Constraint Relaxation for Bayesian Modeling with Parameter Constraints

Leo Duan, Alexander L Young, Akihiko Nishimura, David Dunson

Abstract: Prior information often takes the form of parameter constraints. Bayesian methods include such information through prior distributions having constrained support. By using posterior sampling algorithms, one can quantify uncertainty without relying on asymptotic approximations. However, outside of narrow settings, parameter constraints make it difficult to develop new prior and/or efficient posterior sampling algorithms. In this work, we first describe a general approach to utilize the large pool of unconstrained distributions in constrained space, then we propose to relax the parameter support into the neighborhood surrounding constrained space for convenient posterior estimation. The constraint relaxation can be done using data augmentation technique or with an approximation function. General off the shelf posterior sampling algorithms, such as Hamiltonian Monte Carlo (HMC), can then be used directly. We illustrate this approach through multiple examples involving equality and inequality constraints. While existing methods tend to rely on conjugate families or sophisticated reparameterization, our proposed approach frees us up to define new classes of models for constrained problems. We illustrate this through application to a variety of simulated and real datasets.

KEY WORDS: Simplex, Stiefel Manifold, Parameter Expansion

1 Introduction

It is extremely common to have prior information available on parameter contraints in statistical models. For example, one may have prior knowledge that a vector of parameters lies on the probability simplex or satisfies a particular set of inequality constraints. Other common examples include shape constraints on functions, positive semidefiniteness of matrices and orthogonality. There is a very rich literature on optimization subject to parameter contraints. One common approach is to rely on Lagrange and Karush-Kuhn-Tucker multipliers (Boyd and Vandenberghe, 2004). However, simply producing a point estimate is often insufficient, as uncertainty quantification (UQ) is a key component of most statistical analyses. Usual large sample asymptotic theory, for example showing asymptotic normality of statistical estimators, tends to break down in constrained inference problems. Instead, limiting distributions may have a complex form that needs to be rederived for each new type of constraint, and may be intractable. An appealing alternative is to rely on

Bayesian methods for UQ, including the constraint through a prior distribution having restricted support, and then applying Markov chain Monte Carlo (MCMC) to avoid the need for large sample approximations. Although MCMC is conceptually simple, except for a few limited cases, it is generally difficult to generate random variable strictly inside constrained space.

To overcome this difficulty, one common strategy is to reparameterize with un-/less constrained parameters at equal or less dimension. The new parameters form functions that can always satisfy the constraint. The transformation, if bijective, is known as 'coordinate system' in manifold embedding literature (Nash, 1954; Do Carmo, 2016). Examples include the polar coordinates for data on a hyper-sphere, or stick-breaking construction for Dirichlet distribution on probability simplex (Ishwaran and James, 2001). One can then directly assign prior on the less constrained parameters. Although this strategy has been successful, convenient coordinate system does not always exist; and heavy reparameterzation tends to makes it more difficult to induce prior property on the original space. For example, uniformity of unconstrained parameter in a compact space may not be equivalent to uniformity on the constrained space via transformation. Diaconis et al. (2013) provide a useful tutorial and cautious guide on this subject.

Alternatively, it is typical to rely on customized solution for specific constraints. One popular strategy is to restrict focus to a prior and likelihood such that posterior sampling is tractable. For example, for modeling of data on Stiefel manifolds, von Mises-Fisher and matrix Bingham-von Mises-Fisher distribution (Khatri and Mardia, 1977; Hoff, 2009) are routinely used. Besides limiting consideration to specialized models, another drawback is that the tractable computation, especially posterior conjugacy, tends to break down under common modeling/data complication, such as matrix symmetry, hierarchical structures, etc.

For these reasons, it is appealing to consider approaches that do not rely on conjugate constrained distributions. Early work (Gelfand et al., 1992) suggested using general unconstrained distribution inside a simple truncated space, and running Gibbs sampling ignoring the constraint but only accepting the draws that fall into truncated space. Unfortunately, this method can be highly inefficient if constrained space has a small or zero measure, which will create a low or zero acceptance probability. A recent idea is to run MCMC ignoring the constraint, and then project draws from the unconstrained posterior to the appropriately constrained space. Such an approach was proposed for generalized linear models with order constraints by Dunson and Neelon (2003), extended to functional data with monotone or unimodal constraints (Gunn and Dunson, 2005), and recently modified to nonparametric regression with monotonicity (Lin and Dunson, 2014) or manifold (Lin et al., 2016) constraints. A third independent direction utilizes Hamiltonian Monte Carlo (HMC) that incorporates geometric structure with a Riemannian metric (Girolami and Calderhead, 2011), making proposals strictly inside the constrained space by solving a large linear system. Although simpler algorithms using geodesic flow were proposed for a few selected constrained space (Byrne and Girolami,

2013), compared to the first two strategies that operates in unconstrained space, strictly accommodating the constrained geometry tend to require more customization, such as computing the metric tensor for different manifolds.

The goal of this article is to dramatically expand the families of constrained priors one could use and develop simple computational strategy for more general constraints. We first introduce a general strategy to adapt common existing distributions into constrained space. To enable simple posterior computation, we *relax* the parameter into the neighborhood surrounding the constrained space. This approach enjoys the advantages of unconstrained space sampling while approximately takes into account of the geometry of the constrained space. This relaxation either produces an approximation for posterior under general constraints formed by equality and/or inequality, or an exact solution for several common constrained space such as simplex and Stiefel manifolds. Theoretic studies are conducted and comparison with existing approaches are shown in simulations and data applications.

2 Constraint Relaxation Methodology

In conventional models under constraints, assume that θ is an $\mathcal{R}$-valued random variable with dimensionality $\dim(\mathcal{R}) = r < \infty$ and that θ is subject to some constraints restricting it to a subset $\mathcal{D} \subset \mathcal{R}$. In Bayesian setting, θ is a parameter where constraints arise from some given information.

As a motivating example, consider the case where θ has prior density $\pi_{\mathcal{R}}(\theta)$ with support $\mathcal{R}$ and $\mathcal{D}$ is a measureable subset with positive measure. The posterior density of θ given data Y and $\theta \in \mathcal{D}$ is,

$$\pi(\theta \mid \theta \in \mathcal{D}, Y) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)$$

with likelihood function $\mathcal{L}(\theta; Y)$.

There are two primary problems: posterior inference often requires custom solutions that limit scope of modeling; there often lacks complete confidence in constraints and it is appealing to allow small deviations.

To address these issues, we consider 'relaxing' the constraints by placing a high probability on $\mathcal{D}$ but has support in $\mathcal{R}$. Suppose we used the density

$$\tilde{\pi}(\theta) \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp\left(-\frac{1}{\lambda} v_{\mathcal{D}}(\theta) \right) \tag{1}$$

where $v_{\mathcal{D}}(\theta) = \inf_{x \in \mathcal{D}} \|\theta - x\|$ is a measure of the distance from θ to the constrained space for some metric $\|\cdot\|$.

Note that $\mathbb{1}_{\mathcal{D}}(\theta)$ is the pointwise limit of $\exp\big(-\nu_{\mathcal{D}}(\theta)/\lambda\big)$ (except perhaps on the boundary of $\mathcal{D}$) as $\lambda \to 0^+$. However, (1) has support $\mathcal{R}$ for all $\lambda > 0$, hence 'relaxing' the constraint. We refer to this strategy

as constraint relaxation (**CORE**).

We investigate a number of questions about **CORE** in the article: (i) For what types of distributions and constraints is CORE suitable? (ii) Is there a general approach for constructing the 'relaxed' constraint? (iii) How well do samples from the relaxed constraint represent those from the fully conditioned distribution? (iv) How does the amount of relaxation depend on the tuning parameter λ?

The answers to (ii) - (iv) depend largely upon (i). Therefore, beginning with (i), we assume θ is a continuous random variable (e.g. $\mathcal{R}$ is $[0,\infty)^d$, $\mathbb{R}^{n\times k}$) and θ has an unconstrained distribution, $\pi_{\mathcal{R}}(\theta)$, which is absolutely continuous with respect to Lebesque measure on $\mathcal{R}$ hereby denoted as $\mu_{\mathcal{R}}$. We investigate two general types of constraints.

2.1 Constrained Space with Positive Measure

We first consider the simpler case where $\mathcal{D}$ has positive measure, i.e. $\int_{\mathcal{D}} \mathcal{L}(\theta; y)\pi(\theta)d\mu_{\mathcal{R}}(\theta) > 0$. Generally, inequality constraints (e.g. $a^T\theta < 0$, $\|\theta\|_2^2 < 1$) fall into this category.

The constrained posterior density is

$$\pi_{\mathcal{D}}(\theta \mid Y) = \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\mathcal{D}}(\theta).$$

which is defined with respect to $\mu_{\mathcal{R}}$.

We now consider a relaxed density

$$\tilde{\pi}_\lambda(\theta) = \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\frac{v_{\mathcal{D}}(\theta)}{\lambda}\big)}{\int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\frac{v_{\mathcal{D}}(\theta)}{\lambda}\big)d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\exp\big(-\frac{v_{\mathcal{D}}(\theta)}{\lambda}\big) \tag{2}$$

which is also absolutely continuous with respect to $\mu_{\mathcal{R}}$. Here $\lambda > 0$ and $v_{\mathcal{D}}(\theta)$ is a scalar-valued function which measures the distance from θ to the constrained space $\mathcal{D}$, i.e. $v_{\mathcal{D}}(\theta) = 0 \quad \forall\theta \in \mathcal{D}$ and is positive otherwise. Formally, as $\lambda \to 0^+$, $\exp(-v_{\mathcal{D}}(\theta)/\lambda) \to \mathbb{1}_{\mathcal{D}}(\theta)$ pointwise. If $\mathcal{D}$ is an open subset of $\mathcal{R}$, this limit may not hold on the boundary of $\mathcal{D}$, denoted $\partial\mathcal{D}$; fortunately in general, $\mu_{\mathcal{R}}(\partial\mathcal{D}) = 0$.

There are many possible choices for $v_{\mathcal{D}}$ which may be selected for different reasons. Perhaps the simplest choice is to take

$$v_{\mathcal{D}}(\theta) = \inf_{x\in\mathcal{D}} \|\theta - x\|_k \tag{3}$$

where $\|\cdot\|_k$ denotes the distance using the k-norm. Under this choice of $v_{\mathcal{D}}$, the relaxation is isotropic. More generally, one could use

$$v_{\mathcal{D}}(\theta) = \inf_{x \in \mathcal{D}} \sqrt{(x-\theta)^T A (x-\theta)} \tag{4}$$

for some positive definite matrix A. In this case, the relaxation is anistrophic, and can be viewed as a form of directional relaxation. This choice of distance, $v_{\mathcal{D}}$, allows for a more detailed specification of the rates at which individual components of θ relax to $\mathcal{D}$. In fact, if one seeks to constrain θ within a neighborhood containing $\mathcal{D}$, the matrix A can be chosen to preferentially relax the constraint and limit deviations from $\mathcal{D}$ in certain directions.

As long as $v_{\mathcal{D}}(\theta)$ is zero for $\theta \in \mathcal{D}$ and positive for $\theta \notin \mathcal{D}$, it follows that $\pi_{\mathcal{D}}$ is the pointwise limit of $\tilde{\pi}_\lambda$ for $\mu_{\mathcal{R}}$ a.e. θ in $\mathcal{R}$. Naturally, one could approximately estimate $E[g(\theta) \mid \theta \in \mathcal{D}]$ via relaxed posterior density, $\tilde{\pi}_\lambda$. For small λ, we anticipate

$$\int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta) d\mu_{\mathcal{R}}(\theta) \approx \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}(\theta) d\mu_{\mathcal{R}}(\theta).$$

Rigorous details for the validity of this approximation, a suitable class of functions for which it applies, and rates of convergence in λ are contained in Section 3.1. For now, we discuss the implementation in the case of a bivariate normal distribution constrained to a triangular region in the $\theta_1 - \theta_2$ plane.

