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 by 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 $K_k(v) = \exp(-v/\lambda_k)$, with $\lambda_k > 0$ as the tuning parameter.

To illustrate, 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 $v(\theta)$: $(\theta - 5)_+$, $(\theta - 5)_+^2$, where $(x)_+ = \begin{cases} 0 \text{ if } x \leq 0 \\ x \text{ if } x > 0 \end{cases}$. 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, the first-order $(\theta - 5)_+$ drops faster than second-order $(\theta - 5)_+^2$ when $(\theta - 5)_+ < 1$.

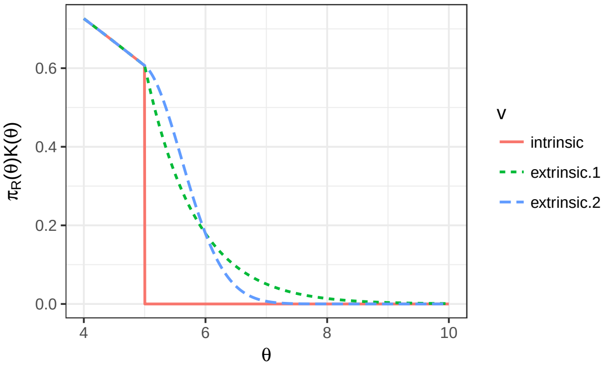


Figure 1: Unnormalized densities for truncated normal $\text{No}_{(-\infty,5)}(0, 5^2)$ under exact intrinsic prior and approximating extrinsic prior. Inside $(-\infty, 5)$, the priors are the same up to a constant difference. The intrinsic prior abruptly drops to 0 on the boundary, while the approximating 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 $\int_\mathcal{D} \pi_{0,\mathcal{R}}(\theta)d\theta > 0$.

Theorem 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*

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

.

proof: via definition of total variation and $K(\theta; \mathcal{D}) = 1$ and $\theta \in \mathcal{D}$

Taking the special 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 DCT

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:

Theorem 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 $\pi_{0,\mathcal{R}}/M_3$. If $f(v) < \infty$ when $v < t$,*

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

proof:

$$\begin{aligned} \int_0^\infty f(v) \exp(-\frac{v}{\lambda}) dv =& \int_0^t f(v) \exp(-\frac{v}{\lambda}) dv + \int_t^\infty \frac{f(v)}{M_3} \exp(-\frac{v}{\lambda}) dv \\ \le& F(t) \exp(-\frac{t}{\lambda}) + \frac{1}{\lambda} \int_0^t F(v) \exp(-\frac{v}{\lambda}) dv + \exp(-\frac{t}{\lambda}) \\ =& (F(t)+1) \exp(-\frac{t}{\lambda}) + \frac{1}{\lambda} \int_0^t f(v^*) v \exp(-\frac{v}{\lambda}) dv \\ \le& (F(t)+1) \exp(-\frac{t}{\lambda}) + \sup_{t^* \in (0,t)} f(t^*) \int_0^t \frac{1}{\lambda} v \exp(-\frac{v}{\lambda}) dv \\ \le& 2 \exp(-\frac{t}{\lambda}) + \sup_{t^* \in (0,t)} f(t^*) \lambda \end{aligned} \quad (5)$$

where 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 $O(\lambda)$.

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

Theorem 3. *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*

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

.

2.4 Selection of Kernel Hyper-Parameters

The hyper-parameters in the kernel control the upper bound of the constraint violation v_k and how fast it approaches zero. In this section, we describe the strategy in selecting those parameters.

As $\lambda_k \to \infty$ in each $K_k(\theta)$, $\mathcal{K}(\theta; \mathcal{D}) \to \mathbb{1}_{\theta \in \mathcal{D}}$, the distance between (**??**) and (**??**) decreases to zero.

