Extrinsic Priors for Bayesian Modeling with Parameter Constraints

Leo Duan, Akihiko Nishimura, David Dunson

Abstract: Prior information often takes for 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 contraints make it difficult to develop efficient posterior sampling algorithms. We propose a general solution, which relaxes the constraint through the use of an *extrinsic prior*, which is concentrated close to the constrained space. 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, our proposed approach frees us up to define new classes of hierarchical models for constrained problems. We illustrate this through application to a variety of simulated and real datasets.

KEY WORDS: Constraint relaxation; Euclidean Embedding; Monotone Dirichlet; Soft Constraint; Stiefel Manifold; Projected Markov chain

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.

Conceptually MCMC can be applied in a broad class of constrained parameter problems without complications (Gelfand et al., 1992). However, in practice, a primary difficulty is designing a Markov transition kernel that leads to an MCMC algorithm with sufficient computational efficiency to be practically useful. Common default transition kernels correspond to Gibbs sampling, random walk Metropolis-Hastings, and (more recently) Hamiltonian Monte Carlo (HMC). Gibbs sampling relies on alternately sampling from the full conditional posterior distributions for the different parameters, ideally in blocks to improve mixing. Gibbs requires the conditional distributions to be available in a form that is tractable to sample from directly, limiting consideration to specialized models. In constrained problems, block updating is typically either not possible or very inefficient (e.g. relying on rejection sampling with a high rejection probability), and one-at-a-time updating can lead to extremely slow mixing. Random walk algorithms provide an alternative, but each step of the random walk must maintain the parameter constraint. A common approach is to apply a normal random walk and simply reject proposals that violate the constraint, but this can have very high rejection rates even if using an adaptive approach that learns the covariance based on the history of the chain. An alternative is to rely on HMC. In simple settings in which a reparameterization can be applied to remove the constraint, HMC can be applied easily. Otherwise, HMC will generate proposals that violate the constraint, and hence face problems with high rejection rates in heavily constrained problems.

Due to the above hurdles, most of the focus in the literature has been on customized solutions developed for specific constraints. One popular strategy is to carefully pick a prior and likelihood such that posterior sampling is tractable. For example, for modeling of data on manifolds, it is typical to restrict attention to specific models, such as the Bingham-von Mises-Fisher distribution for Stiefel manifolds (Khatri and Mardia, 1977; Hoff, 2009). For data on the probability simplex, one instead relies on the Dirichlet distribution. An alternative is to reparameterize the model to eliminate or simplify the constraint. For example, when faced with a monotonicity constraint, one may reparameterize in terms of differences as the resulting positivity constraint leads to much easier sampling. In the literature on modeling of data on manifolds, there are two strategies: (i) *intrinsic* methods that define a statistical model directly on the manifold, and (ii) *extrinsic* methods that indirectly induce a model on the manifold through embedding the manifold in a Euclidean space, defining a model in the Euclidean space, and then projecting back onto the manifold. Essentially all of the current strategies for Bayesian modeling with constraints take an intrinsic-style approach. However, by strictly maintaining the constraint at all stages of the modeling and computation process, one limits the possibilities in terms of defining general methods to deal with parameter constraints.

These drawbacks motivate the development of *extrinsic* approaches that define an unconstrained model and/or computational algorithm, and then somehow adjust for the constraint. A related idea is Gelfand et al. (1992), who suggested running Gibbs sampling ignoring the constraint but only accepting the draws satisfying the constraint. Unfortunately, such an approach is highly inefficient, as motivated above. An alternative 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.

An alternative idea is to *relax* a sharp parameter constraint by defining a prior that has unrestricted support but places small probability outside of the constrained region. Neal (2011) suggested such an approach to apply HMC in settings involving a simple truncation constraint, while Pakman and Paninski (2014) applied a related idea to improve sampling from truncated multivariate normal distributions.

The goal of this article is to dramatically generalize these specific approaches to develop a broad class of *extrinsic priors* for parameter constrained problems. These priors are defined to place small probability outside of the constrained region, while permitting use of efficient and general use MCMC algorithms; in particular, HMC. Unlike intrinsic methods, such as Riemannian and geodesic HMC (Girolami and Calderhead, 2011; Byrne and Girolami, 2013), our approach is simple to implement in general settings using automatic algorithms. The generality frees up a much broader spectrum of Bayesian models, as one no longer needs to focus on very specific computationally tractable models. Theoretic studies are conducted and original models are shown in simulations and data applications.

2 Conditional Constrained Bayes 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 a prior density $\pi_{0,\mathcal{D}}(\theta)$ for θ only having support $\mathcal{D}$. A simple strategy is to reparameterize the constrained space in terms of parameters θ^* in a less constrained space (such as Euclidean space), with the constraint induced. The reparameterization $\theta \to \theta^*$ is often known as 'embedding' in manifold literature (Nash, 1954, 1956). Although this strategy often works, reparameterization makes it difficult to maintain certain property in $\mathcal{D}$, such as uniformity on a manifold as described by Diaconis et al. (2013); also, such a convenient reparameterization is not always available, especially when $\mathcal{D}$ is the intersection of two or more different types of constrained space.

On the other hand, in un/less-constrained space, there is a large family of distributions with well-studied properties. We present a new strategy to adapt them into the constrained space. Starting with a distribution of density $\pi_{\mathcal{R}}(\theta)$ on a encompassing space $\mathcal{R} \supset \mathcal{D}$, we focus on constrained space that can be defined as $\mathcal{D} = \{\theta : v(\theta) = \mathbf{0}\}$ with $v : \mathbb{R}^p \to \mathbb{R}^d$ Lipschitz and measurable with respect to $\pi_{\mathcal{R}}(\theta)$. Although this imposes a restriction, there is a rich class within this category. For each measurable function $f : \mathcal{R} \to \mathbb{R}$, one can use $v(\theta) = f(\theta)$ for equality constraint $f(\theta) = 0$, and $v(\theta) = |f(\theta)|_+ = \begin{cases} 0 \text{ if } f(x) \leq 0 \\ f(x) \text{ if } f(x) > 0 \end{cases}$ for inequality constraint $f(\theta) < 0$.

