Constraint Relaxation for Bayesian Modeling with Parameter Constraints

Leo Duan, Akihiko Nishimura, Alex Young, 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 Constrained Relaxation Methodology

2.1 Deriving Constrained Distribution via Conditioning

Let $\theta \in \mathcal{D}$ denote the parameters of interests. The support $\mathcal{D}$ is a constrained space. The usual Bayesian approach assigns an existing prior density $\pi_{\mathcal{D}}(\theta)$ for θ only having support $\mathcal{D}$, where the available choices are often quite limited. On the other hand, in the un/less-constrained space $\mathcal{R} \supset \mathcal{D}$, there is a large family of distributions with well-studied properties. We denote such a distribution by $\pi_{\mathcal{R}}(\theta)$. It would be appealing to adopt it in the constrained subspace. In this article, we restrict focus on $\mathcal{R}$ being the subspace of p-dimensional Euclidean space, $\mathcal{R} \subset \mathbb{R}^p$, and $\mu^s(A)$ the s-dimensional Lebesgue measure of set A.

Intuitively, one would want to directly apply space truncation to $\theta \in \mathcal{D}$ on $\pi_{\mathcal{R}}(\theta)$, renormalizing by $1/\mu^s(\mathcal{D})$ to obtain the appropriate density. However, this often does not work as many constrained space has $\mu^p(\mathcal{D}) = \int_{\mathcal{D}} \pi_{\mathcal{R}}(\theta)\mu^p(d\theta) = 0$. Therefore, an alternative strategy is needed.

First starting from a derived random variable $w = v(\theta)$, as long as $v : \mathcal{R} \to \mathbb{R}^d$ is measurable with respect to $\pi_{\mathcal{R}}$, one can obtain conditional density given $w = w_0$,

$$\pi_{\mathcal{R}}(\theta \mid w = w_0) = \frac{1}{m^s(w_0)J(v(\theta))}\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=w_0} \tag{1}$$

where $\mathbb{1}_E$ is an indicator equal 1 when condition E holds and 0 otherwise; $J(v(\theta))$ is the Jacobian equal to $\sqrt{\det[D(v(\theta))'D(v(\theta))]}$ with $D(v(\theta))$ the partial derivative matrix and *det* as the determinant; $m^s(w_0) = \int_{v(\theta)=w_0} \frac{1}{J(v(\theta))}\pi_{\mathcal{R}}(\theta)\mu^s(d\theta)$. Note $s \leq p$ is the intrinsic dimension of subspace $\{\theta : v(\theta) = w_0\}$ such that

$m^s(w_0) \in (0, \infty)$. Since the normalizing constant is nonzero, this conditional density is valid provided if such s exists and $J(v(\theta)) > 0$. This is based on the coarea formula in Federer (2014). A more rigorous justification is deferred to the theory section.

A large family of constraints can be associated with such function $v(\theta)$ fixed at certain value w_0, without loss of generality, we take $w_0 = \mathbf{0}$. We have $\mathcal{D} = \{\theta : v(\theta) = \mathbf{0}\}$. For example, one equality constraint $f(\theta) = 0$ can be associated with $v(\theta) = f(\theta)$=0; one inequality $f(\theta) < 0$ can be associated with $v(\theta) = |f(\theta)|_+ = \begin{cases} 0 \text{ if } f(x) \leq 0 \\ f(x) \text{ if } f(x) > 0 \end{cases}$ (assuming $f : \mathcal{R} \to \mathbb{R}$ measurable). Although this imposes a restriction on the suitable constrained space, there is a rich class within this category. Namely, many useful manifolds are simple embedding in $\mathbb{R}^p$ via equality constraint (Do Carmo, 2016).

Omitting normalization constant, the derived constrained density is:

$$\pi_{\mathcal{D}}(\theta) = \pi_{\mathcal{R}}(\theta \mid v(\theta) = \mathbf{0}) \propto \pi_{\mathcal{R}}(\theta) \mathbb{1}_{\theta \in \mathcal{D}} / J(v(\theta)). \tag{2}$$

Obviously, there are potentially more than one suitable $v(\theta)$'s for this derivation. Among those, if there exists $v(\theta)$ with constant $J(v(\theta))$ when $\theta \in \mathcal{D}$, it can be taken as the justified result of doing simple space truncation from $\mathcal{R}$ to $\mathcal{D}$.

To provide more intuition and illustrate the generality, we now derive two distributions on a $(p-1)$-hypersphere, defined as $\mathcal{D} = \{\theta \in \mathbb{R}^p : \theta'\theta = 1\}$. A simple choice of function is $v(\theta) = \theta'\theta - 1$, which has $J(v(\theta)) = \|2\theta\| = 2$ when $\theta \in \mathcal{D}$. We start with a familiar location-scale distribution Gaussian distribution with diagonal covariance $\theta \in \text{No}(F, I\sigma^2)$ as $\pi_{\mathcal{R}}$, where $F \in \mathcal{D}$ and I is the identity matrix. Conditioning on $v(\theta) = \theta'\theta - 1 = 0$ yields

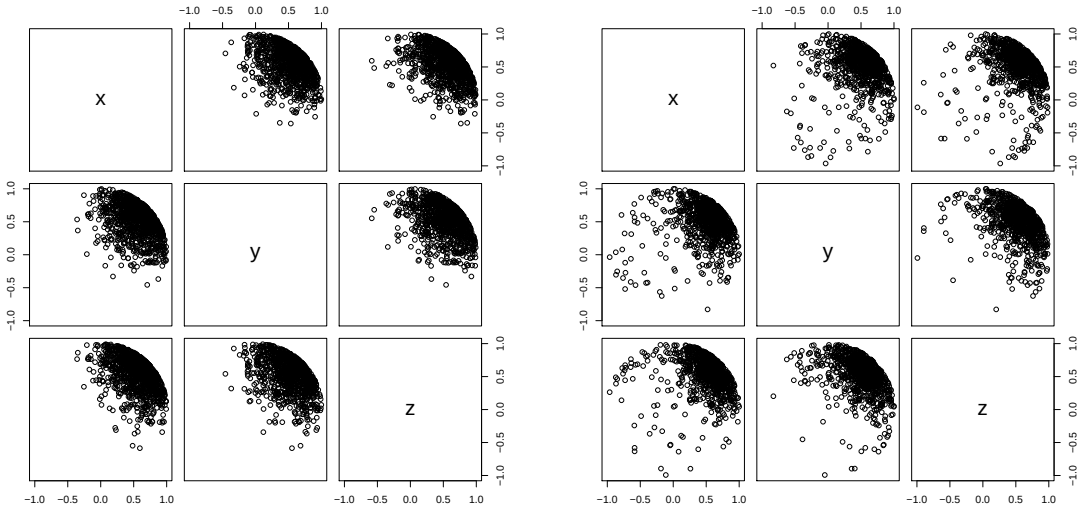
$$\begin{aligned} \pi_{\mathcal{D}}(\theta) &\propto \exp(-\frac{\|F-\theta\|^2}{2\sigma^2}) \mathbb{1}_{\theta'\theta=1}/2 \\ &\propto \exp(\frac{F'}{\sigma^2}\theta) \mathbb{1}_{\theta'\theta=1}, \end{aligned} \tag{3}$$

where the quadratic term $\theta'\theta$ is left out as a constant in the second line. This gives rise to the well-known von Mises–Fisher distribution (Khatri and Mardia, 1977). Although appears more like an exponential in the final form, the behavior of von Mises–Fisher on sphere can be 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 controled by $(2\sigma^2)^{-1}$; as the constrained density $\pi_{\mathcal{D}}(\theta)$ is proportional $\pi_{\mathcal{R}}(\theta)$ on $\mathcal{D}$, it behaves similarly. Figure 1(a) shows this distribution parameterized by $F = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]'$ and $\sigma^2 = 0.1$.