Example: Constrained Gaussian under Linear Inequality

Suppose that $\theta = (\theta_1, \theta_2)$ follows a bivariate Gaussian with mean $\mu \in \mathbb{R}^2$ and covariance matrix, $\Sigma = \sigma^2 I_2$, $\sigma > 0$, and that θ is constrained to the triangular region $\mathcal{D} = \{(\theta_1, \theta_2) \mid \theta_1 + \theta_2 \leq 1, \theta_1 \geq 0, \theta_2 \geq 0\}$. Let $v_{\mathcal{D}}(\theta_1, \theta_2) = \max(\theta_1 + \theta_2 - 1, 0) + \max(-\theta_1, 0) + \max(-\theta_2, 0)$. Then, $v_{\mathcal{D}}(\theta_1, \theta_2) = 0 \ \forall \theta \in \mathcal{D}$. Otherwise, $v_{\mathcal{D}}$ is positive. The fully constrained density of θ given that $\theta \in \mathcal{D}$ is then,

$$\pi_{\mathcal{D}}(\theta_1, \theta_2) = \pi(\theta_1, \theta_2 \mid \theta \in \mathcal{D}) = \frac{\exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2} \right) \mathbb{1}_{\mathcal{D}}(\theta)}{\int_0^1 \int_0^{1-\theta_1} \exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2} \right) d\theta_2 d\theta_1}$$

$$\propto \exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2} \right) \mathbb{1}_{\{\theta_1+\theta_2 \leq 1\} \cap \{\theta_1 \in [0,\infty)\} \cap \{\theta_2 \in [0,\infty)\}}.$$

Alternatively, the relaxed density, $\tilde{\pi}_\lambda(\theta)$, is

$$\tilde{\pi}_\lambda(\theta) = \frac{\exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2} - \frac{1}{\lambda} v_{\mathcal{D}}(\theta) \right)}{\int_{\mathbb{R}^2} \exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2} - \frac{1}{\lambda} v_{\mathcal{D}}(\theta) \right) d\theta_2 d\theta_1} \tag{5}$$

$$\propto \exp\left(-\frac{\|\theta-\mu\|^2}{2\sigma^2} \right) \exp\left(-\frac{1}{\lambda}[\max(\theta_1 + \theta_2 - 1, 0) + \max(-\theta_1, 0) + \max(-\theta_2, 0)] \right).$$

Figure 1 depicts a few plots of the relaxed density as λ decreases. For $\lambda = 10^{-2}$, the relaxed density still places some probability outside of the constrained region, as a smooth decrease in the density can be observed near the boundary of the triangle. For $\lambda = 10^{-4}$ the density rapidly drops to 0 outside of $\mathcal{D}$.

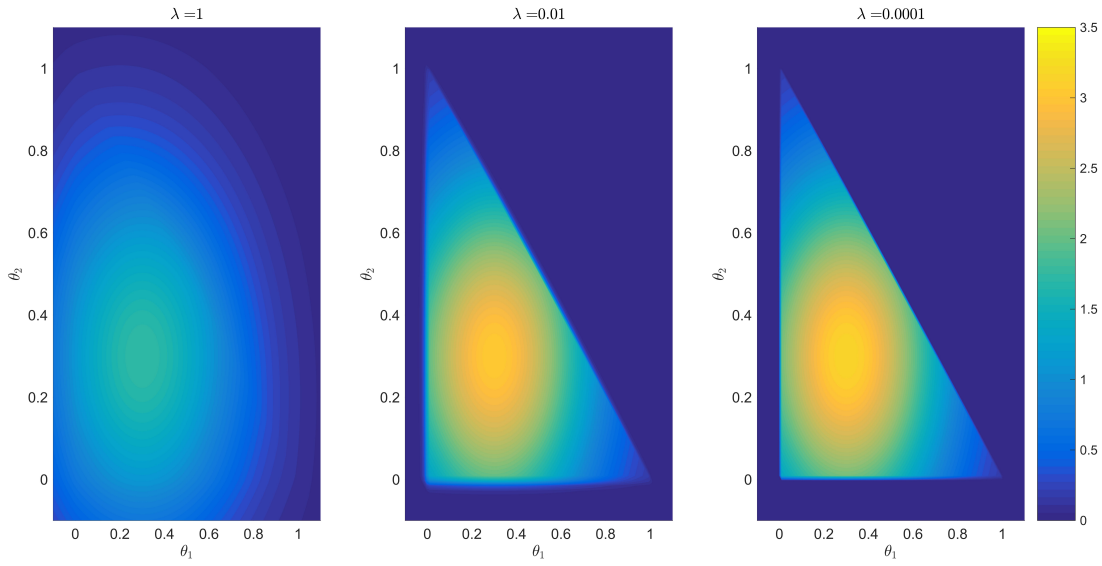


Figure 1: The relaxed distribution from Eq. (5) for decreasing λ is shown above when $\mu = [0.3, 0.3]^T$ and $\sigma^2 = 1/10$. For $\lambda = 1$, the distribution is similar to the unconstrained Gaussian. As λ decreases, the relaxation is reduced and the triangular constrained region becomes more apparent.

2.2 Constrained Space with Zero Measure

In the second case, we consider when $\mathcal{D}$ is a measure zero subset of $\mathcal{R}$, i.e. $\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}}(\theta) = 0$. Letting $\mathcal{R}$ have dimensionality r, we restrict ourselves to the setting where $\mathcal{D}$ can be represented implicitly as the solution set of a consistent system of equations $\{v_j(\theta) = 0\}_{j=1}^{s}$, so that $\mathcal{D} = \{\theta \mid v_j(\theta) = 0, j = 1, \ldots, s\}$ is a $(r - s)$-dimensional submanifold of $\mathcal{R}$. While we impose restrictions, many common constraints (e.g. $\sum_i \theta_i = 1$, $\theta^T\theta = I$) fall into this category.

Due to the zero measure, one cannot obtain conditional probability as before by simply re-normalizing $\left[\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}}(\theta)\right]^{-1}$. Instead, we resort to the more general definition of conditional probability, named *regular conditional probability* (r.c.p.) (Kolmogorov, 1950) to derive a constrained density coherent with $\pi_{\mathcal{R}}$.

More technical definition of the r.c.p. is provided in the appendix. For now, the following intuition is sufficient. While $\mathcal{D}$ has zero r-dimensional volume (i.e. zero Lebesque measure), it has a positive $(r - s)$-dimensional 'surface area', formally known as normalized Hausdorff measure. We can use it as the normalizing constant to obtain a r.c.p density:

$$\pi_{\mathcal{D}}(\theta \mid Y) = \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))\mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))d\bar{\mathcal{H}}^{(r-s)}(\theta)} \propto \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))\mathbb{1}_{\mathcal{D}}(\theta).$$

where $J(\nu_{\mathcal{D}}(\theta)) = \sqrt{(D\nu_{\mathcal{D}})'(D\nu_{\mathcal{D}})}$ is the Jacobian of $\nu_{\mathcal{D}}$, which we assume is positive. This term is introduced as we fix $\{v_j(\theta) = 0\}_{j=1}^{s}$. The density is defined with respect to the $\bar{\mathcal{H}}^{(r-s)}$.

We take a similar strategy by considering $\|\nu_{\mathcal{D}}(\theta)\|_1$ as the distance from θ to $\mathcal{D}$. Thus, $\|v_{\mathcal{D}}(\theta)\|_1 = 0$ implies that $\theta \in \mathcal{D}$ whereas, $\|v_{\mathcal{D}}(\theta)\|_1 > 0$ implies that $\theta \notin \mathcal{D}$. To construct a relaxed density, we expand support in the neighborhood of $\|v_{\mathcal{D}}(\theta)\| = 0$ with a factor $\exp\left(-\frac{1}{\lambda}\|v(\theta)\|\right)\mathbb{1}_{\mathcal{X}}(v(\theta))$, yielding

$$\int_{\mathcal{A}} \tilde{\pi}_\lambda(\theta)d\mu(\theta) = \frac{\int_{\mathcal{X}}\left[\int_{\{\theta:v(\theta)=x\}\cap\mathcal{A}} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))d\bar{\mathcal{H}}^{(r-s)}(\theta)\right]\exp\left(-\frac{1}{\lambda}\|x\|\right)dx}{\int_{\mathcal{X}}\left[\int_{\{\theta:v(\theta)=x\}} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)J^{-1}(\nu_{\mathcal{D}}(\theta))d\bar{\mathcal{H}}^{(r-s)}(\theta)\right]\exp\left(-\frac{1}{\lambda}\|x\|\right)dx}$$

where $\mathcal{X}$ is chosen to ensure the denominator is finite and positive. Using co-area formula (Federer, 2014) to convert double integrals to single Lebesgue integral, then removing the integral over $\mathcal{A}$, this simplifies to a density:

$$\tilde{\pi}_\lambda(\theta) = \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}\|\nu_{\mathcal{D}}(\theta)\|\right)\mathbb{1}_{\mathcal{X}}(v(\theta))}{\int_{\mathcal{R}} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}\|\nu_{\mathcal{D}}(\theta)\|\right)\mathbb{1}_{\mathcal{X}}(v(\theta))d\mu_{\mathcal{R}}(\theta)} \propto \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{1}{\lambda}\|\nu_{\mathcal{D}}(\theta)\|\right)\mathbb{1}_{\mathcal{X}}(v(\theta)), \tag{6}$$

which is defined with respect to $\mu_{\mathcal{R}}$. Note the Jacobian term vanishes during the transformation, yielding a relaxed density very similar to (2) in the last section.

Much like the positive measure case, $\exp\left(-\frac{1}{\lambda}\|\nu_{\mathcal{D}}(\theta)\|\right)$ converges pointwise to $\mathbb{1}_{\mathcal{D}}(\theta)$. As $\lambda \to 0^+$, this multiplicative factor is concentrating the probability to a small layer around the constrained space. As a result, for small λ, one could expect that

$$\int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta)d\mu_{\mathcal{R}} \approx (\theta)\int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}d\bar{\mathcal{H}}^{(r-s)}(\theta)$$

We provide details of this result, and suitable class of functions for which it applies in the next section. Before that, we first provide another example to illustrate.

Example: Constrained Gaussian on Unit Circle

Suppose that $\theta = (\theta_1, \theta_2)$ follows a bivariate Gaussian with mean $\mu \in \mathbb{R}^2$ and covariance matrix, $\Sigma = \sigma^2 I_2$, $\sigma > 0$ which is constrained to the unit circle $\mathcal{D} = \{(\theta_1, \theta_2) \mid \theta_1^2 + \theta_2^2 = 1\}$. Since the unit circle is one-dimensional and $\theta = (\theta_1, \theta_2)$ is two-dimensional, we use a (2-1)=1-dimensional constraint function

$$v_{\mathcal{D}}(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2 - 1.$$

Then, $v_{\mathcal{D}}(\theta_1, \theta_2) = 0 \ \forall \theta \in \mathcal{D}$. Otherwise, $v_{\mathcal{D}}$ is non-zero. Furthermore, $J(\nu_{\mathcal{D}}(\theta)) = 2\|\theta\|_2 = 2$ for $\theta \in \mathcal{D}$. The fully constrained density of θ given that $\theta \in \mathcal{D}$ is then,

$$\begin{aligned}
\pi_{\mathcal{D}}(\theta_1, \theta_2) = \pi(\theta_1, \theta_2 \mid \theta \in \mathcal{D}) &= \frac{\exp\left(- \frac{\|\theta - \mu\|^2}{2\sigma^2} \right) \mathbb{1}_{\mathcal{D}}(\theta)}{\int_{\mathcal{D}} \exp\left(- \frac{\|\theta - \mu\|^2}{2\sigma^2} \right) d\bar{\mathcal{H}}^1(\theta)} \\
&\propto \exp\left(- \frac{\|\theta - \mu\|^2}{2\sigma^2} \right) \mathbb{1}_{\theta_1^2 + \theta_2^2 = 1} \\
&\propto \exp\left(\frac{\theta' \mu}{\sigma^2} \right) \mathbb{1}_{\theta_1^2 + \theta_2^2 = 1}.
\end{aligned}$$

This density can be interpreted with respect to the normalized Hausdorff-1 measure on the unit circle which coincides with arclength in this case. Observe this is the von Mises–Fisher distribution on the unit circle with location $\mu / \|\mu\|_2$ and concentration $\|\mu\|_2 / \sigma^2$.

