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 Langrange 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 (REFs). 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. When the constraints need to upheld strictly, the approximation can be corrected with a simple projection, followed by a Metropolis-Hastings step with high acceptance probability. Unlike intrinsic methods, such as Riemannian and geodesic HMC (Girolami and Calderhead, 2011; Byrne and Girolami, 2013), our approach is relatively efficient and 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 Bayes

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 $\mathcal{D}$. We assume that $\mathcal{D} \subset \mathcal{R}$, with $\mathcal{R}$ denoting a 'less constrained' space. For example, if θ is a p-dimensional vector subject to an inequality constraint, then $\mathcal{R}$ may correspond simply to p-dimensional Euclidean space. Assuming $\pi_{0,\mathcal{D}}(\theta)$ is proper so that $\int_{\mathcal{D}} \pi_{0,\mathcal{D}}(\theta) d\theta = 1$, the constrained prior can be obtained by starting with an unconstrained prior $\pi_{0,\mathcal{R}}(\theta)$ on $\mathcal{R}$, applying the restriction through an indicator function $\mathbb{1}_{\theta \in \mathcal{D}}$, and renormalizing:

$$\pi_{0,\mathcal{D}}(\theta) = \pi_{0,\mathcal{R}}(\theta \mid \theta \in \mathcal{D}) = \frac{\pi_{0,\mathcal{R}}(\theta) \mathbb{1}_{\theta \in \mathcal{D}}}{\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta) d\theta}, \tag{1}$$

if $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta) d\theta > 0$. When $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta) d\theta = 0$, the construction becomes more complicated.

One strategy to overcome the difficulty is using the regular conditional probability for certain set $\mathcal{A}$, via the limit

$$\int_{\mathcal{A}} \pi_{0,\mathcal{D}}(\theta)d\theta = \lim_{\mathcal{D}^+ \supset \mathcal{D}} \frac{\int_{\mathcal{A}} \pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}^+}d\theta}{\int_{\mathcal{D}^+} \pi_{0,\mathcal{R}}(\theta)d\theta} \tag{2}$$

where $\mathcal{D}^+$ is a net converging towards $\mathcal{D}$, with $\int_{\mathcal{D}^+} \pi_{0,\mathcal{R}}(\theta)d\theta > 0$. Based on this probability, one derives the constrained density. To illustrate, consider two independent uniform distribution $\theta_1, \theta_2 \sim \text{Uniform}(0,1)$ under equality constraint $\theta_1 + \theta_2 = w$. One first obtains $\int_{\theta_1<x} \pi_{0,\mathcal{D}}(\theta)d\theta = \lim_{\epsilon\to 0^+} \frac{\int_0^x \int_0^1 \mathbb{1}_{\theta\in\mathcal{D}^+}d\theta_2 d\theta_1}{\int_0^1 \int_0^1 \mathbb{1}_{\theta\in\mathcal{D}^+}d\theta_2 d\theta_1} = \frac{x}{w}$ with $\mathcal{D}^+ = \{\theta : \theta_1 + \theta_2 \in (w-\epsilon, w+\epsilon)\}$, and then obtained constrained density $\pi_{0,\mathcal{D}}(\theta_1) = \frac{1}{w}$ with $\theta_2 = w - \theta_1$.

2.2 Extrinsic Prior

Our extrinsic prior builds on the intrinsic prior in (1) and (2), approximating the sharp indicator function $\mathbb{1}_{\theta\in\mathcal{D}}$ with a *smooth* alternative having less constrained support.

$$\tilde{\pi}_{0,\mathcal{D}}(\theta) = \frac{\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})}{\int_{\mathcal{R}} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})d\theta} \tag{3}$$

where $\mathcal{K}(\theta;\mathcal{D})$ is an approximation to $\mathbb{1}_{\theta\in\mathcal{D}}$ and satisfies $\int_{\mathcal{R}} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})d\theta > 0$.

Let the constraints that define $\mathcal{D}$ be broken into m separable parts, with each corresponding to a constrained space $\mathcal{D}_k$. We have $\mathcal{D} = \bigcap_{k=1}^m \mathcal{D}_k$ and define $\mathcal{K}$ as:

$$\mathcal{K}(\theta;\mathcal{D}) = \prod_{k=1}^m K_k(v_k(\theta)) \tag{4}$$

where v_k is a function $v_k : \mathcal{R} \to [0,\infty)$ that measures the "distance" to the space $\mathcal{D}_k$, with $v_k(\theta) = 0$ when $\theta \in \mathcal{D}_k$; $K_k(v_k(\theta))$ is a function $K_k : [0,\infty) \to [0,1]$, which decreases in $v_k(\theta)$, with $K_k(0) = 1$ and $K_k(\infty) = 0$. Therefore, $\theta \in \mathcal{D}$ and $\mathcal{K}(\theta;\mathcal{D}) = 1$ if and only if all $v_k(\theta) = 0$. In this paper, we focus on a simple exponential function $K_k(v(\theta)) = \exp(-v(\theta)/\lambda_k)$, with $\lambda_k > 0$ as the tuning parameter.

