Extrinsic Prior for Simple and Efficient Bayesian Modeling with Parameter Constraints

Leo Duan, Akihiko Nishimura, David Dunson

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 uncertainty quantification in constrained space. Although customized solutions have been developed in the past, it is still difficult to carry out estimation in general cases, especially when posteriors lack closed-form. In this paper, we propose a simple and general solution by 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 obtain approximate posterior efficiently. If one needs to uphold the constraints, the posterior sample can be projected back to the constrained space to obtain exact solution. We illustrate various cases with equality and inequality constraints. As priors are no longer limited to ones with closed-form posterior, more distribution families can be chosen for the constrained parameters; constraints can be freely adopted for desired property, such as improving convergence. Theory is developed and novel statistical applications under constraints are demonstrated.

KEY WORDS: Constraint relaxation; Euclidean Embedding; Monotone Dirichlet; Soft Constraint; Stiefel Manifold; Projected Markov chain

1 Introduction

Constraints are common 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 that impacts identifiability (Uschmajew, 2010); many manifolds such as simplex can be considered as sub-manifolds of a Euclidean space via some constraint.

When parameters are constrained, challenges often arise in their estimation. 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 via optimization, as asymptotic result for variance estimation in Euclidean space no longer holds in constrained space. In this regard, Bayesian approach is particularly advantageous.

There have been a variety of customized solutions developed for specific constraints. One popular strategy rely on picking constrained prior with posterior easy to sample. 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 with posterior conjugacy, which gives its popularity in matrix and tensor decomposition. To improve its flexibility, Lin et al. (2016) extends the matrix von-Mises-Fisher distribution via non-parametric Bayes approach. When the simple posterior form is not directly available, 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. Both strategies "intrinsically" meet the constraint, therefore are commonly referred as intrinsic approaches. Despite their success, there are several issues. When the parameter is multi-dimensional, one often has to use Gibbs sampling to update one dimension at a time, leading to inefficient computation. Moreover, the closed-form of posterior distribution is prone to break under slightly more advanced model or complicated data assumption. For example, for modeling population of undirected networks, symmetry in each network can disrupt the posterior conjugacy in orthogonal tensor decomposition, demanding new customized sampling algorithm to be developed. As another example, additional structure (such as ordering) on the probability simplex breaks the stick-breaking formulation.

These drawbacks have motivated the development of extrinsic approaches. The key idea is to first generate 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 suggested 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 a deterministic projection to map the unconstrained posterior into the constrained space and obtain monotonicity and manifold-valued regression. In Hamiltonian Monte Carlo, Neal (2011) suggested using large penalty to create an energy wall to induce 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 cases work well, but those settings are very specific and there is a clear lack of general and simple approach.

In this paper, we propose a new 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). 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. 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 adoption of constraints in modeling. Theoretic studies are conducted and original models are shown in simulations and data application.

2 Method

We consider parameters θ in a constrained space $\mathcal{D}$. Both θ and $\mathcal{D}$ can be high or infinite-dimensional. Letting $\mathcal{D}$ be equipped 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 available choices in both priors and constraints are very limited.

Instead, we consider extrinsic approaches by estimating θ in a 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 the constrained prior can be viewed as a conditional density, based on another density $\pi_{0,\mathcal{R}}(\theta)$ in $(\mathcal{R}, \mathscr{R})$ with $\mathscr{R}$ 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. For example, in matrix von Mises–Fisher distribution for θ on Stiefel manifold $\mathcal{D}(\theta) = \mathcal{V}(N, d)$, $\pi_{0,\mathcal{D}} = \frac{\exp(\mathrm{tr}(F'\theta))\mathbb{1}_{\theta'\theta=I_d}}{{}_0G_1(F)}$, with ${}_0G_1(F) = \int_{\mathcal{D}} \exp(\mathrm{tr}(F'\theta))d\theta$ as confluent hypergeometric limit function. 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. Therefore, (2) shows that if $\pi_{0,\mathcal{R}}(\theta)\mathbb{1}_{\theta\in\mathcal{D}}$ is proper, one can use any prior in $\mathcal{R}$ to couple with constraint, and obtain valid posterior in $\mathcal{D}$,

