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 Extrinsic Bayes Methodology

2.1 Intrinsic Distribution Based on Conditional Probability

Let $\theta \in \mathcal{D}$ denote the parameters in likelihood function $L(\theta; y)$, with y the data. The support $\mathcal{D}$ is a constrained space. The usual Bayesian approach assigns a prior density $\pi_{0,\mathcal{D}}(\theta)$ for θ having support $\mathcal{D}$. A common strategy is to reparameterize the model in terms of unconstrained parameters θ^*, with the constraint induced in transforming back to the θ parameterization. Although this strategy often works, such a convenient reparameterization is not always available.

We present a more general construction by inducing intrinsic distribution using a conditional density. Starting from a prior $\pi_{0,\mathcal{R}}(\theta)$ on a 'less constrained' space $\mathcal{R} \supset \mathcal{D}$, we constrain it on $\mathcal{D}$ by inducing its conditional density given $\theta \in \mathcal{D}$, $\pi_{0,\mathcal{D}}(\theta) = \pi_0(\theta \mid \theta \in \mathcal{D})$. For simplicity, we focus on $\mathcal{R}$ being Eucledean space $\mathbb{R}^p$ or its truncated subspace.

It is possible that $\mathcal{D}$ has exactly $\mu^p(\mathcal{D}) = 0$, where μ^p denotes the p-dimensional Lebesgue measure. To circumvent the definition difficulty, *regular conditional probability* (r.c.p.) is commonly used (Kolmogorov, 1950). For this article to be self-contained, we list the definition as below.

Let $(\Omega, \mathscr{B}, P)$ be a probability space and sub-σ-field $\mathscr{A} \subseteq \mathscr{B}$, the function $P(\cdot \mid \mathscr{A})$ is a regular conditional probability on $\mathscr{B}$ given $\mathscr{A}$, if:

1. For each $\omega \in \Omega$, $P(\cdot \mid \mathscr{A})(\omega)$ is a probability measure on $\mathscr{B}$.

2. For each $B \in \mathscr{B}$, the mapping $P(B \mid \mathscr{A})(.)$ is $\mathscr{A}$- measurable.

3. For each $A \in \mathscr{A}$, $B \in \mathscr{B}$, $P(A \cap B) = \int_A P(B \mid \mathscr{A})(\omega) P(d\omega)$.

where ω represents a random variable on $(\Omega, \mathscr{B})$. Therefore, one can have $P(A) = 0$ but a valid $P(B \mid \mathscr{A})(\omega) \in [0, 1]$. Using conventional statistical notation, the above definition is commonly written as $P(B \mid \omega = \omega_0)$, given specific value ω_0.

In our case, we need to define a random variable ω associated with $\theta \in \mathcal{D}$ and $\theta \notin \mathcal{D}$. Letting $v : \mathbb{R}^p \to [0, \infty)^d$ Lipschitz with $p > d$ define a measurable distance to the constrained space $\mathcal{D}$, then $v(\theta) = \mathbf{0}$ is equivalent to $\theta \in \mathcal{D}$, otherwise $\|v(\theta)\| > 0$ indicates $\theta \notin \mathcal{D}$. Taking $\omega = v(\theta)$, a regular conditional probability of $\theta \in B$ given $\theta \in D$ is:

$$P(B \mid v(\theta) = \mathbf{0}) = \int_B \pi_{0,\mathcal{D}}(\theta) d\theta = \frac{1}{m(\mathbf{0})} \int_B \frac{\pi_{0,\mathcal{R}}(\theta) \mathbb{1}_{v(\theta)=\mathbf{0}}}{J_p v(\theta)} \mathcal{H}^{p-d}(d\theta) \tag{1}$$

where $J_p v(\theta) = \sqrt{\det(D\theta' D\theta)}$ with $D\theta$ as the derivative matrix if $\det(D\theta' D\theta) > 0$, or $J_p v(\theta) = 1$ if $\det(D\theta' D\theta) = 0$; $m(x) = \int \frac{\pi_{0,\mathcal{R}}(\theta) \mathbb{1}_{v(\theta)=x}}{J_p v(\theta)} \mathcal{H}^{p-d}(d\theta)$ is the marginal density of v, as the result of p-to-d transform using co-area formula (Federer, 2014); The notation $\mathcal{H}^k(.)$ denotes a k-dimensional Hausdorff measure. The definition (1) is a special case of the regular conditional probability proposed by Diaconis et al. (2013), which is derived based on conditioning a subset A inside a manifold $\mathcal{M}$. In our case, we take $\mathcal{M} = \mathcal{R}$ and $A = \mathcal{D}$.

The Hausdorff measure is defined as $\mathcal{H}^k(A) = \lim_{\delta \to 0} \inf\{\sum \frac{\Gamma(\frac{1}{2})^k}{\Gamma(\frac{k}{2}+1)} \left[\frac{\text{diam}(S_i)}{2}\right]^k : A \subseteq \cup S_i, \text{diam}(S_i) \leq \delta, \text{diam}(S_i) = \sup_{x,y \in S} \|x - y\|\}$. Geometrically, it is the volume of minimum covering of A formed by countably many balls with a uniform diameter δ. Depending on k, $\mathcal{H}^k(A)$ can vary from 0 in high dimesion to possibly ∞ in low dimension (at $k = 0$, it is a re-scaling of counting measure). The reduced dimension $p - d$ makes $0 < m(\mathbf{0}) < \infty$, hence a valid conditional density can be formed. Leo: we may need more justification on why $p - d$ is a perfect dimension to have finite, positve $\mathcal{H}$ measure. Despite the geometric formulation, (1) can be computed by substituting $\mathcal{H}^k(d\theta) = \frac{\Gamma(\frac{k}{2}+1)}{\Gamma(\frac{1}{2})^k} d\theta$, since Hausdorff measure is a rescaling of Lebesgue measure when k is an integer (Federer, 2014).