The distance function $v(\theta)$ is often easy to find. For example, a large class of models can be viewed as constrained under equality $f(\theta) = 0$ and/or inequality $f(\theta) < 0$. The distance to equality constraint can be $v(\theta) = |f(\theta)|$; the distance to inequality can be $v(\theta) = |f(\theta)|_+$, where $(x)_+ = \begin{cases} 0 \text{ if } x \le 0 \\ x \text{ if } x > 0 \end{cases}$. More complicated constraint can be defined similarly. For example, for $\theta \in \mathbb{R}^{n\times p}$ with orthonormality constraint $\theta'\theta = I_p$, the distance can be $v(\theta) = ||\theta'\theta - I_p||_k$ with $||x||_k = (\sum_{ij} |x_{ij}|^k)^{1/k}$.

To illustrate the extrinsic prior, consider a truncated normal prior $\text{No}_{(-\infty,5)}(0, 5^2)$. Figure 1 plots the unnormalized densities of intrinsic prior $\pi_{0,\mathcal{R}}\mathbb{1}_{\theta\in\mathcal{D}} = \exp(-\theta^2/2\cdot 5^2)\mathbb{1}_{\theta\in(-\infty,5)}$ and extrinsic priors

$\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}) = \exp(-\theta^2/2\cdot 5^2)\exp(-v(\theta))$. For the latter, we consider 2 distances $v(\theta)$: $(\theta-5)_+$, $(\theta-5)_+^2$. Inside $\mathcal{D}$, The intrinsic and extrinsic priors are the same up to a constant difference due to normalizing. Outside $\mathcal{D}$, the extrinsic prior decreases continuously towards 0, while intrinsic prior discontinuously drops to 0 at the boundary. With the same λ, first-order distance $(\theta-5)_+$ drops faster than second-order $(\theta-5)_+^2$ when $(\theta-5)_+ < 1$.

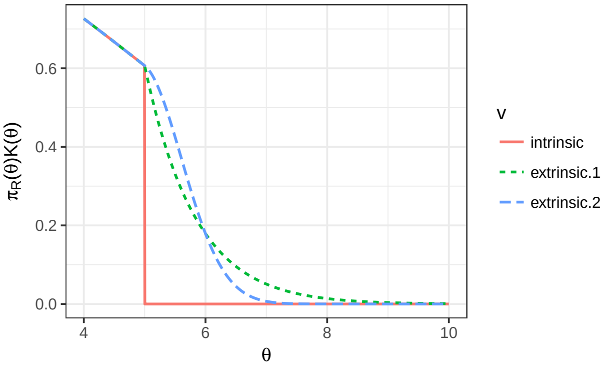


Figure 1: Unnormalized densities for truncated normal $\text{No}_{(-\infty,5)}(0,5^2)$ under exact intrinsic prior and approximating extrinsic prior. Inside $(-\infty,5)$, the priors are the same up to a constant difference. The intrinsic prior abruptly drops to 0 on the boundary, while the approximating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

This smoothing function $\mathcal{K}(\theta;\mathcal{D})$ in (4) is applicable to more general and complicated scenarios. For example, θ can have some parameters constrained and some unconstrained; some parameters can be in multiple constraints simultaneously; constraints can be dependent. In all these cases, one can find proper $\mathcal{D}_k$'s and define $v_k(\theta)$'s accordingly.

2.3 Property of Extrinsic Prior

We now study the properties of the extrinsic prior. One important task is to quantify the difference between extrinsic and intrinsic priors. We first focus on the first case in (1), when $\int_{\mathcal{D}}\pi_{0,\mathcal{R}}(\theta)d\theta > 0$.

Remark 1. *Let* $M_1 = \int_{\mathcal{D}}\pi_{0,\mathcal{R}}(\theta)d\theta$ *and* $M_2 = \int_{\mathcal{R}}\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})d\theta$*, when* $M_1 > 0$*, the total variation distance between the measures of extrinsic and intrinsic prior*

$$||\pi_{0,\mathcal{D}}(\theta),\tilde{\pi}_{0,\mathcal{D}}(\theta)||_{TV} = 1 - \frac{M_1}{M_2} \le \frac{\int_{\theta\in\mathcal{R}\setminus\mathcal{D}}\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})d\theta}{M_1}$$

.

proof: via definition of total variation distance and $K(\theta;\mathcal{D}) = 1$ when $\theta\in\mathcal{D}$, 0 otherwise.

In the case of exponential smoothing function (4), we have:

Corollary 1. *Let $M_1 = \int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta > 0$ and $\mathcal{K}(\theta; D) = \prod_{k=1}^{m} \exp(-v_k(\theta)/\lambda_k)$, one sufficient condition to have*

$$\lim_{all\ \lambda_k \to 0} ||\pi_{0,\mathcal{D}}(\theta), \tilde{\pi}_{0,\mathcal{D}}(\theta)||_{TV} = 0$$

is that $\pi_{0,\mathcal{R}}(\theta)$ is proper, $\int_{\mathcal{R}} \pi_{0,\mathcal{R}}(\theta)d\theta < \infty$.

proof: via dominated convergence theorem

Rewriting $\mathcal{K}(\theta; D) = \exp(-v(\theta)/\lambda)$ with $\lambda = \sup_k \lambda_k$, $v(\theta) = \lambda \sum_{k=1}^{m} \frac{v_k(\theta)}{\lambda_k}$, we obtain the convergence rate:

Remark 2. *Assuming $M_3 = \int_{\mathcal{R}\setminus\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta < \infty$, and $f(v)$ be the density of $v(\theta)$ as the transform of $\pi_{0,\mathcal{R}}(\theta)/M_3$. If there exists an $t < \infty$ such that $f(v) < \infty$ for $v < t$,*

$$\int_0^\infty \pi_{0,\mathcal{R}}(\theta) exp(-\frac{v(\theta)}{\lambda})d\theta \leq 2M_3 \exp(-\frac{t}{\lambda}) + M_3 \sup_{t^* \in (0,t)} f(t^*)\lambda$$

proof:

$$\begin{aligned}\int_0^\infty f(v)\exp(-\frac{v}{\lambda})dv &= \int_0^t f(v)\exp(-\frac{v}{\lambda})dv + \int_t^\infty f(v)\exp(-\frac{v}{\lambda})dv \\ &\leq F(t)\exp(-\frac{t}{\lambda}) + \frac{1}{\lambda}\int_0^t F(v)\exp(-\frac{v}{\lambda})dv + \exp(-\frac{t}{\lambda}) \\ &= (F(t)+1)\exp(-\frac{t}{\lambda}) + \frac{1}{\lambda}\int_0^t f(v^*)v\exp(-\frac{v}{\lambda})dv \qquad (5) \\ &\leq (F(t)+1)\exp(-\frac{t}{\lambda}) + \sup_{t^* \in (0,t)} f(t^*)\int_0^t \frac{1}{\lambda}v\exp(-\frac{v}{\lambda})dv \\ &\leq 2\exp(-\frac{t}{\lambda}) + \sup_{t^* \in (0,t)} f(t^*)\lambda\end{aligned}$$

where $F(t) = \int_0^t f(x)dx$ and the third step is based on mean value theorem with $v^* \in (0, v)$. Rearranging term yields the result. ■

That is, for λ small, the extrinsic prior approaches intrinsic prior in total varation distance in $O(\lambda)$. The rate is quantified under very general assumption. We expect it can be sharpened under special cases.

We now examine the second case in (2) where $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta = 0$.

LD: this needs some more work

Remark 3. *Let $M_1(\mathcal{D}^+) = \int_{\mathcal{D}^+} \pi_{0,\mathcal{R}}(\theta)d\theta$ and $M_2 = \int_{\mathcal{R}} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})d\theta$, with $\mathcal{D}^+$ chosen such that $M_1(\mathcal{D}^+) > 0$ and $M_1(\mathcal{D}^+) < M_2$, the total variation distance between the measures of extrinsic and intrinsic prior*

$$||\pi_{0,\mathcal{D}}(\theta), \tilde{\pi}_{0,\mathcal{D}}(\theta)||_{TV} = 1 - \frac{\lim_{\mathcal{D}^+ \supset \mathcal{D}} \int_{\theta \in \mathcal{D}^+} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta; \mathcal{D}) d\theta}{M_2} = \frac{\lim_{\mathcal{D}^+ \supset \mathcal{D}} \int_{\theta \in \mathcal{R} \setminus \mathcal{D}^+} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta; \mathcal{D}) d\theta}{M_2}$$

.

3 Posterior Computation

Letting the likelihood function be $L(y; \theta)$, the posterior distribution of θ under a extrinsic prior is

$$\tilde{\pi}_{\mathcal{D}}(\theta \mid y) = \frac{L(y; \theta)\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta; \mathcal{D})}{\int_{\mathcal{R}} L(y; \theta)\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta; \mathcal{D}) d\theta}, \tag{6}$$

which we refer as extrinsic posterior from now on. As it is supported on a less restrictive space $\mathcal{R}$, one can exploit conventional sampling approach such as slice sampling, adaptive Metropolis-Hastings and Hamiltonian Monte Carlo (HMC) for posterior sampling. In this section, we focus on HMC for its easiness to use and good performance in sampling with high-dimensional parameter.

3.1 Hamiltonian Monte Carlo for Extrinsic Posterior Sampling

We first provide a short review of Hamiltonian Monte Carlo for continuous θ, although discrete extension is possible (Zhang et al., 2012; Nishimura et al., 2017).

Assuming $\theta \in \mathcal{R}$, with $\mathcal{R}$ in a full or truncated Eucledean space $\mathbb{R}^d$. HMC augments a latent variable "momentum" $p \in \mathbb{R}^d$ (commonly generated from $\text{No}(0, \Sigma)$ with Σ pre-specified), ommiting constant, the negative log-posterior function based on (3) is

$$\begin{aligned} H(\theta, p) &= U(\theta) + M(p), \\ \text{where } U(\theta) &= -\log\left\{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta; \mathcal{D})\right\}, \\ M(p) &= \frac{p'\Sigma^{-1}p}{2}, \end{aligned} \tag{7}$$

Denoting current state as $(\theta^{(0)}, p^{(0)})$, HMC updates θ and p via Hamiltonian dynamics, defined by two partial differential equations:

$$\begin{aligned} \frac{\partial \theta(t)}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = \Sigma^{-1}p, \\ \frac{\partial p(t)}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \end{aligned} \tag{8}$$

By evolving in $t > 0$, this yields $(\theta^{(t)}, p^{(t)})$. Since Hamiltonian system is symplectic, i.e. $H(\theta^{(t)}, p^{(t)}) = H(\theta^{(0)}, p^{(0)})$, one can take $\theta^{(t)}$ as a new posterior sample. However, in most cases, (8) lacks closed-form solution, one has to use an *integrator* that numerically approximates an evolution of the exact solution. With the integrator reversible and volume-preserving (see (Neal, 2011) for details), an Metropolis-Hastings (M-H) step is taken to correct the approximation error, by accepting $(\theta^{(t)}, p^{(t)})$ with probability

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

One common integrator is the leap-frog algorithm (Neal, 2011)that utilizes move $(\theta^{(T\epsilon)}, p^{(T\epsilon)}) \rightarrow (\theta^{((T+1)\epsilon)}, p^{((T+1)\epsilon)})$:

$$p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta}, \quad \theta \leftarrow \theta + \epsilon \Sigma^{-1} p, \quad p \leftarrow p - \frac{\epsilon}{2}\frac{\partial U}{\partial \theta} \tag{9}$$

for $T = 0, \ldots, (L-1)$; where L is the number of leap-frog steps within one iteration and $t = L\epsilon$.