The observation on von Mises-Fisher immediately suggests we can induce different behavior by using a different unconstrained $\pi_{\mathcal{R}}(\theta)$. For example, to induce slower decay in density, one can 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 $I\sigma^2$, and obtain a new constrained density

$$\begin{aligned}\pi_{\mathcal{D}}(\theta) \propto & (1 + \frac{\|F-\theta\|^2}{m\sigma^2})^{-(m+p)/2} \mathbb{1}_{\theta'\theta=1}/2 \\ \propto & (1 - \frac{F'\theta}{1+m\sigma^2/2})^{-(m+p)/2} \mathbb{1}_{\theta'\theta=1}\end{aligned} \tag{4}$$

As in the t-distribution, density decays polynomially as $\|F-\theta\|^2$ increases, at small m, the induced distribution (Figure 1(b) with $m = 3, p = 3$) exhibits less concentration than von Mises–Fisher on the sphere. This can be useful for robust modeling. We will illustrate more general use by adapting shrinakge prior on similar space later in the application section.



(a) Constrained independent Gaussian distribution (b) Constrained independent t_3 distribution

Figure 1: 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 $F = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]'$

In these examples, the conditioning step may not appear explicit due to constant Jacobian, but we did use the result in the normalizing constant. As the sphere $\mathcal{D}$ has measure 0 in $\mathbb{R}^3$ based on either Gaussian or t distribution, the simple space truncation will yield an improper density. This is due to the subspace $\{\theta \in \mathbb{R}^3 : \theta'\theta = 1\}$ has an intrinsic dimension of 2, as can be revealed by its 2-dimensional coordinate system (θ_1^*, θ_2^*), $\theta_1 = \cos(\theta_1^*), \theta_2 = \sin(\theta_1^*)\cos(\theta_2^*), \theta_3 = \sin(\theta_1^*)\sin(\theta_2^*)$. Conditioning of $v(\theta) = \theta'\theta - 1$ on a

2-dimensional measure $m^2(\mathbf{0})$ yields the proper density.

2.2 Constraint Relaxation for Posterior Inference

The conditional distribution (2) allows one to adapt simple unconstrained $\pi_{\mathcal{R}}(\theta)$ into constrained space. When it is used as a prior, with the likelihood as $L(y;\theta)$ and y the data, one can obtain posterior:

$$\begin{aligned}\pi(\theta \mid y) &\propto L(y;\theta)\pi_{\mathcal{D}}(\theta) \\ &\propto L(y;\theta)\pi_{\mathcal{R}}(\theta)/J(v(\theta))\mathbb{1}_{\theta\in\mathcal{D}}.\end{aligned} \tag{5}$$

Clearly, the posterior also has support only inside $\mathcal{D}$ as well due to the inheritance of $\mathbb{1}_{\theta\in\mathcal{D}}$ from $\pi_{\mathcal{D}}(\theta)$. On the hand, the indicator is often inconvenient for posterior inference. We now develop two different strategies to relax the parameter $\theta \in \mathcal{D}$ into the neighborhood of $\mathcal{D}$, obtaining posterior θ^*. Then we directly use θ^* as approximation or project θ^* back to $\mathcal{D}$ for exact inference. We refer this strategy as **COnstraint RElaxation (CORE)**.

2.2.1 Approximation CORE

We first present a general approximation strategy. By putting positive mass around $v(\theta) = \mathbf{0}$, and let it decay exponentially, we relax the support of θ into a neighborhood surrounding $\mathcal{D}$. This generate approximate posterior sample $\theta^* \in \mathcal{R}$ via

$$\tilde{\pi}(\theta^* \mid y) = \frac{1}{m(\lambda)} L(y;\theta^*)\pi_{\mathcal{R}}(\theta^*)/J(v(\theta^*)) \exp(-\sum_{k=1}^{K} |v_k(\theta^*)|^{\alpha}/\lambda_k), \tag{6}$$

where v_k is the kth equation in $v(\theta)$, α is typically chosen as 1 or 2 to mimic Laplace or Gaussian kernel, $m(\lambda)$ is a normalizing constant such that $\int_{\mathcal{R}} \tilde{\pi}(\theta^* \mid y)d\theta^* = 1$; $\lambda_k \geq 0$ is a tuning parameter controling the concentration around $v(\theta^*) = \mathbf{0}$. When $\lambda_k = 0$ for all k, (6) becomes exact; using small $\lambda_k > 0$, (6) conventional Monte Carlo approach can be exploited directly in $\mathcal{R}$.

We use a toy example with closed-form posterior to illustrate the approximation. Consider a posterior from a sum-constrained bivariate Gaussian random vector $[\theta_1, \theta_2]' \mid y \sim \mathrm{No}(0, I)\mathbb{1}_{\theta_1+\theta_2-1=0}$. Using $v(\theta) = \theta_1 + \theta_2 - 1 = 0$, $J(v(\theta)) = \sqrt{2}$, the exact posterior (5) is proportional to

$$\phi(\theta_1)\phi(\theta_2)\mathbb{1}_{v(\theta)=0}$$

where $\phi(.)$ is the standard normal density. The exact posterior density has closed-form

$$\pi(\theta \mid y) = \frac{\sqrt{2}}{\sqrt{2\pi}} \exp(-\frac{(\theta_1 - \frac{1}{2})^2}{2/2})\mathbb{1}_{\theta_2=1-\theta_1}$$

corresponding to $\theta_1 \mid (\theta_1 + \theta_2 = 1) \sim \text{No}(1/2, 1/2)$, $\theta_2 \mid \theta_1 \sim \delta_{1-\theta_1}(.)$, where δ denotes a point mass. Marginally, it is a degenerate Gaussian:

$$\begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} \sim \text{No}_\text{d}\left(\begin{bmatrix}\frac{1}{2} \\ \frac{1}{2}\end{bmatrix}, \begin{bmatrix}\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2}\end{bmatrix}\right).$$

Using $\exp(-(\theta_1^* + \theta_2^* - 1)^2/\lambda)$ to replace $\mathbb{1}_{v(\theta)=0}$, we obtain approximation $\theta_1^* \sim \text{No}(\frac{2}{\lambda+4}, \frac{\lambda+2}{\lambda+4})$, $\theta_2^* \mid \theta_1^* \sim \text{No}(\frac{2}{\lambda+2}(1-\theta_1^*), \frac{\lambda}{\lambda+2})$. Marginally,

$$\begin{bmatrix}\theta_1^* \\ \theta_2^*\end{bmatrix} \sim \text{No}\left(\begin{bmatrix}\frac{2}{\lambda+4} \\ \frac{2}{\lambda+4}\end{bmatrix}, \begin{bmatrix}\frac{\lambda+2}{\lambda+4} & -\frac{2}{\lambda+4} \\ -\frac{2}{\lambda+4} & \frac{\lambda+2}{\lambda+4}\end{bmatrix}\right).$$

As approximation, one can sample from the non-degenerate distribution using small λ. Clearly, the approximation error decreases as λ gets smaller, and intuitively becomes exact when $\lambda \to 0$. We will formalize this notion in the theory section.

2.2.2 Data Augmentation CORE

The previous strategy is generally applicable for approximating the indicator function. In some less general but still common cases, one can relax the parameter θ to θ^* through some relaxing function g first, if there is an inverse function g^{-1} to project it back to $\mathcal{D}$, we can use $\theta = g^{-1}(\theta^*)$ to directly reparameterize $\pi_\mathcal{D}(\theta \mid y)$.