2.5 Control of Constraint Relaxation

We first obtain a control of constraint relaxation, in terms of the posterior values of $|E_k(\theta)|$ and $(G_{k'}(\theta))_+$. Letting v represents their values, the control can be achieved via a bounded prior support near 0 for each kernel $\int_{v<\varepsilon} \mathcal{C}_{i,k}(v)dv = 1$, with $\mathcal{C}_{i,k}(v) = K_{i,k}(v)/\int_{\mathcal{R}} K_{i,k}(v)dv$. The pre-specified constant ε represents the element-wise tolerance for violating each constraint. The bounded prior support allows us to theoretically control the posterior approximation error. With $\mathcal{K}(\theta) \propto \prod_{i,k} \mathcal{C}_{i,k}(x)$ is the joint extrinsic prior density, since $\pi_{\mathcal{K}}(\theta \mid y) \ll \mathcal{K}(\theta)$, the posterior for each constraint relaxation is bounded in $[0, \varepsilon)$ with probability 1. ¡¡¡¡¡¡¡ HEAD

In practice, one may wish to utilize a kernel $K^*_{i,k}(x)$, orginally with unbounded support on $[0, \infty)$ for computing conveniences. To adapt them for bounded support in the relaxation v, one can first choose $\lambda_{i,k}$ to have $\int_{v<\varepsilon} \mathcal{K}^*_{i,k}(x)/\big(\int_{\mathcal{R}} K^*_{\cdot}(v)dv\big) = 1 - \eta$ with η small, then apply truncation $K_{i,k}(v) = K^*_{i,k}(v)\mathbb{1}_{v<\varepsilon}$ to induce $v < \varepsilon$ almost surely. In most cases, the truncation is only nominal for a theoretic guarantee; in computation it is rarely used. For example, in Gaussian kernel $\exp(-\lambda x^2)$ assigns $x < 4/\sqrt{2\lambda}$ with probability 0.99993 apriori; for posterior sampling, one can first do an untruncated sampling, then reject those $x > \varepsilon = 4/\sqrt{2\lambda}$, which is quite rare due to the small prior probability.

To illustrate the control of constraint relaxation, we assume a simple scenario of generating posterior from a truncated Gaussian distribution $\theta \mid y \sim \text{No}_{(\alpha,\beta)}(0, 1)$, with mean 0 and variance 1 and truncation $\theta \in (\alpha, \beta)$. The exact and extrinsic posterior densities are:

=======

In practice, one may wish to utilize a kernel $K^*_{i,k}(x)$, orginally with unbounded support on $[0, \infty)$ for computing conveniences. To adapt them for bounded support in the relaxation v, one can first choose $\lambda_{i,k}$ to have $\int_{v<\varepsilon} \mathcal{K}^*_{i,k}(x)/\big(\int_{\mathcal{R}} K^*_{\cdot}(v)dv\big) = 1 - \eta$ with η small, then apply truncation $K_{i,k}(v) = K^*_{i,k}(v)\mathbb{1}_{v<\varepsilon}$ to induce $v < \varepsilon$ almost surely. In most cases, the truncation is only nominal for a theoretic guarantee; in computation it is rarely used. For example, in Gaussian kernel $\exp(-\lambda x^2)$ assigns $x < 4/\sqrt{2\lambda}$ with probability 0.99993 apriori; for posterior sampling, one can first do an untruncated sampling, then reject those $x > \varepsilon = 4/\sqrt{2\lambda}$, which is quite rare due to the small prior probability.

To illustrate the control of constraint relaxation, we assume a simple scenario of generating posterior from a truncated Gaussian distribution $\theta \mid y \sim \text{No}_{(\alpha,\beta)}(0, 1)$, with mean 0 and variance 1 and truncation $\theta \in (\alpha, \beta)$. The exact and extrinsic posterior densities are:

$$\pi(\theta \mid y) \propto \exp(-\frac{\theta^2}{2})\mathbb{1}_{\theta\in(\alpha,\beta)}, \quad \pi_{\mathcal{K}}(\theta \mid y) \propto \exp(-\frac{\theta^2}{2})K\left((\alpha - \theta)_+\right)K\left((\theta - \beta)_+\right).$$

with $K(x) = \exp(-\lambda x^2)\mathbb{1}_{x<4/\sqrt{2\lambda}}$. We set $(\alpha, \beta) = (1, 2)$. Figure 1 plots the unnormalized densities under the exact and extrinsic posteriors with different λ's. The extrinsic posterior densities inside $\mathcal{D} = (1, 2)$ are the same as the exact one, up to a constant difference due to normalization. Outside $\mathcal{D}$, the larger λ leads to more rapid decline of density and therefore smaller constraint relaxation $(\alpha - \theta)_+$ and $(\theta - \beta)_+$.

