THE AMERICAN STATISTICIAN
2024, VOL. 78, NO. 4, 379–390: General
https://doi.org/10.1080/00031305.2024.2308821

Taylor & Francis
Taylor & Francis Group

GENERAL

Proximal MCMC for Bayesian Inference of Constrained and Regularized Estimation

Xinkai Zhou[a], Qiang Heng[b], Eric C. Chi[c], and Hua Zhou[a]

[a]Department of Biostatistics, UCLA, Los Angeles, CA; [b]Department of Computational Medicine, UCLA, Los Angeles, CA; [c]Department of Statistics, Rice University, Houston, TX

ABSTRACT
This article advocates proximal Markov chain Monte Carlo (ProxMCMC) as a flexible and general Bayesian inference framework for constrained or regularized estimation. Originally introduced in the Bayesian imaging literature, ProxMCMC employs the Moreau-Yosida envelope for a smooth approximation of the total-variation regularization term, fixes variance and regularization strength parameters as constants, and uses the Langevin algorithm for the posterior sampling. We extend ProxMCMC to be fully Bayesian by providing data-adaptive estimation of all parameters including the regularization strength parameter. More powerful sampling algorithms such as Hamiltonian Monte Carlo are employed to scale ProxMCMC to high-dimensional problems. Analogous to the proximal algorithms in optimization, ProxMCMC offers a versatile and modularized procedure for conducting statistical inference on constrained and regularized problems. The power of ProxMCMC is illustrated on various statistical estimation and machine learning tasks, the inference of which is traditionally considered difficult from both frequentist and Bayesian perspectives.

ARTICLE HISTORY
Received March 2023
Accepted January 2024

KEYWORDS
Hamiltonian Monte Carlo; Moreau-Yosida envelope; Proximal mapping

1. Introduction

Many statistical learning tasks are posed as regularized maximum likelihood estimation problems, which require solving optimization problems of the form

$$\text{maximize} \quad \ell(\boldsymbol{\theta}) - \rho P(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ denotes model parameters, $\ell(\boldsymbol{\theta})$ denotes the log-likelihood and quantifies the lack-of-fit between the model and the data, $P(\boldsymbol{\theta})$ is a regularization function that imposes structure on parameter estimates, and ρ is a nonnegative regularization strength parameter that trades off the model fit encoded in $\ell(\boldsymbol{\theta})$ with the desired structure encoded in $P(\boldsymbol{\theta})$. Canonical examples of regularization functions include the ℓ_1-norm that promotes sparsity and the nuclear norm that promotes recovery of low-rank models. To date, most work has focused exclusively on estimating $\boldsymbol{\theta}$ without quantifying the uncertainty in the estimates. Lacking tools for assessing uncertainty in findings from regularized models, practitioners often resort to classical inference tools designed for non-regularized models. This practice will substantially inflate the Type I error and lead to unreproducible scientific discoveries.

This issue has motivated the development of post-selection inference techniques such as simultaneous inference (Berk et al. 2013; Bachoc, Preinerstorfer, and Steinberger 2020; Kuchibhotla et al. 2020) and selective inference (Lee et al. 2016; Choi, Taylor, and Tibshirani 2017; Taylor and Tibshirani 2018). A closely related approach calculates confidence intervals for coefficients of high-dimensional linear models through bias-correction (van de Geer et al. 2014; Zhang and Zhang 2014; Javanmard and Montanari 2014). Most of this literature, however, focuses on variable selection through the ℓ_1-regularization. Extending these strategies to other regularizations and to problems involving constraints is not straightforward. Moreover, caution is warranted when reporting these confidence intervals because their interpretation (e.g., conditional on the selection event) differs from traditional ones.

An alternative is to cast the problem in the Bayesian framework. For example, Park and Casella (2008) introduced the Bayesian lasso, where the ℓ_1-regularization was identified with a Laplace prior and a Gibbs sampler was used to sample from the posterior distribution. This work is part of a large literature on Bayesian variable selection methods, which include sparsity inducing prior distributions such as spike-and-slab (Mitchell and Beauchamp 1988; George and McCulloch 1993), horseshoe (Carvalho, Polson, and Scott 2010; Polson and Scott 2010; Piironen and Vehtari 2017; Bhadra et al. 2019), orthant normal (Hans 2011), correlated Normal-Gamma (Griffin and Brown 2012, 2013), generalized double Pareto (Armagan, Dunson, and Lee 2013), and Dirichlet-Laplace (Bhattacharya et al. 2015). Despite constant innovations in Bayesian techniques for variable selection, incorporating regularizations and constraints beyond sparsity still requires a substantial amount of problem-specific analysis.

More recently, Pereyra (2016) and Durmus, Moulines, and Pereyra (2018, 2022) proposed the proximal Markov chain

CONTACT Hua Zhou huazhou@ucla.edu Department of Biostatistics, UCLA, Los Angeles, CA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

Monte Carlo (ProxMCMC) algorithm for quantifying uncertainty in Bayesian imaging applications where the regularizations of interest include the total-variation semi-norm (Rudin, Osher, and Fatemi 1992) and the ℓ_1-norm. To deal with the nonsmoothness of these regularizations, they employ the Moreau-Yosida envelope to obtain their smooth approximations. Samples from the smooth approximate posterior distribution can be drawn using Langevin dynamics. Their approach offers a framework for conducting statistical inference on regularized regression models whenever the regularization term is convex and admits a proximal map that can be computed efficiently, which holds true for a wide variety of regularizations. The fly in the ointment, however, is that their approach requires manually setting the regularization strength parameter ρ. One solution to this problem is given by Vidal et al. (2020) and De Bortoli et al. (2020), who proposed using an empirical Bayes method called the stochastic approximation proximal gradient (SAPG) to estimate the regularization strength parameter by maximum marginal likelihood. It only provides point estimates of the regularization strength parameter, potentially resulting in suboptimal statistical precision due to the neglect of uncertainty in the regularization strength parameter. In terms of flexibility, the SAPG approach focuses on regularized estimation problems, while constrained estimation problems remain relatively underexplored.

In this article, we address this limitation and extend ProxMCMC to be fully Bayesian by incorporating regularization and constraints through epigraph priors. Our extended ProxMCMC inference framework is suitable for regularized or constrained statistical learning problems and offers three main advantages. First, it provides valid and automatic statistical inference even for problems that involve non-smooth and potentially nonconvex regularization or constraints. The inference for such problems is traditionally considered difficult. Second, it is fully Bayesian, eliminating the need for parameter tuning. This is in contrast to previous ProxMCMC methods (Durmus, Moulines, and Pereyra 2018, 2022) where the regularization strength parameter is either manually fixed or requires tuning. Third, the method is highly modular. Its components—model, prior, proximal map, and sampling algorithm—are independent of each other and can be easily adjusted to address new problems. This feature makes ProxMCMC highly customizable, allowing users to tailor it to their specific problems. The practical significance of the last point cannot be emphasized enough and is exemplified in the constrained lasso example, where the "sum to zero" constraint, imposed by problem-specific considerations, causes existing inference methods to break down, but poses no challenge for the proposed ProxMCMC method. We will save the details for Section 5.1.

Finally, we put the proposed ProxMCMC method on firm foundations by providing guarantees on the properness of the approximate posterior and showing that the approximate posterior can be made arbitrarily close to the target posterior in total-variation under suitable assumptions.

The rest of the article is organized as follows. Section 2 reviews concepts from convex optimization that form the building blocks of the ProxMCMC framework. Section 3 illustrates our method using the familiar lasso problem. Section 4 summarizes the key elements from our case study of lasso to show how the ProxMCMC method can be applied generally. Section 5 presents a variety of illustrative applications, whose numerical results are presented in Section 6. Section 7 provides a brief discussion, while theoretical guarantees can be found in the supplementary materials.

2. Background

We review concepts from convex analysis essential for ProxMCMC, specifically Moreau-Yosida envelopes and proximal mappings. For a more thorough review of proximal mappings and their applications in statistics and machine learning, we refer readers to Combettes and Wajs (2005), Combettes and Pesquet (2011), and Polson, Scott, and Willard (2015). In convex optimization it is often convenient to work with functions that map into the extended reals, $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. The indicator function of a set C, denoted $\delta_C(\boldsymbol{x})$, is defined as

$$\delta_C(\boldsymbol{x}) = \begin{cases} 0 & \boldsymbol{x} \in C \\ \infty & \text{otherwise,} \end{cases} \tag{1}$$

which differs from the familiar $0/1$ indicator function used in statistics. A function $f : \mathbb{E} \to \bar{\mathbb{R}}$ is *lower-semicontinuous* at $\boldsymbol{x} \in \mathbb{E}$ if

$$f(\boldsymbol{x}) \leq \liminf_{n\to\infty} f(\boldsymbol{x}_n) \tag{2}$$

for any sequence $\{\boldsymbol{x}_n\}_{n\geq 1} \subseteq \mathbb{E}$ for which $\boldsymbol{x}_n \to \boldsymbol{x}$ as $n \to \infty$. A function is *proper* if it takes on a finite value for some element in its domain. When the set C is closed and convex, the indicator function $\delta_C(\boldsymbol{x})$ is lower-semicontinuous and convex. Let $\Gamma(\mathbb{R}^m)$ denote the set of all proper, lower-semicontinuous, convex functions from $\mathbb{R}^m$ into $\bar{\mathbb{R}}$. The Euclidean norm of a point $\boldsymbol{x}$ is denoted using the familiar notation $\|\boldsymbol{x}\|$.

2.1. Moreau-Yosida Envelopes and Proximal Maps

Definition 1. Given $g \in \Gamma(\mathbb{R}^m)$ and a positive scaling parameter λ, the *proximal mapping* of g is the operator given by

$$\operatorname{prox}_g^\lambda(\boldsymbol{x}) = \arg\min_{\boldsymbol{\omega}} \left\{ g(\boldsymbol{\omega}) + \frac{1}{2\lambda}\|\boldsymbol{\omega} - \boldsymbol{x}\|^2 \right\}.$$

Definition 2. Given $g \in \Gamma(\mathbb{R}^m)$ and a positive scaling parameter λ, the *Moreau-Yosida envelope* of g is given by

$$g^\lambda(\boldsymbol{x}) = \inf_{\boldsymbol{\omega}} \left\{ g(\boldsymbol{\omega}) + \frac{1}{2\lambda}\|\boldsymbol{\omega} - \boldsymbol{x}\|^2 \right\}.$$

The infimum is always attained at a unique point when $g \in \Gamma(\mathbb{R}^m)$, and the minimizer defines the proximal mapping of g.

