Constrained Bayesian Inference through Posterior Projection with Applications

by

Sayan Patra

Department of Statistical Science
Duke University

Date: ______________________

Approved:

____
David B. Dunson, Supervisor

____
Surya T. Tokdar

____
Rebecca C. Steorts

____
Henry D. Pfister

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
2019

Abstract

Constrained Bayesian Inference through Posterior Projection with Applications

by

Sayan Patra

Department of Statistical Science
Duke University

Date: ____________________________
Approved:

____
David B. Dunson, Supervisor

____
Surya T. Tokdar

____
Rebecca C. Steorts

____
Henry D. Pfister

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
2019

Abstract

In a broad variety of settings, prior information takes the form of parameter restrictions. Bayesian approaches are appealing in parameter constrained problems in allowing a probabilistic characterization of uncertainty in finite samples, while providing a computational machinery for the incorporation of complex constraints in hierarchical models. However, the usual Bayesian strategy of directly placing a prior measure on the constrained space, and then conducting posterior computation with Markov chain Monte Carlo algorithms is often intractable. An alternative is to initially conduct computation for an unconstrained or less constrained posterior, and then project draws from this initial posterior to the constrained space through a minimal distance mapping. This approach has been successful in monotone function estimation but has not been considered in broader settings.

In this dissertation, we develop a unified theory to justify posterior projection in general Banach spaces including for infinite-dimensional functional parameter space. For tractability, in chapter 2 we focus on the case in which the constrained parameter space corresponds to a closed, convex subset of the original space. A special class of non-convex sets called Stiefel manifolds is explored later in chapter 3. Specifically, we provide a general formulation of the projected posterior and show that it corresponds to a valid posterior distribution on the constrained space for particular classes of priors and likelihood functions. We also show how the asymptotic properties of the unconstrained posterior are transferred to the projected posterior. We then illustrate

our proposed methodology via multiple examples, both in simulation studies and real data applications.

In chapter 4, we extend our proposed methodology of posterior projection to that of small area estimation (SAE), which focuses on estimating population parameters when there is little to no area-specific information. “Areas” are often spatial regions, where they might be different demographic groups or experimental conditions. To improve the precision of estimates, a common strategy in SAE methods is to borrow information across several areas. This is generally achieved by using a hierarchical or empirical Bayesian model. However, parameter constraints arising naturally from surveys pose a challenge to the estimation procedure. Examples of such constraints include the variance of the estimate of an area being proportional to the geographic size of the area or the sum of the county level estimates being equal to the state level estimates. We utilize and extend the posterior projection approach to facilitate such computing and reduce parameter uncertainty.

This dissertation develops the fundamental new approaches for constrained Bayesian inference, and there are many possible directions for future endeavors. One such important generalization is considered in chapter 5 to allow for conditional posterior projections; for example, applying projection to a subset of parameters immediately after each update step within a Markov chain Monte Carlo algorithm. We identify several scenarios where such a modified algorithm converges to the underlying true distribution and develop a general theory to ensure consistency. We conclude the dissertation by discussing future directions of research in chapter 6, outlining many directions for continued research on these topics.

To,

my mother,

Iti Mukherjee Patra

and

my father,

Sanmay Kumar Patra.

Couldn't have done it without your genetic material.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

AUC	Area Under the Curve
BMF	Bingham-von Mises-Fisher
DL	Dirichlet-Laplace
DTI	Diffusion Tensor Imaging
MCMC	Markov Chain Monte Carlo
MH	Metropolis Hastings
MSE	Mean Squared Error
ROC	Receiver Operating Characteristic
SAE	Small Area Estimation

Symbols

$\mathcal{X}$	State space
$X^{(n)}$	Data with n points
Θ	Unrestricted parameter space
θ	Unrestricted parameter
$\widetilde{\Theta}$	Restricted parameter space
$\tilde{\theta}$	Restricted parameter
$\mathcal{B}_\Theta$	Borel sets on Θ
P_θ	Distribution of data given parameter θ

$\|\cdot\|$	A norm on Θ
Π_Θ	Prior measure on Θ
$\Pi_\Theta^{(n)}$	Posterior measure on Θ
$\pi_\Theta^{(n)}$	Posterior density on Θ
$\widetilde{\Pi}_\Theta^{(n)}$	Projected posterior measure on Θ
$\tilde{\pi}_\Theta^{(n)}$	Projected posterior density on Θ
$d_\Theta^{(n)}$	Distance measure on Θ
μ_Θ	Base measure on Θ

Acknowledgements

"I BLAME ALL of you. Writing this book has been an exercise in sustained suffering. The casual reader may, perhaps, exempt herself from excessive guilt, but for those of you who have played the larger role in prolonging my agonies with your encouragement and support, well . . . you know who you are, and you owe me." - John Fea.

Sometimes when you pray for a single good mentor in your Ph.D., life blesses you with many. If I have made any significant discoveries or seen a little farther, it is because I was standing on the shoulders of these giants.

First, I would like to thank my advisor David Dunson for all the time and effort he put into my personal development and the betterment of my research. David's enthusiasm and zeal for scholarly pursuits are infectious. Without his continued guidance, constructive criticism and unrelenting persuasion, this dissertation would not be a reality. David understood my capabilities better than me and pushed me beyond my limits to undoubtedly make me a better researcher. His signature early morning page-long emails are a better alarm clock than my mobile, computer and two watches combined. It has also been the single most influential source of my motivation and the influx of new ideas. It was an absolute honor to be David's student and I look forward to collaborating with him throughout my professional career.

Thanks also to Surya Tokdar who was always willing to listen, think and guide me in my projects out of the blue and provide me with valuable perspectives. I will miss our meandering, hour-long meetings where we went countless times to the whiteboard to hash out an idea, though Surya may appreciate having his time returned! He has been gracious enough to help me with work that did not make it into the dissertation, but which I look forward to completing with him in the future.

I am fortunate to work with Rebecca Steorts who introduced me to Small Area Estimation and is a major factor in the completion of the dissertation. Thanks to her for providing guidance, conversation and believing in me on days I did not believe in myself. Many students do not have such an attentive mentor.

Thanks are also due to Henry Pfister, who was considerate to make space for me in his busy schedule and hop on board late into the process. I am grateful to him for serving in the committee and providing helpful comments.

Thank you to the entire DSS staff for their support throughout my time in the program. Special thanks to Lori Rauch who has, with constant patience and kindness, answered every question I had (big and small), and guided me through all the administrative hurdles of being an international student.

My time at Duke would not have been half as fun without the incredible friends and labmates at the department. I would like to thank my first-year cohort, Joe, Jake, Michael, Eric, Liz, Steph, Jody, Willem, and Lu - it is your company that made the transition to a life at the USA a smoother experience. Two of them deserves a special mention though - Joe and Jake, the best roommates I could have ever asked for. Thank you for introducing me to the fascinating world of gym, board games and good beers. Together we enjoyed climbing, foosball, canoeing, several video games, and countless other activities. In return, I hope I have instilled in them a love for spicy foods and incomprehensible measure theory (just joking!). Heartfelt thanks to Sarmistha Nandi for her endless encouragement, patience, and optimism. Lastly, thank you to my friends close and far - whether it was a phone call, Skype chat, game night, or toss-around, your friendship buoyed me and I am more appreciative than I can express.

I am deeply grateful to my family for their constant love and support. A yearly pilgrimage to India kept me sated and focused. Maa (Mother) and Baba (Father), thank you for all your love, patience and sacrifices; for being there when I needed you the most and dutifully leaving me alone when I needed the space. Thank you for supporting my statistics journey from the very first class in school to the final stretch of the dissertation. I wouldn't be where I am without you. Dodo di, I know it hasn't always been clear what I've been up to, but I appreciate you bearing with it and providing me foundation. Finally, shoutout to my best friend and partner Sudarshana, for standing by me through thick and thin. You are the bedrock that I lean on and I couldn't imagine pursuing these past years without you by my side. With you, I am home.

A big thanks also to David Benioff and D.B. Weiss, the creators of Game of Thrones TV series on HBO. The mediocrity of their narrative in season 7 and 8 helped me to not obsessively rewatch the episodes and focus more of my time and energy in finishing my research. Night king, you deserved better!

This research has been funded partially by the National Science Foundation research grant, as well as the Duke University Grad School and I am grateful for the opportunity.

1

Introduction

Constraints on model parameters are abundant in many literatures across several domains. Incorporating such constraints can significantly improve model performance (Dunson and Neelon, 2003; Rao and Molina, 2015; Ghosh and Steorts, 2013). Bayesian methodology is well suited for constrained estimation problems as it naturally provides a framework that allow complex constraints, such as monotonicity and concavity, via hierarchical models. One advantage of obtaining a probabilistic characterization via Bayesian method is the appealing nature of uncertainty quantification via credible intervals in addition to the standard point estimate. However, the usual Bayesian strategy of directly putting a prior on the constrained space, often via a truncated prior, yields an intractable posterior. As a result, the posterior sampling via Markov Chain Monte Carlo (MCMC) or Hamiltonian Monte Carlo (HMC) algorithm is slow and the effective sample size of the output is low.

An alternative approach is to ignore the constraint while conducting posterior computation, and then projecting the samples from the unconstrained posterior onto the constrained parameter space. This approach has been shown to be successful in using ordered parameter (Dunson and Neelon, 2003; Neelon and Dunson,

2004) and monotone function estimation (Lin and Dunson, 2014), but has not been considered in broader settings. In this dissertation, we develop a unified decision-theoretic framework to justify posterior projection in general Banach spaces as well as in small area estimation. In chapter 2, we show that there always exists a base measure in the constrained space with respect to which the posterior projection is a valid Bayesian method. More specifically, we illustrate that there exists a prior density on the constrained space such that the density induced by the projection operator is an appropriate posterior. We generalize the proposed methodology to be applicable to any closed and convex constraint of a separable, uniformly convex and uniformly smooth Banach space. We prove that under these circumstances the projected posterior has a convergence rate at least that of the unconstrained posterior. The usefulness and reliability of the method are illustrated via a simple simulation study. We show even in one dimensional Normally distributed data, the projected posterior provides a better estimation than traditional truncated posterior. Finally, we apply the method to a dose-response study from Agresti and Coull (1998) on patients with trauma due to Subarachnoid Hemorrhage. The patients were divided into four treatment groups and different treatment groups were given either a placebo or increasing doses of a certain drug. The posterior projection method takes into account the orderings among categories of classifications and provides better estimates and credible intervals for every group of patients.

In chapter 3, we extend our proposed methodology to include a special nonconvex structure, namely Stiefel manifold. A Stiefel manifold $St(p, m)$ is the set of all orthonormal p-frames in $\mathbb{R}^m$, that is the set of all p-tuples of orthonormal vectors in $\mathbb{R}^m$. The most common example of Stiefel manifold is the $m-1$ dimensional unit circle in $\mathbb{R}^m$, here $p = 1$. We prove a similar consistency theorem as in chapter 2 and provide insight into how it handles sets with 0 measure. Having laid the theoretical foundations, we apply the method to a set of DTI brain imaging data. The objective of this analysis was to discover "connectome"s, which are functionally connected but anatomically separated brain regions. For identifiability, one restricts the brain subnetwork matrix to be in the Stiefel manifold. Using our proposed methodology we achieve the sampling speed and effective sample size similar to that of an unconstrained Bayesian method, while prominently identifying the brain subnetworks and losing a negligible amount of prediction power.

In chapter 4, we consider posterior projection method in the context of SAE. SAE refers to methodology for improving the precision of the estimates of population parameters each of which is associated with an "area" (Rao and Molina, 2015). An "area", which is either a geographic domain, a demographic group or an experimental condition, is considered small when there is little or no-area specific information for directly estimating the parameters of interest. Traditional survey-based methods provide estimates that are based only on the sample observations corresponding to the particular population characteristic. However, these direct estimates may have an unacceptably large variance for small areas (Prasad and Rao, 1990). To remedy this SAE uses hierarchical and empirical Bayes methods. These model-based estimates improve the accuracy of the estimates by borrowing information from related areas (Louis, 1984; Ghosh, 1992). These estimates often vary wildly and usually do not aggregate to the more reliable direct survey estimates. One way to avoid this is by enforcing the weighted average of the model-based estimates to agree to that of the global direct estimates, this approach is known as "benchmarking" (Datta et al., 2011; Ghosh and Steorts, 2013). Depending on the application it may also be desirable that the estimates of nearby areas have similar parameter values, this is achieved by "smoothing" (Steorts, 2014). From a decision-theoretic perspective the smoothed and benchmarked estimates can be jointly viewed as minimizing the Bayes risk under a pre-specified loss function (usually quadratic) given a set of equality and inequality constraints. Geometrically, this amounts to projecting the model-based estimates onto the feasible set characterized by the constraints (Wehbe et al., 2014). Thus one can use a posterior projection approach to induce a density on the constrained space. In the presence of linear equality constraints, we prove that the posterior projection estimates theoretically converge to the smoothed, benchmarked estimator. In addition, the method naturally provides a probabilistic characterization of the uncertainty of the proposed estimation procedure through a credible interval. The advantage of such a method is then illustrated via an extensive simulation study.

There are many interesting new directions. One such important generalization is considered in chapter 5 to allow for conditional posterior projections; for example, applying projection to a subset of parameters immediately after each update step within a Markov chain Monte Carlo algorithm. For simplicity, we focus on a modified Gibbs sampler Robert and Casella (2013). However, instead of only considering projections we work with the broader class of all linear transformations. The main difference between a general linear transformation and a projection map is that the projection map is not 1-1. Based on a standard bivariate normal distribution with correlation coefficient ρ we identify several scenarios where such a modified algorithm converges to the underlying true distribution and develop a general theory to ensure the consistency.

The remainder of the dissertation is organized as follows. Chapter 2 introduces the idea of the posterior projection and develops the necessary theoretical justification for closed and convex constraints. Chapter 3 extends the methodology to Stiefel manifold, a closed but non-convex space. Chapter 4 shows that using posterior projection one can obtain a credible interval along with a point estimate that is consistent with the benchmarked estimator. Chapter 5 considers projection within the MCMC algorithm, and outlines few scenarios where this converges to the true value. Chapter 6 concludes the dissertation by discussing future work and outlining many directions for continued research on these topics.

2

Posterior Projection onto Closed and Convex Sets

2.1 Introduction

Prior information is commonly available in the form of parameter constraints. For example, one may know that the parameters satisfy some set of linear inequalities, that a regression function is monotone in certain directions, or that parameters lie on a probability simplex or manifold. There is a rich literature on statistical methods for incorporating parameter constraints, both from a frequentist and a Bayesian perspective. The Bayesian viewpoint has some advantages in terms of characterizing uncertainty in a probabilistic manner without appealing to large sample size justifications. In addition, asymptotic results can be highly challenging to obtain in parameter constrained problems, and much of the emphasis has been on proving results for special cases. For example, many papers focus specifically on nonparametric estimation subject to monotonicity (Shively et al., 2011) or concavity (Horowitz and Lee, 2017) constraints. However, even though the Bayesian approach to parameter constraints is conceptually broad, posterior computation is commonly intractable. This has motivated a rich literature on modeling and computational strategies for specific parameter constraints, such as monotonicity in multiple regression functions (Saarela and Arjas, 2011) or Gaussian models subject to linear inequalities (Jidling et al., 2017). There is a clear need for more general approaches.

One intriguing idea that has been proposed in the literature is to initially conduct posterior computation ignoring the parameter constraint, and then project samples from the unconstrained posterior to the constrained space of interest. This posterior projection approach has been specifically developed for ordered parameters (Dunson and Neelon, 2003), unimodality (Gunn and Dunson, 2005) and monotone function estimation (Lin and Dunson, 2014). In such settings, the projected posterior tends to have better performance in several respects than usual Bayesian methods that directly define a posterior on the constrained space. One advantage is computational ease but also the projected posterior tends to reduce bias resulting from parameter constraints. For example, suppose that one wants to estimate a one-dimensional dose response curve characterizing risk of an adverse health response with increasing dose of a chemical exposure. In this setting, it is often reasonable to assume that risk does not decrease with increasing dose, leading to a monotonicity restriction. However, suppose that the sample size is modest and the true dose response curve is close to flat. Usual Bayesian methods, that choose a prior on the space of monotone functions, will tend to badly over-estimate the slope of the dose response curve. The posterior projection approach limits this over-estimation problem by mapping draws from the unconstrained posterior that violate monotonicity to the boundary of the constrained space; this is illustrated by a simpler model in Figure 2.2.

Although this projected posterior approach is intriguing, it has only been implemented in these specific cases in the previous literature, and the general Bayesian justification for the approach is unclear. The main contribution of this article is to provide a general framework for defining projected posteriors in arbitrary problems, while showing that projected posteriors are valid Bayesian posteriors under some conditions. For tractability theoretically and computationally, we mostly focus on cases in which the constrained space corresponds to a closed, convex subset of the original parameter space. General theory for posterior projection on Stiefel manifolds, a special class of non-convex sets, is developed later in the paper.

2.2 Theory & methods

2.2.1 Notation and formulation

Let the data $X^{(n)} = (X_1, ..., X_n)$ be a random sample following the distribution P_θ, which has a density p_θ with respect to a σ-finite dominating measure on the sample space $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)})$. We assume that the unknown parameter θ belongs to a constrained subset $\widetilde{\Theta}$ of the original parameter space Θ, that is $\widetilde{\Theta} \subset \Theta$. Usually this restriction is admitted by a prior density $\pi_{\widetilde{\Theta}}$ with probability measure $\Pi_{\widetilde{\Theta}}$ which is supported on the constrained parameter space $(\widetilde{\Theta}, \mathcal{B}_{\widetilde{\Theta}})$. Thus the posterior distribution of θ given $X^{(n)}$ is

$$\Pi_{\widetilde{\Theta}}(B|X^{(n)}) = \frac{\int_B p_\theta(X^{(n)})\, d\,\Pi_{\widetilde{\Theta}}}{\int_{\widetilde{\Theta}} p_\theta(X^{(n)})\, d\,\Pi_{\widetilde{\Theta}}} \quad \text{for all} \quad B \in \mathcal{B}_{\widetilde{\Theta}}. \tag{2.1}$$

The expression in equation (2.1) is defined if and only if the denominator is positive. For that it suffices for Θ to be Polish (Ghosal and Van der Vaart, 2017). Assuming $\widetilde{\Theta}$ is a non-empty, closed, convex subset of Θ, it is also Polish with its corresponding Borel σ-algebra $\mathcal{B}_{\widetilde{\Theta}} = \mathcal{B}_\Theta \cap \widetilde{\Theta}$.

A popular choice of restricted prior is $\pi_{\widetilde{\Theta}} \propto \pi_\Theta 1_{\widetilde{\Theta}}(\theta)$, where π_Θ is an unconstrained probability density on Θ and $1_{\widetilde{\Theta}}(\theta)$ is an indicator function of $\widetilde{\Theta}$. Quite often this leads to an intractable posterior which is difficult to sample from. In the case where $\widetilde{\Theta}$ has zero measure with respect to the base measure μ_Θ, the posterior in (2.1) is ill-defined as the denominator is 0.

As an alternative method which avoids these problems, we propose the Posterior

Projection approach where initially we ignore the restriction and sample from the unconstrained posterior density $\pi_\Theta(\theta|X^{(n)}) \propto p_\theta(X^{(n)})\pi_\Theta(\theta)$. These samples are then projected to the constrained space $\widetilde{\Theta}$ using a minimal distance mapping $T_{\widetilde{\Theta}}$. In practice, generating samples from the unconstrained posterior is often relatively easy and fast. The details about the choice of the minimal distance mapping is discussed in section 2.2.2.

2.2.2 Metric projection in general spaces

In this section we develop the structure of the underlying topological and measure space necessary for the minimal distance mapping. Because Θ is assumed to be Polish, it is also a complete normed vector space $(\Theta, \|\cdot\|)$. We will use the associated metric $\|\cdot\|$ to define the projection operator. For a nonempty subset $\widetilde{\Theta}$ of Θ, the distance between the point $\theta \in \Theta$ and the set $\widetilde{\Theta}$ is given by the following equation:

$$d(\theta, \widetilde{\Theta}) = \inf\{\| \theta - \tilde{\theta} \| : \tilde{\theta} \in \widetilde{\Theta}\} \tag{2.2}$$

A point $\tilde{\theta} \in \widetilde{\Theta}$ with $\| \theta - \tilde{\theta} \| = d(\theta, \widetilde{\Theta})$ is called a best approximation of θ in $\widetilde{\Theta}$.

Definition 1. *Let $(\Theta, \|\cdot\|)$ be a complete normed space and $\widetilde{\Theta}$ be a nonempty subset of Θ. We define the set valued mapping $T_{\widetilde{\Theta}} : \Theta \to \mathcal{P}(\Theta)$ by*

$$T_{\widetilde{\Theta}}(\theta) = \{\widetilde{\theta} \in \widetilde{\Theta} : \| \theta - \widetilde{\theta} \| = d(\theta, \widetilde{\Theta})\} \tag{2.3}$$

and refer to it as the metric projection operator onto $\widetilde{\Theta}$.

When there is no ambiguity about the constrained space $\widetilde{\Theta}$ we omit the subscript and denote $T_{\widetilde{\Theta}}(\theta)$ by $T\theta$. Clearly T is idempotent, that is $T \circ T = T$. Additionally if we denote the set inverse of T by $T^{-1}B = \{\theta \in \Theta \mid T\theta \in B\}$, we get

$$A \subseteq T^{-1} \circ TA, \quad B = T \circ T^{-1}B, \quad \text{for all } A \in \mathcal{B}_\Theta, B \in \mathcal{B}_{\widetilde{\Theta}}.$$

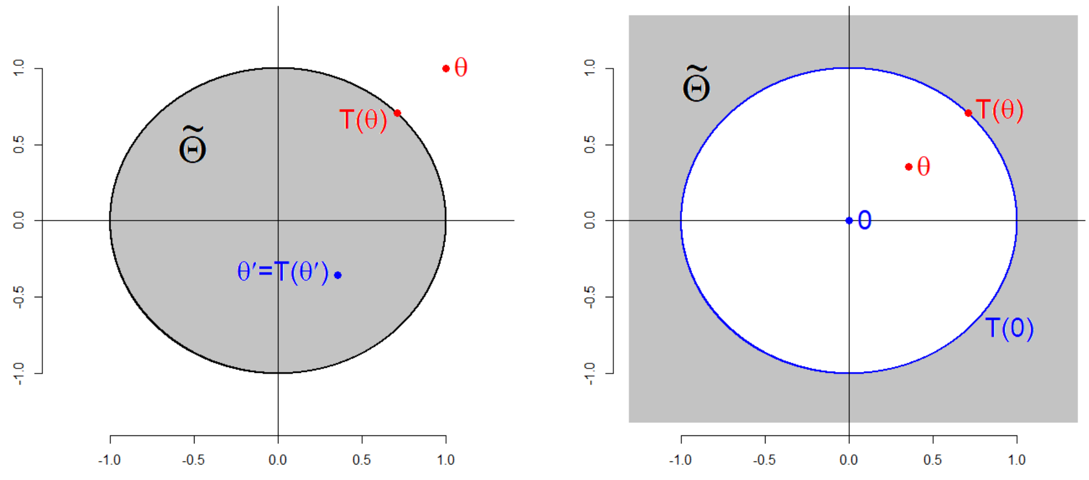
Basic concerns in the metric projection problem include the existence and uniqueness which are dependent on the parameters $(\Theta, d, \widetilde{\Theta}, \theta)$. We say $\widetilde{\Theta}$ is a proximal set if $T\theta \neq \varnothing$ and that $\widetilde{\Theta}$ is a Chebyshev set if $T\theta$ is a singleton for each $\theta \in \Theta$. The proximality condition is vital as it allows one to explicitly compute the post sampling projections while the Chebyshev condition guarantees the uniqueness. We illustrate these concepts with examples in $\mathbb{R}^2$; refer to Fig. 2.1.

Assume that the constrained space is the closed unit ball $B[0,1]$ equipped with the Euclidean norm. It is easy to see that any point within the circle gets mapped back to itself while a point outside gets projected radially, that is $T\theta = \theta/\|\,\theta\,\|$ for $\theta \notin B[0,1]$. Here every point has a unique projection and hence $B[0,1]$ is a Chebyshev set. Now consider that the constrained space is the region outside the open unit ball, that is $\widetilde{\Theta} = \mathbb{R}^2 \backslash B(0,1)$. In this case every element in the boundary of the unit ball is equidistant from the centre 0, therefore $T(0) = S_{\mathbb{R}^2}$, the unit circle. Every other element of $B(0,1)$ has a unique radial projection as before, hence $\widetilde{\Theta}$ is proximal.