It is temping to always induce almost 0 relaxation with very large λ, however, in heavily constrained models such as the ones with equality constraint, the narrow distribution width in $\mathcal{R}$ will cause a adverse effect in some popular algorithms such as Hamiltonian Monte Carlo. In those cases, it is rather useful to have a slightly larger relaxation, then use projection to correct the imperfection. We will illustrate this in the next section. ¡¡¡¡¡¡¡ HEAD

2.6 Hamiltonian Monte Carlo for Extrinsic Posterior Sampling

Extrinsic posterior has support on a less restrictive space $\mathcal{R}$, where conventional sampling approach such as slice sampling, adaptive Metropolis-Hastings and Hamiltonian Monte Carlo (HMC) can be adopted easily. In this paper, we focus on estimation via HMC for its high-level automation aided by software and often good performance due to various adaptive algorithms (Hoffman and Gelman, 2014). To be clear, this is different from Riemannian HMC that requires specific accommodation and heavy computation. The algorithm we use is simply conventional HMC in Euclidean space. In this section, we study the effects of choosing λ on efficiency of Hamiltonian dynamics.

2.7 Hamiltonian Monte Carlo for Extrinsic Posterior Sampling

Extrinsic posterior has support on a less restrictive space $\mathcal{R}$, where conventional sampling approach such as slice sampling, adaptive Metropolis-Hastings and Hamiltonian Monte Carlo (HMC) can be adopted easily. In this paper, we focus on estimation via HMC for its high-level automation aided by software and often good performance due to various adaptive algorithms (Hoffman and Gelman, 2014). To be clear, this is different from Riemannian HMC that requires specific accommodation and heavy computation. The algorithm we use is simply conventional HMC in Euclidean space. In this section, we study the effects of choosing λ on efficiency of Hamiltonian dynamics.

We assume θ is d-dimensional, $\mathcal{R}$ is a full or truncated Euchledean space in $\mathbb{R}^d$, and the constraint functions $E_k(\theta)$'s and $G_k(\theta)$'s are differentiable with respect to θ. We focus on the case where θ is continuous, although discrete extension is possible (Zhang et al., 2012). HMC augments a latent variable named "veolicty" or "momentum" $p \in \mathbb{R}^d$, the negative log-posterior function based on (3) is

$$
\begin{aligned}
H(\theta, p) &= U(\theta) + M(p), \\
\text{where } U(\theta) &= -\log\left\{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)\right\}, \\
M(p) &= \frac{p'\Sigma^{-1}p}{2},
\end{aligned} \tag{6}
$$

with Σ^{-1} a pre-specified positive definite matrix. Instead of taking random walk or Gibbs updating, HMC then update θ and p via Hamiltonian dynamics, satisfying differential equations:

$$
\begin{aligned}
\frac{\partial\theta(t)}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = \Sigma^{-1}p, \\
\frac{\partial p(t)}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}.
\end{aligned} \tag{7}
$$

At the start of each iteration, the current state of θ is viewed as $\theta(0)$ and $p(0)$ randomly generated from $\text{No}(0, \Sigma)$. The solution to (7) yields $\theta(t)$ and $-p(t)$ as the new state. Since Hamiltonian system is symplectic, the negative log-posterior function is unchanged $H(\theta(t), p(t)) = H(\theta(0), p(0))$. However, in most cases, (7) lacks closed-form solution, one has to use discrete approximation, commonly leap-frog algorithm (Neal, 2011):

$$
\begin{aligned}
p(T + \varepsilon/2) &= p(T) - \varepsilon/2\frac{\partial U}{\partial \theta}(\theta(T)), \\
\theta(T + \varepsilon) &= \theta(T) + \varepsilon\Sigma^{-1}p(T + \varepsilon/2), \\
p(T + \varepsilon) &= p(T + \varepsilon/2) - \varepsilon/2\frac{\partial U}{\partial \theta}(\theta(T + \varepsilon)),
\end{aligned} \tag{8}
$$

for $T = 0, \varepsilon, 2\varepsilon, \ldots, (L-1)\varepsilon$, with ε known as the time step, and L as the total leap-frog steps within one iteration. The sequence of $\{(p(T), \theta(T))\}_T$ form a trajectory of length $L + 1$ in the space of $\mathbb{R}^{2d}$. Since this approximating update is deterministic and reversible, an Metropolis-Hastings (M-H) step can be taken at the end to accept $\theta(t)$ and $p(t)$ with probability

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

with $t = L\varepsilon$.