Intuitively, evaluating the proximal mapping of g at $\boldsymbol{x}$ identifies a point $\boldsymbol{\omega}$ that balances between minimizing g and staying close to $\boldsymbol{x}$ in Euclidean distance. The extent to which $\boldsymbol{\omega}$ minimizes g is controlled by the positive scaling parameter λ: larger values of λ pushes $\boldsymbol{\omega}$ closer to the minimum, whereas smaller values keep $\boldsymbol{\omega}$ closer to $\boldsymbol{x}$. From the definition, we can see that the Moreau-Yosida envelope is related to the proximal mapping through the equation $g^\lambda(\boldsymbol{x}) = g(\operatorname{prox}_g^\lambda(\boldsymbol{x})) + \frac{1}{2\lambda}\|\operatorname{prox}_g^\lambda(\boldsymbol{x}) - \boldsymbol{x}\|^2$.

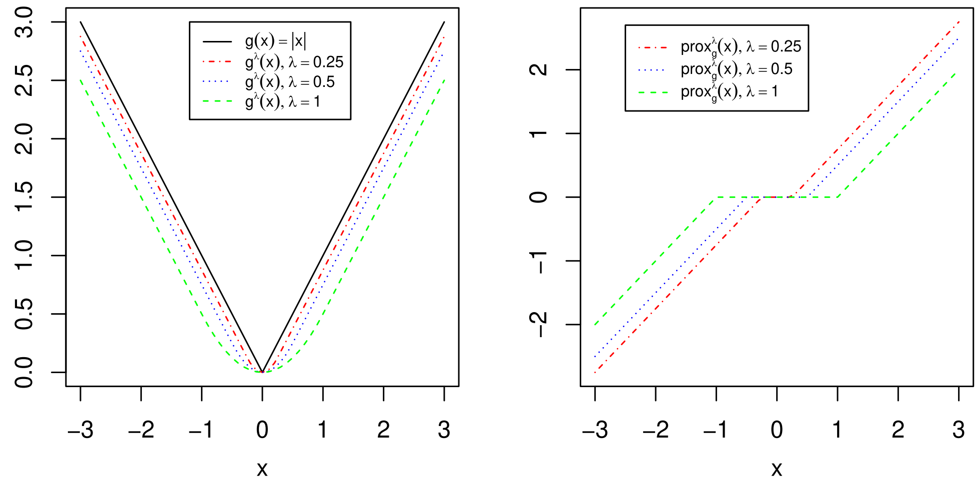

Figure 1. The Moreau-Yosida envelope (left) and proximal mapping (right) of the absolute value function $g(x) = |x|$.

We illustrate these definitions using the well known Huber function

$$g^\lambda(x) = \begin{cases} \frac{1}{2\lambda}x^2 & \text{if } |x| \leq \lambda \\ |x| - \frac{\lambda}{2} & \text{otherwise,} \end{cases}$$

which is the Moreau-Yosida envelope of the absolute value function $g(x) = |x|$. The left panel of Figure 1 shows $g(x)$ and $g^\lambda(x)$ for three different λ values. This familiar example from robust statistics shows that the Moreau-Yosida envelope provides a differentiable approximation to a non-smooth function where the approximation improves as λ gets smaller. The corresponding proximal map is the celebrated soft-thresholding operator $S_\lambda(x)$ defined by

$$S_\lambda(x) = \begin{cases} x - \lambda & \text{if } x > \lambda \\ 0 & \text{if } |x| \leq \lambda \\ x + \lambda & \text{if } x < -\lambda. \end{cases} \tag{3}$$

In the right panel of Figure 1, we show $\text{prox}_g^\lambda(x)$ for the same λ values as in the left panel.

In general, the Moreau-Yosida envelope $g^\lambda(\boldsymbol{x})$ has several important properties. First, $g^\lambda(\boldsymbol{x})$ is convex when $g(\boldsymbol{x})$ is convex. Second, if $g(\boldsymbol{x})$ is convex, then $g^\lambda(\boldsymbol{x})$ is always differentiable even if $g(\boldsymbol{x})$ is not, and its gradient can be expressed in terms of $\text{prox}_g^\lambda(\boldsymbol{x})$, namely,

$$\nabla g^\lambda(\boldsymbol{x}) = \frac{1}{\lambda}\left[\boldsymbol{x} - \text{prox}_g^\lambda(\boldsymbol{x})\right]. \tag{4}$$

Moreover, $\nabla g^\lambda(\boldsymbol{x})$ is λ^{-1}-Lipschitz since proximal mappings are firmly nonexpansive. Finally, $g^\lambda(\boldsymbol{x})$ converges pointwise to $g(\boldsymbol{x})$ as λ tends to zero (Rockafellar and Wets 2009). In summary, the Moreau-Yosida envelope of a non-smooth function $g(\boldsymbol{x})$ is a Lipschitz-differentiable, arbitrarily close approximation to $g(\boldsymbol{x})$.

The closely related proximal mapping plays a prominent role in modern statistical learning since many popular non-smooth regularizations have unique proximal maps that either have explicit formulas or can be computed efficiently (Beck 2017).

In the special case when g is the indicator function $\delta_{\mathcal{E}}$ of a set $\mathcal{E}$, the proximal mapping $\text{prox}_{\delta_{\mathcal{E}}}^\lambda(\boldsymbol{x})$ takes a particularly simple form. From (1) and Definition 1, we can see that it equals the Euclidean projection operator $\mathcal{P}$ onto the set $\mathcal{E}$, that is,

$$\begin{aligned} \text{prox}_{\delta_{\mathcal{E}}}^\lambda(\boldsymbol{x}) &= \underset{\boldsymbol{\omega}\in\mathcal{E}}{\arg\min}\ \frac{1}{2\lambda}\|\boldsymbol{\omega} - \boldsymbol{x}\|^2 \\ &= \underset{\boldsymbol{\omega}\in\mathcal{E}}{\arg\min}\ \|\boldsymbol{\omega} - \boldsymbol{x}\| = \mathcal{P}_{\mathcal{E}}(\boldsymbol{x}), \quad \text{for all } \lambda > 0. \end{aligned}$$

Let $d_{\mathcal{E}}(\boldsymbol{x})$ denote the Euclidean distance from the point $\boldsymbol{x}$ to the set $\mathcal{E}$, namely,

$$d_{\mathcal{E}}(\boldsymbol{x}) = \inf_{\boldsymbol{y}\in\mathcal{E}} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

Since $P_{\mathcal{E}}(\boldsymbol{x})$ is the point in $\mathcal{E}$ that is closest in Euclidean distance to $\boldsymbol{x}$,

$$d_{\mathcal{E}}(\boldsymbol{x}) = \|\boldsymbol{x} - \mathcal{P}_{\mathcal{E}}(\boldsymbol{x})\|.$$

Using Definition 2, the Moreau-Yosida envelope $\delta_{\mathcal{E}}^\lambda(\boldsymbol{x})$ of $\delta_{\mathcal{E}}(\boldsymbol{x})$ is

$$\delta_{\mathcal{E}}^\lambda(\boldsymbol{x}) = \frac{1}{2\lambda}\|\boldsymbol{x} - \mathcal{P}_{\mathcal{E}}(\boldsymbol{x})\|^2 = \frac{1}{2\lambda}d_{\mathcal{E}}^2(\boldsymbol{x}).$$

2.2. Projections onto Epigraphs

The key algorithmic building block in our ProxMCMC framework is the projection onto the set $\mathcal{E}$. For regularized estimation problems, $\mathcal{E}$ is the epigraph of the regularization function $P(\boldsymbol{\theta})$, namely,

$$\mathcal{E} = \text{epi}(P) = \{(\boldsymbol{\theta}, \alpha) : P(\boldsymbol{\theta}) \leq \alpha\}.$$

Projection onto epigraphs is well known (Beck 2017) and is given by

$$\mathcal{P}_{\mathcal{E}}(\boldsymbol{\theta}, \alpha) = \begin{cases} (\boldsymbol{\theta}, \alpha) & P(\boldsymbol{\theta}) \leq \alpha \\ \left(\text{prox}_P^{\nu^*}(\boldsymbol{\theta}), \alpha + \nu^*\right) & P(\boldsymbol{\theta}) > \alpha \end{cases}, \tag{5}$$

where ν^* is any positive root of the auxiliary function $F(\nu) = P\left(\text{prox}_P^\nu(\boldsymbol{\theta})\right) - \nu - \alpha$, and can be found using bisection.

3. An Illustrative Case Study

This section introduces our framework using a canonical example, the lasso regression (Tibshirani 1996). We have chosen the lasso because of its simplicity and familiarity to many readers, rather than as the motivation of this article. The real power of ProxMCMC will be demonstrated on more complex models later. The lasso solves the following minimization problem,

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \rho\|\boldsymbol{\beta}\|_1, \tag{6}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is a vector of continuous responses, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients, and ρ is a nonnegative regularization strength parameter that trades off model fit with sparsity in the estimate of $\boldsymbol{\beta}$. To solve this problem in the ProxMCMC framework, we first write the regularized form (6) in an equivalent constrained form

$$\begin{aligned}&\text{minimize } \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2\\ &\text{subject to } \|\boldsymbol{\beta}\|_1 \le \alpha,\end{aligned}$$

where the constraint parameter α is in one-to-one correspondence with the regularization strength parameter ρ. For this reason, we will also call α the regularization strength parameter. A Bayesian hierarchical model is specified for the constrained formulation of lasso:

- Data likelihood: $\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2 \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$,
- A prior $\pi(\sigma^2)$ for the variance: $\sigma^2 \sim IG(r_{\sigma^2}, s_{\sigma^2})$, where $IG(r, s)$ denotes the Inverse-Gamma distribution with scale parameter r and shape parameter s (mean $= \frac{r}{s-1}$ for $s > 1$),
- A prior $\pi(\boldsymbol{\beta} \mid \alpha)$ for $\boldsymbol{\beta}$ conditional on α, namely

$$\pi(\boldsymbol{\beta} \mid \alpha) = \frac{p!}{\alpha^p 2^p}\exp\left[-\delta_{\mathcal{E}}(\boldsymbol{\beta}, \alpha)\right],$$

  where $\mathcal{E} = \{(\boldsymbol{\beta}, \alpha) : \|\boldsymbol{\beta}\|_1 \le \alpha\}$ and $\frac{p!}{\alpha^p 2^p}$ is the reciprocal of the volume of $\mathcal{E}$. Intuitively, $\pi(\boldsymbol{\beta} \mid \alpha)$ is a flat prior over an ℓ_1-ball of radius α.
- A prior $\pi(\alpha)$ for the ℓ_1-regularization strength parameter α: $\alpha \sim IG(r_\alpha, s_\alpha)$.

