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 via Embedding

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 an embedding 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.

Starting with an unconstrained prior $\pi_{0,\mathcal{R}}(\theta)$ on $\mathcal{R}$, it may or may not have zero measure in $\mathcal{D}$. When $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta) d\theta > 0$, the constrained prior can be obtained by applying the restriction through an indicator function $\mathbb{1}_{\theta \in \mathcal{D}}$, and renormalizing:

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

When $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta = 0$, one may utilize a neighborhood surrounding $\mathcal{D}$, $\mathcal{D}^+_\epsilon = \{x : \text{dist}(x,y) < \epsilon, \forall x \in \mathcal{R}, y \in \mathcal{D}\}$ with dist(.,.) as the appropriate metric in $\mathcal{R}$, such as Euclidean distance in $\mathbb{R}^p$. With $\int_{\mathcal{D}^+_\epsilon} \pi_{0,\mathcal{R}}(\theta)d\theta > 0$ for all $\epsilon > 0$, one obtain the conditional probability for all $\mathcal{A} \subset \mathcal{R}$ with non-trivial probability in the limit:

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

provided the limit exists. Based on the probability, the density $\pi_{0,\mathcal{D}}(\theta)$ can then be derived. Similarly, the conditional expectation for any $f(\theta)$ is $\int f(\theta)\pi_{0,\mathcal{D}}(\theta)d\theta = \lim_{\epsilon \to 0^+} \frac{\int f(\theta)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta \in \mathcal{D}^+_\epsilon} d\theta}{\int_{\mathcal{D}^+_\epsilon} \pi_{0,\mathcal{R}}(\theta)d\theta}$.

Letting the likelihood function be $L(y;\theta)$, the posterior of θ is $\pi_{\mathcal{D}}(\theta \mid y) = \frac{L(y;\theta)\pi_{0,\mathcal{D}}(\theta)}{\int_{\mathcal{D}} L(y;\theta)\pi_{0,\mathcal{D}}(\theta)d\theta}$. When $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta > 0$, substituting in (1) yields:

$$\pi_{\mathcal{D}}(\theta \mid y) = \frac{L(y;\theta)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta \in \mathcal{D}}}{\int_{\mathcal{R}} L(y;\theta)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta \in \mathcal{D}} d\theta}, \tag{3}$$

assuming $\int_{\mathcal{R}} L(y;\theta)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta \in \mathcal{D}} d\theta > 0$.

When $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta = 0$, computing conditional expectation for $f(\theta) = L(y;\theta)\mathbb{1}_{\theta \in \mathcal{A}}$ over $\pi_{0,\mathcal{R}}(\theta)$ yields

$$\int_{\mathcal{A}} \pi_{\mathcal{D}}(\theta \mid y)d\theta = \lim_{\epsilon \to 0^+} \frac{\int_{\mathcal{A}} L(y;\theta)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta \in \mathcal{D}^+_\epsilon} d\theta}{\int_{\mathcal{R}} L(y;\theta)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta \in \mathcal{D}^+_\epsilon} d\theta}. \tag{4}$$

Therefore, both prior and posterior can be obtained intrinsically via the embedding approach. However, its computation is often difficult and (2) and (4) can yield intractable form.

We now use an example to illustrate the embedding approach when $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta = 0$.

Example 1: 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 $\mathcal{D} = \{(\theta_1, \theta_2) : \theta_1 + \theta_2 = 1\}$. Letting $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, clearly $\int_{\mathcal{D}} \phi(\theta_1)\phi(\theta_2)d\theta = 0$. Using neighborhood $\mathcal{D}^+_\epsilon = \{(\theta_1, \theta_2) : \theta_1 + \theta_2 \in (1 - \epsilon, 1 + \epsilon)\}$ and $\mathcal{A} = \{(\theta_1, \theta_2) : \theta_1 < x, \theta_2 \in \mathbb{R}\}$,

$$
\begin{aligned}
\int_{\mathcal{A}} \pi_{0,\mathcal{D}}(\theta)d\theta &= \lim_{\epsilon\to 0^+} \frac{\int_{(1-\epsilon)}^{(1+\epsilon)} \int_{-\infty}^{x} \phi(\theta_1)\phi(z-\theta_1)d\theta_1 dz}{\int_{(1-\epsilon)/\sqrt{2}}^{(1+\epsilon)/\sqrt{2}} \phi(z)dz} \\
&= \lim_{\epsilon\to 0^+} \frac{\frac{\partial}{\partial\epsilon}\int_{(1-\epsilon)}^{(1+\epsilon)} \int_{-\infty}^{x} \phi(\theta_1)\phi(z-\theta_1)d\theta_1 dz}{\frac{\partial}{\partial\epsilon}\int_{(1-\epsilon)/\sqrt{2}}^{(1+\epsilon)/\sqrt{2}} \phi(z)dz} \\
&= \lim_{\epsilon\to 0^+} \frac{\int_{-\infty}^{x} \phi(\theta_1)\left(\phi(1+\epsilon-\theta_1)d\theta_1 + \phi(1-\epsilon-\theta_1)\right)d\theta_1}{\phi(\frac{1+\epsilon}{\sqrt{2}})/\sqrt{2} + \phi(\frac{1-\epsilon}{\sqrt{2}})/\sqrt{2}} \\
&= \int_{-\infty}^{x} \frac{\sqrt{2}}{\sqrt{2\pi}} \exp(-\frac{(\theta_1-\frac{1}{2})^2}{2/2})d\theta_1.
\end{aligned}
$$