The relaxed density, $\tilde{\pi}_\lambda(\theta)$ with a compact $\mathcal{X}$ such that $0 \in \mathcal{X}$ is

$$\begin{aligned}
\tilde{\pi}_\lambda(\theta) &= \frac{\exp\left(- \frac{\|\theta - \mu\|^2}{2\sigma^2} - \frac{1}{\lambda} \|v_{\mathcal{D}}(\theta)\| \right) \mathbb{1}_{\mathcal{X}}(v_{\mathcal{D}}(\theta))}{\int_{\mathbb{R}^2} \exp\left(- \frac{\|\theta - \mu\|^2}{2\sigma^2} - \frac{1}{\lambda} \|v_{\mathcal{D}}(\theta)\| \right) \mathbb{1}_{\mathcal{X}}(v_{\mathcal{D}}(\theta)) d\theta_2 d\theta_1} \\
&\propto \exp\left(- \frac{\|\theta - \mu\|^2}{2\sigma^2} \right) \exp\left(- \frac{1}{\lambda} |\theta_1^1 + \theta_2^2 - 1| \right) \mathbb{1}_{\mathcal{X}}(\theta_1^1 + \theta_2^2 - 1).
\end{aligned} \tag{7}$$

Figure 2 depicts a few plots of the relaxed density as λ decreases. For $\lambda = 10^{-2}$ the constraint along the circle is clear. While the relaxed density still places some small probability outside of the constrained region, the rightmost plot becomes similar to the von-Mises distribution on the circle plotted in two dimensions.

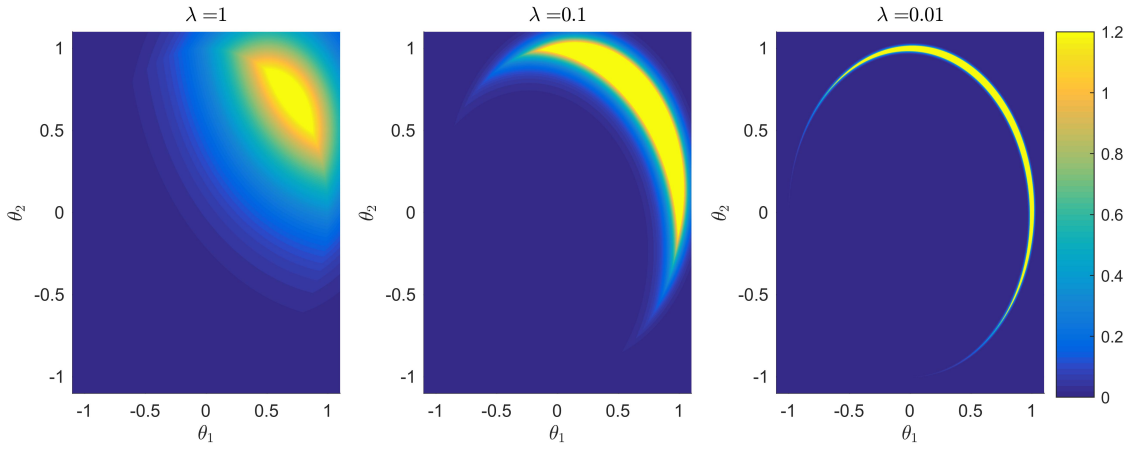


Figure 2: The relaxed distribution from Eq. (7) for decreasing λ is shown above when $\mu = (1/\sqrt{2}, 1/\sqrt{2})$ and $\sigma^2 = 1/25$. As λ decreases, the relaxation is reduced and the circular constraint becomes clear.

3 Theory

In this section, we provide more theoretic justification for the proposed method.

3.1 Constrained Space with Positive Measure

For constrained space with positive measure, we now focus on quantifying the difference between constrained and relaxed densities. Both of these densities are absolutely continuous with respect to Lebesque measure on $\mathcal{R}$. Thus, the expectation of g with respect to constrained density is

$$E[g(\theta)|\theta \in \mathcal{D}] = \int_{\mathcal{D}} g(\theta)\pi_{\mathcal{D}}(\theta)d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}. \tag{8}$$

Similarly, the expected value of g with respect to the relaxed density,

$$E_{\tilde{\pi}_\lambda}[g(\theta)] = \int_{\mathcal{R}} g(\theta)\tilde{\pi}_\lambda(\theta)d\mu_{\mathcal{R}} = \frac{\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}} \mathcal{L}(\theta;Y)\exp(-v_{\mathcal{D}}(\theta)/\lambda)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}. \tag{9}$$

We can now consider the behavior of $E_{\tilde{\pi}_\lambda}[g]$ as $\lambda \to 0^+$.

Lemma 1. *Suppose $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}})$. Then,*

$$\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \leq \frac{\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}{\left[\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right]^2}$$

where $E|g(\theta)| \propto \int_{\mathcal{R}} |g(\theta)|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}d\mu_{\mathcal{R}}(\theta)$ is the expected value of $|g(\theta)|$ with respect to the unconstrained posterior density and $C_{\mathcal{R}} = \int_{\mathcal{R}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$ is the normalizing constant of this unconstrained posterior density. Furthermore, if $v_{\mathcal{D}}(\theta)$ is zero for all $\theta \in \mathcal{D}$ and positive for $\theta \in (\mathcal{R}\setminus\mathcal{D})^o$, it follows from the

dominated convergence theorem that

$$\left| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \right| \to 0 \text{ as } \lambda \to 0^+.$$

Thus, one can obtain sufficiently accurate estimates of $E[g|\theta \in \mathcal{D}]$ by sampling from $\tilde{\pi}_\lambda$ when λ is sufficiently small. From a practical standpoint, it is desirable to understand the rate at which $E_{\tilde{\pi}_\lambda}[g(\theta)]$ converges to $E[g(\theta) \in \mathcal{D}]$. This question is addressed in the following theorem.

Theorem 1. *Suppose $g \in \mathbb{L}^2(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}})$, $v_{\mathcal{D}}(\theta)$ has the form of Eq. (3) with $k = 2$, $\mathcal{D}$ has a piecewise smooth boundary, and that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on a open neighborhood containing $\mathcal{D}$. Then for $0 < \lambda \ll 1$,*

$$\left| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \right| = O(\sqrt{\lambda}).$$

This theorem follows by applying the Cauchy-Schwartz inequality to the term in the numerator of the bound given in Lemma 1. One can attain a bound depending on the surface area of $\mathcal{D}$ when it is bounded. The proofs of Lemma 1 and Theorem 1 are contained in Appendix A.

These results have some important implications both analytically and numerically. First, in addition to point estimates, $E[\theta \mid \theta \in \mathcal{D}]$, it is possible to approximate probabilities $P(\theta \in \mathcal{F}|\theta \in \mathcal{D})$ and higher moments, e.g. $E[\Pi_j \theta_j^{k_j}|\theta \in \mathcal{D}]$, so long as these moments exist for the unconstrained posterior density.

Secondly, these bounds demonstrate that the error in using the relaxed density to approximate $E[g(\theta)|\theta \in \mathcal{D}]$ is proportional to $\sqrt{\lambda}[\int_{\mathcal{D}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)]^{-2}$ although this rate may not be optimal. In practice, λ may need to be very small, particularly in the case where $0 < P(\theta \in \mathcal{D}) \ll 1$. Of course, specific details of the scaling of $\left| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \right|$ will depend upon the choice of $\mathcal{D}$ and $v_{\mathcal{D}}(\theta)$.

In general, one avenue for mitigating numerical difficulties which may arise when $\lambda \ll 1$ is needed is to use (4) to relax the density in directions where accuracy is less important. Fortunately, both Lemma 1 and Theorem 1 will still hold when the matrix A is positive definite since this corresponds to a linear re-scaling of the parameter space.

3.2 Constrained Space with Zero Measure

Before investigating the measure zero case, we review a few important concepts of geometric measure theory which are used throughout this section. First, recall the definition of Hausdorff measure.

Definition - Hausdorff Measure. *Let $A \subset \mathcal{R}^r$. Fix $s \leq r$. Then*

$$\mathcal{H}^s(A) = \lim_{\delta \to 0} \inf \left\{ \sum [diam(S_i)]^s : A \subseteq \cup S_i, diam(S_i) \leq \delta, diam(S_i) = \sup_{x,y \in S} \|x - y\| \right\}$$

.

We denote the normalized Hausdorff measure as $\bar{\mathcal{H}}^s(A) = \frac{\Gamma(\frac{1}{2})^s}{2^s\Gamma(\frac{s}{2}+1)}\mathcal{H}^s(A)$. When $s = r$, Lebesgue and normalized Hausdorff measures coincide $\mu_{\mathbb{R}^m}(A) = \bar{\mathcal{H}}^s(A)$ (Evans and Gariepy, 2015). Additionally, for a subset $\mathcal{D}$, there exists a unique, critical value d such that

$$\bar{\mathcal{H}}^s(\mathcal{D}) = \begin{cases} 0, & s > d \\ \infty, & s < d. \end{cases}$$

The critical value, d, is referred to as the Hausdorff dimension of $\mathcal{D}$. We note that, when $\mathcal{D}$ is a compact, d-dimensional submanifold of $\mathbb{R}^m$, it will have Hausdorff dimension d and $\bar{\mathcal{H}}^d(\mathcal{D})$ is the d-dimensional surface area of A.

We now state the co-area formula which is used to define a regular conditional probability on the measure zero constrained space $\mathcal{D}$ and is pivotal in all of the proofs of the theorems.

Theorem 2. *Co-area formula (Diaconis et al., 2013; Federer, 2014) Suppose $\nu : \mathbb{R}^r \to \mathbb{R}^s$ with $s < r$ is Lipschitz and that $g \in \mathbb{L}^1(\mathbb{R}^r, \mu_{\mathbb{R}^r})$. Assume $J(\nu(\theta)) > 0$, then*

$$\int_{\mathbb{R}^r} g(\theta) J\nu(\theta) d\mu_{\mathbb{R}^r}(\theta) = \int_{\mathbb{R}^s} \left(\int_{\nu^{-1}(y)} g(\theta) d\bar{\mathcal{H}}^{r-s}(\theta) \right) d\mu_{\mathbb{R}^s}(y), \tag{10}$$

The behavior of the pre-images $\nu^{-1}(y)$ in the co-area formula are important for the convergence results presented later in this section. As such, we assume that $\mathcal{D}$ can be defined implicitly as the solution set to a system of s equations, $\{\nu_j(\theta) = 0\}_{j=1}^s$, where

(a) $\nu_j : \mathcal{R} \to \mathbb{R}$ is Lipschitz continuous,

(b) $v_j(\theta) = 0$ only for $\theta \in \mathcal{D}$,

(c) for $k = 1, \ldots, s$, the preimage $v_k^{(-1)}(x)$ is a co-dimension 1 sub-manifold of $\mathcal{R}$ for $\mu_{\mathbb{R}}$-a.e. x in the range of ν_k,

(d) $\nu_j^{(-1)}(0)$ and $\nu_k^{(-1)}(0)$ intersect transversally for $1 \leq j < k \leq s$.

We refer to the functions $\nu_1, \ldots, \nu_s$ as constraint functions. In this case, if we let $\nu : \mathcal{R} \to \mathbb{R}^s$ be the vector-valued function $\nu(\theta) = [\nu_1(\theta), \ldots, \nu_s(\theta)]^T$, then $\mathcal{D} = \ker(v)$ is a co-dimension s submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. x the range of v. Recall, the ambient space, $\mathcal{R}$, is $r-$dimensional. Therefore, it follows that $\mathcal{D}$ is a $(r-s)-$dimensional submanifold of $\mathcal{R}$, and it is natural to discuss the $(r-s)$-dimensional surface area of $\mathcal{D}$.