The relaxation $\theta^* = g(\theta; w)$ often requires an auxillary parameter w that controls the amount of relaxation. Usually w is independent of θ, but can be dependent on θ^*. Treating w as a latent variable from $\pi(w)$ with $\int_\mathcal{W} \pi(w)dw = 1$ and $\mathcal{W}$ as its support, one uses $g^{-1}(\theta^*; w)$ to reparameterize the exact posterior $\pi_\mathcal{D}(\theta \mid y)$. Standard variable transformation yields new parameterization

$$\begin{aligned}\pi(\theta^*, w \mid y) &= \pi_\mathcal{D}(g^{-1}(\theta^*; w) \mid y)\frac{1}{J(g(\theta; w))}\pi(w) \\ &\propto \frac{\pi_\mathcal{R}(g^{-1}(\theta^*; w) \mid y)}{J(v(\theta))\mid_{\theta=g^{-1}(\theta^*;w)}}\frac{1}{J(g(\theta; w))}\pi(w)\end{aligned} \tag{7}$$

where $J(g(\theta; w))$ is the Jacobian of g with respect to θ. The indicator function $\mathbb{1}_{\theta\in\mathcal{D}}$ disappears since $g^{-1}(\theta^*; w) \in \mathcal{D}$ always holds. It can be verified that transforming $\theta^* = g(\theta; w)$ yields $\pi_\mathcal{D}(\theta \mid y)\pi(w)$, which is the exact posterior augmented with latent variable w. Therefore, we refer this strategy as data augmentation constraint relaxation (DA-CORE). Using DA-CORE, one can directly sample θ^* and w in less constrained space, and apply $\theta = g^{-1}(\theta^*; w)$ in the end to obtain the θ-marginal.

One simple example of relaxation would be scaling of θ under norm constraint. For example, in $(p-1)$-simplex, $\|\theta\|_1 = \sum_{i=1}^p \theta_1 = 1$ with $\theta \in \mathcal{R} = [0, \infty]^p$, one can augment an independent latent variable

$w \in [0, \infty)$ have $\theta_i^* = \theta_i w$, and the inverse $\theta_i = \theta_i^*/w$ with $w = \sum_{i=1}^p \theta_i^*$. Suppose exact posterior is a Dirichlet distribution $\theta \mid y \sim \text{Dir}(\alpha)$,

$$\pi_{\mathcal{D}}(\theta \mid y) \propto \prod_{i=1}^{p} \theta_i^{\alpha-1} \mathbb{1}_{\sum_{i=1}^p \theta_i = 1},$$

then the relaxed parameterization is

$$\pi(\theta^*, w \mid y) = \pi(\theta^* \mid y) \propto \prod_{i=1}^{p-1} (\frac{\theta_i^*}{w})^{\alpha-1} w^{-(p-1)} \pi(w), \quad w = \sum_{i=1}^{p} \theta_i^*, \quad \theta^* \in (0, \infty)^p.$$

More generally, one can often choose the relaxation g based on the projection g^{-1} that produces $\theta \in \mathcal{D}$. To illustrate, we consider the random variable θ, a n-by-k matrix in the Stiefel manifold $\theta \in \mathcal{V}(n,k) = \{\theta \in \mathbb{R}^{n \times k} : \theta'\theta = I\}$, where $n \geq k$. The QR-decomposition produces such an orthonormal matrix, for which a n-by-k matrix $X = QR$, with $Q \in \mathcal{V}(n,k)$ and R k-by-k upper triangular and diagonal positive; the QR decomposition is unique if X is of rank k (Gulliksson and Wedin, 1992), which allows us to use it as g^{-1}. Letting w be a random matrix, k-by-k, upper triangular and diagonal positive, we have $\theta^* = \theta w$, with θ^* in rank k. Suppose θ is n-by-k and from the Matrix Bingham–von Mises–Fisher distribution (Hoff, 2009),

$$\pi_{\mathcal{D}}(\theta \mid y) \propto \text{etr}(C'\theta + B\theta' A\theta) \mathbb{1}_{\theta'\theta = I_k}$$

where etr represents the exponential of trace, A is symmetric n-by-n, B is symmetric k-by-k and C is n-by-k. The relaxed parameterization is

$$\pi(\theta^*, w \mid y) = \pi(\theta^* \mid y) \propto \text{etr}(C'\theta^* w^{-1} + B(\theta^* w^{-1})' A \theta^* w^{-1} \text{det}^{-1}(w), \quad \text{rank}(\theta^*) = k, \quad w = \text{QR.R}(\theta^*)$$

where QR.R denotes the function that outputs R matrix in QR decomposition.

In comparison, in other reparameterization such as coordinate system, one can only use equal or less parameters to satisfy the constraint, which can be constringent. In DA-CORE, it is generally more flexible with more parameters, thanks to the data augmentation. Since w is a redundant latent variable, one can assign $\pi(w)$ to allow greater relaxation, which is commonly associated with better performance in posterior sampling. More will be illustrated in the simulation.

2.3 Theoretic Properties

We now present the properties of the proposed approach. We first establish that the conditioning approach yields valid probability measure.

We focus on $\mathcal{R}$ being a p-dimensional Euclidean space and the intrinsic dimension of $\mathcal{D}$, $\dim(\mathcal{D}) = s \leq p$ is integer. Although the p-dimensional Lebesgue measure can have $\mu^p(\mathcal{D}) = 0$, it is still possible to define

a conditional probability given the event $v(\theta) = \mathbf{0}$. We utilize the concept of *regular conditional probability* (r.c.p.) (Kolmogorov, 1950). For this article to be self-contained, we list the definition as below (a more complete review can be found in Leao Jr et al. (2004)).

Let $(X, \mathscr{A}, \mu)$ be a probability space and $(Y, \mathscr{B})$ a measurable space. A function v is measurable if $v : X \to Y$, $v^{-1}(\mathscr{B}) \in \mathscr{A}$. A r.c.p is a function $f : Y \times \mathscr{A} \to [0, 1]$ satisfying:

1. $f(y, .)$ is a measure on $(X, \mathscr{A})$ for each $y \in Y$;
2. $f(., E)$ is a measurable function on $(Y, \mathscr{B})$ for each $E \in \mathscr{A}$;
3. For each $E \in \mathscr{A}$, $F \in \mathscr{B}$, $\mu(E \cap v^{-1}(F)) = \int_F f(y, E)\mu_y(dy)$, with μ_y the induced measure on $(Y, \mathscr{B})$.

Using the previous notation, we write $f(y, E) = P(\theta \in E \mid v(\theta) = y) = \int_E \pi_{\mathcal{R}}(\theta \mid v(\theta) = y)d\theta$

Remark 1. *Assuming $J(v(\theta)) > 0$ and there is a finite and non-negative integer s such that, for some $y \in Y$,*

$$m^s(y) = \int \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}\mu^s(d\theta) \in (0, \infty), \tag{8}$$

then

$$P(E \mid v(\theta) = y) = \begin{cases} \frac{1}{m_s(y)} \int_E \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}\mu^s(d\theta) & , \text{ if } m_s(y) \in (0, \infty) \\ \delta_{x^*}(E) \text{ with fixed } x^* \in \mathbb{R}^p & , \text{ if } m_s(y) \in \{0, \infty\} \end{cases} \tag{9}$$

is a valid r.c.p..

Proof. The first two crieria for r.c.p are trivally satisfied. We use the Hausdorff measure, the standard tool for geometric measure theory (Federer, 2014), defined as $\mathcal{H}^s(A) = \lim_{\delta \to 0} \inf\{\sum [\text{diam}(S_i)]^s : A \subseteq \cup S_i, \text{diam}(S_i) \leq \delta, \text{diam}(S_i) = \sup_{x,y \in S} \|x - y\|\}$. 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 is an integer, Lebesgue and normalized Hausdorff measures coincide $\mu^s(A) = \bar{\mathcal{H}}^s(A)$ (Evans and Gariepy, 2015).

