HHS Public Access

Author manuscript

Bayesian Anal. Author manuscript; available in PMC 2024 May 20.

Published in final edited form as:

Bayesian Anal. 2023 June ; 18(2): 367–390. doi:10.1214/22-ba1308.

Shrinkage with shrunken shoulders: Gibbs sampling shrinkage model posteriors with guaranteed convergence rates

Akihiko Nishimura[*], **Marc A. Suchard**[†]

[*]Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health.

[†]Departments of Biomathematics, Biostatistics, and Human Genetics, University of California – Los Angeles.

Abstract

Use of continuous shrinkage priors — with a "spike" near zero and heavy-tails towards infinity — is an increasingly popular approach to induce sparsity in parameter estimates. When the parameters are only weakly identified by the likelihood, however, the posterior may end up with tails as heavy as the prior, jeopardizing robustness of inference. A natural solution is to "shrink the shoulders" of a shrinkage prior by lightening up its tails beyond a reasonable parameter range, yielding a *regularized* version of the prior. We develop a regularization approach which, unlike previous proposals, preserves computationally attractive structures of original shrinkage priors. We study theoretical properties of the Gibbs sampler on resulting posterior distributions, with emphasis on convergence rates of the Pólya-Gamma Gibbs sampler for sparse logistic regression. Our analysis shows that the proposed regularization leads to geometric ergodicity under a broad range of global-local shrinkage priors. Essentially, the only requirement is for the prior $\pi_{\text{local}}(\cdot)$ on the local scale λ to satisfy $\pi_{\text{local}}(0) < \infty$. If $\pi_{\text{local}}(\cdot)$ further satisfies $\lim_{\lambda \to 0} \pi_{\text{local}}(\lambda)/\lambda^a < \infty$ for $a > 0$, as in the case of Bayesian bridge priors, we show the sampler to be uniformly ergodic.

Keywords

Bayesian inference; sparsity; generalized linear model; Markov chain Monte Carlo; ergodicity

MSC 2010 subject classifications:

Primary 60J20, 62F15; secondary 62J07

1 Introduction

Bayesian modelers are increasingly adopting continuous shrinkage priors to control the effective number of parameters and model complexity in a data-driven manner. These priors are designed to shrink most of the parameters towards zero while allowing for the likelihood to pull a small fraction of them away from zero. To achieve such effects, a shrinkage prior[1]

aki.nishimura@jhu.edu .

[1]We drop the word "continuous" since "shrinkage priors" are commonly understood in the literature as continuous ones, which exclude traditional discrete spike-slab mixtures.

has a density with a “spike” near zero and heavy-tails towards infinity, encoding information that parameter values are likely close to zero but otherwise could be anywhere. Originally developed for the purpose of sparse regression (Carvalho et al., 2009), shrinkage priors have found applications in trend filtering of time series data (Kowal et al., 2019), (dynamic) factor models (Kastner, 2019), graphical models (Li et al., 2019), compression of deep neural networks (Louizos et al., 2017), among others.

Of particular interest in this paper is an application of Bayesian shrinkage to a logistic regression model $y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta} \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right)$ and computational properties of the corresponding posterior inference via Gibbs sampling. Due to the possibility of $\boldsymbol{\beta}$ being only weakly identifiable, use of a shrinkage prior on $\boldsymbol{\beta}$ here warrants proper modification of the prior’s tail in order to ensure reasonable computational and statistical behaviors. Under our tail regularization strategy, we show that the Gibbs sampler achieves geometric ergodicity under a broad range of shrinkage priors. Notably, our proof technique unifies analyses of the Gibbs samplers under various shrinkage priors, providing an easily verifiable condition for geometric and uniform ergodicity.

Shrinkage priors are often expressed as a scale mixture of Gaussians on the unknown parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ (Polson and Scott, 2010):

$$\pi(\beta_j \mid \tau, \lambda_j) \sim \mathcal{N}\left(0, \tau^2 \lambda_j^2\right), \lambda_j \sim \pi_{\text{loc}}(\cdot). \tag{1.1}$$

This *global-local* representation simplifies the posterior conditionals and lead to straightforward inference via Gibbs sampling. The *global scale* τ controls the average magnitude of β_j’s and hence overall sparsity level. The *local scale* λ_j is specific to individual β_j and its density $\pi_{\text{loc}}(\cdot)$ controls the size of the spike and tail behavior of the marginal $\beta_j \mid \tau$. For instance, the popular *horseshoe* prior of Carvalho et al. (2010) uses $\pi_{\text{loc}}\left(\lambda\right) \propto \left(1 + \lambda^2\right)^{-1}$, inducing a marginal $\pi(\beta_j \mid \tau)$ with the spike proportional to $-\log(|\beta_j/\tau|)$ as $|\beta_j/\tau| \to 0$ and the tail proportional to $(\beta_j/\tau)^{-2}$ as $|\beta_j/\tau| \to \infty$. Another notable example is the Bayesian bridge prior of Polson et al. (2014), which generalizes the Bayesian lasso of Park and Casella (2008) with $\pi(\beta_j \mid \tau)$ having a larger spike as $|\beta_j/\tau| \to 0$ and heavier tails as $|\beta_j/\tau| \to \infty$. Most importantly from the computational efficiency perspective, the bridge prior possesses a closed-form expression $\pi(\beta_j \mid \tau) \propto \exp\left(-|\beta_j/\tau|^a\right)$ for $a \in (0,1)$ and thus allows for a collapsed Gibbs update from $\tau \mid \boldsymbol{\beta}$ with λ_j’s marginalized out.

For a simple purpose such as estimating the unknown means of independent Gaussian observations, a broad class of shrinkage priors achieve theoretically optimal performance (van der Pas et al., 2016; Ghosh and Chakrabarti, 2017). The lack of prior information in the tail of the distribution is problematic, however, in more complex models where parameters are only weakly identified. In such models, the posterior may have a tail as heavy as the prior, resulting in unreliable parameter estimates (Ghosh et al., 2018).

To address the above shortcoming of shrinkage priors, we build on the work of Piironen and Vehtari (2017) and propose a computationally convenient way to *regularize* shrinkage priors. The basic idea is to modify the prior so that the marginal distribution of $|\beta_j|$ has light-tails beyond a reasonable range. Our formulation has computational advantages over that of Piironen and Vehtari (2017) due to a subtle yet important difference. By preserving the global-local structure (1.1), our regularized shrinkage priors can benefit from partial marginalization approaches that substantially improve mixing of Gibbs samplers (Polson et al. 2014; Johndrow et al. 2018; Appendix E). In addition, our regularization leaves the posterior conditionals of λ_j's unchanged, allowing their conditional updates via existing specialized samplers (Griffin and Brown 2010; Polson et al. 2014; Appendix F).[2]

Our regularized shrinkage priors allow for posterior inference via Gibbs sampler whose convergence rates often are provably fast. As an illustrative example, we consider Bayesian sparse logistic regression models, whose need for regularization motivated the work of Piironen and Vehtari (2017). Gibbs sampling via the Pólya-Gamma data augmentation of Polson et al. (2013) is a state-of-the-art approach to posterior computation under logistic model. When combined with advanced numerical linear algebra techniques, this Gibbs sampler is highly scalable to large data sets (Nishimura and Suchard, 2018), but its theoretical convergence rate has not been investigated. Assuming that the prior density $\pi_{\text{loc}}(\lambda)$ is continuous and bounded except possibly at $\lambda = 0$, we establish that the Gibbs sampler is geometrically ergodic whenever $\pi_{\text{loc}}(0) < \infty$. Stronger uniform convergence is achieved when $\int \lambda^{-1} \pi_{\text{loc}}(\lambda) \mathrm{d}\lambda < \infty$. The integrability condition holds in particular when $\pi_{\text{loc}}(\lambda) = O(\lambda^a)$ for $a > 0$ as $\lambda \to 0$, which is the case for normal-gamma priors with shape parameter larger than 1/2 (Griffin and Brown, 2010) and for Bayesian bridge priors (Polson et al. 2014 and Appendix E).

Previous studies of the convergence rates under shrinkage models have focused exclusively on linear regression with specific parametric families of shrinkage priors (Pal and Khare, 2014; Johndrow et al., 2018). In contrast, our analysis requires no parametric assumptions on the shrinkage prior, at the same time extending the convergence results to the logistic model and, in Appendix A, to the probit model.

To summarize, this work provides two major contributions to the Bayesian shrinkage literature. First, we propose an effective and Gibbs-friendly approach to suitably modify shrinkage priors for use in weakly-identifiable models (Section 2). Second, we develop theoretical tools to study the behavior of shrinkage model Gibbs samplers near the spike $\beta_j = 0$ without any parametric assumption on $\pi_{\text{loc}}(\cdot)$, thereby unifying convergence analyses of the logistic regression Gibbs samplers under a range of shrinkage priors (Section 3). We conclude the article in Section 4 by demonstrating a practical use case of regularized shrinkage models via simulation study, which emulates increasingly common situations where the sample sizes are large yet the signals are difficult to detect. Our simulation results in particular highlight the dual role of the regularization; by eliminating heavy-tails in the

[2]Appendix F describes a simple and provably efficient rejection-sampler for the conditional distributions of local scale parameter λ_j's under the horseshoe prior. Despite the horseshoe's popularity, we find that no existing algorithm for the conditional update comes with theoretically guaranteed efficiency.

shrinkage model posterior, it induces both more stable parameter estimates and faster mixing of the Gibbs sampler.

2 Regularized shrinkage prior

This section explains how our regularization approach allows us to incorporate prior information on the largest possible parameter values while maintaining the computational tractability of the original shrinkage prior.

Piironen and Vehtari (2017) proposes to control the tail behavior of a global-local shrinkage prior by defining its regularized version with *slab width* $\zeta > 0$ as

$$\beta_j \mid \tau, \lambda_j, \zeta \sim \mathcal{N}\left(0, \left(\frac{1}{\zeta^2} + \frac{1}{\tau^2\lambda_j^2}\right)^{-1}\right), \tag{2.1}$$

with the prior $\pi_{\text{loc}}(\cdot)$ on the local scale λ_j unmodified. This regularization ensures that the variance of $\beta_j \mid \tau, \lambda_j, \zeta$ is upper bounded by ζ^2 and hence $\beta_j \mid \zeta$ marginally has a density with Gaussian tails beyond $|\beta_j| > \zeta$. The slab width ζ can be either given a prior distribution or fixed at a reasonable value.[3]

While beneficial in improving statistical properties (Piironen and Vehtari, 2017), regularization the form (2.1) compromises the posterior conditional structures of shrinkage models. Specifically, the conditional distribution of τ, λ is altered through their dependency on ζ. This structural change is at best an inconvenience and potentially a cause of computational inefficiency, prohibiting the use of common acceleration techniques. For instance, the global scale τ is known to mix slowly when updating from its full conditional, so the state-of-the-art Gibbs samplers for Bayesian sparse regression marginalize out a subset of parameters when updating τ (Johndrow et al., 2018; Nishimura and Suchard, 2018). The analytical tractabilities of the integrals, which these marginalization strategies rely on, is lost when using the regularization as in (2.1).

We propose a more computationally convenient formulation, which induces regularization similar to that of (2.1) while keeping τ and λ conditionally independent of ζ given $\boldsymbol{\beta}$. Our regularized prior $\pi_{\text{reg}}(\cdot)$ defines the distribution of $\beta_j, \lambda_j \mid \tau, \zeta$ as

$$\begin{aligned}\pi_{\text{reg}}(\beta_j, \lambda_j \mid \tau, \zeta) &\propto \exp\left(-\frac{\beta_j^2}{2\zeta^2}\right)\frac{1}{\tau\lambda_j}\exp\left(-\frac{\beta_j^2}{2\tau^2\lambda_j^2}\right)\pi_{\text{loc}}(\lambda_j) \\ &\propto \mathcal{N}\left(\beta_j \mid 0, \left(\frac{1}{\zeta^2} + \frac{1}{\tau^2\lambda_j^2}\right)^{-1}\right)\left(1 + \frac{\tau^2\lambda_j^2}{\zeta^2}\right)^{-1/2}\pi_{\text{loc}}(\lambda_j)\end{aligned} \tag{2.2}$$

[3]While an appropriate choice of ζ is application specific, by way of illustration, we suggest $\zeta = 2$ as a weakly informative and sensible starting point in biomedical applications with standardized predictors. Schuemie et al. (2018) surveys 59,196 published effect estimates in the observational study literature and finds only a small portion of them exceeds 2.

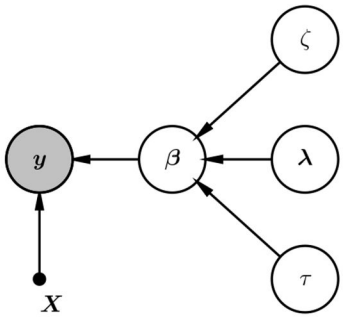
where $\mathcal{N}\left(\cdot \mid 0, \sigma^2\right)$ denotes the centered Gaussian density with variance σ^2. In other words, in addition to defining $\pi(\beta_j \mid \tau, \lambda_j, \zeta)$ as in (2.1), we alter the prior on λ_j as $\pi(\lambda_j \mid \tau, \zeta) \propto \pi_{\text{loc}}(\lambda_j)/\sqrt{1 + \tau^2\lambda_j^2/\zeta^2}$. Incidentally, we see that our regularized prior is very similar to that of Piironen and Vehtari (2017), but has a slightly lighter tail due to the factor $1/\sqrt{1 + \tau^2\lambda_j^2/\zeta^2}$ which, as $\lambda_j \to \infty$, behaves like $\zeta/\tau\lambda_j$.

Alternatively, we can achieve the equivalent regularization through fictitious data that makes values $|\beta_j| \gg \zeta$ unlikely. While it may appear unnatural to introduce an auxiliary likelihood for the purpose of indirectly modifying a prior, this alternative formulation makes the regularization mechanism and resulting posterior properties more transparent. Figure 2.1 schematically describes this alternative construction of our regularized prior as well as the corresponding posterior structure when data $\boldsymbol{y}$ and $\boldsymbol{X}$ inform $\boldsymbol{\beta}$ through the likelihood $L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$.

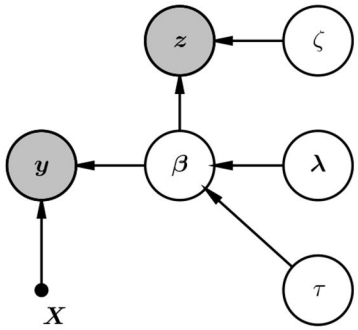
Given a global-local prior $\beta_j \mid \tau, \lambda_j \sim \mathcal{N}\left(0, \tau^2\lambda_j^2\right)$, we introduce fictitious data z_j whose realized value and underlying distribution are assumed to be

$$z_j = 0, z_j \mid \beta_j, \zeta \sim \mathcal{N}\left(\beta_j, \zeta^2\right) \tag{2.3}$$

for $j = 1, \ldots, p$. We then define the regularized prior as the distribution of β_j conditional on $z_j = 0$. Under this model, the distribution of $\beta_j \mid \tau, \lambda_j, \zeta, z_j = 0$ coincides with that of (2.1). On the other hand, the scale parameters τ, λ are conditionally independent of the others given $\boldsymbol{\beta}$, so that the posterior full conditional $\tau, \lambda \mid \boldsymbol{\beta}, \zeta, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{X}\left(\overset{d}{=} \tau, \lambda \mid \boldsymbol{\beta}\right)$ has the same density as in the unregularized version. Our regularization thus allows the Gibbs sampler to update τ, λ with the exact same algorithm as the one designed for the original shrinkage prior. We summarize our discussion as Proposition 2.1 below.



(a) Of the form (2.1) as previously proposed. The posterior conditional of $(\tau, \boldsymbol{\lambda})$ is affected by their dependency on ζ through $\boldsymbol{\beta}$.



(b) Of the form (2.3) as in Proposition 2.1. Regularization does not affect the posterior conditional of $(\tau, \boldsymbol{\lambda})$ as the parameters remains decoupled from ζ.

Figure 2.1:

Directed acyclic graphical model (a.k.a. Bayesian network) representation of regularized shrinkage priors under the two alternative formulations.