The constrained density can then be derived as the conditional density given the value of $v(\theta)$ equal to $\mathbf{0}$

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

where $\mathbb{1}(\theta \in \mathcal{D})$ is an indicator function equal 1 when $\theta \in \mathcal{D}$ and 0 otherwise; $J(v(\theta))$ is induced by the transformation, which equals to $\sqrt{\det D(v(\theta))' D(v(\theta))}$ with $D(v(\theta))$ the partial derivative matrix. This is based on the area formula in (Federer, 2014) and a more rigorous justification is deferred to the theory section.

We now provide some concrete illustration of this approach. Consider prescribing a distribution on a unit hypersphere $\{\theta : \theta'\theta = 1\}$ inside $\mathbb{R}^p$. We start with a familiar location-scale distribution Gaussian distribution with diagonal covariance $\theta \in \text{No}(F, I\sigma^2)$ as $\pi_{\mathcal{R}}$ $(F \in \mathbb{R}^p)$. Conditioning on $v(\theta) = \theta'\theta - 1 = 0$ yields

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

where the quadratic term of θ is left out as a constant under the constraint; $J(v(\theta)) = 2$. This gives rise to the famous von Mises–Fisher distribution (Khatri and Mardia, 1977). As the center and concentration is determined by the location F and scale-inverse σ^{-2} in Gaussian $\pi_{\mathcal{R}}(\theta)$, this pattern is inherited in the conditional (Figure 1(a)); in directional statistics, the normalized $F/\|F\|$ is called the mean direction (when $\|F\| \neq \mathbf{0}$) and $\|F\|/\sigma^2$ is called concentration parameter (Khatri and Mardia, 1977).

This immediately suggests one can choose different $\pi_{\mathcal{R}}(\theta)$ based on its properties in unconstrained space $\mathcal{R}$, and induce similar behavior on the $\mathcal{D}$. Starting from correlated Gaussian $\theta \in \text{No}(F, \Sigma)$, one obtains

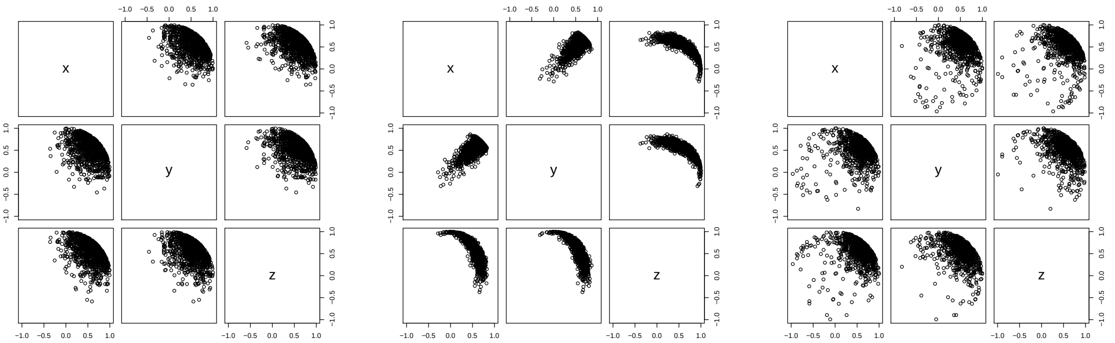
$$\begin{aligned} \pi_{\mathcal{D}}(\theta) &\propto \exp(-\frac{1}{2}\theta'\Sigma^{-1}\theta + F'\Sigma^{-1}\theta)\mathbb{1}_{\theta'\theta=1} \\ &\propto \exp(\sum_{k=2}^{p}(-\frac{1}{2}w_k)\theta'\Psi_k'\Psi_k\theta + (-\frac{1}{2}w_1)\theta'\Psi_1\Psi_1'\theta + + \sum_{k=1}^{p} w_k F'\Psi_k\Psi_k'\theta)\mathbb{1}_{\theta'\theta=1} \\ &\propto \exp(-\frac{1}{2}\sum_{k=2}^{p}(w_k - w_1)\theta'\Psi_k'\Psi_k\theta + \sum_{k=1}^{p} w_k F'\Psi_k\Psi_k'\theta)\mathbb{1}_{\theta'\theta=1} \end{aligned} \tag{3}$$

where Ψ_k and w_k are the kth eigenvector and eigenvalue of Σ^{-1}. This is the p-dimensional generalization of the Fisher-Bingham distribution (Mardia, 1975), which has $p = 3$. Similar to their roles in Gaussian covariance, the eigenvectors control axes of the ellipse on the sphere (Kent, 1982). An illustration with correlated x and y axes is shown in Figure 1(b).

Alternatively, one can start from a multivariate t-distribution $t_m(F, I\sigma^2)$ with m degrees of freedom, mean F and variance $I\sigma^2$, and obtain constrained density

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

Similar to the relation between t-distribution and Gaussian, at small m, the induced distribution (Figure 1(c) with $m = 3$) exhibits less concentration than von Mises–Fisher on the sphere.



(a) Constrained independent Gaussian distribution

(b) Constrained correlated Gaussian distribution

(c) 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) $\text{No}(F, \begin{bmatrix} 0.1 & 0.09 & 0 \\ 0.09 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix})$, (c) $t_3(F, \text{diag}\{0.1\})$, where $F = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]'$.

2.2 Constraint Relaxation for Posterior Inference

When the constrained distribution in last section is used as a prior, one can obtain posterior:

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

where $L(y;\theta)$ denotes the likelihood with y the data. As the normalizing term in prior is canceled, the posterior takes a simple form with the unconstrained density and $J(v(\theta))$. On the other hand, the posterior also has support only inside $\mathcal{D}$ due to the inheritance of $\mathbb{1}_{\theta\in\mathcal{D}}$ from $\pi_{\mathcal{D}}(\theta)$. Unfortunately, the indicator often poses a challenge for posterior inference.