Given data y, the posterior probability of set B is

$$\int_B \pi_{\mathcal{D}}(\theta \mid y)d\theta = \frac{1}{m(\mathbf{0} \mid y)} \int_B \frac{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=\mathbf{0}}}{J_p v(\theta)} \mathcal{H}^{p-d}(d\theta) \tag{2}$$

if $0 < m(\mathbf{0} \mid y) < \infty$, $m(x \mid y) = \int \frac{L(\theta;y)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{v(\theta)=x}}{J_p v(\theta)} \mathcal{H}^{p-d}(d\theta)$ is the marginal density of v given y.

In general, (1) and (2) are analytically intractable and difficult to approximate with Monte Carlo sampling, due to the indicator function $\mathbb{1}_{v(\theta)=\mathbf{0}}$ that constrains $\theta \in \mathcal{D}$; although closed-form may exist for some simple case. To illustrate, we use one toy example throughout this section.

Example 1A: Two Gaussians with Sum Constraint (Intrinsic Approach)

Consider a bivariate Gaussian random vector $[\theta_1, \theta_2]' \sim \text{No}(0, I)$ in constrained space $\mathcal{D} = \{(\theta_1, \theta_2) : \theta_1 + \theta_2 - 1 = 0\}$. Denoting $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, for set $B = \{(\theta_1, \theta_2) : \theta_1 < x, \theta_2 \in \mathbb{R}\}$, $J_2 v(\theta) = 2$.

$$\begin{aligned}
\int_B \pi_{0,\mathcal{D}}(\theta)d\theta =& \frac{\int_{-\infty}^{x} \frac{1}{2}\phi(\theta_1)\phi(\theta_2)\mathbb{1}_{\theta_1+\theta_2-1=0}\frac{\Gamma(\frac{1}{2}+1)}{\Gamma(\frac{1}{2})}d\theta_1}{\int_{-\infty}^{\infty} \frac{1}{2}\phi(\theta_1)\phi(\theta_2)\mathbb{1}_{\theta_1+\theta_2-1=0}\frac{\Gamma(\frac{1}{2}+1)}{\Gamma(\frac{1}{2})}d\theta_1} \\
=& \frac{\int_{-\infty}^{x} \frac{1}{2}\phi(\theta_1)\phi(1-\theta_1)d\theta_1}{\int_{-\infty}^{\infty} \frac{1}{2}\phi(\theta_1)\phi(1-\theta_1)d\theta_1} \\
=& \int_{-\infty}^{x} \frac{\sqrt{2}}{\sqrt{2\pi}} \exp(-\frac{(\theta_1 - \frac{1}{2})^2}{2/2})d\theta_1,
\end{aligned}$$

which corresponds to $\theta_1 \mid (\theta_1 + \theta_2 = 1) \sim \text{No}(1/2, 1/2)$, $\theta_2 \mid \theta_1 \sim \delta_{1-\theta_1}(.)$. Marginally, this is a degenerate bivariate Gaussian distribution:

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

2.2 Extrinsic Bayes

We now propose an extrinsic distribution that builds on (1) and (2), approximating the sharp $\mathbb{1}_{v(\theta)=\mathbf{0}}$ with a *smooth* alternative $\mathcal{K}(\theta; \mathcal{D})$ with less contrained support:

$$\int_B \tilde{\pi}_{0,\mathcal{D}}(\theta)d\theta = \frac{1}{m} \int_B \frac{\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta; \mathcal{D})}{J_p v(\theta)} \mathcal{H}^{p-d}(d\theta) \tag{3}$$

where $m = \int_{\mathcal{R}} \frac{\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})}{J_p v(\theta)} \mathcal{H}^{p-d}(d\theta)$. The posterior takes similar form

$$\int_B \tilde{\pi}_{\mathcal{D}}(\theta \mid y)d\theta = \frac{1}{m} \int_B \frac{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta; \mathcal{D})}{J_p v(\theta)} \mathcal{H}^{p-d}(d\theta) \tag{4}$$

with $m = \int_{\mathcal{R}} \frac{L(\theta;y)\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})}{J_p v(\theta)} \mathcal{H}^{p-d}(d\theta)$.

The function $\mathcal{K}(\theta; \mathcal{D})$ maps from $\mathbb{R}^p \to [0, 1]$ and is defined as:

$$\mathcal{K}(\theta; \mathcal{D}) = \prod_{k=1}^{d} K_k(v_k(\theta)), \tag{5}$$

where $v_k(\theta)$ corresponds to left hand side of the kth equation in $v(\theta) = \mathbf{0}$, as defined in the last section. Instead of assigning point mass at $v(\theta) = \mathbf{0}$, K_k assigns mass *concentrated* at 0 with low but positive density for $v_k(\theta) > 0$. This expands the sampling space from $\mathcal{D}$ to $\mathcal{R}$. For simplicity, from now on we focus on exponential smoothing function $K_k(v(\theta)) = \exp(-\frac{v(\theta)}{\lambda_k})$ with $\lambda_k > 0$ as a tuning parameter.

We now elaborate the notion of distance $v_k(\theta)$. Assuming there are d constraints with each defining a constrained subspace $\mathcal{D}_k$, $\mathcal{D} = \bigcap_{k=1}^{d} \mathcal{D}_k$, then $v_k : \mathcal{R} \to [0, \infty)$ is a measurable function and quantifies the distance to space $\mathcal{D}_k$. For $k = 1, \ldots, m$, $v_k(\theta) = 0$ only if $\theta \in \mathcal{D}_k$. For example, one can use $v(\theta) = |f(\theta)|$ as a distance to equality-constrained space $\{\theta : f(\theta) = 0\}$; $v(\theta) = |f(\theta)|_+$, where $(x)_+ = \begin{cases} 0 \text{ if } x \leq 0 \\ x \text{ if } x > 0 \end{cases}$ as a distance to inequality-constrained space $\{\theta : f(\theta) \leq 0\}$. This idea is flexible to accomodate more complicated scenarios. For example, θ can have only a subset of parameters constrained; parameters can be in multiple constraints simultaneously; constraints can be dependent.

