Extrinsic Prior for Simple and Efficient Bayesian Modeling with Parameter Constraints

Abstract: Parameter constraints are commonly seen in statistical models, such as linear inequality, simplex constraint, parameter ordering, monotonicity, orthogonality, etc. Bayesian approach is useful for uncerntainty quantification in the constrained space. Although customized solutions have been made for different constraints, it is difficult to carry out estimation in general cases, especially when the posterior lacks closed-form or the model is heavily constrained. In this paper, we propose a simple and general solution by first replacing constraints with strongly informative prior. Through this *extrinsic* prior, the parameter support is relaxed to a less restrictive space, where conventional tools such as Hamiltonian Monte Carlo can be exploited to efficiently obtain approximate posterior. These posteriors can then be projected back to the constrained space for exact solution. This approach is very efficient and applicable to a wide range of problems including ones with equality and inequality constraints. Since priors are no longer limited to ones with closed-form posterior, more distribution families can be chosen for the constrained parameters. Multiple constraints can be freely adpoted for desired property such as model identifiability. Theory is developed and novel statistical applications under constraints are demonstrated.

KEY WORDS: Constraint relaxation; Space embedding; Monotone Dirichlet; Orthogonal Gaussian processes; Posterior mixing; Projected Markov chain

1 Introduction

Constraints are commonly assumed in modern statistical models. For example, functional data anaylsis often imposes constraint on shape, such as monotonicity or convexity on curves (Kelly and Rice, 1990); matrix and tensor decomposition utilize orthonormality to remove scaling and rotation problem in indentifiability (Uschmajew, 2010); many manifolds such as simplex can be considered as sub-manifolds of a Euclidean space embedded via constraints.

When data are in constrained space, parameters can enter the likelihood via an integral without closed-form, leading to "doubly intractable" problem. Various successful solutions have been proposed for this issue (Murray et al., 2012; Rao et al., 2016). When parameters are constrained, challenges often arise in estimation.

Frequentist optimization literature often relies on Lagrange and Karush-Kuhn-Tucker multipliers for point estimate under equality and inequality constraints (Boyd and Vandenberghe, 2004). However, uncertainty quantification is difficult since conventional asympotic result on variance estimation often no longer hold in constrained space. In this regard, Bayesian approach is more approriate.

There have been a variety of customized solutions developed for specific constraints. One strategy involves using constrained prior with posterior that can be conveniently sampled. For example, for modeling othornormal matrices on the Stiefel manifold, Bingham-von-Mises-Fisher distribution (Khatri and Mardia, 1977; Hoff, 2009) is a parametric family used in matrix and tensor decomposition. Lin et al. (2016) extends the flexibility of matrix von-Mises-Fisher distribution via non-parametric approach. Another strategy is to bypass the constraint via re-parameterization. The famous example is the stick-breaking construction for Dirichlet distribution and process. The re-parameterization essentially utilizes the coordinate system of the simplex, and circumvents the norm constraint on the probability vertices. As these methods directly satisfy the constraint requirement, we refer them as the intrinsic approaches. Despite the success of intrinsic approaches, posteriors can quickly become very involved in Gibbs sampling, leading to inefficient posterior mixing. The closed-form solution also often breaks under slightly more advanced model or complicated data. For example, in modeling population of undirected networks, the symmetry in each network disrupts the closed form of posterior in orthogonal tensor decomposition proposed by Hoff et al. (2016), demanding new rejection sampling algorithm to be developed. As another example, additional strucutre (such as ordering) on the probability simplex often disrupts the simple form of stick-breaking posterior.

These drawbacks have motivated the development of extrinsic approaches. The key idea is to first sample the proposal freely in a conventional space (such as Euclidean space), then transform it back to the constrained space. One early work can be traced back to Gelfand et al. (1992), who used Gibbs sampling to first generate proposal in unrestricted region, then only accepting those falling inside the constraint space. One critical issue is that unrestricted proposal can have significant mass outside the constraint region, resulting in a high rejection rate. Replacing rejection sampling, Lin and Dunson (2014) and Lin et al. (2016) utilize projection to map the unconstrained posterior into the constrained space and obtain monotonicity and manifold-valued regression. Neal (2011) suggested using large penalty to create a energy wall to guide the Hamiltonian dynamics involving simple space truncation, and accept proposal only when it is inside the truncated space. Pakman and Paninski (2014) applied similar idea in making the generation of truncated multivariate normal more efficient. These specialized cases work well, but such case are often rare and there is a clear lack of general and simple approach.