Property (a), guarantees that ν is itself Lipschitz. The remaining properties (b)-(d) are constructed so that $\nu^{(-1)}(x)$ for $x \in \mathbb{R}^s$ is also a submanifold which is close to $\mathcal{D} = \nu^{(-1)}(0)$ when x is near zero. In particular,

the assumption of transversality ensures that $\nu^{(-1)}(x)$ with also be $r-s$ dimensional for x sufficiently close to 0.

The existence and uniqueness of the constraints must be addressed. In the case where $\mathcal{D}$ is specified by a collection of equality constraints – such as the probability simplex or the Stiefel manifold for example– it is not difficult to find a suitable set of constraint functions. Table 1 contains a number of examples of common constrained spaces and appropriate choices of constraint functions.

$\mathcal{R}$	$\mathcal{D}$	$\dim(R)$	$\dim(D)$	Constraint functions
$[0,1]^r$	Probability simplex, Δ^{r-1}	r	$r-1$	$v(\theta) = \sum(\theta) - 1$
$\mathbb{R}^r$	Line, span$\{\vec{u}\}$ $\vec{u} \neq \vec{0}$	r	1	$\nu_j(\vec{\theta}) = \vec{\theta}^T \vec{b}_j$ $\{\vec{b}_1, \ldots, \vec{b}_{r-1}\}$ a basis for span$\{\vec{u}\}^\perp$
$[-1,1]^r$	Unit sphere, $\mathbb{S}^{r-1}$	r	$r-1$	$v(\theta) = (\|\theta\|^2 - 1)$
$[-1,1]^{n\times k}$	Stiefel manifold, $V_k(\mathbb{R}^n)$ $\theta = [\vec{\theta}_1 \vert \ldots \vert \vec{\theta}_k]$, $\vec{\theta}_j \in \mathbb{R}^n$	nk	$nk - \frac{1}{2}k(k+1)$	$v_{i,j}(\theta) = (\vec{\theta}_i' \vec{\theta}_j - \delta_{i,j})$ $1 \leq i \leq j \leq k$ and $\delta_{i,j} = \mathbb{1}_{i=j}$

Table 1: Table of constraints for some commonly used constrained spaces.

With regards to uniqueness, we note that the constraints cannot be unique in any case. For example, rescaling in each $v_j(\theta)$ will also satisfy (a)-(d). Naturally, an optimal choice will depend largely on the properties of the constrained distribution that one wishes to estimate making the choice of $\{\nu_j\}_{j=1}^s$ context dependent.

Under the given construction of the constrained space, we can now specify the regular conditional probability of θ, given $\theta \in \mathcal{D}$.

Theorem 3. *(Diaconis et al., 2013) Assume that $J(v(\theta)) > 0$ and that for each $z \in \mathbb{R}^s$ there is a finite non-negative p_z such that,*

$$m^{p_z}(z) = \int_{v^{-1}(z)} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} d\bar{\mathcal{H}}^p(\theta) \in (0, \infty).$$

Then, for any Borel subset E, of $\mathcal{R}$, it follows that

$$P(E \mid v(\theta) = z) = \begin{cases} \frac{1}{m^{p_z}(z)} \int_E \frac{\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=z}}{J(v(\theta))} d\bar{\mathcal{H}}^p(\theta) & m^p(z) \in (0,\infty) \\ \delta(E) & m^p(z) \in \{0, \infty\} \end{cases}$$

is a valid regular conditional probability for $\theta \in \mathcal{D}$. Here, $\delta(E) = 1$ if $0 \in E$ and 0 otherwise.

By construction, $\{\theta : v(\theta) = z\}$ is a $(r-s)$ dimensional submanifold of $\mathcal{R}$ for $\mu_{\mathbb{R}^s}$-a.e. z in $\mathcal{X}$, the range of v. As such, it follows that one should take $p_z = r - s$. It is possible that $m^p(z) \in \{0, \infty\}$ for some $z \notin \mathcal{X}$;

however, they are excluded during our construction. See Diaconis et al. (2013) for additional discussion of this issue. Most importantly, $0 \in \mathcal{X}$, therefore, Theorem 3 allows us to define

$$\pi_{\mathcal{D}}(\theta|\theta \in \mathcal{D}, Y) = \frac{1}{m^{r-s}(0)} \frac{\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=0}}{J(v(\theta))} \tag{11}$$

as the constrained posterior density.

As a result, we can define the conditional expectation of $g(\theta)$ given $\theta \in \mathcal{D}$ as

$$E[g(\theta)|\theta \in \mathcal{D}] = E[g(\theta)|\nu(\theta) = 0\,] = \int_{\mathcal{R}} g(\theta)\pi_{\mathcal{D}}(\theta)d\bar{\mathcal{H}}^{r-s}(\theta).$$

The expected value of $g(\theta)$ with respect to the relaxed density, denote $E_{\tilde{\Pi}}[g(\theta)]$, is

$$E_{\tilde{\Pi}}[g(\theta)] = \frac{1}{m_\lambda} \int_{\mathcal{R}} g(\theta)\pi_{\mathcal{R}}(\theta)\mathcal{L}(Y;\theta) \exp\Big(-\frac{1}{\lambda}\|\nu(\theta)\|_1 \Big) d\mu_{\mathcal{R}}(\theta)$$

with $m_\lambda = \int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)\mathcal{L}(Y;\theta) \exp\Big(-\lambda^{-1}\|\nu(\theta)\|_1 \Big) d\mu_{\mathcal{R}}(\theta)$. The primary results of the section are the following statements regarding the use of $E_{\tilde{\Pi}}[g]$ to estimate $E[g|\theta \in \mathcal{D}]$.

Theorem 4. *Let $m : \mathbb{R}^s \to \mathbb{R}$ and $G : \mathbb{R}^s \to \mathbb{R}$ be defined as follows*

$$m(x) = \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\mathcal{R}^{r-s}(\theta)$$

$$G(x) = \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\mathcal{R}^{r-s}(\theta).$$

Suppose that both m and G are continuous on an open interval containing the origin and that $g \in \mathbb{L}^1(\mathcal{R}, \pi_R\mathcal{L}(y;\theta)d\mu_{\mathcal{R}})$. Then,

$$\Big| E_{\tilde{\Pi}}[g] - E[g|\theta \in \mathcal{D}] \Big| \to 0 \text{ as } \lambda \to 0^+.$$

Corollary 1. *In addition to the assumptions of Theorem 4, suppose that both m and G are differentiable at 0. Then*

$$\Big| E_{\tilde{\Pi}}[g] - E[g|\theta \in \mathcal{D}] \Big| = O\Big(\frac{\lambda}{|\log \lambda|^s} \Big)$$

as $\lambda \to 0^+$.

Like the results from Section 3.1, the convergence rates are sub-linear. However, unlike the positive measure case, the convergence rates are dimension dependent.

4 Posterior Computation

Compared to constrained density in space $\mathcal{D}$, relaxed density is supported in $\mathcal{R}$ and can be directly sampled via off–the–shelf tools such as slice sampling, adaptive Metropolis-Hastings and Hamiltonian Monte Carlo (HMC). In this section, we focus on HMC for its easiness to use and good performance in block updating of parameters.

4.1 Hamiltonian Monte Carlo under Constraint Relaxation

We provide a brief overview of HMC for continuous θ^* under constraint relaxation. Discrete extension is possible via recent work of Nishimura et al. (2017).

In order to sample θ, HMC introduces an auxillary momentum variable $p \sim \text{No}(0, M)$. The covariance matrix M is referred to as a *mass matrix* and is typically chosen to be the identity or adapted to approximate the inverse covariance of θ. HMC then sample from the joint target density $\pi(\theta, p) = \pi(\theta)\pi(p) \propto \exp(-H(\theta, p))$ where, in the case of the posterior under relaxation,

$$\begin{aligned} &H(\theta, p) = U(\theta) + K(p), \\ &\text{where } U(\theta) = -\log \pi(\theta), \\ &\qquad K(p) = \frac{p' M^{-1} p}{2}. \end{aligned} \tag{12}$$

with $\pi(\theta)$ is the unnormalized density in (2) or (6).

From the current state $(\theta^{(0)}, p^{(0)})$, HMC generates a proposal for Metropolis-Hastings algorithm by simulating Hamiltonian dynamics, which is defined by a differential equation:

$$\begin{aligned} \frac{\partial \theta^{(t)}}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = M^{-1} p, \\ \frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \end{aligned} \tag{13}$$

The exact solution to (13) is typically intractable but a valid Metropolis proposal can be generated by numerically approximating (13) with a reversible and volume-preserving integrator (Neal, 2011). The standard choice is the *leapfrog* integrator which approximates the evolution $(\theta^{(t)}, p^{(t)}) \to (\theta^{(t+\epsilon)}, p^{(t+\epsilon)})$ through the following update equations:

$$p \leftarrow p - \frac{\epsilon}{2} \frac{\partial U}{\partial \theta}, \quad \theta \leftarrow \theta + \epsilon M^{-1} p, \quad p \leftarrow p - \frac{\epsilon}{2} \frac{\partial U}{\partial \theta} \tag{14}$$

Taking L leapfrog steps from the current state $(\theta^{(0)}, p^{(0)})$ generates a proposal $(\theta^*, p^*) \approx (\theta^{(L\epsilon)}, p^{(L\epsilon)})$, which is accepted with the probability

$$1 \wedge \exp\left(-H(\theta^*, p^*) + H(\theta^{(0)}, p^{(0)}))\right)$$

We refer to this algorithm as CORE-HMC.

4.2 Computing Efficiency in CORE-HMC

Since CORE expands the support from $\mathcal{D}$ to $\mathcal{R}$, it is useful to study the effect of space expansion on the computing efficiency of HMC. In this subsection, we provide some quantification of the effects.

In understanding the computational efficiency of HMC, it is useful to consider the number of leapfrog steps to be a function of ϵ and set $L = \lfloor \tau/\epsilon \rfloor$ for a fixed integration time $\tau > 0$. In this case, the mixing rate of HMC is completely determined by τ in the limit $\epsilon \to 0$ (Betancourt, 2017). In practice, while a smaller stepsize ϵ leads to a more accurate numerical approximation of Hamiltonian dynamics and hence a higher acceptance rate, it takes a larger number of leapfrog steps and gradient evaluations to achieve good mixing. For an optimal computational efficiency of HMC, therefore, the stepsize ϵ should be chosen only as small as needed to achieve a reasonable acceptance rate (Beskos et al., 2013; Betancourt et al., 2014). A critical factor in determining a reasonable stepsize is the *stability limit* of the leapfrog integrator (Neal, 2011). When ϵ exceeds this limit, the approximation becomes unstable and the acceptance rate drops dramatically. Below the stability limit, the acceptance rate $a(\epsilon)$ of HMC increases to 1 quite rapidly as $\epsilon \to 0$ and in fact satisfies $a(\epsilon) = 1 - \mathcal{O}(\epsilon^4)$ (Beskos et al., 2013).

For simplicity, the following discussions assume the mass matrix M is taken to be the identity, and $\mathcal{D} = \cap_{j=1}^{s}\{\theta : v_j(\theta) = 0\}$. We denote $\mathcal{D}_j = \{\theta : v_j(\theta) = 0\}$ and use directional relaxation $\exp(-\sum_j \|v_j(\theta^*)\|\lambda_j^{-1})$. There are generally two factors limiting the efficiency of HMC: (i) the width of support in constrained space; (ii) the largest eigenvalue of the Hessian matrix. For the former, using $\mathcal{Q}$ to denote a support, the width of support is related to the shortest distance to the boundary $\eta(\theta; \mathcal{Q}) = \inf_{\theta' \notin \mathcal{Q}} \|\theta' - \theta\|$. If $\eta(\theta; \mathcal{Q}) \approx 0$ for all $\theta \in \mathcal{Q}$, the proposal would likely be rejected using a large leap-frog step size. In such case, it is useful to utilize CORE to expand support and increase $\eta(\theta; \mathcal{Q})$ for better computing efficiency. For the eigenvalue, let $\mathbf{H}_U(\theta)$ denote the hessian matrix of $U(\theta) = -\log \pi(\theta)$. The linear stability analysis and empirical evidences suggest that, for stable approximation of Hamiltonian dynamics by the leapfrog integrator in $\mathbb{R}^p$, the condition $\epsilon < 2\xi_1(\theta)^{-1/2}$ must hold on most regions of the parameter space (Hairer et al., 2006), with $\xi_1(\theta)$ the largest eigenvalue of $\mathbf{H}_U(\theta)$. The hessian is

$$\mathbf{H}_U(\theta) = -\mathbf{H}_{\log(\mathcal{L}(\theta;y)\pi_{\mathcal{R}}(\theta))}(\theta) + \sum_j \lambda_j^{-1}\mathbf{H}\|v_j(\theta)\|\mathbb{1}_{\theta\notin\mathcal{D}_j}. \tag{15}$$

Note the second term is zero unless θ is outside of $\mathcal{D}_k$. As λ_j^{-1} in the second term often dominates the eigenvalue in the first term, hence the effective eigenvalue often is proportional to $\min_{j:\theta\notin\mathcal{D}_j} \lambda_j^{1/2}$.