The distribution $\pi(\boldsymbol{\beta}, \alpha) = \pi(\boldsymbol{\beta} \mid \alpha) \cdot \pi(\alpha)$ specifies a prior on the epigraph $\mathcal{E} = \{(\boldsymbol{\beta}, \alpha) : \|\boldsymbol{\beta}\|_1 \le \alpha\} \subset \mathbb{R}^{p+1}$. The posterior log-density, up to an irrelevant additive constant, is

$$\begin{aligned}&\log \pi(\boldsymbol{\beta}, \sigma^2, \alpha)\\ &= -\left(\frac{n}{2} + s_{\sigma^2} + 1\right)\log\sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + 2r_{\sigma^2}}{2\sigma^2}\\ &\quad - (s_\alpha + p + 1)\log\alpha - \frac{r_\alpha}{\alpha} - g(\boldsymbol{\beta}, \alpha),\end{aligned}$$

where $g(\boldsymbol{\beta}, \alpha) = \delta_{\mathcal{E}}(\boldsymbol{\beta}, \alpha)$. Unfortunately, the posterior is not differentiable because it contains the non-differentiable indicator function $g(\boldsymbol{\beta}, \alpha)$. As a result, sampling algorithms for smooth log-densities cannot be directly applied.

The key idea of the proposed ProxMCMC method is simple: find a smooth approximation to the non-differentiable posterior so it can be easily sampled from. Specifically, we approximate $g(\boldsymbol{\beta}, \alpha)$ with its Moreau-Yosida envelope $g^\lambda(\boldsymbol{\beta}, \alpha)$ and substitute $g(\boldsymbol{\beta}, \alpha)$ with $g^\lambda(\boldsymbol{\beta}, \alpha)$ in the posterior. As mentioned in Section 2, $g^\lambda(\boldsymbol{\beta}, \alpha)$ approximates $g(\boldsymbol{\beta}, \alpha)$ arbitrarily well as the positive scaling constant λ tends to 0, so the smoothed posterior log-density

$$\begin{aligned}&\log \pi^\lambda(\boldsymbol{\beta}, \sigma^2, \alpha)\\ &= -\left(\frac{n}{2} + s_{\sigma^2} + 1\right)\log\sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + 2r_{\sigma^2}}{2\sigma^2}\\ &\quad - (s_\alpha + p + 1)\log\alpha - \frac{r_\alpha}{\alpha} - g^\lambda(\boldsymbol{\beta}, \alpha),\end{aligned}$$

can be made arbitrarily close to $\log \pi(\boldsymbol{\beta}, \sigma^2, \alpha)$ as λ tends to 0. Since $\log \pi^\lambda(\boldsymbol{\beta}, \sigma^2, \alpha)$ is smooth, it can be readily sampled using any sampling algorithms for smooth log-densities. Hamiltonian Monte Carlo (HMC) (Neal 2011) is used in this article due to its efficiency and generality. The last step of our algorithm is to log-transform nonnegative parameters to make their domains unconstrained, which is a requirement for HMC. The smooth posterior under the parameterization $(\boldsymbol{\beta}, \log\sigma^2, \log\alpha)$ is

$$\begin{aligned}&\log \pi^\lambda(\boldsymbol{\beta}, \log\sigma^2, \log\alpha)\\ &= -\left(\frac{n}{2} + s_{\sigma^2}\right)\log\sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + 2r_{\sigma^2}}{2\sigma^2}\\ &\quad - (s_\alpha + p)\log\alpha - \frac{r_\alpha}{\alpha} - g^\lambda(\boldsymbol{\beta}, \alpha).\end{aligned}$$

Before presenting numerical results, it is worth pausing to reflect on the power of the proposed method. Despite its simplicity, it demonstrates remarkable versatility as its extension beyond sparsity can be readily seen. To develop ProxMCMC algorithms for new regularized problems, one simply needs to find the corresponding Moreau-Yosida envelopes and proximal mappings, both of which are well-known for many non-smooth regularizations (Beck 2017). The same idea can be applied to constrained problems in a similar manner, thus, substantially broadening the range of problems that can be solved by ProxMCMC. Moreover, nothing prevents us from applying ProxMCMC to problems that encompass both regularizations and constraints. Additionally, the regularization strength parameter is seamlessly integrated into the inferential procedure in the proposed ProxMCMC method, rendering it fully Bayesian.

To see whether ProxMCMC gives reasonable results compared with existing methods such as Bayesian lasso and horseshoe prior, we apply them on the diabetes dataset used by Efron et al. (2004). The outcome is a quantitative measure of disease progression over a year, and the covariates are age, sex, body mass index, average blood pressure, and six blood serum measurements. All variables are standardized to have zero mean and unit variance. For Bayesian lasso, we use the `blasso` function from the R package `monomvn` (Gramacy 2019) with default parameters. We show the results of Bayesian lasso with and without using reversible jump MCMC (RJMCMC) to perform model selection. For the horseshoe prior, we use the R package `horseshoe` (van der Pas et al. 2019) and set function parameters `method.tau` and `method.sigma` to be "`truncatedCauchy`" and "`Jeffreys`", respectively. For ProxMCMC, we set $\lambda = 0.001$, $\sigma^2 \sim IG(0.01, 0.01)$, and $\alpha \sim IG(1, 10 + 2)$. We also calculate the 95% selective inference confidence intervals (Lee et al. 2016) using the R package `selectiveInference` (Tibshirani et al. 2019). Since selective inference requires a model to be selected first,

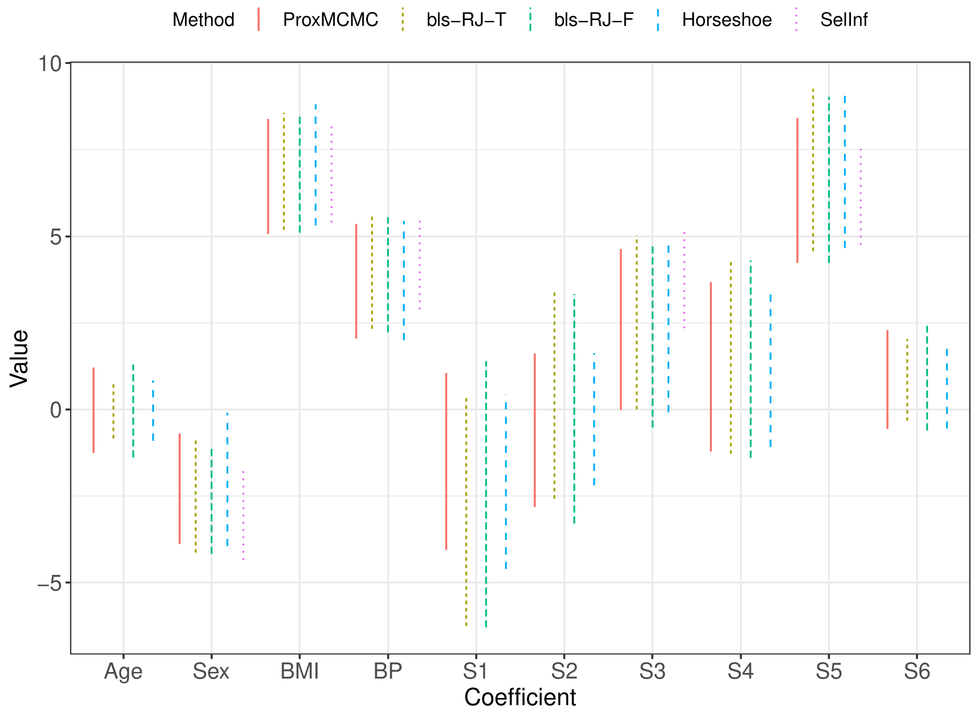

Figure 2. 95% credible intervals calculated by ProxMCMC, Bayesian lasso (bls), and horseshoe prior for the diabetes dataset (Efron et al. 2004). Also shown are the 95% selective inference (SelInf) confidence intervals for the five variables selected by lasso using 10-fold cross-validation. bls-RJ-T refers to Bayesian lasso with reversible jump MCMC (RJMCMC). bls-RJ-F indicates that RJMCMC is not used.

we use lasso with 10-fold cross-validation and choose the largest regularization parameter such that the error is within 1 standard error of the minimum (the `lambda.1se` option from the `glmnet` package). Figure 2 shows the 95% interval estimates of the regression coefficients computed by each method. We see that for null covariates, the credible intervals of Bayesian lasso are narrower when model selection by RJMCMC is used. This is because RJMCMC results in many exact zeros (75% in this example) in the posterior sample, which reduces the width of credible intervals. When RJMCMC is not used, the credible intervals of the null covariates become wider and are similar to those obtained by ProxMCMC. The credible intervals from the horseshoe prior are narrower for null covariates, but for non-null covariates, the widths of the intervals are similar regardless of which method is used. The selective inference confidence intervals are calculated conditional on a selected model, and their coverage guarantee is in the frequentist sense, so they are not directly comparable with credible intervals. Nevertheless, we included them in the plot as a reference. To make sure HMC converged well for ProxMCMC, we also checked the traceplots for α and σ^2, both of which show good mixing. See supplementary materials for details.

4. Methodology

Having seen how to apply ProxMCMC to the special case of lasso, we next present the framework in greater generality. Our proposed ProxMCMC method consists of three steps.

1. Likelihood and prior. The first step is to specify a likelihood model for the data and priors for model parameters, which is a standard step in Bayesian modeling. Let $\boldsymbol{\tau} \in \mathbb{R}^p$ denote parameters that are subject to regularizations or constraints, $\boldsymbol{\eta} \in \mathbb{R}^q$ denote all other parameters including the regularization strength parameter α, and $\boldsymbol{\theta} = (\boldsymbol{\tau}^T, \boldsymbol{\eta}^T)^T \in \mathbb{R}^d$ ($d = p + q$) denote all model parameters. Further let $\ell(\boldsymbol{\theta})$ be the log-likelihood and $\pi(\boldsymbol{\eta})$ be the prior density for $\boldsymbol{\eta}$. The prior for $\boldsymbol{\tau}$ depends on whether the problem involves regularization, constraints, or both.

For regularized problems, the prior for $\boldsymbol{\tau}$, conditional on the regularization strength parameter α, is

$$\pi(\boldsymbol{\tau} \mid \alpha) = c \cdot \exp[-\delta_{\mathcal{E}}(\boldsymbol{\tau}, \alpha)],$$

where c is a normalizing constant, and $\mathcal{E}$ is the epigraph of the regularization (penalty) function $P(\boldsymbol{\tau})$, that is, $\mathcal{E} = \operatorname{epi}(P) = \{(\boldsymbol{\tau}, \alpha) : P(\boldsymbol{\tau}) \leq \alpha\}$. For this reason we refer to $\pi(\boldsymbol{\tau} \mid \alpha)$ as the *epigraph prior*. Since the regularization strength parameter α must be nonnegative, it requires a prior with nonnegative support. We find that placing an inverse Gamma prior on α works well in practice.