We now present an approximation to the sharp indicator, relaxing support into a neighborhood surrounding $\mathcal{D}$,

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

where v_k is the kth equation in $v(\theta)$, $\lambda_k \geq 0$ is a tuning parameter that controls the amount of relaxation, α is typically chosen as 1 or 2, $m(\lambda)$ is an unknown normalizing constant such that $\int_{\mathcal{R}} \tilde{\pi}(\theta \mid y) d\theta = 1$. When $\theta \in \mathcal{D}$, (6) is proportional to (5); when $\theta \notin \mathcal{D}$, (6) has positve density, providing constraint relaxation. When $\lambda_k = 0$ for all k, (5) and (6) are equal.

To illustrate this relaxation, consider a posterior from a sum-constrained bivariate Gaussian random vector $[\theta_1, \theta_2]' \mid y \sim \text{No}(0, I)\mathbb{1}_{\theta_1+\theta_2-1=0}$. Using $v(\theta) = \theta_1 + \theta_2 - 1 = 0$ for the constraint, $J(v(\theta)) = \sqrt{2}$, (5) is proportional to

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

where $\phi(.)$ is the standard normal density. In this simple example, the exact posterior density has closed-form as

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

corresponding to $\theta_1 \mid (\theta_1 + \theta_2 = 1) \sim \text{No}(1/2, 1/2)$, $\theta_2 \mid \theta_1 \sim \delta_{1-\theta_1}(.)$. Equivalently, it follows a degenerate bivariate Gaussian distribution:

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

Using constraint relaxation $\exp(-(\theta_1 + \theta_2 - 1)^2/\lambda)$ 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,

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

Clearly the approximation density becomes exact when $\lambda \to 0$.

2.3 Reparameterization based on Constrain Relaxation

It is possible for some cases to use constraint relaxation for reparameterization of (5) instead of approximation. Although the application is more limited, it still covers a large range of useful constraints, such as simplex. The idea is to consider the constrained θ as the projection of unconstrained parameter θ^*. Coupled with an extra parameter w, one could form a bijective mapping between $\{\theta, w\}$ and $\{\theta^*, w\}$, and obtain variable transformation in the density. For example, θ can be a simple rescaling $\theta^*/\|\theta^*\|_1$ and $w = \|\theta^*\|_1$. Using the mapping, one can reparameterize the exact posterior (5) using less constrained θ^*.

To provide concrete explanation, we illustrate the rescaling case via a $(p-1)$-simplex example $\{\theta : \sum_{i=1}^p \theta_k = 1, \theta_i \in (0,1) \text{ for } i = 1,\dots,p\}$, using Dirichlet distribution

$$\pi_{\mathcal{D}}(\theta) \propto \prod_{i=1}^{p} \left(\theta_i^{\alpha-1} \mathbb{1}_{\theta_i \in (0,1)}\right) \mathbb{1}_{\sum_{i=1}^p \theta_k = 1}. \tag{7}$$

Relaxing the 1-norm constraint into space $\mathcal{R} = [0,1]^p$, one would have $\sum_{i=1}^p \theta_k^* = w$. A bijective mapping exists $\{\theta_1 = \theta_1^*/w, \theta_2 = \theta_2^*/w, \dots, \theta_{p-1} = \theta_{p-1}^*/w, w = w\}$ and substitute into the original density, one would obtain

$$(1 - \frac{\sum_{i=1}^{p-1} \theta_i^*}{w}) \prod_{i=1}^{p-1} (\frac{\theta_i^*}{w})^{\alpha-1} w^{-(p-1)} \tag{8}$$

where $w^{-(p-1)}$ is the Jacobian of transforming from $\{\theta_1, \dots, \theta_{p-1}, w\}$ to $\{\theta_1^*, \dots, \theta_{p-1}^*, w\}$.

avoid incurring approximation error at all by reparameterize the original constrained parameters as the projection of the parameters under constraint relaxation. This

For example, in the previous sum-constraint Gaussian example, one could

original θ in terms of the

2.4 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 $\mathcal{D}$, $\dim(\mathcal{D}) = d \le p$ and is integer. When $d < p$, the p-dimensional Lebesgue measure $\mu^p(\mathcal{D}) = 0$. To maintain the definition of conditional probability, we utilize the concept of *regular conditional probability* (r.c.p.) (Kolmogorov, 1950). For this article to be self-contained, we list the definition (Leao Jr et al., 2004) as below.

Let $(X, \mathscr{A}, \mu)$ be a probability space and $(Y, \mathscr{B})$ a measurable space. With a measurable function $v : X \to Y$, $v^{-1}(\mathscr{B}) \in \mathscr{A}$. 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 $J(v(\theta)) \neq 0$ and there is a finite and non-negtive integer s such that, for some $y \in Y$,*

$$m_s(y) = \int_{\mathbb{R}^s} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))} d\theta \in (0, \infty), \tag{9}$$

then

$$P(E \mid v(\theta) = y) = \begin{cases} \frac{1}{m_s(y)} \int_{E \cap \mathbb{R}^s} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))} d\theta & , \text{ if } m_s(y) \in (0, \infty) \\ \delta_{x^*}(E) \text{ with fixed } x^* \in \mathbb{R}^p & , \text{ if } m_s(y) \in \{0, \infty\} \end{cases} \tag{10}$$

is a valid r.c.p..

Proof. The first two crieria for r.c.p are trivally satisfied. Hausdorff measure is 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 $\mu(A) = \bar{\mathcal{H}}^s(A)$ (Evans and Gariepy, 2015).

