^\top U_0 U_0^\top M_0 \lambda_{0v}| &= \sqrt{n}\Delta_n |\lambda_{0u} M_0^\top M_0 \lambda_{0v}| \\ &\leq \sqrt{n}\Delta_n \|M_0^\top M_0\| \|\lambda_{0u}\|_2 \|\lambda_{0v}\|_2 \lesssim \frac{1}{\sqrt{n}}. \end{aligned}$$

(ii) Next, we start from $\|M_0^\top\|_{2,\infty} \lesssim \sqrt{n}$. Thus,

$$\begin{aligned} \sqrt{nb_n} \|\lambda_{0u}^\top M_0^\top (UU^\top - U_0U_0^\top) M_0 \lambda_{0v}\|_\infty &\lesssim \sqrt{nb_n} \|M_0^\top (UU^\top - U_0U_0^\top) M_0\|_\infty \\ &\leq \sqrt{nb_n} \|M_0^\top\|_{2,\infty}^2 \|UU^\top - U_0U_0^\top\|_\infty \\ &= o_{P_0}(1). \end{aligned}$$

(iii) We now consider $\sqrt{n}\Delta_n |\lambda_{0u}^\top U_0 U_0^\top M_0^\top \epsilon^{(v)} + \lambda_{0v}^\top U_0 U_0^\top M_0^\top \epsilon^{(u)}|$

$$= \sqrt{n}\Delta_n |\lambda_{0u}^\top M_0^\top \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top \epsilon^{(u)}| \lesssim \frac{2}{n^2} (\sqrt{n})^3 \lesssim \frac{1}{\sqrt{n}}.$$

(iv) Finally, we bound

$$\begin{aligned} &\sqrt{nb_n} |\lambda_{0u}^\top M_0^\top (UU^\top - U_0U_0^\top) \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top (UU^\top - U_0U_0^\top) \epsilon^{(u)}| \\ &\lesssim \frac{2}{\sqrt{n}} \|M_0^\top\|_{2,\infty} \sqrt{n} \|UU^\top - U_0U_0^\top\|_{2,\infty} \\ &= o_{P_0}(1). \end{aligned}$$

Putting everything together implies

$$\sqrt{n} |R_{uv}| = o_{P_0}(1).$$

We now look at $\sqrt{n}D_{uv}$. By the central limit theorem, we first observe that

$$\begin{aligned}\sqrt{n}\lambda_{0u}^\top \left(\frac{1}{n}M_0^\top U_0 U_0^\top M_0 - \mathbb{I}_k \right) \lambda_{0v} &= \sqrt{n}\lambda_{0u}^\top \left(\frac{1}{n}M_0^\top M_0 - \mathbb{I}_k \right) \lambda_{0v} \\ &= \sqrt{n}(\bar{V}_n - \lambda_{0u}^\top \lambda_{0v}) \\ &\xrightarrow{D} N(0, \xi_{0,uv}^2),\end{aligned}$$

as $n \rightarrow \infty$, where $\bar{V}_n = (1/n) \sum_{i=1}^n V_i$, with $V_i := (\lambda_{0u}^\top \eta_i)(\lambda_{0v}^\top \eta_i)$ for $i = 1, \dots, n$, and

$$\xi_{0,uv}^2 = \text{var}[(\lambda_{0u}^\top \eta_1)(\lambda_{0v}^\top \eta_1)] = (\lambda_{0u}^\top \lambda_{0v})^2 + \|\lambda_{0u}\|_2^2 \|\lambda_{0v}\|_2^2.$$

We first observe that

$$\sqrt{n} \left[\frac{1}{n}(\lambda_{0u}^\top M_0^\top U_0 U_0^\top \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top U_0 U_0^\top \epsilon^{(u)}) \right] = \sqrt{n} \left[\frac{1}{n}(\lambda_{0u}^\top M_0^\top \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top \epsilon^{(u)}) \right]$$

Let

$$l_{uv}^2(M_0) = \sigma_{0v}^2 \lambda_{0u}^\top \frac{M_0^\top M_0}{n} \lambda_{0u} + \sigma_{0u}^2 \lambda_{0v}^\top \frac{M_0^\top M_0}{n} \lambda_{0v}.$$