Similar to the proof of (2) of Proposition 2 of Diaconis et al. (2013), using co-area formula (Federer, 2014):

$$\begin{aligned} \mu^p(E \cap v^{-1}(F)) &= \int \mathbb{1}_{\theta \in E}\mathbb{1}_{\theta \in v^{-1}(F)}\pi_{\mathcal{R}}(\theta)\mu^p(d\theta) \\ &= \int \left[\int_{v^{-1}(y)} \mathbb{1}_{\theta \in E}\mathbb{1}_{v(\theta) \in F} \frac{\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} \bar{\mathcal{H}}^s(d\theta) \right] \mu^{p-s}(dy) \\ &= \int_F \left[\int_E \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}\mu^s(d\theta) \right] \mu^{p-s}(dy) \end{aligned} \tag{10}$$

For $y \in \{y' : m(y') = 0\}$, $\int_E \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}\mu^s(d\theta) \leq \int_{\mathbb{R}^s} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}\mu^s(d\theta) = 0$; for $y \in \{y' : m(y') = \infty\}$, since $\mu^p(\mathbb{R}^p) = \int \mathbb{1}_{m(y)=\infty} m(y)dy + \int \mathbb{1}_{m(y)<\infty} m(y)dy = 1$, one must have $\int \mathbb{1}_{m(y)=\infty} dy = 0$. Combining parts yields

$$\begin{aligned}
\mu(E \cap v^{-1}(F)) &= \int_F \mathbb{1}_{m(y)\in(0,\infty)} \left[\int_E \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}\mu^s(d\theta)\right] \mu^{p-s}(dy) \\
&= \int_F \mathbb{1}_{m(y)\in(0,\infty)} \left[\int_E \frac{1}{m(y)}\frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))}\mu^s(d\theta)\right] m(y)\mu^{p-s}(dy) \qquad (11) \\
&= \int_F f(y, E)\mu_y(dy)
\end{aligned}$$

□

The above remark gives the definition of r.c.p given all $v(\theta) \in Y$. It is possible $m^s(y) = 0$ or ∞ for some $y \in Y$, but as long as $m^s(\mathbf{0}) \in (0, \infty)$ at certain integer s, we would have a valid constrained density

$$\pi_{\mathcal{D}}(\theta) = \frac{1}{m^s(\mathbf{0})}\frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=\mathbf{0}}}{J(v(\theta))} \qquad (12)$$

The dimension s is often referred as 'intrinsic' dimension of $\mathcal{D}$. Formally, one would use a standard concept in geometric measure theory named Hausdorff measure, (Federer, 2014). The d-dimensional Hausdorff measure is the limit total volume of the d-dimenional balls covering A, $\mathcal{H}^d(A) = \lim_{\delta\to 0} \inf\{\sum [\text{diam}(S_i)]^d : A \subseteq \cup S_i, \text{diam}(S_i) \leq \delta, \text{diam}(S_i) = \sup_{x,y\in S} \|x - y\|\}$. Then the intrinsic dimension is equal to Hausdorff dimension $s = \inf_{d\geq 0}\{H^d(\mathcal{D}) = 0\} = \sup_{d\geq 0}\{H^d(\mathcal{D}) = \infty\}$, at which the Hausdorff measure transitions from 0 to ∞. Although finding s can be challenging (Mardia, 1975; Bowen, 1979), one could could heuristically test $s \in \{p, p-1, \ldots, 0\}$ if s is known to be an integer. Fortunately, for posterior estimation, there is no need for estimating s or the normalizing constant $m^s(\mathbf{0})$ in Monte Carlo sampling.

We now quantify the approximation error of the approximation CORE (6). We use $\Pi(.)$ and $\tilde{\Pi}(.)$ to represent the measures under exact and approximating distributions. For easier notation, we re-parameterize the approximating part as $\exp(-\lambda^{-1}v(\theta))$ where $\lambda = \max_k \lambda_k$ and $v(\theta) = \sum_{k=1}^d \frac{|v_k(\theta)|^\alpha}{\lambda_k^*}$ with $\lambda_k^* = \lambda_k/\lambda$ and define a conditional expectation, $\mathbb{E}(g(\theta) \mid v(\theta) = x) = \int_{\mathbb{R}^s} \frac{g(\theta)\mathbb{1}_{v(\theta)=x}\pi_{\mathcal{R}}(\theta)}{Jv(\theta)}d\theta$.

We now assess the behavior of approximation error in terms of 1-Wasserstein distance, as λ decreases towards 0. The 1-Wasserstein distance $W_1(\Pi, \tilde{\Pi})$ represents the minimal amount of transport needed to transform one distribution to another. Formally, it is defined as

$$W_1(\Pi, \tilde{\Pi}) = \inf_{\gamma \in \Gamma(\Pi, \tilde{\Pi})} \int \|\theta - \theta^*\| d\gamma(\theta, \theta^*)$$

where $\Gamma(\Pi, \tilde{\Pi})$ is the family of all joint measures of with Π and $\tilde{\Pi}$ as the marginals, and $\|\theta - \theta^*\|$ is the Euclidean distance.

Remark 2. *The 1-Wasserstein distance between the measures based on* (5) *and* (6) *has*

$$\lim_{\lambda \to 0} W_1(\Pi, \tilde{\Pi}) = 0.$$

Further, for $\alpha = 1$ *in* (6),

$$W_1(\Pi, \tilde{\Pi}) \leq \lambda(\frac{k_1 k_2}{m(0)^2} + \frac{k_1}{m(0)}) + \exp(-\lambda^{-1} t)(\frac{k_1}{m(0)^2} + \frac{k_3}{m(0)}), \tag{13}$$

where $k_1 = \sup_{g:\|g\|_L \leq 1} \sup_{t^* \in [0,t)} \|\mathbb{E}(g(\theta^*) \mid v(\theta^*) = t^*)\|$, $k_2 = \sup_{t^* \in (0,t)} m(t^*)$ *and* $k_3 = \sup_{g:\|g\|_L \leq 1} \mathbb{E}(\|g(\theta)\|)$.

Proof. Let $g : \mathbb{R}^p \to \mathbb{R}$ be a 1-Lipschitz continuous function, i.e. $\|g(x) - g(y)\| \leq \|x - y\|$, denoted by $\|g\|_L \leq 1$. By Kantorovich-Rubinstein duality, the 1-Wasserstein distance based on Euclidean metric equals to:

$$W_1(\Pi, \tilde{\Pi}) = \sup_{g:\|g\|_L \leq 1} \int g(x) \Pi(dx) - \int g(y) \tilde{\Pi}(dy) \tag{14}$$

Taking $g(\theta) = \exp(-\lambda^{-1} v(\theta))$ in the co-area formula yields

$$\begin{aligned} m_\lambda &= \int_{\mathbb{R}} \left[\int_{v^{-1}(x)} \frac{\exp(-\lambda^{-1} v(\theta)) \pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \bar{\mathcal{H}}^{p-d}(d\theta) \right] \mathbb{1}_{x \geq 0} dx \\ &= \int_{\mathbb{R}} m(x) \exp(-\lambda^{-1} x) \mathbb{1}_{x \geq 0} dx. \end{aligned} \tag{15}$$

Taking $g(\theta) = \mathbb{1}_{v(\theta)=0}$ yields

$$m_0 = \int_{\mathbb{R}} \left[\int_{v^{-1}(y)} \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \bar{\mathcal{H}}^{p-d}(d\theta) \right] \mathbb{1}_{y=0} dy = \int_{v^{-1}(0)} \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \bar{\mathcal{H}}^{p-d}(d\theta) = m(0) \tag{16}$$

Clearly $m_\lambda \geq m_0$.