To provide intuition on how the epigraph prior differs from existing alternatives, consider the simple case where a scalar parameter β is regularized with the ℓ_1-norm. The epigraph is $\mathcal{E} = \{(\beta, \alpha) : |\beta| \leq \alpha\}$. With an $IG(r, s)$ prior on α, the marginal density for β is

$$f_\beta(t) = \int_{|t|}^{\infty} \frac{1}{2\alpha} \pi(\alpha) d\alpha = \frac{s}{2r}\left[1 - F_{IG(r,s+1)}(|t|)\right],$$

where $F_{IG(r,s+1)}(|t|)$ is the cumulative distribution function of $IG(r, s+1)$ evaluated at $|t|$. By comparing the ProxMCMC epigraph prior with Laplacian prior and horseshoe prior, we can see from Figure 3 that it shrinks small β while allowing strong signals to remain large. We would like to reiterate that the main motivation behind ProxMCMC is not to introduce yet another sparsity-inducing prior but rather to address problems

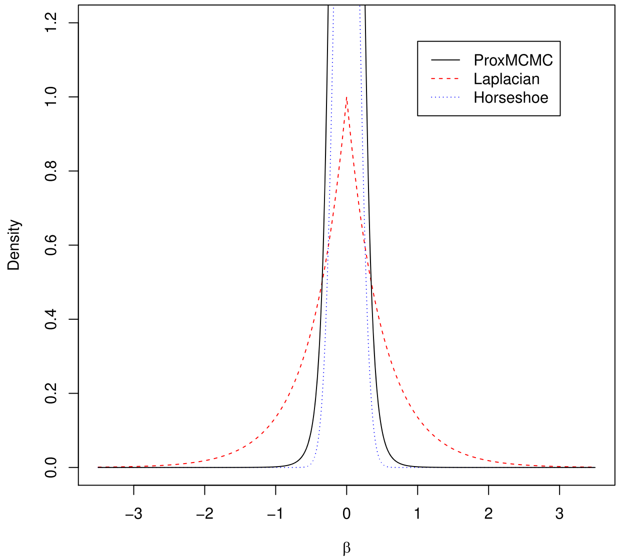

Figure 3. The density of ProxMCMC epigraph prior, Laplacian prior, and horseshoe prior.

that encompass constraints and more complex regularizations. The ℓ_1-norm example is intended to offer intuition.

In the multivariate setting where more than one parameter is regularized, the ProxMCMC epigraph prior enforces negative correlation among components of $\boldsymbol{\beta}$. For example, it is clear from the lasso example, where the epigraph is given by $\mathcal{E} = \{(\boldsymbol{\beta}, \alpha) : \|\boldsymbol{\beta}\|_1 \leq \alpha\}$, that given α, some components of $\boldsymbol{\beta}$ are forced to decrease as others take larger values. This repulsive feature distinguishes the ProxMCMC epigraph prior from other Bayesian priors such as the Laplacian or horseshoe prior, where components of $\boldsymbol{\beta}$ are independent of each other conditional on the hyperparameter, and are marginally positively correlated.

For constrained problems, the set $\mathcal{E}$ refers to the constraint set instead of the epigraph, and the following prior for $\boldsymbol{\tau}$ is used:

$$\pi(\boldsymbol{\tau}) = c \cdot \exp[-\delta_{\mathcal{E}}(\boldsymbol{\tau})],$$

where c is, again, a normalizing constant. For problems that encompass both regularization and constraints, two prior distributions are needed for $\boldsymbol{\tau}$: one to enforce the regularization and the other to enforce the constraints. For simplicity of presentation, we will not distinguish between regularized problems and constrained problems, except in cases where distinction is necessary. We also abuse the notation slightly by using $g(\boldsymbol{\tau})$ to denote either $\delta_{\mathcal{E}}(\boldsymbol{\tau}, \alpha)$ or $\delta_{\mathcal{E}}(\boldsymbol{\tau})$, depending on the problem.

Given the likelihood model and prior distributions, we have the posterior density

$$\pi(\boldsymbol{\theta} \mid Y) = \frac{e^{-U(\boldsymbol{\theta})}}{\int e^{-U(\boldsymbol{s})}\, d\boldsymbol{s}},$$

where $U(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + g(\boldsymbol{\tau})$ and $f(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) - \log \pi(\boldsymbol{\eta})$. The posterior $\pi(\boldsymbol{\theta} \mid Y)$ is not differentiable because $g(\boldsymbol{\tau})$ is not. By substituting $g(\boldsymbol{\tau})$ with its Moreau-Yosida envelope $g^{\lambda}(\boldsymbol{\tau})$, both $U^{\lambda}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}) + g^{\lambda}(\boldsymbol{\tau})$ and

$$\pi^{\lambda}(\boldsymbol{\theta} \mid Y) = \frac{e^{-U^{\lambda}(\boldsymbol{\theta})}}{\int e^{-U^{\lambda}(\boldsymbol{s})}\, d\boldsymbol{s}}$$

become smooth functions.

2. Gradient. The next step is to efficiently evaluate the gradient of the smoothed posterior log-density, which is another standard step in Bayesian modeling. For commonly used likelihood models and priors, the gradient can be computed numerically by auto-differentiation in software packages such as `Stan` (Stan Development Team 2020) and `Turing.jl` (Ge, Xu, and Ghahramani 2018).

As noted earlier, the existence of the gradient of the Moreau-Yosida envelope $g^{\lambda}(\boldsymbol{\tau})$ depends on the convexity of the indicator function $g(\boldsymbol{\tau})$, and thus on the convexity of the epigraph or the constraint set $\mathcal{E}$. When $g(\boldsymbol{\tau})$ is convex, which is the case for many commonly used regularization and constraints, proximal mappings have been extensively studied in the optimization literature (Beck 2017), and efficient implementations are available from mature libraries such as the FOM Matlab toolbox (Beck and Guttmann-Beck 2019), the Python package `PyProximal`, and the Julia package `ProximalOperators.jl`.

When $g(\boldsymbol{\tau})$ is non-convex, $g^{\lambda}(\boldsymbol{\tau})$ is no longer differentiable. Under certain regularity conditions, however, $g^{\lambda}(\boldsymbol{\tau})$ is semidifferentiable and we can calculate a subgradient and use it in place of gradient in sampling algorithms. This approach will be demonstrated on the sparse low rank matrix regression example in Section 5.4.

3. Sampling algorithm. Finally, we invoke a gradient based sampling algorithm such as HMC or the Langevin algorithm to efficiently explore the posterior landscape. Software implementations include `DynamicHMC.jl`, `AdvancedHMC.jl`, and `pyhmc`, to name a few.

Remark. Before proceeding to examples, we pause to highlight ProxMCMC's close connection to distance majorization and proximal distance algorithms (Chi, Zhou, and Lange 2014; Xu, Chi, and Lange 2017; Keys, Zhou, and Lange 2019; Landeros and Lange 2021; Landeros, Wu, and Lange 2022; Landeros et al. 2022). Proximal distance algorithms are used to solve distance penalty problems of the form

$$\text{minimize} \quad f(\boldsymbol{\theta}) + \frac{\rho}{2} d_{\mathcal{E}}(\boldsymbol{\theta})^2, \tag{7}$$

where $f(\boldsymbol{\theta})$ is typically a negative log-likelihood term quantifying model fit, $\mathcal{E}$ is a target constraint set that we wish our estimate of $\boldsymbol{\theta}$ to be close to, and ρ is a nonnegative tuning parameter that trades off model fit with the amount of constraint violation quantified as the distance to $\mathcal{E}$. A solution to (7) is a maximum a posteriori estimate under a distance-to-set prior $\pi(\boldsymbol{\theta}) \propto \exp(-\frac{\rho}{2} d_{\mathcal{E}}(\boldsymbol{\theta})^2)$. Thus, the ProxMCMC method proposed here provides a fully Bayesian framework for generating posterior samples under a distance-to-epigraph set prior. Concurrent work in (Presman and Xu 2022) uses distance-to-set priors to solve constrained Bayesian inference problems and discusses its advantages over prior literature on Bayesian constraint relaxation.

5. Examples

The power of the proposed ProxMCMC method is illustrated on four examples, whose inference is either unknown or regarded as difficult. Since the potential applications of ProxMCMC are innumerable, our examples are not comprehensive. Nevertheless, we hope they serve as a starting point for readers to derive

ProxMCMC algorithms for their own problems. See Heng, Zhou, and Chi (2023) for an application of ProxMCMC to the Bayesian trend filtering problem.

5.1. Constrained Lasso

Constrained lasso is a commonly used technique for analyzing compositional data and has been applied to problems such as consumer spending in economics, topic extraction of documents, and human microbiome analysis (Gaines, Kim, and Zhou 2018; James, Paulson, and Rusmevichientong 2020). The problem is formulated as

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2+\rho\|\boldsymbol{\beta}\|_1 \\ &\text{subject to} \quad \boldsymbol{A\beta}=\boldsymbol{b}, \end{aligned}$$

where $\boldsymbol{y}\in\mathbb{R}^n$ is a vector of continuous responses, $\boldsymbol{X}\in\mathbb{R}^{n\times p}$ is a design matrix, $\boldsymbol{\beta}\in\mathbb{R}^p$ is the vector of regression coefficients, $\boldsymbol{A}$ and $\boldsymbol{b}$ impose constraints $\boldsymbol{\beta}$, and $\boldsymbol{A}$ has full row-rank. In compositional data analysis, for example, where each row of the design matrix $\boldsymbol{X}$ represents proportions of a whole and sums to 1, we can make $\boldsymbol{\beta}$ identifiable by constraining $\sum_i \boldsymbol{\beta}_i=0$, which corresponds to $\boldsymbol{A}=\mathbf{1}_p^T$ (a row of 1s) and $\boldsymbol{b}=0$.