Proposition 2.1. *Consider a global-local shrinkage prior* $\beta_j \mid \tau, \lambda_j \sim \mathcal{N}\left(0, \tau^2\lambda_j^2\right)$, $\lambda_j \sim \pi_{\text{loc}}(\cdot)$ *and* $\tau \sim \pi_{\text{glo}}(\cdot)$. *Introducing the fictitious data* $\boldsymbol{z} = \mathbf{0}$ *as in* (2.3) *is equivalent to using the regularized prior* (2.2) *on* (β_j, λ_j), *yielding*

$$\beta_j \mid \tau, \lambda_j, \zeta, z_j = 0 \sim \mathcal{N}\left(0, \left(\frac{1}{\zeta^2} + \frac{1}{\tau^2\lambda_j^2}\right)^{-1}\right).$$

Or, with λ_j *marginalized out, we have*

$$\pi(\beta_j \mid \tau, \zeta, z_j = 0) \propto \pi(\beta_j \mid \tau)\exp\left(-\frac{\beta_j^2}{2\zeta^2}\right).$$

When the likelihood depends only on $\boldsymbol{\beta}$, the posterior full conditional of $\tau, \boldsymbol{\lambda}$ *has density*

$$\pi(\tau, \boldsymbol{\lambda} \mid \boldsymbol{\beta}) \propto \pi_{\text{glo}}(\tau)\prod_j \frac{1}{\tau\lambda_j}\exp\left(-\frac{\beta_j^2}{2\tau^2\lambda_j^2}\right)\pi_{\text{loc}}(\lambda_j). \tag{2.4}$$

3 Geometric and uniform ergodicity under regularized sparse logistic regression

Shrinkage priors' popularity stems from, to a considerable extent, the ease of posterior computation via Gibbs sampling (Bhadra et al., 2017). As we have shown in Section 2, shrinkage models can incorporate regularization without affecting its computational tractability. We now investigate how fast such Gibbs samplers converge. While regularization was originally motivated to remedy statistically problematic behavior of heavy-tailed shrinkage priors, our results show that it can also improve the Gibbs samplers' convergence rates. The simulation results of Section 4 further corroborate the theory.

As a representative example where regularization is essential, we focus on Bayesian sparse logistic regression (Piironen and Vehtari, 2017; Nishimura and Suchard, 2018). To be explicit, we consider the model

$$\begin{aligned} y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta} &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right), \\ z_j = 0 \text{ for } z_j \mid \beta_j &\sim \mathcal{N}\left(\beta_j, \zeta^2\right), \\ \beta_j \mid \tau, \lambda_j \sim \mathcal{N}\left(0, \tau^2\lambda_j^2\right), \tau \sim \pi_{\text{glo}}(\cdot), \lambda_j &\sim \pi_{\text{loc}}(\cdot). \end{aligned} \tag{3.1}$$

The Pólya-Gamma data-augmentation of Polson et al. (2013) is a widely-used approach to carry out the posterior computation under the logistic model. By introducing an auxiliary

parameter $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ having a Pólya-Gamma distribution, the Gibbs sampler induces a transition kernel: $(\boldsymbol{\omega}^*, \boldsymbol{\beta}^*, \lambda^*, \tau^*) \to (\boldsymbol{\omega}, \boldsymbol{\beta}, \lambda, \tau)$ through the following cycle of conditional updates:

1. Draw $\tau \mid \boldsymbol{\beta}^*, \lambda^*$ from the density proportional to (2.4). When using Bayesian bridge priors, draw from the collapsed distribution $\tau \mid \boldsymbol{\beta}^*$ (Appendix E).
2. Draw $\lambda \mid \boldsymbol{\beta}^*, \tau$ from the density proportional to (2.4).
3. Draw $\omega_i \mid \boldsymbol{\beta}^*, \boldsymbol{X} \sim \text{PolyaGamma}\left(\text{shape} = 1, \text{tilting} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^*\right)$ for $i = 1, \ldots, n$.
4. Draw $\boldsymbol{\beta} \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = 0$ from the multivariate-Gaussian

$$\boldsymbol{\beta} \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = 0 \sim \mathcal{N}\left(\boldsymbol{\Phi}^{-1}\boldsymbol{X}^\top\left(\boldsymbol{y} - \frac{1}{2}\right), \boldsymbol{\Phi}^{-1}\right) \text{ for } \boldsymbol{\Phi} = \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} + \zeta^{-2}\boldsymbol{I} + \tau^{-2}\boldsymbol{\Lambda}^{-2}, \tag{3.2}$$

where $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega})$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda)$.

Note that the transition kernel actually depends neither on $\boldsymbol{\omega}^*$ nor τ^* (nor λ^* in the Bayesian bridge case) because of conditional independence. We refer readers to Polson et al. (2013) for more details on this data augmentation scheme. In our analysis, we do not use any specific properties of the Pólya-Gamma distribution aside from a couple of results from Choi and Hobert (2013) and Wang and Roy (2018).

The Pólya-Gamma Gibbs sampler for the logistic model has previously been analyzed under a Gaussian or flat prior on $\boldsymbol{\beta}$ (Choi and Hobert, 2013; Wang and Roy, 2018), but not under shrinkage priors. We establish geometric and uniform ergodicity — critical properties for any practical Markov chain Monte Carlo algorithms (Jones and Hobert, 2001). These properties imply the Markov chain central limit theorem and enables consistent estimation of Monte Carlo errors, ensuring that the Gibbs sampler reliably estimates quantities of interest (Flegal and Jones, 2011). To avoid cluttering notations and obscuring the main ideas, our analysis below assumes the slab width ζ to be fixed; however, the same conclusions hold if we only assume a prior constraint of the form $\zeta \le \zeta_{\max} < \infty$ (Remark 3.9).

Below are the main ergodicity results we will establish in this section, the uniform rate under Bayesian bridge and geometric rate under more general shrinkage priors:

Theorem 3.1 (Uniform ergodicity in the Bayesian bridge case). *If the prior $\pi_{\text{glo}}(\cdot)$ is supported on $[\tau_{\min}, \infty)$ for $\tau_{\min} > 0$, then the Pólya-Gamma Gibbs sampler for regularized Baysian bridge logistic regression is uniformly ergodic.*

Theorem 3.2 (Geometric ergodicity). *Suppose that the local scale prior satisfies $\| \pi_{\text{loc}} \|_\infty < \infty$ and that the global scale prior $\pi_{\text{glo}}(\cdot)$ is supported on $[\tau_{\min}, \tau_{\max}]$ for $0 < \tau_{\min} \le \tau_{\max} < \infty$. Then the Pólya-Gamma Gibbs sampler for regularized sparse logistic regression is geometrically ergodic.*

Remark. Uniform / geometric ergodicity is an essential requirement for, yet not a guarantee of, practically efficient Markov chains (Roberts and Rosenthal, 2004). In fact, the simulation results of Section 4 show that the benefit of regularization is greatest when ζ is chosen small enough to impose a reasonable prior constraint on the value of β_j's.

3.1 Proof approach: minorization and drift conditions

To establish Theorem 3.1 and 3.2, we verify that each Gibbs sampler satisfies the *minorization* and *drift* conditions, upon on which geometric and uniform ergodicity are immediately implied by the well-known theory of Markov chains (Meyn and Tweedie, 2009; Roberts and Rosenthal, 2004). Here we introduce the relevant notions in terms of a generic transition kernel $P(\theta^*, \mathrm{d}\theta)$.

In the statements to follow, we assume that $P(\theta^*, \mathrm{d}\theta)$ has a corresponding density function which, with slight abuse of notation, we denote by $P(\theta \mid \theta^*)$; in other words, the two satisfy the relation $P(\theta^*, A) = \int_A P(\theta \mid \theta^*)\mathrm{d}\theta$. A chain on the space $\theta \in \mathbf{\Theta}$ with transition kernel $P(\theta^*, \mathrm{d}\theta)$ is said to satisfy a minorization condition with a *small set* S if there are $\delta > 0$ and a probability density $\pi(\cdot)$ such that

$$P(\theta \mid \theta^*) \geq \delta\pi(\theta) \text{ for all } \theta^* \in S.$$

The chain is uniformly ergodic when $S = \mathbf{\Theta}$. Otherwise, the chain is geometrically ergodic if it additionally satisfies a drift condition i.e. there is a *Lyapunov* function $V(\theta) \geq 0$ such that, for $\gamma < 1$ and $b < \infty$,

$$PV(\theta^*) := \int V(\theta) P(\theta \mid \theta^*)\mathrm{d}\theta \leq \gamma V(\theta^*) + b$$

and $S = \{\theta : V(\theta) \leq d\}$ is a small set for some $d > 2b/(1-\gamma)$ (Rosenthal, 1995).

For a two-block Gibbs sampler on the space (θ, ϕ) that alternately samples $\theta \sim \pi(\cdot \mid \phi)$ and $\phi \sim \pi(\cdot \mid \theta)$, geometric and uniform ergodicity of the joint chain follows from that of the marginal chain with transition kernel $P(\theta \mid \theta^*) = \int \pi(\theta \mid \phi)\pi(\phi \mid \theta^*)\mathrm{d}\phi$ (Roberts and Rosenthal, 2001). In establishing the uniform ergodicity under the Bayesian bridge (Theorem 3.1), we decompose the collapsed Gibbs sampler into components $\boldsymbol{\beta}$ and $(\boldsymbol{\omega}, \tau, \lambda)$ and study the marginal chain in $\boldsymbol{\beta}$. In the subsequent analysis establishing the geometric ergodicity under a more general class of regularized shrinkage priors (Theorem 3.2), we decompose the Gibbs sampler into components $(\boldsymbol{\beta}, \lambda)$ and $(\boldsymbol{\omega}, \tau)$ and study the marginal chain in $(\boldsymbol{\beta}, \lambda)$.

3.2 Behavior of shrinkage model Gibbs samplers near $\beta_j = 0$

In many models, establishing minorization and drift condition amounts to quantifying the chain's behavior in the tail of the target. In studying convergence rates under shrinkage models, however, we are faced with an additional and distinctive challenge: the need to establish that the chain does not get "stuck" near the spike at $\beta_j = 0$ (Pal and Khare, 2014;

Johndrow et al., 2018). Regularization effectively eliminates the possibility of the chain meandering to infinity, making it relatively routine to analyze its behavior as $\beta_j \to \infty$. On the other hands, the existing results provide no general insights into the behavior near $\beta_j = 0$. In fact, a careful examination of the proofs by Pal and Khare (2014) and Johndrow et al. (2018) reveals that the analyses under various shrinkage priors could have been unified if we had a more general characterization of shrinkage model Gibbs samplers' behavior near $\beta_j = 0$.

To fill in this theoretical gap, we start our analysis by abstracting key model-agnostic results from our proofs of minorization and drift condition for the sparse logistic regression Gibbs sampler. Our Propositions 3.3 and 3.4 below characterize properties of the distribution of $\lambda_j \mid \beta_j, \tau$ — this distribution, due to conditional independence, typically coincides with the full posterior conditional of λ_j and critically informs behavior of the subsequent update of β_j in a shrinkage model Gibbs sampler. Our proof techniques apply to a broad range of shrinkage priors, essentially requiring only that $\| \pi_{\text{loc}} \|_\infty := \max_\lambda \pi_{\text{loc}}(\lambda) < \infty$.[4]

Proposition 3.3 below plays a critical role in our proof of minorization condition. The proposition tells us that a sample from $\lambda_j \mid \beta_j^*, \tau$ has a uniformly lower-bounded probability of $\lambda_j \geq a$ as long as $|\beta_j^*/\tau|$ is bounded away from zero. In turn, the subsequent update of β_j conditional on λ_j should also have a guaranteed chance of landing away from zero. Intuitively, we can thus interpret the proposition as suggesting that a shrinkage model Gibbs sampler should not get "absorbed" to the spike at $\beta_j = 0$. The difference in the limiting behavior as $|\beta_j^*/\tau| \to 0$, depending on whether $\int \lambda^{-1} \pi_{\text{loc}}(\lambda) \mathrm{d}\lambda < \infty$, is also significant and leads to the difference between geometric and uniform convergence under the sparse logistic regression example through Theorem 3.6.

Proposition 3.3. *For any $a > 0$, the tail probability $\mathbb{P}(\lambda_j \geq a \mid \beta_j^*, \tau)$ is a decreasing function of $|\beta_j^*/\tau|$. If $\int \lambda^{-1} \pi_{\text{loc}}(\lambda) \mathrm{d}\lambda = \infty$, then as $|\beta_j^*/\tau| \to 0$ the tail probability converges to 0, i.e. the conditional $\lambda_j \mid \beta_j^*, \tau$ converges in distribution to a delta measure at 0. If $\int \lambda^{-1} \pi_{\text{loc}}(\lambda) \mathrm{d}\lambda < \infty$, then the conditional $\lambda_j \mid \beta_j^*, \tau$ converges in distribution to $\pi(\lambda_j) \propto \lambda_j^{-1} \pi_{\text{loc}}(\lambda_j)$ as $|\beta_j^*/\tau| \to 0$.*

Another key property of $\lambda_j \mid \beta_j, \tau$, featured prominently in our proof of the drift condition (Theorem 3.8), is provided by Proposition 3.4 below. To briefly provide a context, a Lyapunov function of the form $V(\boldsymbol{\beta}) = \sum_j |\beta_j|^{-\alpha}$ has proven effective in analyzing a shrinkage model Gibbs sampler (Pal and Khare 2014, Johndrow et al. 2018, Section 3.4). And bounding the conditional expectation of $\tau^{-\alpha} \lambda_j^{-\alpha}$ as below often constitutes a critical step in establishing the drift condition.

Proposition 3.4. *Let $R > 0$ and $\alpha \in [0,1)$. If $\| \pi_{\text{loc}} \|_\infty < \infty$, then there is an increasing function $\gamma(r) > 0$ with $\lim_{r \to 0} \gamma(r) = 0$, for which the expectation with respect to $\lambda_j \mid \beta_j^*, \tau$ satisfies*

[4]The results presented in this article, specifically those that depend on Proposition B.2 and Lemma B.3, implicitly assume that $\pi_{\text{loc}}(\lambda)$ is absolutely continuous at $\lambda_{\min} = \inf\{\lambda : \pi_{\text{loc}}(\lambda) > 0\}$. This is a purely technical assumption as any shrinkage prior in practice should satisfy $\pi_{\text{loc}}(\lambda) > 0$ for $\lambda > 0$ and be a differentiable function of λ.

$$\mathbb{E}\left[\tau^{-\alpha}\lambda_j^{-\alpha} \mid \tau, \beta_j^*\right] \leq \gamma(R/\tau)\left(\left|\beta_j^*\right|^{-\alpha} + \left|R\right|^{-\alpha}\right). \tag{3.3}$$

Proposition 3.3 and 3.4 are substantial theoretical contributions on their own, but we defer their proofs to Appendix B so that we can without interruption proceed to establish ergodicity results in the regularized sparse logisitic regression case.

Remark. The assumption $\| \pi_{\text{loc}} \|_\infty < \infty$ is sufficient but not necessary one for the conclusion of Proposition 3.4 and later of Theorem 3.8. Following the analysis by Pal and Khare (2014), we can show that the conclusions also hold under normal-gamma priors with any shape parameter $a > 0$. These priors have the property $\pi_{\text{loc}}(\lambda) \sim O\left(\lambda^{2a-1}\right)$ as $\lambda \to 0$ and hence $\lim_{\lambda \to 0} \pi(\lambda) = \infty$ for $a < 1/2$. We leave it as future work to characterize the behavior of general shrinkage priors with $\| \pi_{\text{loc}} \|_\infty = \infty$.

Remark. In Appendix A, we show that Proposition 3.3 and 3.4 can also be applied to establish uniform/geometric ergodicity of a Gibbs sampler for Bayesian sparse probit regression, demonstrating their relevance beyond the sparse logistic regression example.

3.3 Minorization — with uniform ergodicity in special cases

Having described the noteworthy model-agnostic results within our proofs, from now on we focus exclusively on the regularized sparse logistic regression case. We first consider the Gibbs sampler with fixed τ in Lemma 3.5 and Theorem 3.6. While fixing the global scale parameter is a common assumption in the ergodicity proofs for shrinkage models (Pal and Khare, 2014), we subsequently show that this assumption can be replaced with much weaker ones; we only require $\tau \sim \pi_{\text{glo}}(\cdot)$ to be supported away from 0 in Theorem 3.1 and additionally away from $+\infty$ in Theorem 3.7.

Let $P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau, \lambda)$ denote the transition kernel corresponding to Step 3 and 4 of the Gibbs sampler as described in Page 6 and $P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau)$ corresponding to Step 2 – 4. In other words, we define

$$P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau, \lambda) = \int \pi(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}) \pi(\boldsymbol{\omega} \mid \boldsymbol{\beta}^*, \boldsymbol{X}) \mathrm{d}\boldsymbol{\omega},$$

$$P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau) = \int P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau, \lambda) \pi(\lambda \mid \boldsymbol{\beta}^*) \mathrm{d}\lambda\,.$$

The following lemma builds on a result of Choi and Hobert (2013) and plays a prominent role, along with Proposition 3.3, in our proofs of minorization conditions.