Since $\epsilon^{(u)}$ and $\epsilon^{(v)}$ are independent for $u \neq v$, we have

$$\sqrt{n} \left[\frac{1}{n}(\lambda_{0u}^\top M_0^\top \epsilon^{(v)} + \lambda_{0v}^\top M_0^\top \epsilon^{(u)}) \right] \mid M_0 \sim N(0, l_{uv}^2(M_0)) \stackrel{d}{=} l_{uv}(M_0) Z_{uv},$$

where $Z_{uv} \sim N(0, 1)$ and Z_{uv} is independent of M_0 . Let

$$l_{0,uv}^2 = \sigma_{0v}^2 \|\lambda_{0u}\|_2^2 + \sigma_{0u}^2 \|\lambda_{0v}\|_2^2.$$

Then,

$$\begin{aligned}\sqrt{n}D_{uv} &= \sqrt{n}\lambda_{0u}^\top \left(\frac{1}{n}M_0^\top U_0 U_0^\top M_0 - \mathbb{I}_k \right) \lambda_{0v} + l_{uv}(M_0) Z_{uv} \\ &= \sqrt{n}\lambda_{0u}^\top \left(\frac{1}{n}M_0^\top M_0 - \mathbb{I}_k \right) \lambda_{0v} + l_{0,uv} Z_{uv} + (l_{uv}(M_0) - l_{0,uv}) Z_{uv}.\end{aligned}$$

Since $\|M_0\| \asymp \sqrt{n}$, we have $|l_{uv}(M_0) + l_{0,uv}| = O_{P_0}(1)$. Thus, the third term can be handled by observing

$$\begin{aligned} (l_{uv}(M_0) - l_{0,uv})Z_{uv} &= \frac{l_{uv}^2(M_0) - l_{0,uv}^2}{l_{uv}(M_0) + l_{0,uv}} Z_{uv} \\ &= O_{P_0}\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

as

$$\left\| \frac{M_0^\top M_0}{n} - \mathbb{I}_k \right\| \lesssim \frac{1}{\sqrt{n}}.$$

Since Z_{uv} is independent of M_0 , Lemma 28 immediately implies that

$$\sqrt{n}D_{uv} \stackrel{d}{=} N(0, \xi_{0,uv}^2 + l_{0,uv}^2) + o_{P_0}(1).$$

Let $\mathcal{S}_{0,uv}^2 = l_{0,uv}^2 + \xi_{0,uv}^2$. Putting all the previous results back together, we obtain

$$\sqrt{n}S_{uv} \xrightarrow{d} N(0, \mathcal{S}_{0,uv}^2),$$

by Slutsky's theorem.

When $u = v$, we first observe that δ_u^2 is approximately independent of $\|\mu_u\|_2^2$ by first replacing UU^\top by $U_0U_0^\top$, as $U_0U_0^\top(\mathbb{I}_n - U_0U_0^\top) = \mathbb{O}_{n \times n}$, and then using Proposition 9 to argue the remainder terms are small. Using the central limit theorem on the χ_{n-k}^2 distribution, we obtain the desired result.

C.5 Related Lemmas for Theorems 10 and 11

Lemma 27. *The following inequalities hold for general matrices A, B, X whenever dimensions conform: (1) $\|AB\|_{2,\infty} \leq \|A\|_{2,\infty}\|B\|$. (2) $\|AB^\top\|_\infty \leq \|A\|_{2,\infty}\|B\|_{2,\infty}$. (3) $\|A\|_\infty \leq \|A\|_{2,\infty}$. (4) $\|X^\top AY\|_\infty \leq \|A\|_\infty\|X^\top\|_{2,\infty}\|Y^\top\|_{2,\infty}$. (5) $|x^\top Ay| \leq \|A\|_\infty\|x\|_2\|y\|_2$.*

Proof: (1) Let $A = [a_1, \dots, a_m]^\top$; this implies $\|AB\|_{2,\infty} = \max_i \|a_i^\top B\|_2 \leq (\max_i \|a_i\|)\|B\| = \|A\|_{2,\infty}\|B\|$.

(2) Let $B = [b_1, \dots, b_m]^\top$. Then, $\|AB\|_\infty = \max_{uv} |a_u^\top b_v| \leq \max_{uv} \|a\|_2 \|b\|_2$ by Cauchy-Schwarz inequality, which implies the result.

(3) Follows trivially from their definitions.

(4) Let x_i be the i th column of X and y_j be the j th column of Y . Then, $\|X^\top AY\|_\infty = \max_{ij} |x_i^\top Ay_j| \leq \|A\|_\infty \max_{ij} |x_i^\top y_j| \leq \|A\|_\infty \max_{ij} \|x\|_2 \|y\|_2$ by the Cauchy-Schwarz inequality, which proves the result.