As in the lasso example, we use a normal likelihood model ($\boldsymbol{Y}\mid\boldsymbol{\beta},\sigma^2\sim N(\boldsymbol{X\beta},\sigma^2\boldsymbol{I})$) and inverse Gamma priors for σ^2 and α ($\sigma^2\sim IG(r_{\sigma^2},s_{\sigma^2})$, $\alpha\sim IG(r_\alpha,s_\alpha)$). Let $\mathcal{E}_1=\{(\boldsymbol{\beta},\alpha):\|\boldsymbol{\beta}\|_1\le\alpha\}$ denote the epigraph of the ℓ_1-norm and let $\mathcal{E}_2=\{\boldsymbol{\beta}:\boldsymbol{A\beta}=\boldsymbol{b}\}$ denote the constraint set. With the $(\boldsymbol{\beta},\log\sigma^2,\log\alpha)$ parameterization, the smoothed posterior log-density up to an irrelevant additive constant is

$$\begin{aligned} &\log\pi^\lambda(\boldsymbol{\beta},\log\sigma^2,\log\alpha) \\ &=-\left(\frac{n}{2}+s_{\sigma^2}\right)\log\sigma^2-\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2+2r_{\sigma^2}}{2\sigma^2} \\ &\quad -s_\alpha\log\alpha-\frac{r_\alpha}{\alpha}-g_1^\lambda(\boldsymbol{\beta},\alpha)-g_2^\lambda(\boldsymbol{\beta}), \end{aligned}$$

where $g_1^\lambda(\boldsymbol{\beta},\alpha)$ and $g_2^\lambda(\boldsymbol{\beta})$ are the Moreau-Yosida envelopes of the indicator functions $g_1(\boldsymbol{\beta},\alpha)=\delta_{\mathcal{E}_1}(\boldsymbol{\beta},\alpha)$ and $g_2(\boldsymbol{\beta})=\delta_{\mathcal{E}_2}(\boldsymbol{\beta})$, respectively. From (5), the proximal mapping of $g_1(\boldsymbol{\beta},\alpha)$ is the projection onto the epigraph $\mathcal{E}_1$

$$\text{prox}_{g_1}^\lambda(\boldsymbol{\beta},\alpha)=\begin{cases}(\boldsymbol{\beta},\alpha) & \text{if } \|\boldsymbol{\beta}\|_1\le\alpha \\ (S_{\nu^*}(\boldsymbol{\beta}),\alpha+\nu^*) & \text{if } \|\boldsymbol{\beta}\|_1>\alpha\end{cases},$$

where S_{ν^*} is the soft-thresholding operator, the univariate form of which is given in (3), and ν^* is any positive root of the nonincreasing function $\phi(\nu)=\|S_\nu(\boldsymbol{\beta})\|_1-\nu-\alpha$ (Beck 2017). The proximal mapping of g_2 is the projection onto the hyperplane given by

$$\text{prox}_{g_2}(\boldsymbol{\beta})=\boldsymbol{\beta}-\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}(\boldsymbol{A\beta}-\boldsymbol{b}).$$

The gradient of the posterior log-density is given block-wise by

$$\begin{aligned} \frac{\partial\log\pi^\lambda}{\partial\boldsymbol{\beta}} &=\sigma^{-2}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X\beta})-\lambda^{-1}\left[\boldsymbol{\beta}-\text{prox}_{g_1}^\lambda(\boldsymbol{\beta},\alpha)_{\boldsymbol{\beta}}\right] \\ &\quad -\lambda^{-1}\left[\boldsymbol{\beta}-\text{prox}_{g_2}^\lambda(\boldsymbol{\beta})\right] \\ \frac{\partial\log\pi^\lambda}{\partial\log\sigma^2} &=-\left(\frac{n}{2}+s_{\sigma^2}\right)+\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2+2r_{\sigma^2}}{2\sigma^2} \\ \frac{\partial\log\pi^\lambda}{\partial\log\alpha} &=-s_\alpha+\frac{r_\alpha}{\alpha}-\lambda^{-1}\alpha[\alpha-\text{prox}_g^\lambda(\boldsymbol{\beta},\alpha)_\alpha]. \end{aligned}$$

Numerical results will be presented in Section 6.

5.2. Graphical Lasso

Given iid p-dimensional observations $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$, where $\boldsymbol{x}_i\sim N(\mathbf{0},\Sigma)$ and Σ is a $p\times p$ covariance matrix, graphical lasso infers the underlying conditional dependency among covariates by estimating the precision matrix $\Theta=\Sigma^{-1}$ through maximizing the regularized log-likelihood

$$-\frac{n}{2}\text{tr}(\boldsymbol{S}\Theta)+\frac{n}{2}\text{logdet}(\Theta)-\rho\sum_{j\neq k}|\Theta_{jk}|,$$

where $\boldsymbol{S}$ is the sample covariance and ρ is the regularization strength parameter. Equivalently, we can maximize

$$-\frac{n}{2}\text{tr}(\boldsymbol{S}\Theta)+\frac{n}{2}\text{logdet}(\Theta)-g(\Theta,\alpha),$$

where $g(\Theta,\alpha)=\delta_{\mathcal{E}}(\Theta,\alpha)$ and $\mathcal{E}=\{(\Theta,\alpha):\sum_{j\neq k}|\Theta_{jk}|\le\alpha\}$. The function $g(\Theta,\alpha)$ can be seen as the log-density (up to an additive constant) of the uniform prior for Θ over the ℓ_1-ball $\{\Theta:\sum_{j\neq k}|\Theta_{jk}|\le\alpha\}$. With an $IG(r_\alpha,s_\alpha)$ prior for α, and after smoothing $g(\Theta,\alpha)$ with its Moreau-Yosida envelope $g^\lambda(\Theta,\alpha)$, the smoothed posterior log-density of $(\Theta,\log\alpha)$ is

$$\begin{aligned} \log\pi^\lambda(\Theta,\log\alpha)=&-\frac{n}{2}\text{tr}(\boldsymbol{S}\Theta)+\frac{n}{2}\text{logdet}(\Theta) \\ &-s_\alpha\log\alpha-\frac{r_\alpha}{\alpha}-g^\lambda(\Theta,\alpha). \end{aligned}$$

Since HMC works on unconstrained domains, but Θ needs to be positive definite, we parameterize Θ in terms of its lower Cholesky factor $\boldsymbol{L}$. Adjusting for the log-Jacobian terms, the smoothed posterior log-density becomes

$$\begin{aligned} \log\pi^\lambda(\boldsymbol{L},\log\alpha)=&-\frac{n}{2}\text{tr}(\boldsymbol{SLL}^T)+\frac{n}{2}\text{logdet}(\boldsymbol{LL}^T) \\ &-s_\alpha\log\alpha-\frac{r_\alpha}{\alpha}-g^\lambda(\boldsymbol{LL}^T,\alpha) \\ &+p\log(2)+\sum_{j=1}^p(p-j+2)\boldsymbol{L}_{jj}. \end{aligned}$$

The gradients are

$$\begin{aligned} \nabla_{\text{vech}\boldsymbol{L}}\log\pi^\lambda=&-\big(n\text{vech}(\boldsymbol{SL})\big)^T+n\big(\text{vech}(\boldsymbol{L}^{-1})^T\big)^T \\ &-\frac{2}{\lambda}\Big(\text{vech}\Big([\Theta-\text{prox}_g^\lambda(\Theta,\alpha)_\Theta]\boldsymbol{L}\Big)\Big)^T \\ &+\Big(\text{vech}\Big(\text{diag}(p+1,p,\ldots,2)\Big)\Big)^T \\ \frac{\partial\log\pi^\lambda}{\partial\log\alpha}=&-s_\alpha+\frac{r_\alpha}{\alpha}-\lambda^{-1}\alpha[\alpha-\text{prox}_g^\lambda(\Theta,\alpha)_\alpha], \end{aligned}$$

where $\text{vech}(\boldsymbol{L})$ denotes the vector obtained from stacking the columns of the lower triangular part of the square matrix $\boldsymbol{L}$.

5.3. Matrix Completion

Given a matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ with entries only observed on the index set $\Omega = \{(i,j) : y_{ij} \text{ is observed}\}$, Mazumder, Hastie, and Tibshirani (2010) proposed to complete the matrix by minimizing the convex objective function

$$\frac{1}{2}\|P_\Omega(\boldsymbol{Y}-\boldsymbol{X})\|_{\mathrm{F}}^2 + \alpha\|\boldsymbol{X}\|_*,$$

where $\boldsymbol{X}$ is the completed matrix, $P_\Omega(\boldsymbol{Y}-\boldsymbol{X})$ is the projection of $\boldsymbol{Y}-\boldsymbol{X}$ onto the set of observed entries Ω, namely, the (i,j)th entry of $P_\Omega(\boldsymbol{Y}-\boldsymbol{X})_{ij}$ is $y_{ij}-x_{ij}$ for $(i,j)\in\Omega$ and zero otherwise, α is the regularization strength parameter, and $\|\boldsymbol{X}\|_*$ is the nuclear norm of $\boldsymbol{X}$. The nuclear norm is defined as $\|\boldsymbol{X}\|_* = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_1 = \sum_i \sigma_i(\boldsymbol{X})$, where $\sigma_1(\boldsymbol{X}) \geq \cdots \geq \sigma_m(\boldsymbol{X}) \geq 0$ are the singular values of $\boldsymbol{X}$. To solve the matrix completion problem using ProxMCMC, we use the likelihood model $\mathrm{vec}(\boldsymbol{Y}) \sim N(\mathrm{vec}(\boldsymbol{X}), \sigma^2\boldsymbol{I})$, assume priors $\sigma^2 \sim IG(r_{\sigma^2}, s_{\sigma^2})$ and $\alpha \sim IG(r_\alpha, s_\alpha)$, let $\mathcal{E} = \{(\boldsymbol{X},\alpha) : \|\boldsymbol{X}\|_* \leq \alpha\}$ be the epigraph of $\|\cdot\|_*$, and let $g(\boldsymbol{X},\alpha) = \delta_{\mathcal{E}}(\boldsymbol{X},\alpha)$ be the corresponding indicator function. The smoothed posterior log-density using the $(\boldsymbol{X}, \log\sigma^2, \log\alpha)$ parameterization is

$$\begin{aligned}
&\log\pi^\lambda(\boldsymbol{X},\log\sigma^2,\log\alpha)\\
&= -\left(\frac{|\Omega|}{2} + s_{\sigma^2}\right)\log\sigma^2 - \frac{\sum_{(i,j)\in\Omega}(y_{ij}-x_{ij})^2 + 2r_{\sigma^2}}{2\sigma^2}\\
&\quad - s_\alpha\log\alpha - \frac{r_\alpha}{\alpha} - g^\lambda(\boldsymbol{X},\alpha),
\end{aligned}$$

Let $\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{X}^T$ be the singular value decomposition of $\boldsymbol{X}$, then the proximal mapping of $g(\boldsymbol{X},\alpha)$ is the projection given by

$$\begin{aligned}
&\mathrm{prox}_g^\lambda(\boldsymbol{X},\alpha)\\
&= \begin{cases} (\boldsymbol{X},\alpha) & \text{if } \|\boldsymbol{X}\|_* \leq \alpha \\ (\boldsymbol{U}\mathrm{diag}(S_{\nu^*}(\boldsymbol{\sigma}(\boldsymbol{X})))\boldsymbol{V}^T, \alpha+\nu^*) & \text{if } \|\boldsymbol{X}\|_* > \alpha \end{cases},
\end{aligned}$$

where ν^* is any positive root of the nonincreasing function $\phi(\nu) = \|S_\nu(\boldsymbol{\sigma}(\boldsymbol{X}))\|_1 - \nu - \alpha$. The gradient of the smoothed posterior log-density is

$$\begin{aligned}
\frac{\partial\log\pi^\lambda}{\partial\boldsymbol{X}} &= \sigma^{-2}\left[P_\Omega(\boldsymbol{Y}-\boldsymbol{X})\right] - \lambda^{-1}[\boldsymbol{X} - \mathrm{prox}_g^\lambda(\boldsymbol{X},\alpha)_{\boldsymbol{X}}],\\
\frac{\partial\log\pi^\lambda}{\partial\log\sigma^2} &= -\left(\frac{|\Omega|}{2} + s_{\sigma^2}\right) + \frac{\sum_{(i,j)\in\Omega}(y_{ij}-x_{ij})^2 + 2r_{\sigma^2}}{2\sigma^2},\\
\frac{\partial\log\pi^\lambda}{\partial\log\alpha} &= -s_\alpha + \frac{r_\alpha}{\alpha} - \lambda^{-1}\alpha\left[\alpha - \mathrm{prox}_g^\lambda(\boldsymbol{X},\alpha)_\alpha\right].
\end{aligned}$$

5.4. Sparse Low Rank Matrix Regression

We consider linear regression with matrix covariates, where the rank of the coefficient matrix is subject to regularization. One approach is to regularize the nuclear norm of the coefficient matrix (Zhou and Li 2014), for which the ProxMCMC algorithm is very similar to the matrix completion example above because they share the same proximal mapping. Alternatively, one can constrain the coefficient matrix to have a user-specified rank k (Zhou, Li, and Zhu 2013). Here we explore the second approach to illustrate the potential of ProxMCMC for problems where the regularization or constraints are not convex.