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

$$\begin{aligned} \mu(E \cap v^{-1}(F)) &= \int_{\mathbb{R}^p} \mathbb{1}_{\theta \in E} \mathbb{1}_{\theta \in v^{-1}(F)} \pi_{\mathcal{R}}(\theta) d\theta \\ &= \int_{\mathbb{R}^{p-s}} \left[\int_{v^{-1}(y)} \mathbb{1}_{\theta \in E} \mathbb{1}_{v(\theta) \in F} \frac{\pi_{\mathcal{R}}(\theta)}{J(v(\theta))} \bar{\mathcal{H}}^s(d\theta) \right] dy \\ &= \int_{F \cap \mathbb{R}^{p-s}} \left[\int_{E \cap \mathbb{R}^s} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))} d\theta \right] dy \end{aligned} \tag{11}$$

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

$$\begin{aligned} \mu(E \cap v^{-1}(F)) &= \int_F \mathbb{1}_{m(y) \in (0,\infty)} \left[\int_E \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))} d\theta \right] dy \\ &= \int_F \mathbb{1}_{m(y) \in (0,\infty)} \left[\int_E \frac{1}{m(y)} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=y}}{J(v(\theta))} d\theta \right] m(y) dy \\ &= \int_F f(y, E) \mu_y(dy) \end{aligned} \tag{12}$$

□

The above remark gives the definition of r.c.p given all $v(\theta) \in Y$. Since our primary interest is when $v(\theta) = \mathbf{0}$, as long as $m_s(\mathbf{0}) \in (0, \infty)$ at certain integer s, we would have a valid conditional density $\pi_{\mathcal{D}}(\theta) = \frac{1}{m(\mathbf{0})} \frac{\pi_{\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=\mathbf{0}}}{J(v(\theta))}$.

The dimension s is often refered as 'intrinsic' dimension of $\mathcal{D}$. Formallly, one would use a standard concept in geometric measure theory named Hausdorff measure, (Federer, 2014). The d-dimensional Hausdorff measure is the 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 ∞. Finding s can be challenging (Mardia, 1975; Bowen, 1979). 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 approximation error of the constraint relaxing distribution. It is obvious that the approximating density becomes exact when $\lambda_k = 0$ for all k, we now assess the behavior of approximation error in terms of 1-Wasserstein distance, as λ_k 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

$$W_1(\Pi, \tilde{\Pi}) = \inf_{\gamma \in \Gamma(\Pi, \tilde{\Pi})} \int \|x - y\| d\gamma(x, y)$$

where $\Gamma(\Pi, \tilde{\Pi})$ is the family of all joint measures of the two samples with Π and $\tilde{\Pi}$ as the marginals. We use $\Pi(.)$ and $\tilde{\Pi}(.)$ to represent the measures under exact and approximating distributions. For easier notation, we first re-parameterize the approximating part as $\exp(-\lambda^{-1} v(\theta))$ where $\lambda = \max_k \lambda_k$ and $v(\theta) = \sum_{k=1}^d \frac{|v_k(\theta)|}{\lambda_k^*}$ with $\lambda_k^* = \lambda_k / \lambda$ and define a conditional expectation, $\mathbb{E}(g(\theta) \mid v(\theta) = x) = \int_{\mathbb{R}^s} \frac{g(\theta) \mathbb{1}_{v(\theta)=x} \pi_{\mathcal{R}}(\theta)}{J v(\theta)} d\theta$

Remark 2. *The 1-Wasserstein distance between the extrinsic and intrinsic distributions has*

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

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

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

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 $k_3 = \sup_{g: \|g\|_L \leq 1} \mathbb{E}(\|g(\theta)\|)$.

The first part shows the asymptotic accuracy of the approximation. The second part shows the rate under non-asymptotic λ. The interpretation is that if in a small space expansion based on $\mathcal{D}$, 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 expectation of absolute value of Lipschitz function are bounded by an exponentially increasing number $k_3 = \mathcal{O}(\lambda \exp(t/\lambda))$, then the distance $W_1(\Pi, \tilde{\Pi})$ converges to 0 in $\mathcal{O}(\lambda)$ as $\lambda \to 0$.

3 Posterior Computation

Conditioning the unconstrained density onto $\mathcal{D}$ often disrupts the posterior conjugacy. Fortunately, due to support expansion, the constraint relaxation allows us to sample the posterior directly inside $\mathcal{R}$. 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 sampling with relatively high dimension.

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).

In order to sample from $\theta \in \mathcal{R} \subset \mathbb{R}^p$, 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 (6),

$$\begin{aligned} H(\theta, p) &= U(\theta) + K(p), \\ \text{where } U(\theta) &= -\log\left\{ L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\exp(-\frac{|v_k(\theta)|^\alpha}{\lambda_k})/J(v(\theta)) \right\}, \\ K(p) &= \frac{p' M^{-1} p}{2}. \end{aligned} \tag{14}$$

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:

$$\begin{aligned} \frac{\partial \theta^{(t)}}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = M^{-1} p, \\ \frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \end{aligned} \tag{15}$$

The exact solution to (15) is typically intractable but a valid Metropolis proposal can be generated by numerically approximating (15) 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:

$$p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta}, \quad \theta \leftarrow \theta + \epsilon M^{-1} p, \quad p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta} \tag{16}$$

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

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

3.2 Support Expansion and Computing Efficiency

While an extrinsic distribution more closely approximate the constraint with a smaller λ, computational efficiency of HMC can be negatively impacted by choosing λ too small in certain condition. In this section, we explain and quantify this phenomenon and provide a practical guidance on how to pick a reasonable value of λ.

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 $\mathbf{H}_U(\theta)$ denote the hessian matrix of $U(\theta) = -\log \pi(\theta)$ and let $\omega_1(\theta)$ denotes the first largest eigenvalue of $\mathbf{H}_U(\theta)$. 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 $\epsilon < 2\omega_1(\theta)^{-1/2}$ must hold on most regions of the parameter space θ (Hairer et al., 2006). When θ is strictly constrained in certain region, $\theta \in \mathcal{D}_1^*$, another limiting factor is the shortest distance to the boundary $\eta(\theta; \mathcal{D}_1^*) = \inf_{\theta^* \not\in D_1^*} \|\theta^* - \theta\|$.