Since extrinsic prior replaces the constraint indicator $\mathbb{1}_{\theta\in\mathcal{D}}$ with a continuous function, conventional HMC can be directly run in space $\mathcal{R}$. In HMC, finding optimal time step ε is important. There exists a stability bound for ϵ. When ε is larger than this bound, H diverges and grows exponentially with L, leading to very low acceptance rate in M-H step. When ε is too small, each time step can only generate local update hence low computing efficiency. Since most systems involve nonlinear transition, analytical bound is not available, but one can empirically optimize ε to be close to this bound. This can be achieved via tuning for acceptance rate in the Metropolis-Hastings step. Specifically, given fixed L, one tunes ε so that the acceptance rate is close to but slightly below 1. Despite the technicality, the tuning of ε is implemented in the mature HMC software such as STAN. We instead focus on how λ can affect the stability bound itself.

For multiple-dimensional θ with $\Sigma = I$, the stability bound is roughly determined by the width of distribution in the most constrained direction (Neal, 2011). To provide an intuition, we focus on one time step update $L = 1$. Each update in leap-frog algorithm corresponds to $\theta(\varepsilon) = \theta(0) + \varepsilon p(0) - \varepsilon^2/2 \frac{\partial U}{\partial \theta}(\theta(0)) = \theta(0) + \varepsilon p(0) + O(\varepsilon^2)$. When the support in extrinsic posterior is narrow along certain direction, an move in $\varepsilon p(0)$ can end in region with posterior density 0 (associated with infinite $U(\theta(t))$). This is because we do not constrain $p(0)$, so that it is randomly generated in all direction of $\mathbb{R}^d$. On the other hand, a stable trajectory should approximately preserve $U(\theta(\varepsilon)) + M(p(\varepsilon)) = U(\theta(0)) + M(p(0))$, since $M(p) = p'p/2 \geq 0$, $U(\theta(\varepsilon)) \leq U(\theta(0)) + M(p(0))$. With initial velocity $p(0) \sim N(0, I)$ and finite $U(\theta(0))$, a stable trajectory should never move to region outside of support. Therefore, the stability bound on ε is indeed impacted by the smallest width of posterior support.

Therefore in extrinsic prior, it is important to avoid creating a support too narrow. This could be possible with strong constraints like equality. When embedded in larger space, the approximate hyper-plane specified by equality extrinsic prior has its narrowest width as the amount of relaxation from strict equality. In such cases, very large λ would force small stability bound on ε, creating computing bottleneck; instead, it is more efficient to use smaller λ to induce more relaxation. On the other hand, inequality constraints often do not have this issue, as long as these inequalities do not induce narrow support. Therefore, one can often use large λ in inequality extrinsic prior.

To illustrate, we consider generating posterior $\theta = (x_1, x_2)$ on a unit circle using von Mises–Fisher distribution, $\pi(\theta \mid y) \propto \exp(F'\theta)$ with $\theta'\theta = 1$. This is a simple example of a random variable constraint on a $(2, 1)$-Stiefel manifold $\mathcal{D} = \mathcal{V}(2, 1)$. We set $F = (1, 1)$ to induce a distribution widely spreaded over the manifold, generating great amount of uncertainty for assessing the sampling efficiency. We use extrinsic prior proportional to $K(\theta) = \exp(-\lambda(\theta'\theta - 1)^2)\mathbb{1}_{|\theta'\theta - 1| < 0.1}$. Geometrically, this prior expands the posterior support from a circle to a ring, with its width $|\theta'\theta - 1|$ affected by λ.