When the metric projection T is a measurable mapping between spaces $(\Theta, \mathcal{B}_\Theta)$ and $(\widetilde{\Theta}, \mathcal{B}_{\widetilde{\Theta}})$, which is guaranteed by the Assumption 3 introduced later, T induces a pushforward measure $\widetilde{\Pi}_{\widetilde{\Theta}}\left(\cdot|X^{(n)}\right)$ on $\widetilde{\Theta}$ corresponding to a posterior measure $\Pi_\Theta\left(\cdot|X^{(n)}\right)$ on Θ. For any $B \in \mathcal{B}_{\widetilde{\Theta}}$ this is given by

$$\widetilde{\Pi}^{(n)}_{\widetilde{\Theta}} B = \widetilde{\Pi}_{\widetilde{\Theta}}(\cdot|X^{(n)})(B) = \Pi_\Theta(\cdot|X^{(n)})(T^{-1}B) = \Pi^{(n)}_\Theta(T^{-1}B). \tag{2.4}$$

We refer to $\Pi_\Theta\left(\cdot|X^{(n)}\right)$ and $\widetilde{\Pi}_{\widetilde{\Theta}}\left(\cdot|X^{(n)}\right)$ as the original or unrestricted posterior and the constrained or projected posterior and use the shorthand notations $\Pi^{(n)}_\Theta$ and $\widetilde{\Pi}^{(n)}_{\widetilde{\Theta}}$, respectively.



(a) **(b)**

FIGURE 2.1: Illustration of the Metric projection operator T in $\mathbb{R}^2$ under the Euclidean norm when the constrained space, denoted by the shaded region in the figure, is (a) Chebyshev set and (b) proximal set. A point and its corresponding projection set are indicated by the same color.

2.2.3 Existence of prior on the restricted space

Our aim in this section is to establish a Bayesian justification for our posterior projection scheme. We will prove the existence of a prior $\Pi_{\widetilde{\Theta}}$ such that the resulting posterior $\Pi_{\widetilde{\Theta}}(\cdot|X^{(n)})$ is equivalent to the projected posterior $\widetilde{\Pi}_{\widetilde{\Theta}}^{(n)}$ almost everywhere. Mathematically, for any $B \in \mathcal{B}_{\widetilde{\Theta}}$ we want

$$\widetilde{\Pi}_{\widetilde{\Theta}}^{(n)} B = \Pi_{\widetilde{\Theta}}(B \mid X^{(n)}).$$

This condition is satisfied when the projected posterior has a density $\widetilde{\pi}_{\widetilde{\Theta}}^{(n)}$ with respect to a σ-finite measure $\mu_{\widetilde{\Theta}}$ on $\widetilde{\Theta}$ (Theorem 2).

Assumption 1. *The projected posterior has a density $\widetilde{\pi}_{\widetilde{\Theta}}^{(n)}$, given by the Radon-Nikodym derivative with respect to a σ-finite measure $\mu_{\widetilde{\Theta}}$ on $\widetilde{\Theta}$.*

The question of when such a density exists is a tricky one. For example, suppose $X = f(\theta) + \epsilon$, f is constrained to be non-decreasing, and the original prior is $f \sim \mathrm{GP}(0, K)$ corresponding to a Gaussian process with covariance function K. Then one runs into the difficult problem of defining densities on function spaces. We begin to unravel this mystery with the following lemma which will be instrumental in the subsequent arguments.

Lemma 1. *Let λ_Θ and μ_Θ be two measures on the original parameter space Θ such that $\lambda_\Theta \ll \mu_\Theta$. Then the metric projection T preserves absolute continuity, that is $\widetilde{\lambda}_{\widetilde{\Theta}} \ll \widetilde{\mu}_{\widetilde{\Theta}}$.*

As an immediate corollary, we notice that when the posterior measure is absolutely continuous with respect to the prior (an exception is non-parametric Dirichlet prior, see Ghosal and Van der Vaart (2017) for more details), the corresponding projected posterior is also absolutely continuous with respect to the projected prior measure. The same logic can be applied for the σ-finite base measure μ_Θ to obtain the following result.

$$\Pi_\Theta^{(n)} \ll \Pi_\Theta \ll \mu_\Theta \implies \widetilde{\Pi}_{\widetilde{\Theta}}^{(n)} \ll \widetilde{\Pi}_{\widetilde{\Theta}} \ll \widetilde{\mu}_{\widetilde{\Theta}}$$

We observe that while the projection map preserves absolute continuity (Lemma 1), it does not ensure retention of the σ-finite property. Therefore, even if the unrestricted posterior measure $\Pi_\Theta^{(n)}$ had a density with respect to a base measure μ_Θ on Θ, there is no guarantee that the projected posterior $\widetilde{\Pi}_{\widetilde{\Theta}}^{(n)}$ will have a density with respect to the projected base measure $\widetilde{\mu}_{\widetilde{\Theta}}$. However, one can generally construct a reference measure in $\widetilde{\Theta}$ with respect to which the density exists. One such possibility is using $\widetilde{\Pi}_{\widetilde{\Theta}}$ as a reference measure when $\Pi_\Theta^{(n)} \ll \Pi_\Theta$ is true. Because Π_Θ is finite, $\widetilde{\Pi}_{\widetilde{\Theta}}$ is σ-finite, and thus $\Pi_\Theta^{(n)}$ has a density (Assumption 1). These ideas are thoroughly discussed with respect to an example in Appendix C of the Supplementary Materials.

Theorem 2. *Let the unrestricted posterior measure* $\Pi_\Theta^{(n)}$ *induce the constrained posterior measure* $\widetilde{\Pi}_{\widetilde{\Theta}}^{(n)}$ *through the metric projection* T*. Then under Assumption 1 there exists a prior density* $\pi_{\widetilde{\Theta}}$ *on the constrained parameter space such that the corresponding posterior density is equal to the projected posterior density* $\widetilde{\pi}_{\widetilde{\Theta}}^{(n)}$ *almost everywhere.*

2.3 Metric projection onto closed and convex sets

2.3.1 Convergence properties on closed and convex sets

This section focuses on the frequentist properties of the projected posterior distribution $\widetilde{\Pi}_{\widetilde{\Theta}}^{(n)}$ as $n \to \infty$. We study the rate at which the posterior concentrates on arbitrarily small neighbourhoods of $P_{\theta_0}^{(n)}$, where $\theta_0 \in \Theta$ denotes the true value of the parameter. Let the unrestricted posterior $\Pi_\Theta^{(n)}(\cdot \mid X^{(n)})$ have rate ϵ_n with respect to a semimetric d_Θ on Θ in the following sense:

$$\Pi_\Theta^{(n)}\left(\theta : d_\Theta(\theta, \theta_0) > M_n\epsilon_n \mid X^{(n)}\right) \to 0, \quad \text{for every } M_n \to \infty, \quad P_{\theta_0}^{(n)}(a.s.). \tag{2.5}$$

We need to impose further regulatory conditions on the parameter space to ensure posterior consistency. We list the assumptions below and discuss their necessity and implications.

Assumption 2. $(\Theta, \|\cdot\|)$ *is a separable, uniformly convex and uniformly smooth Banach space.*

We work with Banach spaces as they are a complete normed vector space making calculations more tractable. We also add a separability condition so that Θ is a Polish space which is required for equation (2.1) to be defined. In corollary 4 we give examples of some familiar spaces where this assumption is realized.

Assumption 3. $\widetilde{\Theta}$ *is a nonempty, closed, convex subset of* Θ*.*

Milman-Pettis theorem states that every uniformly convex Banach space is reflexive and strictly convex. Therefore the metric projection operator T is continuous and $\widetilde{\Theta}$ is a Chebyshev set (Li (2004), Theorem E). More importantly, T has the following local Lipschitz continuity property (Li (2004), Theorem G):

$$\| T\theta - \tilde{\theta} \| \leqslant \| \theta - \tilde{\theta} \| \text{ , for any } \theta \in \Theta \text{ and } \tilde{\theta} \in \widetilde{\Theta}.$$

This property is called Lipschitz continuity mod $\widetilde{\Theta}$. It is a relaxation of uniform continuity, which is a very strict global condition and difficult to obtain in practice.

Assumption 4. *The semimetric d_Θ is biLipschitz with respect to $(\Theta, \| \cdot \|)$; that is there exists a constant $C \geqslant 1$ such that*

$$C^{-1} \| \theta - \theta' \| \leqslant d_\Theta(\theta, \theta') \leqslant C \| \theta - \theta' \| \text{ , for any } \theta, \theta' \in \Theta.$$

The implication of this condition becomes evident when considering any norm equivalent to $\| \cdot \|$, specifically all the L_p norms for $p > 0$ in a finite dimensional Euclidean space are equivalent. One might ask why not use d_Θ as the underlying distance function of the projection operator in the first place. Firstly d_Θ is not guaranteed to be a metric. More importantly, the projection map might be easier to compute with respect to $\| \cdot \|$ than d_Θ. For example, in $\mathbb{R}$ the projection operator has a closed form solution with respect to the L_2 norm, but needs to be calculated numerically for L_1 norm.

Theorem 3. *If the true parameter value $\theta_0 \in \widetilde{\Theta}$, then under Assumptions 2-4 the concentration rate of the projected posterior is at least that of the original posterior.*

As an immediate consequence of Theorem 3 we note some of the well-known spaces that satisfy Assumption 2.

Corollary 4. *The statement of Theorem 3 is realized in the following spaces.*

1. *$\ell_p^n, 1 \leqslant p \leqslant \infty$: the space of all n-dimensional sequences with finite p-norm.*

2. *$\ell_p, 1 < p < \infty$: the space of all countably infinite dimensional sequences with finite p-norm.*

3. *$L_p(\mu), 1 < p < \infty$: the space of all functions with finite p-norm with respect to the measure μ.*

The statement of Theorem 3 can be extended to the case where θ_0 is outside the restricted region $\widetilde{\Theta}$ when Θ is a Hilbert space. This is because metric projection onto a closed convex subset of Hilbert space is non expansive (Fitzpatrick and Phelps, 1982), that is

$$\| T\theta - T\theta' \| \leqslant \| \theta - \theta' \|, \quad \text{for any } \theta, \theta' \in \Theta. \tag{2.6}$$

Therefore in Hilbert spaces we achieve the following theorem.

Theorem 5. *Let Θ be a separable Hilbert space. If the original posterior $\Pi_\Theta^{(n)}$ converges to θ_0 at the rate ϵ_n, then under Assumptions 2–4 the projected posterior converges to $T\theta_0$ at the rate at least ϵ_n.*

As an immediate corollary we list a few familiar spaces where the statement of the Theorem 5 is valid.

Corollary 6. *The statement of Theorem 5 is realized in the following spaces.*

1. *$\mathbb{C}^n$: n-dimensional Euclidean space over complex numbers $\mathbb{C}$.*

2. *ℓ_2^n : the space of all n-dimensional square-summable sequences.*

3. *ℓ_2 : the space of all countably infinite dimensional square-summable sequences.*

4. *$L_2(\mu)$: the space of square-integrable functions with respect to the measure μ.*

2.3.2 Gaussian projection onto the non-negative real line

As a simple illustrative example we consider a univariate Gaussian likelihood with a conjugate prior. Let $X^{(n)} = (x_1, x_2, ..., x_n)$ be such that $x_i \sim N(\theta, 1),\ i = (1, \ldots, n)$ and $\theta \geqslant 0$. In such cases it is common to put a truncated prior on the parameter; for example $\theta \sim N_{(0,\infty)}(0, 1000)$; Here $N_{(0,\infty)}$ denotes the normal distribution truncated onto the nonnegative real line. The posterior of θ is then given by

$$\theta|X^{(n)} \sim N_{(0,\infty)}(\theta_n, \sigma_n^2), \quad \theta_n = \frac{n\bar{x}}{1/1000 + n}, \quad \sigma_n^2 = \frac{1}{1/1000 + n} \tag{2.7}$$

which is equivalent to sampling from the unconstrained posterior and discarding negative values. This posterior density assigns no probability to the boundary 0 and has expectation

$$E(\theta \mid X^{(n)}) = \theta_n + \frac{\phi(\alpha)}{1 - \Phi(\alpha)}\sigma_n \tag{2.8}$$

where $\alpha = -\theta_n/\sigma_n$, ϕ and Φ denote the density and distribution function of the standard normal, respectively.

In the posterior projection approach we initially ignore the constraint on θ, and use the unrestricted prior $\theta \sim N(0, 1000)$. The resulting posterior is $\theta \mid X^{(n)} \sim N(\theta_n, \sigma_n^2)$ which has the same form as equation (2.7) but without the truncation. Next we obtain samples θ_s from the unconstrained posterior and set $\tilde{\theta}_s$ to 0 if θ_s is negative, and equal to θ_s otherwise. Thus the projected posterior on $[0, \infty)$ has the following density:

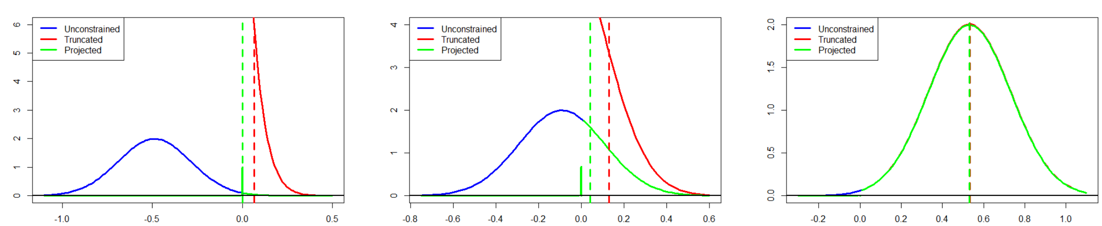
$$\widetilde{\pi}^{(n)}_{[0,\infty)}(\tilde{\theta}) = \Phi\left(\alpha\right) 1_0(\tilde{\theta}) + \left[1 - \Phi\left(\alpha\right)\right] N_{(0,\infty)}(\tilde{\theta}; \theta_n, \sigma_n^2) \tag{2.9}$$

Here $1_0(\tilde{\theta})$ is the indicator of whether $\tilde{\theta}$ is 0 or not. This posterior density (2.9) has expectation

$$E(\tilde{\theta} \mid X^{(n)}) = \left[1 - \Phi\left(\alpha\right)\right] \left(\theta_n + \frac{\phi(\alpha)}{1 - \Phi(\alpha)}\sigma_n\right)$$

which is closer to θ_n than the expression in (2.8). The incorporation of the constraint reduces the uncertainty in the posterior distribution, and projecting the draws that violate the constraint to the boundary allocates higher posterior probability to values that are close to θ_n. This property is appealing since θ_n is a function of $\bar{x}$, which is an unbiased and consistent estimator of θ.

To explore the effect of the projection technique we simulate $n = 25$ datapoints from $N(\theta_0, 1)$ where the value of θ_0 is fixed at $-0.5, 0$ and 0.5, respectively. These cases correspond to being outside, at the boundary and inside the constrained parameter space. For each sample the unconstrained, truncated and projected posterior densities and posterior mean are plotted in Fig. 2.2. When the data are in conflict with the constraint ($\theta_0 = -0.5$) the difference in the approaches is very prominent. Even when the data agree with the constraint, but the truth is at the boundary ($\theta_0 = 0$), the posterior mean of the projection approach is closer to the θ_0 than that of the truncated posterior. The posterior densities coincides when the true value is well inside the constrained region.



$\theta_0 = -0.5$ $\qquad\qquad$ $\theta_0 = 0$ $\qquad\qquad$ $\theta_0 = 0.5$

FIGURE 2.2: Illustration of the effect of the projection technique when the true parameter value θ_0 is outside, at the boundary and inside of the constrained parameter space. In each panel the solid lines represent posterior densities for the Gaussian mean under a noninformative unconstrained $N(0, 1000)$ prior (blue), corresponding projection approach (green) and the truncated $N_{(0,\infty)}(0, 1000)$ prior (blue). The associated posterior means, represented by the dotted vertical lines, are closer to the true parameter value than the truncated one.

It is also possible to determine the prior on the non-negative real line which leads to the projected posterior density in (2.9). This prior density is given by

$$\pi_{[0,\infty)}(\tilde{\theta}) = w_1\, 1_0(\tilde{\theta}) + w_2\, N_{(0,\infty)}(\tilde{\theta};\theta_0,\sigma_0^2)$$

where $w_j = (W_j^0/C_j)/(\sum_{j=1}^2 W_j^0/C_j)$, W_j^0 denotes the corresponding weights in the expression (2.9) and C_j's are given by

$$C_1 = N(0\,;\bar{x},1/n), \quad C_2 = N(0\,;\bar{x},1/n)\frac{N_{(0,\infty)}(0;0,1000)}{N_{(0,\infty)}(0;\theta_n,\sigma_n^2)}.$$

Details of the derivation for a non standard Gaussian is provided in the Appendix C of the Supplementary Materials. It is important to note that the $\pi_{[0,\infty)}(\tilde{\theta})$ is heavily data dependent and this is true in general for the posterior projection approach.

2.4 Projection of a covariance matrix with eigenvalue restriction

Often in practice the covariance matrix of the data is structured. A Linear Mixed Model (LMM) with Y and Z as fixed and random effects

$$X = Y\beta + Z\gamma + \epsilon$$

has covariance structure $\text{var}(X) = ZGZ^T + \psi I$. This can be generalized into $\Sigma = \Phi + \psi I$ where Σ is the covariance of the data and Φ is a positive definite matrix. The usual Bayesian way of putting a conjugate prior on Σ fails to output a closed form posterior, as in the case where the data

$$X \sim N(0,\Sigma) \quad \Phi \sim \mathcal{W}^{-1}(\Psi,\nu).$$

Instead one can view this structure as a restriction where the eigenvalues of the Σ are elementwise greater than ψ. Thus one can ignore the constraint initially and put conjugate prior on Σ i.e. $\Sigma \sim \mathcal{W}^{-1}(\Psi,\nu)$ obtaining a closed form posterior and projecting the MCMC samples after convergence. The details of the projection scheme and extension of the idea is discussed below.

Let us consider the space $\mathcal{S}_n$ of symmetric $n \times n$ matrices equipped with the norm associated with the inner product

$$< a, B >= \text{trace}\left(a^T B\right).$$

The subset $\mathcal{S}_n^{++}$ made of positive definite matrices form a open convex cone of $\mathcal{S}_n$ and the subset $\mathcal{S}_n^{+}$ of positive semidefinite matrices form the convex hull of $\mathcal{S}_n^{+}$. It is well known that given a matrix $a \in \mathcal{S}_n$ there exists an unique projection onto $\mathcal{S}_n^{+}$, say Proj_a. Furthermore the projection has an explicit expression through spectral decomposition. If we consider the decomposition

$$a = U^T \, \text{Diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)\, U.$$

where $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ are the eigenvalues of a and U is a corresponding orthonormal matrix of eigenvectors of a, then the projection onto $\mathcal{S}_n^{+}$ is given by

$$\text{Proj}_{\mathcal{S}_n^{+}}(a) = U^T \, \text{Diag}\,(\max(0, \lambda_1), \max(0, \lambda_2), \ldots, \max(0, \lambda_n))\, U.$$

This result generalizes nicely for spectral sets of matrices. The eigenvalue map $\lambda : \mathcal{S}_n \to \mathbb{R}_n$ maps any symmetric matrix a to its eigenvalues in nonincreasing order, $\lambda_1(a) \geqslant \lambda_2(a) \geqslant \ldots, \lambda_n(a)$. Suppose the matrix $a \in \mathcal{S}_n$ has spectral decomposition $U^T \text{Diag}(\lambda(a)) U$, then it has a unique projection onto the subset $\mathcal{S}_n^{t}$ of matrices with eigenvalues $t_1 \geqslant t_2 \geqslant \ldots t_n$, $t \equiv (t_1, t_2, \ldots t_n) \in \mathbb{R}_n$. The *Isometric Projection* is given by $U^T \text{Diag}(\lambda(a) <> t) U$ where $<>$ denotes elementwise max operation. The uniqueness of the projection follows from the closed and convex property of $\mathcal{S}_n^{t}$ which can be deduced as follows.

$$\mathcal{S}_n^{t} = \{a \in \mathcal{S}_n | z^T a z \geqslant t \,\forall\, z \in \mathbb{R}_n\} = \cap_{z \in \mathbb{R}_n} \{a \in \mathcal{S}_n | z^T a z \geqslant t\} = \cap_{z \in \mathbb{R}_n} H_z$$

Now H_z is a closed halfspace in $\mathcal{S}_n$ for any fixed $z \in \mathcal{S}_n$. So $\mathcal{S}_n^t$ closed and convex as intersection of closed halfspaces.

Another interesting restriction concerns projection onto the set of matrices with maximum eigenvalue having a given multiplicity. The *Maximum Eigenvalue Multiplicity Projection* is given by $U^T \mathrm{Diag}(s) U$ where

$$s_i = \begin{cases} \frac{1}{k} \sum\limits_{j=1}^{k} \lambda_j(a) & \text{if } \ i \leqslant k \\ \lambda_i(a) & \text{if } \ \ i > k \end{cases}$$

2.5 Experiment: Stochastic ordering on a contingency table

Data in the form of a contingency table arise when individuals are classified according to multiple criterion. Often, one or more of the categorical variables have a natural ordering, such as in dose-response studies with the ordered levels corresponding to increasing levels of exposure. Methods that take into account the orderings among categories of classifications benefit in terms of more accurate parameter estimation and better power in hypothesis testing. Agresti and Coull (2002) survey order-restricted statistical methods for contingency tables where these restrictions translate to inequality constraints for a set of probabilities, odds ratios or model parameters. We demonstrate the applicability of our posterior projection method with stochastic ordering constraints and discuss how other constraints can be incorporated.

For simplicity we consider a two way $I \times J$ contingency table with cell entries n_{ij}, $i = (1, \ldots, I)$, $j = (1, \ldots, J)$ and $\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = n$. The rows and columns denote the categories of predictor variable X and response variable Y, respectively. We assume that the rows of the contingency table are independent multinomial samples from different populations with sample sizes $n_{i+} = \sum_{j=1}^{J} n_{ij}$ and probability vector $\theta_{(i)} = (\theta_{i1}, ..., \theta_{iJ})$. Under this setup the condition of Y being stochastically increasing in X translates to

$$\sum_{k=1}^{j} \theta_{ik} \geqslant \sum_{k=1}^{j} \theta_{(i+1),k} \quad i = (1, ..., I), j = (1, ..., J). \tag{2.10}$$

Laudy and Hoijtink (2007) use gamma parametrization of Dirichlet to put a truncated prior on $\theta = (\theta'_{(1)}, ..., \theta'_{(I)})$ which follows the restriction in (2.10). This method is attractive as it allows one to sample from a Dirichlet-multinomial model under a variety of odds ratio constraints. However their sampling technique, which uses the odds ratio inequalities to truncate the gamma prior for every individual parameter at every Gibbs sampling update, is computationally expensive and relatively slow. To circumvent this difficulty we initially ignore the constraint and assign a conjugate Dirichlet prior on $\theta_{(i)}$ with equal hyperparameter α for all the components. The unconstrained posterior distribution of $\theta_{(i)}$ is given by

$$\theta_{(i)} \sim \mathrm{Dir}(n_{i1} + \alpha, ..., n_{iJ} + \alpha). \tag{2.11}$$

We observe that equation (2.10) along with $\sum_{j=1}^{J} \theta_{ij} = 1$ induces linear inequality and equality restrictions on θ. Thus one can use the pooled adjacency violator algorithm Barlow (1972) to find the $\tilde{\theta}$ that minimizes a convex functional of θ under the stochastic order and probability restrictions. We choose Euclidean distance as the convex functional in our application, that is $\tilde{\theta}_E$ minimizes

$$\tau^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} (\theta_{ij} - \tilde{\theta}_{ij})^2. \tag{2.12}$$

Evans et al. (1997) uses the sampled values of τ^2 to construct a test for the hypothesis $H_t = \{\theta | \ \tau(\theta)^2 \leqslant t^{0.5}\}$ against the alternative H_t^c, under stochastic ordering constraint.

We consider the data from Agresti and Coull (1998) which describes the outcome of a sample of patients with a trauma due to subarachnoid Hemorrhage. Patients are divided into four treatment groups. The first group receives a placebo and the rest of the groups are administered increasing doses of a medicine. The outcome of the clinical study is sorted into 5 categories according to the Glasgow Outcome Scale. An interesting way to see if the data exhibits stochastic ordering is to consider the difference of the sampled and projected probabilities under prior and posterior measures. To this end we draw 10000 samples from both the prior and the posterior, calculate the projection and plot the empirical distribution function of τ^2 (2.12). Figure 2.3 shows that the τ^2 value is much smaller in general under the posterior which is an indication of the data following stochastic order.

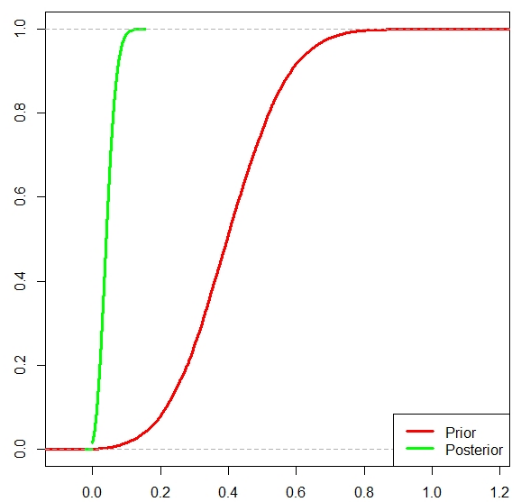


FIGURE 2.3: Empirical cumulative distribution function of τ^2 under prior and posterior measure with $\alpha = 1$

Table 2.1: Sensitivity analysis of hyperparameter α on τ^2.