2.1 Extrinsic Prior

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 ratio of probabilities $Pr(\theta \in \mathcal{D} \mid y)/Pr(\theta \in \mathcal{R} \mid y) \approx 0$, commonly 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 a strongly informative prior with density $\mathcal{K}(\theta)$. This prior has support $\mathcal{S}$ that $\mathcal{D}\subset\mathcal{S}\subset\mathcal{R}$ with its mass concentrated around $\mathcal{D}$. Then one can first obtain posterior based on density proportional to $L(\theta;y)\pi_{0,\mathcal{R}}(\theta)\mathcal{K}(\theta)$.

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. Letting there be m equalities and l inequalities, this leads to embedding $\mathcal{D}=\{\theta\in\mathcal{R}: E_k(\theta)=0 \text{ for } k=1,\ldots,m, \quad G_{k'}(\theta)\le 0 \text{ for } k'=1,\ldots,l\}$, where $E_k(.)$ and $G_{k'}(.)$ are functions that map from $\mathcal{R}$ to real line $\mathbb{R}$. Then the indicator function in (2) can be broken into $\mathbb{1}_{\theta\in\mathcal{D}}=\prod_k \mathbb{1}_{E_k(\theta)=0}\cdot\prod_{k'}\mathbb{1}_{G'_k(\theta)\le 0}$.

We now form $\mathcal{K}$ by replacing each indicator functions with a kernel function $K_.(.)$, this yields posterior:

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

where $(x)_+ = x$ if $x>0$, 0 if $x\le 0$. For example, one simple kernel is the truncated Gaussian kernel $K_{i,k}(x)=\exp(-\lambda_{i,k}x^2)\mathbb{1}_{x<\varepsilon(\lambda_{i,k})}$, where $\varepsilon(\lambda_{i,k})$ is a truncation bound depends on $\lambda_{i,k}$ (more will be explained in next section). Generally, the posterior value of 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 $K_{i,k}(0)=1$ when the constraints are met; the tolerable amount of relaxation is controlled by hyper-parameter $\lambda_{i,k}$. When $\lambda_{i,k}\to\infty$, the kernel becomes a point mass at 0. Therefore, (2) is a special limiting case of (3).

When $\lambda_{i,k}$'s take large but finite values, they give 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 influence of the strongly informative prior $\mathcal{K}(\theta)$. 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 $\mathcal{D}$, if needed. We use subscript $\mathcal{K}$ to denote posterior $\pi_{\mathcal{K}}(\theta\mid y)$, which can be viewed as an approximation to $\pi(\theta\mid y)$ in (2). We will now refer $\pi_{\mathcal{K}}(\theta\mid y)$ as "extrinsic posterior".

2.2 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.

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 $\int_{v<\varepsilon} \mathcal{K}^*_{i,k}(x) / \left(\int_{\mathcal{R}} K^*_{\cdot}(v) dv \right) = 1 - \eta$ 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 $\theta \mid y \sim \mathrm{No}_{(\alpha,\beta)}(0, 1)$, with mean 0 and variance 1 and truncation $\theta \in (\alpha, \beta)$. The exact and extrinsic posterior densities are:

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

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 $(\theta - \beta)_+$.

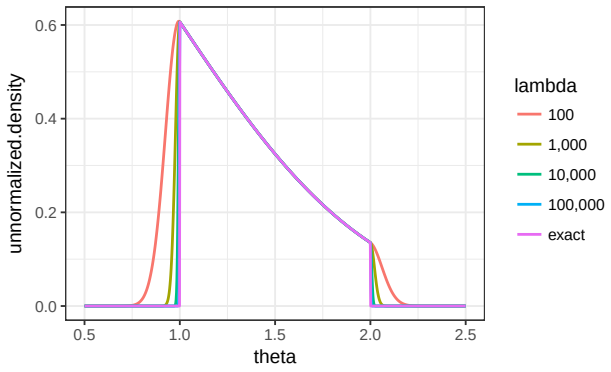


Figure 1: Unnormalized densities for truncated normal $\mathrm{No}_{(1,2)}(0, 1)$, under exact $\pi(\theta \mid y)$ and extrinsic posterior $\pi_{\mathcal{K}}(\theta \mid y)$. The exact density abruptly drops to 0 on the two boundaries, while the approximating ones drop continuously. In the approximation, 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.