Let y_i be the response of the ith sample. Further let $\boldsymbol{Z}_i \in \mathbb{R}^p$ and $\boldsymbol{X}_i \in \mathbb{R}^{q\times r}$ be the corresponding vector and matrix covariates, respectively. The model is

$$y_i = \boldsymbol{Z}_i^T\boldsymbol{\gamma} + \langle\boldsymbol{B},\boldsymbol{X}_i\rangle + \epsilon_i,$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{B}$ are the vector and matrix coefficients, $\langle\boldsymbol{B},\boldsymbol{X}_i\rangle = \mathrm{tr}(\boldsymbol{B}^T\boldsymbol{X}_i) = \langle\mathrm{vec}\boldsymbol{B},\mathrm{vec}\boldsymbol{X}_i\rangle$ is the inner product of the two matrices, and $\epsilon_i \sim N(0,\sigma^2)$. We fix $\mathrm{rank}(\boldsymbol{B})$ at a user-specified value k; the corresponding constraint set and indicator functions are $\mathcal{E}_1 = \{\boldsymbol{B} : \mathrm{rank}(\boldsymbol{B}) = k\}$ and $\delta_{\mathcal{E}_1}(\boldsymbol{B})$. To promote sparsity in $\boldsymbol{B}$, we also incorporate an ℓ_1-regularization on the entries of $\boldsymbol{B}$; the epigraph set and indicator functions are $\mathcal{E}_2 = \{(\boldsymbol{B},\alpha) : \|\mathrm{vec}\boldsymbol{B}\|_1 \leq \alpha\}$ and $\delta_{\mathcal{E}_2}(\boldsymbol{B},\alpha)$. With a flat prior for $\boldsymbol{\gamma}$ ($\pi(\boldsymbol{\gamma}) \propto 1$), an $IG(r_{\sigma^2}, s_{\sigma^2})$ prior for σ^2, and an $IG(r_\alpha, s_\alpha)$ prior for α, the smoothed posterior log-density is

$$\begin{aligned}
&\log\pi(\gamma,\boldsymbol{B},\log\sigma^2,\log\alpha)\\
&= -\frac{\sum_{i=1}^n(y_i - \boldsymbol{Z}_i^T\gamma - \langle\boldsymbol{B},\boldsymbol{X}_i\rangle)^2 + 2r_{\sigma^2}}{2\sigma^2}\\
&\quad - (\frac{n}{2} + s_{\sigma^2})\log\sigma^2 - s_\alpha\log\alpha - \frac{r_\alpha}{\alpha}\\
&\quad - g_1^\lambda(\boldsymbol{B}) - g_2^\lambda(\boldsymbol{B},\alpha),
\end{aligned}$$

where $g_1^\lambda(\boldsymbol{B})$ and $g_2^\lambda(\boldsymbol{B},\alpha)$ are the Moreau-Yosida envelopes of $g_1(\boldsymbol{B}) = \delta_{\mathcal{E}_1}(\boldsymbol{B})$ and $g_2(\boldsymbol{B},\alpha) = \delta_{\mathcal{E}_2}(\boldsymbol{B},\alpha)$, respectively. The proximal mapping of $g_1(\boldsymbol{B})$, given by the projection onto the set $\mathcal{E}_1$, can still be obtained relatively easily through thresholding the singular values of $\boldsymbol{B}$. The gradient formula (4) for the Moreau-Yosida envelope, however, no longer holds because $g_1^\lambda(\boldsymbol{B})$ is not convex. The solution we explore below resorts to the subsmoothness property of Moreau-Yosida envelopes, for which we need the following definitions (Rockafellar and Wets 2009).

Definition 3 (Prox-boundedness). A function $g : \mathbb{R}^n \to \bar{\mathbb{R}}$ is prox-bounded if there exists $\lambda > 0$ such that its Moreau-Yosida envelope $g^\lambda > -\infty$ for some $\boldsymbol{x} \in \mathbb{R}^n$. The supremum of the set of all such λ is the threshold λ_g of prox-boundedness for g.

In the ProxMCMC framework, we only need the Moreau-Yosida envelope of indicator functions, for which we have $g^\lambda(\boldsymbol{x}) > -\infty$ for any $\lambda > 0$, so they are always prox-bounded and the threshold $\lambda_g = \infty$.

Definition 4 (Semidifferentiability). Let $g : \mathbb{R}^n \to \bar{\mathbb{R}}$ and $\bar{\boldsymbol{x}}$ be a point such that $g(\bar{\boldsymbol{x}})$ is finite. If the (possibly infinite) limit

$$\lim_{\tau\downarrow 0, \boldsymbol{w}'\to\boldsymbol{w}} \frac{g(\bar{\boldsymbol{x}} + \tau\boldsymbol{w}') - g(\bar{\boldsymbol{x}})}{\tau}$$

exists, it is the semiderivative of g at $\bar{\boldsymbol{x}}$ for $\boldsymbol{w}$, and g is semidifferentiable at $\bar{\boldsymbol{x}}$ for $\boldsymbol{w}$. If this holds for every $\boldsymbol{w}$, g is semidifferentiable at $\bar{\boldsymbol{x}}$.

By Rockafellar and Wets (2009, Example 10.32), if $g(\boldsymbol{x})$ is lower-semicontinuous, proper, and prox-bounded with threshold λ_g, then for $\lambda \in (0,\lambda_g)$, the Moreau-Yosida envelope $g^\lambda(\boldsymbol{x})$ is semidifferentiable and the subgradient set is

$$\partial g^\lambda(\boldsymbol{x}) \subset \lambda^{-1}\left[\boldsymbol{x} - \mathrm{prox}_g^\lambda(\boldsymbol{x})\right].$$

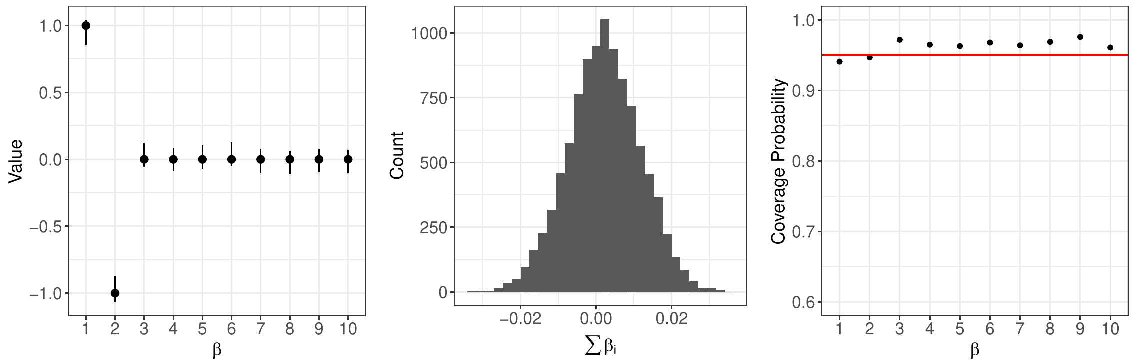

Figure 4. Left: 95% credible intervals for constrained lasso models parameters from one simulated dataset. Dots mark the truth. Middle: Histogram of $\sum_i \beta_i$ over 10,000 samples for the same dataset used on the left. Right: Coverage probability for model coefficients calculated from 1000 simulated datasets. The red line is the nominal level of 95%.

The function $g_1(\boldsymbol{B}) = \delta_{\mathcal{E}_1}(\boldsymbol{B})$ satisfies the above conditions, so we can calculate its subgradient using the above formula and use it in place of the gradient in HMC.

$$
\begin{aligned}
\frac{\partial \log \pi}{\partial \boldsymbol{\gamma}} &= \sigma^{-2} \sum_i (y_i - \boldsymbol{Z}_i^T \boldsymbol{\gamma} - \langle \boldsymbol{B}, \boldsymbol{X}_i \rangle) Z_i, \\
\frac{\partial \log \pi}{\partial \boldsymbol{B}} &= \sigma^{-2} \sum_i (y_i - \boldsymbol{Z}_i^T \boldsymbol{\gamma} - \langle \boldsymbol{B}, \boldsymbol{X}_i \rangle) X_i \\
&\quad - \lambda^{-1} \left[\boldsymbol{B} - \operatorname{prox}_{g_1}^{\lambda}(\boldsymbol{B}) \right] \\
&\quad - \lambda^{-1} \left[\boldsymbol{B} - \operatorname{prox}_{g_2}^{\lambda}(\boldsymbol{B}, \alpha)_{\boldsymbol{B}} \right], \\
\frac{\partial \log \pi}{\partial \log \sigma^2} &= -\left(\frac{n}{2} + s_{\sigma^2} \right) \\
&\quad + \frac{\sum_{i=1}^n (y_i - \boldsymbol{Z}_i^T \boldsymbol{\gamma} - \langle \boldsymbol{B}, \boldsymbol{X}_i \rangle)^2 + 2 r_{\sigma^2}}{2\sigma^2}, \\
\frac{\partial \log \pi}{\partial \log \alpha} &= -s_\alpha + \frac{r_\alpha}{\alpha} - \lambda^{-1} \alpha \left[\alpha - \operatorname{prox}_{g_2}^{\lambda}(\boldsymbol{B}, \alpha)_\alpha \right].
\end{aligned}
$$

Since $g_1^\lambda(\boldsymbol{B})$ is non-convex, $\operatorname{prox}_{g_1}^\lambda(\boldsymbol{B})$ is not unique. Our approach is to pick an arbitrary element in the proximal map set, which works well in practice.

6. Numerical Results

This section demonstrates the proposed ProxMCMC method through either simulation experiments or analysis of publicly available datasets.