(5) Follows directly from bounding the quadratic form using the triangle inequality.

Proof of Proposition 9:

We start with Theorem 4.2 in Chen et al. (2021). By letting $B = \sqrt{\log n}$, we can relax the assumption of bounded errors and assume the errors are Gaussian. Since $M_0 \Lambda_0^\top = U_0 D_0 V_0^\top$ with U_0 having orthogonal columns and $\|(M_0^\top M_0/n) - \mathbb{I}_k\| \rightarrow 0$, the incoherence parameter μ satisfies $\mu = O(1)$. Assumption 5 ensures that the error variances have a common upper bound. Upon using the dilation trick, we obtain the following bound:

$$\|U \operatorname{sgn}(H_U) - U_0\|_\infty \lesssim \frac{1}{n} + \sqrt{\frac{\log(n+p)}{np}},$$

with probability at least $1 - o(1)$. We next observe that

$$\|UU^\top - U_0U_0^\top\|_\infty \leq 2\sqrt{k}\|UR - U_0\|_{2,\infty}$$

for any rotation matrix R . This proves the statement by letting $R = \operatorname{sgn}(H_U)$.

Lemma 28. *Suppose $X_n = Y_n + Z_n$, where $Y_n \xrightarrow{d} N(0, \sigma_1^2)$ and $Z_n \sim N(0, \sigma_2^2)$, with Y_n independent of Z_n . Then, $X_n \xrightarrow{d} N(0, \sigma_1^2 + \sigma_2^2)$.*

Proof: It is immediate from observing the characteristic function of X_n and taking limits.

Lemma 29. *Suppose in the setup of Theorem 10, we have $T_n = Z_n + Y_n$ for random variables T_n, Z_n, Y_n such that the pseudo-posterior of Y_n concentrates around 0 as $n, p \rightarrow \infty$ and $Z_n \sim N(0, 1)$. Then as $n, p \rightarrow \infty$,*

$$\sup_x |\tilde{\Pi}(T_n \leq x) - \Phi(x)| \xrightarrow{P_0} 0.$$

Proof: Fix $\epsilon > 0$. The concentration property of Y_n implies that $\tilde{\Pi}(|Y_n| > \epsilon) \xrightarrow{P_0} 0$ as $n, p \rightarrow \infty$. Fix $x \in \mathbb{R}$. Using the triangle inequality, one obtains

$$|\tilde{\Pi}(T_n \leq x) - \Phi(x)| \leq |\tilde{\Pi}(T_n \leq x, |Y_n| \leq \epsilon) - \Phi(x)| + \tilde{\Pi}(T_n \leq x, |Y_n| > \epsilon),$$

with the second term bounded by $\tilde{\Pi}(|Y_n| > \epsilon) \xrightarrow{P_0} 0$. The first term may be decomposed as

$$|\tilde{\Pi}(T_n \leq x, |Y_n| \leq \epsilon) - \Phi(x)| = S_1 + S_2,$$

where $S_1 = |\tilde{\Pi}(T_n \leq x, |Y_n| \leq \epsilon) - \Phi(x)|\mathbb{1}(Y_n > 0)$ and $S_2 = |\tilde{\Pi}(T_n \leq x, |Y_n| \leq \epsilon) - \Phi(x)|\mathbb{1}(Y_n \leq 0)$. We first consider S_1 :

$$\begin{aligned} S_1 &= |\tilde{\Pi}(T_n \leq x, |Y_n| \leq \epsilon) - \Phi(x)|\mathbb{1}(Y_n > 0) \\ &\leq |\tilde{\Pi}(Z_n \leq x - Y_n, Y_n \leq \epsilon) - \tilde{\Pi}(Z_n \leq x, Y_n \leq \epsilon)|\mathbb{1}(Y_n > 0) \\ &\quad + \tilde{\Pi}(Z_n \leq x, Y_n > \epsilon)\mathbb{1}(Y_n > 0) \\ &\leq \tilde{\Pi}(x - Y_n \leq Z_n \leq x, Y_n \leq \epsilon)\mathbb{1}(Y_n > 0) + \tilde{\Pi}(|Y_n| > \epsilon). \end{aligned}$$

The second term goes to 0 in P_0 -probability, while the first term can be bounded by $\Phi(x) - \Phi(x - \epsilon) \leq L\epsilon$ for some $L > 0$, as Φ is Lipschitz continuous. Similarly, $S_2 \lesssim L\epsilon$ with probability at least $1 - o(1)$. Combining, we have

$$\sup_x |\tilde{\Pi}(T_n \leq x) - \Phi(x)| \lesssim 2L\epsilon$$

with probability at least $1 - o(1)$, for any fixed ϵ . This shows the result.