Value of α	0·1	0·5	1	5	10
Prior mean of τ^2	0·795	0·525	0·402	0·192	0·138
Posterior mean of τ^2	0·043	0·042	0·042	0·040	0·039

Encouraged by this evidence of the constraint, we apply our projection method to the data. After some experimentation the value of α in the posterior (2.11) is fixed at 1. We draw 10000 samples from the posterior and compute the corresponding projections. The projected probabilities along with credible intervals are provided in the Supplementary Materials. We also compare the effect of the projection by comparing the cumulative odds ratios in Table 2.2 and observe that without projection the estimated value of the odds ratio is sometimes less than 1. This is in contradiction to our assumption as when the rows of the categorical table are assumed to be independent Multinomials, the stochastic ordering constraint is equivalent to the cumulative odds ratio being uniformly greater than 1. We also take note that our method produces tighter credible intervals than Laudy and Hoijtink (2007).

Table 2.2: Estimates and credible intervals of cumulative odds ratio under the Dirichlet-multinomial model with $\alpha = 1$. Within each row the upper and lower row contains results using the projection method and sampling from the posterior disregarding the stochastic ordering constraint, respectively. The bold numbers indicate where stochastic ordering fails without projection.

	1–2	2–3	3–4	4–5
1–2	1·147 (1·000 – 1·477)	1·132 (1·000 – 1·372)	1·060 (1·000 – 1·193)	1·014 (1·000 – 1·070)
	1·128 (**0·807** – 1·546)	1·108 (**0·858** – 1·400)	1·043 (**0·889** – 1·215)	1·007 (**0·925** – 1·096)
2–3	1·160 (1·000 – 1·511)	1·291 (1·037 – 1·618)	1·109 (1·000 – 1·261)	1·007 (1·000 – 1·057)
	1·207 (**0·833** – 1·705)	1·302 (**0·970** – 1·716)	1·101 (**0·928** – 1·297)	**0·992** (**0·911** – 1·080)
3–4	1·055 (1·000 – 1·365)	1·219 (1·003 – 1·570)	1·107 (1·000 – 1·299)	1·067 (1·000 – 1·158)
	0·983 (**0·667** – 1·417)	1·175 (**0·839** – 1·619)	1·100 (**0·908** – 1·322)	1·075 (**0·980** – 1·178)

We also investigate the effect of the value of the hyperparameter α. As α increases, more mass is put onto the convex set of probabilities following stochastic order and hence the mean of τ^2 decreases. The posterior mean of τ^2 is unaffected by the change.

2.6 Generalization

Here we discuss briefly how our model can be used in other scenarios.

Shape and order constraints: Isotonic regression is a popular example of order restrictions on a model. Given θ the goal is to minimize the euclidean norm $\frac{1}{2} \| \theta - \tilde{\theta} \|_2^2$ such that $\tilde{\theta}_i$ are non decreasing. The set of $\tilde{\theta}$ satisfying the constraint is a closed convex cone and hence one can use the projection approach. It is evident that our method can be applied when there exists a known order restriction on model parameters that is not necessarily isotone (Dunson and Neelon, 2003), which is the case in many biomedical and financial time series applications. The method is also well suited for more complex restrictions like monotonocity of a function (Lin and Dunson, 2014). The projections are straightforward and can be efficiently calculated via pooled Adjacent violators algorithm (Barlow, 1972).

Inequality restrictions on odds ratio: Odds ratio is a popular measure of association between two categorical variables. Agresti and Coull (2002) mentions four kinds of odds ratio, namely local odds ratio (LO), cumulative odds ratio (CU), continuation odds ratio (CO) and global odds ratio (GO). The condition of uniformly nonnegative log CO is equivalent to the first order stochastic condition which is discussed in previous paragraph. Each of the odds ratio can be viewed as generalization of a particular logit or loglinear model (see El Barmi and Dykstra (1998) for details). As an example the cumulative logit model in Equation 2.13 provides a stochastic ordering of the levels of X when β_is are monotone. Therefore enforcing odds ratio constraint on a categorical data is equivalent to enforcing monotonocity on the parameters of a generalized linear model as discussed before.

$$\text{logit}(\theta_{ij}/\theta_{i+}) = \alpha_j - \beta_i. \tag{2.13}$$

Norm and hyperplane restrictions: Two important examples of norm restrictions on parameters are ridge regression and lasso regression, which fits a linear regression model under the restriction that $\| \beta \|_2 \leqslant \lambda$ and $\| \beta \|_1 \leqslant \lambda$ respectively, for some constant $\lambda > 0$. The norm-0 restriction $\| \beta \|_0 \leqslant \lambda$ has also gained popularity as an effective sparsity inducing method; see (Fu, 1998) for a detailed discussion on these methods. A linear equality or inequality constraint such as lasso is equivalent to specifying a hyperplane whereas a quadratic restriction such as ridge requires a spherical projection. Many such norm regularization and penalty constraints, like elastic net (Zou and Hastie, 2005), are equivalent to simple projection onto closed convex sets. Therefore our method can be applied by choosing suitable conjugate priors and appropriate efficient algorithms for projection. The posterior projection framework has potential advantages in terms of uncertainty quantification and use of a Bayesian approach for tuning parameter choice avoiding cross validation.

2.7 Summary

We have proposed a new methodology to do Bayesian analysis under the constrained parameter space. Often a truncated prior is the go to choice in such scenarios which leads to intractable posteriors, in other cases sampling from the posterior can be challenging. Our method avoids overcomes these problems by initially putting a prior on the parameter space that ignores the constraint. The resultant posterior is usually analytically tractable and easy to sample from. These samples are then projected back to the constrained space via a minimal distance mapping. We discuss the choice of such metric projection operator and provide a subjective Bayesian justification for the projection method. We show that under suitable assumptions this technique achieves equivalent consistency rate as that of the unrestricted posterior. The advantage of the projection method against truncated priors is then illustrated via a simulation example. We proceed to apply our method on a couple of real world data sets that have very different models and constraint structure and show that in both cases the projection method performs well and provides meaningful interpretation. In the next section we consider Stiefel manifolds, a special class of non-convex spaces.

3

Posterior Projection onto Stiefel manifold

3.1 Background and notation

The Stiefel manifold is an important space which arises naturally in modeling of subspaces and in applications such as human activity modelling (Veeraraghavan et al., 2005), video based face recognition (Aggarwal et al., 2004), and shape analysis (Goodall and Mardia, 1999). Edelman et al. (1998) give a detailed overview of the well studied geometric properties of this manifold. For our purpose, we restrict our attention to finite dimensional Stiefel manifolds on $\mathbb{R}^m$.

Definition 2. *Let $p, m \in \mathbb{N}, 1 \leqslant p \leqslant m$. Then the Stiefel manifold $St(p, m) = \widetilde{\Theta}$ is the subset of $\mathbb{R}^{m \times p} = \Theta$ consisting of all orthonormal p-frames. Mathematically,*

$$St(p, m) = \{\theta \in \mathbb{R}^{m \times p} \mid \theta^T \theta = I_p\}. \tag{3.1}$$

For example, when $p = m$, $\mathrm{St}(m, m)$ coincides with $O(m)$, the set of all real orthonormal matrices of size m. When $p = 1$, $\mathrm{St}(1, m) = \mathbb{S}^{m-1}$, the $m-1$ dimensional manifold of unit circle in $\mathbb{R}^m$. It follows that $\mathrm{St}(p, m)$ is a closed subspace of the p-fold Cartesian product of $\mathbb{S}^{m-1}$ as $\mathrm{St}(p, m)$ is given by p unit vectors $\theta_1, \ldots, \theta_p \in \mathbb{R}^m$ such that $\langle \theta_i, \theta_j \rangle = 0$ for all $i \neq j$.

3.2 Properties of the projection

It is evident from the definition 2 that $\mathrm{St}(p, m)$ is not a convex set. However, the Stiefel manifold is a smooth embedded closed submanifold of $\mathbb{R}^{m\times p}$, which itself is a Hilbert space endowed with the inner product $\langle \theta_1, \theta_2 \rangle = \mathrm{trace}\,(\theta_1^T \theta_2)$. Therefore one may define a probability on $\mathbb{R}^{m\times p}$ and project it to $\mathrm{St}(p, m)$. The following result ensures the validity of this technique.

Lemma 7. *For almost every $\theta \in \mathbb{R}^{m\times p}$ there exists a unique projection $T\theta \in St(p, m)$.*

The implication of this proposition is that for any posterior measure $\Pi_{\Theta}^{(n)}$ that is locally absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{m\times p}$, the projected posterior measure $\widetilde{\Pi}_{\tilde{\Theta}}^{(n)}$ is well defined on $\mathrm{St}(p, m)$. It is also possible to characterize the set for which the projection is unique.

Lemma 8. *Only the set of matrices with full column rank p, denoted Θ_p, has unique projection on $St(p, m)$.*

$$\Theta_p = \{\theta \in \mathbb{R}^{m\times p} \mid rank(\theta) = p\}. \tag{3.2}$$

This follows immediately from Bardelli and Mennucci (2017), proposition 4.8. This result depicts the importance of choosing a lower value of p, if possible, while modeling using the posterior projection method because the dimension of Θ_p, given by mp, grows linearly with p. Intuitively it becomes harder to find p independent vectors in $\mathbb{R}^m$ as p grows larger. Trivially $\mathrm{St}(p, m) \subset \Theta_p$.

Proposition 9. *Θ_p is an open set containing $St(p, m)$.*

Because Θ_p is an open set every matrix close to a matrix in the Stiefel manifold has full rank. Specifically, if $\| \theta - \tilde{\theta} \| < 1$, for some $\tilde{\theta} \in \mathrm{St}(p, m)$ then θ has rank p (Absil and Malick, 2012). Thus if our posterior converges to a point on the Stiefel manifold every point in the vicinity is guaranteed to have a unique projection.

As we noted earlier, $\mathrm{St}(p,m)$ is not convex. Therefore we are going to focus on its convex hull, which is defined as the smallest convex set of the vector space $\mathbb{R}^{m\times p}$ that contains $\mathrm{St}(p,m)$ and is denoted as $\mathrm{conv}(\mathrm{St}(p,m))$. The closed convex hull is the closure of its convex hull.

Theorem 10 (Journe et al. (2010))**.** *The convex hull of the Stiefel manifold is the closed unit spectral ball.*

The unit spectral ball is given by

$$\mathbb{B}_{\mathrm{Sp}}(p,m) = \{\theta \in \mathbb{R}^{m\times p} : \|\,\theta\,\|_2 \leqslant 1\} = \{\theta \in \mathbb{R}^{m\times p} : \theta^T\theta \preceq I_p\}, \tag{3.3}$$

where the spectral norm $\|\,\theta\,\|_2$ is the largest singular value of the matrix θ. The boundary of the spectral ball is given by all matrices with largest singular value equal to 1. Because any element of $\mathrm{St}(p,m)$ has exactly p singular values all of which are equal to 1, the Stiefel manifold is a subset of the boundary of its convex hull. Interestingly, it is also the subset of the boundary of $\mathbb{B}_{\sqrt{p}}$, the closed centered ball of matrices in $\mathbb{R}^{m\times p}$ with Frobenius norm less than or equal to $p^{1/2}$. When $p=1$, these two balls coincide. This is important as the metric projection onto $\mathrm{St}(p,m)$ tries to minimize the Frobenius norm, for any $\theta \in \Theta$,

$$T\theta = \underset{\tilde{\theta}\in\mathrm{St}(p,m)}{\arg\min}\|\,\theta-\tilde{\theta}\,\|^2 = \underset{\tilde{\theta}\in\mathrm{St}(p,m)}{\arg\max}\ \mathrm{trace}(\theta^T\tilde{\theta}).$$

From this equivalent representation one can obtain a closed form solution for the projection which ties the spectral and Frobenius ball.

Proposition 11. *Let UDV^T be a thin singular value decomposition (SVD) of $\theta \in \Theta_p$, that is U is a $m \times p$ matrix. Then the projection is given by $T\theta = UV^T$.*

A proof of this proposition can be found in Absil and Malick (2012), proposition 7. An immediate consequence of the result is that the projection is unaltered by a change in scale. This is because for a non-negative constant $c > 0$, a thin SVD of $c\theta$ is given by $U(cD)V^T$. In particular $\theta / \| \theta \|_2$ has the same projection as θ and belongs to $\mathbb{B}_{\mathrm{Sp}}(p, m)$ as its maximum eigenvalue is 1. We will use this trick to scale our samples of U prior to projection in the brain analysis example (section 3.4) for better estimation and prediction.

3.2.1 Convergence on Stiefel manifold

Even though Θ is a Hilbert space in this case, the projection operator T on $\mathrm{St}(p, m)$ is not non-expansive. This property was a key component of establishing posterior consistency in a convex subset $\widetilde{\Theta}$ of a Hilbert space Θ (section 2.3.1). Let us consider the following example.

Example 1. *Consider three matrices in* $\mathbb{R}^{2\times 2}$

$$\theta_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \theta_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \textit{and} \quad \theta_3 = \begin{bmatrix} 0.25 & 0.75 \\ 0.75 & 0.25 \end{bmatrix}$$

It is easy to see that $\theta_1, \theta_2 \in St(2, 2)$. *The matrix* θ_3 *has singular values* 1 *and* 0.5 *and therefore does not belong to* $St(2, 2)$. *In fact* $T\theta_3 = \theta_2$ *which satisfies the following inequality.*

$$\| T\theta_3 - T\theta_1 \| = \| \theta_2 - \theta_1 \| = 2 > 1.5 = \| \theta_3 - \theta_1 \|.$$

Therefore, T *is not non expansive.*

However, we can use the triangle inequality to establish that T is Lipschitz continuous mod $\mathrm{St}(p, m)$.

Lemma 12. $\| T\theta - \tilde{\theta} \| \leqslant 2 \| \theta - \tilde{\theta} \|$ *for any* $\theta \in \mathbb{R}^{m\times p}$, $\tilde{\theta} \in St(p, m)$.

Coupled with Lemma 7 and 8 this gives us our desired consistency result. Interestingly, even when θ does not have a unique projection, Lemma 12 holds for every element of $T\theta$. Therefore, for convergence purposes any projection of $\theta \notin \Theta_p$ suffices.

Theorem 13. *If the true matrix $\theta_0 \in St(p,m)$, then under Assumption 4 the concentration rate of the projected posterior is at least that of the original posterior.*

3.3 Metric projection onto measure 0 sets

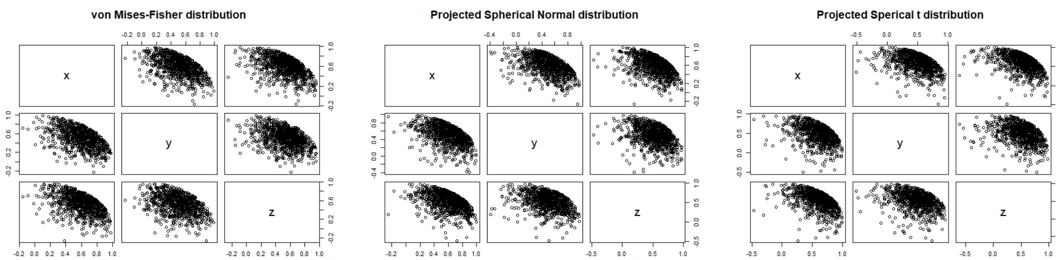
It is clear from Fig. 2.1 that the metric projection T maps elements outside the constrained region to the boundary. One might wonder what happens when the constrained parameter space has measure 0, so that $\Pi_{\Theta}^{(n)}\widetilde{\Theta} = 0$. However, our object of interest is the projected posterior under which the constrained space has measure 1, as $\widetilde{\Pi}_{\widetilde{\Theta}}^{(n)}\widetilde{\Theta} = \Pi_{\Theta}^{(n)}\Theta = 1$. This argument extends to any subset $B \in \widetilde{\Theta}$, that is even if $\Pi_{\Theta}^{(n)}B = 0$, $\widetilde{\Pi}_{\widetilde{\Theta}}^{(n)}B \geqslant 0$. Therefore it is of clear importance to properly identify the underlying probability measure. We illustrate this via modeling on $\mathrm{St}(1,m)$, the $(m-1)$ dimensional unit sphere $\widetilde{\Theta} = \mathbb{S}^{m-1}$ on $\Theta = \mathbb{R}^m$, with $m = 3$. Let

$$x_i \sim N(\theta, 10I), i = (1, .., 100), \quad \theta \sim \mathrm{VMF}(\mu, \phi^{-1}I),$$

where $\mathrm{VMF}(\mu, \phi^{-1}I)$ is the von Mises-Fisher distribution with $\mu, \theta \in \widetilde{\Theta}$. The resulting posterior is $\mathrm{VMF}(\mu_n, \phi_n^{-1}I)$ where

$$\phi_n = \| n\bar{x} + \phi\mu \|_2, \quad \mu_n = (n\bar{x} + \phi\mu)/\phi_n$$

Alternatively, one can leverage our projection approach by using an unconstrained Gaussian prior $N(\mu, \phi^{-1}I)$ on θ and projecting the posterior samples onto $\widetilde{\Theta}$ via the Euclidean norm. Additionally, projection makes it easy to consider different unconstrained prior distributions, such as t-densities to allow heavier tails. Figure 3.1 demonstrates posterior samples using von Mises–Fisher, projected spherical normal and projected spherical t distributions.

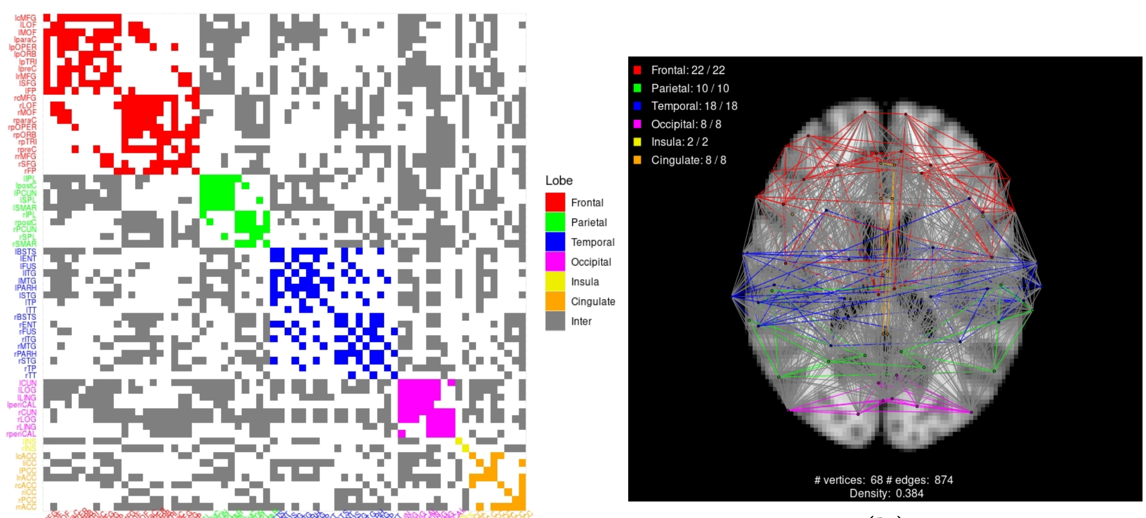


(a) (b) (c)

FIGURE 3.1: Pairs plot of 1000 random samples from different constrained posterior distributions on a unit sphere on $\mathbb{R}^3$. Data are generated from a Gaussian N(θ, $10I$) with $\theta = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]$. The samples are drawn from the posterior with (a) the prior distribution VMF(μ, $0.1I_3$), projecting onto unit sphere after sampling with the prior distribution (b) N(μ, $0.1I_3$), (c) t_3(μ, $0.1I_3$) where $\mu = \theta$.

3.4 Experiment: Analyzing brain network

An emerging paradigm in neuroscience is that cognitive tasks are performed by brain networks that consist of several anatomically separate but functionally linked regions, sharing information with each other. The study of the physical and functional connections between different brain regions, called the "connectome" coined by Sporns et al. (2005) is facilitated by modern noninvasive imaging techniques. These allow the mapping of anatomical regions and their interconnecting pathways, at near-millimeter spatial resolution (Shi and Toga, 2017; Wang et al., 2014). Analysis of fMRI time series data is a popular approach for constructing the brain network. Several such studies have shown a high level of functional connectivity between well known brain regions, like primary visual and auditory network (Cordes et al., 2002; Fox and Raichle, 2007). Structurally this large amount of functional data is believed to be transported by white matter tracts, which are bundles of millions of long distance axons that directly interconnect large groups of spatially separated neurons. Free water molecules are known to diffuse along a strong preferred direction in white matter tracts. Therefore data from diffusor tensor imaging (DTI), which maps the diffusion profile of free water molecules in brain tissue have become increasingly popular in this field. For the current study we use the DTI sequence data from the well known KKI-42 (Landman et al., 2011) dataset. It consists of two scans for $N = 21$ volunteers without any prior history of neuropsychiatric diseases. Following the routine of scan-rescan methods we use the first scan to estimate model parameters and validate the model on the second scan. Recent connectome preprocessing pipelines (Craddock et al., 2013; Roncal et al., 2013) were used on the DTI imaging to produce an $m \times m$ symmetric adjacency matrix X_n for every individual $n = 1, .., N$. In our application $m = 68$ and a particular node $i = 1, ..., m$ characterizes a specific brain region according to the Desikan et al. (2006). $X_{nij} \in \{0, 1\}$ is an indicator of a white fiber connection between brain regions $i = 2, ..., m$ and $j = 1, ..., i - 1$ for the nth individual. $X_{nii} = 0$ for all $i = 1, ..., m$ and we use only the lower triangular adjacency matrices for model fitting and validation.



(b)

(a)

FIGURE 3.2: Illustration of a sample of the data displaying (a) adjacency matrix and (b) brain network obtained from the first scan of the first volunteer. The brain regions and the associated lobe membership are chosen according to the Desikan atlas. Both the plots were generated by the BrainGraph package in R.

Following Hoff (2008) we model the brain network as the following reduced-rank logit model:

$$X_{nij} \sim Bern(\pi_{nij}), \quad \log\left(\frac{\pi_{nij}}{1-\pi_{nij}}\right) = Z_{ij} + u_i^T \Lambda_n u_j \ , \tag{3.4}$$

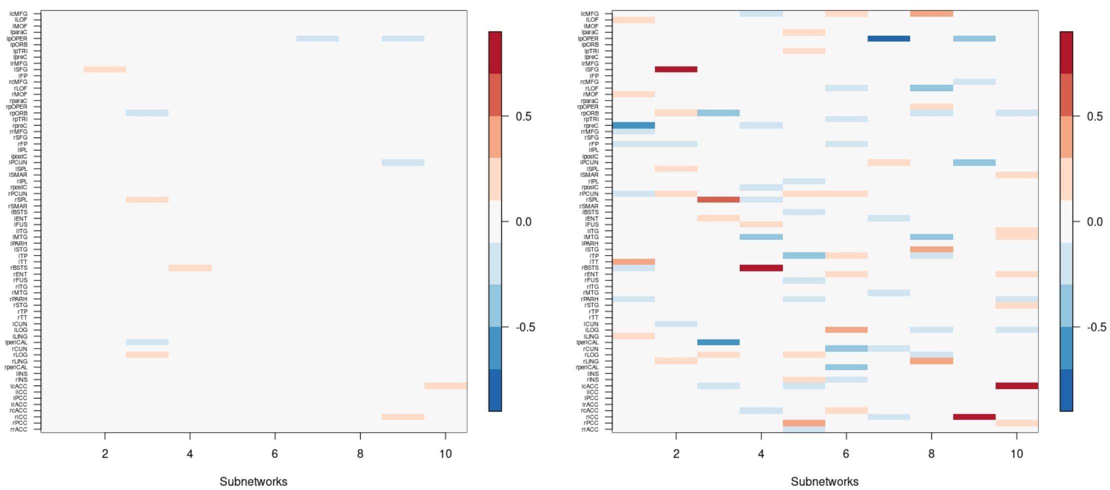
where $\Lambda_n = \text{diag}(\Lambda_{n1}, ..., \Lambda_{np})$ and u_i are p-dimensional column vectors. We choose $p = 10$ as an upper bound on the number of subnetworks consistent with our discussion in section 3.1. The probability of a link between node i and j in the connectivity graph for the nth individual depends on their similarity in a space of unobserved latent characteristics. The latent factor $\Lambda_{nk}, k = (1, ..., p)$ is a person specific effect on brain subnetwork k and u_{ik} represents the effect of brain region i on subnetwork k. Z_{ij} is the baseline log-odds of a connection between the (i, j) pair of brain regions. The singular value decomposition like representation of the dataset in equation (3.4) naturally suggests a constraint of orthonormality on u_i. This restriction is necessary for the model to be identifiable which is important for interpretation purposes (Hoff, 2016). We remove rotational and scaling ambiguity by assuming that the matrix U with row vectors $u_1, ..., u_m$ lies on the Stiefel manifold $\text{St}(p, m) = \{U : U^T U = I_p\}$. Instead of enforcing this sharp constraint, Duan et al. (arXiv:1801.01525v2) considers a constraint relaxed model on the data and therefore does not have exact orthogonality on U.