Lemma 3.5. *Whenever* $\min_j \tau\lambda_j \geq R > 0$, *there is* $\delta' > 0$ — *independent of* τ *and* λ except through R — *such that the following minorization condition holds*:

$$P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau, \lambda) \geq \delta' \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}_R, \boldsymbol{\Phi}_R^{-1}\right),$$

where $\boldsymbol{\Phi}_R = \frac{1}{2}\boldsymbol{X}^\top \boldsymbol{X} + \zeta^{-2}\boldsymbol{I} + R^{-2}\boldsymbol{I}$ *and* $\boldsymbol{\mu}_R = \boldsymbol{\Phi}_R^{-1}\boldsymbol{X}^\top\left(\boldsymbol{y} - \mathbf{1}/2\right)$.

We defer the proof to Appendix C.

We now establish a minorization condition for the Gibbs sampler with fixed τ.

Theorem 3.6 (Minorization). *Let* $\epsilon, R > 0$. *On a small set* $\left\{\boldsymbol{\beta}^* : \min_j \left|\beta_j^*/\tau\right| \geq \epsilon\right\}$, *the marginal transition kernel satisfies a minorization condition*

$$P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau) \geq \delta(\tau) \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}_R, \boldsymbol{\Phi}_R^{-1}\right),$$

where $\boldsymbol{\mu}_R$ *and* $\boldsymbol{\Phi}_R$ *are defined as in Lemma 3.5, and* $\delta(\tau) > 0$ *is increasing in* τ *and otherwise depends only on* ϵ, R, *and* π_{loc}. *Moreover, the minorization holds uniformly on* $\boldsymbol{\beta}^* \in \mathbb{R}^p$ *in case the prior satisfies* $\int_0^\infty \lambda^{-1} \pi_{\text{loc}}\left(\lambda\right) \mathrm{d}\lambda < \infty$.

Proof. Using Lemma 3.5, we have

$$\begin{aligned} P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau) &= \int P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau, \lambda) \pi(\lambda \mid \boldsymbol{\beta}^*, \tau) \mathrm{d}\lambda \\ &\geq \int_{\{\min_j \tau\lambda_j \geq R\}} P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau, \lambda) \pi(\lambda \mid \boldsymbol{\beta}^*, \tau) \mathrm{d}\lambda \\ &\geq \delta' \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}_R, \boldsymbol{\Phi}_R^{-1}\right) \prod_j \int_{R/\tau}^\infty \pi\left(\lambda_j \mid \beta_j^*, \tau\right) \mathrm{d}\lambda_j, \end{aligned}$$

for $\delta' > 0$ depending only on R. Also, Proposition 3.3 implies that whenever $\left|\beta_j^*/\tau\right| \geq \epsilon$

$$\int_{R/\tau}^\infty \pi\left(\lambda_j \mid \beta_j^*, \tau\right) \mathrm{d}\lambda_j \geq \int_{\frac{R}{\tau}}^\infty \pi(\lambda \,|\, |\beta^*/\tau| = \epsilon) \mathrm{d}\lambda > 0.$$

Hence, $\prod_j \int_{R/\tau}^\infty \pi\left(\lambda_j \mid \beta_j^*, \tau\right) \mathrm{d}\lambda_j$ is lower bounded by a positive constant depending only on ϵ and R/τ. In case $C = \int_0^\infty \lambda^{-1} \pi_{\text{loc}}\left(\lambda\right) \mathrm{d}\lambda < \infty$, we can forgo the assumption $\left|\beta_j^*/\tau\right| \geq \epsilon$ and obtain a uniform lower bound since

$$\int_R^\infty \pi\left(\lambda_j \mid \beta_j^*, \tau\right) \mathrm{d}\lambda_j \geq \frac{1}{C} \int_R^\infty \lambda^{-1} \pi_{\text{loc}}\left(\lambda\right) \mathrm{d}\lambda > 0.$$

We now relax the assumption of fixed τ. The results of van der Pas et al. (2017) suggest that a constraint of the form $0 < \tau_{\min} \leq \tau \leq \tau_{\max} < \infty$ can improve the statistical property of shrinkage priors. As it turns out, such a constraint also enables us to establish minorization conditions for the full Gibbs sampler under sparse logistic regression with τ update incorporated. We can in fact take $\tau_{\max} = \infty$ in case of the Bayesian bridge prior, whose unique structure allows us to marginalize out λ_j's when updating τ (Polson et al. 2014;

Appendix E). This collapsed update of τ from $\tau \mid \boldsymbol{\beta}$ makes it possible to deduce the uniform ergodicity result of Theorem 3.1 as an immediate consequence of Theorem 3.6 by studying the marginal transition $\boldsymbol{\beta}^* \to \boldsymbol{\beta}$ with kernel

$$P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*) = \int_{\tau_{\min}}^{\infty} P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau)\pi(\tau \mid \boldsymbol{\beta}^*)\mathrm{d}\tau\,. \tag{3.4}$$

Proof of Theorem 3.1. It suffices to establish uniform minorization for the marginal transition kernel (3.4). Under the Bayesian bridge prior, we have $\pi_{\mathrm{loc}}(\lambda) \propto O(\lambda^{2a})$ as $\lambda \to 0$ (Appendix E) and hence $\int \lambda^{-1}\pi_{\mathrm{loc}}(\lambda) < \infty$. The minorization condition of Theorem 3.6 thus holds uniformly in $\boldsymbol{\beta}^*$, yielding

$$\int_{\tau_{\min}}^{\infty} P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau)\pi(\tau \mid \boldsymbol{\beta}^*)\mathrm{d}\tau \geq \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}_R, \boldsymbol{\Phi}_R^{-1}\right) \int_{\tau_{\min}}^{\infty} \delta(\tau)\pi(\tau \mid \boldsymbol{\beta}^*)\mathrm{d}\tau, \tag{3.5}$$

for $R > 0$. Theorem 3.6 further tells us that $\delta(\tau) > 0$ is increasing in τ, so we have

$$\int_{\tau_{\min}}^{\infty} \delta(\tau)\pi(\tau \mid \boldsymbol{\beta}^*)\mathrm{d}\tau \geq \delta(\tau_{\min}) > 0. \tag{3.6}$$

The inequalities (3.5) and (3.6) together establish uniform minorization. □

For more general shrinkage priors, the global scale τ must be updated from the full conditional $\tau \mid \boldsymbol{\beta}, \lambda$. This makes it necessary to study the marginal transition $(\boldsymbol{\beta}^*, \lambda^*) \to (\boldsymbol{\beta}, \lambda)$, jointly in regression coefficients and local scales, with kernel

$$P(\boldsymbol{\beta}, \lambda \mid \boldsymbol{\beta}^*, \lambda^*) = \int_{\tau_{\min}}^{\tau_{\max}} P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau, \lambda) \prod_j \pi(\lambda_j \mid \beta_j^*, \tau)\pi(\tau \mid \boldsymbol{\beta}^*, \lambda^*)\mathrm{d}\tau\,. \tag{3.7}$$

We establish a minorization condition for this general case in Theorem 3.7.

Theorem 3.7. *If the prior* $\pi_{\mathrm{glo}}(\cdot)$ *is supported on* $[\tau_{\min}, \tau_{\max}]$ *for* $0 < \tau_{\min} \leq \tau_{\max} < \infty$, *then the marginal transition kernel* $P(\boldsymbol{\beta}, \lambda \mid \boldsymbol{\beta}^*, \lambda^*)$ *of the Pólya-Gamma Gibbs sampler for regularized sparse logistic regression satisfies a minorization condition on a small set* $\{(\boldsymbol{\beta}^*, \lambda^*): 0 < \epsilon \leq |\beta_j^*| \leq E < \infty \ \text{for all } j\}$.

Proof. By Lemma 3.5 and the fact $\tau\lambda_j \geq \tau_{\min}\lambda_j$, we know that for $R > 0$

$$P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau, \lambda) \geq \mathbb{1}\{\min_j \tau_{\min}\lambda_j \geq R\}\delta'\mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}_R, \boldsymbol{\Phi}_R^{-1}\right). \tag{3.8}$$

To lower bound the term $\prod_j \pi(\lambda_j \mid \beta_j^*, \tau)$ in (3.7), we first recall that

$$\pi(\lambda_j \mid \beta_j^*, \tau) = \frac{\lambda_j^{-1}\exp\left(-\beta_j^{*2}/2\tau^2\lambda_j^2\right)\pi_{\mathrm{loc}}(\lambda_j)}{\int_0^\infty \lambda^{-1}\exp\left(-\beta_j^{*2}/2\tau^2\lambda^2\right)\pi_{\mathrm{loc}}(\lambda)\mathrm{d}\lambda}.$$

When $\tau_{\min} \le \tau \le \tau_{\max}$ and $\epsilon \le |\beta_j^*| \le E$, we have

$$\exp\left(-E^2/2\tau_{\min}^2\lambda^2\right) \le \exp\left(-\beta_j^2/2\tau^2\lambda^2\right) \le \exp\left(-\epsilon^2/2\tau_{\max}^2\lambda^2\right).$$

It follows from the above inequalities that

$$\pi(\lambda_j \mid \beta_j^*, \tau) \ge \frac{\lambda_j^{-1}\exp\left(-E^2/2\tau_{\min}^2\lambda_j^2\right)\pi_{\mathrm{loc}}(\lambda_j)}{\int_0^\infty \lambda^{-1}\exp\left(-\epsilon^2/2\tau_{\max}^2\lambda^2\right)\pi_{\mathrm{loc}}(\lambda)\mathrm{d}\lambda} := \eta\pi_{\mathrm{lower}}(\lambda_j) \tag{3.9}$$

for $\eta > 0$ and density $\pi_{\mathrm{lower}}(\cdot)$ independent of β_j^* and τ. Combining (3.8) and (3.9), we can lower bound the transition kernel (3.7) as

$$\begin{aligned} &P(\boldsymbol{\beta}, \lambda \mid \boldsymbol{\beta}^*, \lambda^*) \\ &\ge \delta'\eta\mathbb{1}\left\{\min_j \lambda_j \ge \frac{R}{\tau_{\min}}\right\}\mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}_R, \boldsymbol{\Phi}_R^{-1}\right)\prod_j \pi_{\mathrm{lower}}(\lambda_j)\int_{\tau_{\min}}^{\tau_{\max}} \pi(\tau \mid \boldsymbol{\beta}^*, \lambda)\mathrm{d}\tau \\ &= \delta'\eta\mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}_R, \boldsymbol{\Phi}_R^{-1}\right)\prod_j \mathbb{1}\left\{\lambda_j \ge \frac{R}{\tau_{\min}}\right\}\pi_{\mathrm{lower}}(\lambda_j). \end{aligned}$$

□

3.4 Drift condition and geometric ergodicity

Here we establish a drift condition for geometric ergodicity under sparse logistic regression. As discussed in Section 3.2, the regularization prevents the Markov chain from meandering to infinity, so the main question is whether the chain can get "stuck" for a long time near $\beta_j^* = 0$. The following result shows that this does not happen as long as the global scale τ is bounded away from zero.

Theorem 3.8. *Suppose that the local scale prior satisfies* $\|\pi_{\mathrm{loc}}\|_\infty < \infty$ *and that the global scale prior* $\pi_{\mathrm{glo}}(\cdot)$ *is supported on* $[\tau_{\min}, \infty)$ *for* $\tau_{\min} > 0$. *Then the marginal transition kernel* $P(\boldsymbol{\beta}, \lambda \mid \boldsymbol{\beta}^*, \lambda^*)$ *satisfies a drift condition with a Lyapunov function* $V(\boldsymbol{\beta}) = \sum_j |\beta_j|^{-\alpha}$ *for any* $0 \le \alpha < 1$.

Proof. Note that $PV(\boldsymbol{\beta}^*)$ can be expressed as a series of iterated expectations with respect to (1) $\boldsymbol{\beta} \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = 0$, (2) $\boldsymbol{\omega} \mid \boldsymbol{\beta}^*$, (3) $\lambda \mid \boldsymbol{\beta}^*, \tau$, and (4) $\tau \mid \boldsymbol{\beta}^*, \lambda^*$. We will bound the iterated expectations of $|\beta_j|^{-\alpha}$ one by one.

Since $\boldsymbol{\beta} \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \mathbf{0}$ is distributed as Gaussian, denoting by μ_j and σ_j^2 the conditional mean and variance of β_j, Proposition 3.10 below tells us that

$$\mathbb{E}\left[\left|\beta_j\right|^{-\alpha} \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \mathbf{0}\right] \leq C_\alpha(\mu_j/\sigma_j)\sigma_j^{-\alpha} \text{ where } \sup_t C_\alpha(t) \leq \frac{\Gamma\left(\frac{1-\alpha}{2}\right)}{2^{\alpha/2}\sqrt{\pi}} \text{ and } C_\alpha(t) = O\left(|t|^{-\alpha}\right) \text{ as } |t| \to \infty .$$

For the purpose of this proof, we can simply set C_α to be its global upper bound; however, a tighter bound may be obtained when the posterior concentrates away from zero and thereby resulting in $|\mu_j/\sigma_j| \to \infty$ and $C_\alpha(\mu_j/\sigma_j) \to 0$ as the sample size increases. Combined with Proposition 3.11 below, the above inequality implies

$$\frac{1}{C_\alpha}\mathbb{E}\left[\left|\beta_j\right|^{-\alpha} \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \mathbf{0}\right] \leq \tau^{-\alpha}\lambda_j^{-\alpha} + \zeta^{-\alpha} + 1 - \frac{\alpha}{2} + \frac{\alpha}{2}\sum_{i=1}^{n}\omega_i x_{ij}^2 . \tag{3.10}$$

In taking the expectation of (3.10) with respect to $\boldsymbol{\omega} \mid \boldsymbol{\beta}^*$, we use the result $\mathbb{E}[\omega_j \mid \boldsymbol{\beta}^*] \leq 1/4$ of Wang and Roy (2018) to obtain

$$\frac{1}{C_\alpha}\mathbb{E}\left[\left|\beta_j\right|^{-\alpha} \mid \tau, \lambda\right] \leq \tau^{-\alpha}\lambda_j^{-\alpha} + \zeta^{-\alpha} + 1 - \frac{\alpha}{2} + \frac{\alpha}{8}\sum_{i=1}^{n} x_{ij}^2 . \tag{3.11}$$

Taking the expectation of (3.11) with respect to $\lambda \mid \tau, \boldsymbol{\beta}^*$, we have

$$\frac{1}{C_\alpha}\mathbb{E}\left[\left|\beta_j\right|^{-\alpha} \mid \tau, \boldsymbol{\beta}^*\right] \leq \mathbb{E}\left[\tau^{-\alpha}\lambda_j^{-\alpha} \mid \tau, \beta_j^*\right] + C'\left(\alpha, \boldsymbol{X}\right) \text{ where } C'(\alpha, \boldsymbol{X}) = \zeta^{-\alpha} + 1 - \frac{\alpha}{2} + \frac{\alpha}{8}\sum_{i=1}^{n} x_{ij}^2 . \tag{3.12}$$

Now choose $R > 0$ small enough that $\gamma(R/\tau) \leq \gamma(R/\tau_{\min}) < C_\alpha^{-1}$ in Proposition 3.4. Then we have the following inequality for $\gamma' \coloneqq C_\alpha\gamma(R/\tau_{\min}) < 1$:

$$C_\alpha\mathbb{E}\left[\tau^{-\alpha}\lambda_j^{-\alpha} \mid \tau, \beta_j^*\right] \leq \gamma'\left(\left|\beta_j^*\right|^{-\alpha} + \left|R\right|^{-\alpha}\right)$$

for all $\tau \geq \tau_{\min}$. Incorporating the above inequality into (3.12), we obtain

$$\mathbb{E}\left[\left|\beta_j\right|^{-\alpha} \mid \tau, \boldsymbol{\beta}^*\right] \leq \gamma'\left|\beta_j^*\right|^{-\alpha} + \gamma'\left|R\right|^{-\alpha} + C_\alpha C'(\alpha, \boldsymbol{X}) .$$

Since $\pi(\tau \mid \boldsymbol{\beta}^*, \lambda^*)$ is supported on $\tau \geq \tau_{\min}$ by our assumption, taking the expectation with respect to $\tau \mid \boldsymbol{\beta}^*, \lambda^*$ yield

$$\mathbb{E}\left[|\beta_j|^{-\alpha} \mid \boldsymbol{\beta}^*, \lambda^*\right] \leq \gamma' \left|\beta_j^*\right|^{-\alpha} + \gamma' \left|R\right|^{-\alpha} + C_\alpha C'(\alpha, \boldsymbol{X}).$$