Bibliography

- Abramson, I. S. (1982), “On bandwidth variation in kernel estimates—a square root law,” *The Annals of Statistics*, 10, 1217–1223.
- Attias, H. (2013), “Inferring parameters and structure of latent variable models by variational Bayes,” *arXiv preprint arXiv:1301.6676*.
- Avalos-Pacheco, A., Rossell, D., and Savage, R. S. (2022), “Heterogeneous large datasets integration using Bayesian factor regression,” *Bayesian Analysis*, 17, 33–66.
- Azzalini, A. (2005), “The skew-normal distribution and related multivariate families,” *Scandinavian Journal of Statistics*, 32, 159–188.
- Bandeira, A. S. and Van Handel, R. (2016), “Sharp nonasymptotic bounds on the norm of random matrices with independent entries,” *Annals of Probability*, 44, 2479–2506.
- Barr, D. B., Wilder, L. C., Caudill, S. P., Gonzalez, A. J., Needham, L. L., and Pirkle, J. L. (2005), “Urinary creatinine concentrations in the US population: implications for urinary biologic monitoring measurements,” *Environmental Health Perspectives*, 113, 192–200.
- Baxmann, A. C., Ahmed, M. S., Marques, N. C., Menon, V. B., Pereira, A. B., Kirsztajn, G. M., and Heilberg, I. P. (2008), “Influence of muscle mass and physical activity on serum and urinary creatinine and serum cystatin C,” *Clinical Journal of the American Society of Nephrology*, 3, 348–354.
- Betancourt, M. and Girolami, M. (2015), “Hamiltonian Monte Carlo for hierarchical models,” *Current Trends in Bayesian Methodology with Applications*, 79, 2–4.
- Bhattacharya, A. and Dunson, D. (2011), “Sparse Bayesian infinite factor models,” *Biometrika*, 98, 291–306.
- Biau, G. and Devroye, L. (2015), *Lectures on the Nearest Neighbor Method*, Springer.
- Bickel, P. J. and Levina, E. (2008), “Regularized estimation of large covariance matrices,” *Annals of Statistics*, 36, 199–227.

- Bien, J. and Tibshirani, R. J. (2011), “Sparse estimation of a covariance matrix,” *Biometrika*, 98, 807–820.
- Bien, J., Taylor, J., and Tibshirani, R. (2013), “A lasso for hierarchical interactions,” *Annals of Statistics*, 41, 1111.
- Blei, D. M. and Jordan, M. I. (2006), “Variational inference for Dirichlet process mixtures,” *Bayesian Analysis*, 1, 121–143.
- Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. (2015), “Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures,” *Biostatistics*, 16, 493–508.
- Bolfarine, H., Carvalho, C. M., Lopes, H. F., and Murray, J. S. (2022), “Decoupling Shrinkage and Selection in Gaussian Linear Factor Analysis,” *Bayesian Analysis*, 1, 1–23.
- Bowman, A. W. (1984), “An alternative method of cross-validation for the smoothing of density estimates,” *Biometrika*, 71, 353–360.
- Breiman, L., Meisel, W., and Purcell, E. (1977), “Variable kernel estimates of multivariate densities,” *Technometrics*, 19, 135–144.
- Brezger, A. and Lang, S. (2006), “Generalized structured additive regression based on Bayesian P-splines,” *Computational Statistics & Data Analysis*, 50, 967–991.
- Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009), “Handling sparsity via the horseshoe,” in *Artificial Intelligence and Statistics*, pp. 73–80, PMLR.
- Chen, Y., Chi, Y., Fan, J., Ma, C., et al. (2021), “Spectral methods for data science: A statistical perspective,” *Foundations and Trends® in Machine Learning*, 14, 566–806.
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), “BART: Bayesian additive regression trees,” *The Annals of Applied Statistics*, 4, 266–298.
- Czarnota, J., Gennings, C., and Wheeler, D. C. (2015), “Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk,” *Cancer Informatics*, 14, CIN-S17295.
- De Boor, C. (1978), *A Practical Guide to Splines*, vol. 27, Springer-Verlag New York.
- De Vito, R., Bellio, R., Trippa, L., and Parmigiani, G. (2021), “Bayesian multi-study factor analysis for high-throughput biological data,” *The Annals of Applied Statistics*, 15, 1723–1741.