1. Asymptotic result:

We have

$$
\begin{aligned}
&\sup_{g:\|g\|_L\le 1}\int g(\theta)\left[\frac{\exp(-\lambda^{-1}v(\theta))}{m_\lambda}-\frac{\mathbb{1}_{v(\theta)=0}}{m_0}\right]\frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)}\bar{\mathcal{H}}^{p-d}(d\theta)\\
&=\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\mathbb{E}(g(\theta)\mid x)\left[\frac{\exp(-\lambda^{-1}x)\mathbb{1}_{x\ge 0}}{m_\lambda}-\frac{\mathbb{1}_{x=0}}{m_0}\right]dx\\
&=\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\mathbb{E}(g(\theta)\mid x)\left[\frac{1}{m_\lambda}-\frac{1}{m_0}\right]\mathbb{1}_{x=0}dx+\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\mathbb{E}(g(\theta)\mid x)\frac{\exp(-\lambda^{-1}x)}{m_\lambda}\mathbb{1}_{x>0}dx\\
&\le\sup_{g:\|g\|_L\le 1}\|\mathbb{E}(g(\theta)\mid 0)\|\left[\frac{1}{m_0}-\frac{1}{m_\lambda}\right]+\frac{1}{m_0}\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx
\end{aligned}
\tag{17}
$$

Note $m_\lambda\le\int_{\mathbb{R}}m(x)\mathbb{1}_{x\ge 0}dx=\int_{\mathbb{R}}\pi_{\mathcal{R}}(\theta)=1$. By dominated convergence theorem,

$$
\lim_{\lambda\to 0}m_\lambda=\int_{\mathbb{R}}m(x)\lim_{\lambda\to 0}\exp(-\lambda^{-1}x)\mathbb{1}_{x\ge 0}dx=m_0. \tag{18}
$$

Since $\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)dx\le\int_{\mathbb{R}}\sup_{g:\|g\|_L\le 1}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)dx$, letting $q_\lambda=\sup_{g:\|g\|_L\le 1}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}$, we have $0\le q_1-q_{\lambda_1}\le q_1-q_{\lambda_2}$ for $1\ge\lambda_1\ge\lambda_2$, by monotone convergence theorem, $\lim_{\lambda\to 0}\int[q_1(x)-q_\lambda(x)]dx=\int[q_1(x)-q_0(x)]dx$ hence $\lim_{\lambda\to 0}\int q_\lambda(x)dx=0$. Combining the results yields

$$
\lim_{\lambda\to 0}W_1(\Pi,\tilde{\Pi})=0. \tag{19}
$$

2. Non-asymptotic result:

$$
\begin{aligned}
\frac{1}{m_0}-\frac{1}{m_\lambda}&\le\frac{\int_{\mathbb{R}}m(x)\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx}{m_0^2}\\
&=\frac{1}{m_0^2}\left[\int_0^t m(x)\exp(-\lambda^{-1}x)dx+\int_t^\infty m(x)\exp(-\lambda^{-1}x)dx\right]\\
&\le\frac{1}{m_0^2}\left[\sup_{t^*\in(0,t)}m(t^*)\int_0^t\exp(-\lambda^{-1}x)dx+\exp(-\lambda^{-1}t)\int_t^\infty m(x)dx\right]\\
&\le\frac{1}{m_0^2}\left[\lambda\sup_{t^*\in(0,t)}m(t^*)+\exp(-\lambda^{-1}t)\right]
\end{aligned}
\tag{20}
$$

$$\sup_{g:\|g\|_L\leq 1}\int_{\mathbb{R}}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx$$

$$\leq \sup_{g:\|g\|_L\leq 1}\sup_{t^*\in(0,t)}\|\mathbb{E}(g(\theta)\mid t^*)\|\int_0^t\exp(-\lambda^{-1}x)dx+\exp(-\lambda^{-1}t)\sup_{g:\|g\|_L\leq 1}\int_t^\infty\|\mathbb{E}(g(\theta)\mid x)\|dx \tag{21}$$

$$\leq \sup_{g:\|g\|_L\leq 1}\sup_{t^*\in(0,t)}\|\mathbb{E}(g(\theta)\mid t^*)\|\lambda+\exp(-\lambda^{-1}t)\sup_{g:\|g\|_L\leq 1}\mathbb{E}(\|g(\theta)\|)$$

Combining (17)(20)(21), $k_1=\sup_{g:\|g\|_L\leq 1}\sup_{t^*\in[0,t)}\|\mathbb{E}(g(\theta)\mid t^*)\|$, $k_2=\sup_{g:\|g\|_L\leq 1}\mathbb{E}(\|g(\theta)\|)$, $k_3=\sup_{t^*\in(0,t)}m(t^*)$

$$\begin{aligned}&\sup_{g:\|g\|_L\leq 1}\int g(x)\Pi(dx)-\int g(x)\tilde{\Pi}(dx)\\&\leq\lambda(\frac{k_1k_3}{m_0^2}+\frac{k_1}{m_0})+\exp(-\lambda^{-1}t)(\frac{k_1}{m_0^2}+\frac{k_2}{m_0})\end{aligned} \tag{22}$$

□

The first part shows the asymptotic accuracy of the approximation. The second part shows the rate with non-asymptotic λ under mild assumptions. The interpretation for these assumptions is that if in a small space expansion of $\mathcal{D}$, defined as $\{\theta^* : v(\theta^*)\in[0,t]\}$, the marginal density of $v(\theta^*)$ and the conditional expectation of Lipschitz functions are bounded $k_1,k_2=\mathcal{O}(1)$, and the expected norm of Lipschitz function are smaller than a bound that grows near exponentially $k_3=\mathcal{O}(\lambda\exp(t/\lambda))$, then the distance $W_1(\Pi,\tilde{\Pi})$ converges to 0 in $\mathcal{O}(\lambda)$ as $\lambda\to 0$.

3 Posterior Computation

Adapting unconstrained density into space $\mathcal{D}$ often disrupts its posterior conjugacy. Since one can now sample the posterior in $\mathcal{R}$ using CORE, one can exploit conventional sampling 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.

3.1 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). For easy notation, we use q to represent θ^* under approximation-CORE (6), and $\{\theta^*, w\}$ under DA-CORE (7).

In order to sample q, 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 q. HMC then sample from the joint target density $\pi(q,p) = \pi(q)\pi(p) \propto \exp(-H(q,p))$ where, in the case of the posterior under relaxation,

$$\begin{aligned} &H(q,p) = U(q) + K(p), \\ &\text{where } U(q) = -\log \pi(q), \\ &\qquad K(p) = \frac{p' M^{-1} p}{2}. \end{aligned} \tag{23}$$

with $\pi(q)$ is the unnormalized density in (6) or (7).

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

$$\begin{aligned} \frac{\partial q^{(t)}}{\partial t} &= \frac{\partial H(q,p)}{\partial p} = M^{-1} p, \\ \frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(q,p)}{\partial q} = -\frac{\partial U(q)}{\partial q}. \end{aligned} \tag{24}$$

The exact solution to (24) is typically intractable but a valid Metropolis proposal can be generated by numerically approximating (24) with a reversible and volume-preserving integrator (Neal, 2011). The standard choice is the *leapfrog* integrator which approximates the evolution $(q^{(t)}, p^{(t)}) \to (q^{(t+\epsilon)}, p^{(t+\epsilon)})$ through the following update equations:

$$p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial q}, \quad q \leftarrow q + \epsilon M^{-1} p, \quad p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial q} \tag{25}$$

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

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

3.2 Computing Efficiency and Support Expansion

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. In this section, we provide some quantification of the effects and provide a practical guidance on choosing $\pi(w)$ or λ in the two strategies.