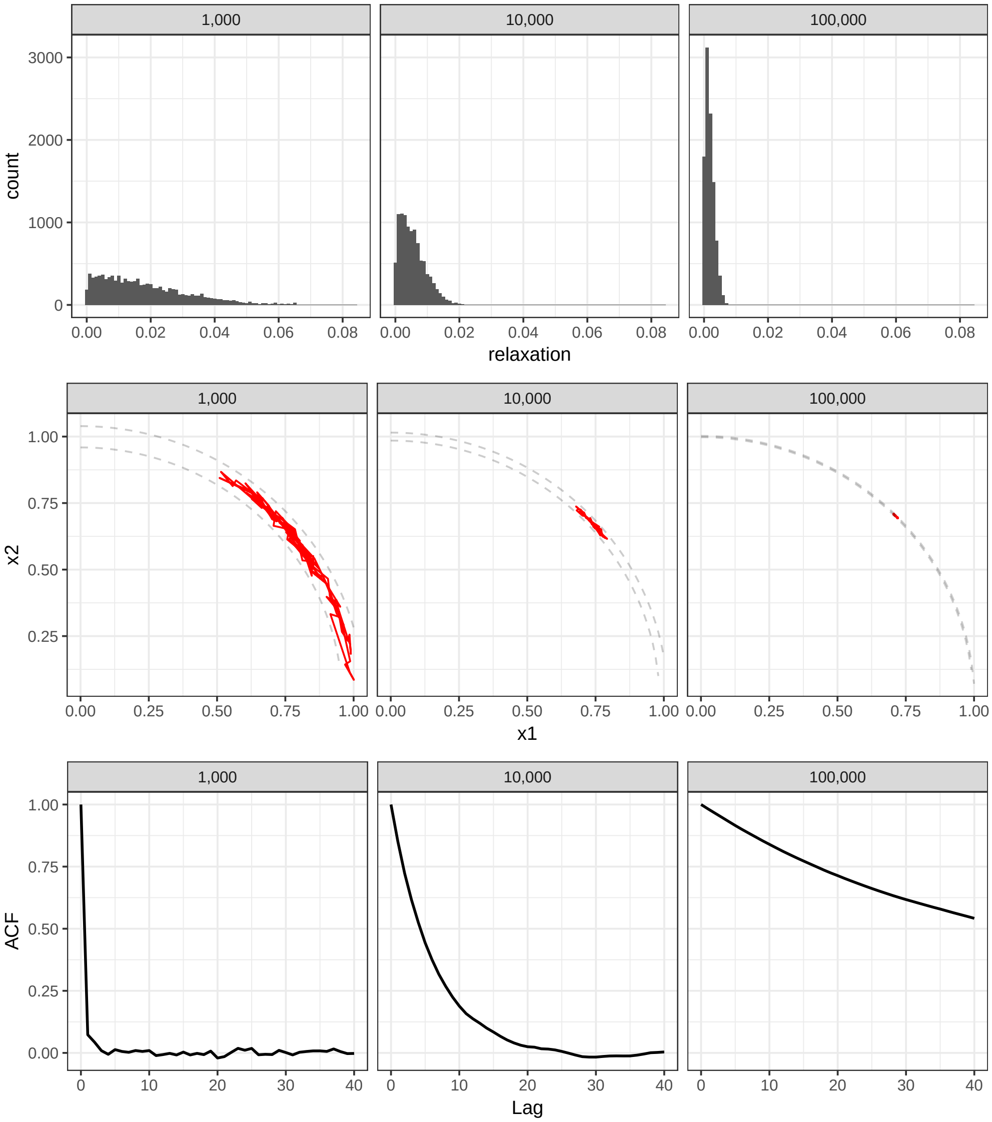


Figure 2: Sampling posterior from a von Mises–Fisher distribution on a unit circle, using HMC with extrinc prior under $\lambda = 10^3, 10^4, 10^5$. Row 1 shows the posterior distribution of the constraint relaxation $|\theta'\theta - 1|$; Row 2 shows the path of 100 leap-frog steps; Row 3 shows the autocorrelation plot (ACF). Large λ gives very small constraint relaxation, but suffers from slow mixing due to inefficient local update; smaller λ increases the relaxation but results in excellent mixing.

We tested three different values of $\lambda = 10^3, 10^4, 10^5$. For each λ, we ran HMC for $10,000$ iterations, with $L = 100$ leap-frog steps in each iteration. We set $\Sigma = \text{diag}(1,1)$ in generating velocity p. During the initial $2,000$ iterations, the leap-frog step size ε is tuned for an acceptance rate close to 0.8, then it is fixed during the remaining part of Markov chain. The last $5,000$ iterations are used as posterior samples. Figure 2 plots the posterior distribution of constraint relaxation $|\theta'\theta - 1|$, the sampling path and the autocorrelation function (ACF) for each Markov chain. Very large $\lambda = 10^5$ has much less constraint relaxation; however, due to the small ring width, the Hamiltonian dynamics has to use small ε and can only explore local space for each 100 time steps. This results in a very slow mixing (large autocorrelation even at 40 lags). On the other hand, smaller $\lambda = 10^3$ has slightly larger constraint relaxation, but allows much more efficient exploration of the space and excellent mixing performance. In general, we find that $\lambda = 10^3$ is a good empirical value for all the equality constraints used in this paper.