Differentiating with respect to x yields $\theta_1 \mid (\theta_1+\theta_2=1) \sim \text{No}(1/2,1/2)$, $\theta_2 \mid \theta_1 \sim \delta_{1-\theta_1}$, where δ_x denotes a mass point at x. Marginally, this corresponds to a degenerate bivariate Gaussian distribution:

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

2.2 Extrinsic Bayes

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.

$$
\tilde{\pi}_{0,\mathcal{D}}(\theta) = \frac{\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})}{\int_{\mathcal{R}} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})d\theta} \tag{5}
$$

where $\mathcal{K}(\theta;\mathcal{D})$ is an approximation to $\mathbb{1}_{\theta\in\mathcal{D}}$ and satisfies $\int_{\mathcal{R}} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})d\theta > 0$. This approximation applies regardless if $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta > 0$ or $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta = 0$. For the latter, starting from $\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)$ an approximation to $\mathbb{1}_{\theta\in\mathcal{D}_\epsilon^+}$, the limit can be obtained $\int_{\mathcal{A}} \tilde{\pi}_{0,\mathcal{D}}(\theta)d\theta = \lim_{\epsilon\to 0^+} \frac{\int_{\mathcal{A}} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}{\int_{\mathcal{R}} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta} = \int_{\mathcal{A}} \frac{\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})}{\int_{\mathcal{R}} \pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D})d\theta}d\theta$.

Its posterior, referred as extrinsic posterior from now on, takes the form

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

Assuming there are m constraints with each defining a constrained subspace $\mathcal{D}_k$, $\mathcal{D} = \bigcap_{k=1}^m \mathcal{D}_k$, we have

$$
\mathcal{K}(\theta;\mathcal{D}) = \prod_{k=1}^{m} K_k(v_k(\theta)), \tag{7}
$$

where $v_k : \mathcal{R} \to [0,\infty)$ is a measurable function and quantifies the distance to space $\mathcal{D}_k$. 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 $\{\theta : f(\theta) \leq 0\}$. The function $K_k : [0, \infty) \to [0, 1]$ penalizes $v_k(\theta) > 0$ and decreases in $v_k(\theta)$. For example, $K_k(v) = \exp(-v)$ or $K_k(v) = \frac{1}{1+v^2}$. For illustration simplicity, in this paper, we focus on exponential smoothing function $K_k(v(\theta)) = \exp(-\frac{v(\theta)}{\lambda_k})$ with $\lambda_k > 0$ as a tuning parameter.

This framework is applicable to a variety of complicated case. For example, θ can have only some of parameters constrained; parameters can be in multiple constraints simultaneously; constraints can be dependent. Regardless, it is generally simple to adapt (7) via appropriate mapping through $v(\theta)$.

We now provide some additional regularity conditions on v_k and K_k: for $k = 1, \ldots, m$, $v_k(\theta) = 0$ only if $\theta \in \mathcal{D}_k$ and $K_k(v) = 1$ only if $v = 0$. Therefore, $\theta \in \mathcal{D} \Leftrightarrow$ all $v_k(\theta) = 0 \Leftrightarrow \mathcal{K}(\theta; \mathcal{D}) = 1$, this ensures $\mathcal{K}(\theta; \mathcal{D})$ is the same as the indicator function for $\theta \in \mathcal{D}$.

To illustrate, consider a truncated normal prior $\text{No}_{(-\infty,5)}(0, 5^2)$. The unnormalized density of intrinsic prior is $\exp(-\theta^2/2 \cdot 5^2)\mathbb{1}_{\theta \in (-\infty,5)}$, which can be approximated by extrinsic prior $\exp(-\theta^2/2 \cdot 5^2)\exp(-v(\theta))$. We plot two distances $v(\theta) = (\theta - 5)_+$ and $v(\theta) = (\theta - 5)_+^2$ in Figure 1. 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.

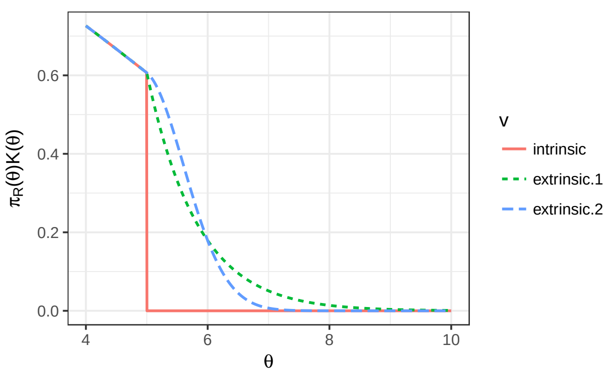


Figure 1: Unnormalized densities for truncated normal $\text{No}_{(-\infty,5)}(0, 5^2)$ under exact intrinsic prior and approximating extrinsic prior. Inside $(-\infty, 5)$, the priors are the same up to a constant difference. The intrinsic prior abruptly drops to 0 on the boundary, while the approximating 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 apply extrinsic Bayes in the previous example of two Gaussians under sum constraint.