In this paper, we propose a general extrinsic approach, by parameterizing constraints as a limiting case of strongly informative prior. We refer them as extrisic priors. We then relax the effective support of the prior

to a neighborhood of constraint space, obtaining posterior via efficient tools such as conventional Hamiltonian Monte Carlo (HMC). These posteriors are approximate to the canonical formulation, with approximation error bounded during the prior specification. The imperfection of approximation can be corrected with a simple projection and a Metropolis-Hastings step with high acceptance probability, leading to a Markov chain corresponding to the exact formulation. Compared to other manifold based methods such as Riemannian and geodesic HMC (Girolami and Calderhead, 2011; Byrne and Girolami, 2013), our approach is efficient in computation and simple to implement via highly automatic software like STAN. The simplicity enables a larger spectrum of prior to be chosen and more free adaoption of constraints in modeling. Theoretic studies are conducted and original models are shown in simulations and data application.

2 Method

We consider a parameter θ in a constrained space $\mathcal{D}$. The space $\mathcal{D}$ can be high- or infinite-dimensional. Letting this space be equiped with a σ-field $\mathscr{D}$, the standard Bayesian approach assigns a prior for θ in $\mathcal{D}$, based on a density $\pi_{0,\mathcal{D}}(\theta)$ in a separable space $(\mathcal{D}, \mathscr{D})$. In intrinsic approaches, priors are chosen for computational conveniences so that the posterior can be easily sampled strictly inside $\mathcal{D}$. Clearly, the choices of priors and constraints one can impose are very limited. Instead, we consider extrinsic approaches by estimating θ in the larger space $\mathcal{R}$ where $\mathcal{D} \in \mathcal{R}$. We first provide a probabilistic justification.

Assuming $\pi_{0,\mathcal{D}}(\theta)$ is proper $\int_{\mathcal{D}} \pi_{0,\mathcal{D}}(\theta)d\theta < \infty$, then this prior can be viewed as a conditional density, based on another density $\pi_{0,\mathcal{R}}(\theta)$ in $(\mathcal{R}, \mathscr{C})$ with $\mathscr{C}$ as the σ-field of $\mathcal{R}$:

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

where $\mathbb{1}_{\theta\in\mathcal{D}} = 1$ when $\theta \in \mathcal{D}$, 0 otherwise. Note as long as $\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}}$ is proper, $\pi_{0,\mathcal{R}}(\theta)$ can be improper. Letting $L(\theta; y)$ be the likelihood function and y be the observed data, the posterior can be obtained via:

$$\pi(\theta \mid y, \theta \in \mathcal{D}) = \frac{L(\theta; y)\pi_{0,\mathcal{D}}(\theta)}{\int_{\mathcal{D}} L(\theta; y)\pi_{0,\mathcal{D}}(\theta)d\theta} = \frac{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}}}{\int_{\mathcal{D}} L(\theta; y)\pi_{0,\mathcal{R}}(\theta)d\theta}, \tag{2}$$

where the last equality holds because $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta$ is a finite constant.

2.1 Extrinsic Prior for Constraints

One obvious extrinsic approach utilizing (2) is to first generate proposal in $\mathcal{R}$ based on $L(\theta; y)\pi_{0,\mathcal{R}}(\theta)$ (assuming it is proper), then accepting it when it falls in $\mathcal{D}$ (Gelfand et al., 1992). However, when the probability $Pr(\theta \in \mathcal{D} \mid y)/Pr(\theta \in \mathcal{R} \setminus D \mid y) \approx 0$, especially common in equality constraint, this would lead to most of the proposals being rejected.

We propose a different strategy. Instead of ignoring $\mathbb{1}_{\theta\in\mathcal{D}}$ in the first step, we approximate it with an additional strongly informative prior $\mathcal{E}(\theta)$. The prior has its support in $\mathcal{R}$, but its mass concentrated near $\mathcal{D}$. When $\theta \in \mathcal{D}$, $\mathcal{E}(\theta)$ is constant; when θ is outside of D, $\mathcal{E}(\theta)$ quickly drops to 0. Then one can first obtain approximate posterior based on density proportional to $L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)$ (the conditions for posterior properity is postponed to the theory section).