To summarize, $\theta \in \mathcal{D} \Leftrightarrow$ all $v_k(\theta) = 0 \Leftrightarrow \mathcal{K}(\theta; \mathcal{D}) = 1$, leading to $\mathcal{K}(\theta; \mathcal{D}) = \mathbb{1}_{\theta \in \mathcal{D}}$ if $\theta \in \mathcal{D}$, but $\mathcal{K}(\theta; \mathcal{D}) > 0$ if $\theta \notin \mathcal{D}$. To provide some intuition about the smoothing, Figure 1 plots the densities of a truncated normal prior $\text{No}_{(-\infty,5)}(0, 5^2)$ and the extrinsic approximation. We use $\mathcal{K}(\theta; \mathcal{D}) = \exp(-v(\theta))$ with $v(\theta) = (\theta - 5)_+$ and $v(\theta) = (\theta - 5)_+^2$ as two example distances. Inside $\mathcal{D} = (-\infty, 5)$, both intrinsic and extrinsic distribution are the same, except for a different normalizing constant; outside $\mathcal{D}$, intrinsic one drops directly to 0 at the boundary, whereas the extrinsic one decreases smoothly.

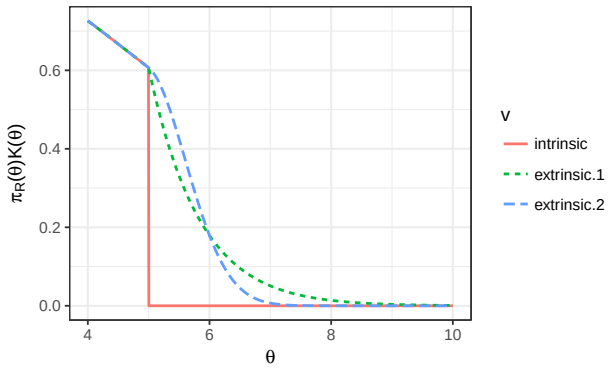


Figure 1: Unnormalized densities for truncated normal $\text{No}_{(-\infty,5)}(0, 5^2)$ under exact intrinsic prior and approximating extrinsic prior. Inside $(-\infty, 5)$, the priors are the same up to a constant difference. The intrinsic prior abruptly drops to 0 on the boundary, while the approximating ones drop smoothly. Intrinsic prior based on first-order $v(\theta)$ drops faster than the one based on second order when $v(\theta) \in (0, 1)$.

We now return to the previous example of two Gaussians under sum constraint, applying extrinsic Bayes technique.

Example 1B: Two Gaussians with Sum Constraint (Extrinsic Approach)

We use extrinsic prior $\tilde{\pi}_{0,\mathcal{D}}(\theta) \propto \exp(-\frac{\theta_1^2+\theta_2^2}{2})\exp(-\frac{v(\theta)}{\lambda})$. Choosing $v(\theta) = (\theta_1+\theta_2-1)^2$ allows us to obtain closed-form for the extrinsic prior $\theta_1 \sim \text{No}(\frac{2}{\lambda+4}, \frac{\lambda+2}{\lambda+4})$, $\theta_2 \mid \theta_1 \sim \text{No}(\frac{2}{\lambda+2}(1-\theta_1), \frac{\lambda}{\lambda+2})$. 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 $\lambda \to 0$, the extrinsic prior becomes the same as the degenerate bivariate Gaussian in intrinsic approach.

For more general cases, (3) and (4) do not have closed-form; however, it is now easy to use conventional Monte Carlo techniques since $\mathcal{K}(\theta;\mathcal{D})$ expands the space from $\mathcal{D}$ to $\mathcal{R}$.

2.3 Approximation Error

We now quantify approximation error of extrinsic distribution. Due to similar form in prior and posterior, we now introduce some general notation. Let $\pi_{\mathcal{R}}(\theta)$ be the normalized density in $\mathcal{R}$ such that $\int \pi_{\mathcal{R}}(\theta) = 1$, which is $\pi_{0,\mathcal{R}}(\theta)$ for prior and $L(y;\theta)\pi_{0,\mathcal{R}}(\theta)$ for posterior; $\Pi(.)$ and $\tilde{\Pi}(.)$ to represent the measures under intrinsic and extrisic distributions.

Suppose $\Phi : \mathbb{R}^p \to \mathbb{R}^d$ with $p > d$ is Lipschitz, then the co-area formula (Federer, 2014) is,

$$\int_{\mathbb{R}^n} f(\theta) J_N \Phi(\theta) \mu^n(d\theta) = \int_{\mathbb{R}^m} \int_{\Phi^{-1}(y)} f(\theta) \mathcal{H}^{n-m}(d\theta)\mu^m(dy), \tag{6}$$

where $\mu^k(d\theta)$ a k-dimensional Lebesgue measure. We first re-parameterize $\mathcal{K}(\theta;\theta)$ as $\exp(-\lambda^{-1}v(\theta))$ where $\lambda = \max_k \lambda_k$ and $v(\theta) = \prod_{k=1}^d \frac{v_k(\theta)}{\lambda_k^*}$ with $\lambda_k^* = \lambda_k/\lambda$. Having $\Phi(\theta) = v(\theta)$ yields $Jv(\theta) = \|\nabla v(\theta)\|$ if $\|\nabla v(\theta)\| > 0$ and $Jv(\theta) = 1$ if $\|\nabla v(\theta)\| = 0$.

Having $f(\theta) = \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)}$ and $f(\theta) = \frac{\pi_{\mathcal{R}}(\theta)g(\theta)}{Jv(\theta)}$ yield the marginal density and conditional expectation, respectively:

$$\begin{aligned} m(x) &= \int_{v^{-1}(x)} \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta),\\ \mathbb{E}(g(\theta) \mid x) &= \int_{v^{-1}(x)} \frac{g(\theta)\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta). \end{aligned} \tag{7}$$

Comparing the posterior samples from intrinsic and extrinsic distributions, 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.