2.8 Soft and Hard Constraints

We now introduce two new notions "soft" and "hard" constraints. Often, some model constraints are included as an extra means to improve convergence and identifiability. For example, the ordering of parameters are often used to address multi-modality under parameter permutation. In such cases, one can allow those constraints to be slightly relaxed without obviously impacting these objectives. We refer such relaxed constraint as soft constraint; in our framework, the extrinsic prior generates a soft constraint. One obvious benefit of soft constraint is that one can directly replaces the inconvenient model constraint by soft constraint, and use extrinsic posterior for statistical inference. Another benefit is to introduce some uncertainty on some constraint, and allow the posterior to mildly violate these constraint if the data strongly suggests so. ¡¡¡¡¡¡ HEAD

On the other hand, there are some constraints that need to be upheld strictly, such the constraints embedding manifold in Euclidean space. We refer those as hard constraints. In the last example, the 2-norm constraint needs to be always met in order to have parameter on the unit circle. Under this scenario, the extrinsic posterior is an approximation to posterior under hard constraint, hence needs to be corrected to have valid inference. We now describe a simple procedure in the next section.

On the other hand, there are some constraints that need to be upheld strictly, such the constraints embedding manifold in Euclidean space. We refer those as hard constraints. In the last example, the 2-norm constraint needs to be always met in order to have parameter on the unit circle. Under this scenario, the extrinsic posterior is an approximation to posterior under hard constraint, hence needs to be corrected to have valid inference. We now describe a simple procedure in the next section.

2.9 Correcting Projection for Hard Constraint

The extrinsic posterior $\pi_{\mathcal{K}}(\theta \mid y)$ is an approximation to (**??**) under hard constraint. We now introduce a step to correct the approximation error, by projecting θ back to constrained space $\mathcal{D}$.

The Markov chain produced by HMC is geometrically ergodic under very general conditions (Livingstone et al., 2016). Letting θ^* be a random sample collected based on $\pi_{\mathcal{K}}(\theta \mid y)$, there exists deterministic projection $P : \mathcal{R} \to \mathcal{D}$ and obtain $\theta^*_{\mathcal{D}} = P(\theta^*)$. Using this as proposal machineary, one can construct another Markov chain with $\pi(\theta_{\mathcal{D}} \mid y)$ as the target distribution. Letting the current state be $\theta_{\mathcal{D}} = P(\theta)$, we generate proposal $\theta^*_{\mathcal{D}} = P(\theta^*)$ and accept it with probability:

$$1 \wedge \frac{\pi(\theta^*_{\mathcal{D}} \mid y)\pi_{\mathcal{K}}(\theta \mid y)}{\pi(\theta_{\mathcal{D}} \mid y)\pi_{\mathcal{K}}(\theta^* \mid y)} = 1 \wedge \frac{L(\theta^*_{\mathcal{D}}; y)\pi_{0,\mathcal{R}}(\theta^*_{\mathcal{D}}) \cdot L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta)}{L(\theta_{\mathcal{D}}; y)\pi_{0,\mathcal{R}}(\theta_{\mathcal{D}}) \cdot L(\theta^*; y)\pi_{0,\mathcal{R}}(\theta^*)\mathcal{K}(\theta^*)}. \tag{9}$$

This procedure converts a set of extrinsic posterior samples into a Markov chain with exact constrained posterior as the target. Simple projection often exists for common constraints and yields high acceptance rate. Because the extrinsic prior allows only very small relaxation of hard constraints, the projection gives very little change from θ to θ^*_D. In the last example of unit circle, one can project by simply normalizing each $P(\theta^*) = \theta^*/||\theta^*||_2$. Since in the extrinsic posterior $||\theta^*||_2$ is very close to 1, the change is very small. We obtained the exact chain with acceptance rate of 0.98.

3 Theory

4 Examples and Application

In this section, we demonstrate the utility of extrinsic prior via three examples.