On average we found only 36% of the brain regions to be connected across all the subjects in the dataset, a significant portion of which are intra-lobe connections. Because of the relatively low number of connections, it is reasonable to assume only a few brain regions will contribute to a particular subnetwork. Therefore, we induce sparsity on the brain subnetworks, parameterised by the factor matrix U, through shrinkage priors. However, the well known matrix Bingham-von Mises–Fisher (BMF) family of priors on $\text{St}(p, m)$ does not induce much shrinkage. Therefore, we take advantage of our method by initially ignoring the Stiefel manifold restriction and putting independent Dirichlet-Laplace (DL) shrinkage priors on the columns of U (Bhattacharya et al., 2015). The hyperparameters for the inverse gamma are chosen so that the priors are weakly informative. As advised in the source literature, the value of the parameter a in the DL prior was fixed at $1/2$ and $1/m$ in two different runs of the model.

$$\begin{aligned} Z_{ij} &\sim \mathrm{N}(0, \sigma_Z^2), \qquad \Lambda_{nk} \sim \mathrm{N}(0, \sigma_k^2), \quad u_k \sim \mathrm{DL}_a \\ \sigma_Z^2 &\sim \mathrm{IG}(2,1), \qquad\quad \sigma_k^2 \sim \mathrm{IG}(2,1), \quad k = (1, ..., p) \end{aligned} \tag{3.5}$$

The model was fit in RStan using the No U-Turn Sampler algorithm. We discarded the first 5000 draws of 10000 iterations as burn-in. The convergence and mixing rates of the sampler were adequate. To remove scaling ambiguity, posterior samples of U were normalized to the unit spectral ball prior to projection. These norm values were absorbed in the factor loading matrix Λ to obtain $U_l^{(s)} = U^{(s)}/\| U^{(s)} \|_2$ and $(\Lambda_n)_l^{(s)} = \Lambda_n^{(s)} * \| U^{(s)} \|_2^2$, respectively. These $U_l^{(s)}$ samples are then projected onto the Stiefel manifold and multiplied by $\| U^{(s)} \|$ to get $\widetilde{U}^{(s)}$. In the event $U_l^{(s)}$ does not have a unique projection we choose any one of its projections as per our discussion in Lemma 12. The projection routine was implemented using the Manopt package from Matlab (Nicolas et al., 2014). The projection estimate for the logit function of π_{nij} from the sth sample is given by $Z_{ij}^{(s)} + \tilde{u}_i^{(s)T} (\Lambda_n)_l^{(s)} \tilde{u}_j^{(s)}$.

The brightly colored regions along the diagonal in Fig. 3.2(a) represent the intra-lobe connections. Brain regions within a single lobe, represented by a particular color, are spatially close and therefore are generally connected among most subjects. These, along with some frequent inter-lobe connections, are accounted for by the Z matrix in (3.4). The role of the factor matrix U is to reveal the more obscure relationship present in the data beyond the spatial association. With strong shrinkage our method clearly detects subnetwork membership as opposed to the unrestricted method (Fig. 3.3). In the absence of the orthogonality constraint the posterior sample means of the elements of the factor matrix u_{ik} are scattered around 0. The projection technique maps the very small values to 0 while inflating others, bringing out the subnetwork structure. The differences in the results is depicted in Fig. 3.3.



(a) Unrestricted **(b) Projected**

FIGURE 3.3: Posterior sample mean of the factor matrix U (a) before projection and (b) after projection under $\mathrm{DL}_{1/m}$ prior. The columns are sorted according to decreasing value of σ_k^2. The rows are rearranged to match Desikan atlas as in Fig. 3.2.

The prominent subnetwork structure in the factor matrix of the projection method allows us to interpret the connectomes that are not spatially associated. As an example, the 3rd factor under the projected model puts brain regions Superior Parietal, Pars Orbitalis, Pericalcarine and Lateral Occipital Gyrus in the same subnetwork. These brain regions have been associated with spatial orientation, visual and sensory input from hands, processing of syntax in the oral and sign language, musical syntax, primary visual cortex (Coullon et al., 2015) and face perception (Nagy et al., 2012) and hence form a sensible subnetwork. With $\mathrm{DL}_{1/2}$ prior the projection still retains its utility by forcing small values of u_{ik} towards 0, but it fails to indicate any strong membership.

Table 3.1: Estimation and cross validation area under the curve (AUC) measure for the KKI-42 dataset under different Dirichlet–Laplace priors.

Prior	$\mathrm{DL}_{1/m}$		$\mathrm{DL}_{1/2}$	
	Unrestricted	Projected	Unrestricted	Projected
Fitted AUC	96·8	93·8	98·8	88·2
Prediction AUC	95·8	91·3	95·0	86·1

One can build a classifier algorithm based on the posterior probabilities π_{nij} by estimating X_{nij} to be 1 above a threshold value and 0 otherwise.We validate our model by testing the accuracy of this binary classifier against the held out second scan. As the results depend on the chosen cutoff level, we estimate the receiver operating characteristic curves (ROC) and compare methods based on area under the curve (AUC). With strong shrinkage the projected model has similar predictive performance as the unrestriced model, but falls of substantially otherwise (Table 3.1). This is expected as the projection method is designed to enhance interpretability through constraints, and such interpretability may have a cost.

3.5 Summary & future Directions

This article has developed the fundamental underpinning for a general new approach for constrained Bayesian inference. There are many interesting new directions. One important generalization is to allow conditional posterior projections; for example, applying projection to a subset of parameters immediately after each update step within a Markov chain Monte Carlo algorithm. Another direction is studying when projected posteriors are expected to produce better results than traditional approaches; for example, obtaining lower risk parameter estimates. Also of considerable interest is the construction of formal goodness-of-fit test of the constraint, as well as adaptive procedures that avoid projecting when the cost to goodness-of-fit is too high.

4

Posterior Projection in Small Area Estimation

4.1 Introduction

The steady increase in popularity of small area estimation (SAE) is due to a growing demand for more precise and accurate estimation of population characteristics of interest from survey data (Rao and Molina, 2015). Here an "area" refers to a geographic domain, a demographic group or an experimental condition. A typical SAE problem involves simultaneous estimation of several parameters, each of which is associated with an "area". Traditional survey based methods provide estimates that are based only on the sample observations corresponding to the particular population characteristic. For a domain with a small sample size, the direct small area estimators have large standard errors and coefficients of variation (Prasad and Rao, 1990). On the other hand, some domains might have no observations in the sample. One way to increase the precision for these small area estimates is to use hierarchical and empirical Bayesian models that borrow information from related areas. The need for such methods have been successfully applied in many areas such as sociology, epidemiology, political science, business and insurance (Rao and Molina, 2015), with recent advances in well-known approaches like smoothed ridge regression and elastic net (Wehbe et al., 2014).

One problem that arises in practice is that the model-based estimates vary wildly from the direct estimates. Moreover the model based estimates for small areas usually do not aggregate to the more reliable direct survey estimates for a larger area. This is because original survey was designed to achieve a specific accuracy at a higher level of aggregation than that of the small areas. More severe complications can arise in the event of model failures as often there is little to no check for validity of the assumed model. This agreement is often a political necessity to convince the legislators of the utility of such models (Fay and Herriot, 1979). The process of adjusting model-based estimates by imposing consistency constraints is known as "benchmarking". One simple example of such a constraint is to require that the weighted average of the model-based small area estimates to match a known value, typically the direct estimates at a higher aggregation level. The benchmarked estimators have been shown to adapt quicker than model-based estimates in the event of an external shock not accounted for in the model, we refer to Pfeffermann and Tiller (2006) for an example in the context of time series models. In many applications, it is also desirable that the estimates vary smoothly across areas, so that nearby areas will have similar parameters. This process is known as"smoothing".

Steorts (2014) provides a unified decision-theoretic approach to smoothing and benchmarking by regarding them both as constraints on Bayesian estimates. Benchmarking refers to equality constraints on weighted averages and variances and smoothing corresponds to an inequality constraint on the "roughness" of the model-based estimates. The constraints are incorporated into the model by adding appropriate penalty terms to the objective function. This approach is similar in nature to that of the ridge regression and LASSO, where one imposes a L_2 and L_1 penalty on the coefficients respectively. The smoothed, benchmarked estimators are the minimizer of the Bayes risk subject to equality and inequality constraints. For a quadratic objective functions and linear constraints the estimators can be solved in closed form, using only basic matrix operations on the unconstrained Bayes estimates. For more complex constraints such as a quadratic equality restriction on variance of the model estimates, no closed form solution exists in general. However, as long as the objective function is convex, a solution exists which can be obtained using numerical optimization techniques.

Geometrically, the smoothed and benchmarked Bayes estimates are projection of the model-based estimates onto the feasible set characterized by the equality and inequality constraints. Therefore, the benchmarking and smoothing constraints can be equivalently viewed as restrictions on the parameter space itself. This is the basis of the posterior projection framework. Instead of projecting the model-based estimates, we project the posterior draws from the model. The projected samples induce a density on the constrained space which can be used to not only get point estimates but also a credible interval for the population parameters of interest. For linear constraints one can directly show that the estimates obtained from posterior projection approach converge almost surely to the smoothed, benchmarked estimator. Similar to benchmarking and smoothing, this method does not need any distributional assumptions on the data or on the unconstrained estimators. Leveraging on the generality of the posterior projection approach, the method can be extended to include mode complex restrictions as long as the constrained set is closed and convex. Additionally, one can choose a different loss function other than quadratic/ Euclidean, such as the absolute deviation, if it is more suitable for the task at hand. For simplicity, in this chapter we focus on illustrating and comparing the methods with quadratic loss and smoothing functions under linear equality constraints. This extends the approaches of Datta et al. (2011); Ghosh and Steorts (2013) which themselves were influenced by Louis (1984) and Ghosh (1992).

The rest of the chapter is organized as follows. In section 2, we introduce the general framework for smoothing. Section 2.1 provides the notations and terminologies used for the rest of this chapter. Section 2.2 illustrates the equivalency of the penalized objective function and smoothing constraints. In section 3, we extend the proposed framework to include linear benchmarking constraints. Section 3.1 derives the closed form estimator under equality constraints and section 3.2 illustrates methods to reduce an overdetermined equality constraint to that of section 3.1. In section 4, we introduce the posterior projection framework and derive the theoretical results justifying such an approach. Section 5 illustrates and compares the two methods via an extensive simulation study based on area-level model of Fay and Herriot (1979). Finally, in section 6 we apply the method on a real dataset. Section 7 concludes with a discussion and future directions.

4.2 Loss function and smoothing

In this section, we introduce the decision-theoretic framework for smoothing and Benchmarking in small area estimation along with relevant notations which is used throughout the paper. This generalized constrained Bayesian approach builds upon the work of Datta et al. (2011) and Ghosh and Steorts (2013) and extends the methodology without imposing distributional assumptions. We also demonstrate the theoretical findings under the setup of area-level Fay-Herriot model.

4.2.1 Notation and terminology

We assume m areas, and for each area i, we want to estimate the associated scalar quantity θ_i. The unknown quantity $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ represent our parameter of interest. We denote the response variables, which are often available in the form of direct estimators, by $\boldsymbol{y} = (y_1, \ldots, y_m)$. For each area, we also have a $p \times 1$ vector of covariates denoted by $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Collectively, we denote the matrix of covariates as follows:

$$X_{m\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mp} \end{pmatrix}.$$

Our goal is to estimate the unknown parameter $\boldsymbol{\theta}$, by an estimator $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_m)$. From a decision-theoretic perspective this is achieved by defining a loss function $L(\boldsymbol{\theta}, \boldsymbol{\delta})$ and minimizing the corresponding posterior risk $E(L(\boldsymbol{\theta}, \boldsymbol{\delta})| \, \boldsymbol{y})$ with respect to $\boldsymbol{\delta}$. The optimal estimator is denoted by $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \hat{\delta}_2, \dots, \hat{\delta}_m)$. We discuss the loss functions that are used in section 4.2.2.

4.2.2 Loss functions

In this section, we consider loss functions under the assumption that there are no benchmarking constraints. Most loss functions combine piecewise deviance measures to determine the optimal estimator. The choice of the deviance measures, and thus the loss function is application specific, see Williams and Savitsky (2019) for a detailed discussion on this topic. For convenience, we assume a weighted squared error loss. Let $\phi_i > 0$ be the weight for area i. Then, the weighted quadratic loss function has the following form:

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{i=1}^{m} \phi_i (\theta_i - \delta_i)^2 = (\boldsymbol{\theta} - \boldsymbol{\delta})^T \Phi (\boldsymbol{\theta} - \boldsymbol{\delta}), \tag{4.1}$$

where $\Phi = \mathrm{diag}(\phi_1, \dots, \phi_m)$, is a measure of "closeness" between the parameter and the estimator. The choice of the loss function weights is application specific. In many SAE applications, the weights Φ are unknown and can be estimated from the survey design (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis et al., 2018). Isaki et al. (2004) proposed inverse-variance estimator, where each weight is the reciprocal of the posterior variance of the Bayes estimator. We refer to Datta et al. (2011) for a complete review of popular choices of loss function weights in practice. In our proposed methodology, instead of choosing a specific set of weights, we simply assume that a Bayes estimator exists. The informative choice of not imposing additional distributional assumptions makes the modeling structure quite flexible.

4.2.3 Smoothing

In many applications, it is desirable that the estimates vary smoothly over neighbouring areas. To achieve this, Steorts et al. (2019) introduced an additional penalty term in the loss function of (4.1). This reduces the overall variance of the estimator and will be referred to as "smoothing" throughout this chapter. Let $q_{ij} \geqslant 0$ denote a set of smoothing weights for area i and j. We choose a weighted sum of squared differences between elements, $\sum_{i,j=1}^{m} (\delta_i - \delta_j)^2 q_{ij}$, as the smoothing penalty. The smoothing weights $q_{ij} \geqslant 0$ gauges how important it is that the estimate of θ_i be close to the estimate of θ_j. Similar to loss function weights, the choice of smoothing weights q_{ij} are also application specific. It can be discrete corresponding to clustering of areas, or continuous-valued, corresponding to a metric space of areas. It is often the case that the smoothing weights are directly dependent on the covariates, that is $q_{ij} = q(x_i, x_j)$.

Note that the smoothing penalty does not depend on the parameter $\boldsymbol{\theta}$. This specific choice allows us to get a closed form solution of our proposed estimator, which greatly facilitates computation. The smoothing term can also be written as $\sum_{i,j} (\delta_i - \delta_j)^2 q_{ij} = \boldsymbol{\delta}^T \Omega \boldsymbol{\delta}$, for a suitable symmetric positive semi-definite matrix Ω (see Steorts and Louis (2014) for the equivalence). Therefore, we want to minimize the posterior risk of the smoothed objective function

$$
\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{\delta}) &= \sum_{i=1}^{m} \phi_i(\theta_i - \delta_i)^2 + \gamma \sum_{i,j} (\delta_i - \delta_j)^2 q_{ij} \\
&= (\boldsymbol{\theta} - \boldsymbol{\delta})^T \Phi (\boldsymbol{\theta} - \boldsymbol{\delta}) + \gamma \boldsymbol{\delta}^T \Omega \boldsymbol{\delta}.
\end{aligned}
\tag{4.2}
$$

The scalar quantity $\gamma \geqslant 0$ is called the penalty factor and is chosen to specify the overall importance of smoothness. Minimizing posterior risk of the loss function in equation (4.1) is equivalent to minimizing

$$
E\left[(\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B)^T \Phi (\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B) \mid \boldsymbol{y}\right] + \gamma \boldsymbol{\delta}^T \Omega \tag{4.3}
$$

Datta et al. (2011); Steorts (2013) show that this can be simplified to as the minimization of

$$
(\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B)^T \Phi (\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B) + \gamma \boldsymbol{\delta}^T \Omega \boldsymbol{\delta} \tag{4.4}
$$

where $\hat{\boldsymbol{\theta}}^B = (\hat{\theta}_1^B, \ldots \hat{\theta}_m^B)$ is the mean of the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. In practice $\hat{\boldsymbol{\theta}}^B$ is unknown and approximated by MCMC techniques. This equivalence makes the quadratic loss function one of the most popular choices among the class of smooth deviance measures. It is differentiable everywhere and thus admits a closed-form solution.

4.3 Benchmarking

We now focus on the situation where the estimates have to satisfy benchmarking constraints, in addition to being smooth as in section 4.2.2. The process of modifying model-based estimates so that they agree with direct estimators on an aggregated level is known as benchmarking. This is generally achieved by adding equality constraints on the weighted means or weighted variances of subsets (possibly all) of the estimates to the loss function in (4.4). Our proposed method extend linear, weighted mean constraints, as in Datta et al. (2011); Ghosh and Steorts (2013) to allow for linear inequality constraints in the context of both smoothing and posterior projections.

Definition 3. *(Benchmarking constraints, benchmarked Bayes estimator). Benchmarking constraints are equality or inequality constraints on the weighted means or weighted variances of subsets (possibly all) of the estimates. The benchmarked Bayes estimator is the minimizer of the posterior risk subject to the benchmarking constraints.*

Let us denote the k linear equality constraints as $\sum_{i=1}^{m} w_{ji}\delta_i = \ell_i$ for $1 \leqslant j \leqslant k$. In matrix notation this can be written as $W\boldsymbol{\delta} = \boldsymbol{\ell}$ where (j,i)th element of $W \in \mathbb{R}^{k\times m}$ is w_{ji}. Without loss of generality we assume $k \leqslant m$, otherwise the system of linear equation is overdetermined and we reduce the number of equality constraints as discussed in section 4.3.2. Similarly, we can represent the inequality constraints as $F\boldsymbol{\delta} \leqslant h$ with $F \in \mathbb{R}^{p\times m}$. Hence, our new objective function to incorporate both smoothing and benchmarking is the following:

$$\underset{\boldsymbol{\delta}}{\text{minimize}} \quad (\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B)^T \Phi(\boldsymbol{\delta} - \hat{\boldsymbol{\theta}}^B) + \gamma \boldsymbol{\delta}^T \Omega \boldsymbol{\delta} \qquad \text{subject to} \quad W\boldsymbol{\delta} = \boldsymbol{\ell}, F\boldsymbol{\delta} \leqslant h \tag{4.5}$$

where $W \in \mathbb{R}^{k\times m}$, $F \in \mathbb{R}^{p\times m}$, $\boldsymbol{\ell} \in \mathbb{R}^k$ and $f \in \mathbb{R}^p$. The benchmarked Bayes estimator $\hat{\boldsymbol{\delta}}^{BM}$ is the solution of the equation (4.5). The levels to which we benchmark, i.e., the values of the $W, \boldsymbol{\ell}, F, h$, are assumed to be given *externally* from some other data source. (For internal benchmarking, we refer to Bell et al. (2013).)

4.3.1 Equality constraints

In the absence of linear inequalities, (4.5) simplifies to a Linear Equality Constrained Quadratic (LECQ) problem which has been considered in Steorts et al. (2019). In the context of our approach, we consider a more generalized constrained quadratic optimization problem:

$$\underset{\boldsymbol{\delta}}{\text{minimize}} \quad \frac{1}{2}\boldsymbol{\delta}^T B\boldsymbol{\delta} - \boldsymbol{\delta}^T\boldsymbol{b} \quad \text{subject to} \quad W\boldsymbol{\delta} = \boldsymbol{\ell} \tag{4.6}$$

where $B \in \mathbb{R}^{m\times m}$ is symmetric, $\boldsymbol{b} \in \mathbb{R}^m$, $W \in \mathbb{R}^{k\times m}$ and $\boldsymbol{\ell} \in \mathbb{R}^k$. The Karush-Kuhn-Tucker (KKT) conditions for the optimal solution $\hat{\boldsymbol{\delta}}$ to (4.6) give rise to the following system of linear equations:

$$\underbrace{\begin{pmatrix} B & W^T \\ W & 0 \end{pmatrix}}_{K} \begin{pmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\ell} \end{pmatrix} \tag{4.7}$$

where $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^k$ is the associated Langrange multiplier from solving (4.7). We make the following assumptions.

Assumption 5. *$W \in \mathbb{R}^{k\times m}$ has full row rank $k \leqslant m$.*

Check section 4.3.2 to see what happens when W is not full row rank. Let us denote by Z the matrix whose columns span the null space of W, that is $WZ = 0$.

Assumption 6. *B is positive definite in the null space of W, that is the reduced hessian $Z^T BZ$ is positive definite.*

Under these assumptions the KKT matrix K is non-singular and hence the linear system has a unique solution $(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\lambda}})$. It is easy to see that if B is positive definite then $Z^T BZ$ is also positive definite, and hence $\hat{\boldsymbol{\delta}}$ is a global minimizer.

Theorem 14. *In the presence of only linear equality constraints, there exists a unique solution to the benchmarked and smoothed objective function in* (4.5). *The solution is given by*

$$\hat{\boldsymbol{\delta}}^{BM} = B^{-1}\left[\boldsymbol{b} + W^T(WB^{-1}W^T)^{-1}(\boldsymbol{\ell} - WB^{-1}\boldsymbol{b})\right]. \tag{4.8}$$

Proof. Notice that when $B = \Phi + \gamma\,\Omega$ and $\boldsymbol{b} = \Phi\hat{\boldsymbol{\theta}}^B$, (4.6) reduces to the benchmarked objective function of (4.5). The matrix B is positive definite by construction and assumption 6 is satisfied. Thus, the KKT matrix K is invertible. The inverse can be derived by block inversion technique, and is given by:

$$K^{-1} = \begin{pmatrix} B & W^T \\ W & 0 \end{pmatrix}^{-1} = \begin{pmatrix} B^{-1} + B^{-1}W^T ZWB^{-1} & -B^{-1}W^T Z \\ ZWB^{-1} & Z \end{pmatrix}$$

where $Z = -(WB^{-1}W^T)^{-1}$. Pre-multiplying both sides of equation (4.7) by K^{-1}, we get:

$$\begin{pmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} B^{-1}\boldsymbol{b} + B^{-1}W^T ZWB^{-1}\boldsymbol{b} - B^{-1}W^T Z\boldsymbol{\ell} \\ ZWB^{-1}\boldsymbol{b} + Z\boldsymbol{\ell} \end{pmatrix}$$

Hence

$$\begin{aligned} \hat{\boldsymbol{\delta}} &= B^{-1}\boldsymbol{b} + B^{-1}W^T ZWB^{-1}\boldsymbol{b} - B^{-1}W^T Z\boldsymbol{\ell} \\ &= B^{-1}\left[\boldsymbol{b} + W^T Z(WB^{-1}\boldsymbol{b} - \boldsymbol{\ell})\right] \\ &= B^{-1}\left[\boldsymbol{b} + W^T(WB^{-1}W^T)^{-1}(\boldsymbol{\ell} - WB^{-1}\boldsymbol{b})\right]. \end{aligned}$$

Because the KKT matrix K is invertible, this estimator is the global minimizer of equation (4.6), and thus unique.

4.3.2 Linearly dependent equality constraints

In the case when equality constraints are linearly dependent W is not full row rank, say $r < min(k, m)$. One can proceed by choosing any of the r independent rows of W to reduce it to the full-rank problem. This can be achieved as follows. If we use a QR decomposition of W^T, by permuting the columns, we may assume that

$$W^T = Q^T \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi \tag{4.9}$$

where $R \in \mathbb{R}^{r\times r}$ is an invertible upper triangular matrix, $S \in \mathbb{R}^{r\times(k-r)}$ and $\Pi \in \mathbb{R}^{k\times k}$ is a permutation matrix. Then define

$$W_1^T = Q^T \begin{pmatrix} R \\ 0 \end{pmatrix}$$

which is a $m \times r$ matrix with full column rank. The matrix W_1 and consequently the solution to the equality constrained problem depends on the choice of the basis matrix Q, however any of the solution suffice as long as the Benchmarked estimator is unique. The uniqueness criterion is particularly important for the almost sure convergence to hold in Theorem 16.

4.4 Posterior projection

Instead of viewing (4.5) as an optimization problem in $\boldsymbol{\delta}$ we can think of it as a restriction on the parameter space Θ itself, i.e. we want to model $\boldsymbol{\theta}$ such that $\boldsymbol{\theta} \in \widetilde{\Theta}$ where

$$\widetilde{\Theta} = \{\boldsymbol{\theta} \in \Theta \mid W\boldsymbol{\theta} = \boldsymbol{\ell}, F\boldsymbol{\theta} \leqslant h\}.$$

When the parameter constraint is simple one can directly incorporate the restriction into an appropriate prior distribution. For example consider the benchmarking constraint $\theta_1 = \ldots = \theta_m$. One can simply use a univariate prior on θ_1 and set the estimate of $\theta_2, \ldots, \theta_m$ to that of θ_1. However, in practical applications $\widetilde{\Theta}$ usually does not have a simple form, more so in the presence of inequality constraints. Trying to incorporate this constraint directly via a restricted prior such as the truncated prior $\pi(\tilde{\boldsymbol{\theta}}) \propto \pi(\boldsymbol{\theta}) 1_{\widetilde{\Theta}}(\theta)$ often leads to an intractable posterior and inefficient posterior sampling. Here $\pi(\boldsymbol{\theta})$ is an unconstrained probability density on Θ and $1_{\widetilde{\Theta}}(\theta)$ is an indicator function of $\widetilde{\Theta}$.