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.

2.3 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

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

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:

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

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, the negative log-posterior function is unchanged $H(\theta(t), p(t)) = H(\theta(0), p(0))$. 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 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.

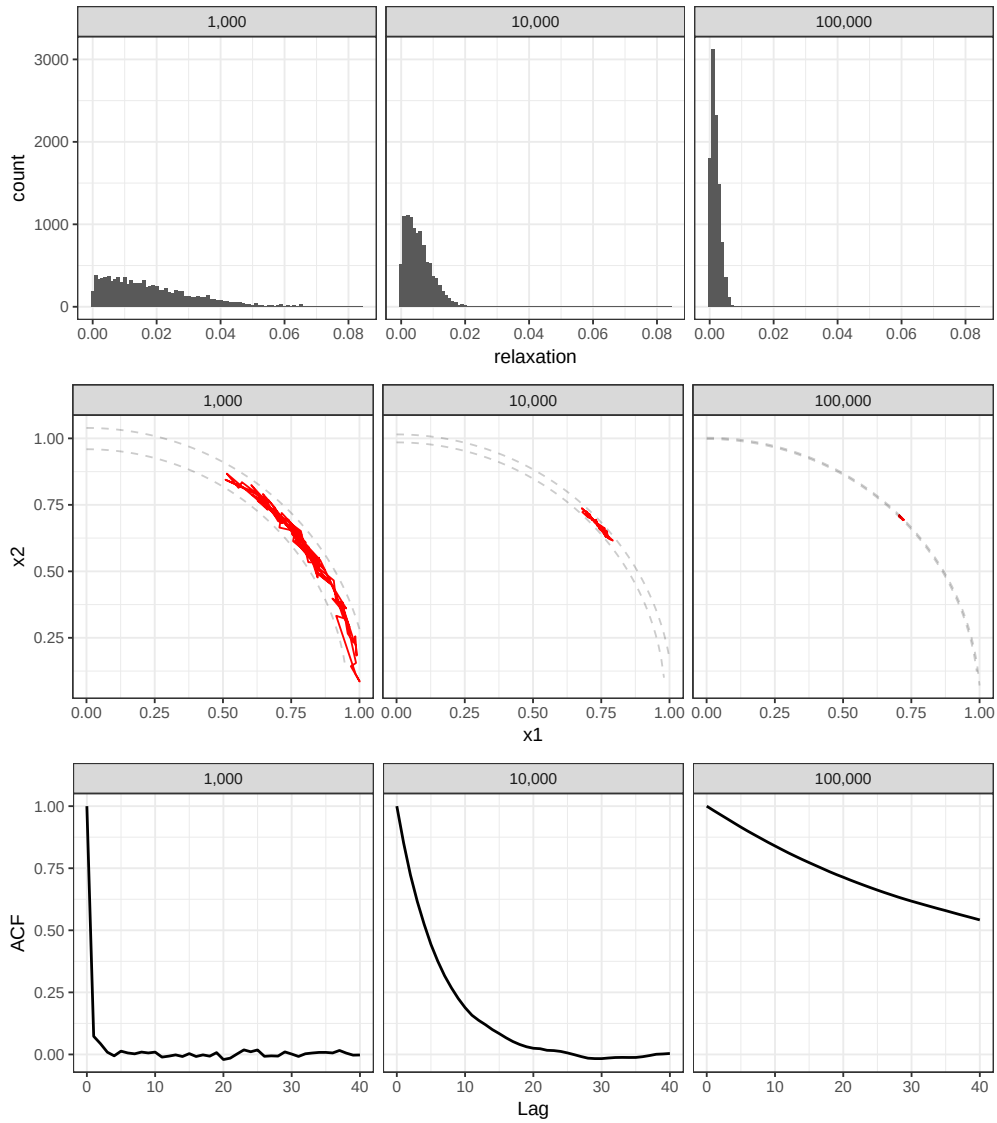


Figure 2: Sampling posterior from a von Mises–Fisher distribution on a unit circle, using HMC with extrinc prior under $\lambda = 10^3, 10^4, 10^5$. Row 1 shows the posterior distribution of the constraint relaxation $|\theta'\theta - 1|$; Row 2 shows the path of 100 leap-frog steps; Row 3 shows the 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.