3.2 Optimizing Computing Efficiency

As a Markov chain Monte Carlo, one would hope to use HMC to build a chain that rapidly converges. The convergence rate is associated with the maximal correlation (Liu, 2008), $\gamma(\theta, \theta^*) = \sup_{g \in L^2(\Pi)} \operatorname{corr}(g(\theta), g(\theta^*))$, where $L^2(\Pi) = \{g(\theta) : \operatorname{var}(g(\theta)) < \infty\}$. Smaller $\gamma(\theta, \theta^*)$ corresponds to less correlation and faster convergence. At fixed ϵ, one can adaptively choosing L so that $\gamma(\theta, \theta^*)$ falls under certain desired rate (Hoffman and Gelman, 2014).

We now further examine this via a view of computing efficiency. Since each iterator step is often the bottleneck, one would use as few steps as possible to reach below the desired convergence rate. Given a desired convergence rate γ^*, an optimal ϵ would be:

$$\epsilon^* = \operatorname*{arg\,inf}_{\epsilon:\epsilon \leq \epsilon_{max}} \inf_L \{L : \gamma\left(\theta^{(0)}, \theta^{(\epsilon L)}\right) \leq \gamma^*\},$$

where ϵ_{max} is the stability bound for the integrator step size. Given fixed integrator time ϵL, to minimize L, often the optimality occurs when $\epsilon \approx \epsilon_{max}$.

The stability bound ϵ_{max} can be influenced by the specification in extrinsic prior. The stability bound is roughly determined by the width of support in the most constrained direction (Neal, 2011). Formally, letting $\mathcal{S}_\epsilon$ be the contour region of the support $\mathcal{S}_\epsilon = \{x : \pi(x) \in (0, \epsilon)\}$ with $\pi(x)$ as the density, the minimum support width is:

$$w_\pi = \inf_{x_1 \in \mathcal{S}_\epsilon} \sup_{x_2 \in \mathcal{S}_\epsilon} \{d(x_1, x_2) : \pi(\alpha x_1 + (1-\alpha) x_2) > 0 \text{ for } \alpha \in [0, 1]\}$$

where $d(.,.)$ is a metric, and is Eucledean distance in continous HMC.

To provide an intuition for why ϵ_{max} depends on w_π, we focus on one-step update $L = 1$. Each update in leap-frog algorithm corresponds to $\theta^{(\epsilon)} = \theta^{(0)} + \epsilon p^{(0)} - \epsilon^2/2\frac{\partial U}{\partial \theta} = \theta^{(0)} + \varepsilon p^{(0)} + O(\epsilon^2)$. If $\epsilon \gg w_\pi$, a random move in $\varepsilon p(0)$ can end outside the support with $U(\theta^{(\epsilon)}) = \infty$. This would violate the condition that discrete integrator approximately preserves $U(\theta^{(\varepsilon)}) + M(p^{(\varepsilon)}) = U(\theta^{(0)}) + M(p^{(0)}) < \infty$.

Therefore to efficiently utilize HMC, if the constrained space $\mathcal{D}$ has narrow support via certain direction in its embedded space $\mathcal{R}$, one needs to create more relaxation in extrinsic prior. For example, one equality constraint $x+y+z = 1$ in $\mathbb{R}^3$ creates a hyperplane. Extrinsic prior with $\exp(-\frac{|x+y+z-1|}{\lambda})$ relaxes the support along the normal vector of the plane, creating a slab with some thickness greater than 0. To have decent ϵ_{max}, one needs to avoid choosing λ too small. Empirically, we found $\lambda = 10^{-3}$ is often a good value for such constraint.

4 Examples and Application

We now use examples to illustrate the properties of extrisinc priors and their utility in complex real data application.

4.1 Simulations

Example 1: Unit Circle

When the constrained support in $\mathcal{D}$ has small minimum support width w_π, it limits the stability bound. Therefore it is important to allow some support expansion to the corresponding constraint. We first illustrate such case via estimation on a unit circle. The circle has $w_\pi = 0$ and therefore needs some support expansion to run HMC.