Lastly, one often may want to use CORE for obtaining approximate estimate under constrained model. For this purpose, we now provide a practical guidance on choosing λ_j. For $\mathcal{D}_j$'s with very small distance to support boundary $\eta(\theta;\mathcal{D}_j) \approx 0$, one should use moderate λ_j to increase support width; for $\mathcal{D}_j$'s without this issue, one should use very small $\lambda_j \approx 0$ to keep $\theta \in \mathcal{D}_j$, so that it has almost no influence on the hessian eigenvalue. To prevent high rejection of HMC near the boundary of $\mathcal{D}_j$, we suggest random step size ϵ at each iteration to reduce the error, which is typically recommended in HMC (Neal, 2011). When space expansion is exploited, a trade-off between approximation accuracy and computational efficiency is involved. Empirically, we found reducing λ_j 10 to times smaller requires approximately 3 times more computing time.

5 Simulated Examples

The simple computation of CORE frees up the modeling flexibility. We now illustrate more utility of the method via simulated examples.

Example: Sphere t Distribution

We now derive a new distribution on a $(p-1)$-sphere $\mathcal{D} = \{\theta \in \mathbb{R}^p : \|\theta\|_2 = 1\}$. Recall that von Mises–Fisher distribution (Khatri and Mardia, 1977) is the result of constraining a multivariate Gaussian $\theta \sim \text{No}(F, I\sigma^2)$ to $\mathcal{D}$

$$\pi_{\mathcal{D}}(\theta) \propto \exp(-\frac{\|F-\theta\|^2}{2\sigma^2})\mathbb{1}_{\theta'\theta=1} \propto \exp(\frac{F'}{\sigma^2}\theta)\mathbb{1}_{\theta'\theta=1}.$$

Although the final form appears more like an exponential, the behavior of von Mises–Fisher on sphere can be largely explained by its unconstrained parent Gaussian. In the Gaussian $\pi_{\mathcal{R}}(\theta)$, θ is symmetrically distributed around F, with density decaying exponentially as $\|\theta - F\|^2$ increases with rate $(2\sigma^2)^{-1}$; as the constrained density $\pi_{\mathcal{D}}(\theta)$ is proportional $\pi_{\mathcal{R}}(\theta)$, it concentrates similarly.

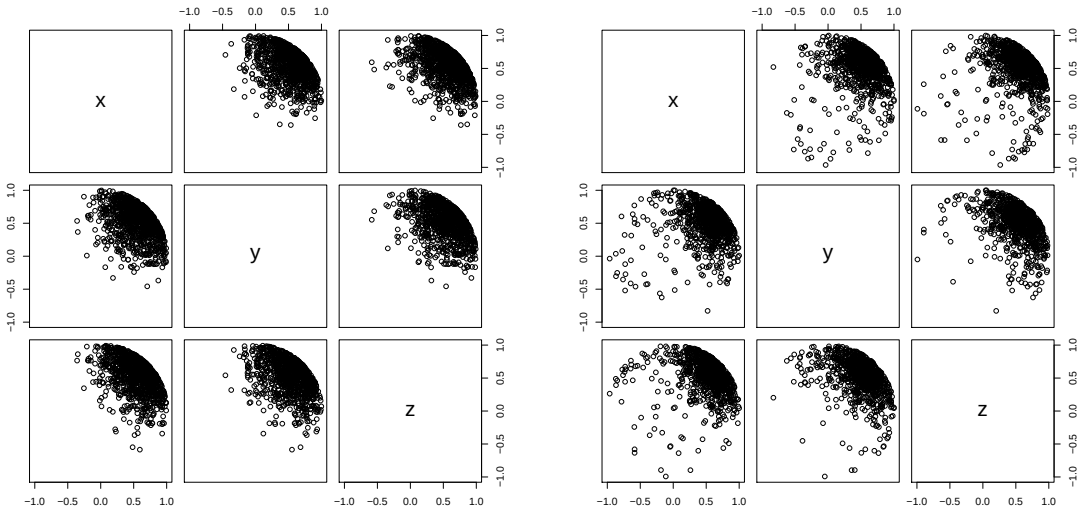
This naturally suggests we could use another distribution to induce different behavior on the sphere; then one could use CORE to generate approximate sample. We start from a multivariate t-distribution $\pi_{\mathcal{R}}(\theta)$,

$t_m(F, I\sigma^2)$ with m degrees of freedom, mean $F \in \mathcal{D}$ and variance $I\sigma^2$, using (11) to generate a density

$$\begin{aligned}\pi_{\mathcal{D}}(\theta) \propto & (1 + \frac{\|F-\theta\|^2}{m\sigma^2})^{-\frac{(m+p)}{2}} \mathbb{1}_{\theta'\theta=1} \\ \propto & (1 - \frac{F'\theta}{1+m\sigma^2/2})^{-\frac{(m+p)}{2}} \mathbb{1}_{\theta'\theta=1}\end{aligned} \tag{16}$$

As in the t-distribution, the density decays polynomially as $\|F-\theta\|^2$ increases, as opposed to the exponential decay in Gaussian. We refer to this new distribution as sphere t-distribution.

CORE allows us to easily obtain approximate sample via relaxation function $\exp(-\lambda^{-1}\|\theta'\theta-1\|)$. Figure 3 shows that the sphere t-distribution with $m = 3$ exhibits much less concentration than von Mises–Fisher on the sphere,. This can be useful for robust modeling when there could be 'outlier' on the sphere.



(a) von Mises–Fisher distribution.

(b) Sphere t-distribution with $m = 3$.

Figure 3: Sectional view of random samples from constrained distributions on a unit sphere inside $\mathbb{R}^3$. The distributions are derived through conditioning on $\theta'\theta = 1$ based on unconstrained densities of (a) $\text{No}(F, \text{diag}\{0.1\})$, (b) $t_3(F, \text{diag}\{0.1\})$, where $F = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]'$. The samples are generated via CORE-HMC with $\lambda = 10^{-3}$.

Example: Ordered Dirichlet Distribution

We derive an ordered Dirichlet distribution. We build it upon the canonical Dirichlet distribution $\text{Dir}(\alpha)$ with $\pi_{\mathcal{D}}(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1} \mathbb{1}_{\sum_{j=1}^{J}\theta_j=1}$ and further impose order constraint, $1 > \theta_1 \geq \ldots \geq \theta_J > 0$, yielding

$$\pi_{\mathcal{D}}(\theta) \propto \prod_{j=1}^{J} \theta_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^{J}\theta_j=1} \cdot \prod_{j=1}^{J-1} \mathbb{1}_{\theta_j \geq \theta_{j+1}}. \tag{17}$$

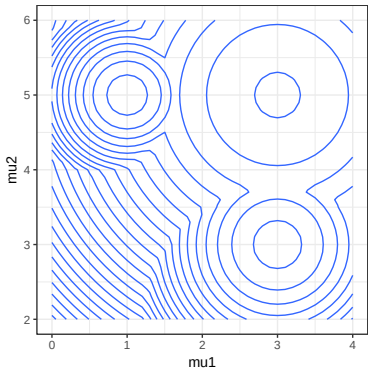
As commonly used in mixture model, canonical Dirichlet prior has its index j exchangeable. Since its

permutation does not change the likelihood, label-switching problem often occurs (reviewed in Jasra et al. (2005)). Naturally, order constraint in θ can alleviate this problem, especially in preventing the switch between large θ_j and small $\theta_{j'}$.

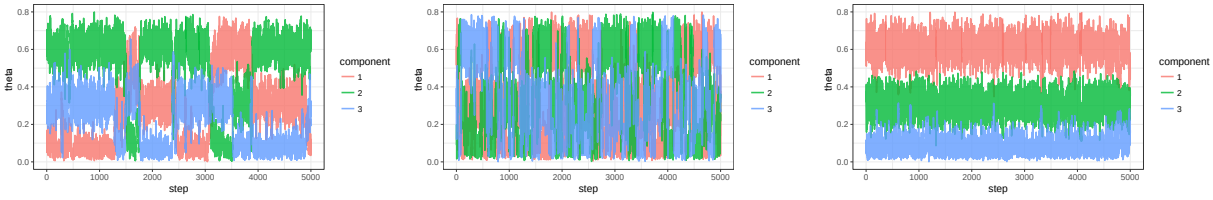
To illustrate, we consider a hierarchical normal distribution with a common variance but the mean from a mixture, for data $y_i \in \mathbb{R}^2$ indexed by $i = 1, \ldots, n$:

$$y_i \overset{indep}{\sim} \text{No}(\mu_i, \Sigma), \qquad \mu_i \overset{iid}{\sim} G, \qquad G(.) = \sum_{j=1}^{J} \theta_j \delta_{\mu_j}(.),$$

We generate $n = 100$ samples from 3 components with $\{\theta_1, \theta_2, \theta_3\} = \{0.6, 0.3, 0.1\}$, $\{\mu_1, \mu_2, \mu_3\} = \{[1, 5], [3, 3], [3, 5]\}$ and $\Sigma = I_2$. We assign weakly informative priors $\text{No}(0, 10I_2)$ for each μ_j and inverse–Gamma prior for the diagonal element in $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$ with $\sigma_1^2, \sigma_2^2 \sim \text{IG}(2, 1)$. Figure 4(a) shows the contour of posterior density of μ. The small component sample size leads to large overlap among the posterior.



(a) Posterior density of the component means $\{\mu_j\}_{j=1}^3$.



(b) Gibbs sampling of unordered Dirichlet

(c) HMC sampling of unordered Dirichlet

(d) HMC sampling of ordered Dirichlet

Figure 4: Contour of the posterior density of component means and traceplot of the posterior sample for the component weights w, in a 3-component normal mixture model. Panel (a) shows that there is significant overlap among component means $\{\mu_j\}_{j=1}^3$. Without ordering in θ, its traceplot shows label-switching issue in both Gibbs (b) and HMC (c) sampling of Dirichlet distribution. The ordered Dirichlet distribution has significantly less label-switching issue (d), where we utilize CORE to obtain approximate posterior sample.

The ordering disrupts the traditional Gibbs sampling (Ishwaran and James, 2001), however, one could still obtain approximate posterior using CORE. We use $\exp(-{\lambda_1}^{-1} \| \sum_{j=1}^{J} \theta_j - 1) \|) \prod_{j=1}^{J-1} \exp[-{\lambda_2}^{-1} \max(\theta_{j+1} -$

$\theta_j, 0)]$ to relax the constraints. We use $\lambda_1 = 10^{-3}$ on simplex constraint to allow efficient sampling and $\lambda_2 = 10^{-6}$ to induce almost no relaxation on the ordering.

We compare the traceplots of ordered Dirichlet and unordered Dirichlet. Without the order constraint, significant label-switching occur in both Gibbs and HMC (traceplots in Figure 4(b,c)), whereas ordered Dirchlet has almost no label-switching(Figure 4(d)).