Example 1: Ordered Dirichlet Prior in Mixture Model

We first consider a simplex modeling problem, where a $(J-1)$–simplex $w = \{w_1, \dots w_J\}$ has all $w_j \in (0,1)$ and $\sum_{j=1}^J w_j = 1$. We illustrate its use via a normal mixture model with mixture means and common variance, for data $y_i \in \mathbb{R}^d$ indexed by $i = 1, \dots, n$:

$$y_i \overset{indep}{\sim} \text{No}(\mu_i, \Sigma),$$

$$\mu_i \overset{iid}{\sim} G,$$

$$G(.) = \sum_{j=1}^J w_j \delta_{\mu_j}(.),$$

which is associated with likelihood

$$L(y) = |\Sigma|^{-n/2} \prod_{i=1}^n \sum_{j=1}^J w_j \exp\left(-\frac{1}{2}(y_i - \mu_j)'\Sigma^{-1}(y_i - \mu_j)\right).$$

Standard practice assigns Dirichlet distribution on the simplex in finite mixture $Dir(\alpha)$ and Dirichlet process $DP(\alpha)$ for infinite mixture when J is unknown. For simplicity, we focus on finite mixture case with J finite and known. The prior $Dir(\alpha)$ can be viewed as a prior $\pi_{0,\mathcal{R}}(w) = \prod_{j=1}^J w_j^{\alpha-1}$ with $\mathcal{R} = (0,1)^J$, under additional hard constrained of $1-$norm equality:

$$\pi_{0,\mathcal{D}}(w) \propto \prod_{j=1}^{J} w_j^{\alpha-1}\, \mathbb{1}_{\sum_{j=1}^{J} w_j=1} \tag{10}$$

This can be easily approximated with extrinsic prior. However, one known issue for mixture modeling under canonical Dirichlet prior is the label-switching problem. With parameter $\{\mu_j, w_j\}$ indexed by $j = 1, \ldots, J$, due to exchangability, one can switch any two j and j' without changing likelihood. It is a controversial topic whether the occurrence of label-switching or the lack thereof is more ideal (see review in Jasra et al. (2005)) in general; but in the case that posterior distribution is symmetric about any permutation in j's, as our normal mixture example, sampling over all permutations of j is redundant. Therefore, it is rather useful to avoid label-switching and have convergence in such cases. Unfortunately, sometimes the switching issue can be impossible to avoid, even with very local update in Gibbs sampling. This is because when sample size n is small, posterior variances of μ_j's can be quite large, with significant overlap among their high posterior regions. In early work, Diebolt and Robert (1994) suggested ordering in μ_j's, but it is not clear how it would work with multi-dimensional $\mu_j \in \mathbb{R}^d$ with $d \geq 2$.

Observing that each w_j is one-dimensional, we apply order constraint on $w_1 \geq w_2 \geq \ldots \geq w_J$, yielding an ordered Dirichlet prior:

$$\pi_{0,\mathcal{D}}(w_1, \ldots w_J) \propto \prod_{j=1}^{J} w_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^{J} w_j=1} \cdot \prod_{j=1}^{J-1} \mathbb{1}_{w_j \geq w_{j+1}} . \tag{11}$$

where $w_j \in (0, 1)$. Unlike early post-hoc relabeling algorithm (Stephens, 2000), we remove exchangability directly to reduce label-switching. Strictly speaking, label-switching could still happen when any two w_j's are very close; nevertheless, this help prevent label-switching between large and small components.

The ordered Dirichlet no longer has closed-form posterior, however it is easy to approximately estimate with the help of extrinsic prior:

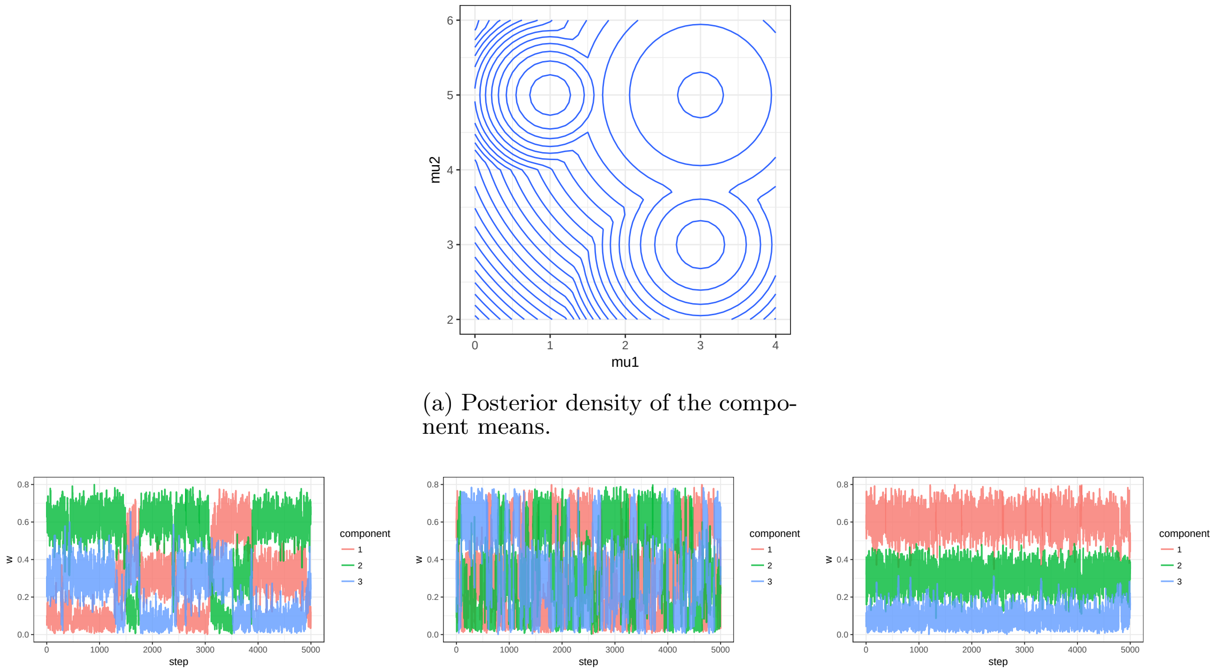
$$\pi_{0,\mathcal{R}}(w) \cdot \mathcal{K}(w) \propto \prod_{j=1}^{J} w_j^{\alpha-1} \cdot \prod_{j=1}^{J-1} K_1((w_{j+1} - w_j)_+) \cdot K_2(|\sum_{j=1}^{J} w_j - 1|)$$

where $K_k(x) = \exp(-\lambda_k x^2)\, \mathbb{1}_{x<4/\sqrt{2\lambda_k}}$ for $k = 1, 2$. 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 $\mathbb{R}^J$. For comparison, we also test with $\lambda_1 = 0$ to remove the order constraint and allow HMC to run on a canonical Dirichlet prior in (10).

We generate $n = 100$ samples from 3 components with true $\{w_1, w_2, w_3\} = \{0.6, 0.3, 0.1\}$, with corresponding two-dimensional means $\{\mu_1, \mu_2, \mu_3\} = \{[1, 5], [3, 3], [3, 5]\}$ and identity covariance $\Sigma = I_2$. We assign 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 $\sigma_1^2, \sigma_2^2 \sim IG(2,1)$.

Figure 3 shows the contour of true posterior density of μ_j's and the traceplot of w_j's in three approaches: standard Gibbs sampling with augmented component assignment (Diebolt and Robert, 1994) under canonical prior (10), HMC using extrinsic prior associated under canonical prior (10) and and HMC using extrinsic prior under ordered prior (11). Each approach runs $10,000$ iterations with first $5,000$ discarded as burn-in. For the posterior extrinsic collected under extrinsic prior, a simple projection $P(w^*) = w^*/||w^*||_1$ is used as proposal in M-H correction, yielding acceptance rate of 0.95. Due to small sample size and relatively overlap of means, significant label-switching is shown in both Gibbs and HMC under canonical Dirichlet prior; while HMC with ordered Dirichlet prior does not suffer this issue.



(a) Posterior density of the component means.

(b) Gibbs sampling under canonical Dirichlet

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

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

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

Example 2: Orthonormal Tucker Factorization in Multiple Network Analysis

We now consider another application of constrained model in network analysis. What a a

$$A_i \sim \text{Bern}(\frac{1}{1+\exp(-\psi_i)})$$

$$\psi_i = UD_iU$$

$$D_i = \text{diag}(d_{i1}, d_{i2}, d_{i3})$$

$$vec(U) \sim N(vec(X), \sigma^2)$$

for $k = 1, \ldots d$. The orthonormailty restricts rotation and scaling and $g_k(x_1) \geq 0$ restricts column-wise sign change.

5 Discussion

References

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.

Diebolt, J. and C. P. Robert (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 363–375.

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.

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).

Livingstone, S., M. Betancourt, S. Byrne, and M. Girolami (2016). On the geometric ergodicity of hamiltonian monte carlo. *arXiv preprint arXiv:1601.08057*.

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

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

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

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.