Theorem 3.7 and 3.8 together imply the geometric ergodicity result of Theorem 3.2:

Proof of Theorem 3.2. We show that $V(\boldsymbol{\beta}) = \sum_j |\beta_j|^{-\alpha} + \| \boldsymbol{\beta} \|^2$ is a Lyapunov function for the marginal transition kernel $P(\boldsymbol{\beta}, \lambda \mid \boldsymbol{\beta}^*, \lambda^*)$. Note that

$$\begin{aligned}
&\mathbb{E}\left[\| \boldsymbol{\beta} \|^2 \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \mathbf{0} \right] \\
&= \| \mathbb{E}[\boldsymbol{\beta} \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \mathbf{0}] \|^2 + \sum_j \mathrm{var}(\beta_j^2 \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \mathbf{0}) \\
&= \| \boldsymbol{\Sigma} \boldsymbol{X}^\top \left(\boldsymbol{y} - \frac{1}{2} \right) \|^2 + \sum_j \boldsymbol{e}_j^\top \boldsymbol{\Sigma} \boldsymbol{e}_j
\end{aligned}$$

for $\boldsymbol{\Sigma} = \left(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} + \zeta^{-2} \boldsymbol{I} + \tau^{-2} \boldsymbol{\Lambda}^{-2} \right)^{-1}$. Since $\boldsymbol{\Sigma} \prec \zeta^2 \boldsymbol{I}$, we have $\boldsymbol{e}_j^\top \boldsymbol{\Sigma} \boldsymbol{e}_j \leq \zeta^2$ and $\| \boldsymbol{\Sigma} \boldsymbol{X}^\top \left(\boldsymbol{y} - \frac{1}{2} \right) \|^2 \leq \zeta^2 \| \boldsymbol{X}^\top \left(\boldsymbol{y} - \frac{1}{2} \right) \|^2$. Thus we have

$$\mathbb{E}\left[\| \boldsymbol{\beta} \|^2 \mid \boldsymbol{\omega}, \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \mathbf{0} \right] \leq \zeta^2 \| \boldsymbol{\Sigma} \boldsymbol{X}^\top \left(\boldsymbol{y} - \frac{1}{2} \right) \|^2 + n\zeta^2. \tag{3.13}$$

Since the right-hand side does not depend on $\boldsymbol{\omega}, \tau, \lambda$, the expectation with respect to $P(\boldsymbol{\beta}, \lambda \mid \boldsymbol{\beta}^*, \lambda^*)$ satisfies the same bound:

$$\mathbb{E}\left[\| \boldsymbol{\beta} \|^2 \mid \boldsymbol{\beta}^*, \lambda^* \right] \leq \zeta^2 \| \boldsymbol{\Sigma} \boldsymbol{X}^\top \left(\boldsymbol{y} - \frac{1}{2} \right) \|^2 + n\zeta^2.$$

In addition to the above bound, we know that $\sum_j |\beta_j|^{-\alpha}$ is a Lypunov function by Theorem 3.8. Hence, $V(\boldsymbol{\beta}) = \sum_j |\beta_j|^{-\alpha} + \| \boldsymbol{\beta} \|^2$ is again a Lyapunov function. Moreover, by Theorem 3.7, we know that the Gibbs sampler satisfies a minorization condition on the set $\{\boldsymbol{\beta}^* : 0 < \epsilon \leq |\beta_j^*| \leq E < \infty \text{ for all } j\}$ for $\epsilon > 0$ and $E < \infty$. Thus the sampler is geometrically ergodic. □

Remark 3.9. As mentioned earlier, the geometric and uniform ergodicity as well as analogues of the intermediate results continue to hold when we relax the assumption of fixed ζ to a prior constraint of the form $\zeta \leq \zeta_{\max} < \infty$. The proof goes as follows. Due to the conditional independence, the Gibbs sampler on the joint space draws alternately from $\zeta \mid \boldsymbol{\beta}, \boldsymbol{z} = 0$ and $\boldsymbol{\beta}, \boldsymbol{\omega}, \tau, \lambda \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = 0, \zeta$. By repeating all the previous arguments with $\zeta_{\max}$ in place of ζ, we obtain essentially the identical minorization and drift bounds that hold for all $\zeta \leq \zeta_{\max}$. Since the bounds hold uniformly on the support $\zeta \leq \zeta_{\max}$, the identical bounds again hold when taking the expectation over $\zeta \mid \boldsymbol{\beta}, \boldsymbol{z} = \mathbf{0}$.

Auxiliary results for proof of geometric ergodicity—Proposition 3.10 and 3.11 below are used in the proof of Theorem 3.8 and are proved in Appendix D. Proposition 3.10

is a refinement of Proposition A1 in Pal and Khare (2014) and of Equation (41) in Johndrow et al. (2018), neither of which have the $D(\mu/\sigma)$ term.

Proposition 3.10. *For* $\alpha \in (0,1)$ *and* $\beta \sim \mathcal{N}(\mu, \sigma^2)$, *we have*

$$\mathbb{E}\left|\beta\right|^{-\alpha} \leq \frac{\Gamma\left(\frac{1-\alpha}{2}\right)}{2^{\alpha/2}\sqrt{\pi}}\sigma^{-\alpha}\min\left\{1, D\left(\mu/\sigma\right)\right\},$$

where $D(t) = O\left(|t|^{-\alpha}\right) \to 0$ *as* $|t| \to \infty$ *and can be chosen as*

$$D(t) = \frac{1}{B\left(\frac{\alpha}{2}, \frac{1-\alpha}{2}\right)}\left[\frac{2^{\frac{5}{2}-\alpha}}{1-\alpha}\exp\left(-\frac{t^2}{4}\right) + 2^{\frac{1}{2}+\alpha}\Gamma\left(\frac{\alpha}{2}\right)|t|^{-\alpha}\right]. \tag{3.14}$$

Proposition 3.11. *The diagonals* σ_j of $\boldsymbol{\Sigma} = \left(\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} + \zeta^{-2}\boldsymbol{I} + \tau^{-2}\boldsymbol{\Lambda}^{-2}\right)^{-1}$ *satisfy the following inequality for* $0 \leq \alpha < 1$:

$$\sigma_j^{-\alpha} \leq \tau^{-\alpha}\lambda_j^{-\alpha} + \zeta^{-\alpha} + 1 - \frac{\alpha}{2} + \frac{\alpha}{2}\sum_{i=1}^{n}\omega_i x_{ij}^2.$$

4 Simulation

We run a simulation study to assess the computational and statistical properties of the regularized sparse logistic regression model. We use the Bayesian bridge prior $\pi(\beta_j \mid \tau) \propto \tau^{-1}\exp\left(-|\beta_j/\tau|^a\right)$ to take advantage of the efficient global scale parameter update scheme. This prior also allows us to experiment with a range of spike and tail behavior by varying the exponent a, inducing larger spikes and heavier tails as $a \to 0$. For the global scale parameter, we chose the objective prior $\pi_{\text{glo}}(\tau) \propto \tau^{-1}$ (Berger et al., 2015, Appendix E) with the range restriction $10^{-6} \leq \mathbb{E}[|\beta_j| \mid \tau] \leq 1$ to ensure posterior propriety, though in practice we never observe a posterior draw of τ outside this range. For the posterior computations, we use the Pólya-Gamma Gibbs sampler provided by the *bayesbridge* package available from Python Package Index (pypi.org); the source code is available at the GitHub repository https://github.com/OHDSI/bayes-bridge.

4.1 Data generating process: "large n, but weak signal" problems

Piironen and Vehtari (2017) demonstrate the benefits of regularizing shrinkage priors in the "$p > n$" case, when the number of predictors p exceeds the sample size n. To complement their study, we consider the case of rare outcomes and infrequently observed features, another common situation in which regularizing shrinkage priors becomes essential. For example in healthcare data, many outcomes of interests have low incidence rates and many treatments are prescribed to only a small fraction of patients (Tian et al., 2018). This

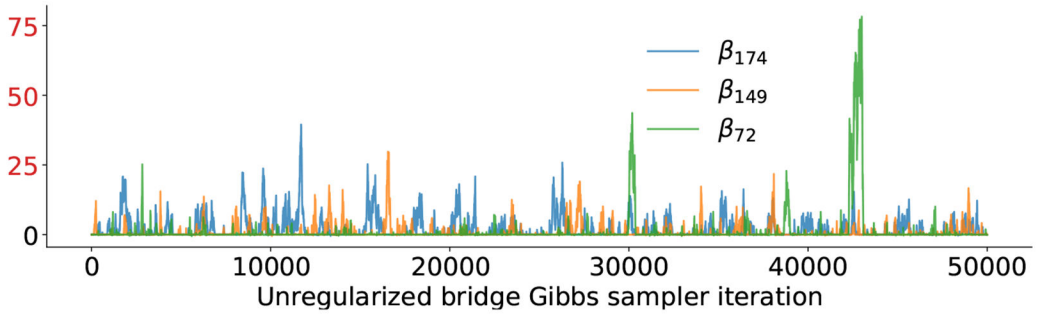
results in binary outcomes $\boldsymbol{y}$ and features $\boldsymbol{x}_j$ filled mostly with 0's, making the amount of information much less than otherwise expected (Greenland et al., 2016).

To simulate under these "large n, but weak signal" settings, we generate synthetic data with $n = 2{,}500$ and $p = 500$ as follows. We construct binary features with a range of observed frequencies by first drawing $2w_j \sim \text{Beta}(1/2, 2)$ for $j = 1, \ldots, 500$; this in particular means $0 \leq w_j \leq 0.5$ and $\mathbb{E}[w_j] = 0.1$. For each j, we then generate $x_{ij} \sim \text{Bernoulli}(w_j)$ for $i = 1, \ldots, n$. We choose the true signal to be $\beta_j = 1$ for $j = 1, \ldots, 10$ and $\beta_j = 0$ for $j = 11, \ldots, 500$. To simulate an outcome with low incidence rate, we choose the intercept to be $\beta_0 = 1.5$ and draw $y_i \sim \text{Bernoulli}\left(\text{logit}\left(-\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right)$, resulting in $y_i = 1$ for approximately 5% of its entries.

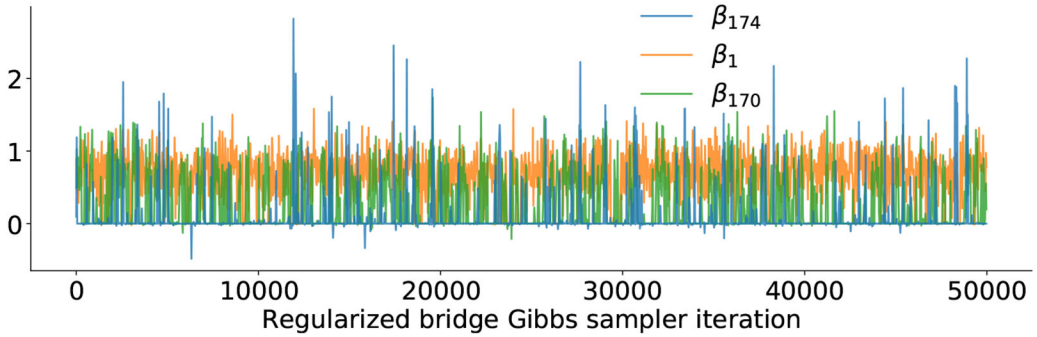
4.2 Convergence and mixing: with and without regularization

With the above data generating process, outcome $\boldsymbol{y}$ and design matrix $\boldsymbol{X}$ barely contain enough information to estimate all the coefficients β_j's. In particular, sparse logistic model without regularization can lead to a heavy-tailed posterior, for which uniform and geometric ergodicity of the Pólya-Gamma Gibbs sampler becomes questionable.

These potential convergence and mixing issues are evidenced by the traceplot (Figure 4.1a) of the posterior samples based on bridge exponent $a = 1/16$. As we are particularly concerned with the Markov chain wandering off to the tail of the target, we examine the estimated credible intervals to identify the coefficients with potential convergence and mixing issues. Plotted in Figure 4.1 are the coefficients with the widest 95% credible intervals; these coefficients also have some of the smallest estimated effective sample sizes, though the accuracy of such estimates is not guaranteed without geometric ergodicity. When regularizing the shrinkage prior with a slab width $\zeta = 1$, the posterior samples indicate no such convergence or mixing issues (Figure 4.1b) and yield more sensible posterior credible intervals (Figure 4.2).



(a) Without regularization, the Markov chain takes multiple "excursions" — each lasting over hundreds of iterations — into the unreasonable value range of the coefficients. The deviation in β_{172} is particularly prominent around the 42,000th iteration. More severe deviations may occur if the chain is run longer.



(b) With regularization, the Markov chain does not display any serious mixing issues. The noticeable auto-correlation is due to the multi-modality of the posterior, an unavoidable feature of shrinkage models. Note that the coefficients with widest credible intervals do not coincide with the unregularized setting.

Figure 4.1:
Traceplot under the Bayesian bridge logistic regression with exponent 1/16. Shown are the three coefficients with most potentially problematic mixing behaviors; see the main text for the details on our criteria.

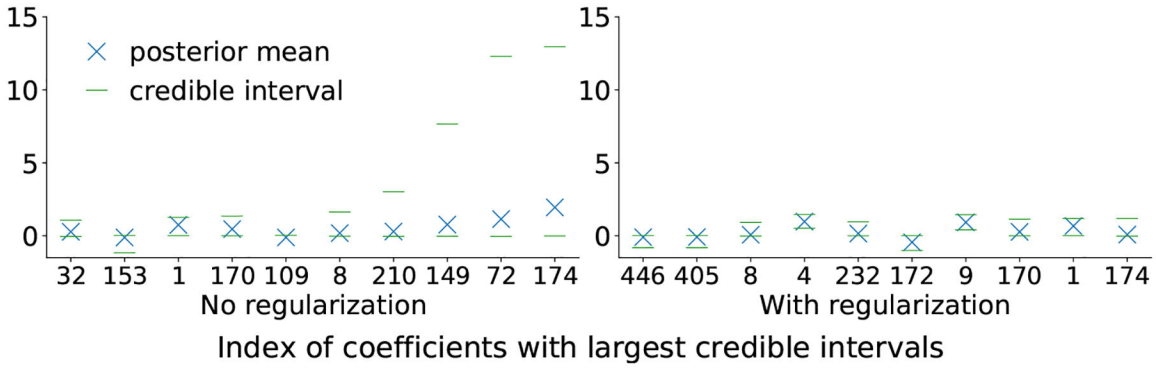


Figure 4.2:

Ten widest 95% posterior credible intervals under the Bayesian bridge logistic regression with (right) and without (left) regularization. Without regularization, the intervals are unrealistically large compared to the signal size of $\beta_j = 1$ for $j = 1, \ldots, 10$.

We emphasize that there is no fundamental change in the Gibbs sampler itself when incorporating the regularization, the only change being the addition of the $\zeta^{-2}\boldsymbol{I}$ term in the conditional precision matrix (3.2) when updating $\boldsymbol{\beta}$. It is the change in the posterior - more specifically the guaranteed light tails of the $\boldsymbol{\beta}$ marginal — that induces faster convergence and mixing.

We also assess sensitivity of convergence and mixing rates on the slab width ζ. The regularized prior recovers the unregularized one as $\zeta \to \infty$. This means that, as seen from the problematic computational behavior of the unregularized model, ζ cannot be taken too large in this limited data setting. In other words, the choice of ζ has to reflect some degree of prior information on β_j's. We need not assume strong prior information, however; Figure 4.3 demonstrates that even small amount of regularization (e.g. $\zeta = 2$ or 4) can noticeably improve the computational behavior over the unregularized case.