2.4 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.

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.

2.5 Correcting Projection for Hard Constraint

The extrinsic posterior $\pi_{\mathcal{K}}(\theta \mid y)$ is an approximation to (2) 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:

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

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:

$$\pi_{0,\mathcal{D}}(w) \propto \prod_{j=1}^{J} w_j^{\alpha-1} \, \mathbb{1}_{\sum_{j=1}^{J} w_j = 1} \tag{8}$$

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 $j = 1, \dots, J$, 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:

$$\pi_{0,\mathcal{D}}(w_1, \ldots w_J) \propto \prod_{j=1}^{J} w_j^{\alpha-1} \cdot \mathbb{1}_{\sum_{j=1}^{J} w_j = 1} \cdot \prod_{j=1}^{J-1} \mathbb{1}_{w_j \geq w_{j+1}} \,. \tag{9}$$

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:

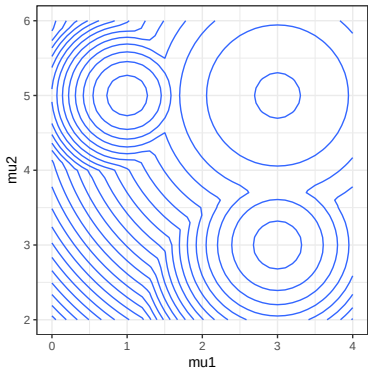
$$\pi_{0,\mathcal{R}}(w) \cdot \mathcal{K}(w) \propto \prod_{j=1}^{J} w_j^{\alpha-1} \cdot \prod_{j=1}^{J-1} K_1((w_{j+1} - w_j)_+) \cdot K_2(|\sum_{j=1}^{J} w_j - 1|)$$

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 (8).

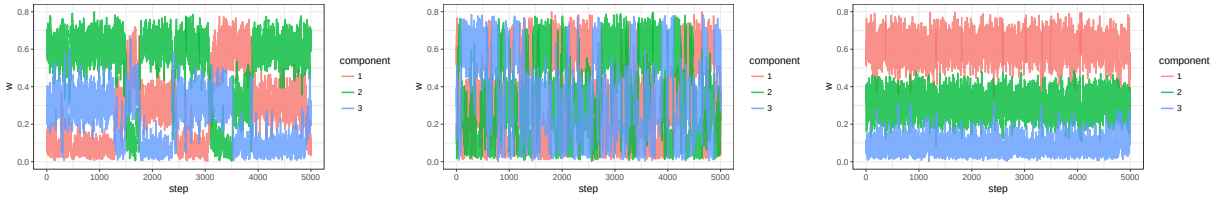
We generate $n = 100$ samples from 3 components with true $\{w_1, w_2, w_3\} = \{0.6, 0.3, 0.1\}$, with corresponding two-dimensional means $\{\mu_1, \mu_2, \mu_3\} = \{[1, 5], [3, 3], [3, 5]\}$ and identity covariance $\Sigma = I_2$. We assign informative priors $\mathrm{No}(0, 10I_2)$ for each μ_j and inverse Gamma prior for the digonal element in $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$ 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 (8), HMC using extrinsic prior associated under canonical prior (8) and and HMC using extrinsic prior under ordered prior (9). 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 Gaussian Processes in Functional Principle Component Analysis

leo: There are some issues with this model. Skip it for now.