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

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

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

$$W_1(\Pi, \tilde{\Pi}) \leq \lambda(\frac{k_1 k_3}{m_0^2} + \frac{k_1}{m_0}) + \exp(-\lambda^{-1} t)(\frac{k_1}{m_0^2} + \frac{k_2}{m_0}), \tag{8}$$

if there exists a t-ball surrounding $\mathcal{D}$, $\{\theta : v(\theta) < t\}$ *having bounded marginal density for* $v(\theta)$ *and conditional expectation for any Lipschitz functions,* $k_1, k_3 = \mathcal{O}(1)$ *and the expectation over* $\mathcal{R}$ *has* $k_2 = \mathcal{O}(\lambda \exp(\lambda^{-1} t))$, $W_1(\Pi, \tilde{\Pi})$ *converges to* 0 *in* $\mathcal{O}(\lambda)$.

The proof is provided in the appendix.

3 Posterior Computation

One particular appeal of the extrinsic approach is its advantage in posterior computation. As it is supported on a less restrictive space $\mathcal{R}$, 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 sampling with high-dimensional parameters.

3.1 Hamiltonian Monte Carlo for Extrinsic Posterior Sampling

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

In order to sample from $\theta \in \mathcal{R} \subset \mathbb{R}^d$, 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 an extrinsic posterior (4),

$$
\begin{aligned}
&H(\theta, p) = U(\theta) + K(p), \\
&\text{where } U(\theta) = -\log\left\{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta; \mathcal{D})/J(v(\theta))\right\}, \\
&K(p) = \frac{p' M^{-1} p}{2}.
\end{aligned} \tag{9}
$$

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:

$$
\begin{aligned}
\frac{\partial \theta^{(t)}}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = M^{-1} p, \\
\frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}.
\end{aligned} \tag{10}
$$

The exact solution to (10) is typically intractable but a valid Metropolis proposal can be generated by numerically approximating (10) 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:

$$
p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta}, \quad \theta \leftarrow \theta + \epsilon M^{-1} p, \quad p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta} \tag{11}
$$

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 $\eta(\theta; \mathcal{D}_1^*) = \inf_{\theta^* \notin D_1^*} \|\theta^* - \theta\|$. Therefore,

$$\epsilon < \eta(\theta; \mathcal{D}_1^*) \wedge 2\omega_1(\theta)^{-1/2} \tag{12}$$

In the case of an extrinsic posterior with $\mathcal{K}(\theta; \mathcal{D}) = \prod_{k=1}^d \exp(-\lambda_k^{-1} v_k(\theta))$, the Hessian $\mathbf{H}_U(\theta)$ is given by

$$\mathbf{H}_U(\theta) = -\mathbf{H}_{\log \pi_{\mathcal{R}}}(\theta) + \sum_k \lambda_k^{-1} \mathbf{H}_{v_k}(\theta) \mathbb{1}_{\theta \notin \mathcal{D}_k}, \tag{13}$$

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 $\lambda_k^{-1} \mathbf{H}_{v_k}(\theta) \mathbb{1}_{\theta \notin \mathcal{D}_k}$ is 0 unless $\theta \notin \mathcal{D}_k$. When $\theta \notin \mathcal{D}_k$, the eigenvalue of (13) is commonly dominated by λ_k^{-1}.

The key for computational efficiency is to prevent the bound in (12) 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 $\theta \notin \mathcal{D}_{(j')}$ to happen. As the result, the effective stability bound is:

$$\epsilon < \eta(\theta; \cap_{j=1}^m \mathcal{D}_{(j)}) \wedge \left(2 \min_{j' \in \{m+1, \ldots, d\}} \lambda_{(j')}^{1/2} \right) \tag{14}$$

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; whereas equality constraints need relaxation with moderate $\lambda_{j'}$ in the second term, as they commonly define hyper-plane that has $\eta(.) \approx 0$.

For $\lambda_{j'}$ not very close to 0, a trade-off between approximation accuracy and computational efficiency is involved. Fortunately, the approximation error $\mathcal{O}(\max_{j'\in\{m+1,\ldots,d\}}\lambda_{(j')})$ decreases faster than the efficiceny cap $\mathcal{O}(\min_{j'\in\{m+1,\ldots,d\}}\lambda_{(j')}^{1/2})$. 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.

4 Simulated Examples and Application

We now use examples to illustrate the properties of extrisinc priors and their utility in common scenarios.

4.1 Simulations

We first use two examples to demonstrate how to choose λ_k's, depending on the geometry of $\mathcal{D}$.

Example 2: Linear Regression Under Linear Inequality

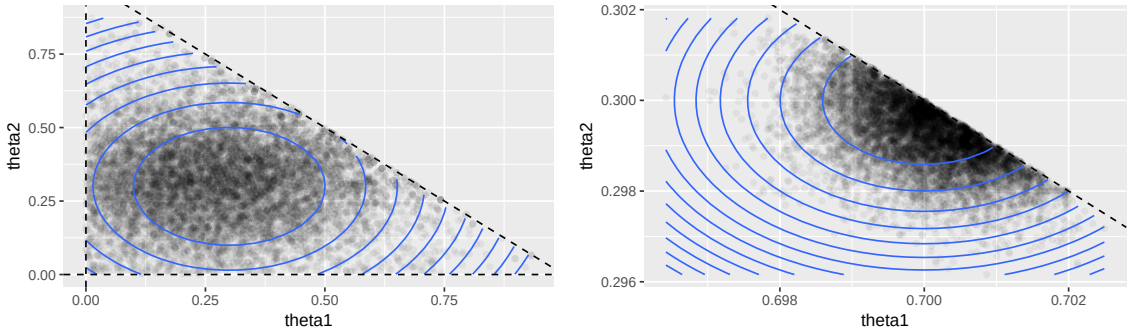
Consider a linearly constrained regression model:

$$y_i \sim \mathrm{No}(x_i\theta, \sigma^2) \text{ for } i = 1, \ldots n, \quad \text{with } A\theta \leq c$$

where parameter θ is a p-dimensional vector; the constraint parameters A, a $d \times p$ matrix, and c, a d-dimensional vector, are both fixed and given. The constrained space $\mathcal{D}$ is the polyhedrons in $\mathbb{R}^p$ formed by d lines. Unless inequalities form a very tight space, often $\eta(\theta; \mathcal{D})$ is not very close to 0.