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

We now 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.

Obviously, the derivation based on extrinsic approach is much simpler than intrinsic one.

2.3 Approximation Error

We now study the properties of the extrinsic prior and posterior. One important task is to quantify the difference between extrinsic and intrinsic ones. Due to similar reasoning for prior and posterior, we use some general notation. Let $\pi_{\mathcal{R}}(\theta)$ be an unnormalized density in $\mathcal{R}$, which is $\pi_{0,\mathcal{R}}(\theta)$ when studying prior and $L(y;\theta)\pi_{0,\mathcal{R}}(\theta)$ when studying posterior; $\Pi(.)$ and $\tilde{\Pi}(.)$ to represent the measures under intrinsic and extrisic methods associated with either prior or posterior distribution.

We first explore the case case in (1) and (3), where $\int_{\mathcal{D}} \pi_{\mathcal{R}}(\theta)d\theta > 0$.

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

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

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

Using exponential smoothing function $\mathcal{K}(\theta;D) = \prod_{k=1}^m \exp(-v_k(\theta)/\lambda_k)$, as all $\lambda_k \to 0$, the total variation $||\Pi(.), \tilde{\Pi}(.)||_{TV} \to 0$, if $\int_{\mathcal{R}} \pi_{\mathcal{R}}(\theta)d\theta < \infty$. This is a direct result of dominated convergence theorem and $\mathcal{K}(\theta;\mathcal{D}) \leq 1$.

With some mild conditions, we further quantify the non-asymptotic rate. Letting $\lambda = \sup_k \lambda_k$, $M = \int_{\mathcal{R}\setminus\mathcal{D}} \pi_{\mathcal{R}}(\theta)d\theta < \infty$. As a linear combination of measurable functions, $v(\theta) = \lambda\sum_{k=1}^m \frac{v_k(\theta)}{\lambda_k}$ is measurable. Let $f(v)$ be the density of $v(\theta)$.

Remark 2. *If there exists a $t < \infty$ such that $f(v) < \infty$ for any $v < t$,*

$$\int_{\mathcal{R}\setminus\mathcal{D}} \pi_{\mathcal{R}}(\theta)\prod_{k=1}^m \exp(-v_k(\theta)/\lambda_k)d\theta \leq M\exp(-\frac{t}{\lambda}) + M \sup_{t^*\in(0,t)} f(t^*)\lambda$$

proof:

$$\begin{aligned}
\frac{1}{M}\int_{\mathcal{R}\setminus\mathcal{D}}\pi_{\mathcal{R}}(\theta)\prod_{k=1}^{m}\exp(-v_k(\theta)/\lambda_k)d\theta =&\mathbb{E}\exp(-\frac{v}{\lambda}) \\
=&\mathbb{E}\mathbb{1}_{(0,t)}\exp(-\frac{v}{\lambda})+\mathbb{E}\mathbb{1}_{(t,\infty)}\exp(-\frac{v}{\lambda}) \\
\leq&\int_0^t f(v)\exp(-\frac{v}{\lambda})dv+\exp(-\frac{t}{\lambda}) \\
\leq&\sup_{t^*\in(0,t)} f(t^*)\int_0^t\exp(-\frac{v}{\lambda})dv+\exp(-\frac{t}{\lambda}) \\
=&\sup_{t^*\in(0,t)} f(t^*)\lambda(1-\exp(-t/\lambda))+\exp(-\frac{t}{\lambda}) \\
\leq&\sup_{t^*\in(0,t)} f(t^*)\lambda+\exp(-\frac{t}{\lambda})
\end{aligned} \tag{8}$$

Rearranging terms yields the result. ■

For λ close to 0, the extrinsic measure approaches intrinsic one in total varation distance in $O(\lambda)$. This rate is quantified under very general assumption. We expect it can be sharpened under special cases.

We now examine the second case in (2) and (4) where $\int_{\mathcal{D}}\pi_{\mathcal{R}}(\theta)d\theta=0$.