In this paper, we focus on $\mathcal{D}$ that can be embedded in $\mathcal{R}$ via equality and inequality constraints, although other types of constraints can be incorporated similarly. There are m equalities and l inequalities, leading to $\mathcal{D} = \{\theta \in \mathcal{R} : E_k(\theta) = 0 \text{ for } k = 1, \ldots, m, \quad G_{k'}(\theta) \leq 0 \text{ for } k' = 1, \ldots, l\}$, where $E_k(.)$ and $G_{k'}(.)$ are functions that map from $\mathcal{R}$ to real line $\mathbb{R}$. Then the indicator function is $\mathbb{1}_{\theta\in\mathcal{D}} = \prod_k \mathbb{1}_{E_k(\theta)=0} \cdot \prod_{k'} \mathbb{1}_{G'_k(\theta)\leq 0}$.

We now replace the indicator functions with $\mathcal{E}(\theta)$, represented as a product of $(m + l)$ kernel functions $K_.(.)$, leading to posterior:

$$\begin{aligned} \pi_K(\theta \mid y) &\propto L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta) \\ &\propto L(\theta; y)\pi_{0,\mathcal{R}}(\theta) \cdot \prod_{k=1}^{m} K_{1,k}\Big(|E_k(\theta)|\Big) \cdot \prod_{k'=1}^{l} K_{2,k'}\Big(\big(G_{k'}(\theta)\big)_+\Big) \end{aligned} \tag{3}$$

where $(x)_+ = x$ if $x > 0$, 0 if $x \leq 0$. The posterior $\pi_K(\theta \mid y)$ is an approximation to $\pi(\theta \mid y)$ in (2). We will now refer $\pi_K(\theta \mid y)$ as "extrinsic posterior". The functions $|E_k(\theta)| \in [0, \infty)$ or $(G_{k'}(\theta))_+ \in [0, \infty)$ represent the amount of relaxation for each constraint, where 0 represents no relaxation. Each kernel $K_{i,k}$ satisfies $K_{i,k}(0) = 1$; the tolerable amount of relaxation is controlled by a hyper-parameter $\lambda_{i,k}$. When $\lambda_{i,k} \to \infty$, the kernel becomes a point mass at 0. Therefore, (2) is a limiting case of (3). For example, one simple and useful kernel is the truncated Gaussian $K_{i,k}(x) = \exp(-\lambda_{i,k}x^2)\mathbb{1}_{x<\varepsilon}$.

Instead of taking infinite values for λ's, we assign large but finite ones. This gives rise to a continuous relaxation of the sharp boundary of the indicator function. The relaxation allows the posterior θ to be easily sampled in $\mathcal{R}$ under the guidance of the strongly informative prior $\mathcal{E}(\theta)$. For example, one can carry out conventional HMC for constrained parameters directly in Euchledean space. At the same time, since posteriors are generated in a tight neighborhood of $\mathcal{D}$, they can be easily projected back to $\mathcal{D}$ as to produce exact posterior in $\mathcal{D}$.

2.2 Control of Constraint Relaxation

In the extrinsic posterior (3), when $\theta \in \mathcal{D}$, the density is the same as (2), up to a constant difference. However, since we induce positive support in $\mathcal{R} \setminus \mathcal{D}$, it is important to control the approximate posterior close to the space $\mathcal{D}$. This can be measured by the posterior distribution of the constraint relaxation $|E_k(\theta)|$ and $(G_{k'}(\theta))_+$.

We control the amount of constraint relaxation via a bounded prior support near 0 for each kernel. That is $\int_{x<\varepsilon} \mathcal{C}_{i,k}(x)dx = 1$, with $\mathcal{C}_{i,k}(x) = K_{i,k}(x)/\int_{\mathcal{R}} K_{i,k}(x)dx$. 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. As $\mathcal{E}(\theta) \propto \prod_{i,k} \mathcal{C}_{i,k}(x)$ is the joint extrinsic prior density, since $\pi_K(\theta \mid y) \ll \mathcal{E}(\theta)$, the posterior for each constraint relaxation is bounded in $[0, \varepsilon)$ with probability 1.