We now consider functional principle component analysis. Letting x_i be the input for $i = 1, \ldots, n$ in functions $f_j(x_i)$ for $j = 1, \ldots, p$. We observe functional data $y_{i,j}$ as noisy realizations of $f_j(x_i)$. Commonly, p is very large and it is useful to view the functions as linear combination of d functional factors g_k.

$$y_{ij} = f_j(x_i) + \epsilon_{ij},$$

$$f_j(x_i) = \sum_{k=1}^{d} \eta_{jk} g_k(x_i),$$

$$[g_k(x_1), \ldots, g_k(x_n)] \sim \text{No}(0, \Sigma_k)$$

$$\Sigma_{k,(i,i')} = \phi_k \exp(-\frac{||x_i - x_{i'}||^2}{2\rho_k^2})$$

for $k = 1, \ldots d$ with $d < p$, and $\epsilon_{ij} \sim \text{No}(0, \sigma^2)$ is the random measurement error; $\Sigma_{k,(i,i')}$ is the (i, i')th

element in matrix Σ_k. Using matrix notation $G = [g_1, \ldots, g_d]$ and $\eta = [\eta_{.1}, \ldots, \eta_{.d}]$, the $n \times p$ function matrix $[f_j(x_i)]_{ij}$ can be written as $G\eta'$. We utilize a squared exponential Gaussian process to model each latent factor g_k.

We first assign a shrinkage prior on loadings $\eta_{jk} \sim \text{No}(0, \tau_k)$, $\tau_k \sim IG(aq^{3(k-1)}, q^{2(k-1)})$ with $q > 1$. This prior ensures the shrinkage grows stronger as k increases (Bhattacharya et al., 2011), avoiding arbitrary specification of d and exchangability in permuting k. However, G and η are still not identifiable. Any orthonormal matrix P that $P'P = I_d$ can produce another set of factors $G^* = GP'$ and loadings $\eta^* = P\eta$. Since this projection P is associated with rotation, scaling or column-wise sign change, we apply the following constraint on G:

$$\sum_{i=1}^{n} g_k(x_i) g_{k'}(x_i) = \begin{cases} 1 \text{ if } k = k' \\ 0 \text{ if } k \neq k' \end{cases}$$

$$g_k(x_1) \geq 0$$

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

As the result, these constraints create a Gaussian process prior on a Stiefel manifold $\mathcal{V}(N, d)$. To our best knowledge, this is the first application of Gaussian process in this manifold. We approximate these constraints with extrinsic prior:

$$\pi_{\mathcal{K}}(G) \propto \prod_{k=1}^{d} K_1((-g_k(x_1))_+) \cdot \prod_{k'=1}^{d} \prod_{k=1}^{k'} K_2(|\sum_{i=1}^{n} g_k(x_i) g_{k'}(x_i) - \delta_{k,k'}|)$$

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 strongly enforce the sign of $g_k(x_1)$ to be positive, while $\lambda_2 = 10^3$ to allow efficient mixing in embedding Stiefel manifold in $\mathbb{R}^{N \times d}$. To compare with unconstrained Gaussian processes for g_k's, we also test with $\lambda_1 = 0$ and $\lambda_2 = 0$.

We generate $n = 50$ inputs $x_i \sim \text{U}(0, 1)$ from uniform $(0, 1)$ and three smooth functions $g_1^*(x) = \sin(16x)/x$, $g_2^*(x) = \sin(25x) \cdot x$ and $g_3^*(x) = \cos(20x)/x$. The functions are combined via $f_j(x_i) = \sum_{i=1}^{3} \eta_{jk}^* g_k^*(x_i)$ with $\eta_{jk}^* \overset{iid}{\sim} \text{No}(0, 1)$ for $j = 1, \ldots, 20$ and $i = 1, \ldots, n$. We add random noise to create a 50×20 data points $y_{ij} \sim \text{No}(f_j(x_i), 0.1^2)$, and randomly remove 20% of data to mimic the unbalanced data in real world. We set $a = q = 2$ in the shrinkage prior for all η_{jk}'s, uniform prior $\text{U}(0, 20)$ for all ρ_k's, ϕ_k's and σ^2.

Example 3: Orthonormal Tucker Factorization in Multiple Network Analysis

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.

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.