- Devroye, L. and Györfi, L. (1985), *Nonparametric Density Estimation: the L_1 view*, Wiley Series in Probability and Statistics.
- Dinari, O., Yu, A., Freifeld, O., and Fisher, J. (2019), “Distributed MCMC Inference in Dirichlet process mixture models Using Julia,” in *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pp. 518–525.
- Duong, T. (2020), *ks: Kernel Smoothing*, R package version 1.11.7.
- Evans, D. (2008), “A law of large numbers for nearest neighbour statistics,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 464, 3175–3192.
- Evans, D., Jones, A. J., and Schmidt, W. M. (2002), “Asymptotic moments of near-neighbour distance distributions,” *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 458, 2839–2849.
- Fan, J., Lou, Z., and Yu, M. (2023), “Are Latent Factor Regression and Sparse Regression Adequate?” *Journal of the American Statistical Association*, pp. 1–77.
- Ferguson, T. S. (1973), “A Bayesian analysis of some nonparametric problems,” *The Annals of Statistics*, 1, 209–230.
- Ferrari, F. and Dunson, D. B. (2020), “Identifying main effects and interactions among exposures using Gaussian processes,” *The Annals of Applied Statistics*, 14, 1743.
- Ferrari, F. and Dunson, D. B. (2021), “Bayesian factor analysis for inference on interactions,” *Journal of the American Statistical Association*, 116, 1521–1532.
- Ferraro, P. M., Costanzi, S., Naticchia, A., Sturniolo, A., and Gambaro, G. (2010), “Low level exposure to cadmium increases the risk of chronic kidney disease: analysis of the NHANES 1999-2006,” *BMC Public Health*, 10, 1–8.
- Forbes, G. and Bruining, G. J. (1976), “Urinary creatinine excretion and lean body mass,” *The American Journal of Clinical Nutrition*, 29, 1359–1366.
- Frühwirth-Schnatter, S. (2023), “Generalized cumulative shrinkage process priors with applications to sparse Bayesian factor analysis,” *Philosophical Transactions of the Royal Society A*, 381, 20220148.
- Gelman, A., Meng, X.-L., and Stern, H. (1996), “Posterior predictive assessment of model fitness via realized discrepancies,” *Statistica Sinica*, pp. 733–760.
- Ghosal, S. and van der Vaart, A. (2007), “Posterior convergence rates of Dirichlet mixtures at smooth densities,” *The Annals of Statistics*, 35, 697–723.

- Ghosal, S. and van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference*, vol. 44, Cambridge University Press.
- Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999), “Posterior consistency of Dirichlet mixtures in density estimation,” *The Annals of Statistics*, 27, 143–158.
- Gneiting, T. and Raftery, A. E. (2007), “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, 102, 359–378.
- Golub, G. H. and van Loan, C. F. (1996), *Matrix Computations*, John Hopkins University Press, 3rd edn.
- Hahn, P. R., Martin, R., and Walker, S. G. (2018), “On recursive Bayesian predictive distributions,” *Journal of the American Statistical Association*, 113, 1085–1093.
- Hall, P. (1987), “On Kullback-Leibler loss and density estimation,” *The Annals of Statistics*, 15, 1491–1519.
- Hao, N., Feng, Y., and Zhang, H. H. (2018), “Model selection for high-dimensional quadratic regression via regularization,” *Journal of the American Statistical Association*, 113, 615–625.
- Haris, A., Witten, D., and Simon, N. (2016), “Convex modeling of interactions with strong heredity,” *Journal of Computational and Graphical Statistics*, 25, 981–1004.
- Hays, S. M., Aylward, L. L., and Blount, B. C. (2015), “Variation in urinary flow rates according to demographic characteristics and body mass index in NHANES: potential confounding of associations between health outcomes and urinary biomarker concentrations,” *Environmental Health Perspectives*, 123, 293–300.
- Hjort, N. L. and Jones, M. C. (1996), “Locally parametric nonparametric density estimation,” *The Annals of Statistics*, pp. 1619–1647.
- Hoffman, M. D., Gelman, A., et al. (2014), “The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.” *J. Mach. Learn. Res.*, 15, 1593–1623.
- Hughes, M. C. and Sudderth, E. B. (2014), “Bnpy: Reliable and scalable variational inference for Bayesian nonparametric models,” in *Proceedings of the NIPS Probabilistic Programming Workshop, Montreal, QC, Canada*, pp. 8–13.
- J. Ross, G. and Markwick, D. (2019), *dirichletprocess: Build Dirichlet Process Objects for Bayesian Modelling*, R package version 0.3.1.
- Jain, R. (2016), “Associated complex of urine creatinine, serum creatinine, and chronic kidney disease,” *Epidemiology (Sunnyvale)*, 6, 2161–1165.