Remark 3. *If* $\lim_{\epsilon\to0^+}\frac{\int_{\mathcal{A}}\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}_\epsilon^+}d\theta}{\int_{\mathcal{D}_\epsilon^+}\pi_{\mathcal{R}}(\theta)d\theta}$ *and* $\lim_{\epsilon\to0^+}\frac{\int_{\mathcal{A}}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}{\int_{\mathcal{R}}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}$ *converge uniformly in* $\mathcal{A}$*, the total variation distance between the measures of extrinsic and intrinsic distribution*

$$\begin{aligned}
||\Pi(.),\tilde{\Pi}(.)||_{TV} =&\sup_{\mathcal{A}}||\Pi(\mathcal{A})-\tilde{\Pi}(\mathcal{A})|| \\
=&\sup_{\mathcal{A}}||\lim_{\epsilon\to0^+}\frac{\int_{\mathcal{A}}\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}_\epsilon^+}d\theta}{\int_{\mathcal{D}_\epsilon^+}\pi_{\mathcal{R}}(\theta)d\theta}-\lim_{\epsilon\to0^+}\frac{\int_{\mathcal{A}}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}{\int_{\mathcal{R}}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}|| \\
=&\lim_{\epsilon\to0^+}\sup_{\mathcal{A}}||\frac{\int_{\mathcal{A}}\pi_{\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}_\epsilon^+}d\theta}{\int_{\mathcal{D}_\epsilon^+}\pi_{\mathcal{R}}(\theta)d\theta}-\frac{\int_{\mathcal{A}}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}{\int_{\mathcal{R}}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}|| \\
=&\lim_{\epsilon\to0^+}\frac{\int_{\mathcal{R}\setminus\mathcal{D}_\epsilon^+}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}{\int_{\mathcal{R}\setminus\mathcal{D}}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}.
\end{aligned} \tag{9}$$

Let $g(\epsilon,\lambda)=\frac{\int_{\mathcal{R}\setminus\mathcal{D}_\epsilon^+}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}{\int_{\mathcal{R}\setminus\mathcal{D}}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}$, using exponential smoothing function and $\lambda=\sup_k\lambda_k$, $\lim_{\lambda\to0^+}g(\epsilon,\lambda)=0$ pointwise in ϵ. And we have $||\Pi(.),\tilde{\Pi}(.)||_{TV}\leq\lim_{\epsilon\to0^+}\frac{M_\epsilon\exp(-\frac{t}{\lambda})+M_\epsilon\sup_{t^*\in(0,t)}f(t^*)\lambda}{\int_{\mathcal{R}\setminus\mathcal{D}}\pi_{\mathcal{R}}(\theta)\mathcal{K}(\theta;\mathcal{D}_\epsilon^+)d\theta}$ assuming $M_\epsilon=\int_{\mathcal{R}\setminus\mathcal{D}_\epsilon^+}\pi_{\mathcal{R}}(\theta)d\theta<\infty$. For every ϵ, there is a small λ to make the total variation distance 0. However, with

$\lambda = \lambda_0$ fixed and non-negligble, $\lim_{\epsilon \to 0^+} g(\epsilon, \lambda) = 1$. This is an inherent weakness of total variation distance in comparing continuous and degenerate distributions. Therefore, to quantify distance with non-zero λ, we use Wasserstein distance instead:

$$W_p(\Pi, \tilde{\Pi}) = \left(\inf_{\gamma \in \Gamma(\Pi, \tilde{\Pi})} \int d(x, y)^p d\gamma(x, y) \right)^{1/p}$$

where $d(x, y)$ is the metric on $\mathcal{R}$ and $\Gamma(\Pi, \tilde{\Pi})$ the set of couplings of Π and $\tilde{\Pi}$.

We continue in Example 1, denoting the bivariate Gaussian distribution obtained extrinsically as $\text{No}(\mu_\lambda, \Sigma_\lambda)$, the intrinsic one can be viewed as $\text{No}(\mu_0, \Sigma_0)$. The Wasserstein distance between the intrinsic and extrinsic distributions in Example 1 (Dowson and Landau, 1982) is $W_2(\Pi, \tilde{\Pi}) = \left(||\mu_\lambda - \mu_0||_2^2 + \text{tr}(\Sigma_\lambda + \Sigma_0 - 2(\Sigma_\lambda^{1/2} \Sigma_0 \Sigma_\lambda^{1/2})^{1/2}) \right)^{1/2}$. Figure 2 plots W_2 under different values of λ, with $W_2 \approx 0$ with $\lambda = 10^{-3}$.

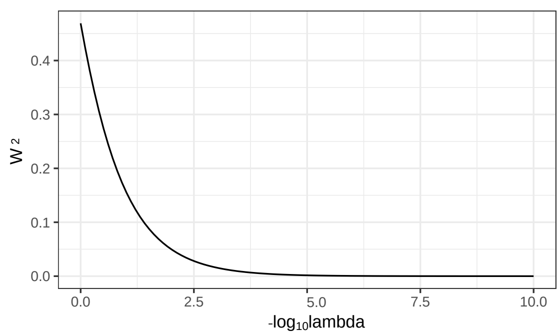


Figure 2: Wasserstein distance of 2nd order W_2 between the intrinsic and extrinsic prior for two normal random variables under sum constraint. W_2 declines rapidly as λ (shown as $-\log_{10}(\lambda)$) decreases.

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).