The Hessian $\mathbf{H}_U(\theta)$ is given by

$$\mathbf{H}_U(\theta) = -\mathbf{H}_{\log \pi_{\mathcal{R}}}(\theta) + \sum_k \lambda_k^{-1} \mathbf{H}_{v_k}(\theta) \mathbb{1}_{\theta \not\in \mathcal{D}_k}, \tag{17}$$

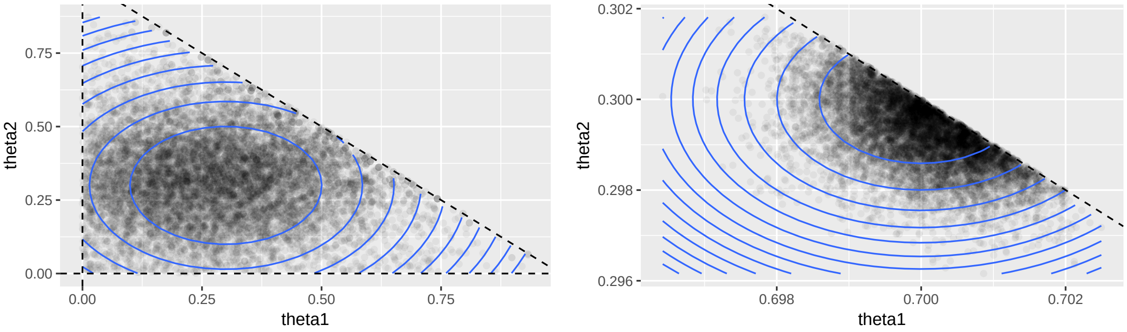
where in the first term $\pi_{\mathcal{R}}(\theta) = \pi_{0,\mathcal{R}}(\theta) L(\theta; y)/Jv(\theta)$ is defined on all $\mathcal{R}$, while in the second term $\lambda_k^{-1}\mathbf{H}_{v_k}(\theta)\mathbb{1}_{\theta\not\in\mathcal{D}_k}$ is 0 unless $\theta \not\in \mathcal{D}_k$. When $\theta \not\in \mathcal{D}_k$, the eigenvalue of (17) is commonly dominated by λ_k^{-1}.

The key for computational efficiency is to prevent the bound $2\omega_1(\theta)^{-1/2}$ and $\eta(\theta, \mathcal{D}_1^*)$ from being too close to 0. This can be achieved by strictly upholding certain constraints while relaxing more constrigent ones. Formally, recall $\mathcal{D} = \cap_{k=1}^d \mathcal{D}_k$, $\{\mathcal{D}_k\}$ can be into two sets, $\{\mathcal{D}_{(1)}, \mathcal{D}_{(2)}, \ldots, \mathcal{D}_{(m)}\}$ and $\{\mathcal{D}_{(m+1)}, \mathcal{D}_{(m+2)}, \ldots, \mathcal{D}_{(d)}\}$, such that for most region in $\mathcal{D}_1^* = \cap_{j=1}^m \mathcal{D}_{(j)}$, $\eta(\theta; \mathcal{D}_1^*)$ is away from 0, but $\mathcal{D}_1^* \cap D_{(j')}$ for any $j' = m+1, \ldots, d$ has $\eta(\theta; \mathcal{D}_1^* \cap D_{(j')}) \approx 0$. As $\lambda_{(j)}$ controls the amount of relaxation, one can use $\lambda_{(j)} \approx 0$ for $j = 1, \ldots, m$ to force $\theta \in \mathcal{D}_1^*$ for most of the time, while moderately small $\lambda_{(j')}$ for $j' = (m+1), \ldots, d$ to allow $\theta \not\in \mathcal{D}_{(j')}$ to happen.

As the result, the effective stability bound is affected by $\eta(\theta; \cap_{j=1}^m \mathcal{D}_{(j)})$ and $\Big(\min_{j' \in \{m+1,\ldots,d\}} \lambda_{(j')}^{1/2}\Big)$. Generally, often one can use very small λ_j to almost perfectly uphold inequality constraints, as they do not lead to small $\eta(.)$ in the first term . This is also used by Neal (2011) in creating a high energy 'wall'. Equality constraints need relaxation with moderate $\lambda_{j'}$ in the second term, as they commonly define hyper-plane that has $\eta(.) \approx 0$. To reduce inaccuracy near the boundary of $\cap_{j=1}^m \mathcal{D}_{(j)}$, we use random step size ϵ at each iteration.

For $\lambda_{j'}$ not very close to 0, a trade-off between approximation accuracy and computational efficiency is involved. Fortunately, the approximation error $\mathcal{O}(\max\limits_{j' \in \{m+1,\ldots,d\}} \lambda_{(j')})$ decreases faster than the efficiceny cap $\mathcal{O}(\min\limits_{j' \in \{m+1,\ldots,d\}} \lambda_{(j')}^{1/2})$. For example, empirically we found $\lambda_{j'} = 10^{-4}$ often yields a very low approximation error; reducing the error tolerance 10 times lower only requires approximately 3 times of computing budget.

To illustrate, we first run HMC on models with inequality constraints. We generate a bivariate Gaussian $\theta \sim \text{No}\left(\mu, I\sigma^2\right)$ subject to linear inequality $\theta \in (0,1)^2, \theta_1 + \theta_2 < 1$. We choose (μ, σ^2) as $([0.3, 0.3], 1/10)$ and $([0.7, 0.3]', 1/10^4)$ in two separate settings, inducing a wide-spread distribution centered in the interior of $\mathcal{D}$ and a concentrated distribution on the boundary of $\mathcal{D}$. For constraint relaxation, we use $\exp(-\frac{|v(\theta)|}{\lambda})$, with $v(\theta) = |\theta_1 + \theta_2 - 1|_+ + |-\theta_1|_+ + |-\theta_2|_+ + |\theta_1 - 1|_+ + |\theta_2 - 1|_+$. Using $\lambda = 10^{-8}$ ensures almost no approximation error, while the distance to boundary $\eta(\theta; \mathcal{D})$ is large enough for efficient sampling. Figure 2 plots the posterior sample and its contour. In comparison, simple rejection sampling with untruncated normal proposal $\text{No}(\mu, I\sigma^2)$ suffers from increasing rejection rate from 12% to 51%, as μ moves further away from the center of $\mathcal{D}$.



(a) Constrained $\text{No}([0.3, 0.3], 1/10)$

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

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.