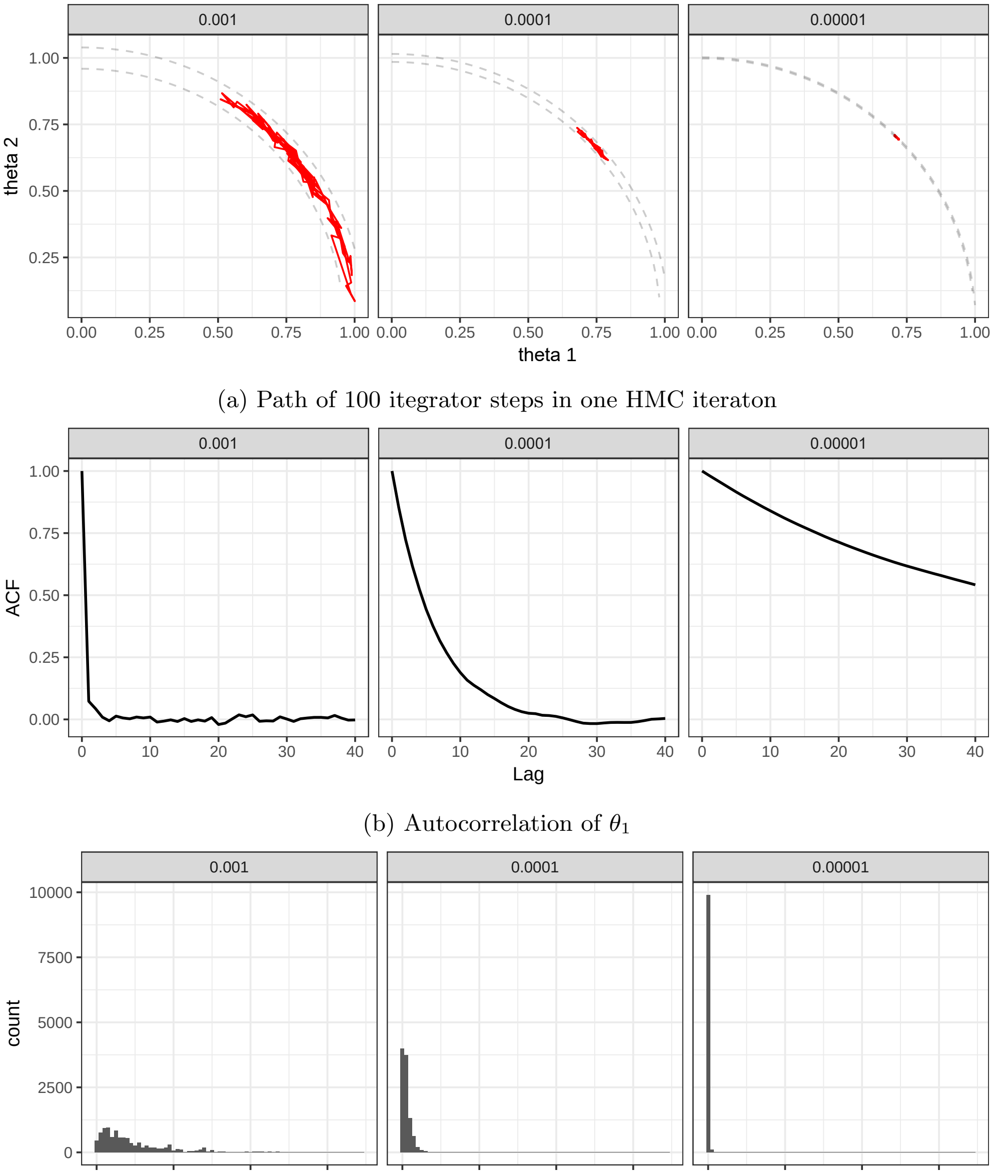
Consider data $y_i \in \mathbb{R}^2$ for $i = 1, \ldots, n$ that are noisy realization from one point on unit circle:

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

where $\theta \in \mathcal{V}(2, 1)$, a $(2, 1)$–Stiefel manifold, is assigned a von Mises–Fisher prior $\pi_{0,\mathcal{D}}(\theta) \propto \exp(F'\theta_i)$.

To generate data, we use $\theta = (\sqrt{3}/2, 1/2)$, $\sigma^2 = 0.5^2$ and small $n = 5$, in order to induce widely spread-out posterior θ on the manifold. For posterior sampling, we use $F = (1, 1)$ to induce a weakly informative prior for θ and an inverse-Gamma prior IG$(2, 1)$ for σ^2 . To allow extrinsic posterior sampling, we use $v(\theta) = |\theta'\theta - 1|$ as the distance to constrained space and extrinsic prior $\tilde{\pi}_{0,\mathcal{D}}(\theta) = \exp(F'\theta_i)\exp(-\frac{|\theta'\theta-1|}{\lambda})$.

We test three different tuning parameters $\lambda = 0.001$, 0.0001 and 0.00001. We run each algorithm for $20,000$ steps, with first $10,000$ discarded. To compare the computing efficiency, we restrict the maximum leap-frog steps L to be 100 and visualize how much space each algorithm can explore within one HMC iteration. Figure 2(a) plots the path of $L = 100$ leap-frog steps. Larger λ leads to wider support expansion and larger stability bound ϵ_{max}. This makes $\theta^{(\epsilon L)}$ much less correlated with $\theta^{(0)}$ at each iteration, measured by autocorrelation (ACF) based on the posterior sample of θ_1 (Figure 2(b)). The posterior distribution of the distance $v(\theta)$ is small for all three settings (Figure 2(c)), with smaller λ associated with smaller $v(\theta)$



(a) Path of 100 itegrator steps in one HMC iteraton

(b) Autocorrelation of θ_1

(c) Posterior distribution of $|\theta'\theta - 1|$

Figure 2: HMC Sampling on a unit circle, using extrinc prior with $\mathcal{K}(\theta) = \exp(-\frac{|\theta'\theta-1|}{\lambda})$, with $\lambda = 0.001$, 0.0001 and 0.00001. Panel (a) shows the larger relaxation in the narrowest direction of support (orthogonal vector to the circle) can result in more efficient space exploration within 100 leap-frog steps; panel (b) shows the autocorrelation of the posterior sample; panel (c) shows the posterior distribution of the distance to the constraint.