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. Let $\mathbf{H}_U(q)$ denote the hessian matrix of $U(q) = -\log \pi(q)$ and let $\xi_1(q)$ denotes the first largest eigenvalue of $\mathbf{H}_U(q)$. While analyzing stability and accuracy of an integrator is highly problem specific, 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). Besides the eigenvalue, if the support of q is a constrained space $\mathcal{Q}$, another limiting factor is roughly the shortest distance to the boundary $\eta(\theta; \mathcal{Q}) = \inf_{q' \notin \mathcal{Q}} \|q' - q\|$. If either $\eta(\theta; \mathcal{Q})$ or $\xi_1(\theta)^{-1/2}$ is close to 0, the upper bound would be too low to obtain efficient computation. In constrained model, the parameter space $\mathcal{D}$ can have very small $\eta(\theta; \mathcal{D})$. Constraint relaxation can reduce this problem via support expansion.

For approximation-CORE (6), to control approximation error, one can choose to relax a subset of constringent constraints. Observing $\mathcal{D} = \cap_k D_k$, each approximation $\exp(-\frac{|v_k(\theta^*)|^\alpha}{\lambda_k})$ corresponds to a constrained space $\mathcal{D}_k$. One practical strategy is that, for $\mathcal{D}_k$'s with $\eta(\theta; \mathcal{D}_k) \approx 0$, one uses moderate λ_k to induce some support expansion (denoted by $\lambda_k \geq \zeta$ with ζ moderately small but not too close to 0); for $\mathcal{D}_k$'s without this issue, one uses very small $\lambda_k \approx 0$ to almost always uphold the constraint. The latter was also suggested by Neal (2011) as creating a high 'energy wall'. Noting this could create inaccuracy of HMC near the boundary with $\lambda_k \approx 0$ under fixed step size, we use random step size ϵ at each iteration to reduce the error.

The Hessian $\mathbf{H}_U(q)$ under approximation-CORE is given by

$$\mathbf{H}_U(q) = -\mathbf{H}_{\log L(y;\theta^*)\pi_{\mathcal{R}}(\theta^*)/J(v(\theta^*))}(\theta^*) + \sum_k \lambda_k^{-1} \mathbf{H}_{|v_k|^\alpha}(\theta^*) \mathbb{1}_{\theta \notin \mathcal{D}_k}, \tag{26}$$

where the second term $\lambda_k^{-1} \mathbf{H}_{v_k}(\theta) \mathbb{1}_{\theta \notin \mathcal{D}_k}$ is 0 unless $\theta \notin \mathcal{D}_k$. Since the λ_k^{-1}'s in the second term often

dominate the eigenvalue, hence $\xi_1^{-1/2}(\theta^*) \approx \min_{\lambda_k \geq \zeta} \lambda_k^{1/2}$. A trade-off between approximation accuracy and computational efficiency is involved. Fortunately, as quantified above, the approximation error is often $\mathcal{O}(\max_{\lambda_k \geq \zeta} \lambda_k)$ and decreases faster than the efficiceny cap $\mathcal{O}(\min_{\lambda_k \geq \zeta} \lambda_k^{1/2})$, as λ_k decreases. In our experiments, we did find changing from $\lambda_k = 10^{-4}$ to 10^{-5} requires approximately 3 times of computing budget, due the effect on stability bound.

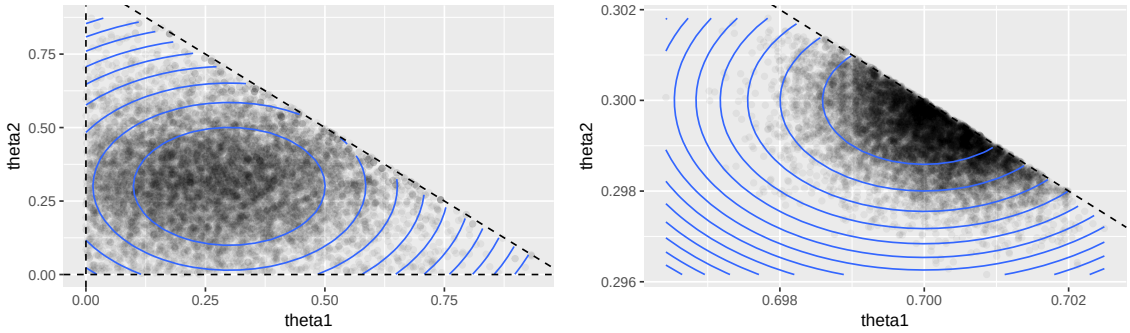
On the other hand, since DA-CORE (7) does not involve such error trade-off, it is preferred when it is applicable. Letting $\mathcal{Q}_\theta \subset \mathcal{D}$ denote the support for the constrained $\theta \in \mathcal{D}$, the reparameterization changes the support to $Q_{\theta^*} = \{g(\theta; w) : \theta \in \mathcal{Q}\}$. Therefore, one could choose $\pi(w)$ to substantially increase $\eta(\theta^*; Q_{\theta^*})$. Since the augmented variable w is redundant, DA-CORE can be considered as one type of parameter expansion discovered by Liu and Wu (1999), who originally focused on accelerating Gibbs sampling of probit regression. Although for greater space expansion, it is possible to use diffuse or even improper prior for $\pi(w)$ on Q_{θ^*}, we recommend assigning $\pi(w)$ loosely centered at $w_0 : g(\theta; w_0) = \theta$ (corresponding to when $\theta^* = \theta$), which makes θ^* a mild relaxation of θ. This ensures no substantial change in $\xi_1^{-1/2}(\theta^*)$ in HMC.

4 Simulations

In order to compare against existing approaches on computing efficiency and provide empirical evidence supporting our previous result, we run simulations on several toy examples in this section.

4.1 Gaussian under Linear Inequality

We first consider linear models under linear inequality constraints. Although recent work proposed a new customized prior with posterior conjugacy (Danaher et al., 2012), via our framework, one can simply exploit general Gaussian prior. We sample a bivariate Gaussian $\theta \sim \mathrm{No}\left(\mu, I\sigma^2\right)$ subject to linear inequality $\theta \in (0,1)^2, \theta_1 + \theta_2 < 1$, which forms a triangle. In two separate settings, choosing (μ, σ^2) as $([0.3, 0.3], 1/10)$ induces a wide-spread distribution centered in the interior of $\mathcal{D}$, while $([0.7, 0.3]', 1/10^4)$ induces a distribution concentrated on the boundary of $\mathcal{D}$. As DA-CORE does not appear straightforward in this case, we use the approximation-CORE (6) with $\exp(-\frac{|v(\theta)|}{\lambda})$, $v(\theta) = |\theta_1+\theta_2-1|_+ + |-\theta_1|_+ + |-\theta_2|_+ + |\theta_1-1|_+ + |\theta_2-1|_+$. Since the triangle has wide support with $\eta(\theta; \mathcal{D})$ away from 0, small $\lambda = 10^{-8}$ guarantees almost no approximation error. Figure 2 plots the posterior sample and its contour. Clearly, all posterior fall inside $\mathcal{D}$. To compare, we ran simple rejection sampling with untruncated normal proposal $\mathrm{No}(\mu, I\sigma^2)$. As expected, it suffers from a rapidly growing rejection rate from 12% to 51%, as μ moves further away from the center of $\mathcal{D}$.



(a) Constrained $\mathrm{No}([0.3, 0.3], 1/10)$ (b) Constrained $\mathrm{No}([0.7, 0.3], 1/10^4)$

Figure 2: Posterior sample of bivariate normal distribution subject to linear inequality constraints $\theta \in (0,1)^2, \theta_1 + \theta_2 < 1$, using HMC with constraint relaxation. Posterior is spread out around the center (panel (a)) or concentrated on the boundary (panel (b)) of the region.