6 Application: Finding Sparse Basis in a Population of Networks

We now consider a real data application in brain network analysis. The brain connectivity structures are obtained in the data set KKI-42 (Landman et al. 2011), which consists of $n = 21$ healthy subjects without any history of neurological disease. We take the first scan out of the scan-rescan data as the input. Each observation is a $V \times V$ symmetric network, recorded as an adjacency matrix A_i for $i = 1, \ldots, n$. The regions are constructed via the Desikan et al. (2006) atlas, for a total of V = 68 nodes. For the ith matrix A_i, $A_{i,k,l} \in \{0, 1\}$ is the element on the kth row and lth column of A_i, with $A_{(i,k,l)} = 1$ indicating there is an connection between kth and lth region, $A_{(i,k,l)} = 0$ if there is no connection. The matrix is symmetric due to the undirectedness of the network, but the diagonal records $A_{(i,k,k)}$ for all i and k are missing due to the lack of meaning for self-connectivity.

One scientific interest in neuroscience is to quantify the variation of brain networks and identify the regions (nodes) that contribute to it. Extending factor analysis to multiple matrices, one appealing approach is to have the networks share a common factor matrix but let the loadings vary across subjects. This can be considered as a simplified equivalent of three-way tensor factorization (Kolda and Bader, 2009). Then to selectively identify the important nodes, one natural way is to apply shrinkage on the elements of factor matrix.

Geometrically, the factor matrix, denoted by $\{U_1, \ldots, U_d\}$, reside on a Stiefel manifold $\mathcal{V}(n, d) = \{U : U'U = I_d\}$, where $U = [U_1, \ldots, U_d]$ is the $n \times d$ matrix. Using r to index $1, \ldots, d$, each frame U_r represents a $(n-1)$-hypersphere. Applying shrinkage forces some of its sub-coordinates to be close to 0, which is reducing each U_r onto a lower-dimensional hypersphere. Although previous work was done using sparse PCA (Zou et al., 2006) for continuous outcome, little work has been done in a probabilistic model for binary matrices.

To apply shrinkage in the constrained space, we adopt the induced prior as common in Bayesian literature (reviewed by Polson and Scott (2012)), which usually takes the form hierarchical structure $\theta_i \mid \kappa_i, \sigma \sim \text{No}(0, \kappa_i\sigma), \quad \kappa_i \sim G_1, \quad \sigma \sim G_2$ with κ_i, σ as the local and global scale parameters. However, when constraining θ_i, one caveat would be only adapting the conditional density $\text{No}(\theta_i; \kappa_i\sigma)$, which yields intractable normalizing constant involving $\kappa_i\sigma$ in the conditional. This difficulty can be avoided by reparameterizing $\theta_i = \eta_i\kappa_i\sigma$ with $\eta_i \sim \text{No}(0, 1)$, and adapting the *joint* density of $\{\eta_i, \kappa_i, \sigma\}$ on constrained space instead. The

joint density will not have intractable constant as long as the hyper-parameters in G_1 and G_2 are fixed.

We now take the Dirichlet-Laplace prior (Bhattacharya et al., 2015) as unconstrained distribution $\pi_{\mathcal{R}}$ and adapt it onto Stiefel manifold via (**??**).

$$
\begin{aligned}
&A_{(i,k,l)} \sim \text{Bern}(\frac{1}{1+\exp(-\psi_{(i,k,l)} - z_{(k,l)})}) \\
&\psi_{(i,k,l)} = \sum_{r=1}^{d} v_{(i,r)} u_{(k,r)} u_{(l,r)} \\
&U'U = I_d \text{ with } U = \{u_{(k,r)}\}_{k=1,\ldots,n;r=1,\ldots,d} \\
&u_{(k,r)} = \eta_{(k,r)} \kappa_{(k,r)} \sigma_u \\
&\eta_{(k,r)} \sim \text{Lap}(0,1), \quad \{\kappa_{(1,r)} \ldots \kappa_{(V,r)}\} \sim \text{Dir}(\alpha), \quad \sigma_u^2 \sim \text{IG}(2,1) \\
&z_{(k,l)} \sim \text{No}(0, \sigma_z^2), \quad \sigma_z^2 \sim \text{IG}(2,1) \\
&v_{(i,r)} \sim \text{No}(0, \sigma_{v,(r)}^2), \quad \sigma_{v,(r)}^2 \sim \text{IG}(2,1)
\end{aligned}
$$

for $k > l$, $k = 2, \ldots, V$, $i = 1, \ldots, n$; $\text{Lap}(0,1)$ denotes the Laplace distribution centered at 0 with scale 1; $Z = \{z_{(k,l)}\}_{k=1,\ldots,V;l=1,\ldots,V}$ is a symmetric unstructured matrix that serves as the latent mean; $\{v_{(i,r)}\}_{r=1,\ldots,d}$ is the loading for the ith network, with each $v_{(i,r)} > 0$; for all other scale parameters $\sigma_\cdot^2$, we choose weakly informative prior inverse Gamma $\text{IG}(2,1)$, as appropriate for the scale under the logistic link. To induce sparsity in each Dirichlet, we use $\alpha = 0.1$ as suggested by Bhattacharya et al. (2015).

There are two types of constraints in the model, $U'U = I_d$ and $\sum_{k=1}^{V} \kappa_{(k,r)} = 1$ for $r = 1, \ldots, d$. Taking $v_1(U) = U'U - I_d$ and $v_2(\kappa_{(k,r)}) = \sum_{k=1}^{V} \kappa_{(k,r)} - 1$ for each r, the Jacobian is constant in (**??**). For posterior computation, we use DA-CORE as described above. Using latent variable w_U d-by-d upper triangular and positive diagonal matrix, and $w_{\kappa,(r)} > 0$ for $r = 1, \ldots, d$, we relax the parameters to

$$
U^* = U w_U, \quad \kappa^*_{(k,r)} = \kappa_{(k,r)} w_{\kappa,(r)},
$$

which yields re-parameterization via projection

$$
\begin{aligned}
&U = U^* w_U^{-1}, \quad w_U = \text{QR.R}(U^*), \\
&\kappa_{(k,r)} = \frac{\kappa^*_{(k,r)}}{w_{\kappa,(r)}}, \quad w_{\kappa,(r)} = \sum_{k=1}^{V} \kappa^*_{(k,r)} \\
&\eta_{k,r} = \frac{u_{(k,r)}}{\kappa_{(k,r)} \sigma_u},
\end{aligned}
\tag{18}
$$

where QR.R denotes the function that outputs R matrix in QR decomposition. To control the amount of

relaxation, we assign w_U near I_d via $\pi(w_U) \propto \text{etr}\left[-\frac{(w_U-I_d)'(w_U-I_d)}{\lambda}\right]$ and $w_{\kappa,(r)}$ near 1 via $\pi(w_{\kappa,(r)}) \propto \exp\left[-\frac{(w_{\kappa,(r)}-1)^2}{\lambda}\right]$ and set $\lambda = 10^{-3}$.

For comparison, we test with the specified model (i) against (ii) the same model except with simple $u_{(k,r)} \sim \text{No}(0, \sigma_u^2)$ instead of the shrinkage prior and (iii) the same model except without the orthonormality constraint $U'U = I$ and the shrinkage prior. We run all models for $10,000$ iterations and discard the first $5,000$ iteration as burn-in. For each iteration, we run 300 leap-frog steps. For efficient computing, we truncated $d = 20$.

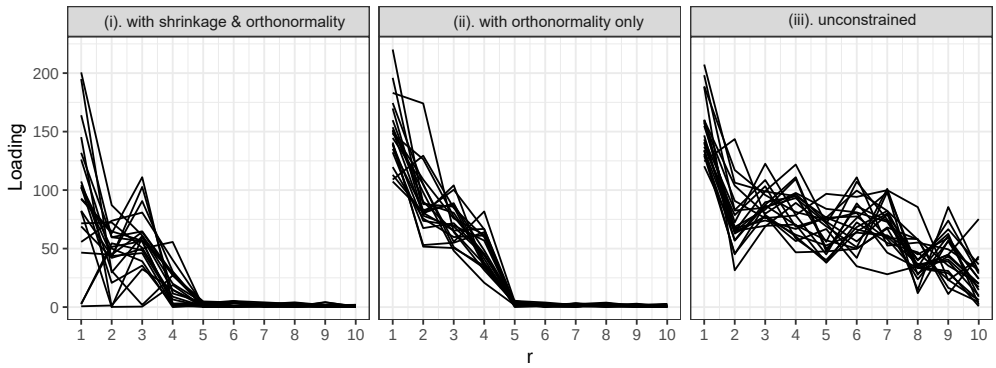
Table 2 lists the benchmark results. Compared to (i) and (ii), the unconstrained model (iii) suffers from very low effective sample size, due to the serious convergence issue in the factor matrix U. As explained by previous findings in matrix/tensor factorization (Hoff et al., 2016), the factor matrix could scale and rotate without changing the likelihood, and substantial improvement could be obtained by applying orthonormality constraint.

Figure 5(a) plots the posterior mean loadings $v_{(i,r)}$, with each line representing one subject. For all $i = 1, \ldots, 21$, the lines drop quickly to near 0 after $r \geq 5$ in model (i) and (ii), but only do so until $r \geq 10$ in model (iii). This indicates that independent factors are more effective representation of the span, compared to non-orthogonal ones. Clearly, (i) shows more variability than (ii) in the loading $v_{(i,r)}$. We validate these models by calculating area under the receiver operating characteristic curve (AUC) based on the mean predicted probability and the binary outcome $A_{(i,k,l)}$, using the fitted data and the other unused rescan data from the 21 subjects. The models (i) and (ii) with orthonomality constraint perform similarly well, and clearly better than the unconstrained model (iii) in prediction AUC.

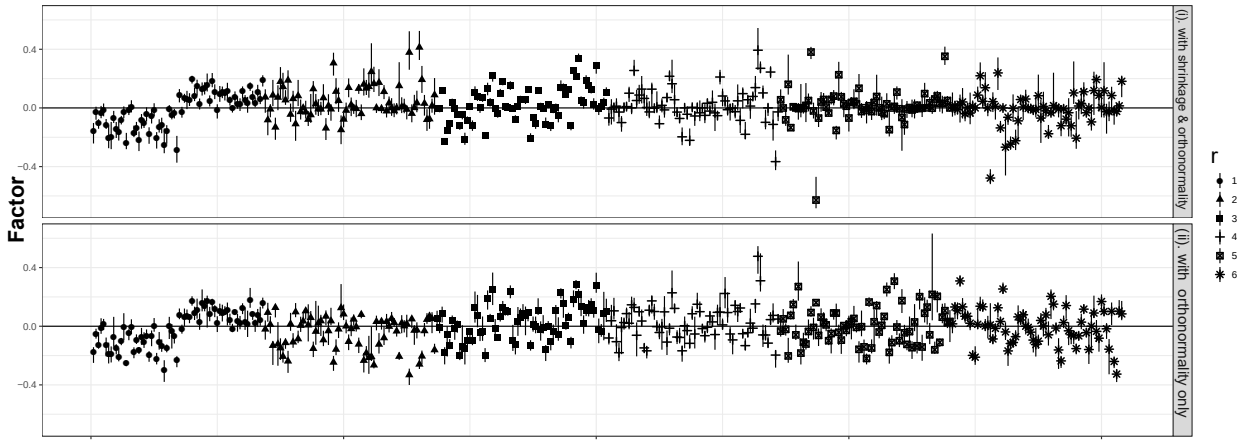
Model	(i).with shrinkage & orthonormality	(ii).with orthonormality only	(iii).unconstrained
Fitted AUC	97.9%	97.1%	96.9%
Prediction AUC	96.2%	96.2%	93.6%
ESS /1000 Iterations	193.72	188.10	8.15

Table 2: Comparing 3 models for 21 brain networks

Figure 5(b) compares the models (i) and (ii) over the top 6 frames of U_r, with r re-ordered such that $\sigma^2_{v,(1)} \geq \sigma^2_{v,(2)} \geq \ldots \geq \sigma^2_{v,(d)}$. The posterior of U_1, U_2, U_3 look very similar between the two, whereas U_4, U_5, U_6 have a considerable subset of points close to 0 in the model with shrinkage prior.