Example 2: Linear Regression Under Inequality Constraint

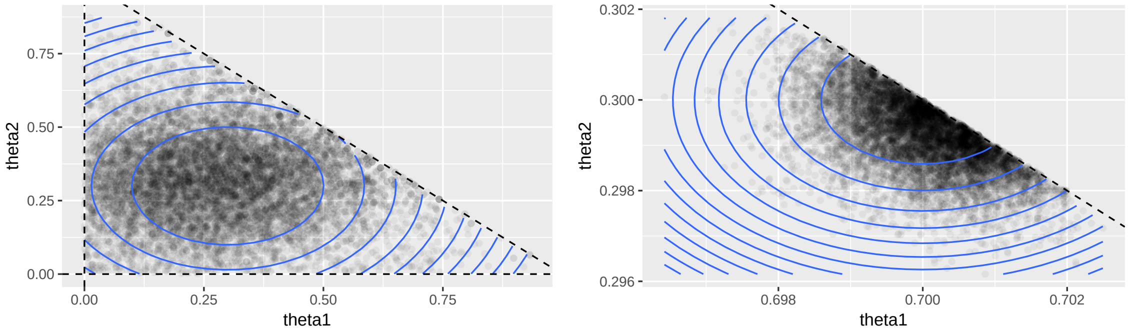
If the support on constrained space $\mathcal{D}$ does not have small minimum support width w_π, one can use extrinsic prior with almost no relaxation. This applies a large class of inequality constraints that has wide support. To illustrate, consider a linear regression

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

where both θ is a p-dimensional vector; d many linear equalities are defined by $d \times p$ matrix A and a d-dimensional vector c. The inequalities form one or multiple polyhedrons in $\mathbb{R}^p$, with $\mathcal{D}$ as the interior or exterior space.

We consider simple bivariate case $\theta \in (0,1)^2$ subject to $\theta_1 + \theta_2 \le 1$, making $\mathcal{D}$ a triangle. To simulate data, we use $\sigma^2 = 0.1^2$, $x_i \sim \text{No}([0,0]', I)$ for $i = 1, \ldots, n$. We then generate two datasets with two different sets of (θ, n). In the first set, we use $\theta = [0.3, 0.3]'$ with $n = 10$, so that the posterior has wide spread in the constrained space; in the second set, we use $\theta = [0.7, 0.3]'$ with $n = 10^4$ so that the posterior is concentrated near the boundary. In both cases, we assign weakly informative prior for $\theta \sim \text{No}([0.5, 0.5]', I10^2)$ and $\sigma^2 \sim IG(2, 1)$.

Unlike the unit circle example, the constrained space $\mathcal{D}$ has large w_π. This allows us to choose $\lambda = 10^{-8}$ for $\mathcal{K}(\theta) = \exp(-\frac{|\theta_1+\theta_2-1|}{\lambda})$ in the extrinsic prior, leading to almost no constraint relaxation. We collect 10,000 posterior samples efficiently via HMC. Figure 3 plots the posterior sample and its contour. There is no posterior outside $\mathcal{D}$ due to the small λ.



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

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

Figure 3: Extrinsic posterior distribution of the normal mean θ, with approximation to constraint $\theta_1 + \theta_2 \le 1$. 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.

4.2 Applications

We now illustrate the utilitiy of extrinsic prior in two applications.

Application 1: Ordered Dirichlet Prior

We first use ordered simplex in finite mixture model. A $(J-1)$–simplex is a vector $w = \{w_1, \ldots w_J\}$ with $1 > w_1 \ge \ldots \ge w_J > 0$ and $\sum_{j=1}^J w_j = 1$.

For probablity simplex, standard practice assigns Dirichlet prior $Dir(\alpha)$, with $\pi_{0,\mathcal{D}}(w) = \prod_{j=1}^J w_j^{\alpha-1} \mathbb{1}_{\sum_{j=1}^J w_j=1}$. However, this does not accomodate ordering; therefore, the index j is exchangeable and permutating j's does not change the density. This commonly leads to label-switching problem in mixture model estimation (reviewed in Jasra et al. (2005)).

Imposing order constraint yields an ordered Dirichlet prior:

$$\pi_{0,\mathcal{D}}(w_1, \dots w_J) = \prod_{j=1}^J w_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^J w_j=1} \cdot \prod_{j=1}^{J-1} \mathbb{1}_{w_j \ge w_{j+1}} \cdot \tag{10}$$

where $w_j \in (0, 1)$ for $j = 1, \dots, J$. The ordered Dirichlet prior can be approximated by extrinsic prior:

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

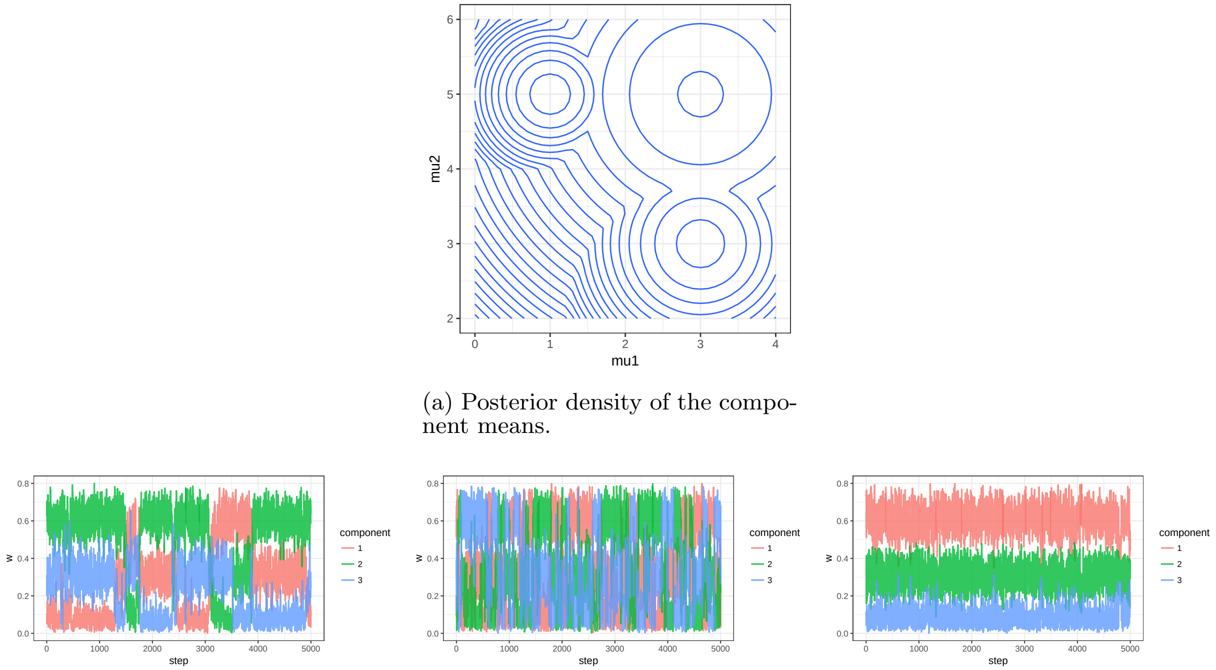
We now adopt this simplex distribution in a normal mixture model with mixture means and common variance, for data $y_i \in \mathbb{R}^d$ indexed by $i = 1, \dots, n$:

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

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 $\lambda_2 = 10^{-3}$ to allow efficient mixing in embedding a simplex in $(0, 1)^J$. To illustrate the benefit of ordered Dirichlet, we also test Gibbs sampling and extrinsic prior method on canonical Dirichlet prior without order constraint.

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, generting 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 clearly better convergence due to the 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).

Application 2: Orthonormal Tucker 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 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 large $V(V-1)/2 = 2,278$ compared to sample size n, with potential observational error in recording connectivity. The diagonal in each A_i is missing due to the lack of meaning in self-connectivity in brain network. These facts merit using a low-rank Bayesian approach.

We consider a symmetric Bernoulli Tucker decomposition model:

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

$$\psi_{i,k,l} = \sum_{r_3=1}^{d1} \sum_{r_2=1}^{d1} \sum_{r_1=1}^{d3} D_{r_1,r_2,r_3} W_{i,r_1} U_{k,r_2} U_{l,r_3}$$

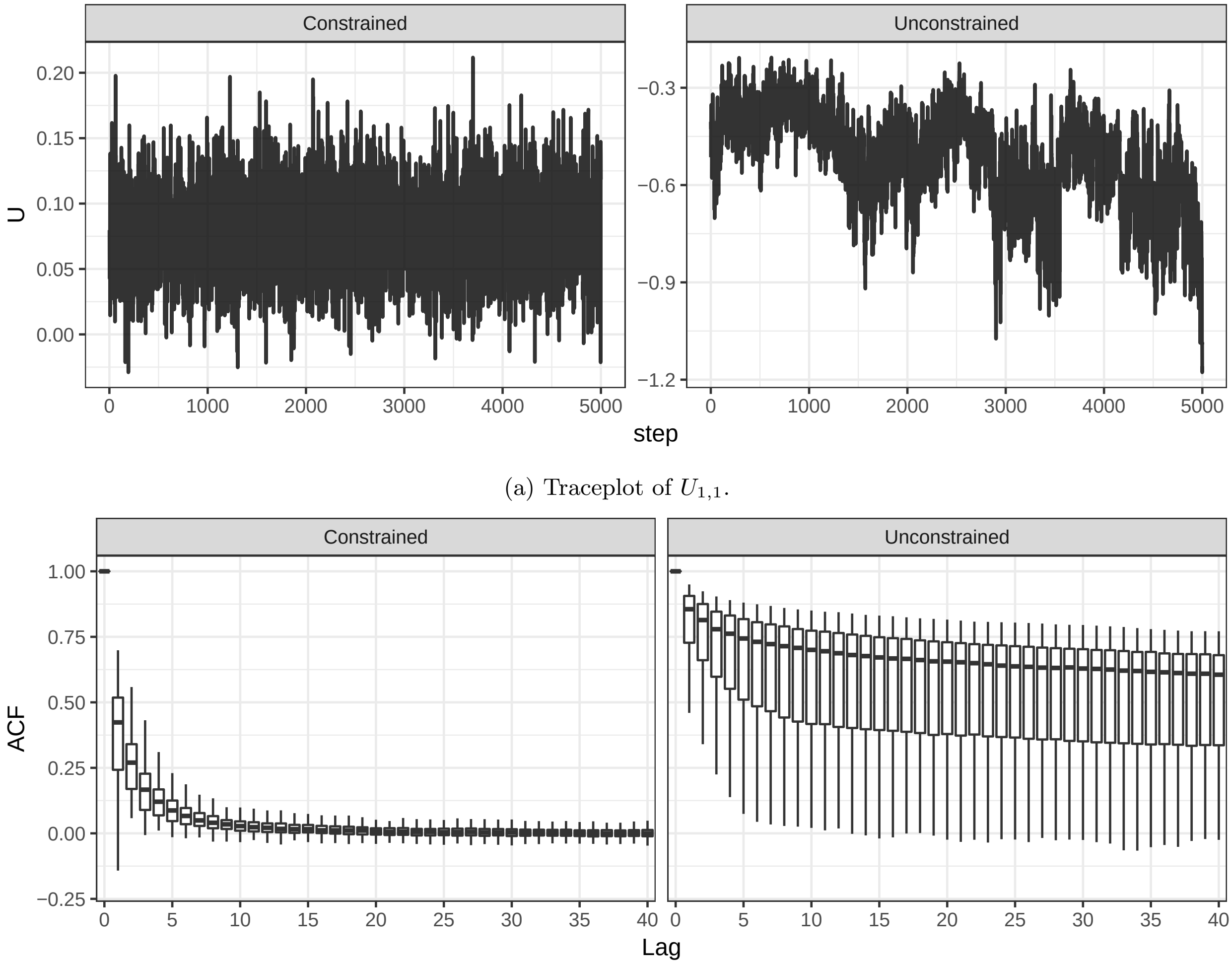
for $k > l$, $k = 2, \dots, V$, $i = 1, \dots, n$; U is $V \times d_1$ matrix, W is $n \times d_2$ matrix. The symmetry of each network is induced by using U twice in the decomposition. The core tensor D is a $d_1 \times d_1 \times d_3$ array, whose entries correpond to the interaction between different factor; $D_{r_1,k,l} = D_{r_1,l,k}$ for all r_1, k, l. 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).