- James, G. D., Sealey, J. E., Alderman, M., Ljungman, S., Mueller, F. B., Pecker, M. S., and Laragh, J. H. (1988), “A longitudinal study of urinary creatinine and creatinine clearance in normal subjects: race, sex, and age differences,” *American Journal of Hypertension*, 1, 124–131.
- Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011), “DPpackage: Bayesian Semi- and Nonparametric Modeling in R,” *Journal of Statistical Software*, 40, 1–30.
- Jeng, P.-H., Huang, T.-R., Wang, C.-C., and Chen, W.-L. (2021), “Clinical relevance of urine flow rate and exposure to polycyclic aromatic hydrocarbons,” *International Journal of Environmental Research and Public Health*, 18, 5372.
- Joubert, B. R., Kioumourtzoglou, M.-A., Chamberlain, T., Chen, H. Y., Gennings, C., Turyk, M. E., Miranda, M. L., Webster, T. F., Ensor, K. B., Dunson, D. B., et al. (2022), “Powering Research through Innovative Methods for Mixtures in Epidemiology (PRIME) Program: Novel and Expanded Statistical Methods,” *International Journal of Environmental Research and Public Health*, 19, 1378.
- Kashani, K., Rosner, M. H., and Ostermann, M. (2020), “Creatinine: from physiology to clinical application,” *European Journal of Internal Medicine*, 72, 9–14.
- Kass, R. E. and Raftery, A. E. (1995), “Bayes factors,” *Journal of the American Statistical Association*, 90, 773–795.
- Keith, M., Jameson, A., Van Straten, W., Bailes, M., Johnston, S., Kramer, M., Possenti, A., Bates, S., Bhat, N., Burgay, M., et al. (2010), “The High Time Resolution Universe Pulsar Survey I, System configuration and initial discoveries,” *Monthly Notices of the Royal Astronomical Society*, 409, 619–627.
- Kim, N. H., Hyun, Y. Y., Lee, K.-B., Chang, Y., Rhu, S., Oh, K.-H., and Ahn, C. (2015), “Environmental heavy metal exposure and chronic kidney disease in the general population,” *Journal of Korean Medical Science*, 30, 272–277.
- Kurihara, K., Welling, M., and Vlassis, N. (2006), “Accelerated Variational Dirichlet Process Mixtures,” in *Advances in Neural Information Processing Systems*, eds. B. Schölkopf, J. Platt, and T. Hoffman, vol. 19, MIT Press.
- Lang, S. and Brezger, A. (2004), “Bayesian P-splines,” *Journal of Computational and Graphical Statistics*, 13, 183–212.
- Laurent, B. and Massart, P. (2000), “Adaptive estimation of a quadratic functional by model selection,” *Annals of Statistics*, pp. 1302–1338.
- Lavine, M. (1992), “Some aspects of Polya tree distributions for statistical modelling,” *Annals of Statistics*, 20, 1222–1235.

- Lavine, M. (1994), “More aspects of Polya tree distributions for statistical modelling,” *The Annals of Statistics*, 22, 1161–1176.
- Loader, C. (2006), *Local regression and likelihood*, Springer Science & Business Media.
- Loader, C. R. (1996), “Local likelihood density estimation,” *The Annals of Statistics*, 24, 1602–1618.
- Loftsgaarden, D. O. and Quesenberry, C. P. (1965), “A nonparametric estimate of a multivariate density function,” *The Annals of Mathematical Statistics*, 36, 1049–1051.
- Lopes, H. F. and West, M. (2004), “Bayesian model assessment in factor analysis,” *Statistica Sinica*, pp. 41–67.
- Lorimer, D. R. and Kramer, M. (2012), *Handbook of Pulsar Astronomy*, Cambridge University Press.
- Luo, J. and Hendryx, M. (2020), “Metal mixtures and kidney function: An application of machine learning to NHANES data,” *Environmental Research*, 191, 110126.
- Luo, Y., Han, R., and Zhang, A. R. (2021), “A Schatten-q low-rank matrix perturbation analysis via perturbation projection error bound,” *Linear Algebra and its Applications*, 630, 225–240.
- Lyon, R. J. (2016), “Why are pulsars hard to find?” Ph.D. thesis, The University of Manchester (United Kingdom).
- Ma, H. and Li, J. (2019), “A True $O(n \log n)$ Algorithm for the All-k-Nearest-Neighbors Problem,” in *International Conference on Combinatorial Optimization and Applications*, pp. 362–374, Springer.
- Ma, Y. and Liu, J. S. (2022), “On Posterior Consistency of Bayesian Factor Models in High Dimensions,” *Bayesian Analysis*, 17, 901–929.
- Mack, Y. and Rosenblatt, M. (1979), “Multivariate k-nearest neighbor density estimates,” *Journal of Multivariate Analysis*, 9, 1–15.
- Makalic, E. and Schmidt, D. F. (2015), “A simple sampler for the horseshoe estimator,” *IEEE Signal Processing Letters*, 23, 179–182.
- Man, A. X. and Culpepper, S. A. (2022), “A mode-jumping algorithm for Bayesian factor analysis,” *Journal of the American Statistical Association*, 117, 277–290.