To circumvent this difficulty we break the problem into two parts. First we ignore the constraints in the modeling stage and use an unrestricted prior $\pi(\boldsymbol{\theta})$ which leads to an unrestricted posterior $\pi(\boldsymbol{\theta}|y)$, $\boldsymbol{\theta} \in \Theta$. Next we project the posterior samples $\boldsymbol{\theta}^s$ onto the constrained space $\widetilde{\Theta}$ through a minimal distance mapping. In the context of small area estimation, the chosen loss function such as (4.1) naturally defines a projection operator from Θ to $\widetilde{\Theta}$ (see lemma 15 for details), which induces the push-forward measure $\tilde{\pi}(\tilde{\boldsymbol{\theta}}|y)$ onto the restricted space $\widetilde{\Theta}$. Thus, the posterior projection method reduces to minimizing the same benchmarked objective function as in (4.5) with $\hat{\boldsymbol{\theta}}^B$ replaced by $\boldsymbol{\theta}^s$ for $s = 1, ..., S$ where S is the number of posterior samples. The benchmarked and smoothed objective function for the posterior projection estimator is given by the following:

$$\underset{\boldsymbol{\delta}}{\text{minimize}} \quad (\boldsymbol{\delta} - \boldsymbol{\theta}^s)^T \Phi (\boldsymbol{\delta} - \boldsymbol{\theta}^s) + \gamma \boldsymbol{\delta}^T \Omega \qquad \text{subject to} \quad W\boldsymbol{\delta} = \boldsymbol{\ell}, F\boldsymbol{\delta} \leqslant h \tag{4.10}$$

The minimizer of the optimization problem in (4.10) is the projected sample $\hat{\boldsymbol{\delta}}^s$. The projected samples can be viewed as draws from the induced pushforward measure $\tilde{\pi}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})$ onto the restricted space $\widetilde{\Theta}$ (see section 2.2.2 for details). The posterior projection estimator is given by the average of the projected samples.

$$\hat{\boldsymbol{\delta}}^{PP} = \frac{1}{S} \sum_{s=1}^{S} \hat{\boldsymbol{\delta}}^s \tag{4.11}$$

This estimator converges to the benchmarked estimator (theorem 16), and additionally provides an credible interval. In addition, using the posterior projection approach, we obtain a natural Bayesian uncertainty quantification such as posterior credible intervals, whereas, under the approach of Steorts et al. (2019) only a point estimate is available. The projected measure $\tilde{\pi}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})$ also converges to the true value $\boldsymbol{\theta}^0$ as the number of responses increases.

4.4.1 Theoretical results

Lemma 15. *Under the assumptions 5-6, minimizing the objective function in* (4.5) *is a projection problem.*

Proof. Because B is positive definite it is invertible. Let us denote the inverse matrix by B^{-1}.

$$
\begin{aligned}
&\frac{1}{2}\boldsymbol{\delta}^T B\boldsymbol{\delta} - \boldsymbol{b}^T\boldsymbol{\delta} \\
&= \frac{1}{2}(\boldsymbol{\delta} - B^{-1}\boldsymbol{b})^T B(\boldsymbol{\delta} - B^{-1}\boldsymbol{b}) - \frac{1}{2}\boldsymbol{b}^T B^{-1}\boldsymbol{b} \qquad (4.12)\\
&= \frac{1}{2}\| \boldsymbol{\delta} - B^{-1}\boldsymbol{b} \| - \frac{1}{2}\boldsymbol{b}^T B^{-1}\boldsymbol{b}
\end{aligned}
$$

Therefore minimizing the quadratic objective function (4.6) is equivalent to minimizing $\| \boldsymbol{\delta} - B^{-1}\boldsymbol{b} \|$. Thus $\boldsymbol{\delta}$ is the projection of $B^{-1}\boldsymbol{b}$ via the weighted inner product given in (4.12).

Theorem 16. *Let $W \in \mathbb{R}^{k\times m}$ be of full row rank $k \leqslant m$ and the reduced hessian $Z^T BZ$ be positive definite. Then the posterior projection estimator converges almost surely to the benchmarked estimator.*

Under the assumptions the KKT matrix K is nonsingular. Let us partition the inverse KKT matrix K^{-1} in the following way:

$$
K^{-1} = \begin{pmatrix} K_{11}^{-1} & K_{12}^{-1} \\ K_{21}^{-1} & K_{22}^{-1}, \end{pmatrix}
$$

where $K_{11}^{-1} \in \mathbb{R}^{m\times m}$, $K_{12}^{-1} \in \mathbb{R}^{m\times k}$, $K_{21}^{-1} \in \mathbb{R}^{k\times m}$ and $K_{22}^{-1} \in \mathbb{R}^{k\times k}$. Thus the posterior projection samples are given by $\hat{\boldsymbol{\delta}}^s = K_{11}^{-1}\boldsymbol{\theta}^s + K_{12}^{-1}\boldsymbol{\ell}$. Hence

$$\begin{aligned}
\hat{\boldsymbol{\delta}}^{PP} = \frac{1}{S}\sum_{s=1}^{S}\hat{\boldsymbol{\delta}}^s &= \frac{1}{S}\sum_{s=1}^{S}(K_{11}^{-1}\boldsymbol{\theta}^s + K_{12}^{-1}\boldsymbol{\ell}) \\
&= K_{11}^{-1}\frac{1}{S}\sum_{s=1}^{S}\boldsymbol{\theta}^s + K_{12}^{-1}\boldsymbol{\ell} \\
&\xrightarrow{a.s.} K_{11}^{-1}\hat{\boldsymbol{\theta}}^B + K_{12}^{-1}\boldsymbol{\ell} = \hat{\boldsymbol{\theta}}^{BM}
\end{aligned}$$

4.5 Simulation study

In this section we illustrate the Benchmarked and Posterior Projection estimators via simulation study. Here we assume the setup of the area-level Fay-Herriot model (Fay and Herriot, 1979)

$$\begin{aligned}
\boldsymbol{y}_{m\times 1} &= \boldsymbol{\theta}_{m\times 1} + \boldsymbol{e}_{m\times 1} \\
\boldsymbol{\theta}_{m\times 1} &= X_{m\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{u}_{m\times 1},
\end{aligned} \tag{4.13}$$

where e_i and u_i are mutually independent with

$$\boldsymbol{e}_{m\times 1} \overset{ind}{\sim} \text{MVN}(\mathbf{0}, D) \quad \text{and} \quad \boldsymbol{u}_{m\times 1} \overset{ind}{\sim} \text{MVN}(\mathbf{0}, \sigma_u^2 I_m), \tag{4.14}$$

and

$$D_{m\times m} = \text{Diag}(D_1, \ldots, D_m),$$

$$B_{m\times m} = \text{Diag}(D_1(\sigma_u^2 + D_1)^{-1}, \ldots, D_m(\sigma_u^2 + D_m)^{-1}).$$

Note that $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients and the $\text{rank}(X) = p(< m)$. Equation (4.13) was first considered in the context of estimating income for small areas by Fay and Herriot (1979).

We assume that the covariance matrix of the sampling error D is known and variance of the modeling error σ_u^2 is unknown. For convenience we fix $D_1 = \ldots =$

$D_m = \sigma_e^2 = 1$, however the inferences from the simulation holds under other choices. The value of σ_u^2 was also chosen to be 1. We do not have any prior information on the covariate parameter $\boldsymbol{\beta}$, thus we use an uninformative uniform prior. We will assume the following priors on the unknown parameters:

$$\boldsymbol{\beta} \sim \text{Uniform}(\mathbb{R}^p), \quad \sigma_e^2 \sim \text{InverseGamma}(a/2, b/2)$$

We choose $a = b = 2$ so that the prior mode of the σ_e^2 is close to the true value. We also experimented with other values of a and b, and it does not have any significant impact on the outcome of the simulation experiment. The full conditionals are given by:

$$\begin{aligned}
&\boldsymbol{\beta} \mid \boldsymbol{\theta}, \sigma_u^2, \boldsymbol{y} \sim \text{MVN}\left((X^T X)^{-1} X^T \boldsymbol{\theta},\, \sigma_u^2 (X^T X)^{-1}\right) \\
&\sigma_u^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{y} \sim \text{InverseGamma}\left(\frac{1}{2}(a+p),\, \frac{1}{2}[b + (\boldsymbol{\theta} - X\boldsymbol{\beta})^T(\boldsymbol{\theta} - X\boldsymbol{\beta})]\right) \\
&\boldsymbol{\theta} \mid \boldsymbol{\beta}, \sigma_u^2, \boldsymbol{y} \sim \text{MVN}\left(\left(D^{-1} + \sigma_u^{-2} I_m\right)^{-1}\left(D^{-1}\boldsymbol{y} + \sigma_u^{-2} X\boldsymbol{\beta}\right),\, \left(D^{-1} + \sigma_u^{-2} I_m\right)^{-1}\right)
\end{aligned} \tag{4.15}$$

The simulation study is repeated for $m = 20, 50$ and 100. In an area-level model we have a single observation for each area and thus p/m represents the amount of information present in the raw dataset. So we let the number of covariates scale proportionately with the number of areas and choose $p/m = 10\%, 50\%$ and 90%. For every choice of m and p we simulated 100 different covariates $X_{m \times p}$. The columns of the covaraite matrix were generated from a multivariate normal with mean $\mathbf{0}$, covariance matrix I_m and correlation coefficient ρ. The value of the correlation coefficient ρ is fixed at 0.3 so that the covariates are independent for lower values of p and shows linear dependency for higher values. $\boldsymbol{\theta}_{m\times 1}$ and data $\boldsymbol{y}_{m\times 1}$ were simulated according to (4.13). Aggregating the results over the simulated datasets removes any effect due to the randomness of the covariates in comparing the performance of the Benchmarked and the posterior projection estimators. The full conditionals in (4.15) were used to draw samples from the posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_u^2|\boldsymbol{y})$ according to a Gibbs sampling scheme. For each iteration, we drew 10^5 samples with a burn in value of 10^3 and a thinning parameter equal to 100.

The linear equality constraint on the estimators is chosen to be $\boldsymbol{w}^T\boldsymbol{\delta} = \boldsymbol{w}^T\boldsymbol{y}$, where $\boldsymbol{w} = (1, \ldots, 1)^T$. More specifically, we require the sum of the estimators to agree with the sum of the observed responses. This can also be viewed as a restriction on the parameter space, such that $\boldsymbol{w}^T\boldsymbol{\theta} = \boldsymbol{w}^T\boldsymbol{y}$. We consider two different set of weights ϕ_i for the loss function. When $\phi_i = 1$ we refer to the estimator as constant estimator, and when $\phi_i = V(\theta_i|\boldsymbol{y})^{-1}$ we call it the inverse variance estimator. The posterior samples of $\boldsymbol{\theta}$ are projected according to the objective function (4.6) and (4.10) with only the equality constraints to derive the benchmarked and posterior projection estimators respectively.

Fig. 4.5 illustrates that in the presence of linear equality constraints the MSE between the posterior projection estimator and the benchmarked estimator is almost 0, thus reinforcing the statement of the theorem 16. We also check the effect of the number of areas and the covariates onto the quality of the estimation procedure. For every iteration we calculate the MSE between the projected posterior estimator $\hat{\boldsymbol{\delta}}^{PP}$ and the projected true value of the parameter $\hat{\boldsymbol{\delta}}^0$ as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{\boldsymbol{\delta}}_i^{PP} - \hat{\boldsymbol{\delta}}_i^0)^2 = \frac{1}{m}(\hat{\boldsymbol{\delta}}^{PP} - \hat{\boldsymbol{\delta}}^0)^T\,(\hat{\boldsymbol{\delta}}^{PP} - \hat{\boldsymbol{\delta}}^0) \tag{4.16}$$

where $\hat{\boldsymbol{\delta}}^0$ is the projection of $\boldsymbol{\theta}^0$ in the constrained space $\boldsymbol{w}^T\boldsymbol{\theta} = \boldsymbol{w}^T\boldsymbol{y}$ using the objective function (4.10). It can be seen from Fig. 4.5 that the average MSE decreases as the number of areas increase. Surprisingly, for a constant loss function number of covariates has little to no effect on the MSE.

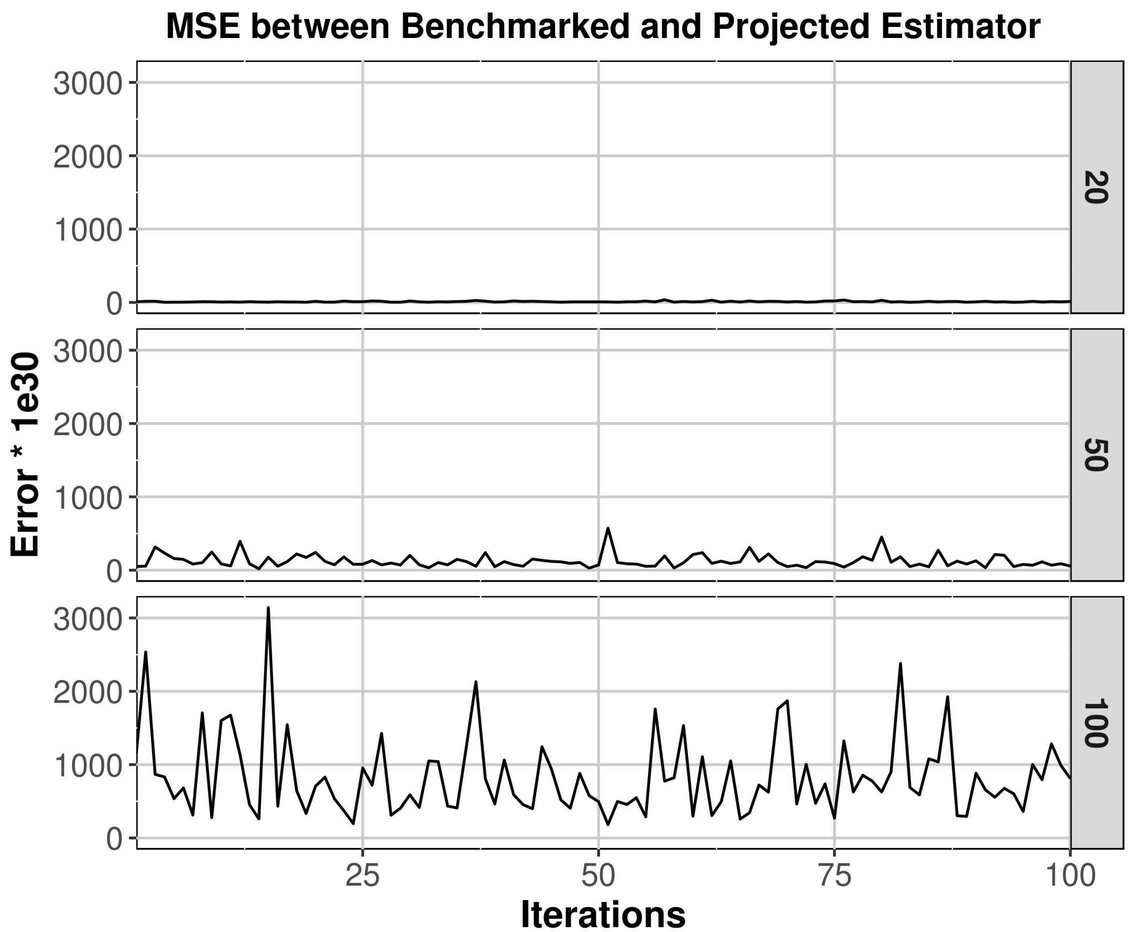


FIGURE 4.1: MSE between the posterior projection estimator and the benchmarked estimator is almost 0, independent of the number of areas and covariates. This illustrates that the posterior projection estimator always converge to the benchmarked estimator in the presence of linear equality constraints.

4.6 Summary & future research

This chapter develops a generalized approach to area-level SAE by projecting posterior samples in the feasible set characterized by benchmarking and smoothing constraints. This extends the applicability of the decision-theoretic framework to include more complex constraints, such as non-linearity and inequality. The proposed method provides estimates and credible intervals for the parameters that has similar performance to that of the state of the art methods. This performance is illustrated

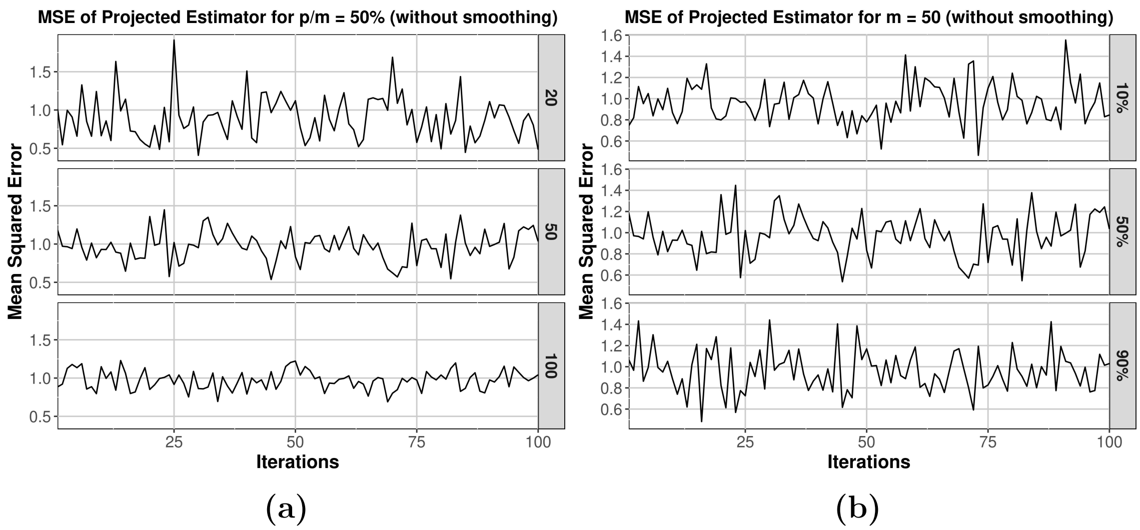


FIGURE 4.2: Effect of number of areas (m) and covariates (p/m) on MSE between the posterior projection estimator and the projected true parameter value. Here we have chosen a loss function with constant weights and no smoothing. It can be seen that across simulations (a) the average MSE decreases as the number of areas increase and (b) there is little to no effect of the number of covariates on the average MSE.

via an extensive simulation study.

Even with the need for more research, the framework and supporting methods developed in this chapter show extensive applicability. While the above discussion and examples focus on area-level model adjustments, unit-level benchmarking is also popular (Ghosh and Steorts, 2013). Our current method can be naturally extended to include this setup. One possible direction is to include generalized variability constraints on parameters such that the model is still tractable. This however makes the problem intractable and requires numerical optimization methods to obtain a solution. Replacing the quadratic objective function with another convex loss function is also a feature of particular interest, but is beyond the scope of this chapter.

5

Conditional Gibbs Sampler

5.1 Introduction

In chapter 2 we developed the theory for the posterior projection mechanism for generalized Banach spaces and extended it to Stiefel manifold, a special non-convex scenario, in chapter 3. The approximate posterior distribution is guaranteed to converge to the true parameter value under suitable assumptions. However the method can only be applied after the unconstrained sampler has converged which increases the time requirement of the entire procedure. Often in statistical experiments it is the case that the parameter constraints are a modeling choice, such as the brain analysis example in section 3.4. In these situations the true parameter value is not guaranteed to be within the constrained region and the method shows sensitivity to the prior hyperparameters. If these values are not chosen carefully, the estimation and prediction power of the model can be a lot lower than that of the competitive models.

Additionally, consider the case when the parameter set is (θ, ϕ), where ϕ is the parameter of interest and the constraint is on the nuisance parameter θ . Projecting the unconstrained joint posterior distribution $\pi(\theta, \phi \mid X)$ after convergence does not change the marginal distribution of the target parameter (see section 5.2.1 for details). Towards this we develop a conditional projection sampler, an algorithm that projects the samples to the constrained space within the MCMC sampler framework. For simplicity we focus on a modified Gibbs algorithm and demonstrate the effectiveness of the method via several examples. We also recognize scenarios when the algorithm fails to converge and develop general theorems to ensure the posterior consistency of such an approximate sampler.

5.2 Notation and formulation

Let the parameter set be $\eta = (\theta, \phi) \in \theta \times \Phi$ where $\widetilde{\theta} \subset \theta \subseteq \mathbb{R}^m$ is the restricted nuisance parameter space and $\Phi \subseteq \mathbb{R}^n$ is the parameter space of interest. We denote the joint unrestriced posterior by $\pi(\theta, \phi \mid \theta)$ where θ is the data generated by the underlying process. $T : \theta \times \Phi \to \widetilde{\theta} \times \Phi$ represent a linear transformation with respect to the underlying norm $\| \cdot \|$. Notice that this also constitute any projection operator as it is a linear transformation. It is easier to build guiding examples for one-one, onto linear transformations and this is the reason we consider the entire class of linear estimators instead of focusing only on projection. Before moving forward we prove a small result which will be instrumental for navigating towards the right direction.

Lemma 17. *Let the original parameter space be $\eta = (\theta, \phi) \in \theta \times \Phi \subset \mathbb{R}^d$. If $T : \Theta \times \Phi \to \widetilde{\Theta} \times \Phi$ denotes the projection operator with respect to the underlying norm $\| \cdot \|$ then*

$$T(\theta, \phi) = (T_\Theta(\theta, \phi), \phi) \tag{5.1}$$

where T_Θ is the restriction of T on Θ.

Proof. The proof is trivial for Euclidean norm. For any $(\theta', \phi') \in \widetilde{\Theta} \times \Phi$

$$\| (\theta, \phi) - (\theta', \phi') \|^2 = (\theta - \theta')^2 + (\phi - \phi')^2$$

So $\phi' = \phi$ minimizes the norm, therefore $T(\theta, \phi) = (T_\Theta(\theta, \phi), \phi)$.

For an arbitrary norm on $\mathbb{R}^d$ we can deduce the following:

$$\| (\theta, \phi) - (\theta', \phi') \| = \| (\theta - \theta', 0) + (0, \phi - \phi') \| \leqslant \| \theta - \theta' \|_\Theta + \| \phi - \phi' \|_\Phi$$

Without any restriction on Φ the $\| \phi - \phi' \|_\Phi$ can be minimized to 0. Hence $T(\theta, \phi) = (T_\Theta(\theta, \phi), \phi)$.

The implication of this lemma is that as long as the parameter constraint is independent of ϕ it is enough to consider linear transformations only on θ.

5.2.1 A motivating example

Let us consider the following example where data $x_i \sim N(\theta, \sigma^2)$, $i = (1, \ldots, n)$ and the mean $\theta \geqslant c \in \mathbb{R}$. Our objective is to estimate $\phi = \sigma^2$. In the traditional Bayesian approach one completes the prior specification as follows.

$$\theta \sim \mathrm{N}_{[c,\infty)}(\theta_0, \sigma_0^2), \quad \sigma^2 \sim \mathrm{IG}(\alpha, \beta).$$

where $\mathrm{N}_{[c,\infty)}(\theta_0, \sigma_0^2)$ denotes a normal density with mean θ_0 and variance σ_0^2 truncated to $[c, \infty)$. One can use Gibbs sampling to alternately sample from the full conditionals $\pi(\theta \mid \phi, X)$ and $\pi(\phi \mid \theta, X)$. This produces a Markov chain whose stationary distribution is the full posterior $\pi(\theta, \phi \mid \theta)$. We have seen the potential problems for this truncated approach in chapter 2. Using posterior projection method we ignore the non-negativity constraint on θ during the modeling stage and take care of it during post-sampling. Under this scheme one proceeds by putting unconstrained priors on θ and ϕ, that is,

$$\theta \sim \mathrm{N}(\theta_0, \sigma_0^2), \quad \sigma^2 \sim \mathrm{IG}(\alpha, \beta).$$

The posteriors are given by

$$\theta \mid \sigma^2, \theta \sim \mathrm{N}\left((1/\sigma_0^2 + n/\sigma^2)^{-1}(\theta_0/\sigma_0^2 + \sum_{i=1}^{n} x_i/\sigma^2), (1/\sigma_0^2 + n/\sigma^2)^{-1}\right)$$

$$\sigma^2 \mid \theta, \theta \sim IG\left(\alpha + n/2, \beta + \sum_{i=1}^{n}(x_i - \theta)^2/2\right)$$

The projection of a posterior sample (θ, σ^2) to the constrained space is the map

$$T(\theta, \sigma^2) = (\max(\theta, c), \sigma^2)$$

As it is evident from the last equation that the constrained space does not affect the marginal distribution of σ^2. As a consequence, the algorithm loses prediction power when the true value of θ is way outside the constrained region. However a conditional projection algorithm where we include the projection step within the sampling steps efficiently eliminates this problem.

5.3 Markov chain & Gibbs sampling

A Markov Chain Monte Carlo (MCMC) algorithm generates sample from a target distribution by constructing a Markov chain that has the target distribution as its stationary distribution. We wish to modify an existing MCMC algorithm such that the corresponding stationary distribution converges to the true value. Gibbs sampling along with Metropolis-Hastings (MH) are two of the most widely used MCMC algorithms. Gibbs sampling samples from the joint posterior by alternatingly sampling from the full conditionals. This particular structure makes it easy to modify Gibbs sampling for our purpose.