We consider simple bivariate case $\theta \in (0,1)^2$ subject to $\theta_1+\theta_2 \leq 1$, making $\mathcal{D}$ a triangle. To simulate data, we use $\sigma^2 = 0.1^2$, $x_i \sim \mathrm{No}([0,0]', I)$ for $i = 1, \ldots, n$. We then generate two datasets using different values of θ and n. In the first experiment, we use $\theta = [0.3, 0.3]'$ with $n = 10$, so that the posterior has wide spread and centered in the interior of $\mathcal{D}$; in the second experiment, we use $\theta = [0.7, 0.3]'$ with $n = 10^4$ so that the posterior is concentrated on the boundary. In both cases, we assign weakly informative prior for $\theta \sim \mathrm{No}([0.5, 0.5]', I10^2)$ and inverse-Gamma prior $\sigma^2 \sim \mathrm{IG}(2, 1)$; we use $\mathcal{K}(\theta) = \exp(-\frac{v(\theta)}{\lambda})$ with $v(\theta) = |\theta_1 + \theta_2 - 1|_+$ in the extrinsic prior. This yields posterior $\theta \mid \sigma^2, y \sim \mathrm{No}_{\theta\in(0,1)^2, \theta_1+\theta_2<1}(\mu_\theta, \Sigma_\theta)$, where $\Sigma_\theta = (x'x/\sigma^2 + I/10^2)^{-1}$ and $\mu_\theta = \Sigma_\theta(x'y/\sigma^2 + [0.5, 0.5]'/10^2)$.

The chosen $\lambda = 10^{-8}$ leads to almost no support expansion, while the distance to boundary $\eta(\theta; \mathcal{D})$ is large enough to allow efficient sampling. We collect $10,000$ posterior samples. Figure 2 plots the posterior sample and its contour. There is no posterior fallen outside $\mathcal{D}$, thanks to $\lambda \approx 0$. To compare, we also use simple rejection sampling by proposing θ from untruncated normal $\mathrm{No}(\mu_\theta, \Sigma_\theta)$ and reject if $\theta \notin \mathcal{D}$. The first experiement has rejection rate 12% and the second has 51%; obviously, the rejection rate will continue to rise when μ_θ is further away from $\mathcal{D}$.



(a) $n = 10$ with true $\theta = [0.3, 0.3]'$ (b) $n = 10^4$ with true $\theta = [0.7, 0.3]'$

Figure 2: Extrinsic posterior distribution of the normal mean θ, with approximation to constraint $\theta_1 + \theta_2 \leq 1$. Posterior is either loosely distributed near the center (panel (a)) or concentrated on the boundary (panel (b)) of the region. The extrinsic posterior has no samples outside of the region due to almost no relaxation.

Example 3: Unit Circle

Consider $\mathcal{D}$ as a two-dimensional unit cirlce $\{(\theta_1, \theta_2) : \theta_1^2 + \theta_2^2 = 1\}$, equivalently a $(2, 1)$–Stiefel manifold. Within this space, $\eta(\theta; \mathcal{D}) = 0 \forall \theta \in \mathcal{D}$ hence support expansion is necessary with moderate λ.

Let data $y_i \in \mathbb{R}^2$ for $i = 1, \ldots, n$ be noisy realization from one point on unit circle:

$$y_i \sim \text{No}(\theta, I_2\sigma^2), \text{ with } \theta'\theta = 1,$$

where $\theta \in \mathcal{D}$ is assigned a von Mises–Fisher prior $\pi_{0,\mathcal{D}}(\theta) \propto \exp(F'\theta)$ (with corresponding unconstrained prior $\pi_{0,\mathcal{R}}(\theta) \propto \exp(F'\theta)\|\theta\|$).

To generate data, we use $\theta = (\sqrt{3}/2, 1/2)$, $\sigma^2 = 0.5^2$ and small $n = 10$, in order to induce widely spread-out posterior θ on the manifold. We then use $F = (1, 1)$ to induce a weakly informative prior for θ and an inverse-Gamma prior $\text{IG}(2, 1)$ for σ^2. To assign extrinsic prior, we use $v(\theta) = |\theta'\theta - 1|$ as the distance to circle and extrinsic prior $\tilde{\pi}_{0,\mathcal{D}}(\theta) = \exp(F'\theta)\exp(-\frac{|\theta'\theta-1|}{\lambda})$.

We test $\lambda = 10^{-3}$, 10^{-4} and 10^{-5} in three experiments, each corresponding to a different stability bound in leap-frog algorim. To visualize effects of stability bound, we restrict the maximum leap-frog steps L to be 100 and show how much space each setting can explore within one HMC iteration. Figure 3 plots its path of $L = 100$ leap-frog steps.

Since the intrinsic posterior has closed-form $\pi_{\mathcal{D}}(\theta) \propto \exp[(F + \sum_i y_i/\sigma^2)'\theta]$, we collect the sample and compare the result. Each experiment is repeated 10 times. Table 1 shows the numeric 1-Wasserstein distance from its extrinsic approximate sample to exact posterior, and the posterior distance to $\mathcal{D}$, $v(\theta)$. All three settings have low approximation errors.