6.1. Constrained Lasso: Simulated Microbiome Data

We illustrate the ProxMCMC method for constrained lasso using a simulated microbiome dataset. The 16S microbiome sequencing technology measures the number of various organisms called operational taxonomic units (OTUs) in a biological sample. For statistical analysis, counts are normalized into proportions for each sample, resulting in a design matrix $\boldsymbol{X}$ where each row sums to 1, which makes it necessary to constrain regression parameters so that they are identifiable. We use the popular sum-to-zero constraint ($\sum_j \beta_j = 0$) in this example. We set sample size $n = 1000$ and number of OTUs $p = 10$. The design matrix $\boldsymbol{X}$ is generated as follows. First, each entry in $\boldsymbol{X}$ is sampled iid from a uniform distribution ($U_{[0,1]}$). Second, the rows of $\boldsymbol{X}$ are scaled so that each row sums to 1. We set $\beta_1 = 1$, $\beta_2 = -1$ and the remaining β_j to 0 so that 20% of the entries in $\boldsymbol{\beta}$ are nonzero. The noise is generated from a normal distribution with mean 0 and $\sigma = 0.1$ so that the sample signal-to-noise ratio $\operatorname{var}(\boldsymbol{X\beta})/\sigma^2$ is approximately 0.7. We use $IG(0.01, 0.01)$ as a prior for σ^2 and $IG(1, p+1)$ as a prior for α, set $\lambda = 10^{-5}$, and ran HMC for 10,000 iterations. The experiment is repeated 1000 times to estimate the coverage probability. Figure 4 (left) shows the 95% credible intervals and the true values (black dots) for the regression parameters for the first simulated dataset. We can see that credible intervals provide good coverage of the truth. Figure 4 (middle) shows the histogram of $\sum_j \beta_j$ for posterior samples from the first simulated dataset. The histogram is highly concentrated around 0, which shows that the posterior samples satisfy the sum-to-zero constraint well. To measure the sampling efficiency of our algorithms, we calculate the effective sample size of the slowest moving component of the multivariate posterior samples. The slowest moving component can be obtained by first performing a principal components analysis on the posterior covariance matrix and then projecting the posterior samples onto the most prominent eigenvector (Durmus, Moulines, and Pereyra 2018). After obtaining the slowest moving component, which is a vector of the same length as the number of posterior samples, we can calculate its effective sample size with the `ess_rhat` function from the `MCMCDiagnosticTools.jl` package. Using this method, the effective sample size of the slowest $\boldsymbol{\beta}$ component is 7044. Finally, Figure 4 (right) shows the coverage probability of model parameters. Results indicate that the coverage probability of ProxMCMC credible intervals are very close to the nominal level of 95%.

6.2. Graphical Lasso: Cytometry Data

We compare ProxMCMC with Bayesian graphical lasso (Wang 2012) on the cell-signalling data from Sachs et al. (2005), which was used in the original graphical lasso paper (Friedman, Hastie, and Tibshirani 2008). The dataset contains flow cytometry measurements on $p = 11$ proteins and $n = 7466$ cells. We first use the R package `CVglasso` to compute 5-fold cross-validated graphical lasso estimates for Θ, which are used as references for the comparison between ProxMCMC and Bayesian graph-

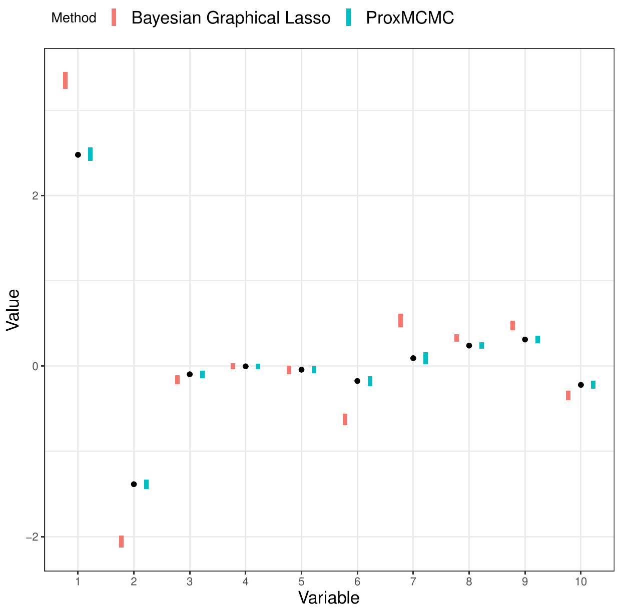

Figure 5. Comparing the 95% credible intervals of Bayesian graphical lasso versus ProxMCMC on the cytometry data. Black dots are estimates obtained from 5-fold cross-validated graphical lasso.

ical lasso. For Bayesian graphical lasso, we use the R package `BayesianGLasso` (Wang 2012). We experimented with both the default prior and other prior settings but found little difference, so we report the results using the default prior (Gamma distribution with shape parameter 1 and scale parameter 0.1). For ProxMCMC, we use an $IG(1, p+1)$ prior for α and set $\lambda = 0.01$. We ran 10,000 iterations for both methods. Figure 5 displays the 95% credible intervals. Due to the large number of parameters, we only show the results for the first 10 parameters in the plot, but the same pattern is observed for other parameters. We can see that ProxMCMC credible intervals are consistently narrower and provide good coverage of the graphical lasso estimates, whereas those provided by Bayesian graphical lasso can be wide or fail to cover the cross-validated estimates. Among all 66 parameters, all ProxMCMC credible intervals cover the reference values whereas only 24% of Bayesian graphical lasso credible intervals do. The effective sample size of the slowest Θ component is 5540.

6.3. Matrix Completion: Simulated Matrix

We simulate the true low-rank matrix as $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{X}_2 + \sigma\boldsymbol{E}$, where $\boldsymbol{X}_1 \in \mathbb{R}^{50\times 2}$, $\boldsymbol{X}_2 \in \mathbb{R}^{2\times 50}$, $\sigma = 0.5$, and entries of $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{E}$ are generated from the standard normal distribution. We randomly mask 25%, 50%, and 75% of the entries and apply ProxMCMC to calculate the posterior median and 95% credible intervals for the missing entries. We use an $IG(0.01, 0.01)$ prior for σ^2 and an $IG(1, 50\times 50+1)$ prior for α, and set $\lambda = 0.01$. The number of HMC samples is set at 1000. For comparison, we also try an empirical Bayesian method called the stochastic approximation proximal gradient (SAPG) (De Bortoli et al. 2020; Vidal et al. 2020), and use the SK-ROCK method (Pereyra, Mieles, and Zygalakis 2020) for posterior sampling. The details of this approach is left to the supplementary materials. Table 1 displays the mean absolute deviation (MAD) averaged over missing entries and the percentage of missing entries covered by their 95% credible intervals for the two methods at different missing rate. As expected, the posterior average MAD increases as the missing rate increases. We also see that, for a given missing rate, the average MAD of ProxMCMC is lower than that of SAPG, and the credible intervals provided by ProxMCMC cover an equal or higher percentage of missing entries than that provided by SAPG. The results indicate that ProxMCMC has superior statistical precision, likely because ProxMCMC is fully Bayesian and accounts for the uncertainty of α and σ^2, while SAPG commits to a single point estimate of α and σ^2 after hyperparameter calibration. We also emphasize that it is not straightforward to apply SAPG to problems with constraints, such as the constrained lasso or the sparse low rank matrix regression problem. Therefore, ProxMCMC offers greater flexibility in model formulation.

Table 1. Comparison between ProxMCMC and stochastic approximation proximal gradient (SAPG) for the matrix completion example.

	ProxMCMC		SAPG	
Percent missing	Average MAD	Percent covered	Average MAD	Percent covered
25%	0.23	100%	0.24	100%
50%	0.34	100%	0.42	99%
75%	0.74	97%	0.79	93%

NOTE: "Average MAD" is the mean absolute deviation averaged over missing entries; "Percent covered" is the percentage of missing entries covered by their 95% credible intervals.

6.4. Sparse Low Rank Matrix Regression: Detecting the Butterfly Signal

We simulate data from the following model: the mean response for the ith sample is $\mu_i = \boldsymbol{Z}_i^T\boldsymbol{\gamma} + \langle \boldsymbol{B}, \boldsymbol{X}_i\rangle$, where $\boldsymbol{Z}_i \in \mathbb{R}^2$ and $\boldsymbol{X}_i \in \mathbb{R}^{25\times 25}$ are vector and matrix covariates, whose entries are generated from iid standard normal. We set the true $\boldsymbol{\gamma} = (1,1)^T$ and let $\boldsymbol{B}$ be the 25×25 butterfly signal shown in Figure 6 (left), where black pixels equal 0, white pixels 1, and grey pixels between 0 and 1. The response for the ith sample, y_i, equals $\mu_i + \epsilon_i$, where ϵ_i is generated from iid standard normal. We use an $IG(0.01, 0.01)$ prior for σ^2 and an $IG(\sum_i \sigma(\boldsymbol{B}_0)_i, 2)$ prior for α, where $\sigma(\boldsymbol{B}_0)_i$ is the ith singular value of $\boldsymbol{B}_0$, and $\boldsymbol{B}_0$ is the least squares estimate of $\boldsymbol{B}$ obtained without regularization or constraints. We set the Moreau-Yosida envelope parameter $\lambda = 0.001$. Figure 6 shows the true signal $\boldsymbol{B}$ (left) and the posterior mean from 10,000 HMC samples at sample size $N = 2000$ (middle) and $N = 5000$ (right). For inference, we calculated the 95% credible intervals for entries of $\boldsymbol{B}$ and found that among the 625 ($= 25\times 25$) entries, 94% are covered by their 95% credible intervals at both sample sizes. The effective sample size of the slowest component of $\boldsymbol{B}$ is 361 at $N = 2000$, and 2054 at $N = 5000$.

7. Discussion

The examples above demonstrate that the ProxMCMC method is a highly flexible tool for obtaining statistical inference on regularized or constrained statistical learning problems. We find that it works well when the regularization or constraints are non-

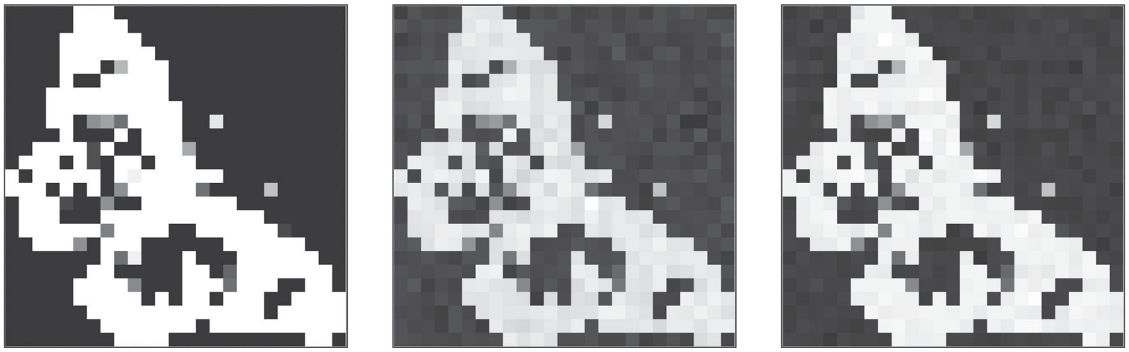

Figure 6. ProxMCMC for sparse low rank matrix regression on the butterfly signal. Left: true signal; Middle: posterior mean at sample size 2000; Right: posterior mean at sample size 5000. Black pixels equal 0, white pixels 1, and grey pixels between 0 and 1.

smooth and even non-convex. In addition, by adopting epigraph priors, our method is fully Bayesian, eliminating the need for tuning the regularization strength parameter.