Throughout the rest of this chapter we forget about posterior distributions and focus on building a modified Gibbs algorithm, coupled with linear transformations that has a stationary distribution. They key towards this venture is to understand the mechanism of the Gibbs sampler. Let the unconstrained joint density be denoted as $\pi_{\Theta,\Phi}(\theta,\phi)$. The unconstrained marginals and full conditionals are given by

$$\begin{aligned}
\pi_\Theta(\theta) &= \int \pi_{\Theta,\Phi}(\theta,\phi)\, d\phi, \\
\pi_\Phi(\phi) &= \int \pi_{\Theta,\Phi}(\theta,\phi)\, d\theta, \\
\pi_{\Theta|\Phi}(\theta \mid \phi) &= \frac{\pi_{\Theta,\Phi}(\theta,\phi)}{\pi_\Phi(\phi)}, \\
\pi_{\Phi|\Theta}(\phi \mid \theta) &= \frac{\pi_{\Theta,\Phi}(\theta,\phi)}{\pi_\Theta(\theta)}.
\end{aligned}$$

Then the two stage Gibbs sampler proceeds as follows. At t-th iteration,

1. Sample $\theta^{t+1} \sim \pi_{\Theta|\Phi}(\theta \mid \phi^t)$,

2. Sample $\phi^{t+1} \sim \pi_{\Phi|\Theta}(\phi \mid \theta^{t+1})$.

So the transition kernel is

$$K(\theta^{t+1},\phi^{t+1} \mid \theta^t,\phi^t) = \pi_{\Phi|\Theta}(\phi^{t+1} \mid \theta^{t+1})\, \pi_{\Theta|\Phi}(\theta^{t+1} \mid \phi^t).$$

5.3.1 Proof of convergence

The two stage Gibbs sampler does not satisfy detailed balance in general. Therefore the best way to show that $\pi_{\Theta,\Phi}(\theta,\phi)$ is the stationary distribution of this Markov chain is to prove $K\pi_{\Theta,\Phi} = \pi_{\Theta,\Phi}$.

$$
\begin{aligned}
K\pi_{\Theta,\Phi}(\theta', \phi') &= \int\int \pi_{\Theta,\Phi}(\theta, \phi) K(\theta', \phi' \mid \theta, \phi)\, d\theta\, d\phi \\
&= \int\int \pi_{\Theta,\Phi}(\theta, \phi)\pi_{\Phi|\Theta}(\phi' \mid \theta')\pi_{\Theta|\Phi}(\theta' \mid \phi)\, d\theta\, d\phi \\
&= \pi_{\Phi|\Theta}(\phi' \mid \theta') \int \pi_{\Phi}(\phi)\pi_{\Theta|\Phi}(\theta' \mid \phi)\, d\phi \\
&= \pi_{\Phi|\Theta}(\phi' \mid \theta') \int \pi_{\Theta,\Phi}(\theta', \phi)\, d\phi \\
&= \pi_{\Phi|\Theta}(\phi' \mid \theta')\pi_{\Theta}(\theta') \\
&= \pi_{\Theta,\Phi}(\theta', \phi')
\end{aligned}
$$

Hence the most important aspect of designing a Gibbs type sampler is to carefully construct its transition kernel.

5.4 Order dependency

The Gibbs sample is symmetric by construction. By introducing a linear transformation of only θ within the Gibbs sampler induces asymmetry in the system. Therefore it is important to decide which ordering works best for our purpose. In this section we scrutinize the effect of different orderings on the outcome of the algorithm. We only consider linear transformations of the form (5.1).

5.4.1 Sampling scheme 1

In this sampling scheme we project the samples at the end of a complete iteration. The scheme is as follows:

At the t-th iteration,

1. Sample $\theta^{t+1} \sim \pi_{\Phi|\Theta}(\theta \mid \tilde{\phi}^t)$.

2. Sample $\phi^{t+1} \sim \pi_{\Theta|\Phi}(\phi \mid \theta^{t+1})$.

3. Project to the constrained space,
   $(\tilde{\theta}^{t+1}, \tilde{\phi}^{t+1}) = T(\theta^{t+1}, \phi^{t+1}) = (T_\Theta\,(\theta^{t+1}, \phi^{t+1}), \phi^{t+1})$.

Observe that ϕ gets projected back to itself. Therefore at step 1 of $(t+1)$-th iteration the projection does not have any effect, that is $\pi(\theta \mid \tilde{\phi}^{t+1}, \theta) = \pi(\theta \mid \phi^{t+1}, \theta)$. Step 2 is independent of the projection by construction and therefore the projection map does not have any effect on the $(t+1)$th iteration of the sampling scheme. This method is equivalent to the posterior projection method where we project the samples obtained from the unconstrained sampler after the convergence. We have already developed the theory for that method in section 2.3.1.

5.4.2 Sampling scheme 2

If we switch step 1 and 2 of the sampler in the previous section, we get the following sampler:

At the t-th iteration,

1. Sample $\phi^{t+1} \sim \pi_{\Phi|\Theta}(\phi \mid \tilde{\theta}^{t})$.

2. Sample $\theta^{t+1} \sim \pi_{\Theta|\Phi}(\theta \mid \phi^{t+1})$.

3. Project to the constrained space,
   $(\tilde{\theta}^{t+1}, \tilde{\phi}^{t+1}) = T(\theta^{t+1}, \phi^{t+1}) = (T_\Theta\,(\theta^{t+1}, \phi^{t+1}), \phi^{t+1})$.

The projection becomes relevant under this algorithm as $\tilde{\theta}^{t+1} \neq \theta^{t+1}$. This results in a conditional distribution that is influenced by the constraints in step 1 of the $(t+1)$-th iteration of the sampler. However this does not extend easily to multi-stage Gibbs sampler, which is a common choice in Bayesian Hierarchical models. For constraints on more than two parameters one needs to worry about which ordering will ensure that projecting at the end induces a constrained conditional on every parameter of interest in the next iteration of the sampler.

5.4.3 *Sampling scheme 3*

In this sampling scheme we enforce the symmetry by coupling a sampling step with an immediate transformation step.

At the t-th iteration,

1. Sample $\theta^{t+1} \sim \pi_{\Theta|\Phi}(\theta \mid \phi^t)$.

2. Project to the constrained space, $\tilde{\theta}^{t+1} = T_\Theta\left(\theta^{t+1}, \phi^{t+1}\right)$.

3. Sample $\phi^{t+1} \sim \pi_{\Phi|\Theta}(\phi \mid \tilde{\theta}^{t+1})$.

Note that because we assume that the linear transformation does not affect ϕ, we can essentially ignore the associated transformation step. This also generalizes well to multi-stage Gibbs sampler due to the imposed symmetry. Hence we will use this ordering for our modified Gibbs sampler and refer to it as the conditional Gibbs sampler.

5.5 Properties of the conditional Gibbs sampler

It is evident that the sampler 5.4.3 produces a Markov chain. The projection at step 2 of the sampler induces a pushforward measure on the constrained nuisance parameter space $\widetilde{\Theta}$. We argued in chapter 2 that one can construct a reference measure $\tilde{\mu}(\tilde{\theta})$ with respect to which the pushforward measure will have a density. Let us denote this density be $\tilde{\pi}_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi)$. Notice that step 1 and 2 together implies $\tilde{\theta}^{t+1} \sim \tilde{\pi}_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi^t)$. Therefore the transition kernel is

$$K(\tilde{\theta}^{t+1}, \phi^{t+1} \mid \tilde{\theta}^t, \phi^t) = \pi_{\Phi|\Theta}(\phi^{t+1} \mid \tilde{\theta}^{t+1})\, \tilde{\pi}_{\widetilde{\Theta}|\Phi}(\tilde{\theta}^{t+1} \mid \phi^t). \tag{5.2}$$

We need to decide when does this transition kernel have a stationary distribution.

Theorem 18. *The transition kernel* $K : \widetilde{\Theta} \times \Phi \to \widetilde{\Theta} \times \Phi$ *defined in equation* (5.2) *has a stationary distribution if and only if there exists function* $u(\tilde{\theta})$ *and* $v(\phi)$ *such that*

$$\pi_{\Phi|\Theta}(\phi \mid \tilde{\theta})/\tilde{\pi}_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi) = u(\tilde{\theta})\, v(\phi) \tag{5.3}$$

where $\int u(\tilde{\theta})\, d\tilde{\mu}(\tilde{\theta}) < \infty$.

Proof. A Gibbs sampler ensures a joint stationary distribution by alternately sampling from the full conditionals. Therefore the Gibbs like kernel K in equation (5.2) has a stationary distribution if and only if there exists a joint distribution corresponding to the conditionals $\pi_{\Phi|\Theta}(\phi \mid \tilde{\theta})$ and $\tilde{\pi}_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi)$. Theorem 4.1 in Arnold and Press (1989) states this can happen if and only if the condition (5.3) is satisfied. Combining these two facts the conclusion is immediate.

5.6 Effect of 1-1 linear transformation

In this section and the next we look at several examples to figure out a pattern in the stationarity of the conditional Gibbs sampler. For simplicity we assume a bivariate standard normal distribution with correlation coefficient ρ as the unconstrained density for the parameters (θ, ϕ), that is

$$\begin{pmatrix} \theta \\ \phi \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right] \tag{5.4}$$

Then the unconstrained marginals and conditionals are given by

$$\begin{aligned} \pi_{\Theta}(\theta) &= \mathrm{N}(\theta;\, 0, 1), \\ \pi_{\Phi}(\phi) &= \mathrm{N}(\phi;\, 0, 1), \\ \pi_{\Theta|\Phi}(\theta \mid \phi) &= \mathrm{N}(\theta;\, \rho\phi, 1-\rho^2), \\ \pi_{\Phi|\Theta}(\phi \mid \theta) &= \mathrm{N}(\phi;\, \rho\theta, 1-\rho^2). \end{aligned} \tag{5.5}$$

5.6.1 *Example 1*

In our first example we consider the linear transformation $(\tilde{\theta}, \phi) = T(\theta, \phi) = (2\theta, \phi)$. Thus T is 1-1, onto and the transformation of θ is independent of ϕ. Also notice that $\widetilde{\Theta} = \Theta = \mathbb{R}$ in this case. From the normal distribution theory we know that

$$\begin{pmatrix} \tilde{\theta} \\ \phi \end{pmatrix} = \begin{pmatrix} 2\theta \\ \phi \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 2\rho \\ 2\rho & 1 \end{pmatrix}\right] \tag{5.6}$$

and the corresponding marginals and conditionals are given by,

$$\begin{aligned}
&\pi_{\widetilde{\Theta}}(\tilde{\theta}) = \mathrm{N}(\tilde{\theta};\, 0, 4), \\
&\pi_{\Phi}(\phi) = \mathrm{N}(\phi;\, 0, 1), \\
&\pi_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi) = \mathrm{N}(\tilde{\theta};\, 2\rho\phi, 4(1-\rho^2)), \\
&\pi_{\Phi|\widetilde{\Theta}}(\phi \mid \tilde{\theta}) = \mathrm{N}(\phi;\, \rho\tilde{\theta}/2, 1-\rho^2).
\end{aligned} \tag{5.7}$$

The conditional Gibbs sampler in this case proceeds as follows. At the t-th iteration:

1. Sample $\theta^{t+1} \sim \pi_{\Theta|\Phi}(\theta \mid \phi^t) = \mathrm{N}(\theta;\, \rho\phi^t, 1-\rho^2)$
2. Calculate $\tilde{\theta}^{t+1} = T(\theta^{t+1}, \phi^t) = 2\theta^{t+1}$.

   Thus $\tilde{\theta}^{t+1} \sim \mathrm{N}(\tilde{\theta};\, 2\rho\phi^t, 4(1-\rho^2))$, same as $\pi_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi^t)$.
3. Sample $\phi^{t+1} \sim \pi_{\Phi|\Theta}(\phi \mid \tilde{\theta}^{t+1}) = \mathrm{N}(\phi;\, \rho\tilde{\theta}^{t+1}, 1-\rho^2)$.

Notice that if one uses $\phi^{t+1} \sim \pi_{\Phi|\Theta}(\phi \mid \theta^{t+1}) = \mathrm{N}(\phi;\, \rho\theta^{t+1}/2, 1-\rho^2)$ in step 3, we get back the original Gibbs sampler. The stationary distribution under this sampler is given by

$$\begin{pmatrix} \tilde{\theta} \\ \phi \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4\frac{1-\rho^2}{1-2\rho^2} & 2\sqrt{2}\rho\frac{1-\rho^2}{1-2\rho^2} \\ 2\sqrt{2}\rho\frac{1-\rho^2}{1-2\rho^2} & \frac{1-\rho^2}{1-2\rho^2} \end{pmatrix}\right] \tag{5.8}$$

which is different than the joint distribution in (5.6). The stationary distribution is valid only when $2\rho^2 < 1$ or equivalently $|\rho| < 0.7071$. When $2\rho^2 >= 1$ the sample path diverge to infinity. A contour plot of the different sampling procedures is provided in Fig. 5.6.1 when $\rho = 0.5$.

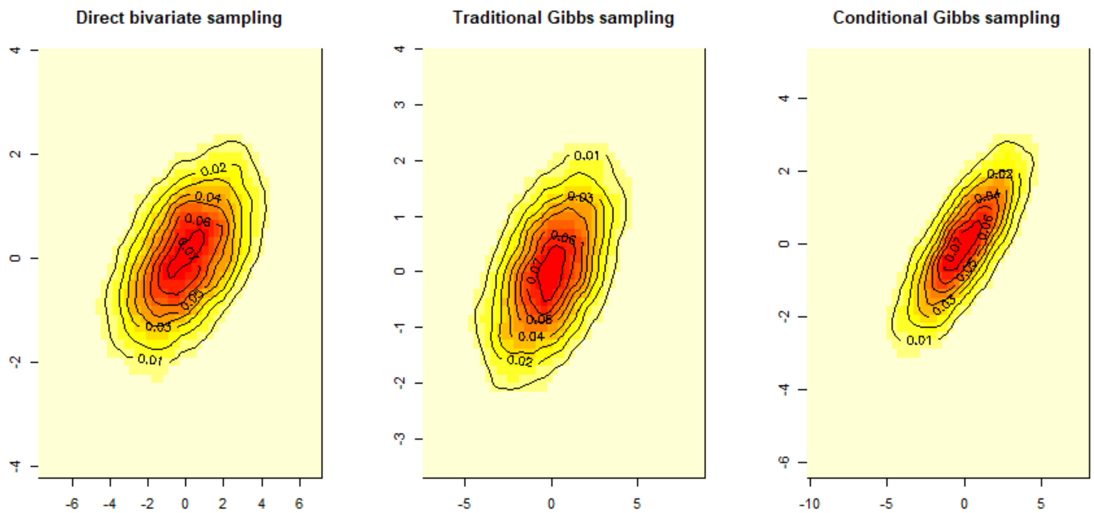


FIGURE 5.1: Illustration of stationary distributions of different sampling schemes under a 1-1, onto, independent transformation, here $\rho = 0.5$

The contours of the direct bivariate sampling and the traditional Gibbs sampling are similar, as expected. The samples of θ and ϕ from the conditional Gibbs sampler are more correlated than other sampling schemes. The marginal distribution of ϕ from conditional Gibbs sampler and traditional Gibbs sampler are $\mathrm{N}(0, \frac{1-\rho^2}{1-2\rho^2})$ and $\mathrm{N}(0, 1)$ respectively. Thus, even though the distributions are different in both cases the samples almost surely converge to their mean 0, which is identical. As any bivariate normals with arbitrary mean and covariance structure can be transformed to (5.4), the conclusion remains true in general.

5.6.2 *Example 2*

Next we consider the linear transformation $(\tilde{\theta}, \phi) = T(\theta, \phi) = (\theta + \phi, \phi)$. Here T is 1-1 and onto but the transformation of θ is dependent on ϕ. Similar as before $\widetilde{\Theta} = \Theta = \mathbb{R}$. From the normal distribution theory we know that

$$\begin{pmatrix} \tilde{\theta} \\ \phi \end{pmatrix} = \begin{pmatrix} \theta + \phi \\ \phi \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2(1+\rho) & (1+\rho) \\ (1+\rho) & 1 \end{pmatrix}\right] \tag{5.9}$$

and the corresponding marginals and conditionals are given by,

$$\pi_{\widetilde{\Theta}}(\tilde{\theta}) = \mathrm{N}(\tilde{\theta};\, 0, 2(1+\rho)),$$

$$\pi_{\Phi}(\phi) = \mathrm{N}(\phi;\, 0, 1),$$

$$\pi_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi) = \mathrm{N}\left(\tilde{\theta};\, (1+\rho)\phi, (1 - \frac{1+\rho}{2})2(1+\rho)\right) = \mathrm{N}(\tilde{\theta};\, (1+\rho)\phi, 1-\rho^2), \tag{5.10}$$

$$\pi_{\Phi|\widetilde{\Theta}}(\phi \mid \tilde{\theta}) = \mathrm{N}\left(\phi;\, \frac{1+\rho}{2(1+\rho)}\tilde{\theta}, (1 - \frac{1+\rho}{2})\right) = \mathrm{N}\left(\phi;\, \frac{\tilde{\theta}}{2}, \frac{1-\rho}{2}\right).$$

The conditional Gibbs sampler in this case proceeds as follows. At the t-th iteration:

1. Sample $\theta^{t+1} \sim \pi_{\Theta|\Phi}(\theta \mid \phi^t) = \mathrm{N}(\theta;\, \rho\phi^t, 1-\rho^2)$

2. Calculate $\tilde{\theta}^{t+1} = T(\theta^{t+1}, \phi^t) = \theta^{t+1} + \phi^t$.

   Thus $\tilde{\theta}^{t+1} \sim \mathrm{N}(\tilde{\theta};\, (1+\rho)\phi^t, 1-\rho^2)$, same as $\pi_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi^t)$.

3. Sample $\phi^{t+1} \sim \pi_{\Phi|\Theta}(\phi \mid \tilde{\theta}^{t+1}) = \mathrm{N}(\phi;\, \rho\tilde{\theta}^{t+1}, 1-\rho^2)$.

The stationary distribution under this sampler is given by the bivariate normal with mean 0, $\mathrm{var}(\theta) = \mathrm{var}(\phi) = \frac{1-\rho^2}{1-\rho(1+\rho)}$ and correlation coefficient $\sqrt{\rho(1+\rho)}$. Again notice that the stationary distribution is different than the joint distribution in (5.6). This stationary distribution is valid when $\rho(1+\rho) < 1$, that is $\rho < 0.618$. Otherwise

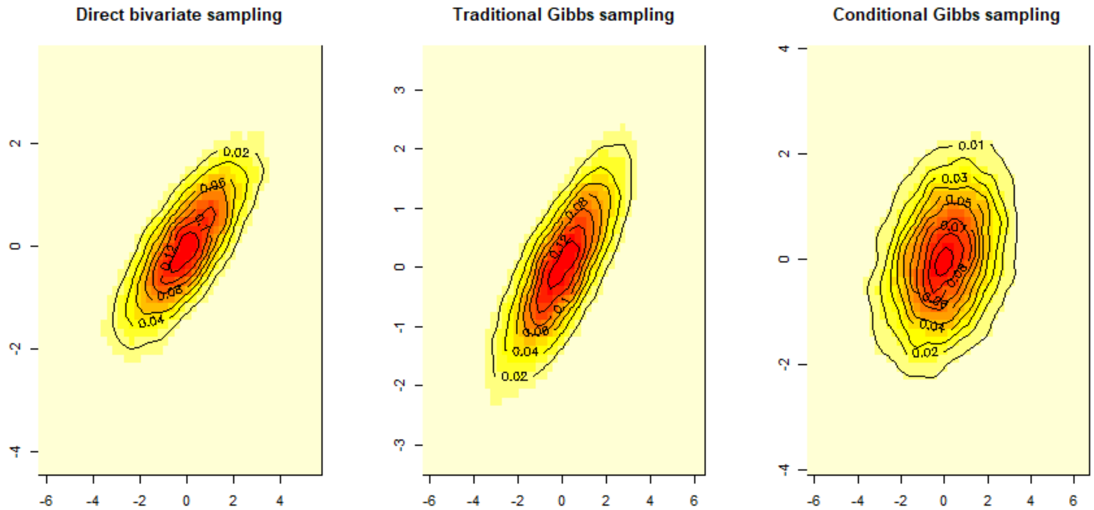


FIGURE 5.2: Illustration of stationary distributions of different sampling schemes under a 1-1, onto, dependent transformation, here $\rho = 0.2$.

the sample paths diverge to infinity. A contour plot of the different sampling procedures is provided in Fig. 5.6.2 when $\rho = 0.2$. In this case, however, there is no way of getting back the original Gibbs sampler.

The contours of the direct bivariate sampling and the traditional Gibbs sampling in Fig 5.6.2 are similar, as expected. In an exact opposite manner to the previous scenario, the samples of θ and ϕ from the conditional Gibbs sampler are less correlated than other sampling schemes. The marginal distribution of ϕ from conditional Gibbs sampler and traditional Gibbs sampler are $\mathrm{N}(0, \frac{1-\rho^2}{1-\rho(1+\rho)})$ and $\mathrm{N}(0, 1)$ respectively. Thus, even though the distributions are different in both cases the samples almost surely converge to their mean 0, which is identical. As any bivariate normals with arbitrary mean and covariance structure can be transformed to (5.4), the conclusion remains true in general.

5.7 Effect of projection

In this section we are checking the effect of projection against more traditional methods. The difference between a projection map and the linear maps considered in the previous section is that depending on the restricted set, a projection can be not 1-1. Additionally, the restricted nuisance parameters space $\widetilde{\Theta}$ is strictly a subset of Θ. Let us consider the following scenarios.

5.7.1 Example 3

We consider the linear transformation $(\tilde{\theta}, \phi) = T(\theta, \phi) = (10, \phi)$. Clearly T is neither 1-1 nor onto, but the transformation of θ is independent of ϕ. The restricted space $\widetilde{\Theta} = \{10\}$ contains the single point 10. By a little abuse of notation we can write the distribution on the projected variables as

$$\begin{pmatrix} \tilde{\theta} \\ \phi \end{pmatrix} = \begin{pmatrix} 10 \\ \phi \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 10 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\right]$$

and the corresponding marginals and conditionals are given by,

$$\begin{aligned} \pi_{\widetilde{\Theta}}(\tilde{\theta}) &= \mathrm{N}(\tilde{\theta};\, 10, 0) = 1_{10}, \\ \pi_{\Phi}(\phi) &= \mathrm{N}(\phi;\, 0, 1), \\ \pi_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi) &= \mathrm{N}(\tilde{\theta};\, 10, 0) = 1_{10}, \\ \pi_{\Phi|\widetilde{\Theta}}(\phi \mid \tilde{\theta}) &= \mathrm{N}(\phi;\, 0, 1). \end{aligned} \tag{5.11}$$

The conditional Gibbs sampler in this case proceeds as follows. At the t-th iteration:

1. Sample $\theta^{t+1} \sim \pi_{\Theta|\Phi}(\theta \mid \phi^t) = \mathrm{N}(\theta;\, \rho\phi^t, 1 - \rho^2)$

2. Calculate $\tilde{\theta}^{t+1} = T(\theta^{t+1}, \phi^t) = 10$.

   Thus $\tilde{\theta}^{t+1} \sim \mathrm{N}(\tilde{\theta};\, 10, 0)$, same as $\pi_{\widetilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi^t)$.

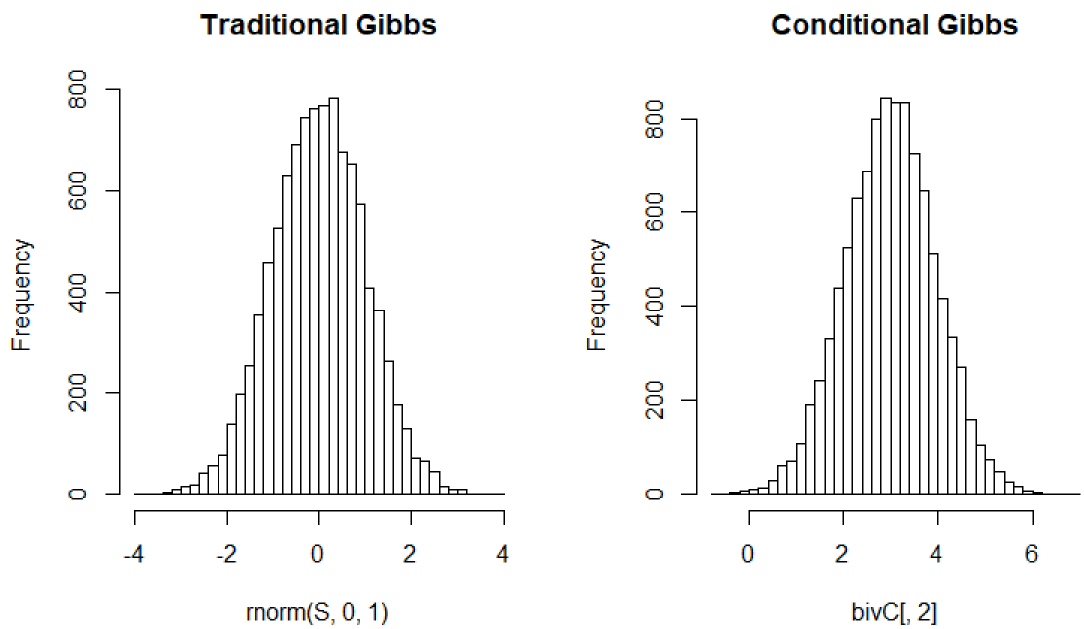


FIGURE 5.3: Illustration of stationary distributions of a traditional Gibbs sampler vs conditional Gibbs sampler when $\rho = 0.3$ when projecting to a single boundary point 10.