(a) Posterior mean of the loadings $v_{i,r}$ for 21 subjects using three models. Each line represents the loadings for one subject over $r = 1, \ldots 10$.



(b) Posterior mean and pointwise 95% credible interval of the factors $U_1, \ldots, U_6$ in the two constrained models.

Figure 5: Loadings and factors estimates of the network models. Panel (a) compares the varying loadings of the subjects in three models; Panel (b) compares the estimated shared factors with and without the shrinkage prior (model (iii) is omitted due to non-convergence in the factors).

7 Discussion

Parameter constraint often limits the flexibility to develop new model and creates huge burden in developing efficient posterior sampling algorithms. In this article, we develop a formal strategy to utilize the large pool of distributions in the constrained space, and propose a constraint relaxation approach to allow simple implementation for posterior estimation. For common constrained space that can be projected to via a function, we propose an exact algorithm based on data augmentation; for more general problem, we propose an approximation approach. This strategy works well for general equality and inequality constraints.

The future work of this research may include tackling the 'doubly intractable' problem. This issue is common when the data is on the constrained space, or the constrained prior has hyper-parameters to estimate. In the data application, we show that a reparameterization strategy works for some shrinkage priors, but clearly, more general treatment is needed. We expect our work to be compatible to the existing solutions (Murray et al., 2012; Rao et al., 2016; Stoehr et al., 2017).

A Proofs for Section 3.1

Proof. Proof of Lemma 1

Recall, that the distance function $v_{\mathcal{D}}(\theta)$ is chosen so that $v_{\mathcal{D}}(\theta)$ is zero for all $\theta \in \mathcal{D}$. It follows that for any function g

$$\begin{aligned}&\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\\&= \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta).\end{aligned} \tag{19}$$

Then,

$$\begin{aligned}&\left|E[g(\theta)|\theta\in\mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right|\\&= \left|\frac{\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)} - \frac{\int_{\mathcal{R}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{R}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp\big(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}\right|\\&= \left|\frac{\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}{\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\left[\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}\right.}\right|\end{aligned}$$

where the second equality follows from combining the fractions and making use of (3). We can bound the denominator from below by $\left[\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right]^2 > 0$ so that

$$\begin{aligned}&\left|E[g(\theta)|\theta\in\mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right|\\&\leq \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\right.}{C_{\mathcal{D}}^2}\end{aligned}$$

where $C_{\mathcal{D}} = \int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$. If we add and subtract

$$\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\cdot\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)$$

within the numerator, we can apply the triangle inequality. Thus,

$$\begin{aligned}&\left|E[g(\theta)|\theta\in\mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right|\\&\leq \frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\right|\cdot\left|\int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)\right.}{C_{\mathcal{D}}^2}\\&\quad+\frac{\left|\int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)\right|\cdot\left|\int_{\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\right.}{C_{\mathcal{D}}^2}\end{aligned}$$

Since $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}})$, we can then bound the numerators as follows. First,

$$\left| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta) \right| \cdot \left| \int_{\mathcal{D}} g(y_i)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu \right.$$

$$\leq \left| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta) \right| \cdot \left(\left| \int_{\mathcal{D}} g(\theta)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) \right| + \left| \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta) \right.$$

$$\leq \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta) \cdot \left(\int_{\mathcal{D}} |g(y_i)| \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/ \right.$$

$$\leq \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta) \cdot \int_{\mathcal{R}} |g(x_i)| \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) = C_{\mathcal{R}} E|g(x_i)| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_1$$

Here, $C_{\mathcal{R}} = \int_{\mathcal{R}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta)$ is the normalizing constant of $\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}$. Secondly,

$$\left| \int_{\mathcal{R}\setminus\mathcal{D}} g(\theta)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta) \right| \cdot \left| \int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) - \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta) \right.$$

$$\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) dx \cdot \left| \int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) \right| + \left| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta) \right.$$

$$\leq \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta) \left(\int_{\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) + \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) d\mu_{\mathcal{R}}(\theta) \right)$$

$$= \int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta).$$

Thus, we have the bounds specified by the theorem,

$$\left| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \right|$$

$$\leq \frac{C_{\mathcal{R}} E|g(\theta)| \int_{\mathcal{R}\setminus\mathcal{D}} \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2} + \frac{\int_{\mathcal{R}\setminus\mathcal{D}} |g(\theta)| \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2}$$

$$= \frac{\int_{\mathcal{R}\setminus\mathcal{D}} (C_{\mathcal{R}} E|g(\theta)| + |g(\theta)|) \mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda) d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2}.$$

It remains to be shown that

$$\left| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \right| \to 0 \text{ as } \lambda \to 0^+.$$

Again, by the assumptions that $g \in \mathbb{L}^1(\mathcal{R}, \mathcal{L}(\theta; Y)\pi_{\mathcal{R}} d\mu_{\mathcal{R}})$ and $v_{\mathcal{D}}(\theta) > 0$ for $\mu_{\mathcal{R}}$ a.e. $\theta \in \mathcal{R}\setminus\mathcal{D}$, it follows that $(C_{\mathcal{R}} E|g(x_i)| + |g(x_i)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta)$ is a dominating function of $(C_{\mathcal{R}} E|g(x_i)| + |g(x_i)|)\mathcal{L}(\theta; Y)\pi_{\mathcal{R}}(\theta) \exp(-v_{\mathcal{D}}(\theta)/\lambda)$ which converges to zero for $\mu_{\mathcal{R}}$-a.e. $\theta \in \mathcal{R} \setminus \mathcal{D}$ as $\lambda \to 0^+$. Thus, by the dominated convergence theorem, $\left| E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)] \right| \to 0$ as $\lambda \to 0^+$.

□

Proof. Proof of Theorem 1

We begin with the bound from Lemma 1.

$$\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \leq \frac{\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta)}{C_{\mathcal{D}}^2}.$$

For the moment, let us focus on the numerator of the previous expression. By the Cauchy-Schwartz inequality,

$$\begin{aligned}
&\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta) \\
&\leq \left(\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)^2\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2}\left(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2} \\
&\leq \left(\int_{\mathcal{R}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)^2\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2}\left(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2}
\end{aligned}$$

By assumption, $g \in \mathbb{L}^2(\mathcal{R}, \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}\mu_{\mathcal{R}})$. Thus,

$$\begin{aligned}
&\int_{\mathcal{R}\setminus\mathcal{D}}(C_{\mathcal{R}}E|g(\theta)| + |g(\theta)|)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)\exp(-v_{\mathcal{D}}(\theta)/\lambda)d\mu_{\mathcal{R}}(\theta) \\
&= \underbrace{\left([C_{\mathcal{R}}^3 + 2C_{\mathcal{R}}^2](E|g|)^2 + C_{\mathcal{R}}E[|g|^2]\right)^{1/2}}_{C_{\mathcal{R},g}<\infty}\left(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2} \\
&= C_{\mathcal{R},g}\left(\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2}
\end{aligned}$$

We separate the integral

$$\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$$

over the sets $\{\theta : v_{\mathcal{D}}(\theta) > -\lambda\log\lambda\}$ and $\{\theta : 0 < v_{\mathcal{D}}(\theta) < -\lambda\log\lambda\}$.

$$\begin{aligned}
&\int_{\mathcal{R}\setminus\mathcal{D}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \\
&= \int_{\{\theta:v_{\mathcal{D}}(\theta)>-\lambda\log\lambda\}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) + \int_{\{\theta:\,0<v_{\mathcal{D}}(\theta)<-\lambda\log\lambda\}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \\
&\leq \lambda^2\int_{\{\theta:v_{\mathcal{D}}(\theta)>-\lambda\log\lambda\}}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) + \int_{\{\theta:\,0<v_{\mathcal{D}}(\theta)<-\lambda\log\lambda\}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \\
&\leq C_{\mathcal{R}}\lambda^2 + \int_{\{\theta:\,0<v_{\mathcal{D}}(\theta)<-\lambda\log\lambda\}}\exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)
\end{aligned}$$

To review, to this point we have shown that

$$\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \leq \frac{C_{\mathcal{R},g}}{D_{\mathcal{D}}^2}\left(C_{\mathcal{R}}\lambda^2 + \int_{\{\theta:\, 0<v_{\mathcal{D}}(\theta)<-\lambda\log\lambda\}} \exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)\right)^{1/2} \tag{20}$$

From the requirements of Theorem 1, we now let $v_{\mathcal{D}}(\theta) = \inf_{x\in\mathcal{D}} \|\theta - x\|_2$ and assume that $\mathcal{D}$ has a piecewise smooth boundary. In this case, the set $\{\theta :\, 0 < v_{\mathcal{D}}(\theta) < -\lambda\log\lambda\}$ forms a 'shell' of thickness $-\lambda\log\lambda$ which encases $\mathcal{D}$.

For the moment, suppose that $\mathcal{D}$ is a bounded subset of $\mathcal{R}$. Furthermore, suppose we take λ sufficiently small so that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on $V_\lambda = \{\theta :\, 0 < v_{\mathcal{D}}(\theta) < -\lambda\log\lambda\}$. Observe that

$$\int_{\{\theta:\, 0<v_{\mathcal{D}}(\theta)<-\lambda\log\lambda\}} \exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) \leq \sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| \int_{V\lambda} \exp(-2v_{\mathcal{D}}(\theta)/\lambda)d\mu_R(\theta)$$

$$\leq \sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| \int_{V\lambda} d\mu_R(\theta) = \sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| \cdot Vol(V_\lambda)$$

$$= \sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| S_{\mathcal{D}} \cdot \lambda|\log\lambda|$$

Here,$S_{\mathcal{D}}$ is the surface area of boundary of $\mathcal{D}$, which is finite by the assumptions that $\mathcal{D}$ is bounded and has a piecewise smooth boundary. Additionally, since V_λ is relatively compact, it follows that $\sup_{V_\lambda}|\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)| < \infty$.

Consider the more general case where $\mathcal{D}$ is not a bounded subset of $\mathcal{R}$. Since $\int_{\mathcal{R}} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$, there exists a radius ρ such that $\int_{\|\theta\|_2>\rho} \mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) < \lambda^2$. Note that, for $\theta \in V_\lambda$, $J(v_{\mathcal{D}}(\theta)) = \sqrt{(Dv_{\mathcal{D}})'(Dv_{\mathcal{D}})} = 2$. By the co-area formula Diaconis et al. (2013); Federer (2014)

$$\int_{\{\theta:\, 0<v_{\mathcal{D}}(\theta)<-\lambda\log\lambda\}} \exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta) = \int_0^{-\lambda\log\lambda} e^{-\frac{x}{\lambda}}\left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)dx$$

Again, we may take λ sufficiently small so that $\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)$ is continuous on V_λ. As such, the function $\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)$ is a continuous map from the closed interval, $[0, -\lambda\log\lambda]$, to $\mathbb{R}$. Hence it is bounded. As a result,

$$\int_{\{\theta:\, 0<v_{\mathcal{D}}(\theta)<-\lambda\log\lambda\}} \exp(-2v_{\mathcal{D}}(\theta)/\lambda)\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\mu_{\mathcal{R}}(\theta)$$

$$\leq \sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right) \cdot \int_0^{-\lambda\log\lambda} e^{-\frac{x}{\lambda}}dx$$

$$= \sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_{\mathcal{D}}^{-1}(x)} \frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right) \cdot (\lambda - \lambda^2) = O(\lambda)$$

This result also applies to the case where $\mathcal{D}$ is bounded. Thus, we may conclude that

$$\begin{aligned}
&\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| \\
&\leq \frac{C_{\mathcal{R},g}}{D_{\mathcal{D}}^2}\left(C_{\mathcal{R}}\lambda^2 + \sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_{\mathcal{D}}^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)\cdot(\lambda-\lambda^2)\right)^{1/2} \\
&= \frac{C_{\mathcal{R},g}}{D_{\mathcal{D}}^2}\cdot\sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_{\mathcal{D}}^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)\cdot\sqrt{\lambda}+o(\sqrt{\lambda})
\end{aligned}$$

