Extrinsic Prior for Simple and Efficient Bayesian Modeling with Parameter Constraints

Abstract: Models

KEY WORDS: Bayesian probit; Bayesian logit; Big n; Data Augmentation; Maximal Correlation; Polya-Gamma.

1 Introduction

Constraints are very common in statistical modeling. In applied domain, modeling assumptions often require some constraint. For example, in functional data analysis related to degenerative disease, it is common to assume the curve need to satisfy certain shape constraint such as monotonicty (Lin and Dunson, 2014). In stastical optimization, constraints such as orthonormality are also routinely used to ensure identifiability of the model (Uschmajew, 2010). In manifold modeling, a large class of manifolds can be viewed as submanifolds of a more conventional space (e.g. Euchlean space), embedded via different constraints.

These constraints can cause substantial modeling difficulty. When constraints are applied on the data, they could generate an intractable integral with the parameter in the likelihood, leading to a "doubly intractable" problem. Several successful solutions have been proposed to address this issue (Murray et al., 2012; Rao et al., 2016). On the other hand, there is a clear lack of general and simple solution, when the constraints are applied on the parameters. In frequentist optimization literature, the use of Lagrange and Karush-Kuhn-Tucker multipliers provides a means to obtain point estimate under the equality and inequality constraints (Boyd and Vandenberghe, 2004). But due to the space constraint, the standard asympotic approximation for variance estimation usually do not hold. Therefore, a Bayesian approach would be more approriate to quantify the uncertainty. Ideally, one would assign prior on a constraint support, then utilize standard toolbox such as Markov chain Monte carlo (MCMC) to obtain posterior sample on this space. However, this turns out to be very challenging.

To assign prior on the constraint space, the available families of distribution are often quite limited. For example, for othornormal matrices on the Stiefel manifold, the matrix von Mises-Fisher distribution (Khatri and Mardia, 1977) is one of the only few choices. For regression under linear inequality constraints, only until recently a tractable prior is proposed for the polyhedral region set by the inequailities (Danaher et al., 2012). Alternatively, Gelfand et al. (1992) proposed a truncation strategy by first considering common unrestricted distribution, then assigning zero support outside the constraint region. Accordingly, the posterior estimation proceeds in first generating unrestricted proposals using Gibbs sampling, then only accepting those inside the constraint space. Although this approach allows using a more general class of prior distribution, the drawback is that the unrestricted proposal can have significant mass outside the constraint region, resulting in a high rejection rate.

To meet the constraints, efficient computation is elusive and often demands substantial efforts to develop. And often it can be disrupted by slight complication such as hierarchical structure or additional constraint. For example, stick-breaking parameterization is commonly used in probability simplex modeling, in order to circumvent the constraint of vertices summing to 1. However, its computational efficency can be broken by additional structure constraint, such as the ordering of the simplex, which is useful in reducing the label switching problem in mixture modeling (Diebolt and Robert, 1994). As another example, in multiway tensor factorization, orthonormality is useful to induce good posterior mixing in estimating factor matrices. This largely relies on the sampler of Bingham-von Mises-Fisher distribution (Hoff et al., 2016). However, when there is symmetry in the slices of tensor (commonly in population of undirected networks), at least two factors would be the same. This disrupts the closed form of the posterior, demanding new rejection sampling algorithm to be developed. The Bayesian manifold modeling also faces the same quadmire. Hamiltonian Monte Carlo accomodating the geometric structure of the manifold have been developed (Girolami and Calderhead, 2011; Byrne and Girolami, 2013), but the computation is intensive and mixing is suboptimal. These challenges prohibit a good utilization of constraints in statistics.

We propose to solve this problem by viewing the constrigent constraints as a limiting case of a strongly informative prior, referred as extrisic prior. We then relax the effective support of the prior to the neighbor of the constraint space, allowing approximate posterior to be collected efficiently via Hamiltonian Monte Carlo directly in Euchledean space. The imperfection of approximation can be corrected by an efficient reconstruction of an exact Markov chain after projection. The proposed approach is simple to implement and can be automatically carried out via software such as STAN. Theoretic studies are conducted and substantial improvement is shown in simulations and data application.