- Middleton, D. R., Watts, M. J., Lark, R. M., Milne, C. J., and Polya, D. A. (2016), “Assessing urinary flow rate, creatinine, osmolality and other hydration adjustment methods for urinary biomonitoring using NHANES arsenic, iodine, lead and cadmium data,” *Environmental Health*, 15, 1–13.
- Mildenberger, T. and Weinert, H. (2012), “The benchden Package: Benchmark Densities for Nonparametric Density Estimation,” *Journal of Statistical Software*, 46, 1–14.
- Miller, J. W. and Dunson, D. B. (2019), “Robust Bayesian inference via coarsening,” *Journal of the American Statistical Association*, 114, 1113–1125.
- Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010), “Bayesian profile regression with an application to the National Survey of Children’s Health,” *Biostatistics*, 11, 484–498.
- Neal, R. M. et al. (2011), “MCMC using Hamiltonian dynamics,” *Handbook of Markov Chain Monte Carlo*, 2, 2.
- Newton, M. A. (2002), “On a nonparametric recursive estimator of the mixing distribution,” *Sankhyā: The Indian Journal of Statistics, Series A*, 64, 306–322.
- Newton, M. A. and Zhang, Y. (1999), “A recursive algorithm for nonparametric analysis with missing data,” *Biometrika*, 86, 15–26.
- 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.
- Pollack, A. Z., Mumford, S. L., Mendola, P., Perkins, N. J., Rotman, Y., Wactawski-Wende, J., and Schisterman, E. F. (2015), “Kidney biomarkers associated with blood lead, mercury, and cadmium in premenopausal women: a prospective cohort study,” *Journal of Toxicology and Environmental Health, Part A*, 78, 119–131.
- Polson, N. G. and Scott, J. G. (2010), “Shrink globally, act locally: sparse Bayesian regularization and prediction,” *Bayesian Statistics*, 9, 501–538.
- 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.
- Pólya, G. (1920), “Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem,” *Mathematische Zeitschrift*, 8, 171–181.
- Poworoznek, E., Ferrari, F., and Dunson, D. (2021), “Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching,” *arXiv preprint arXiv:2107.13783*.

- R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Rahman, H. H., Niemann, D., and Munson-McGee, S. H. (2022), “Association of albumin to creatinine ratio with urinary arsenic and metal exposure: evidence from NHANES 2015–2016,” *International Urology and Nephrology*, 54, 1343–1353.
- Ramsay, J. O. (1988), “Monotone regression splines in action,” *Statistical Science*, 3, 425 – 441.
- Rao, V., Lin, L., and Dunson, D. B. (2016), “Data augmentation for models based on rejection sampling,” *Biometrika*, 103, 319–335.
- Richard, E., Savalle, P.-A., and Vayatis, N. (2012), “Estimation of simultaneously sparse and low rank matrices,” *arXiv preprint arXiv:1206.6474*.
- Ročková, V. and George, E. I. (2016), “Fast Bayesian factor analysis via automatic rotations to sparsity,” *Journal of the American Statistical Association*, 111, 1608–1622.
- Rohe, K. and Zeng, M. (2020), “Vintage factor analysis with varimax performs statistical inference,” *arXiv preprint arXiv:2004.05387*.
- 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, A., Lavine, I., Herring, A. H., and Dunson, D. B. (2021), “Perturbed factor analysis: Accounting for group differences in exposure profiles,” *The Annals of Applied Statistics*, 15, 1386–1404.
- Rue, H. (2001), “Fast sampling of Gaussian Markov random fields,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 325–338.
- Schiavon, L., Canale, A., and Dunson, D. B. (2022), “Generalized infinite factorization models,” *Biometrika*, 109, 817–835.
- Scott, D. W. (2015), *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons.
- Sheather, S. J. and Jones, M. C. (1991), “A reliable data-based bandwidth selection method for kernel density estimation,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 53, 683–690.
- Shikhaliev, A. P., Potter, L. C., and Chi, Y. (2019), “Low-rank structured covariance matrix estimation,” *IEEE Signal Processing Letters*, 26, 700–704.