In practice, one may wish to utilize a kernel $K^*_{i,k}(x)$, orginally with unbounded support on $[0, \infty)$ for computing conveniences. To adopt them for bounded support in the relaxation x, one can first choose $\lambda_{i,k}$ to have $\int_{x<\varepsilon} \mathcal{K}^*_{i,k}(x)/\big(\int_{\mathcal{R}} K^*_{\cdot}(x)dx\big) = 1-\eta$ with η small, then apply truncation $K_{i,k}(x) = K^*_{i,k}(x)\mathbb{1}_{x<\varepsilon}$ to ensure $x < \varepsilon$ almost surely. In most cases, the truncation is only nominal for a theoretic guarantee; in computation it is often not needed. For example, in Gaussian kernel $\exp(-\lambda x^2)$ setting $\lambda = \frac{1}{2(\varepsilon/4)^2}$ ensures the relaxation $x < \varepsilon$ with probability 0.99993 apriori; for posterior sampling, one can first do an untruncated sampling, then reject those $x < \varepsilon$, which quite rare due to the small prior probability.

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

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

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 posterior and approximation with different λ's. The approximate 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 is associated with more rapid decline of density and therefore smaller constraint relaxation.

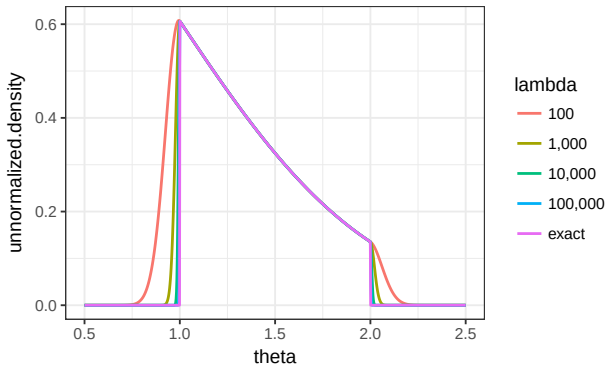


Figure 1: Unnormalized densities for truncated normal $\mathrm{No}_{(1,2)}(0,1)$, under exact and approximating densities. The exact density abruptly drops to 0 on the two boundaries, while the approximating ones drop continuously. In the approximation, larger λ is associated with lower tolerance for constraint relaxation ($(1-\theta)_+$ and $(\theta-2)_+$). All densities inside $(1, 2)$ are the same up to a constant difference.

Although it is temping to always induce almost 0 relaxation with very large λ, in heavily constrained models such as the ones with equality constraint, the narrow distribution width in $\mathcal{R}$ will cause a detrimental 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 two sections.

2.3 Posterior Sampling for Extrinsic Posterior

Extrinsic posterior is the approximation to the ones under exact formulation. As it is defined on a less restrictive space $\mathcal{R}$, it can be sampled easily. The tradiational sampling tools such as slice sampling, adaptive Metrohepolis-Hastings can be utilized. In this section, we present the sampling algorithm using Hamiltonian Monte Carlo (HMC), due to its high-level automation aided by software and excellent performance in convergence and posterior mixing. Various adaptive algorithms such as Hoffman and Gelman (2014) have been developed for making the new state less correlated with the current state.

In using the conventional HMC, we assume θ is d-dimensional and the space $\mathcal{R}$ is an Euchledean space $\mathbb{R}^d$ 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 is essentially a data augmentation based MCMC. Using a latent variable named "veolicty" $p \in \mathbb{R}^d$, the negative log-posterior function based on (3) is

$$\begin{aligned} &H(\theta, p) = U(\theta) + M(p), \\ &\text{where } U(\theta) = -\log\left\{L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)\right\}, \\ &M(p) = \frac{p'\Sigma^{-1}p}{2}, \end{aligned} \tag{4}$$

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

$$\begin{aligned} \frac{\partial\theta(t)}{\partial t} &= \frac{\partial H(\theta, p)}{\partial p} = \Sigma^{-1}p, \\ \frac{\partial p(t)}{\partial t} &= -\frac{\partial H(\theta, p)}{\partial \theta} = -\frac{\partial U(\theta)}{\partial \theta}. \end{aligned} \tag{5}$$

At the start of each iteration, the current state of θ is viewed as $\theta(0)$ and $p(0)$ randomly generated from $\mathrm{No}(0, \Sigma)$. The solution to (5) yields $\theta(t)$ and $-p(t)$ as the new state. Since Hamiltonian system is symplectic, $H(\theta(t), p(t)) = H(\theta(0), p(0))$. However, in most cases, (5) lacks closed-form solution, one has to use discrete approximation, commonly leap-frog algorithm (Neal, 2011):

$$
\begin{aligned}
p(T+\varepsilon/2) &= p(T) - \varepsilon/2 \frac{\partial U}{\partial \theta}(\theta(T)), \\
\theta(T+\varepsilon) &= \theta(T) + \varepsilon \Sigma^{-1} p(T+\varepsilon/2), \\
p(T+\varepsilon) &= p(T+\varepsilon/2) - \varepsilon/2 \frac{\partial U}{\partial \theta}(\theta(T+\varepsilon)),
\end{aligned} \tag{6}
$$

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 reversible, an Metropolis-Hastings step is 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 constraints $\theta \in \mathcal{D}$ is now replaced by prior $\mathcal{E}(\theta)$, the derivative $\frac{\partial U}{\partial \theta}$ can be computed easily and simple HMC can be directly run in space $\mathcal{R}$.