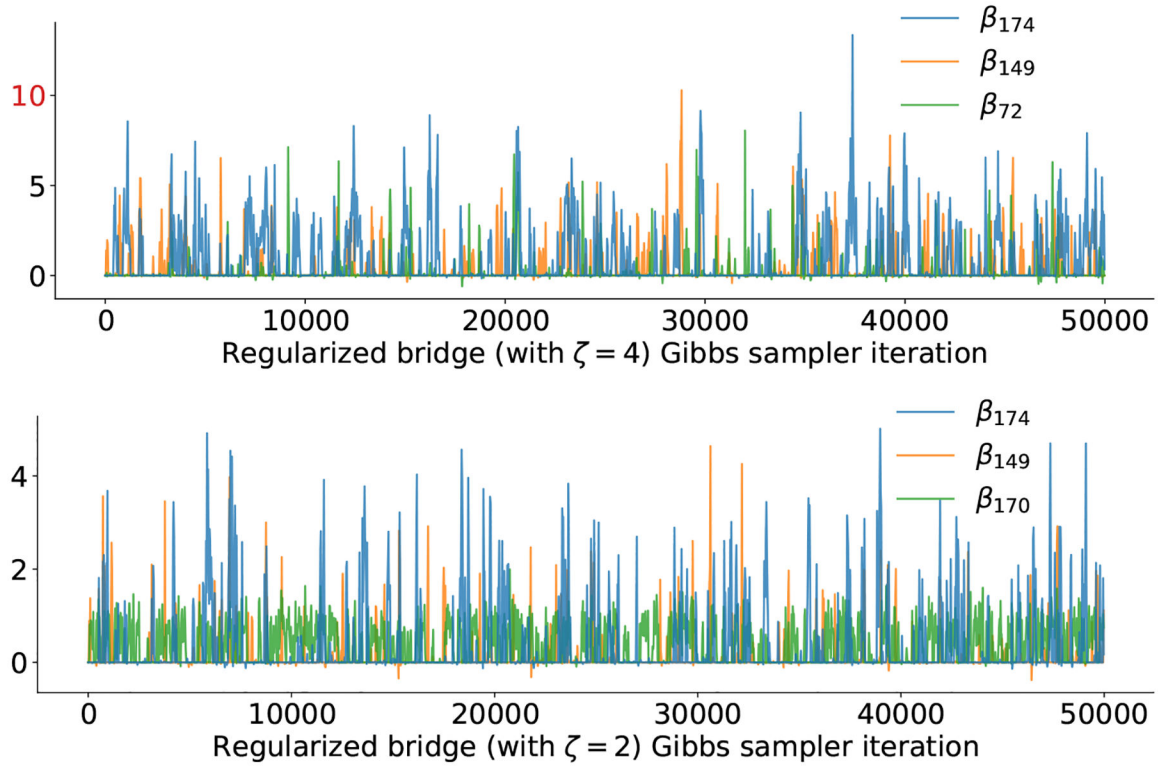


Figure 4.3:
Traceplots under different slab widths: $\zeta = 2$ (bottom) and $\zeta = 4$ (top). The settings are otherwise identical to those of Figure 4.1. As before, the three coefficients with most problematic mixing behaviors do not always coincide across different slab widths.

4.3 Statistical properties of shrinkage model for weak signals

To study the shrinkage model's ability to separate out the non-zero β_j from the $\beta_j = 0$, we simulate 10 replicate data sets and estimate the posterior for each of them. In total, we obtain 5,000 marginal posterior distributions — 10 independent replication for each of the $p = 500$ regression coefficients — with 100 for the signal $\beta_j = 1$ and 4,900 for the non-signal $\beta_j = 0$. As all the predictors $\boldsymbol{x}_j$'s are simulated in an exchangeable manner, the 100 (and 4,900) posterior marginals for the signal (and non-signal) are also exchangeable.

Figure 4.4 show the posterior credible intervals. Due to the low incidence rate and infrequent binary features, many of the signals are too weak to be detected. We also find that the credible intervals seemingly do not achieve their nominal frequentist coverage for signals below detection strength. This finding is consistent with the existing theoretical results on shrinkage priors and is unsurprising in light of the impossibility theorem by Li (1989) — confidence intervals cannot be optimally tight and have nominal coverage at the same time. Credible intervals produced by Bayesian shrinkage models tend to be optimally tight and thus require appropriate manual adjustments to achieve the nominal coverage (van der Pas et al., 2017). No statistical procedure is immune to this tightness-coverage trade-off; therefore, the apparent under-coverage should be seen not as a flaw but more as a feature of Bayesian shrinkage models.

We benchmark the signal detection capability of the posterior against the frequentist lasso, arguably the most widely-used approach to feature selection. Obtaining the lasso point estimates requires a selection of the hyper-parameter commonly referred to as the *penalty* parameter. For its choice, we first follow the standard practice of minimizing the ten-fold cross-validation errors (Hastie et al., 2009). However, this approach yields inconsistent and poor overall performance, detecting only 13 out of the 100 signals (Figure 4.4). Cross-validation likely fails here because each fold does not capture the characteristics of the whole data when the signals are so weak. As a more stable alternative for calibrating the penalty parameter, we try an empirical Bayes procedure based on the Bayesian interpretation of the lasso (Park and Casella, 2008). We first estimate the posterior marginal mean of the penalty parameter from the Bayesian lasso Gibbs sampler. Conditionally on this value, we then find the posterior mode of $\boldsymbol{\beta}$. This procedure seems to yield more consistent performance, detecting 39 out of the 100 signals albeit with the estimates more shrunk towards null than the Bayesian posterior means. The empirical Bayes procedure demonstrates more consistent behavior for the non-signals as well (Figure 4.5).

We also assess how the spike size and (pre-regularization) tail behavior of the prior influence statistical properties of the resulting posterior. For this purpose, we fit the regularized bridge model with the exponent $a^{-1} \in \{2,4,8,16\}$ to the same data sets. Figure 4.6 summarizes the credible intervals under the $a = 1/4$ case. The credible intervals are centered around the values similar to the $a = 1/16$ case (Figure 4.4), but are much wider overall. We observe the same pattern throughout the range of the exponent values: similar median values, but tighter intervals for the smaller exponents. In particular, as can be seen in Figure 4.7, more "extreme" shrinkage priors with larger spikes and heavier-tails seem to yield tighter credible intervals for the same coverage.

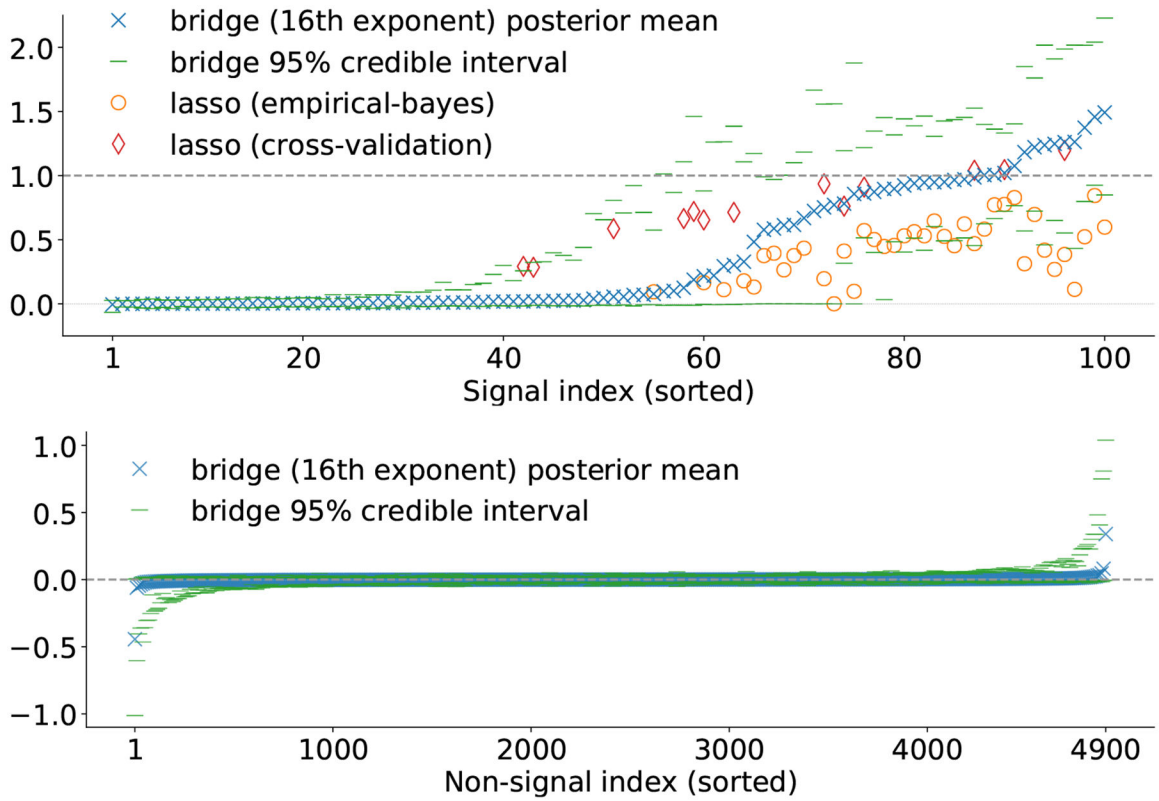


Figure 4.4:

The 95% posterior credible intervals for the signals $\beta_j = 1$ (top) and non-signals $\beta_j = 0$ (bottom) under the Bayesian bridge logistic regression with the bridge exponent 1/16. The intervals are sorted by the posterior means. To avoid clutter, the top plot shows only the non-zero values of the lasso estimates. The lasso estimates for the non-signals are summarized in Figure 4.5 and are not shown in the bottom plot.

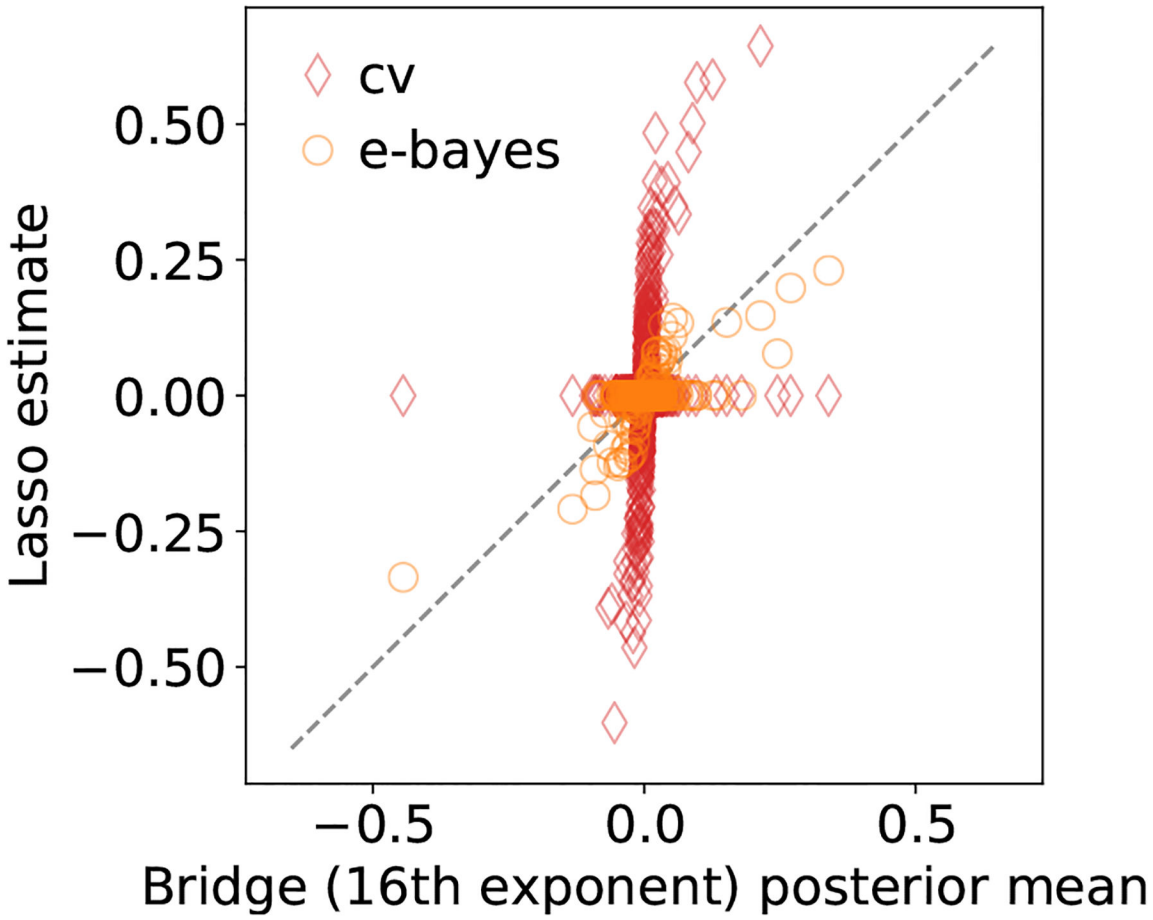


Figure 4.5:
Comparison of the 4,900 Bayesian bridge posterior means and lasso estimates for the non-signals $\beta_j = 0$. Lasso with cross-validation produces a larger number of false positives. Lasso with the empirical Bayes calibration yields the estimates more in line with the bridge posterior.

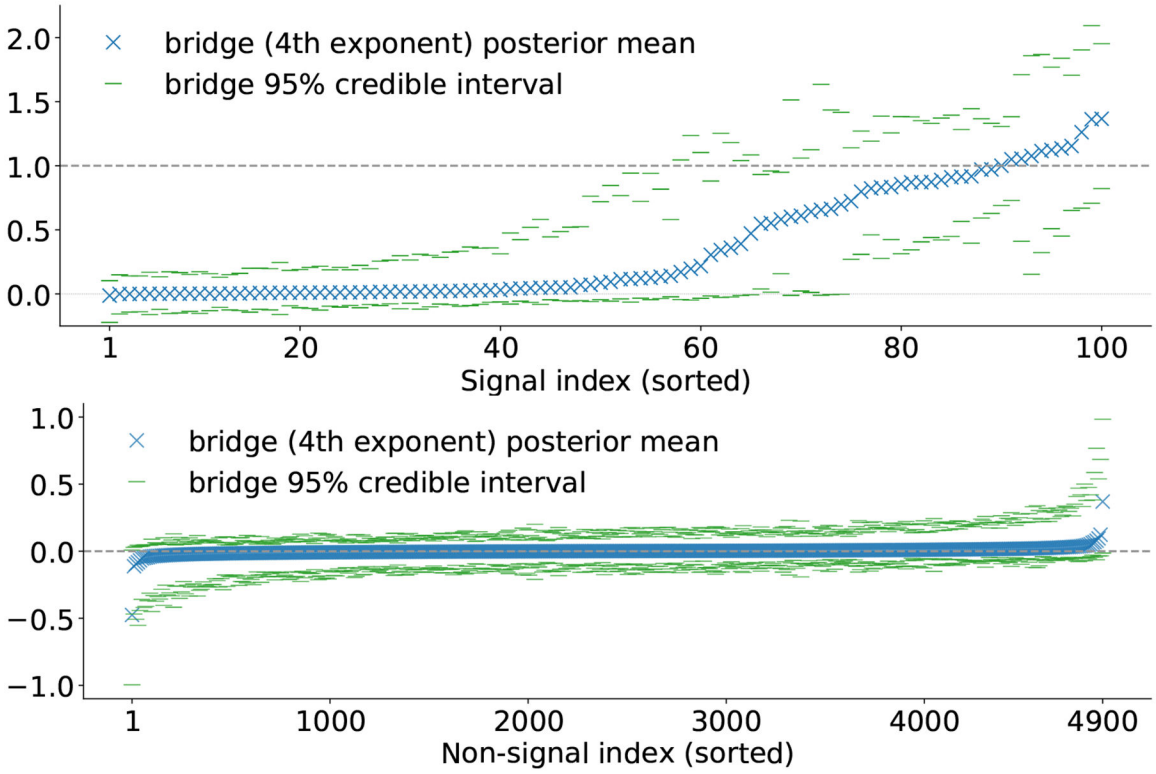


Figure 4.6:
The 95% posterior credible intervals under the Bayesian bridge logistic regression with the bridge exponent 1/4. Compared with the 1/16 exponent case (Figure 4.4), the posterior distributions have similar means but much wider credible intervals.

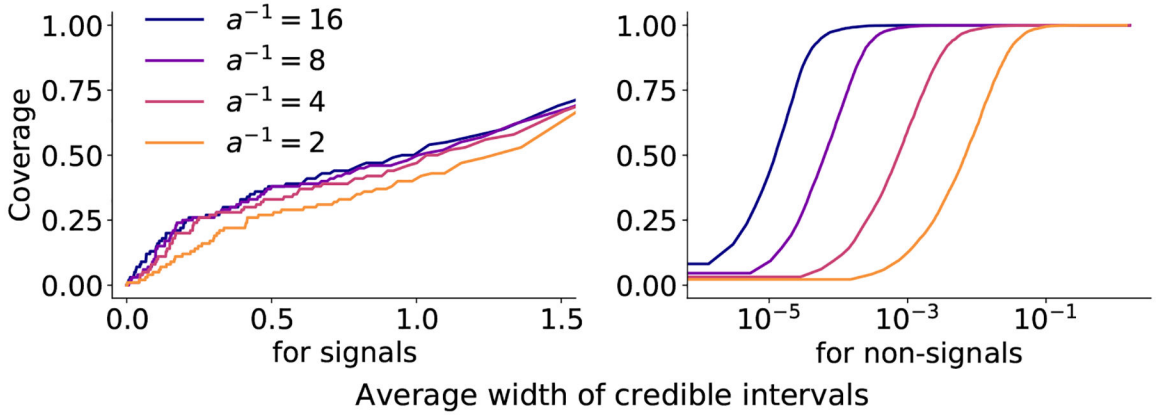


Figure 4.7:
Average width v.s. coverage of the credible intervals. The plots are produced by computing the equal-tailed credible intervals at a range of credible levels. The x-axis is in the $\log_{10}$ scale for the non-signals.