Assuming $\theta \in \mathcal{R}$ with $\mathcal{R}$ being 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 (6) is

$$\begin{aligned} &H(\theta, p) = U(\theta) + M(p), \\ &\text{where } U(\theta) = -\log\left\{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta; \mathcal{D})\right\}, \\ &M(p) = \frac{p'\Sigma^{-1}p}{2}. \end{aligned} \tag{10}$$

At the beginning of each iteration, from current state $(\theta^{(0)}, p^{(0)})$, HMC updates θ and p via Hamiltonian dynamics, defined by two partial differential equations:

$$\begin{aligned} \frac{\partial \theta^{(t)}}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = \Sigma^{-1}p, \\ \frac{\partial p^{(t)}}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \end{aligned} \tag{11}$$

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, (11) lacks closed-form solution, one has to use an *integrator* that numerically approximates the exact evolution. With a reversible and volume-preserving integrator (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 popular integrator with reversible and volume-preserving properties is the leap-frog algorithm (Neal, 2011), which utilizes discrete time-step move $(\theta^{(T\eta)}, p^{(T\eta)}) \to (\theta^{((T+1)\eta)}, p^{((T+1)\eta)})$:

$$p \leftarrow p - \frac{\eta}{2}\frac{\partial U}{\partial \theta}, \quad \theta \leftarrow \theta + \eta\Sigma^{-1}p, \quad p \leftarrow p - \frac{\eta}{2}\frac{\partial U}{\partial \theta} \tag{12}$$

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

3.2 Support Expansion and 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), $\gamma(\theta, \theta^*) = \sup_{g \in L^2(\Pi)} \text{corr}(g(\theta), g(\theta^*))$, where $L^2(\Pi) = \{g(\theta) : \text{var}(g(\theta)) < \infty\}$. Smaller $\gamma(\theta, \theta^*)$ corresponds to less correlation and faster convergence. Given η, one can adaptively choosing L so that $\gamma(\theta, \theta^*)$ falls under certain desired rate (Hoffman and Gelman, 2014).

Since each integrator step is often the bottleneck, one would use L as small as possible; this minimum step number depends on the step size η. Given a desired convergence rate γ^*, an optimal η would be:

$$\eta^* = \underset{\eta:\eta\leq\eta_{max}}{\arg\inf} \inf_L \{L : \gamma\left(\theta^{(0)}, \theta^{(\eta L)}\right) \leq \gamma^*\},$$

where η_{max} is the stability bound to prevent the integrator from diverging in approximation, often $\eta^* \approx \eta_{max}$.

The stability bound η_{max} is roughly determined by the width of support in the most constrained direction (Neal, 2011). Formally, letting $\mathcal{S}_\eta$ be the outer contour region of the effective support $\mathcal{S}_\delta = \{x : \pi(x) \in (0, \delta)\}$ with δ arbitrarily small and $\pi(x)$ denoting the density, the minimum support width is:

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

where $d(.,.)$ is the appropriate metric in space $\mathcal{R}$ and is Eucledean distance in continous HMC.

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

Since HMC with extrinsic prior operates in $\mathcal{R}$, whether constrained space $\mathcal{D}$ has a small w_π or not should determine how to specify $\mathcal{K}(\theta; \mathcal{D})$ in extrinsic prior. When w_π is away from 0, one can let $\mathcal{K}(\theta; \mathcal{D})$ rapidly fall to 0 when $\theta \notin \mathcal{D}$. In this case, one can simply use very large λ in exponential smoothing function $K(v(\theta)) = \exp(-v(\theta)/\lambda)$, without affecting computing efficiency. On the other hand, if $w_\pi \approx 0$ for $\mathcal{D}$, one needs to let $\mathcal{K}(\theta; \mathcal{D})$ induce some support expansion, so that the computation can be efficiently carried out. This involves some trade-off between computing time and approximation accuracy. Empirically, we found that $\lambda = 10^{-3}$ provide a good balance point. These will be illustrated via two examples in the next section.

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

Example 2: Linear Regression Under Inequality Constraint

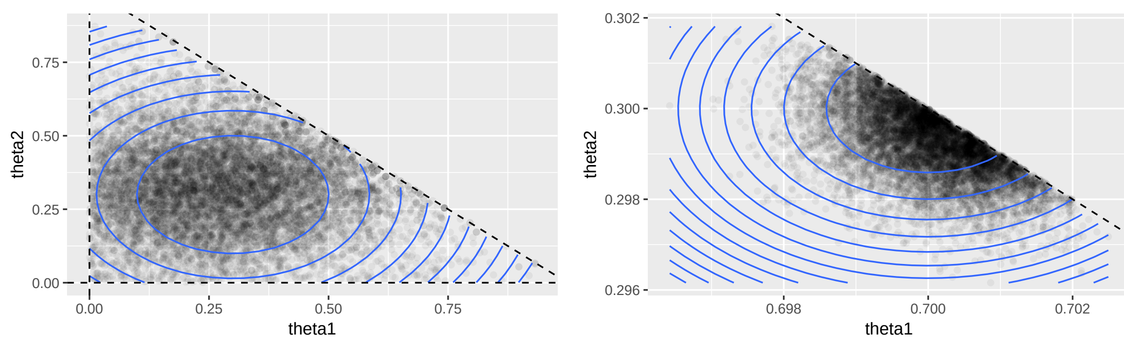
When the support on constrained space $\mathcal{D}$ has w_π away from 0, one can use extrinsic prior with almost no support expansion. This applies a large class of inequality constraints that has wide support. For example, consider a linearly constrained regression model:

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

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 given. 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 $\theta_1 + \theta_2 \leq 1$, 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 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 $\theta \sim \text{No}([0.5, 0.5]', I10^2)$ and inverse-Gamma prior $\sigma^2 \sim \text{IG}(2, 1)$.