We now focus on finding the optimal time step ε, which can potentially impact the choice of λ. The time step ε controls the stability of trajectory. When ε is too large, H grows exponentially with L, leading to very small acceptance rate. When ε is too small, it leads to wasteful computation in making very local update. Therefore, it is useful to set large ϵ near a stability bound. For simple system such as $U(\theta) = \theta^2/2\sigma^2$, one can write (6) as a linear transfromation of $[\theta(T+\varepsilon), p(T+\varepsilon)]' = Q[\theta(T), p(T)]'$ (Q is a $2d \times 2d$ transition matrix), bounding the eigenvalues in Q below magnitude 1 determines the bound for ε. Since most systems involve nonlinear transition, analytical bound is not available, but one can empirically optimize ε to be close to this bound, by 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.

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 $L = 1$. Each update in leap-frog algorithm corresponds to $\theta(\varepsilon) = \theta(0) + \varepsilon p(0) - \varepsilon^2/2 \frac{\partial U}{\partial \theta}(\theta(0)) = \theta(0) + \varepsilon p(0) + O(\varepsilon^2)$. Even with moderate ε, $\theta(\varepsilon)$ can be outside the support, when the support is narrow in certain direction. This is because $p(0)$ is randomly generated in all direction of $\mathbb{R}^d$. However, 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 would never move to position with infinite $U(\theta(t))$, which corresponds to 0 posterior density. Therefore, given $\theta(0)$, the stability bound on ε is impacted by the smallest width of posterior support.

In extrinsic posterior, since the width of support is determined by λ in the prior, it is important to avoid creating a support too narrow. This is especially common with strong constraints like equality, a very large

λ would force small bound on ε, creating inefficient bottleneck. Instead, it is rather useful to use smaller λ to induce more relaxation, allowing discrete Hamiltonian dynamics to efficiently explore the space.

To illustrate, consider generating a random variable $\theta = (x_1, x_2)$ 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 determined by λ.