5 Discussion

Shrinkage priors have been adopted in a variety of Bayesian models, but the potential issues arising from their heavy-tails are often overlooked. Our method provides a simple and convenient way to regularize shrinkage priors, making the posterior inference more robust. Both the theoretical and empirical results demonstrate the benefits of regularization in improving the statistical and computational properties when parameters are only weakly identified. Much of the systematic investigations into the shrinkage prior properties has so far focused on rather simple models and situations in which signals are reasonable strong. Our work adds to the emerging efforts to better understand the behavior of shrinkage models in more complex settings.

Acknowledgments

We are indebted to Andrew Holbrook for the alliteration in the article title. This work was partially supported through National Institutes of Health grants R01 AI107034, U19 AI135995 and R01 AI153044 and through Food and Drug Administration grant HHS 75F40120D00039.

Appendix A: Further results on behavior of shrinkage model Gibbs samplers: probit regression as example

As we discussed in Section 3.2, Propositions 3.3 and 3.4 are quite general in scope and can provide insight into behavior of shrinkage model Gibbs samplers more broadly.

Here we demonstrate the broader relevance of these results, as well as of a few additional results, by applying them to establish uniform/geometric ergodicity of a Gibbs sampler for regularized Bayesian sparse probit regression. More explicitly, we consider the model

$$y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta} \sim \text{Bernoulli}\left(\boldsymbol{\Phi}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right),$$

$$z_j = 0 \text{ for } z_j \mid \beta_j \sim \mathcal{N}\left(\beta_j, \zeta^2\right),$$

$$\beta_j \mid \tau, \lambda_j \sim \mathcal{N}\left(0, \tau^2\lambda_j^2\right), \tau \sim \pi_{\text{glo}}\left(\cdot\right), \lambda_j \sim \pi_{\text{loc}}\left(\cdot\right),$$

where $\boldsymbol{\Phi}(t)$ denotes the cumulative distribution function of the standard Gaussian. The corresponding Gibbs sampler induces a transition kernel $(\boldsymbol{\beta}^*, \boldsymbol{\lambda}^*, \tau^*) \to (\boldsymbol{\beta}, \boldsymbol{\lambda}, \tau)$ through the following cycle of conditional updates:

1. Draw $\tau \mid \boldsymbol{\beta}^*, \boldsymbol{\lambda}^*$ from the density proportional to (2.4). When using Bayesian bridge priors, draw from the collapsed distribution $\tau \mid \boldsymbol{\beta}^*$ (Appendix E).
2. Draw $\boldsymbol{\lambda} \mid \boldsymbol{\beta}^*, \tau$ from the density proportional to (2.4).
3. Draw $\boldsymbol{\beta} \mid \tau, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}$ from the density proportional to

$$\pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}) \propto L_{\text{probit}}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) L(\boldsymbol{z} = \boldsymbol{0} \mid \boldsymbol{\beta}) \pi(\boldsymbol{\beta} \mid \tau, \lambda) \propto L_{\text{probit}}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) \pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{z} = \boldsymbol{0}) \tag{A.1}$$

where $L_{\text{probit}}\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}\right) = \prod_i \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})^{y_i} (1 - \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta}))^{1 - y_i}$ is the probit likelihood. The density (A.1) belongs to a unified skew-normal family, from which we can draw independent samples by the algorithm of Durante (2019).

Borrowing terminology from Durante (2019), we refer to the above Gibbs sampler as the *conjugate Gibbs sampler* for probit models to distinguish it from the more traditional one based on the data augmentation scheme of Albert and Chib (1993).

Theorems A.1 and A.2 below provide uniform and geometric ergodicity results for the conjugate Gibbs sampler and are exact analogues of the corresponding results Theorems 3.1 and 3.2 for the logistic case.

Theorem A. 1 (Uniform ergodicity for probit model). *If the prior* $\pi_{\text{glo}}(\cdot)$ *is supported on* $[\tau_{\text{min}}, \infty)$ *for* $\tau_{\text{min}} > 0$, *then the conjugate Gibbs sampler for regularized Baysian bridge probit regression is uniformly ergodic*.

Theorem A. 2 (Geometric ergodicity for probit model). *Suppose that the local scale prior satisfies* $\| \pi_{\text{loc}} \|_\infty < \infty$ *and that the global scale prior* $\pi_{\text{glo}}(\cdot)$ *is supported on* $[\tau_{\text{min}}, \tau_{\text{max}}]$ *for* $0 < \tau_{\text{min}} \leq \tau_{\text{max}} < \infty$. *Then the conjugate Gibbs sampler for regularized sparse probit regression is geometrically ergodic*.

A.1 Proofs of Theorem A.1 and A.2

The proof of Theorem A.1 (and A.2) above follows a path essentially identical to the proof of Theorem 3.1 (and 3.2) with most arguments carrying through verbatim or with trivial modifications; we only need to replace a few model-specific inequalities with the corresponding ones for the probit model. For establishing minorization conditions, Lemma A.3 below replaces Lemma 3.5. For establishing drift conditions, the bound on the conditional expectation of $|\beta_j|^{-\alpha}$ in Lemma A.4 replaces Eq. (3.11), and the bound on the conditional expectation of $\| \beta \|^2$ in Lemma A.5 replaces Eq. (3.13). Remarkably, Lemma A.4 and A.5 only requires a likelihood $L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$ to be a bounded function of $\boldsymbol{\beta}$ and thus may be applicable beyond the probit case.

We sketch out the proofs of Theorem A.1 and A.2 below. Again, the omitted details are essentially identical to the logistic case or, in fact, simpler because the probit case does not involve the additional Pólya-Gamma parameter.

Proof of Theorem A.1. A minorization result analogous to Theorem 3.6 follows from Proposition 3.3 and Lemma A.3. This minorization result straightforwardly implies a uniform minorization under Bayesian bridge priors as in Theorem 3.1. See the proofs of Theorem 3.6 and 3.1 for details. □

Proof of Theorem A.2. A minorization result analogous to Theorem 3.7 follows from Lemma A.3. Proposition 3.4, Lemma A.4, and Lemma A.5 together imply that $V(\boldsymbol{\beta}) = \sum_j |\beta_j|^{-\alpha} + \| \boldsymbol{\beta} \|^2$ is a Lyapunov function as in the proofs of Theorem 3.8 and 3.2. The geometric ergodicity then follows from the minorization and drift condition. See the proofs of Theorem 3.7, 3.8, and 3.2 for details. □

A. 2 Minorization lemma for probit model

Lemma A.3. *Whenever* $\min_j \tau\lambda_j \geq R > 0$, *there are* $\tilde{\delta}, \tilde{\delta}' > 0$ — *independent of* τ *and* λ *except through* R — *such that the following minorization condition holds*:

$$\begin{aligned}
&\pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}) \\
&\geq \tilde{\delta} L_{probit}\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}\right) \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{0}, \left(\zeta^{-2} + R^{-2}\right)^{-1} \boldsymbol{I}\right) \\
&\geq \tilde{\delta}' \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{0}, \left[\boldsymbol{X}^\top \boldsymbol{X} + \left(\zeta^{-2} + R^{-2}\right)\boldsymbol{I}\right]^{-1}\right).
\end{aligned} \tag{A.2}$$

Proof. The conditional distribution of $\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}$ is given by

$$\pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}) = \frac{L_{\text{probit}}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) \pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{z} = \boldsymbol{0})}{\int L_{\text{probit}}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}') \pi(\boldsymbol{\beta}' \mid \tau, \lambda, \boldsymbol{z} = \boldsymbol{0}) \mathrm{d}\boldsymbol{\beta}'}. \tag{A.3}$$

Since $\Phi(t) = 1 - \Phi(-t) \leq 1$ for all t, we have $\| L_{\text{probit}} \|_\infty \leq 1$ and

$$\int L_{\text{probit}}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}') \pi(\boldsymbol{\beta}' \mid \tau, \lambda, \boldsymbol{z} = \boldsymbol{0}) \mathrm{d}\boldsymbol{\beta}' \leq \int \pi(\boldsymbol{\beta}' \mid \tau, \lambda, \boldsymbol{z} = \boldsymbol{0}) \mathrm{d}\boldsymbol{\beta}' = 1. \tag{A.4}$$

Also, we can easily verify that the following inequality holds whenever $\min_j \tau\lambda_j \geq R$:

$$\begin{aligned}
\pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{z} = 0) &= \prod_j \frac{1}{\sqrt{2\pi}} \left(\zeta^{-2} + \tau^{-2}\lambda_j^{-2}\right)^{1/2} \exp\left(-\frac{1}{2}\left(\zeta^{-2} + \tau^{-2}\lambda_j^{-2}\right)\beta_j^2\right) \\
&\geq \prod_j \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\zeta^{-2} + R^{-2}\right)\beta_j^2\right).
\end{aligned} \tag{A.5}$$

Combining (A.4) and (A.5), we can lower bound (A.3) with $\tilde{\delta} > 0$ as

$$\pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}) \geq \tilde{\delta} L_{\text{probit}}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{0}, \left(\zeta^{-2} + R^{-2}\right)^{-1} \boldsymbol{I}\right), \tag{A.6}$$

establishing the first inequality in (A.2).

To establish the second inequality in (A.2), we will show that

$$\min\{\Phi(t), 1-\Phi(t)\} \geq \min\left\{1-\Phi(1), \frac{1}{2\sqrt{2\pi}}\right\}\exp\left(-t^2\right); \tag{A.7}$$

this will imply $L_{\text{probit}}\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}\right) \geq \min\left\{1-\Phi(1), \left(2\sqrt{2\pi}\right)^{-1}\right\}\exp\left(-\parallel \boldsymbol{X}\boldsymbol{\beta} \parallel^2\right)$ and complete the proof. Eq 7.1.13 of Abramowitz and Stegun (1965) tells us that

$$1-\Phi(t) \geq \frac{1}{\sqrt{2\pi}}\frac{t}{t^2+1}\exp\left(-\frac{t^2}{2}\right). \tag{A.8}$$

We therefore have

$$1-\Phi(t) \geq \frac{1}{2\sqrt{2\pi}}\frac{1}{t}\exp\left(-\frac{t^2}{2}\right) \geq \frac{1}{2\sqrt{2\pi}}\exp\left(-t^2\right) \text{ for } t \geq 1; \tag{A.9}$$

the latter inequality follows from the fact that $t^{-1} \geq \exp\left(-t^2/2\right)$ for $t \geq 1$, which can be proven, for example, by noting that $\frac{\mathrm{d}}{\mathrm{d}t}\left(t\exp\left(-t^2/2\right)\right) \leq 0$ for $t \geq 1$. For $t \leq 1$, we have $1-\Phi(t) \geq 1-\Phi(1)$ since $\Phi(t)$ is increasing in t. Combining the lower bounds for $t \geq 1$ and $t \leq 1$, we obtain

$$1-\Phi(t) \geq \min\left\{1-\Phi(1), \frac{1}{2\sqrt{2\pi}}\exp\left(-t^2\right)\right\} \geq \min\left\{1-\Phi(1), \frac{1}{2\sqrt{2\pi}}\right\}\exp\left(-t^2\right).$$

Since $\Phi(t) = 1-\Phi(-t)$, the same lower bound also holds for $\Phi(t)$, yielding (A.7). □

A. 3 Drift condition lemmas for bounded likelihood models

As we mentioned in Section A.1, Lemma A.4 and A.5 here apply not only to the probit case but also to any model whose likelihood is a bounded function of $\boldsymbol{\beta}$. Lemma A.4 in particular holds with or without the fictitious likelihood $L(\boldsymbol{z}=0 \mid \boldsymbol{\beta})$ for regularization. While stated in terms of a generic bounded likelihood $L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$, Lemma A.4 can be applied to regularized models simply by replacing the likelihood $\boldsymbol{\beta} \to L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$ in its statement with the regularized one $\boldsymbol{\beta} \to L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})L(\boldsymbol{z}=\boldsymbol{0} \mid \boldsymbol{\beta})$.

Lemma A.4. Let $\alpha \in [0,1)$. *Suppose the likelihood satisfies* $\parallel L \parallel_\infty := \sup_\beta L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) < \infty$ *and is strictly positive and continuous at* $\boldsymbol{\beta}=0$. *Then the following inequality holds for the conditional expectation under* $\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}$ *with constants* $C, C' < \infty$ *depending only on* α *and functionals of the likelihood* $\boldsymbol{\beta} \to L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$:

$$\mathbb{E}\left[|\beta_j|^{-\alpha} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}\right] \leq C|\tau\lambda_j|^{-\alpha} + C'. \tag{A.10}$$

Proof. The conditional distribution of $\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}$ is given by

$$\pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}) = \frac{L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})\pi(\boldsymbol{\beta} \mid \tau, \lambda)}{\int L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}')\pi(\boldsymbol{\beta}' \mid \tau, \lambda)\mathrm{d}\boldsymbol{\beta}'}. \tag{A.11}$$

We consider the conditional expectation (A.10) under two separate cases: $\max_j \tau\lambda_j \leq \epsilon$ and $\min_j \tau\lambda_j \geq \epsilon$, where $\epsilon > 0$ is any value small enough to guarantee the likelihood to be positive on the set $\| \boldsymbol{\beta}' \|_\infty = \max_j |\beta_j'| \leq \epsilon$.

When $\max_j \tau\lambda_j \leq \epsilon$, we have

$$\begin{aligned}
\int L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}')\pi(\boldsymbol{\beta}' \mid \tau, \lambda)\mathrm{d}\boldsymbol{\beta}' &\geq \int_{\| \boldsymbol{\beta}' \|_\infty \leq \epsilon} L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}')\pi(\boldsymbol{\beta}' \mid \tau, \lambda)\mathrm{d}\boldsymbol{\beta}' \\
&\geq \left(\min_{\| \boldsymbol{\beta}' \|_\infty \leq \epsilon} L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}')\right)\prod_j \int_{-\epsilon}^{\epsilon} \pi(\beta_j' \mid \tau, \lambda_j)\mathrm{d}\beta_j' \\
&\geq \left(\min_{\| \boldsymbol{\beta}' \|_\infty \leq \epsilon} L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}')\right)\left(\Phi(1) - \Phi(-1)\right)^p,
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian. Using the above lower bound on the numerator, we can bound (A.11) as

$$\pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}) \leq C_\epsilon \pi(\boldsymbol{\beta} \mid \tau, \lambda) \tag{A.12}$$

for $C_\epsilon = \| L \|_\infty / (\min_{\| \boldsymbol{\beta}' \|_\infty \leq \epsilon} L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}'))(\Phi(1) - \Phi(-1))^p$. It now follows that

$$\mathbb{E}\left[|\beta_j|^{-\alpha} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}\right] \leq C_\epsilon \mathbb{E}\left[|\beta_j|^{-\alpha} \mid \tau, \lambda\right] = C_\alpha C_\epsilon |\tau\lambda_j|^{-\alpha}, \tag{A.13}$$

where the latter equality with $C_\alpha = \Gamma\left(\frac{1-\alpha}{2}\right)/2^{\alpha/2}\sqrt{\pi}$ derives from the formula for negative moments of Gaussians (Winkelbauer, 2012).