We use $\mathcal{K}(\theta) = \exp(-\frac{v(\theta)}{\lambda})$ with $v(\theta) = |\theta_1 + \theta_2 - 1|_+$ in the extrinsic prior. We choose $\lambda = 10^{-8}$ leading to almost no support expansion. We collect $10,000$ posterior samples efficiently via HMC. The Markov chain mixes rapidly , generating $10,000$ effective sample size in both experiments. Figure 3 plots the posterior sample and its contour. There is no posterior that fall outside $\mathcal{D}$, thanks to 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 $\theta_1 + \theta_2 \leq 1$. 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

If the constrained support in $\mathcal{D}$ has a minimum support width w_π close to 0, it limits the stability bound and computing efficiency in continuous HMC. In this example, we consider $\mathcal{D}$ is a two-dimensional unit cirlce $\{(\theta_1, \theta_2) : \theta_1^2 + \theta_2^2 = 1\}$ (alternatively, $V(2,1)$, a $(2,1)$–Stiefel manifold). In this space, $w_\pi = 0$ hence support expansion is necessary.

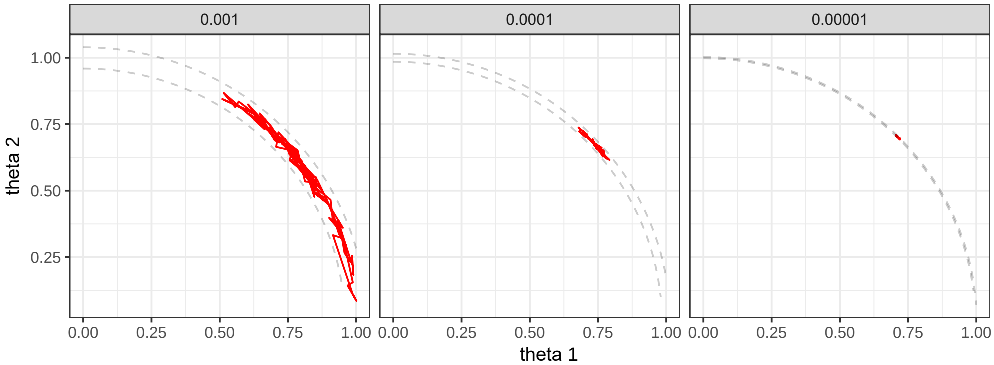
Let data $y_i \in \mathbb{R}^2$ for $i = 1, \ldots, n$ be noisy realization from one point on unit circle:

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

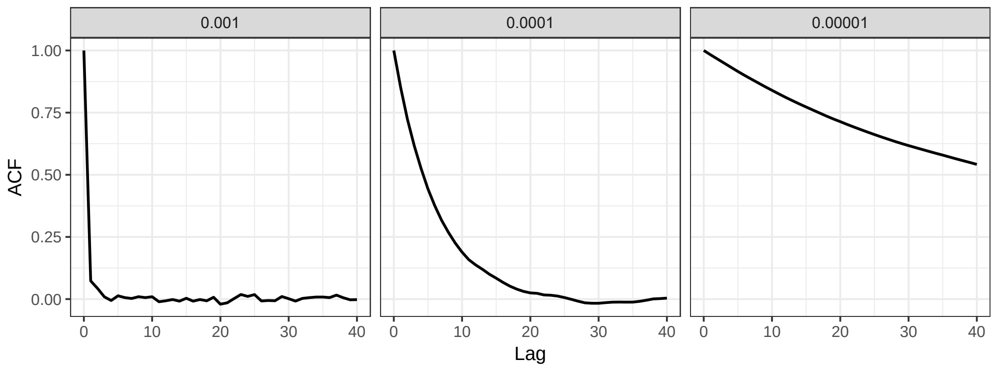
where $\theta \in \mathcal{D}$ is assigned a von Mises–Fisher prior $\pi_{0,\mathcal{D}}(\theta) \propto \exp(F'\theta_i)$.

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. We then use $F = (1, 1)$ to induce a weakly informative prior for θ and an inverse-Gamma prior 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 $\tilde{\pi}_{0,\mathcal{D}}(\theta) = \exp(F'\theta_i)\exp(-\frac{|\theta'\theta-1|}{\lambda})$.

