Extrinsic Prior for Simple and Efficient Bayesian Modeling Under Constraint

1 Introduction

Literature review:

- History: Hoffman 1993, Gelfand, Smith and Lee 1992, Danaher et al 2012 – limited to linear constraint
- Manifold literature: Gibbs/ rejection sampling – slow mixing; HMC geodesics method – difficult and computationally intensive
- Lagrange/ KKT multiplier, Constraint Bayesian optimization – point estimate only, no uncertainty

Summary of proposed method: Replacing the inequality and equality constraints with strongly informative prior, allowing carrying out efficient Monte Carlo in a convenient space (e.g. Euchledian space). Obtain posterior, then use projection to correct the imperfection if needed.

Strength: Large support near the constraint space. Simple and efficient computation. Extremely easy to play, encouraging the community to plug-and-play more constraints in the model.

2 Method

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 use:

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

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 $K_\cdot(.)$, 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

2.2 Posterior Sampling

2.3 Correcting Projection

2.4 Illustration: Ordered Simplex

3 Theory

4 Application

5 Discussion