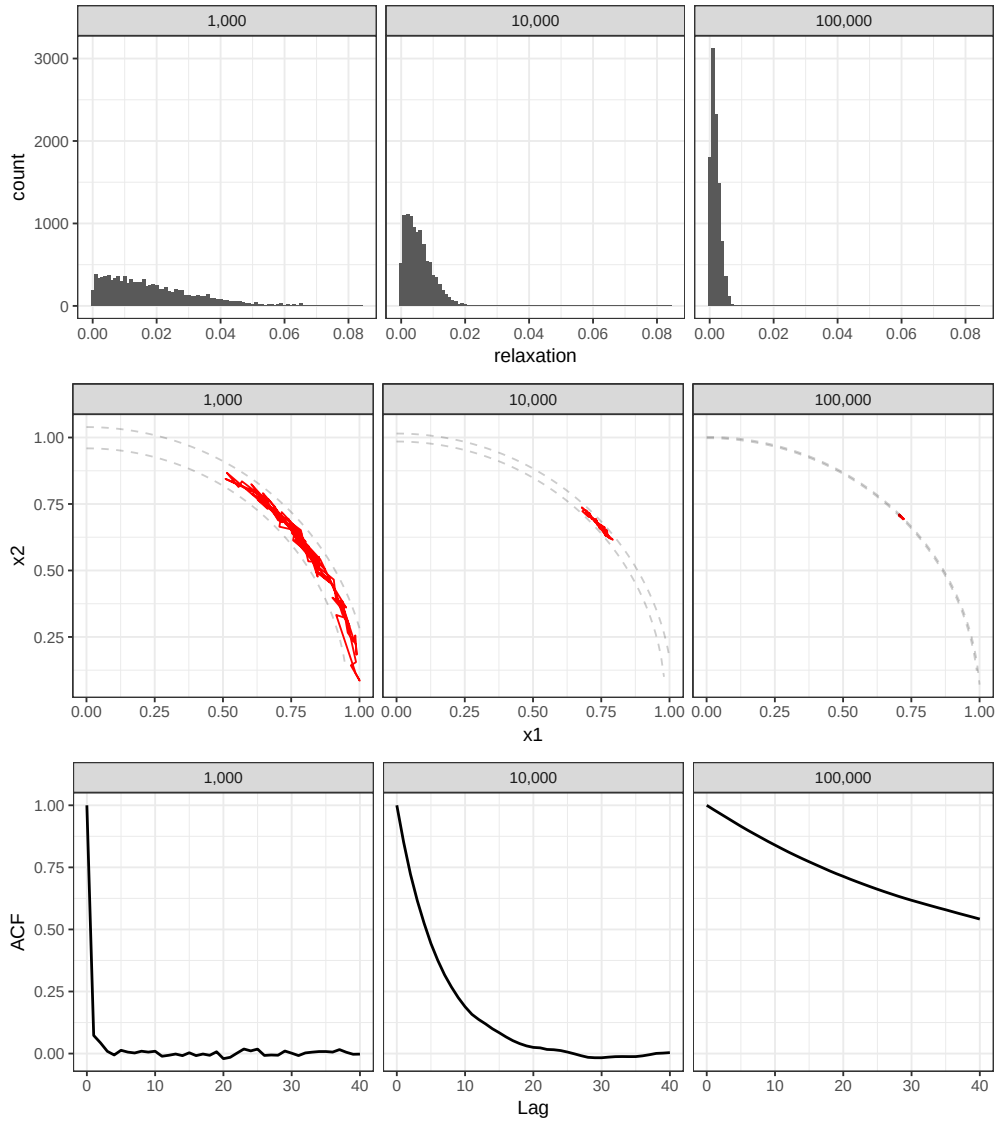


Figure 2: Sampling posterior from a von Mises–Fisher distribution on a unit circle, using HMC with extrinc prior under $\lambda = 10^3, 10^4, 10^5$. Row 1 shows the posterior distribution of the constraint relaxation $|\theta'\theta - 1|$; Row 2 shows the path of 100 leap-frog steps; Row 3 shows the 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 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. We find that $\lambda = 10^3$ is a good emipirical value for all the equality constraints in this paper.

2.4 Correcting Projection to Constrained Space

The Markov chain produced by HMC is geometrically ergodic under very general conditions (Livingstone et al., 2016). With the extrinsic posterior $\pi_K(\theta \mid y)$ as approximation to (2), one may be interested in further obtaining exact posterior in $\mathcal{D}$, likely for two reasons: (i) to strictly uphold the constraints; (ii) to ease the strict relaxation control on extrisinc prior.

Letting θ^* be a random sample collected based on $\pi_K(\theta \mid y)$, 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:

$$1 \wedge \frac{\pi(\theta^*_{\mathcal{D}} \mid y)\pi_K(\theta \mid y)}{\pi(\theta_{\mathcal{D}} \mid y)\pi_K(\theta^* \mid y)} = 1 \wedge \frac{L(\theta^*_{\mathcal{D}}; y)\pi_{0,\mathcal{R}}(\theta^*_{\mathcal{D}}) \cdot L(\theta; y)\pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)}{L(\theta_{\mathcal{D}}; y)\pi_{0,\mathcal{R}}(\theta_{\mathcal{D}}) \cdot L(\theta^*; y)\pi_{0,\mathcal{R}}(\theta^*)\mathcal{E}(\theta^*)}. \tag{7}$$

The remaining task is then to optimize the projection with respect to the acceptance rate. Noting

$$|\log(\frac{\pi(\theta^*_{\mathcal{D}} \mid y)\pi_K(\theta \mid y)}{\pi(\theta_{\mathcal{D}} \mid y)\pi_K(\theta^* \mid y)})| \leq |\log(\pi(\theta^*_{\mathcal{D}} \mid y)) - \log(\pi_K(\theta^* \mid y))| + |\log(\pi(\theta_{\mathcal{D}} \mid y)) - \log(\pi_K(\theta \mid y))|, \tag{8}$$

it is sensible choose $\theta_{\mathcal{D}} \in \mathcal{D}$ to minimize the difference $Q(\theta_{\mathcal{D}}) = |\log(\pi(\theta_{\mathcal{D}} \mid y)) - \log(\pi_K(\theta \mid y))|$ towards 0 for each sample in the extrinsic posterior. Obviously, when the approximate $\theta \in \mathcal{D}$ exactly, the optimal projection would be the identity function; when $\theta \notin \mathcal{D}$, standard constrained optimization technique can be used.