3. Sample $\phi^{t+1} \sim \pi_{\Phi|\Theta}(\phi \mid \tilde{\theta}^{t+1}) = \mathrm{N}(\phi;\, 10\rho, 1-\rho^2)$.

In this very particular point projection scenario the conditional distributions does not change between iterations. Thus the transition kernel (5.2) is fixed and the stationary distribution for the process is given by $1_{10}(\tilde{\theta}) \times \mathrm{N}(\phi;\, 10\rho, 1-\rho^2)$. While in traditional Gibbs the marginal of ϕ is $\mathrm{N}(\phi; 0, 1)$, in conditional Gibbs the marginal is $\mathrm{N}(\phi;\, 10\rho, 1-\rho^2)$, which is directly dependent on the projection boundary. Also the credible interval for ϕ becomes narrower as the value of correlation coefficient increases (see Fig. 5.7.1). This in someway confirms the fact that the accuracy of the conditional Gibbs as an estimation procedure depends on how much mass the unconstrained density has on the constraint set. Additionally the conditional Gibbs can preserve the correlation structure between variables which otherwise would have been lost in traditional MCMC methods.

5.7.2 *Example 4*

We now consider the nearest distance mapping $(\tilde{\theta}, \phi)$ of (θ, ϕ) such that $\tilde{\theta} \geqslant c$ where $c \in \mathbb{R}$ is a constant. This can be represented as a linear transformation $(\tilde{\theta}, \phi) = T(\theta, \phi) = (\max(\theta, c), \phi)$. It is hard to derive the density of this transformation under the same distributional assumption of (5.4). The conditional Gibbs sampler in this case proceeds as follows. At the t-th iteration:

1. Sample $\theta^{t+1} \sim \pi_{\Theta|\Phi}(\theta \mid \phi^t) = \mathrm{N}(\theta;\, \rho\phi^t, 1-\rho^2)$
2. Calculate $\tilde{\theta}^{t+1} = T(\theta^{t+1}, \phi^t) = \max(\theta^{t+1}, c)$.
   Thus $\tilde{\pi}_{\tilde{\Theta}|\Phi}(\tilde{\theta} \mid \phi^t) = 1(\tilde{\theta} = c)\int_{-\infty}^{c} \mathrm{N}(x;\, \rho\phi^t, 1-\rho^2)\, dx + 1(\tilde{\theta} > c)\mathrm{N}(\tilde{\theta};\, \rho\phi^t, 1-\rho^2)$.
3. Sample $\phi^{t+1} \sim \pi_{\Phi|\Theta}(\phi \mid \tilde{\theta}^{t+1}) = \mathrm{N}(\phi;\, \rho\tilde{\theta}^{t+1}, 1-\rho^2)$.

We compare the stationary distributions of the different sampling procedures by varying the cutoff point c in Fig. **??**. The correlation coefficient is fixed at $\rho = 0.5$ with several cutoff values. The difference is stark for $c = 0$. The conditional Gibbs sampler puts more mass on the boundary of θ (x-axis) than that of the traditional Gibbs sampler. However the distribution of ϕ (y-axis) is similar. However the methods converge to different points as the conditional Gibbs sampler puts more mass on the cross-section of ϕ corresponding to the boundary of θ, which is c in this case. As the value of c approaches $-\infty$, outcomes from all the sampling procedures coincide.

5.8 An Ar(1) process

As an opposite spectrum of the scenario in section 5.6.1 we consider the linear transformation $(\tilde{\theta}, \phi)$ such that $\tilde{\theta} = \phi$ where $\theta \sim \mathrm{N}(0, 1)$. This is equivalent to a projection from $\mathbb{R}^2$ to the one dimensional line $\theta = \phi$, that is $(\tilde{\theta}, \phi) = T(\theta, \phi) = (\phi, \phi)$. One simple way would be to sample θ and set $\phi = \theta$ directly. But what if we sample

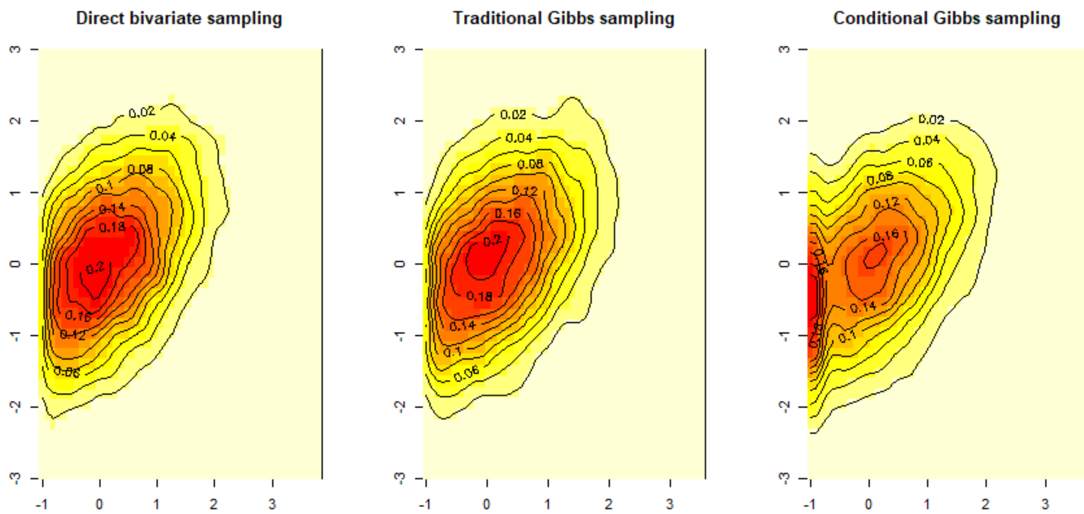


FIGURE 5.4: Illustration of effect of boundary points on the stationary distributions under different sampling scheme. Here $\rho = 0.5$ and $c = -1$.

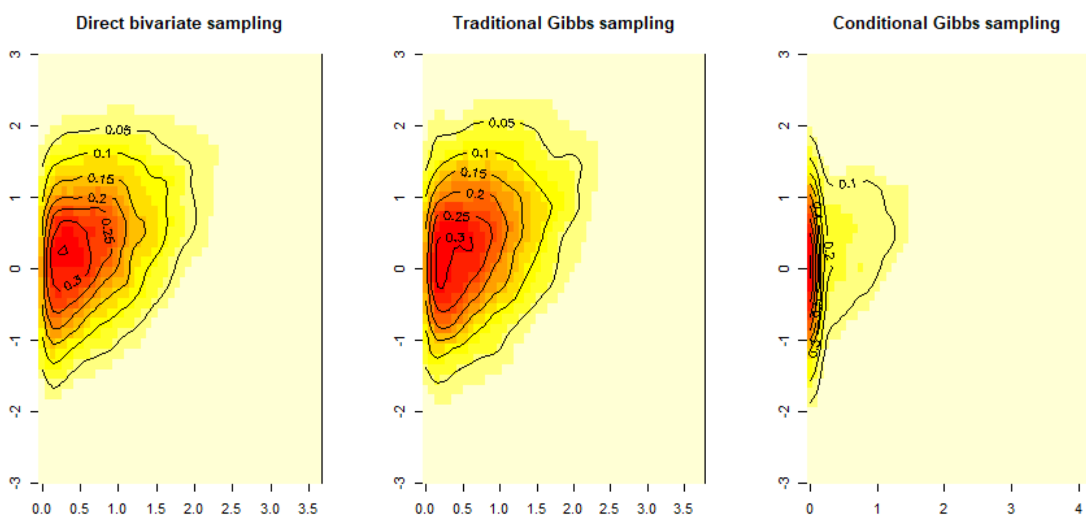


FIGURE 5.5: Illustration of effect of boundary points on the stationary distributions under different sampling scheme. Here $\rho = 0.5$ and $c = 0$.

(θ, ϕ) from a bivariate normal as in equation (5.4) and then project the samples onto the $\theta = \phi$ line.

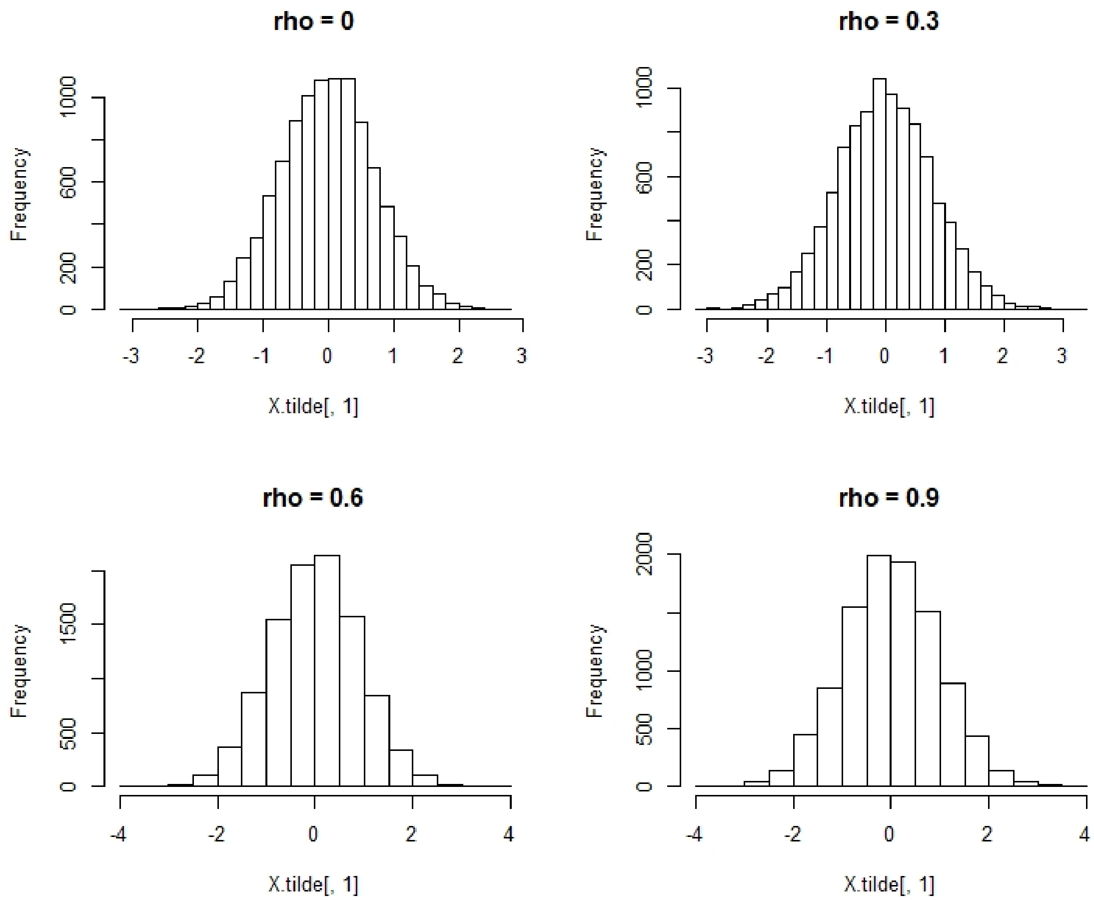


FIGURE 5.6: Comparison of the stationary distributions of conditional Gibbs and posterior projection sampler. The case $\rho = 0$ corresponds to the conditional Gibbs sampler. The variance of the stationary distribution of the posterior projection sampler increases with ρ.

The conditional Gibbs sampler in this case proceeds as follows. At the t-th iteration:

1. Sample $\theta^{t+1} \sim \pi_{\Theta|\Phi}(\theta \mid \phi^t) = \mathrm{N}(\theta;\, \rho\phi^t, 1-\rho^2)$

2. Calculate $\tilde{\theta}^{t+1} = T(\theta^{t+1}, \phi^t) = \phi^t$. Thus $\tilde{\theta}^{t+1} \sim \mathrm{N}(\phi;\, \rho\phi^{t-1}, 1-\rho^2)$.

3. Sample $\phi^{t+1} \sim \pi_{\Phi|\Theta}(\phi \mid \tilde{\theta}^{t+1}) = \mathrm{N}(\phi;\ \rho\phi^t, 1-\rho^2)$.

The step 1 and 2 of the algorithm combined implies that the sampling of θ can essentially be ignored. A careful scrutiny reveals that the output of the algorithm can also be expressed as

$$\phi^t = \rho\phi^{t-1} + \epsilon^t, \quad \epsilon^t \sim N(0, 1-\rho^2) \tag{5.12}$$

which is an AR(1) process. From the theory of Autoregressive models we know this is a stationary process. The stationary distribution of (5.12) is given by $\mathrm{N}(\phi; 0, \frac{1-\rho^2}{1-\rho^2}) = \mathrm{N}(\phi; 0, 1)$. Hence the stationary distribution of the conditional Gibbs sampler is independent of the correlation coefficient and same as that of the traditional Gibbs sampler.

It is also interesting to observe the difference between the conditional Gibbs sampler in section 5.4.3 and posterior projection sampler in section 5.4.1. Outcome of the t-th iteration of the conditional Gibbs looks like (ϕ^t, ϕ^{t+1}) whereas it is (ϕ^{t+1}, ϕ^{t+1}) for the posterior projection. Due to this the conditional Gibbs sampler results in a stationary distribution which does not depend on the correlation coefficient ρ. However for the posterior projection sampler the stationary distribution of the samples is more spread out with increasing value of ρ (Fig **??**). Here the case $\rho = 0$ corresponds to the outcome of the conditional Gibbs sampler. In both cases the posterior samples converge almost surely to the true mean.

5.9 Summary

In this chapter, we have described the conditional Gibbs sampler and when it might be useful. We explored some specific examples and derived conditions when it works. There are many interesting directions. One might use a rejection like step within the sampler to ensure consistency. We can also choose to ignore the nuisance parameter and focus on the Markov chain of the parameter of interest. This is however very hard to do in practice, as integrating over the nuisance parameter θ is analytically intractable. One might also try to focus on the case where we have a large number of data, so that the posterior converges to an arbitrary small neighbourhood of the true value of the parameters. In that scenario an approximate algorithm like conditional Gibbs sampler might produce the same result.

6

Conclusion

We have proposed a new methodology to perform Bayesian analysis under the constrained parameter space. Often a truncated prior is the go to choice in such scenarios, which leads to intractable posteriors. In other cases sampling from the posterior can be challenging. Our method overcomes these problems by initially putting a prior on the parameter space that ignores the constraint. The resultant posterior is usually analytically tractable and easy to sample from. These samples are then projected back to the constrained space via a minimal distance mapping. We discuss the choice of such metric projection operators and provide a subjective Bayesian justification for the projection method. We first focus on the closed and convex constraint scenario, and show that under suitable assumptions the proposed methodology achieves equivalent consistency rate as that of the unrestricted posterior. We also extend the underlying theoretical justification to a special class of non-convex constraints, Stiefel manifold. The advantage of the projection method against truncated priors is then illustrated via a simulation example. We proceed to apply our method on two real data sets, a drug-response study and a brain subnetwork identification problem. These experiments have drastically different models and constraint structure. We show that in both cases, the projection method provides accurate predictions and meaningful interpretation.

There are many interesting new directions to our proposed methodology. One such important generalization is considered, where we allow for conditional posterior projections; for example, applying projection to a subset of parameters immediately after each update step within a MCMC. For simplicity, we focus on a modified Gibbs sampler. However, instead of only considering projections, we work with the broader class of all linear transformations. Based on a standard bivariate normal distribution with correlation coefficient ρ, we identify several scenarios where such a modified algorithm converges to the underlying true distribution, and develop a general theory to ensure the consistency. Another direction is studying when projected posteriors are expected to produce better results than traditional approaches; for example, obtaining lower risk parameter estimates. Also of considerable interest is the construction of formal goodness-of-fit test of the constraint, as well as adaptive procedures that avoid projecting when the cost to goodness-of-fit is too high.

We applied the proposed methodolgy in the context of SAE to derive posterior projection point estimates that converge to the traditional benchmarked estimate. In addition, the method naturally provides a probabilistic characterization of the uncertainty in the estimation procedure via credible intervals. The advantage of using the proposed methodology is then illustrated via theoretical justification and simulation studies. While building this decision-theoretic framework, we did not impose any distributional assumption on our model. This flexibility opens numerous avenues for continued research. One can venture beyond the Fay-Herriot setup to more complex semiparameteric and nonparametric models, and check the convergence and mean square error in estimation via our proposed methodology. Another interesting direction for future research is to study the behaviour of the projected posterior estimates under more complex constraints. This consists of quadratic constraints, such as benchmarking the variability of the estimates, as well as non-convex constraints such as, Stiefel manifolds. Although we considered only area level models in our work, the underlying theoretical framework naturally extends to unit level models. With unit level data, one has more information per area, and thus variance in the estimate decreases. The asymptotic behaviour the posterior projection method in unit level models, in terms of convergence rate and consistency, is also a promising area of future research.

Appendix A

Proofs, additional tables and figures for Posterior Projection in closed and convex space

A.1 Existence of prior on constrained parameter space

Proof of Lemma 1. Let $B \in \mathcal{B}_{\widetilde{\Theta}}$ such that $\widetilde{\mu}_{\widetilde{\Theta}}(B) = 0$. Then

$$\widetilde{\mu}_{\widetilde{\Theta}}(B) = 0 \implies \mu_{\Theta}(T^{-1}B) = 0 \implies \lambda_{\Theta}(T^{-1}B) = 0 \implies \widetilde{\lambda}_{\widetilde{\Theta}}(B) = 0.$$

This implies $\widetilde{\lambda}_{\widetilde{\Theta}} \ll \widetilde{\mu}_{\widetilde{\Theta}}$. $\square$

Proof of Theorem 2. We want a prior density $\pi_{\widetilde{\Theta}}(\tilde{\theta})$ on $\widetilde{\Theta}$ such that the corresponding posterior is equal to the projected posterior density $\widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)})$, that is we want the following to hold.

$$\widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)}) = p_{\tilde{\theta}}(X^{(n)})\ \pi_{\widetilde{\Theta}}(\tilde{\theta}) \left(\int_{\widetilde{\Theta}} p_{\tilde{\theta}}(X^{(n)})\ \pi_{\widetilde{\Theta}}(\tilde{\theta})\ d\mu_{\widetilde{\Theta}}(\tilde{\theta})\right)^{-1} \tag{A.1}$$

$$\implies \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)}) \left(\int_{\widetilde{\Theta}} p_{\tilde{\theta}}(X^{(n)})\ \pi_{\widetilde{\Theta}}(\tilde{\theta})\ d\mu_{\widetilde{\Theta}}(\tilde{\theta})\right) = p_{\tilde{\theta}}(X^{(n)})\ \pi_{\widetilde{\Theta}}(\tilde{\theta}).$$

We choose $\pi_{\widetilde{\Theta}}(\tilde{\theta}) = p_{\tilde{\theta}}(X^{(n)})^{-1}\ \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)}) \left(\int_{\widetilde{\Theta}} p_{\tilde{\theta}}(X^{(n)})^{-1}\ \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)})\ d\mu_{\widetilde{\Theta}}(\tilde{\theta})\right)^{-1}$ which is well defined because of Assumption 1. This also defines a density on $\widetilde{\Theta}$ as

$\int_{\widetilde{\Theta}} \pi_{\widetilde{\Theta}}(\tilde{\theta})\ d\mu_{\widetilde{\Theta}}(\tilde{\theta}) = 1$. Substituting our chosen $\pi_{\widetilde{\Theta}}(\tilde{\theta})$ in the left hand side of (A.1) we get

$$
\begin{aligned}
&\widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)}) \left[\int_{\widetilde{\Theta}} \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)}) \left\{ \int_{\widetilde{\Theta}} p_{\tilde{\theta}}(X^{(n)})^{-1}\ \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)})\ d\mu_{\widetilde{\Theta}}(\tilde{\theta}) \right\}^{-1} d\mu_{\widetilde{\Theta}}(\tilde{\theta}) \right] \\
&= \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)}) \ \left(\int_K \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)}) d\mu_{\widetilde{\Theta}}(\tilde{\theta}) \right) \left\{ \int_K p_{\tilde{\theta}}(X^{(n)})^{-1}\, \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)})\, d\mu_{\widetilde{\Theta}}(\tilde{\theta}) \right\}^{-1} \\
&= \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)}) \left\{ \int_{\widetilde{\Theta}} p_{\tilde{\theta}}(X^{(n)})^{-1}\ \widetilde{\pi}_{\widetilde{\Theta}}(\tilde{\theta}|X^{(n)})\ d\mu_{\widetilde{\Theta}}(\tilde{\theta}) \right\}^{-1} \\
&= p_{\tilde{\theta}}(X^{(n)}) \pi_{\widetilde{\Theta}}(\tilde{\theta}).
\end{aligned}
$$

□

A.2 Posterior convergence of projection method

We study the rate at which the posterior concentrates on arbitrary small neighbourhoods of $P_{\theta_0}^{(n)}$, where $\theta_0 \in \Theta$ denotes the true value of the parameter. Let the unrestricted posterior $\Pi_{\Theta}^{(n)}(\cdot \mid X^{(n)})$ have rate ϵ_n with respect to a semimetric d_Θ on Θ in the following sense:

$$\Pi_{\Theta}^{(n)} \left(\theta : d_\Theta(\theta, \theta_0) > M_n \epsilon_n \mid X^{(n)} \right) \to 0, \quad \text{for every } M_n \to \infty, \quad P_{\theta_0}^{(n)} (a.s.). \tag{A.2}$$

Proof of Theorem 3. From Assumptions 2–4 we can deduce the following inequality. For every $\theta \in \Theta$ and $\widetilde{\theta} \in \widetilde{\Theta}$,

$$C^{-2} d_\Theta(T\theta, \tilde{\theta}) \leqslant C^{-1} \| \, T\theta - \tilde{\theta} \, \| \leqslant C^{-1} \| \, \theta - \tilde{\theta} \, \| \leqslant d_\Theta(\theta, \tilde{\theta}). \tag{A.3}$$

Hence, for every $M_n \to \infty$, $\epsilon_n > 0$ and $\theta_0 \in \widetilde{\Theta}$, we have

$$T^{-1} \left\{ \tilde{\theta} : d_\Theta(\tilde{\theta}, \theta_0) \geqslant C^2 M_n \epsilon_n \right\} = \left\{ \theta : d_\Theta(T\theta, \theta_0) \geqslant C^2 M_n \epsilon_n \right\} \subseteq \{ \theta : d_\Theta(\theta, \theta_0) \geqslant M_n \epsilon_n \}$$

Therefore,

$$\begin{aligned}
&\widetilde{\Pi}^{(n)}_{\widetilde{\Theta}} \left\{\tilde{\theta} : d_\Theta(\tilde{\theta}, \theta_0) > C^2 M_n \epsilon_n \mid X^{(n)}\right\} \\
&= \Pi^{(n)}_{\Theta} \left[T^{-1} \left\{\tilde{\theta} : d_\Theta(\tilde{\theta}, \theta_0) > C^2 M_n \epsilon_n\right\} \mid X^{(n)}\right] \\
&= \Pi^{(n)}_{\Theta} \left\{\theta : d_\Theta(T\theta, \theta_0) > C^2 M_n \epsilon_n \mid X^{(n)}\right\} \\
&\leqslant \Pi^{(n)}_{\Theta} \left\{\theta : d_\Theta(\theta, \theta_0) > M_n \epsilon_n \mid X^{(n)}\right\} \to 0
\end{aligned}$$

Because C is a fixed constant the projected posterior converges to θ_0 and has concentration rate at least that of the original posterior. □

Proof of Corollary 4. All these spaces are separable, uniformly convex and uniformly smooth Banach spaces. □

Metric projection onto a closed convex subset of Hilbert space is non expansive, that is

$$\| T\theta - T\theta' \| \leqslant \| \theta - \theta' \|, \quad \text{for any } \theta, \theta' \in \Theta. \tag{A.4}$$

Therefore in Hilbert spaces we do not need to assume that $\theta \in \widetilde{\Theta}$ and we achieve the following theorem.