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 $d_1 = d_3$ and $D_{r_1,r_2,r_3} = 0$ unless $r_1 = r_2 = r_3$. The additional flexibility in the Tucker is appealing, as one would 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 to rotation and scaling. For example, one can multiply a $d_1 \times d_1$ orthonormal matrix R, to U and obtain $U^* = UR$ and each slice of $D_{r_1,\cdot,\cdot}$ to obtain $D^*_{r_1,\cdot,\cdot} = R' D_{r_1,\cdot,\cdot} R$. This leaves the likelihood unchanged.

Therefore, we consider applying some constraint on the Tucker decomposition, while still maintaining its varying rank property over different modes. Motivated by high-order singular value decomposition, we first impose orthonormality constraints $U'U = I_{d_1}$ and $W'W = I_{d_3}$ and diagonal strucutre for all $D_{r_1,\cdot,\cdot}$, with $D_{r_1,r_2,r_3} = 0$ unless $r_2 = r_3$.

We assign normal prior for $U_{k,r_2} \sim \text{No}(0, \phi_1)$, $W_{i,r_1} \sim \text{No}(0, \phi_2)$, $Z_{k,l} \sim \text{No}(0, \phi_3)$, $D_{r_1,r_2,r_2} \sim No(0, \phi_{4,r_1,r_2})$ for all i, k, l, r_1, r_2, and inverse-Gamma prior $\phi_1, \phi_2, \phi_3 \overset{indep}{\sim} \text{IG}(2, 1)$. For ϕ_{4,r_1,r_2}, we assign multiplicative inverse gamma distribution (Bhattacharya et al., 2011) $\phi_{4,r_1,r_2} = \prod_{m_1=1}^{r_1} \nu_{1,m_1} \prod_{m_2=1}^{r_2} \nu_{2,m_2}$ with $\nu_{1,1}, \nu_{2,1} \overset{indep}{\sim} \text{IG}(a_1, 1)$ and $\nu_{1,m}, \nu_{2,m} \overset{indep}{\sim} \text{IG}(a_2, 1)$ for $m \geq 2$. This induces increasing concentration towards zero as r_1, r_2 increase, allowing adaptive choosing of latent dimensions. We set $a_1 = a_2 = 5$ in this application.

To allow estimation for model with orthonormality constraint, we use extrinsic prior with $\mathcal{K}(\theta) = \exp(-\frac{|U'U - I_{d_1}| + |W'W - I_{d_2}|}{\lambda})$ and set $\lambda = 10^{-3}$. To compare, we also test with the same model configuration without the orthonormality constraint, and refer to it as unconstrained model. 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 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: 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).

Ignoring those D_{r_1,r_2,r_3} corresponding to $\phi_{4,r_1,r_2} < 10^{-3}$, the multiplicative prior results in effective $\tilde{d}_1 = 5$ and $\tilde{d}_3 = 3$ dimensions.

5 Discussion

References

Bhattacharya, A., D. B. Dunson, et al. (2011). Sparse bayesian infinite factor models. *Biometrika 98*(2), 291.

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.

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).

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.

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.

Hoffman, M. D. and A. Gelman (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

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.

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

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

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

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.

Zhang, Y., Z. Ghahramani, A. J. Storkey, and C. A. Sutton (2012). Continuous relaxations for discrete hamiltonian monte carlo. In *Advances in Neural Information Processing Systems*, pp. 3194–3202.