- Shively, T. S., Sager, T. W., and Walker, S. G. (2009), “A Bayesian approach to non-parametric monotone function estimation,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 159–175.
- Song, P. X.-K. (2000), “Multivariate dispersion models generated from Gaussian copula,” *Scandinavian Journal of Statistics*, 27, 305–320.
- Srivastava, S., Engelhardt, B. E., and Dunson, D. B. (2017), “Expandable factor analysis,” *Biometrika*, 104, 649–663.
- Srivastava, S., Li, C., and Dunson, D. B. (2018), “Scalable Bayes via barycenter in Wasserstein space,” *The Journal of Machine Learning Research*, 19, 312–346.
- Terrell, G. R. and Scott, D. W. (1992), “Variable kernel density estimation,” *The Annals of Statistics*, pp. 1236–1265.
- Teschl, G. (2009), “Mathematical methods in quantum mechanics,” *Graduate Studies in Mathematics*, 99, 106.
- Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*, Springer New York, NY, 1st edn.
- Vaidya, P. M. (1986), “An optimal algorithm for the all-nearest-neighbors problem,” in *27th Annual Symposium on Foundations of Computer Science*, pp. 117–122.
- Vershynin, R. (2010), “Introduction to the non-asymptotic analysis of random matrices,” *arXiv preprint arXiv:1011.3027*.
- Vershynin, R. (2011), “Spectral norm of products of random and deterministic matrices,” *Probability theory and related fields*, 150, 471–509.
- Wand, M. P. and Jones, M. C. (1994), “Multivariate plug-in bandwidth selection,” *Computational Statistics*, 9, 97–116.
- Wang, C., Jiang, B., and Zhu, L. (2019), “Penalized interaction estimation for ultra-high dimensional quadratic regression,” *arXiv preprint arXiv:1901.07147*.
- Wang, L. and Dunson, D. B. (2011), “Fast Bayesian inference in Dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, 20, 196–216.
- Wang, W. and Stephens, M. (2021), “Empirical bayes matrix factorization,” *The Journal of Machine Learning Research*, 22, 5332–5371.
- Wei, R., Reich, B. J., Hoppin, J. A., and Ghosal, S. (2020), “Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures,” *Statistica Sinica*, 30, 55–79.

- West, M. (1992), *Hyperparameter estimation in Dirichlet process mixture models*, Duke University ISDS Discussion Paper# 92-A03.
- Williams, C. K. and Rasmussen, C. E. (2006), *Gaussian Processes for Machine Learning*, vol. 2, MIT press Cambridge, MA.
- Wong, W. H. and Ma, L. (2010), “Optional Polya tree and Bayesian inference,” *The Annals of Statistics*, 38, 1433–1459.
- Xie, F., Cape, J., Priebe, C. E., and Xu, Y. (2022), “Bayesian sparse spiked covariance model with a continuous matrix shrinkage prior,” *Bayesian Analysis*, 17, 1193–1217.
- Zhang, A. R. and Zhou, Y. (2020), “On the non-asymptotic and sharp lower tail bounds of random variables,” *Stat*, 9, e314.
- Zhang, T. and Zou, H. (2014), “Sparse precision matrix estimation via lasso penalized D-trace loss,” *Biometrika*, 101, 103–120.
- Zhang, X., Nott, D. J., Yau, C., and Jasra, A. (2014), “A sequential algorithm for fast fitting of Dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, 23, 1143–1162.
- Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2016), “Bayesian group factor analysis with structured sparsity,” *The Journal of Machine Learning Research*.

Biography

Shounak Chattopadhyay completed his Bachelor's and Master's degrees in Statistics from the Indian Statistical Institute, Kolkata. In his Master's dissertation, he worked on developing approaches for directional statistics. He embarked on his graduate studies at the Department of Statistical Science, Duke University in the Fall of 2018 and plans to graduate in the Summer of 2023. Following his graduation, Shounak will be a postdoctoral researcher at the University of California, Los Angeles, working under the supervision of Dr. Marc Suchard.