To illustrate equality constraint relxation, we generate a simple von Mises–Fisher distribution $\pi_{\mathcal{D}}(\theta) \propto \exp(F'\theta)$ on a unit cirlce $\{(\theta_1, \theta_2) : \theta_1^2 + \theta_2^2 = 1\}$. We use $F = (1, 1)$ to induce a relatively spread-out θ on the manifold and $\exp(-\frac{|\theta'\theta - 1|}{\lambda})$ for constraint relxation. As $\eta_{\mathcal{D}} = 0$ on the circle, limiting the stability bound of leap–frog, some support expansion through moderate λ is needed. This involves a trade-off between accuracy and speed. We test $\lambda = 10^{-3}$, 10^{-4} and 10^{-5} in three experiments. Table 1 shows the computing time per $1,000$ effective sample size, the effective 'violation' $|v(\theta)| = |\theta_1 + \theta_2 - 1|$ and the 1-Wasserstein distance W_1 from its relxation approximate to the exact posterior, obtained using 'movMF' package. As W_1 is numerically computed, we calculate the average distance comparing two independent samples from the same exact distribution, using it as reference. The distance W_1 based on $\lambda = 10^{-5}$ approximation is indistinguishable from this low numerical error, while the other has slightly larger error using shorter computing time.

λ	10^{-3}	10^{-4}	10^{-5}	Exact
W_1	0.050 (0.019, 0.095)	0.034 (0.027, 0.037)	0.014 (0.013,0.025)	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
Avg. Time (sec/1000 eff. sample)	4.6	15.3	40.5	

Table 1: Average approximation error (with 95% credible interval, out of 10 repeated experiments) of sampling using constraint relaxtion for a von–Mises Fisher distribution on a unit circle. At 10^{-5}, the approximation is close to the limit of numeric error of numeric W_1. The needed computing time increases about 3 times when accuracy increases 10 times.

4 Simulations and Applications

4.1 Benchmark against Existing Approaches

We first benchmark the computing performance of the proposed constraint relaxation approach with existing state-of-art competitors. Bingham–von Mises–Fisher distribution is routinely used for such purpose, with $\pi_{\mathcal{D}}(\theta) \propto \exp(B'\theta + \theta' A\theta)$. We compare the performance with the Gibbs sampler (Hoff, 2009) and the geodesic HMC (Byrne and Girolami, 2013). The distribution is set to $A = \text{diag}(-1000, -600, -200, 200, 600, 1000)$ and $B = (100, 0, 0, 0, 0, 0)$.

We show that

4.2 Application: Sparse Bases Learning

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 21 healthy subjects without any history of neurological disease. Each subject has two brain network observations from scan–rescan, yielding a total of n = 42. Each observation is a $V \times V$ symmetric network A_i, recorded as adjacency matrix A_i for $i = 1, \ldots, n$. 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 betwen kth and lth region, $A_{i,k,l} = 0$ if there is no connection. The regions are constructed via the Desikan et al. (2006) atlas, for a total of V = 68 nodes.

The ambient dimension of observation is $V(V-1)/2 = 2,278$, which is significantly larger than sample size $n = 40$. They potentially contain observational error in recording connectivity, and the diagonal in each A_i is missing due to the lack of self-connectivity. These facts motivate a Bayesian low-rank approach. We consider a symmetric tensor decomposition model:

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

$$\psi_{i,k,l} = \sum_{r_1=1}^{d_1} \sum_{r_2=1}^{d_2} D_{r_1,r_2} W_{i,r_2} U_{k,r_1} U_{l,r_1}$$

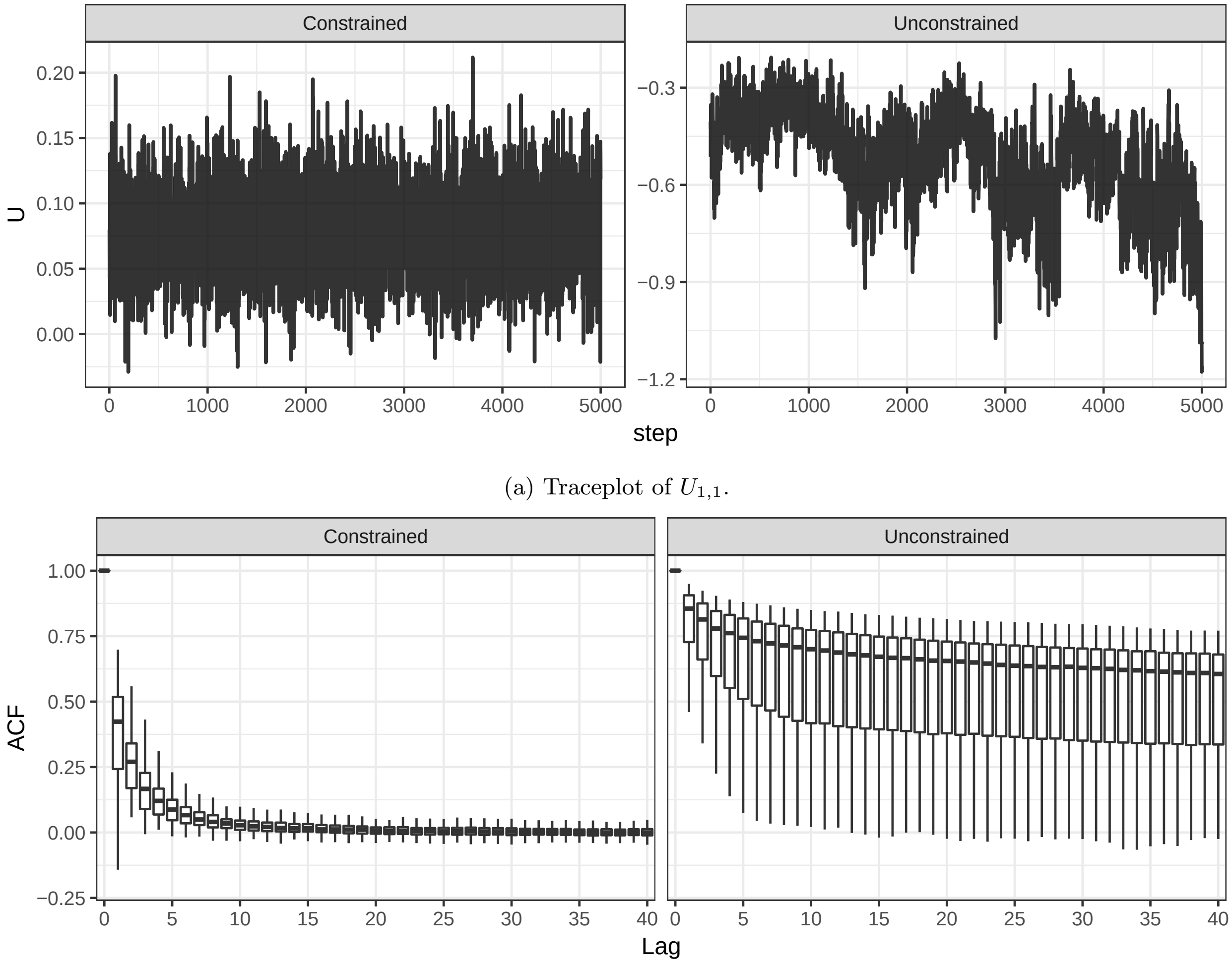
for $k > l$, $k = 2, \ldots, V$, $i = 1, \ldots, n$; U is $V \times d_1$ matrix, W is $n \times d_2$ matrix; D is a $d_1 \times d_2$ array. The $V \times V$ matrix Z is almost unstructural except symmetric $Z_{k,l} = Z_{l,k}$, which is commonly used to induce low-rank in the decomposition (Durante et al., 2016).