2 Method

unless the likelihood involve

intractable problem.

can usually be considered as assigning a prior in a

The task under consideration involves a p-dimensional parameter θ in a constrained space $\mathcal{D}$. Without loss of generality, the space $\mathcal{D}$ is assumed to be embedded in another space $\mathcal{R}$ (e.g. Euchledean space $\mathbb{R}^p$) via m equalities and l inequalities, $\mathcal{D} = \{\theta \in \mathcal{R} : E_k(\theta) = 0 \text{ for } k = 1, \ldots, m, \quad G_{k'}(\theta) \leq 0 \text{ for } k' = 1, \ldots, l\}$, where $E_k(.)$ and $G_{k'}(.)$ are functions that map from $\mathcal{R}$ to real line $\mathbb{R}$. These functions are differentiable with respect to θ and not necessarily linear.

Throughout this section, we use one example to illustrate the embedding and the method. Consider an *ordered* $(d-1)$-simplex. The parameter is a d-dimensional probability vector $\theta = \{p_1, \ldots, p_d\}$ with $p_1 \geq p_2 \geq \ldots \geq p_d$. Its space $\mathcal{D}$ is embedded in $[0,1]^d$ via $d-1$ inequality constraints $p_{i+1} - p_i \leq 0$ for $i = 1, \ldots, d-1$ and one equality constraint $\sum_{i=1}^{d} p_i - 1 = 0$. Alternatively, one can view the space $\mathcal{D}$ as embedded in a broader space $\mathbb{R}^d$, via additional d identity inequalities $p_i \geq 0$ for $i = 1, \ldots, d$; however this is not necessary since in general, constraints via identity functions as such are trivial to handle. Therefore, from now on we assume that all chosen space $\mathcal{R}$ has already accommodated the simple identity constraints, using space truncation.

We hope to obtain statistical inference on θ in this constrained space $\mathcal{D}$. Letting $L(\theta; y)$ be the likelihood and $\pi_0(\theta \mid \theta \in \mathcal{D})$ be the prior, given observed data y, we are interested in the posterior distribution:

$$\pi(\theta \mid y, \theta \in \mathcal{D}) = \frac{L(\theta; y)\pi_0(\theta \mid \theta \in \mathcal{D})}{\int_{\mathcal{D}} L(\theta; y)\pi_0(\theta \mid \theta \in \mathcal{D}) d\theta}, \tag{1}$$

where the prior $\pi_0(\theta \mid \theta \in \mathcal{D}) = \frac{\pi_0(\theta)}{\int_{\mathcal{D}} \pi_0(\theta) d\theta}$ with $\pi_0(\theta)$ defined in $\mathcal{R}$. Due to the space integrated over, $\int_{\mathcal{D}} \pi_0(\theta) d\theta$ often lacks closed-form; but since it is a constant, one commonly utilize:

$$\pi(\theta \mid y, \theta \in \mathcal{D}) \propto L(\theta; y)\pi_0(\theta), \quad \theta \in \mathcal{D} \tag{2}$$

To satisfy $\theta \in \mathcal{D}$, it often demands substantial efforts. One costly strategy is to first propose $\theta \in \mathcal{R}$, then reject those violating any of the constraints (i.e. $\theta \notin \mathcal{D}$) (CITES Gelfand et al 1992). Alternatively, one relies on a skillful re-parameterization of θ to meet the constraint implicitly. For example, in manifold modeling, one often switches to the coordinate system instead of using θ directly; in unordered simplex modeling, one uses stick-breaking construction instead to meet the fixed 1-norm constraint. However, any complication like the ordered simplex example would disrupt the solution, making the estimation substantially more difficult.