Since $\sup_{x\in[0,-\lambda\log\lambda]}\left(\int_{v_{\mathcal{D}}^{-1}(x)}\frac{1}{2}\mathcal{L}(\theta;Y)\pi_{\mathcal{R}}(\theta)d\bar{\mathcal{H}}^{r-1}(\theta)\right)$ is a decreasing function in λ, we may conclude that $\left|E[g(\theta)|\theta \in \mathcal{D}] - E_{\tilde{\pi}_\lambda}[g(\theta)]\right| = O(\sqrt{\lambda})$. □

B Proofs from Section 3.2

Proof. Recall that we have two densities. The first is the fully constrained density for $\theta \in \mathcal{D}$.

$$\pi_{\mathcal{D}}(\theta) = \frac{1}{m_0}\frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))}\mathbb{1}_{\mathcal{D}}(\theta)$$

where the normalizing constant m_0 is calculated w.r.t. Hausdorff measure

$$m_0 = \int_{\mathcal{R}}\frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))}\mathbb{1}_{\mathcal{D}}(\theta)d\bar{\mathcal{H}}^{r-s}(\theta).$$

Secondly, we have the relaxed distribution

$$\tilde{\pi}_{\mathcal{D}}(\theta) = \frac{1}{m_\lambda}\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{\|v(\theta)\|_1}{\lambda}\right)$$

where the normalizing constant is calculated w.r.t. Lebesgue measure on $\mathcal{R}$, denote by $\mu_{\mathcal{R}}$,

$$m_\lambda = \int_{\mathcal{R}}\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)\exp\left(-\frac{\|v(\theta)\|_1}{\lambda}\right)d\mu_{\mathcal{R}}(\theta).$$

For a given function, $g : \mathcal{R} \to \mathbb{R}$, we can define the exact and approximate expectations of g, respectively

E_Π and $E_{\tilde{\Pi}}$, as

$$E_\Pi[g(\theta)] = E[g(\theta)|\theta \in \mathcal{D}] = \int_{\mathcal{R}} \frac{g(\theta)}{m_0} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} \mathbb{1}_{\mathcal{D}}(\theta) d\bar{\mathcal{H}}^{r-s}(\theta)$$

$$= \int_{\mathcal{D}} \frac{g(\theta)}{m_0} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta)$$

$$E_{\tilde{\Pi}}[g(\theta)] = \int_{\mathcal{R}} \frac{g(\theta)}{m_\lambda} \mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta) \exp\left(- \frac{\|v(\theta)\|_1}{\lambda} \right) d\mu_{\mathcal{R}}(\theta)$$

$$= \int_{\mathbb{R}^s} \frac{1}{m_\lambda} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} \exp\left(- \frac{\|\nu(\theta)\|_1}{\lambda} \right) d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}$$

$$= \int_{\mathbb{R}^s} \frac{\exp\left(- \frac{\|x\|_1}{\lambda} \right)}{m_\lambda} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}$$

Let,

$$m(x) = m^{r-s}(x) = \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta).$$

By construction, $m(x) > 0$ for $\mu_{\mathbb{R}^s}$-a.e. $x \in Range(\nu)$. In particular, $m_0 = m(0) > 0$. By Theorem 1,

$$E[g(\theta)|\nu(\theta) = x] = \frac{1}{m(x)} \int_{\nu^{-1}(x)} g(\theta) \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta). \tag{21}$$

As such, we may may express $E_{\tilde{\Pi}}[g(\theta)]$ as

$$E_{\tilde{\Pi}}[g(\theta)] = \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\left(- \frac{\|x\|_1}{\lambda} \right) E\big[g(\theta)|\nu(\theta) = x\big] d\mu_{\mathbb{R}^s}(x). \tag{22}$$

Let us first consider the small λ behavior of m_λ. We begin by re-expressing m_λ in terms of $m(x)$ through the co-area formula.

$$m_\lambda = \int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta) \exp\left(- \frac{\|v(\theta)\|_1}{\lambda} \right) d\mu_{\mathcal{R}}(\theta)$$

$$= \int_{\mathbb{R}^s} \exp\left(- \frac{\|x\|_1}{\lambda} \right) \int_{\nu^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)\mathcal{L}(y;\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x)$$

$$= \int_{\mathbb{R}^s} m(x) \exp\left(- \frac{\|x\|_1}{\lambda} \right) d\mu_{\mathbb{R}^s}(x)$$

Split the above integral into two regions: the interior and exterior of $B_1(0; \lambda|\log(\lambda^{s+1})|)$. Note that

outside of B_1, $\exp(-\|x\|_1/\lambda) \leq \lambda^{s+1}$.

$$
\begin{aligned}
m_\lambda &= \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\bigg(- \frac{\|x\|_1}{\lambda}\bigg) d\mu_{\mathbb{R}^s}(x) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \exp\bigg(- \frac{\|x\|_1}{\lambda}\bigg) d\mu_{\mathbb{R}^s}(x) \\
&= O\bigg(\lambda^{s+1} \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) d\mu_{\mathbb{R}^s}(x)\bigg) \\
&\qquad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \bigg[1 + O\bigg(\frac{1}{\lambda} \exp\bigg(- \frac{\|x\|_1}{\lambda}\bigg)\bigg)\bigg] d\mu_{\mathbb{R}^s}(x) \\
&= O\bigg(\lambda^{s+1} \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) d\mu_{\mathbb{R}^s}(x)\bigg) \\
&\qquad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} m(x) \bigg[1 + O(\lambda^s)\bigg] d\mu_{\mathbb{R}^s}(x)
\end{aligned}
$$

Since $m(x)$ is continuous on an open neighborhood containing the origin, we may choose λ small enough so that $m(x)$ is uniformly continuous on $B_1(0;\lambda|\log\lambda^{s+1}|)$. Then,

$$
\begin{aligned}
m_\lambda &= O\bigg(\lambda^{s+1}\bigg) + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} [m(0) + o(1)][1 + O(\lambda^s)] d\mu_{\mathbb{R}^s}(x) \\
&= O(\lambda^{s+1}) + [m(0) + o(1)][1 + O(\lambda^s)] \underbrace{\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)}}_{Vol(B_1(0;\lambda|\log(\lambda^{s+1})|)} \\
&= m(0)\frac{|2(s+1)\lambda\log\lambda|^s}{\Gamma(s+1)} + o(|\lambda\log\lambda|^s)
\end{aligned}
$$

at leading order as $\lambda \to 0^+$.

We know turn to the small λ behavior of $\tilde{E}[g(\theta)]$. Again, we may choose λ sufficient small so that both

$$
\begin{aligned}
& m(x) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) \\
& G(x) = \int_{\nu^{(-1)}(x)} g(\theta) \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) = m(x) E[g|\nu(\theta) = x]
\end{aligned}
$$

are continuous on $B_1(0;\lambda|\log\lambda^{s+1}|)$ and hence uniformly continuous at $x = 0$.

Similar to the study of m_λ, separate the $\tilde{E}[g(\theta)]$ into integrals over the interior and exterior of $B_1(0, \lambda|\log(\lambda)^{s+1}|)$. Again, we assume λ is taken to be sufficiently small so that both $m(x)$ and $G(x)$ are uniformly continuous

on B_1. Then

$$
\begin{aligned}
E_{\tilde{\Pi}}[g(\theta)] &= \int_{\mathbb{R}^s} \frac{m(x)}{m_\lambda} \exp\left(- \frac{\|\nu(x)\|_1}{\lambda} \right) E[g(\theta)|\nu(\theta) = x] d\mu_{\mathbb{R}^s}(x) \\
&= \int_{\mathbb{R}^s} \frac{1}{m_\lambda} \exp\left(- \frac{\|\nu(x)\|_1}{\lambda} \right) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&= \int_{\mathbb{R}^s \setminus B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda} \exp\left(- \frac{\|x\|_1}{\lambda} \right) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&\quad + \int_{B_1(0;\lambda|\log(\lambda^{s+1})|)} \frac{1}{m_\lambda} \exp\left(- \frac{\|x\|_1}{\lambda} \right) \int_{\nu^{(-1)}(x)} \frac{\mathcal{L}(y;\theta)\pi_{\mathcal{R}}(\theta)}{J(\nu(\theta))} d\bar{\mathcal{H}}^{r-s}(\theta) d\mu_{\mathbb{R}^s}(x) \\
&= O\left(\frac{\lambda^{s+1}}{m_\lambda} \right) + \int_{B_1} \frac{m(0) + o(1)}{m_\lambda} (1 + O(\lambda^s)] \Big(E[g(\theta)|\nu(\theta) = 0] + o(1) \Big) d\mu_{\mathbb{R}^s}(x) \\
&= O\left(CE|g| \frac{\lambda}{|\log \lambda|^s} \right) + E[g(\theta)|\theta \in \mathcal{D}] + o(1).
\end{aligned}
$$

And we may conclude that

$$
\left| E[g|\theta \in \mathcal{D} - E_{\tilde{\Pi}}[g] \right| \to O \text{ as } \lambda \to 0^+.
$$

The proof of the corollary follows from changing the $o(1)$ correction within the integrals over $B_1(0;\lambda|\log \lambda^{s+1}|)$ with $O(\lambda|\log \lambda^{s+1}|)$ corrections. □

References

Beskos, A., N. Pillai, G. Roberts, J. M. Sanz-Serna, and A. Stuart (2013, 11). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli 19*(5A), 1501–1534.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*.

Betancourt, M., S. Byrne, and M. Girolami (2014). Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv:1411.6669*.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association 110*(512), 1479–1490.

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Byrne, S. and M. Girolami (2013). Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics 40*(4), 825–845.

Diaconis, P., S. Holmes, M. Shahshahani, et al. (2013). Sampling from a manifold. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, pp. 102–125. Institute of Mathematical Statistics.

Do Carmo, M. P. (2016). *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition*. Courier Dover Publications.

Dunson, D. B. and B. Neelon (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics 59*(2), 286–295.

Evans, L. C. and R. F. Gariepy (2015). *Measure theory and fine properties of functions*. CRC press.

Federer, H. (2014). *Geometric measure theory*. Springer.

Gelfand, A. E., A. F. Smith, and T.-M. Lee (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association 87*(418), 523–532.

Girolami, M. and B. Calderhead (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(2), 123–214.

Gunn, L. H. and D. B. Dunson (2005). A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics 6*(3), 434–449.

Hairer, E., C. Lubich, and G. Wanner (2006). *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer-Verlag.

Hoff, P. D. (2009). Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics 18*(2), 438–456.

Hoff, P. D. et al. (2016). Equivariant and scale-free tucker decomposition models. *Bayesian Analysis 11*(3), 627–648.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 50–67.

Khatri, C. and K. Mardia (1977). The von mises-fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 95–106.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM review 51*(3), 455–500.

Kolmogorov, A. N. (1950). Foundations of the theory of probability.

Lin, L. and D. B. Dunson (2014). Bayesian monotone regression using gaussian process projection. *Biometrika*.

Lin, L., B. St Thomas, H. Zhu, and D. B. Dunson (2016). Extrinsic local regression on manifold-valued data. *Journal of the American Statistical Association* (just-accepted).

Murray, I., Z. Ghahramani, and D. MacKay (2012). Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

Nash, J. (1954). C1 isometric imbeddings. *Annals of mathematics*, 383–396.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 2*, 113–162.

Nishimura, A., D. Dunson, and J. Lu (2017). Discontinuous hamiltonian monte carlo for sampling discrete parameters. *arXiv preprint arXiv:1705.08510*.

Polson, N. G. and J. G. Scott (2012). Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(2), 287–311.

Rao, V., L. Lin, and D. B. Dunson (2016). Data augmentation for models based on rejection sampling. *Biometrika 103*(2), 319–335.

Stoehr, J., A. Benson, and N. Friel (2017). Noisy hamiltonian monte carlo for doubly-intractable distributions. *arXiv preprint arXiv:1706.10096*.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.