Continuing the unit circle example, we first obtained posterior sample from $\pi_K(\theta \mid y)$ with $\lambda = 10^3$. The almost immediate drop to 0 in ACF indicates a rapid convergence to the target posterior $\pi_K(\theta \mid y)$. We then

obtain $\hat{\theta}_{\mathcal{D}} = \underset{\theta_{\mathcal{D}}:\theta'_{\mathcal{D}}\theta_{\mathcal{D}}=1}{\operatorname{argmin}} |F'\theta_{\mathcal{D}} - \{F'\theta - \lambda(\theta'\theta - 1)^2\}|$ and construct the exact Markov chain. The acceptance rate is 97.9%.

3 Theory

In this section, we establish the conditions for extrinsic posterior to achieve a good and efficient approximation to the true posterior. First it is important to ensure the posterior under constraint relaxation remains proper. Assuming the original constrained prior is proper, $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta < \infty$, one sufficient condition for the relaxed prior to be proper in $\pi_K(\theta)$ is obviously $\int_{\mathcal{R}\setminus\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)d\theta < \infty$. In the following, we show several general scenarios that is achieved.

Lemma 1. *Assuming the constrained prior $\pi_{0,\mathcal{D}}(\theta)$ is proper, let $g_k(\theta)$ be the relaxation for each constraint with $g_k(\theta) \in [0, \varepsilon_k)$ a.s. for $k = 1, \ldots, l$ and $l < \infty$. If $\pi_{0,\mathcal{R}}(\theta)$ is Lipschitz continuous with respect to every $g_k(\theta)$, then the posterior is proper.*

Proof. Since $\pi_{0,\mathcal{D}}(\theta)$ is proper, $\int_{\mathcal{D}} \pi_{0,\mathcal{R}}(\theta)d\theta \leq M_1$ with $M_1 < \infty$. Let $\mathcal{E}(\theta) = CK(\theta)$ where C is a finite constant and $K(\theta)$ is the kernel function such that $K(\theta) = 1$ when $\theta \in \mathcal{D}$, and $K(\theta) < 1$ when $\theta \notin \mathcal{D}$.

$$\int \pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)d\theta = C\int \pi_{0,\mathcal{R}}(\theta)K(\theta)\mathbb{1}_{\theta\in\mathcal{D}}d\theta + \int \pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)\mathbb{1}_{\theta\notin\mathcal{D}}d\theta$$

$$\leq CM_1 + \int \pi_{0,\mathcal{R}}(\theta)\mathcal{E}(\theta)\mathbb{1}_{\theta\notin\mathcal{D}}d\theta$$

□

4 Simulated Examples

To illustrate, we illustrate the extrinsic prior via three constrained models.

Example 1: Ordered Dirichlet Prior in Mixture Model

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

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

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

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

which is associated with the likelihood

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

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 space for prior $Dir(\alpha)$ can be viewed as embedded in $\mathcal{R} = (0,1)^J$ via an equality constraint $\sum_{j=1}^{J} w_j = 1$.

However, one known issue for mixture modeling under canonical Dirichlet prior is the label-switching problem. As the parameter $\{\mu_j, w_j\}$ is indexed by $j = 1, \ldots, J$, one can switch any two j and j' due to exchangability. It is controversial whether the occurence of label-switching or the lack thereof is more perferable (see review in Jasra et al. (2005)); on the other hand, if the posterior distribution is symmetric about any permutation in j's, as the above normal mixture model, sampling over all permutations is redundant. It is useful to avoid label-switching for better convergence. Unfortunately, even Gibbs sampling with local update sometimes cannot avoid label-switching, especially when sample size n is small, posterior variances of μ_j's can be quite large with overlap across components. To mitigate this problem, we put order constraint on $w_1 \geq w_2 \geq \ldots \geq w_J$, yielding an ordered Dirichlet prior:

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

where $w_j \in (0,1)$. This order constraint is particularly useful to prevent label-switching between large and small components. Unlike early post-hoc relabeling algorithm (Stephens, 2000), we remove exchangability directly to reduce label-switching. In early work, Diebolt and Robert (1994) also suggested ordering in μ_j's, although it is not clear how it would work with multi-dimensional $\mu_j \in \mathbb{R}^d$ with $d \geq 2$.