We now propose a different strategy by replacing the constringent embedding with a set of strongly informative priors. Specifically, instead of modeling $\theta \in \mathcal{D}$, we relax its space to an emcompassing and tight neighborhood in $\mathcal{R}$. This is achieved through adding $(m + l)$ kernel functions $K_{.}(.)$, leading to posterior:

$$\pi(\theta \mid y) \propto L(\theta; y)\pi_0(\theta) \cdot \prod_{k=1}^{m} K_{1,k}\Big(E_k(\theta)\Big) \cdot \prod_{k'=1}^{l} K_{2,k'}\Big(\big(G_{k'}(\theta)\big)_+\Big), \quad \theta \in \mathcal{R} \tag{3}$$

where $(x)_+ = x$ if $x > 0$, 0 if $x \leq 0$. All the kernel functions $K_\cdot(.)$ satisfy $K_\cdot(0) = 1$, so that when $\theta \in \mathcal{D}$ exactly (recall $\mathcal{D} \subseteq \mathcal{R}$), the posterior density $\pi(\theta \mid y)$ is the same as the strict embedding case, except for a constant proportional difference. And $K_\cdot(x)$ declines repaidly when $x \neq 0$. For example, one simple kernel for this purpose is the Gaussian kernel $K_{i,k}(x) = \exp(-\lambda_{i,k} x^2)$ with large λ_{ik}.

The kernels are part of prior densities that handle the constraints via an *extrinsic* approach. We therefore refer them as extrinsic priors. The posterior obtained using (3) is an approximation to those obtained in a strict embedding approach. One can simply use them for the approximate statistical inference, or use correcting projection to map them back to $\mathcal{D}$. In this article, we focus on the latter.

2.1 Prior Specification

The tightness of the neighborhood is governed by the hyper-parameters in the extrinsic prior. It has mainly two effects on the posterior estimation: (1) the approximation accuracy for the constraints; (2) the posterior mixing, which is related to the convergence speed and posterior autocorrelation of the Markov chain. In general, tighter neighborhood leads to slower mixing. Therefore, a balance needs to be struck when choosing the hyper-parameters.

Let $\mathcal{E}_{i,k}(\theta) = c_{i,k} K_{i,k}(x)$ be the normalized extrinsic prior for the (i,k)th constraint ($i = 1$ equality, $i = 2$ inequality), where $c_\cdot = 1/\big(\int_{\mathcal{R}} K_\cdot(x) dx\big)$. To construct a tight neighborhood, we require $\int_{|x|<\epsilon} \mathcal{E}_{i,k}(\theta) = 1 - \eta$ with $\eta > 0$ negligibly small. The constant ϵ is pre-specified and represents the element-wise tolerance for violating each constraint. The amount of violation is reflected in the posterior values of $E_k(\theta)$ and $\big(G_k(\theta)\big)_+$. For example, in the Gaussian kernel, setting $\lambda_\cdot = \frac{1}{2(\epsilon/3)^2}$ ensures each error is contained within a radius of ϵ from 0 with each marginal probability 0.997.

leo: We probably need some regulartiy condition on $\pi_0(\theta)$ and $L(y;\theta)$, originally defined on $\mathcal{D}$: when the space is extended to $\mathcal{R}$, they should not have a big increase outside of $\mathcal{D}$ (e.g. $\pi_0(\theta) \to \infty$ would be bad).

So far we have taken an element-wise approach for specifying the kernel $K_\cdot$'s. It is possible to specify $K_\cdot$'s in a dependent way and contain the approximation error with probability better than $(1-\eta)^{(l+m)}$. However, unless $l + m$ is very large, this is often unnecessary since a correcting projection will be made to erase the approximation error, after the posterior is collected. Therefore, we focus on controlling the error in acceptable range, allowing good mixing and easy projection of θ back to $\mathcal{D}$.

leo: more work is to be done on assessing the effects on mixing

Example 1: Ordered Simplex

Example 2: Monotone Spline

Example 3: Orthonormal Gaussian Processes

2.2 Posterior Sampling

2.3 Correcting Projection

2.4 Illustration: Ordered Simplex

3 Theory

4 Application

5 Discussion

References

Boyd, S. and L. Vandenberghe (2004). *Convex optimization.* Cambridge university press.

Byrne, S. and M. Girolami (2013). Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics 40*(4), 825–845.

Danaher, M. R., A. Roy, Z. Chen, S. L. Mumford, and E. F. Schisterman (2012). Minkowski–weyl priors for models with parameter constraints: an analysis of the biocycle study. *Journal of the American Statistical Association 107*(500), 1395–1409.

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

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*.

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

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

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