Turning to the case $\min_j \tau\lambda_j \geq \epsilon$, we have

$$\pi(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}) = \frac{L\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}\right)\prod_j \exp\left(-\frac{\beta_j^2}{2\tau^2\lambda_j^2}\right)}{\int L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}')\prod_j \exp\left(-\frac{\beta_j'^2}{2\tau^2\lambda_j^2}\right)\mathrm{d}\boldsymbol{\beta}'} \leq \frac{\| L \|_\infty}{\int L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}')\prod_j \exp\left(-\beta_j'^2/2\epsilon^2\right)\mathrm{d}\boldsymbol{\beta}'} \coloneqq C_\epsilon'. \tag{A.14}$$

Using the above bound on the conditional density, we obtain

$$\begin{aligned}\mathbb{E}\left[|\beta_j|^{-\alpha} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}\right] &\leq 1 + \mathbb{E}\left[|\beta_j|^{-\alpha}\mathbb{1}\{|\beta_j| \leq 1\} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}\right] \\ &\leq 1 + C_\epsilon'\int_{-1}^{1}|\beta_j|^{-\alpha}\mathrm{d}\beta_j \\ &= 1 + 2C_\epsilon'/\left(1-\alpha\right).\end{aligned} \tag{A.15}$$

The bounds (A.13) and (A.15) together show that an inequality of the form (A.10) holds for any value of τ and λ, whether in $\{\max_j \tau\lambda_j \leq \epsilon\}$ or $\{\min_j \tau\lambda_j \geq \epsilon\}$. □

Lemma A.5. *Suppose the likelihood satisfies the assumptions as in Lemma A.4. Then the conditional expectation of β_j^2 under $\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}$ is bounded by a constant which depends only on ζ and functionals of the likelihood $\boldsymbol{\beta} \to L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$.*

Proof. We will derive the following bound on the conditional density

$$\pi\left(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}\right) \leq \widetilde{C}\mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{0}, \zeta^2\boldsymbol{I}\right)\left(1 + \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{0}, \tau^2\boldsymbol{\Lambda}^2\right)\right) = \widetilde{C}\mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{0}, \zeta^2\boldsymbol{I}\right) + \widetilde{C}\left(\tau^2\lambda_j^2 + \zeta^2\right)^{-1/2}\mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{0}, \left(\tau^{-2}\boldsymbol{\Lambda}^{-2} + \zeta^{-2}\boldsymbol{I}\right)^{-1}\right), \tag{A.16}$$

which will imply the desired bound on the conditional expectation:

$$\begin{aligned}\mathbb{E}\left[\beta_j^2 \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}\right] &\leq \tilde{C}\zeta^2 + \tilde{C}\left(\tau^2\lambda_j^2 + \zeta^2\right)^{-1/2}\left(\tau^{-2}\lambda_j^{-2} + \zeta^{-2}\right)^{-1} \\ &= \tilde{C}\zeta^2 + \tilde{C}\zeta^2\tau^2\lambda_j^2\left(\tau^2\lambda_j^2 + \zeta^2\right)^{-3/2} \\ &\leq \tilde{C}\zeta^2 + \tilde{C}\zeta^2\left(\tau^2\lambda_j^2 + \zeta^2\right)^{-1/2} \\ &\leq \tilde{C}\zeta^2 + \tilde{C}\zeta.\end{aligned}$$

To complete the proof, therefore, it remains to establish (A.16). Our argument here closely follows those we use in deriving the bounds (A.12) and (A.14) in the proof of Lemma A.4. The conditional distribution of $\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z} = \boldsymbol{0}$ is given by

$$\pi\left(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}=\mathbf{0}\right)=\frac{L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) L(\boldsymbol{z}=\mathbf{0} \mid \boldsymbol{\beta}) \pi(\boldsymbol{\beta} \mid \tau, \lambda)}{\int L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}') L(\boldsymbol{z}=\mathbf{0} \mid \boldsymbol{\beta}) \pi(\boldsymbol{\beta}' \mid \tau, \lambda) \mathrm{d}\boldsymbol{\beta}'}. \tag{A.17}$$

As before, we choose $\epsilon > 0$ to be any value small enough to guarantee the likelihood to be positive on the set $\| \boldsymbol{\beta}' \|_\infty = \max_j \left|\beta_j'\right| \leq \epsilon$. We can repeat an argument analogous to the derivation of the bound (A.12) to conclude that, when $\max_j \tau\lambda_j \leq \epsilon$,

$$\pi\left(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}=\mathbf{0}\right) \leq \widetilde{C}_\epsilon L\left(\boldsymbol{z}=\mathbf{0} \mid \boldsymbol{\beta}\right) \pi\left(\boldsymbol{\beta} \mid \tau, \lambda\right) \tag{A.18}$$

for $\widetilde{C}_\epsilon = \| L\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}\right) \|_\infty / (\min_{\| \boldsymbol{\beta}' \|_\infty \leq \epsilon} L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}') L(\boldsymbol{z}=\mathbf{0} \mid \boldsymbol{\beta})) \left(\Phi\left(1\right) - \Phi\left(-1\right)\right)^p$ with the $\| \cdot \|_\infty$ norm taken with respect to $\boldsymbol{\beta}$. For the case $\min_j \tau\lambda_j \geq \epsilon$, we follow the derivation of the bound (A.14) to conclude that

$$\pi\left(\boldsymbol{\beta} \mid \tau, \lambda, \boldsymbol{y}, \boldsymbol{X}\right) = \widetilde{C}_\epsilon L\left(\boldsymbol{z}=0 \mid \boldsymbol{\beta}\right) \text{ where } \widetilde{C}_\epsilon = \frac{\| L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) \|_\infty}{\int L(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}') L\left(\boldsymbol{z}=\mathbf{0} \mid \boldsymbol{\beta}\right) \prod_j \exp\left(-\beta_j'^2 / 2\epsilon^2\right) \mathrm{d}\boldsymbol{\beta}'}. \tag{A.19}$$

Combining (A.18) and (A.19) yields the desired bound (A.16). □

Appendix B: Proofs for Section 3.2

B. 1 Proof of Proposition 3.3

The key ingredient in our proof of Proposition 3.3 is the following general result on the stochastic ordering of tilted densities. The result allows us to study the behavior of $\pi(\lambda \mid \beta^*, \tau)$ viewed as a product of $f\left(\lambda\right) = \lambda^{-1} \pi_{\text{loc}}\left(\lambda\right)$ and $G\left(\lambda\right) = \exp\left(-\beta^{*2} / 2\tau^2\lambda^2\right)$.

Proposition B.1. *Consider probability densities $\pi_G(\lambda) \propto G(\lambda) f(\lambda)$ and $\pi_H(\lambda) \propto H(\lambda) f(\lambda)$ on $\lambda \in [0, \infty)$ for $f, G, H \geq 0$. Suppose that f satisfies $\int_u^\infty f(\lambda) \mathrm{d}\lambda < \infty$ for $u > 0$. Suppose also that G and H are absolutely continuous and increasing, $G \leq H$, and $\lim_{\lambda \to \infty} G(\lambda) = \lim_{\lambda \to \infty} H(\lambda)$. Then π_G is stochastically dominated by π_H i.e.*

$$\int_a^\infty \pi_G(\lambda) \mathrm{d}\lambda \leq \int_a^\infty \pi_H(\lambda) \mathrm{d}\lambda \text{ for any } a \in \mathbb{R}. \tag{B.1}$$

Proof. Multiplying G and H with an appropriate constant if necessary, without loss of generality we can assume $\lim_{\lambda \to \infty} G(\lambda) = \lim_{\lambda \to \infty} H(\lambda) = 1$ so that G and H can be interpreted as cumulative distribution functions.

We first deal with the case $G(0) = H(0) = 0$; when $\int f(\lambda)\mathrm{d}\lambda = \infty$, this assumption is in fact implied by the integrability of $G(\lambda)f(\lambda)$ and $H(\lambda)f(\lambda)$. In this case, we have $G(\lambda) = \int_0^\lambda g(u)\mathrm{d}u$ and $H(\lambda) = \int_0^\lambda h(u)\mathrm{d}u$ for density functions $g, h \geq 0$. As can be verified using Fubini's theorem for positive functions, we can express π_G and π_H as

$$\pi_G(\cdot) = \int f(\cdot \mid u)g(u)\mathrm{d}u \text{ and } \pi_H(\cdot) = \int f(\cdot \mid u)h(u)\mathrm{d}u,$$

where $f(\cdot \mid u)$ for $u > 0$ denote a probability density

$$f(\cdot \mid u) = \frac{f(\lambda)\mathbb{1}\{\lambda > u\}}{\int_u^\infty f(\lambda)\mathrm{d}\lambda}.$$

Again by Fubini's theorem for positive functions, we have

$$\int_a^\infty \pi_G(\lambda)\mathrm{d}\lambda = \int F_a(u)g(u)\mathrm{d}u \text{ and } \int_a^\infty \pi_H(\lambda)\mathrm{d}\lambda = \int F_a(u)h(u)\mathrm{d}u \tag{B.2}$$

where

$$F_a(u) = \int_a^\infty f(\lambda \mid u)\mathrm{d}\lambda = \frac{\int_{\max\{a,u\}}^\infty f(\lambda)\mathrm{d}\lambda}{\int_u^\infty f(\lambda)\mathrm{d}\lambda}.$$

Note that the integrals in (B.2) can be represented as expectations with respect to distributions G and:

$$\int_a^\infty \pi_G(\lambda)\mathrm{d}\lambda = \mathbb{E}_{U\sim G}[F_a(U)] \text{ and } \int_a^\infty \pi_H(\lambda)\mathrm{d}\lambda = \mathbb{E}_{U\sim H}[F_a(U)]. \tag{B.3}$$

Since F_a is an increasing function and G is stochastically dominated by H by our assumption, the representation (B.3) implies the desired inequality (B.1).

Earlier, we made a simplifying assumption $G(0) = H(0) = 0$. More generally, we have the relation $G(\lambda) - G(0) = \int_0^\lambda g(u)\mathrm{d}u$ and $H(\lambda) - H(0) = \int_0^\lambda h(u)\mathrm{d}u$ for integrable functions $g, h \geq 0$. Essentially the identical arguments as before show that the identity (B.3) and hence the conclusion (B.1) still hold in this case. □

Proof of Proposition 3.3. Note that

$$\pi(\lambda_j \mid \beta_j^*, \tau) \propto \exp\left(-c^2/\lambda_j^2\right)\lambda_j^{-1}\pi_{\mathrm{loc}}(\lambda_j) \text{ for } c = c(\beta_j^*/\tau) = \frac{\beta_j^*}{\sqrt{2}\tau}.$$

Applying Proposition B.1 with $f(\lambda) = \lambda^{-1}\pi_{\text{loc}}(\lambda)$, we see that

$$\mathbb{P}(\lambda_j > a \mid \beta_j^*, \tau) \leq \mathbb{P}(\lambda_j > a \mid \beta_j^{*'}, \tau)$$

whenever $|\beta_j^*/\tau| \geq |\beta_j^{*'}/\tau|$.

Suppose now that $\int \lambda^{-1}\pi_{\text{loc}}(\lambda)\mathrm{d}\lambda = \infty$. For any β_j^*/τ, we have

$$\int_a^\infty \exp\left(-\frac{\beta_j^{*2}}{2\tau^2\lambda_j^2}\right)\lambda_j^{-1}\pi_{\text{loc}}(\lambda_j)\mathrm{d}\lambda_j \leq \int_a^\infty \lambda_j^{-1}\pi_{\text{loc}}(\lambda_j)\mathrm{d}\lambda_j \leq 1/a\,. \tag{B.4}$$

On the other hand, by Fatou's lemma,

$$\liminf_{|\beta_j^*/\tau| \to 0} \int \exp\left(-\frac{\beta_j^{*2}}{2\tau^2\lambda^2}\right)\lambda^{-1}\pi_{\text{loc}}(\lambda)\mathrm{d}\lambda \geq \int \lambda^{-1}\pi_{\text{loc}}(\lambda)\mathrm{d}\lambda = \infty\,. \tag{B.5}$$

From (B.4) and (B.5), we conclude that for any $a > 0$

$$\mathbb{P}(\lambda_j > a \mid \beta_j^*, \tau) = \frac{\int_a^\infty \exp\left(-\frac{\beta_j^{*2}}{2\tau^2\lambda_j^2}\right)\lambda_j^{-1}\pi_{\text{loc}}(\lambda_j)\mathrm{d}\lambda_j}{\int \exp\left(-\frac{\beta_j^{*2}}{2\tau^2\lambda^2}\right)\lambda^{-1}\pi_{\text{loc}}(\lambda)\mathrm{d}\lambda} \to 0 \text{ as } |\beta_j^*/\tau| \to 0,$$

i.e. $\pi(\lambda_j \mid \beta_j^*, \tau)$ converges in distribution to a delta measure at 0.

We now turn to quantifying the limiting behavior when $\int \lambda^{-1}\pi_{\text{loc}}(\lambda)\mathrm{d}\lambda < \infty$. For any $a \in [0, \infty]$, the dominated convergence theorem yields

$$\lim_{|\beta_j^*/\tau| \to 0} \int_0^a \exp\left(-\frac{\beta_j^{*2}}{2\tau^2\lambda_j^2}\right)\lambda_j^{-1}\pi_{\text{loc}}(\lambda_j)\mathrm{d}\lambda_j = \int_0^a \lambda^{-1}\pi_{\text{loc}}(\lambda)\mathrm{d}\lambda\,.$$

The above convergence result implies the point-wise convergence of the cumulative distribution function:

$$\lim_{|\beta_j^*/\tau| \to 0} \mathbb{P}(\lambda_j \leq a \mid \beta_j^*, \tau) = \frac{\int_0^a \lambda_j^{-1}\pi_{\text{loc}}(\lambda_j)\mathrm{d}\lambda_j}{\int \lambda^{-1}\pi_{\text{loc}}(\lambda)\mathrm{d}\lambda}\,.$$

B. 2 Proof of Proposition 3.4

Proof. In upper-bounding $\mathbb{E}[\lambda_j^{-\alpha} \mid \tau, \boldsymbol{\beta}^*]$, we can without loss of generality assume that $\pi(0) > 0$ by virtue of Proposition B.2 below. In terms of the constants ϵ and $C''(\alpha, \pi_{\mathrm{loc}})$ as defined in Lemma B.3 below, let

$$\gamma(r) = C''(\alpha, \pi_{\mathrm{loc}}) / \log\left(1 + \frac{4\epsilon^2}{r^2}\right). \tag{B.6}$$

By Lemma B.3 and the monotonicity of $\gamma(r)$, we then have

$$\mathbb{E}\left[\tau^{-\alpha}\lambda_j^{-\alpha} \mid \tau, \beta_j^*\right] \leq \gamma(R/\tau)\left|\beta_j^*\right|^{-\alpha} \text{ whenever } \left|\beta_j^*\right| \leq R.$$

On the other hand, since the distribution $\lambda_j \mid \tau, \beta_j^*$ stochastically dominates $\lambda_j \mid \tau, \beta_j^{*'}$ whenever $\beta_j^* \geq \beta_j^{*'}$ (Proposition 3.3), we have

$$\mathbb{E}\left[\tau^{-\alpha}\lambda_j^{-\alpha} \mid \tau, \beta_j^*\right] \leq \mathbb{E}\left[\tau^{-\alpha}\lambda_j^{-\alpha} \middle| \tau, \left|\beta_j^{*'}\right| = R\right] \text{ whenever } \left|\beta_j^*\right| \geq R. \tag{B.7}$$

Combining (B.6) and (B.7) yields the inequality (3.3).

Proposition B.2. *Given a prior* $\pi_{\mathrm{loc}}(\cdot)$ *such that* $\pi_{\mathrm{loc}}(0) = 0$ *and* $\| \pi_{\mathrm{loc}} \|_\infty < \infty$, *there is a density* $\pi'_{\mathrm{loc}}(\cdot)$ *such that* $\pi'_{\mathrm{loc}}(\lambda)$ *is continuous at* $\lambda = 0, \pi'_{\mathrm{loc}}(0) > 0, \| \pi'_{\mathrm{loc}} \|_\infty < \infty$, *and* $\pi_{\mathrm{loc}}(\lambda) \propto G(\lambda)\pi'_{\mathrm{loc}}(\lambda)$ *for a bounded increasing function* $G \geq 0$. *Consequently, a density* $\pi(\cdot)$ *stochastically dominates* $\pi'(\cdot)$ *when* $\pi(\lambda) \propto f(\lambda)\pi_{\mathrm{loc}}(\lambda)$ *and* $\pi'(\lambda) \propto f(\lambda)\pi'_{\mathrm{loc}}(\lambda)$ *for* $f \geq 0$. *By taking* $f(\lambda) = \lambda^{-1}\exp\left(-\beta_j^{*2}/2\tau^2\lambda_j^2\right)$ *in particular, we have the following inequality between the expectations with respect to* $\pi(\cdot)$ *and* $\pi'(\cdot)$:

$$\mathbb{E}\left[\lambda_j^{-\alpha} \mid \tau, \beta_j^*\right] \leq \mathbb{E}'\left[\lambda_j^{-\alpha} \mid \tau, \beta_j^*\right] \text{ for } \alpha \geq 0. \tag{B.8}$$

Proof. Redefining $\pi_{\mathrm{loc}}(\lambda)$ as $\pi_{\mathrm{loc}}(\lambda - \lambda_{\min})$ for $\lambda_{\min} = \inf\{\lambda : \pi_{\mathrm{loc}}(\lambda) > 0\}$ if necessary, we can without loss of generality assume that $\pi_{\mathrm{loc}}(\lambda) > 0$ for all sufficiently small $\lambda > 0$. Define

$$G(\lambda) = \min\left\{\|\pi_{\mathrm{loc}}\|_\infty, \int_0^\lambda \max\left\{0, \frac{\mathrm{d}\pi_{\mathrm{loc}}}{\mathrm{d}\lambda}(u)\right\}\mathrm{d}u\right\}. \tag{B.9}$$

Then G is clearly increasing and bounded. The definition (B.9) further guarantees that $\lim_{\lambda \to 0}\pi_{\mathrm{loc}}(\lambda)/G(\lambda) = 1, \pi_{\mathrm{loc}} \leq G$, and $\lim_{\lambda \to \infty} G(\lambda) = \| \pi_{\mathrm{loc}} \|_\infty$. Define $\pi'_{\mathrm{loc}}(\cdot)$ via the

relation $\pi'_{\text{loc}}(\lambda) \propto \pi_{\text{loc}}(\lambda)/G(\lambda)$ for $\lambda > 0$ and $\pi'_{\text{loc}}(0) \coloneqq \lim_{\lambda \to 0}\pi'_{\text{loc}}(\lambda)$. Then $\pi'_{\text{loc}}(\cdot)$ satisfy $\| \pi'_{\text{loc}} \|_\infty = \pi'_{\text{loc}}(0) = \left(\int \pi(\lambda)/G(\lambda)\mathrm{d}\lambda\right)^{-1} > 0$, as well as all the other desired properties.