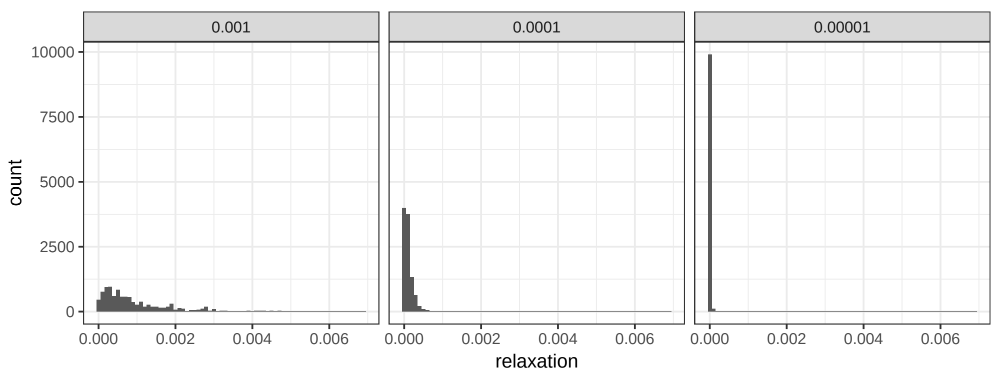
We test $\lambda = 0.001$, 0.0001 and 0.00001 in three experiments. We run each algorithm for $20,000$ steps, with first $10,000$ discarded. To compare the computing efficiency, we restrict the maximum leap-frog steps L to be 100 and visualize how much space each algorithm can explore within one HMC iteration. Figure 4(a) plots the path of $L = 100$ leap-frog steps. Larger λ leads to wider support expansion and larger stability bound η_{max}. This makes $\theta^{(\eta L)}$ much less correlated with $\theta^{(0)}$ at each iteration, measured by autocorrelation (ACF) based on the posterior sample of θ_1 (Figure 4(b)). Even though we use somewhat small λ, the posterior distance $v(\theta)$ is small for all three settings (Figure 4(c)), with smaller λ associated with smaller $v(\theta)$ but lower computing efficiency.



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



(b) Autocorrelation of θ_1



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

Figure 4: 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.

4.2 Applications

In statistical modeling, it is common to encounter parameters or latent variables that are non-identifiable. Imposing constraints can often solve or reduce such issue, although they tend to make the computation difficult. We now illustrate utility of extrinsic prior for two applications.

Example 3: Ordered Dirichlet Prior

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

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:

$$\pi_{0,\mathcal{D}}(w_1, \dots w_J) \propto \prod_{j=1}^{J} w_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^{J} w_j = 1} \cdot \prod_{j=1}^{J-1} \mathbb{1}_{w_j \geq w_{j+1}} . \tag{13}$$

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

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

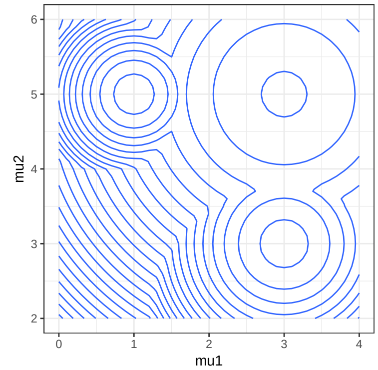
We now adopt this simplex distribution in a normal mixture model with mixture means and common variance, for data $y_i \in \mathbb{R}^d$ indexed by $i = 1, \dots, n$:

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

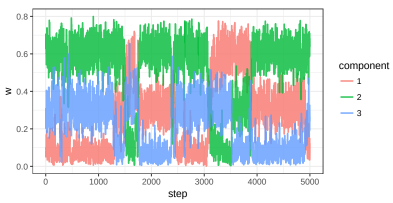
where $\delta_b(a) = 1$ if $a = b$ and 0 otherwise.

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 $\lambda_2 = 10^{-3}$ 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.

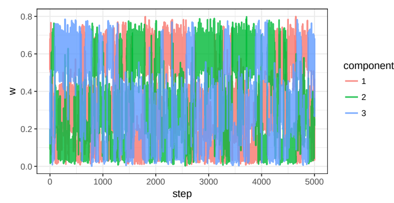
Figure 5(a) shows the contour of true posterior density of μ_j's. The small component sample size leads to large overlap among the posterior of μ_j's, generting in significant label-switching in both Gibbs and HMC under canonical Dirichlet prior. Figure 5(b,c,d) show the traceplot of w. Ordered Dirchlet has clearly better convergence due to ordering.



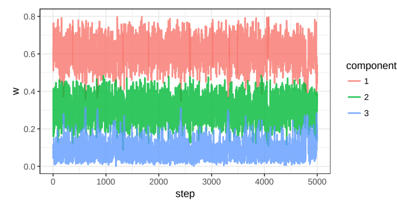
(a) Posterior density of the component means.



(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 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).