This model is a special Tucker decomposition with a sparse core tensor, whose diagonal plane is equal to D and 0 for other elements. The Tucker decomposition is more flexible than another routinely used decomposition, namely parallel factor analysis (PARAFAC). The PARAFAC assumes all ranks are equal and the core tensor D only has non-zero value when all its sub-indices are equal. In this case, PARAFAC would assume $d_1 = d_2$. The additional flexibility in the Tucker is appealing, as one could utilize the varying rank over different sub-direction (mode) of the tensor. On the other hand, a completely unconstrained Tucker decomposition is not identifiable in the matrices and core tensor, due scaling. For example, one can multiply a $d_1 \times d_1$ non-zero diagonal matrix R, to U and obtain $U^* = UR$ obtain $D^*_{.,r_2,.} = R^{-1} D_{.,r_2,.} R^{-1}$ for $r_2 = 1, \ldots, d_2$. This leaves the likelihood unchanged, creating identifiability issue.

Therefore, we consider applying some constraint on the Tucker decomposition. Motivated by high-order singular value decomposition, we impose orthonormality constraints $U'U = I_{d_1}$ and $W'W = I_{d_2}$. Hoff et al. (2016) previously obtained conjugated posterior for Tucker decomposition under orthonormality constraint, however, the symmetry in undirectedness of networks breaks the conjugacy.

We assign normal prior for $U_{k,r_2} \sim \text{No}(0, \phi_1)$, $W_{i,r_1} \sim \text{No}(0, \phi_2)$, $Z_{k,l} \sim \text{No}(0, \phi_3)$, $D_{r_1,r_2} \sim No(0, \phi_{4,r_1,r_2})$ for all i, k, l, r_1, r_2, and inverse-Gamma prior $\phi_1, \phi_2, \phi_3 \overset{indep}{\sim} \text{IG}(2,1)$, $\phi_{4,r_1,r_2} = \tau_{r_1}\tau_{r_2}$, with $\tau_{r_1}, \tau_{r_2} \overset{indep}{\sim}$ $\text{IG}(2,1)$ for all r_1, r_2. We

To allow estimation for model with orthonormality constraint, we use extrinsic prior with $\mathcal{K}(\theta) = \exp(-\frac{(U'U - I_{d_1})^2 + (W'W - I_{d_2})^2}{\lambda})$ and set $\lambda = 10^{-3}$. To compare, we also test with the same model configuration without the orthonormality constraint. We run both models for $10,000$ steps and discard the first $5,000$ steps. Figure 3 plots the traceplot and autocorrelation for matrix U. Unconstrained base model has severe convergence issue due to the non-identifiability, while constrained model converges and show low autocorrelation for all the parameters.



(a) Traceplot of $U_{1,1}$.

(b) ACF of all elements in U

Figure 3: Orthonormality constraint in the tensor decomposition modelallows convergence and rapid mixing on the factor matrix (left column); whereas unconstrained model does not converge due to free scaling. Traceplot for one parameter in factor matrix U and boxplot for autocorrelations of all parameters are shown.

5 Discussion

The estimation difficulty associated with parameter constraint often hinders the development of new models. Often one needed to carefully avoid models without conjugate posteriors, or skillfully re-parameterize the model for a more tractable algorithm. The extrinsic approach we introduced significantly reduces the burden. Through space expansion, it allows conventional toolbox such as HMC to be easily adopted to sample posterior without closed-forms. This allows researchers to impose constraints more freely in modeling and simplifies the way to incorporate constraint information about the functional of parameters.

We show the approximation error of the extrinsic approach can controlled via tuning parameter, with some trade-off between computing time and accuracy. A potentially more efficient strategy would be obtaining a rough approximate first in $\mathcal{R}$, then projecting into $\mathcal{D}$. Lin et al. (2016) developed algorithms similar to this idea and obtained consistency result for point estimation. A useful task would be to find an optimal projection also quantifying the uncertainty associated with finite sample. Lastly, the normalization of parameters over constrained space can sometime yield intractable integral, known as 'doubly stochastic' problem. We expect that the proposed extrinsic prior can be adapted and used together with the various existing solutions (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*.

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.

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.

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

Durante, D., D. B. Dunson, and J. T. Vogelstein (2016). Nonparametric bayes modeling of populations of networks. *Journal of the American Statistical Association* (In press).

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.

Kent, J. T. (1982). The fisher-bingham distribution on the sphere. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71–80.

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.

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).

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

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

Nash, J. (1956). The imbedding problem for riemannian manifolds. *Annals of mathematics*, 20–63.

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*.

Pakman, A. and L. Paninski (2014). Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics 23*(2), 518–542.

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*.