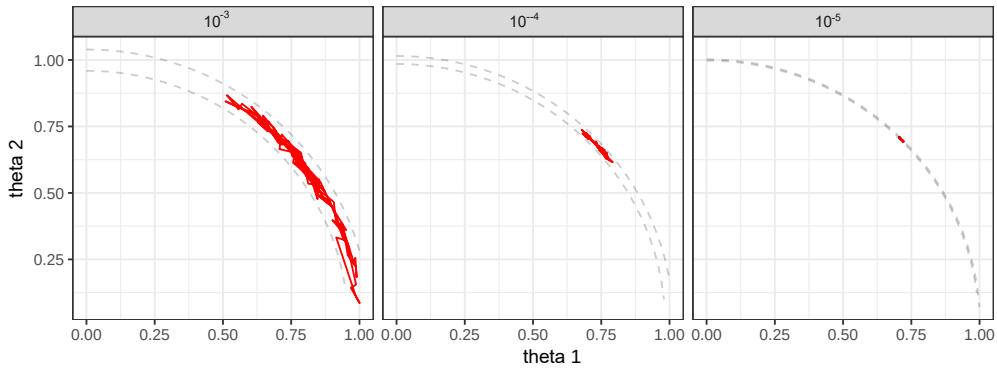


Figure 3: Path of 100 itegrator steps in one HMC iteraton, sampling on a unit circle via extrinc prior with $\mathcal{K}(\theta) = \exp(-\frac{|\theta'\theta-1|}{\lambda})$, with $\lambda = 10^{-4}$, $= 10^{-5}$ and $= 10^{-6}$. Larger relaxation in the narrowest direction of support (orthogonal vector to the circle) result in more efficient space exploration.

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)
$v(\theta) \mid y$	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) of sampling from extrinsic distribution for a von–Mises Fisher distribution on a unit circle. The numeric 1-Wasserstein distance W_1 is low for all three settings and close to the numeric error in comparing two independent samples from the exact distribution; the posterior distance to $\mathcal{D}$, $v(\theta) = |\theta'\theta - 1|$ is quite small. The computing time needed for generating every 1000 effective samples is shown.

So far we have compared extrinsic approach with intrinsic one in settings with conjugate posterior. More advanced models often do not have simple form in constrained space. We now show two such examples where constrained models substantially outperform the unconstrained ones. Although these models create computing difficulty in intrinsic approach, they can be easily estimated extrinsically.

4.2 Applications

Example 3: Ordered Dirichlet Prior

We first consider an ordered simplex in finite mixture model. The $(J-1)$–simplex is a vector $w = \{w_1, \dots w_J\}$ with $1 > w_1 \geq \dots \geq w_J > 0$ and $\sum_{j=1}^{J} w_j = 1$. This is built on the standard Dirichlet prior $Dir(\alpha)$, with $\pi_{0,\mathcal{D}}(w) = \prod_{j=1}^{J} w_j^{\alpha-1} \mathbb{1}_{\sum_{j=1}^{J} w_j = 1}$ with additional order constraint. In the base model Dirichlet prior, index j is exchangeable and its permutation does not change the likelihood. This leads to label-switching problem in mixture model estimation (reviewed in Jasra et al. (2005)).

Imposing order constraint yields an ordered Dirichlet prior:

$$\pi_{0,\mathcal{D}}(w_1, \dots w_J) \propto \prod_{j=1}^{J} w_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^{J} w_j = 1} \cdot \prod_{j=1}^{J-1} \mathbb{1}_{w_j \geq w_{j+1}} \,. \tag{15}$$

where $w_j \in (0,1)$ for $j = 1, \dots, J$. Although the standard Gibbs sampling algorithm (Ishwaran and James, 2001) no longer applies in this setting, the ordered Dirichlet prior can be approximated by extrinsic prior:

$$\tilde{\pi}_{0,\mathcal{D}}(w_1, \dots w_J) \propto \prod_{j=1}^{J} w_j^{\alpha-1} \cdot \exp(-\frac{\sum_{j=1}^{J}(w_{j+1} - w_j)_+}{\lambda_1}) \exp(-\frac{|\sum_{j=1}^{J} w_j - 1|}{\lambda_2})$$

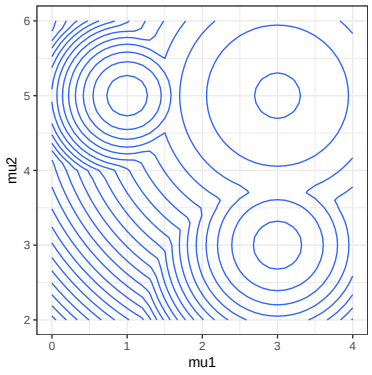
We now adopt this simplex distribution in a normal mixture model with mixture means and common variance, for data $y_i \in \mathbb{R}^d$ indexed by $i = 1, \dots, n$:

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

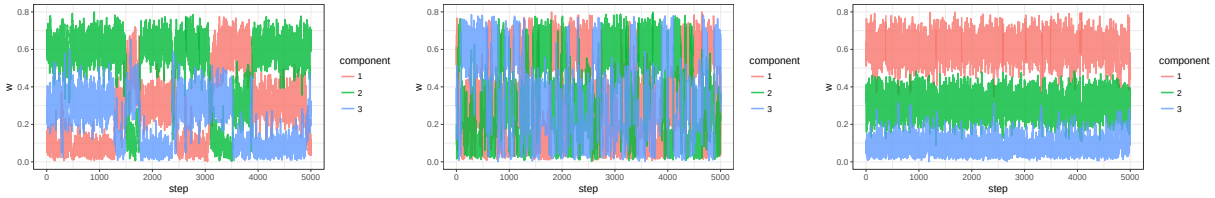
where $\delta_b(a) = 1$ if $a = b$ and 0 otherwise.

We generate $n = 100$ samples from 3 components with true $\{w_1, w_2, w_3\} = \{0.6, 0.3, 0.1\}$ and two-dimensional means $\{\mu_1, \mu_2, \mu_3\} = \{[1,5], [3,3], [3,5]\}$ with identity covariance $\Sigma = I_2$. We assign weakly informative priors $\text{No}(0, 10I_2)$ for each μ_j and inverse Gamma prior for the digonal element in $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$ with $\sigma_1^2, \sigma_2^2 \sim IG(2,1)$. We use $\lambda_1 = 10^{-6}$ to induce almost no relaxation on the ordering and $\lambda_2 = 10^{-3}$ on simplex constraint. To demonstrate the benefit of ordered Dirichlet, we compare the base model Dirichlet prior without order constraint, estimating via Gibbs sampling and HMC.