Example 4: Orthonormal Tensor 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 $V(V-1)/2 = 2,278$, 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 interpretable self-connectivity. These facts motivate a probabilitic low-rank model approach. We consider a symmetric tensor decomposition model:

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

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

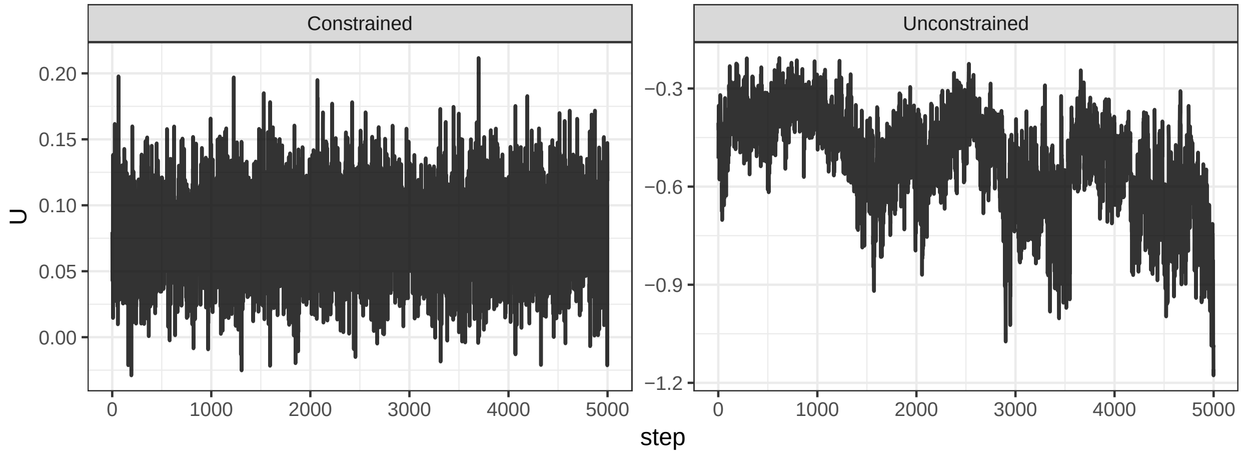
for $k > l$, $k = 2, \ldots, V$, $i = 1, \ldots, n$; U is $V \times d_1$ matrix, W is $n \times d_2$ matrix; D is a $d_1 \times d_2$ array. The $V \times V$ matrix Z is almost unstructural except symmetric $Z_{k,l} = Z_{l,k}$, which is commonly used to induce low-rank in the decomposition (Durante et 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 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 scaling. For example, one can multiply a $d_1 \times d_1$ non-zero diagonal matrix R, to U and obtain $U^* = UR$ obtain $D = R^{-2} D_{r_1,\cdot,\cdot}$. This leaves the likelihood unchanged, creating identifiability issue.

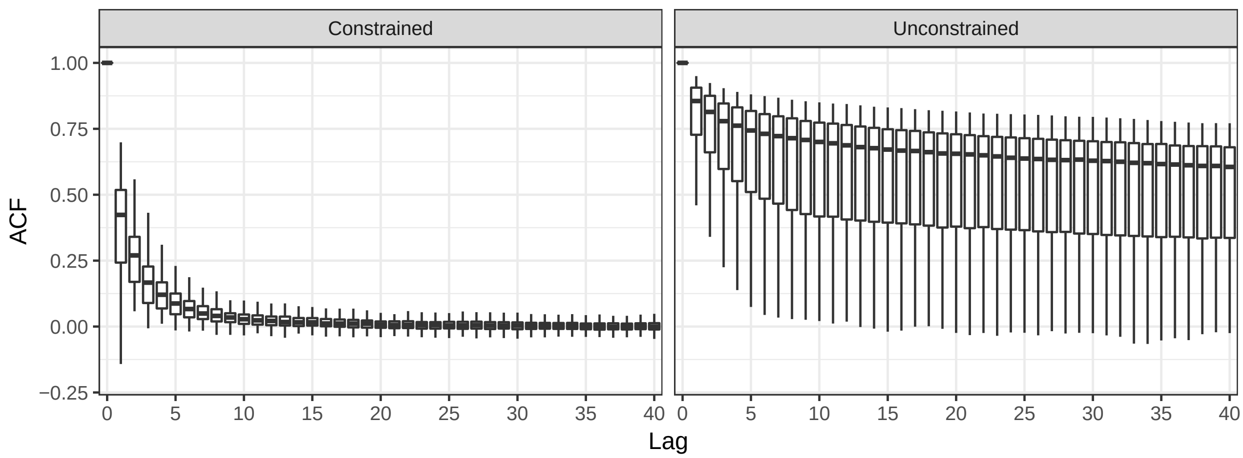
Therefore, we consider applying some constraint on the Tucker decomposition, while still maintaining its varying rank property over different modes. Motivated by high-order singular value decomposition, we impose orthonormality constraints $U'U = I_{d_1}$ and $W'W = I_{d_2}$.

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 $\tau_{r_1}, \tau_{r_2} \overset{indep}{\sim}$ IG$(2, 1)$ for all r_1, r_2.

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 6 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 6: 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.

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.

Dowson, D. and B. Landau (1982). The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis 12*(3), 450–455.

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.