6 Appendix

Remark 1 proof:

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:

$$W_1(\Pi, \tilde{\Pi}) = \sup_{g:\|g\|_L \leq 1} \int g(x)\Pi(dx) - \int g(y)\tilde{\Pi}(dy) \tag{18}$$

Taking $g(\theta) = \exp(-\lambda^{-1}v(\theta))$ yields

$$\begin{aligned} m_\lambda &= \int_{\mathbb{R}} \left[\int_{v^{-1}(x)} \frac{\exp(-\lambda^{-1}v(\theta))\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) \right] \mathbb{1}_{x\geq 0} dx \\ &= \int_{\mathbb{R}} m(x) \exp(-\lambda^{-1}x) \mathbb{1}_{x\geq 0} dx. \end{aligned} \tag{19}$$

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

$$m_0 = \int_{\mathbb{R}} \left[\int_{v^{-1}(y)} \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) \right] \mathbb{1}_{y=0} dy = \int_{v^{-1}(0)} \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) = m(0) \tag{20}$$

Clearly $m_\lambda \geq m_0$.

1. Asymptotic result:

We have

$$\begin{aligned} &\sup_{g:\|g\|_L \leq 1} \int g(\theta) \left[\frac{\exp(-\lambda^{-1}v(\theta))}{m_\lambda} - \frac{\mathbb{1}_{v(\theta)=0}}{m_0} \right] \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) \\ &= \sup_{g:\|g\|_L \leq 1} \int_{\mathbb{R}} \mathbb{E}(g(\theta) \mid x) \left[\frac{\exp(-\lambda^{-1}x)\mathbb{1}_{x\geq 0}}{m_\lambda} - \frac{\mathbb{1}_{x=0}}{m_0} \right] dx \\ &= \sup_{g:\|g\|_L \leq 1} \int_{\mathbb{R}} \mathbb{E}(g(\theta) \mid x) \left[\frac{1}{m_\lambda} - \frac{1}{m_0} \right] \mathbb{1}_{x=0} dx + \sup_{g:\|g\|_L \leq 1} \int_{\mathbb{R}} \mathbb{E}(g(\theta) \mid x) \frac{\exp(-\lambda^{-1}x)}{m_\lambda} \mathbb{1}_{x>0} dx \\ &\leq \sup_{g:\|g\|_L \leq 1} \|\mathbb{E}(g(\theta) \mid 0)\| \left[\frac{1}{m_0} - \frac{1}{m_\lambda} \right] + \frac{1}{m_0} \sup_{g:\|g\|_L \leq 1} \int_{\mathbb{R}} \|\mathbb{E}(g(\theta) \mid x)\| \exp(-\lambda^{-1}x) \mathbb{1}_{x>0} dx \end{aligned} \tag{21}$$

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,

$$\lim_{\lambda\to 0} m_\lambda = \int_{\mathbb{R}} m(x) \lim_{\lambda\to 0}\exp(-\lambda^{-1}x)\mathbb{1}_{x\ge 0}dx = m_0. \tag{22}$$

Since $\sup\limits_{g:\|g\|_L\le 1} \int_{\mathbb{R}} \|\mathbb{E}(g(\theta) \mid x)\|\exp(-\lambda^{-1}x)dx \le \int_{\mathbb{R}} \sup\limits_{g:\|g\|_L\le 1} \|\mathbb{E}(g(\theta) \mid x)\|\exp(-\lambda^{-1}x)dx$, letting $q_\lambda = \sup\limits_{g:\|g\|_L\le 1} \|\mathbb{E}(g(\theta) \mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}$, 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

$$\lim_{\lambda\to 0} W_1(\Pi, \tilde{\Pi}) = 0. \tag{23}$$

2. Non-asympotic result:

$$\begin{aligned}\frac{1}{m_0} - \frac{1}{m_\lambda} &\le \frac{\int_{\mathbb{R}} m(x)\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx}{m_0^2} \\ &= \frac{1}{m_0^2}\left[\int_0^t m(x)\exp(-\lambda^{-1}x)dx + \int_t^\infty m(x)\exp(-\lambda^{-1}x)dx\right] \\ &\le \frac{1}{m_0^2}\left[\sup_{t^*\in(0,t)} m(t^*)\int_0^t \exp(-\lambda^{-1}x)dx + \exp(-\lambda^{-1}t)\int_t^\infty m(x)dx\right] \\ &\le \frac{1}{m_0^2}\left[\lambda \sup_{t^*\in(0,t)} m(t^*) + \exp(-\lambda^{-1}t)\right]\end{aligned} \tag{24}$$

$$\begin{aligned}&\sup_{g:\|g\|_L\le 1}\int_{\mathbb{R}} \|\mathbb{E}(g(\theta) \mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx \\ &\le \sup_{g:\|g\|_L\le 1}\sup_{t^*\in(0,t)} \|\mathbb{E}(g(\theta) \mid t^*)\| \int_0^t \exp(-\lambda^{-1}x)dx + \exp(-\lambda^{-1}t)\sup_{g:\|g\|_L\le 1}\int_t^\infty \|\mathbb{E}(g(\theta) \mid x)\|dx \\ &\le \sup_{g:\|g\|_L\le 1}\sup_{t^*\in(0,t)} \|\mathbb{E}(g(\theta) \mid t^*)\|\lambda + \exp(-\lambda^{-1}t)\sup_{g:\|g\|_L\le 1}\mathbb{E}(\|g(\theta)\|)\end{aligned} \tag{25}$$

Combining (21)(24)(25), $k_1 = \sup\limits_{g:\|g\|_L\le 1}\sup\limits_{t^*\in[0,t)} \|\mathbb{E}(g(\theta) \mid t^*)\|$, $k_2 = \sup\limits_{g:\|g\|_L\le 1}\mathbb{E}(\|g(\theta)\|)$, $k_3 = \sup\limits_{t^*\in(0,t)} m(t^*)$)

$$\begin{aligned}&\sup_{g:\|g\|_L\le 1}\int g(x)\Pi(dx) - \int g(x)\tilde{\Pi}(dx) \\ &\le \lambda(\frac{k_1k_3}{m_0^2} + \frac{k_1}{m_0}) + \exp(-\lambda^{-1}t)(\frac{k_1}{m_0^2} + \frac{k_2}{m_0})\end{aligned} \tag{26}$$

□