Proof of Theorem 5. A similar argument as in the proof of Theorem A.2 gives us the following relation. This is similar to equation (A.3) in nature, but appropriated to Hilbert space using the non-expansive property (A.4). For any $\theta, \theta' \in \Theta$,

$$C^{-2} d_\Theta(T\theta, T\theta') \leqslant C^{-1} \| T\theta - T\theta' \| \leqslant C^{-1} \| \theta - \theta' \| \leqslant d_\Theta(\theta, \theta').$$

Hence, for every $M_n \to \infty$, $\epsilon_n > 0$ and $\theta, \theta_0 \in \Theta$, we have

$$T^{-1}\left\{\theta : d_\Theta(\theta, T\theta_0) \geqslant C^2 M_n \epsilon_n\right\} = \left\{\theta : d_\Theta(T\theta, T\theta_0) \geqslant C^2 M_n \epsilon_n\right\} \subseteq \{\theta : d_\Theta(\theta, \theta_0) \geqslant M_n \epsilon_n\}$$

Therefore,

$$
\begin{aligned}
&\widetilde{\Pi}^{(n)}_{\widetilde{\Theta}}\left\{\theta : d_{\Theta}(\theta,\, T\theta_0) > C^2 M_n \epsilon_n \mid X^{(n)}\right\} \\
&= \Pi^{(n)}_{\Theta}\left[T^{-1}\left\{\theta : d_{\Theta}(\theta,\, T\theta_0) > C^2 M_n \epsilon_n\right\} \mid X^{(n)}\right] \\
&= \Pi^{(n)}_{\Theta}\left\{\theta : d_{\Theta}(T\theta,\, T\theta_0) > C^2 M_n \epsilon_n \mid X^{(n)}\right\} \\
&\leqslant \Pi^{(n)}_{\Theta}\left\{\theta : d_{\Theta}(\theta, \theta_0) > M_n \epsilon_n \mid X^{(n)}\right\} \to 0
\end{aligned}
$$

Because C is a fixed constant the projected posterior converges to $T\theta_0$ and has concentration rate at least that of the original posterior. □

Proof of Corollary 6. All these spaces are separable Hilbert spaces. □

A.3 Generalization of projection of conjugate Gaussian likelihood-prior

As a generalization of the illustrative example we consider projection of a univariate non-standard Gaussian likelihood with conjugate prior onto the closed interval $[c, d]$ on the real line where c and d can be $-\infty$ and ∞, respectively. Let $X^{(n)} = (x_1, x_2, ..., x_n)$ be such that $x_i \sim N(\theta, \sigma^2),\ i = (1, \dots, n)$. When the constrained parameter space is a closed interval on the real line it is common to put a truncated prior on the parameter, that is $\theta \sim N_{(c,d)}(\theta_0, \sigma_0^2)$. Here $N_{(c,d)}$ denotes the normal distribution truncated between c and d. The posterior of θ is then given by

$$
\theta | X^{(n)}, \sigma^2 \sim N_{(c,d)}(\theta_n, \sigma_n^2), \quad \theta_n = \sigma_n^2 \left(\theta_0/\sigma_0^2 + n\bar{x}/\sigma^2\right), \quad \sigma_n^2 = \left(1/\sigma_0^2 + n/\sigma^2\right)^{-1}. \tag{A.5}
$$

This is equivalent to sampling from the unconstrained posterior and discarding values outside c and d. This posterior density assigns 0 probability to the boundaries and has expectation

$$
E\left(\theta \mid X^{(n)}, \sigma^2\right) = \theta_n + \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} \sigma_n \tag{A.6}
$$

where $\alpha = (c - \theta_n)/\sigma_n$, $\beta = (d - \theta_n)/\sigma_n$ and ϕ and Φ respectively denotes the density and distribution function of standard normal.

In the posterior projection approach we initially ignore the constraint on θ, and use the unrestricted prior $\theta \sim N(\theta_0, \sigma_0^2)$. The resulting posterior is $\theta \mid X^{(n)} \sim N(\theta_n, \sigma_n^2)$ which has the same form as equation (A.5) but without the truncation. Let us denote this probability measure by $\Pi_{\mathbb{R}}^{(n)}$. Next we obtain samples θ_s from the unconstrained posterior and set $\tilde{\theta}_s$ to either c, θ_s or d according to whether θ_s is less than c, between c and d, or greater than d, respectively. Thus for any $B \in \mathcal{B}_{[c,d]}$ the projected posterior measure is given by

$$\widetilde{\Pi}_{[c,d]}^{(n)}(B) = \Pi_{\mathbb{R}}^{(n)}(-\infty, c)\, 1_B(c) + \Pi_{\mathbb{R}}^{(n)}((c,d) \cap B) + \Pi_{\mathbb{R}}^{(n)}(d, \infty)\, 1_B(d) \tag{A.7}$$

Here $1_B(c)$ is 1 if $c \in B$ and 0 otherwise. It is easy to see that this posterior (A.7) does not have a density with respect to $\mu_{\mathbb{R}}$, the Lebesgue measure in $\mathbb{R}$, as it has point mass at c and d. One might think that because the metric projection preserves absolute continuity (Lemma 1), $\widetilde{\Pi}_{[c,d]}\left(\cdot \mid X^{(n)}, \sigma^2\right)$ will have a density with respect to the projected Lebesgue measure in $[c,d]$, which is given by

$$\begin{aligned}\widetilde{\mu}_{[c,d]}(B) &= \mu_{\mathbb{R}}(-\infty, c)\, 1_B(c) + \mu_{\mathbb{R}}((c,d) \cap B) + \mu_{\mathbb{R}}(d, \infty)\, 1_B(d) \\ &= \Phi(\alpha)\, 1_B(c) + \mu_{\mathbb{R}}((c,d) \cap B) + \Phi(-\beta)\, 1_B(d).\end{aligned}$$

But $\widetilde{\mu}_{[c,d]}$ is not σ-finite as $\widetilde{\mu}_{[c,d]}(c) = \infty$, therefore the condition for Radon-Nikodym derivative is not satisfied. However, one can construct a σ-finite reference measure λ by putting point mass $t_c, t_d \in \mathbb{R}$ at c and d.

$$\lambda_{[c,d]}(B) = t_c 1_B(c) + \mu_{\mathbb{R}}((c,d) \cap B) + t_d 1_B(d) \tag{A.8}$$

It can be checked that the projected prior measure is a version of the reference measure when the measure in the interior of the set (c,d) is chosen as the prior measure $\Pi_{\mathbb{R}}$ instead of $\mu_{\mathbb{R}}$. If we fix $t_c = t_d = 1$, then the projected posterior on the interval $[c,d]$ has the following density with respect to (A.8).

$$\widetilde{\pi}^{(n)}_{[c,d]}(\tilde{\theta}) = \Phi(\alpha)\, 1_c(\tilde{\theta}) + [\Phi(\beta) - \Phi(\alpha)]\, N_{(c,d)}(\tilde{\theta}; \theta_n, \sigma_n^2) + \Phi(-\beta)\, 1_d(\tilde{\theta}) \tag{A.9}$$

This posterior density (A.9) has expectation

$$E\left(\tilde{\theta} \mid X^{(n)}, \sigma^2\right) = c\ \Phi(\alpha) + [\Phi(\beta) - \Phi(\alpha)] \left(\theta_n + \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\sigma_n\right) + d\ \Phi(-\beta)\,.$$

It is also possible to determine the prior on the interval $[c,d]$ which leads to the projected posterior density in (A.9). Any such prior density must have point masses at the boundary. Let the density be

$$\pi_{[c,d]}(\tilde{\theta}) = w_1\, 1_c(\tilde{\theta}) + w_2\, N_{(c,d)}(\tilde{\theta}; \theta_0, \sigma_0^2) + w_3\, 1_d(\tilde{\theta})$$

A simple calculation shows that the posterior under the Gaussian likelihood with this prior leads to the following posterior density.

$$\pi^{(n)}_{[c,d]}(\tilde{\theta}) = W_1\, 1_c(\tilde{\theta}) + W_2\, N_{(c,d)}(\tilde{\theta}; \theta_n, \sigma_n^2) + W_3\, 1_d(\tilde{\theta}),$$

where $W_j = \frac{w_j C_j}{\sum_{j=1}^3 w_j C_j}$ for $j \in \{1,2,3\}$ and C_j are derived as follows:

$$C_1 = \int_c^d \frac{\sqrt{n}}{\sigma}\phi\left(\frac{\sqrt{n}(\bar{x} - \tilde{\theta})}{\sigma}\right) 1_c(\tilde{\theta})\ d\tilde{\theta} = N(c; \bar{x}, \sigma^2/n).$$

$$C_2 = \int_c^d \frac{\sqrt{n}}{\sigma}\phi\left(\frac{\sqrt{n}(\bar{x} - \tilde{\theta})}{\sigma}\right) 1_d(\tilde{\theta})\ d\tilde{\theta} = N(d; \bar{x}, \sigma^2/n).$$

$$C_3 = \int_c^d \frac{\sqrt{n}}{\sigma}\phi\left(\frac{\sqrt{n}(\bar{x} - \tilde{\theta})}{\sigma}\right) N_{(c,d)}(\tilde{\theta}; \theta_n, \sigma_n^2)\ d\tilde{\theta} = N(\bar{x}; 0, \sigma^2/n)\frac{N_{(c,d)}(0; \theta_0, \sigma_0^2)}{N_{(c,d)}(0; \theta_n, \sigma_n^2)}.$$

If we denote the weights of the projected posterior density in (A.9) as W_j^0, then our goal is to find w_j such that $W_j = W_j^0$. If one selects $w_j = (W_j^0/C_j)/(\sum_{j=1}^3 W_j^0/C_j)$, then this is satisfied.

A.4 Stochastic ordering on contingency table

We model the data from Agresti and Coull (1998) via a Dirichlet-Multinomial model. The unconstrained posterior is given by equation (2.11) with $\alpha = 1$. We then apply the posterior projection approach to account for the stochastic ordering. The estimated cell probabilities and confidence intervals of the resultant method is provided in table A.1.

Table A.1: Estimates and credible intervals of the cell probability under the projection approach with $\alpha = 1$.

Treatment group	Outcome				
	Death	Vegetative state	Major disability	Minor disability	Good recovery
Placebo	0·285 (0·236 – 0·342)	0·124 (0·088 – 0·166)	0·221 (0·171 – 0·275)	0·227 (0·178 – 0·278)	0·143 (0·105 – 0·177)
Low dose	0·251 (0·206 – 0·300)	0·113 (0·074 – 0·157)	0·232 (0·179 – 0·288)	0·249 (0·200 – 0·304)	0·155 (0·124 – 0·189)
Medium dose	0·218 (0·178 – 0·258)	0·065 (0·035 – 0·102)	0·256 (0·202 – 0· 312)	0·301 (0·245 – 0·359)	0·161 (0·130 – 0·199)
High dose	0·207 (0·165 – 0·248)	0·027 (0·009 – 0·051)	0·252 (0·197 – 0·310)	0·299 (0·242 – 0·362)	0·214 (0·165 – 0·290)

Appendix B

Proofs, additional tables and figures for Posterior Projection in Stiefel manifold

B.1 Metric projection onto Stiefel manifold

Proof of Lemma 7. The proof of this proposition is immediate from the statement of the proposition 4.2 of Bardelli and Mennucci (2017) as $\mathrm{St}(p, m)$ is a closed set of the Riemanninan manifold $\mathbb{R}^{m\times p}$. A detailed proof can be found in references therein. $\square$

Proof of Proposition 9. We note a result from linear algebra that states that an element of $\mathbb{R}^{m\times p}$ is in Θ_p if and only if it has a $p \times p$ invertible submatrix. Therefore we define the function $f(\theta) = \sum_B |\det(B)|$ for $\theta \in \mathbb{R}^{m\times p}$ where the sum is over all $p \times p$ submatrices. Clearly f is continuous, so $\Theta_p = f^{-1}(\mathbb{R}^{m\times p}\backslash\{0\}$ is an open set. $\square$

Proof of Lemma 12. For any $\theta \in \mathbb{R}^{m\times p}$, $\tilde{\theta} \in \mathrm{St}(p, m)$ we have

$$\begin{aligned} \| T\theta - \tilde{\theta} \| &\leqslant \| T\theta - \theta \| + \| \theta - \tilde{\theta} \|, \quad \text{triangle inequality} \\ &\leqslant 2 \| \theta - \tilde{\theta} \| \quad \text{from definition of } T\theta. \end{aligned}$$

$\square$

Proof of Theorem 13. We present an argument similar to that of Theorem A.2. However we focus on Θ_p, the set of all rank p matrices instead of $\mathbb{R}^{m\times p}$, as we showed in Lemma 7 that Θ_p has probability 1. This is similar to equation (A.3) in nature, but characterized for Stiefel manifolds. For any $\theta \in \Theta_p$, $\tilde{\theta} \in \mathrm{St}(p, m)$,

$$\frac{1}{2}C^{-2}d_\Theta(T\theta, \tilde{\theta}) \leqslant \frac{1}{2}C^{-1}\| T\theta - \tilde{\theta} \| \leqslant C^{-1}\| \theta - \tilde{\theta} \| \leqslant d_\Theta(\theta, \tilde{\theta}).$$

Hence, for every $M_n \to \infty$, $\epsilon_n > 0$ and $\theta_0 \in \widetilde{\Theta} = \mathrm{St}(p, m)$, we have

$$T^{-1}\left\{\tilde{\theta} : d_\Theta(\tilde{\theta}, \theta_0) \geqslant 2\, C^2 M_n \epsilon_n\right\} = \left\{\theta : d_\Theta(T\theta, \theta_0) \geqslant 2\, C^2 M_n \epsilon_n\right\} \subseteq \{\theta : d_\Theta(\theta, \theta_0) \geqslant M_n \epsilon_n\}$$

Therefore,

$$\begin{aligned}
&\widetilde{\Pi}^{(n)}_{\widetilde{\Theta}}\left\{\tilde{\theta} : d_\Theta(\tilde{\theta}, \theta_0) > 2\, C^2 M_n \epsilon_n \mid X^{(n)}\right\} \\
&= \Pi^{(n)}_{\Theta}\left[T^{-1}\left\{\tilde{\theta} : d_\Theta(\tilde{\theta}, \theta_0) > 2\, C^2 M_n \epsilon_n\right\} \mid X^{(n)}\right] \\
&= \Pi^{(n)}_{\Theta}\left\{\theta \in \Theta_p : d_\Theta(T\theta, \theta_0) > 2\, C^2 M_n \epsilon_n \mid X^{(n)}\right\} \\
&\leqslant \Pi^{(n)}_{\Theta}\left\{\theta \in \Theta_p : d_\Theta(\theta, \theta_0) > M_n \epsilon_n \mid X^{(n)}\right\} \to 0
\end{aligned}$$

Because C is a fixed constant the projected posterior converges to θ_0 and has concentration rate at least that of the original posterior. $\square$

Bibliography

Absil, P.-A. and Malick, J. (2012), "Projection-like retractions on matrix manifolds," *SIAM Journal on Optimization*, 22, 135–158.

Aggarwal, G., Chowdhury, A. R., and Chellappa, R. (2004), "A system identification approach for video-based face recognition," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4, pp. 175–178, IEEE.

Agresti, A. and Coull, B. A. (1998), "Order-restricted inference for monotone trend alternatives in contingency tables," *Computational Statistics & Data Analysis*, 28, 139–155.

Agresti, A. and Coull, B. A. (2002), "The analysis of contingency tables under inequality constraints," *Journal of Statistical Planning and Inference*, 107, 45–73.

Arnold, B. C. and Press, S. J. (1989), "Compatible Conditional Distributions," *Journal of the American Statistical Association*, 84, 152–156.

Bardelli, E. and Mennucci, A. C. G. (2017), "Probability measures on infinite-dimensional Stiefel manifolds," *Journal of Geometric Mechanics*, 9, 291–316.

Barlow, R. E. (1972), *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*, Chichester : Wiley.

Bell, W., Datta, G., and Ghosh, M. (2013), "Benchmarked Small Area Estimators," *Biometrika*, 100, 189–202.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015), "Dirichlet-Laplace Priors for Optimal Shrinkage," *Journal of the American Statistical Association*, 110, 1479–1490.

Cordes, D., Haughton, V., Carew, J. D., Arfanakis, K., and Maravilla, K. (2002), "Hierarchical clustering to measure connectivity in fMRI resting-state data," *Magn Reson Imaging*, 20, 305–17.

Coullon, G. S., Emir, U. E., Fine, I., Watkins, K. E., and Bridge, H. (2015), "Neurochemical changes in the pericalcarine cortex in congenital blindness attributable to bilateral anophthalmia," *J Neurophysiol*, 114, 1725–33.

Craddock, R. C., Jbabdi, S., Yan, C.-G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., Kelly, C., Heberlein, K., Colcombe, S., and Milham, M. P. (2013), "Imaging human connectomes at the macroscale," *Nature Methods*, 10, 524.

Datta, G. S., Ghosh, M., Steorts, R., and Maples, J. (2011), "Bayesian benchmarking with applications to small area estimation," *TEST*, 20, 574–588.

Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006), "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *Neuroimage*, 31, 968–80.

Dunson, D. B. and Neelon, B. (2003), "Bayesian Inference on Order-Constrained Parameters in Generalized Linear Models," *Biometrics*, 59, 286–295.

Edelman, A., Arias, T. A., and Smith, S. T. (1998), "The geometry of algorithms with orthogonality constraints," *SIAM journal on Matrix Analysis and Applications*, 20, 303–353.

El Barmi, H. and Dykstra, R. (1998), "Maximum Likelihood Estimates via Duality for Log-Convex Models When Cell Probabilities are Subject to Convex Constraints," *The Annals of Statistics*, 26, 1878–1893.

Evans, M., Gilula, Z., Guttman, I., and Swartz, T. (1997), "Bayesian Analysis of Stochastically Ordered Distributions of Categorical Variables," *Journal of the American Statistical Association*, 92, 208–214.

Fay, R. and Herriot, R. (1979), "Estimates of income from small places: an application of James-Stein procedures to census data," *Journal of the American Stastical Association*, 74, 269–277.

Fitzpatrick, S. and Phelps, R. R. (1982), "Differentiability of the metric projection in Hilbert space," *Transactions of the American Mathematical Society*, 270, 483–501.

Fox, M. D. and Raichle, M. E. (2007), "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging," *Nat Rev Neurosci*, 8, 700–11.

Fu, W. J. (1998), "Penalized regressions: the bridge versus the lasso," *Journal of computational and graphical statistics*, 7, 397–416.

Ghosal, S. and Van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference*, vol. 44, Cambridge University Press.

Ghosh, M. (1992), "Constrained Bayes estimation with applications," *Journal of the American Statistical Association*, 87, 533–540.

Ghosh, M. and Steorts, R. C. (2013), "Two-Stage Bayesian Benchmarking as Applied to Small Area Estimation," *TEST*, 22, 670–687.

Goodall, C. R. and Mardia, K. V. (1999), "Projective shape analysis," *Journal of Computational and Graphical Statistics*, 8, 143–168.

Gunn, L. H. and Dunson, D. B. (2005), "A transformation approach for incorporating monotone or unimodal constraints," *Biostatistics*, 6, 434–449.

Hoff, P. (2008), "Modeling homophily and stochastic equivalence in symmetric relational data," in *Advances in Neural Information Processing Systems 20*, eds. J. Platt, D. Koller, Y. Singer, and S. Roweis, pp. 657–664, MIT Press, Cambridge, MA.

Hoff, P. D. (2016), "Equivariant and Scale-Free Tucker Decomposition Models," *Bayesian Anal.*, 11, 627–648.

Horowitz, J. L. and Lee, S. (2017), "Nonparametric estimation and inference under shape restrictions," *Journal of Econometrics*, 201, 108–126.

Isaki, C., Tsay, J., and Fuller, W. (2004), "Weighting sample data subject to independent controls," *Survey Methodology*, 20, 35–44.

Jidling, C., Wahlstrm, N., Wills, A., and Schn, T. B. (2017), "Linearly constrained Gaussian processes," in *Advances in Neural Information Processing Systems*, pp. 1215–1224.

Journe, M., Nesterov, Y., Richtrik, P., and Sepulchre, R. (2010), "Generalized power method for sparse principal component analysis," *Journal of Machine Learning Research*, 11, 517–553.

Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A. D., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., Smith, S. A., Joel, S., Mori, S., Pekar, J. J., Barker, P. B., Prince, J. L., and van Zijl, P. C. M. (2011), "Multi-Parametric Neuroimaging Reproducibility: A 3T Resource Study," *NeuroImage*, 54, 2854–2866.

Laudy, O. and Hoijtink, H. (2007), "Bayesian methods for the analysis of inequality constrained contingency tables," *Stat Methods Med Res*, 16, 123–38.

Li, J. (2004), "The metric projection and its applications to solving variational inequalities in Banach spaces," *Fixed Point Theory*, 5, 285–298.

Lin, L. and Dunson, D. B. (2014), "Bayesian monotone regression using Gaussian process projection," *Biometrika*, 101, 303–317, 10.1093/biomet/ast063.

Louis, T. (1984), "Estimating a population of parameter values using Bayes and empirical Bayes methods," *Journal of the American Stastical Association*, 79.

Nagy, K., Greenlee, M. W., and Kovcs, G. (2012), "The Lateral Occipital Cortex in the Face Perception Network: An Effective Connectivity Study," *Frontiers in Psychology*, 3, 141.

Neelon, B. and Dunson, D. B. (2004), "Bayesian Isotonic Regression and Trend Analysis," *Biometrics*, 60, 398–406.

Nicolas, B., Bamdev, M., Absil, P.-A., and Rodolphe, S. (2014), "Manopt, a Matlab Toolbox for Optimization on Manifolds," *Journal of Machine Learning Research*, 15, 1455–1459.

Pfeffermann, D. (2013), "New important developments in small area estimation," *Statistical Science*, 28, 40–68.

Pfeffermann, D. and Tiller, R. (2006), "Small area estimation with state-space models subject to benchmark constraints," *Journal of the American Stastical Association*, 101, 1387–1397.

Prasad, N. and Rao, J. (1990), "The estimation of the mean squared error of small-area estimators," *Journal of the American Stastical Association*, 85, 163–171.

Rao, J. and Molina, I. (2015), *Small Area Estimation*, John Wiley & Sons, New York.

Robert, C. and Casella, G. (2013), *Monte Carlo statistical methods*, Springer Science & Business Media.

Roncal, W. G., Koterba, Z. H., Mhembere, D., Kleissas, D. M., Vogelstein, J. T., Burns, R., Bowles, A. R., Donavos, D. K., Ryman, S., Jung, R. E., Wu, L., Calhoun, V., and Vogelstein, R. J. (2013), "MIGRAINE: MRI Graph Reliability Analysis and Inference for Connectomics," in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 313–316.

Saarela, O. and Arjas, E. (2011), "A Method for Bayesian Monotonic Multiple Regression," *Scandinavian Journal of Statistics*, 38, 499–513.

Shi, Y. and Toga, A. W. (2017), "Connectome imaging for mapping human brain pathways," *Molecular Psychiatry*, 22, 1230.

Shively, T. S., Walker, S. G., and Damien, P. (2011), "Nonparametric function estimation subject to monotonicity, convexity and other shape constraints," *Journal of Econometrics*, 161, 166–181.

Sporns, O., Tononi, G., and Ktter, R. (2005), "The Human Connectome: A Structural Description of the Human Brain," *PLoS Computational Biology*, 1, e42.

Steorts, R. (2013), "Divergence losses under benchmarking for small area estimation," *Working Paper*.

Steorts, R. and Louis, T. (2014), "Constrained Bayesian Benchmarking," *In Preparation*.

Steorts, R., Schmid, T., and Tzavidis, N. (2019), "Smoothing and Benchmarking for Small Area Estimation," .

Steorts, R. C. (2014), "Smoothing, Clustering, and Benchmarking for Small Area Estimation," *arXiv e-prints*, p. arXiv:1410.7056.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., and Rojas-Perilla, N. (2018), "From start to finish: a framework for the production of small area official statistics," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 927–979.

Veeraraghavan, A., Roy-Chowdhury, A. K., and Chellappa, R. (2005), "Matching shape sequences in video with applications in human movement analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1896–1909.

Wang, J., He, L., Zheng, H., and Lu, Z.-L. (2014), "Optimizing the Magnetization-Prepared Rapid Gradient-Echo (MP-RAGE) Sequence," *PLoS ONE*, 9, e96899.

Wehbe, L., Ramdas, A., Steorts, R., and Shalizi, C. (2014), "Regularized Brain Reading with Shrinkage and Smoothing ," *Submitted*.

Williams, M. R. and Savitsky, T. D. (2019), "Optimization of Survey Weights under a Large Number of Conflicting Constraints," *arXiv e-prints*, p. arXiv:1901.03791.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

Biography

Sayan Patra received his undergraduate degree in 2012 from Indian Statistical Institute (ISI), Kolkata, with a major in Statistics and minor in Econometrics. He also completed his masters degree at ISI, with a concentration in Mathematics, Statistics and Probability, graduating in first class with distinction, in July 2014.

He joined Duke University to pursue a PhD in Statistical Science, graduating in the summer of 2019, and earning a Masters of Science in Statistical Science in 2016 en route. He received best student paper award for his paper "Constrained Bayesian Inference through Posterior Projection" from The Section on Bayesian Statistical Science (SBSS) of the American Statistical Association (ASA) in 2019. Next, he will be moving onto a position at the Walt Disney Parks and Resorts U.S., Inc. in Orlando, Florida, to work as a decision science consultant.