The Moreau-Yosida envelope parameter λ controls how well the smoothed posterior approximates the original posterior. For constrained problems, a smaller λ leads to better satisfaction of the constraints. For example, the histogram of $\sum_j \boldsymbol{\beta}_j$ from the constrained lasso simulation experiment is more concentrated around 0 when λ is smaller. Choosing λ values that are too small, however, renders slow mixing of the sampling algorithm. We leave a more in-depth investigation of this phenomenon to future work. For practical purposes, we recommend using smaller λ when computational resources allow. Setting $\lambda = 0.001$ seems to work well in most applications as the examples show.

Finally, we emphasize that the four examples are meant to whet readers' appetites, not to satiate them. As demonstrated through these examples, the proposed ProxMCMC method is highly modular and can be readily extended to other problems. We hope that this article offers sufficient detail for readers to explore new applications of the ProxMCMC algorithm.

Supplementary Materials

Supplementary materials are available online and include the theoretical properties of ProxMCMC and additional experiment results. The Julia code for reproducing the numerical results are available at *https://github.com/xinkai-zhou/ProxMCMCExamples*.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This research was partially funded by grants from the National Science Foundation (DMS-2054253, IIS-2205441) and the National Institutes of Health (R35GM141798, R01HG006139, R01GM135928).

ORCID

Xinkai Zhou http://orcid.org/0000-0001-8564-3491
Qiang Heng http://orcid.org/0000-0002-4042-6773
Hua Zhou http://orcid.org/0000-0003-1320-7118

References

Armagan, A., Dunson, D., and Lee, J. (2013), "Generalized Double Pareto Shrinkage," *Statistica Sinica*, 23, 119–143. [379]

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2020), "Uniformly Valid Confidence Intervals Post-Model-Selection," *The Annals of Statistics*, 48, 440–463. [379]

Beck, A. (2017), *First-Order Methods in Optimization* volume 25 of MOS-SIAM Series on Optimization, Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM); Philadelphia, PA: Mathematical Optimization Society. [381,382,384,385]

Beck, A., and Guttmann-Beck, N. (2019), "FOM–A MATLAB Toolbox of First-Order Methods for Solving Convex Optimization Problems," *Optimization Methods and Software*, 34, 172–193. [384]

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid Post-selection Inference," *The Annals of Statistics*, 41, 802–837. [379]

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019), "Lasso Meets Horseshoe: A Survey," *Statistical Science*, 34, 405–427. [379]

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015), "Dirichlet-Laplace Priors for Optimal Shrinkage," *Journal of the American Statistical Association*, 110, 1479–1490. [379]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [379]

Chi, E. C., Zhou, H., and Lange, K. (2014), "Distance Majorization and its Applications," *Mathematical Programming*, 146, 409–436. [384]

Choi, Y., Taylor, J., and Tibshirani, R. (2017), "Selecting the Number of Principal Components: Estimation of the True Rank of a Noisy Matrix," *The Annals of Statistics*, 45, 2590–2617. [379]

Combettes, P. L., and Pesquet, J.-C. (2011), *Proximal Splitting Methods in Signal Processing*, pp. 185–212, New York: Springer. [380]

Combettes, P. L., and Wajs, V. R. (2005), "Signal Recovery by Proximal Forward-Backward Splitting," *Multiscale Modeling & Simulation*, 4, 1168–1200. [380]

De Bortoli, V., Durmus, A., Pereyra, M., and Vidal, A. F. (2020), "Maximum Likelihood Estimation of Regularization Parameters in High-Dimensional Inverse Problems: An Empirical Bayesian Approach. Part II: Theoretical Analysis," *SIAM Journal on Imaging Sciences*, 13, 1990–2028. [380,388]

Durmus, A., Moulines, E., and Pereyra, M. (2018), "Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin meets Moreau," *SIAM Journal on Imaging Sciences*, 11, 473–506. [379,380,387]

——— (2022), "A Proximal Markov Chain Monte Carlo Method for Bayesian Inference in Imaging Inverse Problems: When Langevin Meets Moreau," *SIAM Review*, 64, 991–1028. [379,380]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [382,383]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9, 432–441. [387]

Gaines, B. R., Kim, J., and Zhou, H. (2018), "Algorithms for Fitting the Constrained Lasso," *Journal of Computational and Graphical Statistics*, 27, 861–871. [385]

Ge, H., Xu, K., and Ghahramani, Z. (2018), "Turing: A Language for Flexible Pobabilistic Inference," in *Proceedings of the Twenty-First International*

Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pp. 1682–1690, PMLR. [384]
George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [379]
Gramacy, R. B. (2019), *monomvn: Estimation for MVN and Student-t Data with Monotone Missingness*, R package version 1.9-13. [382]
Griffin, J. E., and Brown, P. J. (2012), "Structuring Shrinkage: Some Correlated Priors for Regression," *Biometrika*, 99, 481–487. [379]
——— (2013), "Some Priors for Sparse Regression Modelling," *Bayesian Analysis*, 8, 691–702. [379]
Hans, C. (2011), "Elastic Net Regression Modeling with the Orthant Normal Prior," *Journal of the American Statistical Association*, 106, 1383–1393. [379]
Heng, Q., Zhou, H., and Chi, E. C. (2023), "Bayesian Trend Filtering via Proximal Markov Chain Monte Carlo," *Journal of Computational and Graphical Statistics*, 32, 938–949. [385]
James, G. M., Paulson, C., and Rusmevichientong, P. (2020), "Penalized and Constrained Optimization: an Application to High-Dimensional Website Advertising," *Journal of the American Statistical Association*, 115, 107–122. [385]
Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [379]
Keys, K. L., Zhou, H., and Lange, K. (2019), "Proximal Distance Algorithms: Theory and Practice," *Journal of Machine Learning Research*, 20, 1–38. [384]
Kuchibhotla, A. K., Brown, L. D., Buja, A., Cai, J., George, E. I., and Zhao, L. H. (2020), "Valid Post-Selection Inference in Model-Free Linear Regression," *The Annals of Statistics*, 48, 2953–2981. [379]
Landeros, A., and Lange, K. (2021), "Algorithms for Sparse Support Vector Machines," arXiv:2110.07691 [stat.ME]. [384]
Landeros, A., Padilla, O. H. M., Zhou, H., and Lange, K. (2022), "Extensions to the Proximal Distance Method of Constrained Optimization," arXiv:2009.00801 [math.OC]. [384]
Landeros, A., Wu, T. T., and Lange, K. (2022), "Feature Selection for Vertex Discriminant Analysis," arXiv:2203.11168 [stat.CO]. [384]
Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), "Exact Post-selection Inference, with Application to the Lasso," *The Annals of Statistics*, 44, 907–927. [379,382]
Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322. [386]
Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1036. With comments by James Berger and C. L. Mallows and with a reply by the authors. [379]
Neal, R. M. (2011),""MCMC Using Hamiltonian Dynamics," in *Handbook of Markov Chain Monte Carlo* (Vol. 2), eds. S. Brooks, A. Gelman, G. L. Jones, X.-L. Meng, pp. 113–162, Boca Raton, FL: CRC Press. [382]
Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [379]
Pereyra, M. (2016), "Proximal Markov Chain Monte Carlo Algorithms," *Statistics and Computing*, 26, 745–760. [379]
Pereyra, M., Mieles, L. V., and Zygalakis, K. C. (2020), "Accelerating Proximal Markov Chain Monte Carlo by Using an Explicit Stabilized Method," *SIAM Journal on Imaging Sciences*, 13, 905–935. [388]
Piironen, J., and Vehtari, A. (2017), "Sparsity Information and Regularization in the Horseshoe and Other Shrinkage Priors," *Electronic Journal of Statistics*, 11, 5018–5051. [379]
Polson, N. G., and Scott, J. G. (2010), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," in *Bayesian Statistics* (Vol. 9), eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, pp. 501–538, Oxford: Oxford University Press. [379]
Polson, N. G., Scott, J. G., and Willard, B. T. (2015), "Proximal Algorithms in Statistics and Machine Learning," *Statistical Science*, 30, 559–581. [380]
Presman, R., and Xu, J. (2022), "Distance-to-Set Priors and Constrained Bayesian Inference," arXiv:2210.12258 [stat.ME]. [384]
Rockafellar, R. T., and Wets, R. J.-B. (2009), *Variational Analysis* (Vol. 317), Berlin: Springer. [381,386]
Rudin, L. I., Osher, S., and Fatemi, E. (1992), "Nonlinear Total Variation based Noise Removal Algorithms," *Physica D: Nonlinear Phenomena*, 60, 259–268. [380]
Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005), "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data," *Science*, 308, 523–529. [387]
Stan Development Team. (2020), "Stan Modeling Language Users Guide and Reference Manual." [384]
Taylor, J., and Tibshirani, R. (2018), "Post-Selection Inference for-Penalized Likelihood Models," *Canadian Journal of Statistics*, 46, 41–61. [379]
Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [382]
Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., Reid, S., and Markovic, J. (2019), *selectiveInference: Tools for Post-Selection Inference*. R package version 1.2.5. [382]
van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [379]
van der Pas, S., Scott, J., Chakraborty, A., and Bhattacharya, A. (2019), *horseshoe: Implementation of the Horseshoe Prior*, R package version 0.2.0. [382]
Vidal, A. F., De Bortoli, V., Pereyra, M., and Durmus, A. (2020), "Maximum Likelihood Estimation of Regularization Parameters in High-Dimensional Inverse Problems: An Empirical Bayesian Approach. Part I: Methodology and Experiments," *SIAM Journal on Imaging Sciences*, 13, 1945–1989. [380,388]
Wang, H. (2012), "Bayesian Graphical Lasso Models and Efficient Posterior Computation," *Bayesian Analysis*, 7, 867–886. [387,388]
Xu, J., Chi, E., and Lange, K. (2017), "Generalized Linear Model Regression Under Distance-to-Set Penalties," in *Advances in Neural Information Processing Systems* (Vol. 30), eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Curran Associates, Inc. [384]
Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [379]
Zhou, H., and Li, L. (2014), "Regularized Matrix Regression," *Journal of the Royal Statistical Society*, Series B, 76, 463–483. [386]
Zhou, H., Li, L., and Zhu, H. (2013), "Tensor Regression with Applications in Neuroimaging Data Analysis," *Journal of the American Statistical Association*, 108, 540–552. [386]