4.2 von Mises–Fisher on Unit Circle

To illustrate equality constraint relaxation, we generate a simple von Mises–Fisher distribution $\pi_{\mathcal{D}}(\theta) \propto \exp(F'\theta)$ on a unit circle $\{(\theta_1, \theta_2) : \theta_1^2 + \theta_2^2 = 1\}$. We use $F = (5, 5)$ to induce a relatively spread-out θ on the manifold. For sampling, we compare three strategies: approximate-CORE using $\exp(-\frac{|\theta'\theta - 1|}{\lambda})$ for approximating the indicator, DA-CORE using $\theta_1 = \frac{\theta_1^*}{w}, w = \sqrt{(\theta_1^*)^2 + (\theta_2^*)^2}$ and $\pi(w) \sim \mathrm{No}(1, 1)\mathbb{1}_{w>0}$ and exact von Mises–Fisher obtained using 'movMF' package.

Unlike the previous linear inequality constraint, the unit circle has narrow $\eta(\theta; \mathcal{D}) = 0$ for all $\theta \in \mathcal{D}$, therefore, some support expansion is needed for HMC. We test $\lambda = 10^{-3}$, 10^{-4} and 10^{-5} for approximation-CORE. To compare the efficiency of HMC, we fix the number of leap-frog steps to 20 within one iteration HMC, and let software STAN automatically tune for stable step size. Table 1 shows the effective sample size per 1000 iterations, the effective 'violation' $|v(\theta)| = |\theta_1 + \theta_2 - 1|$ and the 1-Wasserstein distance W_1 as the approximation error. As W_1 is numerically computed, to provide a baseline error, we also calculate the average W_1 comparing two independent samples from the same exact distribution. The approximation error W_1 based on $\lambda = 10^{-5}$ approximation is indistinguishable from this low numerical error, while the other approximations have slightly larger error but more effective samples. As expected, the DA-CORE is exact and has high effective sample size.

	HMC based on CORE				Exact
	Approximation			DA-CORE	
	$\lambda = 10^{-3}$	$\lambda = 10^{-4}$	$\lambda = 10^{-5}$		
W_1	0.050 (0.019, 0.095)	0.034 (0.027, 0.037)	0.014 (0.013,0.025)	0.017 (0.0012,0.026)	0.015 (0.0014,0.025)
$\lvert v(\theta)\rvert \mid y$	9×10^{-4} $(2.6 \cdot 10^{-5}, 3.3 \cdot 10^{-3})$	9×10^{-5} $(2.0 \cdot 10^{-6}, 3.4 \cdot 10^{-4})$	9×10^{-6} $(2.7 \cdot 10^{-7}, 3.5 \cdot 10^{-5})$	0	0
ESS /1000 Iterations	751.48	260.54	57.10	788.30	

Table 1: Benchmark of constraint relaxation methods on sampling von–Mises Fisher distribution on a unit circle. For each approximation CORE, average approximation error (with 95% credible interval, out of 10 repeated experiments) is computed, and numeric error of W_1 is shown under column 'exact' as comparing two independent copies from the exact distribution. Effective sample size shows DA-CORE and approximation-CORE with relatively large λ have high computing efficiency.

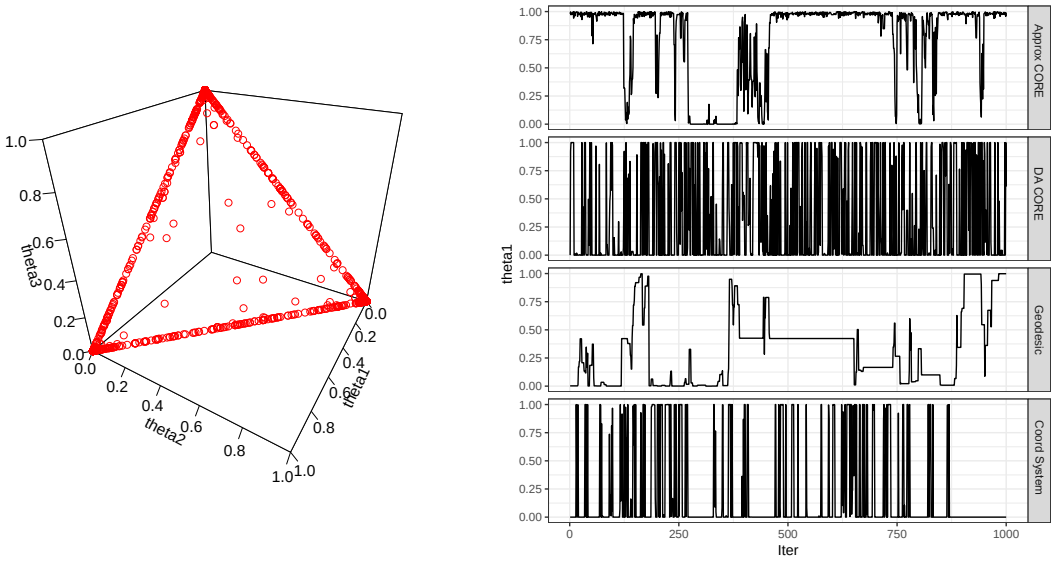
4.3 Dirichlet on a Simplex

Lastly, we experiment with a particularly challenging distribution on a $(p-1)$-simplex, defined by $\{\theta : \theta \in (0,\infty)^p, \sum_{i=1}^p \theta_i = 1\}$. We consider Dirichlet distribution $\mathrm{Dir}(\alpha)$, with $\pi_{\mathcal{D}}(\theta) \propto \prod_{i=1}^p \theta_i^{\alpha-1}$. When the concentration parameter $\alpha < 1$, $\mathrm{Dir}(\alpha)$ exhibits sparse property that some θ_i's become very close to 0, which is exploited in topic modeling (Wang and Blei, 2009) and shrinkage (Bhattacharya et al., 2015) literature. Despite the simple form, the computation can be quite difficult if there is large uncerntainty associated with θ on top of sparsity. The distribution will be multi-modal with distribution scattered along the boundary of the simplex (Figure 3(a)).

To illustrate, we consider $p = 3$ and various values of $\alpha \in \{1, 0.5, 0.1, 0.01\}$. We test the performance of approximation-CORE and DA-CORE. To compare, we also test the standard HMC using coordinate system $\theta_1 = \cos^2(\theta_1^*), \theta_2 = \sin^2(\theta_1^*)\cos^2(\theta_2^*), \theta_3 = \sin^2(\theta_1^*)\sin^2(\theta_2^*)$ for $\theta^* \in (0, 2\pi)^2$, which is equivalent to stick-breaking representation (Ishwaran and James, 2001); and the geodesic HMC utilizing the geometric flow directly on the simplex (Byrne and Girolami, 2013). For all HMCs, we fix the number of steps in each iteration to be 30 and tune the step size to have effective sample size as large as possible.

Table 2 lists the effective sample sizes under different α's. As α becomes smaller than 1, approximation-CORE and geodesic HMC become worse in performance, while DA-CORE and coordinate system are much less impacted. Figure 3(b) shows at $\alpha = 0.01$, the approximation-CORE and geodesic HMC are stuck for a long time, while DA-CORE works substantially better. As a well-tested reparameterization, HMC based on coordinate system still works acceptably well in this case.

The difficulty that approximation-CORE encountered was anticipated. Byrne and Girolami (2013) have previously reported similar slow-down of geodesic HMC computing on hyper-Dirichlet distribution (Hankin et al., 2010) with $\alpha < 1$. Comparing these two approaches, geodesic HMC relies on restricting the kinetic flow on $\mathcal{D}$ via its product with the metric tensor, and approximation-CORE relies on creating high energy wall in the potential energy. The latter can be viewed as an approximation to the former, which explains the similarity in performance.



(a) 2,000 samples from Dir(0.01) on 2-simplex.

(b) Traceplot of θ_1 using 4 types of HMCs.