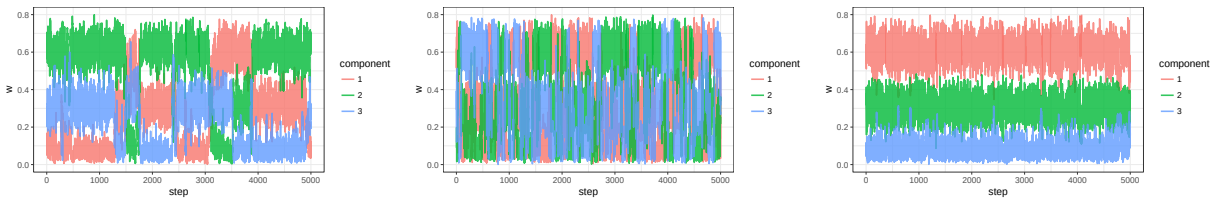
The ordering on simplex vertex disrupts the closed-form to be used in Gibbs sampling. Instead, observe that the ordered Dirichlet prior can be approximated by:

$$\pi_K(w_1, \ldots w_J) \propto \prod_{j=1}^{J} w_j^{\alpha-1} \cdot \prod_{j=1}^{J-1} K_1((w_{j+1} - w_j)_+) \cdot K_2(|\sum_{j=1}^{J} w_j - 1|)$$

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 $\mathbb{R}^J$. In comparison, we also test with $\lambda_1 = 0$ to remove the order constraint and allow HMC to run on a canonical Dirichlet prior without ordering.

We first generate $n = 100$ samples from 3 components with true $\{w_1, w_2, w_3\} = \{0.6, 0.3, 0.1\}$, 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 traceplot of w_j's in three approaches: standard Gibbs sampling with augmented component assignment under canonical Dirichlet prior, HMC using extrinsic prior associated under canonical Dirichlet prior and and HMC using extrinsic prior under ordered Dirichlet. Each approach runs $10,000$ iterations with first $5,000$ discarded as burn-in. For the posterior extrinsic collected under extrinsic prior, a projection is used as proposal in Metropolis-Hastings correction. Due to small sample size and relatively close component means, significant label-switching is shown in both Gibbs and HMC with canonical Dirichlet prior; while HMC with ordered Dirichlet prior does not suffer this issue.



(a) Gibbs sampling under unordered Dirichlet (b) HMC sampling under unordered Dirichlet, using extrinsic prior (c) HMC sampling under ordered Dirichlet, using extrinsic prior

Figure 3: Traceplot of the posterior sample for the simplex vertices, used as the component weights in a 3-component normal mixture model. Extrinsic prior allows the inclusion of order constraint in the simplex, reducing the label-switching issue.

Example 2: Convex Curve

We now illustrate a curve fitting problem where the shape of curve is convex. Convexity is common in real life such as a trajectory of projectile or accelerated decreasing of organ functions in disase monitinoring. Consider a cubic spline function $f(t)$ for data y_t with $t \in [0, 1]$

$$y_t = f(t) + \epsilon_t,$$

$$f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{j=1}^{J} b_j (t - \tau_j)_+^3,$$

where $\epsilon_t \sim \text{No}(0, \sigma^2)$ and τ_j's are pre-specified knots in $(0, 1)$. To induce the convexity, it suffices to have the second derivative:

$$f''(t) = 2\beta_2 + 6\beta_3 t + \sum_{j=1}^{J} 6 b_j (t - \tau_j)_+ \geq 0.$$

Given data y_t at observed time $t = t_1, \ldots, t_n$, the posterior estimation of $\beta_{\cdot}$ and $b_{\cdot}$ is estimated under n linear inequality constraints.

5 Application

6 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.

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

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.

Kelly, C. and J. Rice (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 1071–1085.

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., V. Rao, and D. B. Dunson (2016). Bayesian nonparametric inference on the stiefel manifold. *Statistica Sinica*.

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*.

Murray, I., Z. Ghahramani, and D. MacKay (2012). Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

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.

Rao, V., L. Lin, and D. B. Dunson (2016). Data augmentation for models based on rejection sampling. *Biometrika 103*(2), 319–335.

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

Uschmajew, A. (2010). Well-posedness of convex maximization problems on stiefel manifolds and orthogonal tensor product approximations. *Numerische Mathematik 115*(2), 309–331.

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.