Figure 4(a) shows the contour of true posterior density of μ_j's. The small component sample size leads to large overlap among the posterior of μ_j's, genering in significant label-switching in both Gibbs and HMC under canonical Dirichlet prior. Figure 4(b,c,d) show the traceplot of w. Ordered Dirchlet has almost no label-switching due to ordering.



(a) Posterior density of the component means.



(b) Gibbs sampling under canonical Dirichlet

(c) HMC sampling under canonical Dirichlet, with extrinsic prior

(d) HMC sampling under ordered Dirichlet, with extrinsic prior

Figure 4: Contour of the posterior density of component means and traceplot of the posterior sample for the component weights w, in a 3-component normal mixture model. Panel (a) shows that there is significant overlap among component means μ_j's, creating label-switching issues in both Gibbs sampling (b) and HMC using canonical prior (c). The ordered Dirichlet prior significantly reducing label-switching (d).

Example 4: Orthonormal Low Rank Factorization of Multiple Undirected 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 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:

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

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

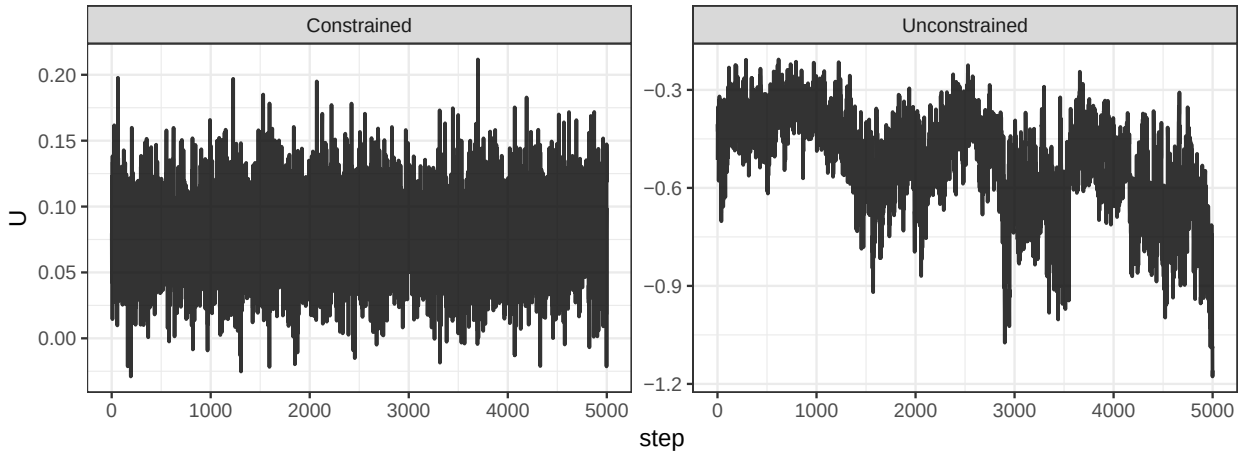
for $k > l$, $k = 2, \ldots, V$, $i = 1, \ldots, n$; U is $V \times d_1$ matrix, W is $n \times d_2$ matrix; D is a $d_1 \times d_2$ array. The $V \times V$ matrix Z is almost unstructural except symmetric $Z_{k,l} = Z_{l,k}$, which is commonly used to induce low-rank in the decomposition (Durante et 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 $D^*_{\cdot,r_2,\cdot} = R^{-1} D_{\cdot,r_2,\cdot} R^{-1}$ for $r_2 = 1, \ldots, d_2$. This leaves the likelihood unchanged, creating identifiability issue.

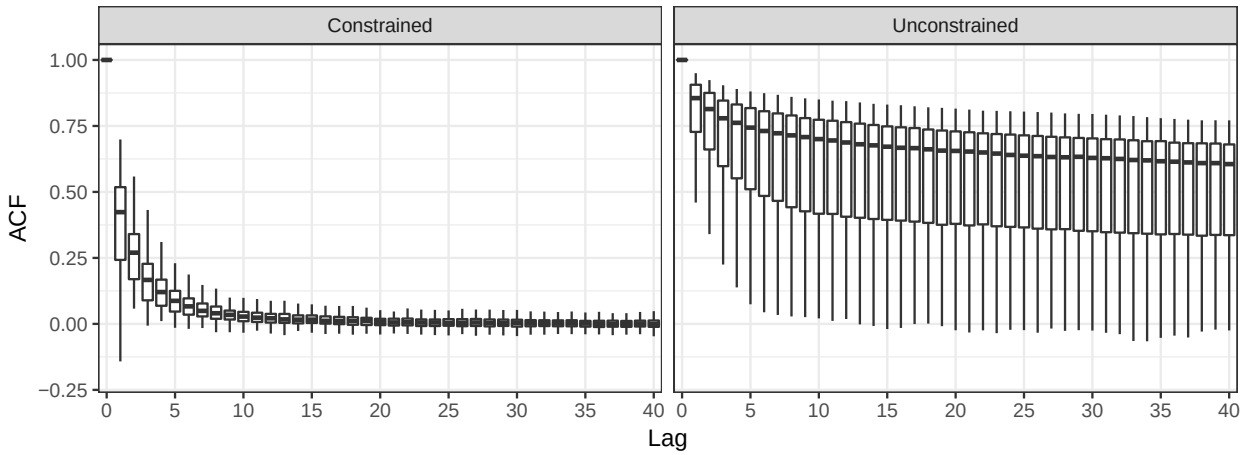
Therefore, we consider applying some constraint on the Tucker decomposition. Motivated by high-order singular value decomposition, we impose orthonormality constraints $U'U = I_{d_1}$ and $W'W = I_{d_2}$. Hoff et al. (2016) previously obtained conjugated posterior for Tucker decomposition under orthonormality 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 5 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 5: 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*.

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).