Figure 3: Sampling of Dirichlet on an simplex with distribution concentrated on the boundaries. Panel(a) illustrates the distribution under Dir(0.01); Panel(b) compares the traceplots of 4 different types of HMCs, which are based on: approximation-CORE with $\lambda = 10^{-3}$, DA-CORE, geodesic flow on simplex (Byrne and Girolami, 2013) and coordinate system.

	HMC based on CORE			Geodesic HMC	Coord System HMC
	Approx $\lambda = 10^{-3}$	Approx $\lambda = 10^{-4}$	DA		
ESS /1000 Iter. ($\alpha = 1$)	511.43	146.07	947.53	174.14	**961.08**
ESS /1000 Iter. ($\alpha = 0.5$)	145.15	33.16	**912.94**	31.47	846.92
ESS /1000 Iter. ($\alpha = 0.1$)	88.32	26.88	**992.75**	28.70	875.83
ESS /1000 Iter. ($\alpha = 0.01$)	20.54	3.91	**722.44**	17.26	128.55

Table 2: Average effective sample size per 1000 iterations in Dir(α), under different α.

5 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 global-local shrinkage 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 (2).

$$A_{(i,k,l)} \sim \text{Bern}(\frac{1}{1+\exp(-\psi_{(i,k,l)} - z_{(k,l)})})$$

$$\psi_{(i,k,l)} = \sum_{r=1}^{d} v_{(i,r)} u_{(k,r)} u_{(l,r)}$$

$$U'U = I_d \text{ with } U = \{u_{(k,r)}\}_{k=1,\ldots,n;r=1,\ldots,d}$$

$$u_{(k,r)} = \eta_{(k,r)}\kappa_{(k,r)}\sigma_u$$

$$\eta_{(k,r)} \sim \text{Lap}(0,1), \quad \{\kappa_{(1,r)} \ldots \kappa_{(V,r)}\} \sim \text{Dir}(\alpha), \quad \sigma_u^2 \sim \text{IG}(2,1)$$

$$z_{(k,l)} \sim \text{No}(0, \sigma_z^2), \quad \sigma_z^2 \sim \text{IG}(2,1)$$

$$v_{(i,r)} \sim \text{No}(0, \sigma_{v,(r)}^2), \quad \sigma_{v,(r)}^2 \sim \text{IG}(2,1)$$

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 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 (2). 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

$$\begin{aligned} &U = U^* w_U^{-1}, \quad w_U = \text{QR.R}(U^*), \\ &\kappa_{(k,r)} = \frac{\kappa^*_{(k,r)}}{w_{\kappa,(r)}}, \quad w_{\kappa,(r)} = \sum_{k=1}^{V} \kappa^*_{(k,r)} \\ &\eta_{k,r} = \frac{u_{(k,r)}}{\kappa_{(k,r)} \sigma_u}, \end{aligned} \tag{27}$$

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 3 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.

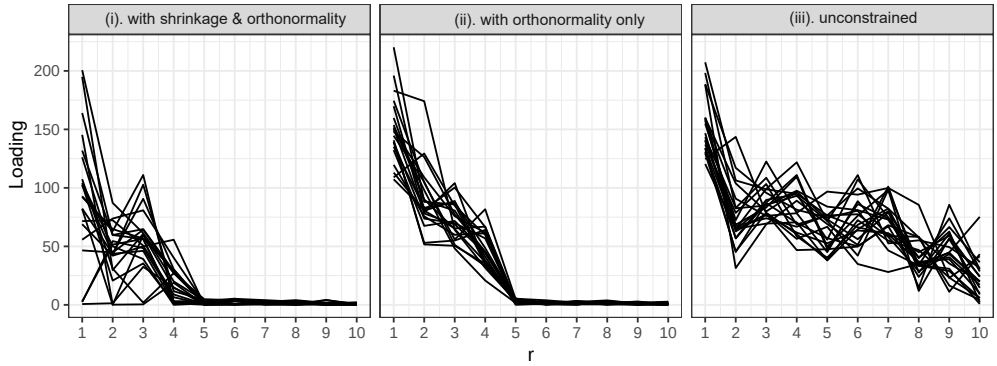
Figure 4(a) plots the posterior mean loadings $v_{(i,r)}$, with each line representing one subject. For all $i = 1, \ldots, 21$, the lines drop quickly to near 0 after $r \geq 5$ in model (i) and (ii), but only do so until $r \geq 10$ in model (iii). This indicates that independent factors are more effective 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.

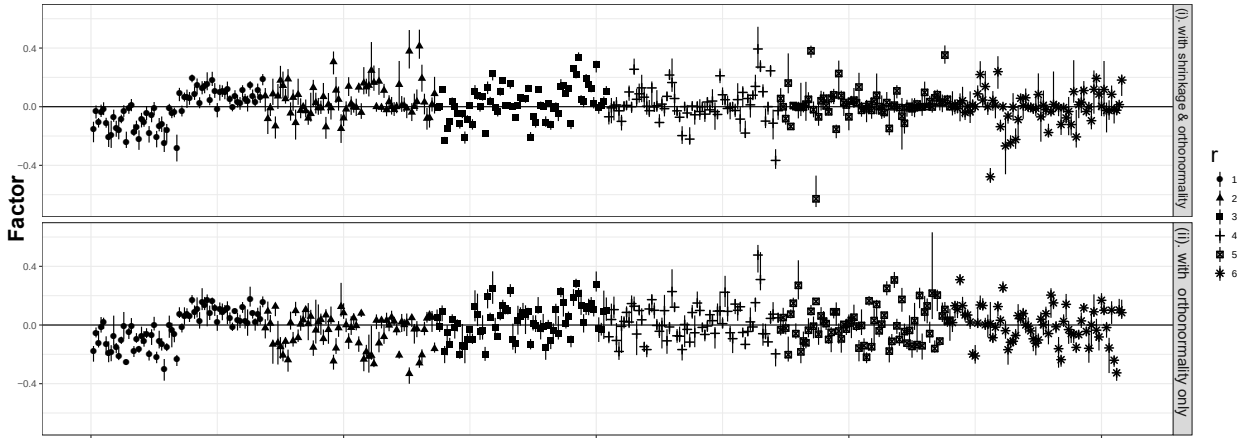
Model	(i).with shrinkage & orthonormality	(ii).with orthonormality only	(iii).unconstrained
Fitted AUC	97.9%	97.1%	96.9%
Prediction AUC	96.2%	96.2%	93.6%
ESS /1000 Iterations	193.72	188.10	8.15

Table 3: Comparing 3 models for 21 brain networks

Figure 4(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 4: 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).

6 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).

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.

Bowen, R. (1979). Hausdorff dimension of quasicircles. *Publ. math. IHES 50*(1), 11–25.

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.

Danaher, M. R., A. Roy, Z. Chen, S. L. Mumford, and E. F. Schisterman (2012). Minkowski–weyl priors for models with parameter constraints: an analysis of the biocycle study. *Journal of the American Statistical Association 107*(500), 1395–1409.

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.

Gulliksson, M. and P.-Å. Wedin (1992). Modifying the qr-decomposition to constrained and weighted linear least squares. *SIAM Journal on Matrix Analysis and Applications 13*(4), 1298–1313.

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.

Hankin, R. K. et al. (2010). A generalization of the dirichlet distribution. *Journal of Statistical Software 33*(11), 1–18.

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.

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.

Leao Jr, D., M. Fragoso, and P. Ruffino (2004). Regular conditional probability, disintegration of probability and radon spaces. *Proyecciones (Antofagasta) 23*(1), 15–29.

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).

Liu, J. S. and Y. N. Wu (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association 94*(448), 1264–1274.

Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 349–393.

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*.

Wang, C. and D. M. Blei (2009). Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *Advances in neural information processing systems*, pp. 1982–1989.

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