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

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

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

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

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

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

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

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

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

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

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

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

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

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:

$$W_1(\Pi, \tilde{\Pi}) = \sup_{g:\|g\|_L \leq 1} \int g(x)\Pi(dx) - \int g(y)\tilde{\Pi}(dy) \tag{16}$$

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

$$\begin{aligned} m_\lambda &= \int_{\mathbb{R}} \left[\int_{v^{-1}(x)} \frac{\exp(-\lambda^{-1}v(\theta))\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) \right] \mathbb{1}_{x\geq 0} dx \\ &= \int_{\mathbb{R}} m(x) \exp(-\lambda^{-1}x) \mathbb{1}_{x\geq 0} dx. \end{aligned} \tag{17}$$

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

$$m_0 = \int_{\mathbb{R}} \left[\int_{v^{-1}(y)} \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) \right] \mathbb{1}_{y=0} dy = \int_{v^{-1}(0)} \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) = m(0) \tag{18}$$

Clearly $m_\lambda \geq m_0$.

1. Asymptotic result:

We have

$$\begin{aligned} &\sup_{g:\|g\|_L \leq 1} \int g(\theta) \left[\frac{\exp(-\lambda^{-1}v(\theta))}{m_\lambda} - \frac{\mathbb{1}_{v(\theta)=0}}{m_0} \right] \frac{\pi_{\mathcal{R}}(\theta)}{Jv(\theta)} \mathcal{H}^{p-d}(d\theta) \\ &= \sup_{g:\|g\|_L \leq 1} \int_{\mathbb{R}} \mathbb{E}(g(\theta) \mid x) \left[\frac{\exp(-\lambda^{-1}x)\mathbb{1}_{x\geq 0}}{m_\lambda} - \frac{\mathbb{1}_{x=0}}{m_0} \right] dx \\ &= \sup_{g:\|g\|_L \leq 1} \int_{\mathbb{R}} \mathbb{E}(g(\theta) \mid x) \left[\frac{1}{m_\lambda} - \frac{1}{m_0} \right] \mathbb{1}_{x=0} dx + \sup_{g:\|g\|_L \leq 1} \int_{\mathbb{R}} \mathbb{E}(g(\theta) \mid x) \frac{\exp(-\lambda^{-1}x)}{m_\lambda} \mathbb{1}_{x>0} dx \\ &\leq \sup_{g:\|g\|_L \leq 1} \|\mathbb{E}(g(\theta) \mid 0)\| \left[\frac{1}{m_0} - \frac{1}{m_\lambda} \right] + \frac{1}{m_0} \sup_{g:\|g\|_L \leq 1} \int_{\mathbb{R}} \|\mathbb{E}(g(\theta) \mid x)\| \exp(-\lambda^{-1}x) \mathbb{1}_{x>0} dx \end{aligned} \tag{19}$$

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

$$\lim_{\lambda \to 0} m_\lambda = \int_{\mathbb{R}} m(x) \lim_{\lambda \to 0} \exp(-\lambda^{-1}x) \mathbb{1}_{x\geq 0} dx = m_0. \tag{20}$$

Since $\sup_{g:\|g\|_L\leq 1}\int_{\mathbb{R}}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)dx \leq \int_{\mathbb{R}}\sup_{g:\|g\|_L\leq 1}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)dx$, letting $q_\lambda = \sup_{g:\|g\|_L\leq 1}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}$, we have $0 \leq q_1 - q_{\lambda_1} \leq q_1 - q_{\lambda_2}$ for $1 \geq \lambda_1 \geq \lambda_2$, 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

$$\lim_{\lambda\to 0} W_1(\Pi,\tilde{\Pi}) = 0. \tag{21}$$

2. Non-asympotic result:

$$\begin{aligned}
\frac{1}{m_0} - \frac{1}{m_\lambda} &\leq \frac{\int_{\mathbb{R}} m(x)\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx}{m_0^2} \\
&= \frac{1}{m_0^2}\left[\int_0^t m(x)\exp(-\lambda^{-1}x)dx + \int_t^\infty m(x)\exp(-\lambda^{-1}x)dx\right] \\
&\leq \frac{1}{m_0^2}\left[\sup_{t^*\in(0,t)} m(t^*)\int_0^t \exp(-\lambda^{-1}x)dx + \exp(-\lambda^{-1}t)\int_t^\infty m(x)dx\right] \\
&\leq \frac{1}{m_0^2}\left[\lambda \sup_{t^*\in(0,t)} m(t^*) + \exp(-\lambda^{-1}t)\right]
\end{aligned} \tag{22}$$

$$\begin{aligned}
&\sup_{g:\|g\|_L\leq 1}\int_{\mathbb{R}}\|\mathbb{E}(g(\theta)\mid x)\|\exp(-\lambda^{-1}x)\mathbb{1}_{x>0}dx \\
&\leq \sup_{g:\|g\|_L\leq 1}\sup_{t^*\in(0,t)}\|\mathbb{E}(g(\theta)\mid t^*)\|\int_0^t \exp(-\lambda^{-1}x)dx + \exp(-\lambda^{-1}t)\sup_{g:\|g\|_L\leq 1}\int_t^\infty \|\mathbb{E}(g(\theta)\mid x)\|dx \\
&\leq \sup_{g:\|g\|_L\leq 1}\sup_{t^*\in(0,t)}\|\mathbb{E}(g(\theta)\mid t^*)\|\lambda + \exp(-\lambda^{-1}t)\sup_{g:\|g\|_L\leq 1}\mathbb{E}(\|g(\theta)\|)
\end{aligned} \tag{23}$$

Combining (19)(22)(23), $k_1 = \sup_{g:\|g\|_L\leq 1}\sup_{t^*\in[0,t)}\|\mathbb{E}(g(\theta)\mid t^*)\|$, $k_2 = \sup_{g:\|g\|_L\leq 1}\mathbb{E}(\|g(\theta)\|)$, $k_3 = \sup_{t^*\in(0,t)} m(t^*)$

$$\begin{aligned}
&\sup_{g:\|g\|_L\leq 1}\int g(x)\Pi(dx) - \int g(x)\tilde{\Pi}(dx) \\
&\leq \lambda(\frac{k_1k_3}{m_0^2} + \frac{k_1}{m_0}) + \exp(-\lambda^{-1}t)(\frac{k_1}{m_0^2} + \frac{k_2}{m_0})
\end{aligned} \tag{24}$$

□