When $\pi(\lambda) \propto f(\lambda)\pi_{\text{loc}}(\lambda)$ and $\pi'(\lambda) \propto f(\lambda)\pi'_{\text{loc}}(\lambda)$, the densities satisfies the relation $\pi'(\lambda) \propto G(\lambda)\pi(\lambda)$. By applying Proposition B.1 with $H = \| G \|_\infty$, we conclude that $\pi(\cdot)$ stochastically dominates $\pi'(\cdot)$. The inequality (B.8) is an immediate consequence of this stochastic ordering. □

Lemma B.3. *Suppose that* $\pi_{\text{loc}}(\lambda)$ *is continuous at* $\lambda = 0$ *and* $\pi_{\text{loc}}(0) > 0$. *For* $\alpha \in [0,1)$ *and* $\epsilon > 0$ *small enough that* $\min_{\lambda \in [0,\epsilon]}\pi_{\text{loc}}(\lambda) \geq \pi_{\text{loc}}(0)/2$, *we have the following inequality*:

$$\mathbb{E}[\tau^{-\alpha}\lambda_j^{-\alpha} \mid \tau, \boldsymbol{\beta}^*] \leq C''(\alpha, \pi_{\text{loc}})\left|\beta_j^*\right|^{-\alpha}/\log\left(1 + \frac{4\tau^2\epsilon^2}{\left|\beta_j^*\right|^2}\right),$$

where $C''(\alpha, \pi_{\text{loc}}) > 0$ is a constant depending only on α and $\pi_{\text{loc}}(\cdot)$ *given by*

$$C''(\alpha, \pi_{\text{loc}}) = 2^{2+\alpha/2}\frac{\| \pi_{\text{loc}} \|_\infty}{\pi_{\text{loc}}(0)}\int_0^\infty \frac{1}{\lambda^{1+\alpha}}\exp\left(-\frac{1}{\lambda^2}\right)\mathrm{d}\lambda\,.$$

Proof. Observe that

$$\mathbb{E}[\lambda_j^{-\alpha} \mid \tau, \boldsymbol{\beta}^*] = \int_0^\infty \frac{1}{\lambda^{1+\alpha}}\exp\left(-\frac{c_j^2}{\lambda^2}\right)\pi_{\text{loc}}(\lambda)\mathrm{d}\lambda \Big/ \int_0^\infty \frac{1}{\lambda}\exp\left(-\frac{c_j^2}{\lambda^2}\right)\pi_{\text{loc}}(\lambda)\mathrm{d}\lambda, \tag{B.10}$$

where $c_j = c(\tau, \beta_j) = |\beta_j|/\sqrt{2}\tau$. With the change of variable $\lambda \to \lambda/c_j$, we can write the right-hand side of (B.10) as

$$\frac{1}{c_j^\alpha}\int_0^\infty \frac{1}{\lambda^{1+\alpha}}\exp\left(-\frac{1}{\lambda^2}\right)\pi_{\text{loc}}(c_j\lambda)\mathrm{d}\lambda \Big/ \int_0^\infty \frac{1}{\lambda}\exp\left(-\frac{1}{\lambda^2}\right)\pi_{\text{loc}}(c_j\lambda)\mathrm{d}\lambda\,. \tag{B.11}$$

We can upper bound the numerator as

$$\frac{1}{c_j^\alpha}\int_0^\infty \frac{1}{\lambda^{1+\alpha}}\exp\left(-\frac{1}{\lambda^2}\right)\pi_{\text{loc}}(c_j\lambda)\mathrm{d}\lambda \leq \frac{1}{c_j^\alpha}\|\pi_{\text{loc}}\|_\infty\int_0^\infty \frac{1}{\lambda^{1+\alpha}}\exp\left(-\frac{1}{\lambda^2}\right)\mathrm{d}\lambda\,. \tag{B.12}$$

To lower bound the denominator, we restrict the range of integration to $[0, \epsilon/c_j]$ for $\epsilon > 0$ and apply the change of variable $\phi = \lambda^{-2}$:

$$\int_0^\infty \frac{1}{\lambda}\exp\left(-\frac{1}{\lambda^2}\right)\pi_{\mathrm{loc}}(c_j\lambda)\mathrm{d}\lambda \geq \left(\min_{[0,\epsilon]}\pi_{\mathrm{loc}}\right)\int_0^{\epsilon/c_j}\frac{1}{\lambda}\exp\left(-\frac{1}{\lambda^2}\right)\mathrm{d}\lambda$$
$$= \left(\min_{[0,\epsilon]}\pi_{\mathrm{loc}}\right)\int_{c_j^2/\epsilon^2}^\infty \phi^{-1}\exp(-\phi)\mathrm{d}\phi\,.$$

The inequality of Gautschi (1959) tells us that $\int_a^\infty \phi^{-1}\exp\left(-\phi\right)\mathrm{d}\phi \geq \log\left(1+2a^{-1}\right)/2$, so we obtain

$$\int_0^\infty \frac{1}{\lambda}\exp\left(-\frac{1}{\lambda^2}\right)\pi_{\mathrm{loc}}(c_j\lambda)\mathrm{d}\lambda \geq \left(\min_{[0,\epsilon]}\pi_{\mathrm{loc}}\right)\frac{1}{2}\log\left(1+2\frac{\epsilon^2}{c_j^2}\right). \tag{B.13}$$

From the upper bound (B.12) of the numerator and lower bound (B.13) of the denominator, it follows that the ratio (B.11) is upper bounded by

$$c_j^{-\alpha}\frac{2\|\,\pi_{\mathrm{loc}}\,\|_\infty}{\left(\min_{[0,\epsilon]}\pi_{\mathrm{loc}}\right)\log\left(1+2\epsilon^2c_j^{-2}\right)}\int_0^\infty \frac{1}{\lambda^{1+\alpha}}\exp\left(-\frac{1}{\lambda^2}\right)\mathrm{d}\lambda\,.$$

Substituting $c_j = |\beta_j|/\sqrt{2}\tau$ into the above expression completes the proof. □

Appendix C: Proof of Lemma 3.5

Our proof of Lemma 3.5 builds on the known fact below.

Proposition C.1 (Choi and Hobert, 2013). *For fixed τ and λ, the marginal transition kernel satisfies the minorization condition*

$$P(\boldsymbol{\beta}\mid\boldsymbol{\beta}^*,\tau,\lambda) \geq \delta_{\tau\lambda}\mathcal{N}\left(\boldsymbol{\beta};\boldsymbol{\mu}_{\tau\lambda},\boldsymbol{\Phi}_{\tau\lambda}^{-1}\right)$$

where $\boldsymbol{\Phi}_{\tau\lambda} = \frac{1}{2}\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \zeta^{-2}\boldsymbol{I} + \tau^{-2}\boldsymbol{\Lambda}^{-2}$, $\boldsymbol{\mu}_{\tau\lambda} = \boldsymbol{\Phi}_{\tau\lambda}^{-1}\boldsymbol{X}^\mathsf{T}\left(\boldsymbol{y}-\mathbf{1}/2\right)$, *and*

$$\delta_{\tau\lambda} = C_n\frac{\left|\zeta^{-2}\boldsymbol{I}+\tau^{-2}\boldsymbol{\Lambda}^{-2}\right|^{1/2}}{\left|\boldsymbol{\Phi}_{\tau\lambda}\right|^{1/2}}\exp\left\{\frac{1}{2}\boldsymbol{w}^\mathsf{T}\left[\boldsymbol{\Phi}_{\tau\lambda}^{-1}-\left(\zeta^{-2}\boldsymbol{I}+\tau^{-2}\boldsymbol{\Lambda}^{-2}\right)^{-1}\right]\boldsymbol{w}\right\} \tag{C.1}$$

for $\boldsymbol{w} = \boldsymbol{X}^\mathsf{T}\left(\boldsymbol{y}-\mathbf{1}/2\right)$ *and* $C_n > 0$ *depending only on* n.

Proposition C.2 and C.3 below are the main workhorses for our proof of Lemma 3.5 along with Proposition C.1. We first state the results and use them to prove Lemma 3.5, before proceeding to prove the results themselves.

Proposition C.2. *As a function of* $\tau\lambda$*, the minorization constant* (C.1) *is uniformly bounded below by a positive constant on the set* $\min_j \tau\lambda_j \geq R > 0$.

Proposition C.3. *If two precision matrices* $\boldsymbol{\Phi}$ *and* $\boldsymbol{\Phi}'$ *satisfy* $\boldsymbol{\Phi} \prec \boldsymbol{\Phi}'$*, then a minorization* $\mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}, \boldsymbol{\Phi}^{-1}\right) \geq \delta \mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}', \boldsymbol{\Phi}'^{-1}\right)$ *holds for* $\delta > 0$ *given by*

$$\begin{aligned}\delta &= \inf_{\boldsymbol{\beta}} \frac{\mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}, \boldsymbol{\Phi}^{-1}\right)}{\mathcal{N}\left(\boldsymbol{\beta}; \boldsymbol{\mu}', \boldsymbol{\Phi}'^{-1}\right)} \\ &= \frac{|\boldsymbol{\Phi}|^{1/2}}{|\boldsymbol{\Phi}'|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}' - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Phi}\left[(\boldsymbol{\Phi}' - \boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}'\boldsymbol{\mu}' - \boldsymbol{\Phi}\boldsymbol{\mu}) - \boldsymbol{\mu}\right]\right\}. \end{aligned} \tag{C.2}$$

When the means take the form $\boldsymbol{\mu} = \boldsymbol{\Phi}^{-1}\boldsymbol{w}$ *and* $\boldsymbol{\mu}' = \boldsymbol{\Phi}^{-1}\boldsymbol{w}$, (C.2) *simplifies to*

$$\delta = \frac{|\boldsymbol{\Phi}|^{1/2}}{|\boldsymbol{\Phi}'|^{1/2}} \exp\left\{\frac{1}{2}\boldsymbol{w}^{\top}\left(\boldsymbol{\Phi}^{-1} - \boldsymbol{\Phi}'^{-1}\right)\boldsymbol{w}\right\} \geq \frac{|\boldsymbol{\Phi}|^{1/2}}{|\boldsymbol{\Phi}'|^{1/2}}.$$

Proof of Lemma 3.5. On the set $\{\lambda : \min_j \tau\lambda_j \geq R\}$, Proposition C. 1 implies that

$$P(\boldsymbol{\beta} \mid \boldsymbol{\beta}^*, \tau, \lambda) \geq \left(\min_{\tau\lambda_j \geq R} \delta_{\tau\lambda}\right) \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_{\tau\lambda}, \boldsymbol{\Phi}_{\tau\lambda}^{-1}),$$

where $\min_{\tau\lambda_j \geq R} \delta_{\tau\lambda}$ is guaranteed to be strictly positive by Proposition C.2.

We complete the proof by showing that the following inequality holds whenever $\min_j \tau\lambda_j \geq R$:

$$\mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_{\tau\lambda}, \boldsymbol{\Phi}_{\tau\lambda}^{-1}) \geq \frac{|\boldsymbol{\Phi}_{\infty}|^{1/2}}{|\boldsymbol{\Phi}_R|^{1/2}} \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_R, \boldsymbol{\Phi}_R^{-1}) \tag{C.3}$$

for $\boldsymbol{\Phi}_{\infty} = \frac{1}{2}\boldsymbol{X}^{\top}\boldsymbol{X} + \zeta^{-2}\boldsymbol{I}$. When $\min_j \tau\lambda_j > R$, we have $R^{-2} - \tau^{-2}\lambda_j^{-2} > 0$ and hence

$$\boldsymbol{\Phi}_R - \boldsymbol{\Phi}_{\tau\lambda} = \left(R^{-2}\boldsymbol{I} - \tau^{-2}\boldsymbol{\Lambda}^{-2}\right) \succ 0.$$

By Proposition C.3, it follows that

$$\mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_{\tau\lambda}, \boldsymbol{\Phi}_{\tau\lambda}^{-1}) \geq \frac{|\boldsymbol{\Phi}_{\tau\lambda}|^{1/2}}{|\boldsymbol{\Phi}_R|^{1/2}} \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_R, \boldsymbol{\Phi}_R^{-1}). \tag{C.4}$$

The above inequality in fact holds not only on the set $\{\lambda : \tau\lambda_j > R\}$ but also on the closure $\{\lambda : \min_j \tau\lambda_j \geq R\}$ since all the quantities depend continuously on $\tau\lambda_j$. The inequality (C.3) follows from (C.4) by observing that $\boldsymbol{\Phi}_{\tau\lambda} \succ \boldsymbol{\Phi}_\infty$ and hence $|\boldsymbol{\Phi}_{\tau\lambda}| \geq |\boldsymbol{\Phi}_\infty|$. □

Proof of Proposition C.2 and C.3

In the proofs to follow, we will make use of the following elementary linear algebra facts about positive definite matrices. We will denote the largest, ith largest, and smallest eigenvalue of a matrix $\boldsymbol{A}$ as $\nu_{\max}(\boldsymbol{A})$, $\nu_i(\boldsymbol{A})$, and $\nu_{\min}(\boldsymbol{A})$. The determinant of $\boldsymbol{A}$ is denoted by $|\boldsymbol{A}|$ and the trace by $\mathrm{tr}(\boldsymbol{A})$. The notation $\boldsymbol{A} \prec \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive definite or, equivalently, $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} < \boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{v}$ for any vector $\boldsymbol{v} \neq 0$.

Proposition C.4. *Given positive definite matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$*, we have*

1. $\left(\boldsymbol{A} + \boldsymbol{B}\right)^{-1} \prec \boldsymbol{A}^{-1}$.
2. $\left(\boldsymbol{A} + \boldsymbol{B}\right)^{-1} \succ \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$
3. $\nu_i(\boldsymbol{A}) + \nu_{\min}(\boldsymbol{B}) \leq \nu_i(\boldsymbol{A} + \boldsymbol{B}) \leq \nu_i(\boldsymbol{A}) + \nu_{\max}(\boldsymbol{B})$ *for all* i.
4. $|\boldsymbol{A}| < |\boldsymbol{A} + \boldsymbol{B}|$.
5. $\left|\boldsymbol{A} + \boldsymbol{B}\right| \leq \left|\boldsymbol{A}\right| \exp\left\{\nu_{\max}\left(\boldsymbol{B}\right)\mathrm{tr}\left(\boldsymbol{A}^{-1}\right)\right\}$.

When $\boldsymbol{A} \prec \boldsymbol{C}$ *for another positive definite matrix* $\boldsymbol{C}$*, we can apply above results with* $\boldsymbol{B} = \boldsymbol{C} - \boldsymbol{A} \succ 0$ *to obtain analogous inequalities*.

Proof. The eigenvalues of $\boldsymbol{I} + \boldsymbol{B}$ are given by $1 + \nu_i(\boldsymbol{B})$ and those of $(\boldsymbol{I} + \boldsymbol{B})^{-1}$ by $1/(1 + \nu_i(\boldsymbol{B})) < 1$, so we have $(\boldsymbol{I} + \boldsymbol{B})^{-1} \prec \boldsymbol{I}$. This result holds when $\boldsymbol{B}$ is replaced by $\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}$ and thus implies that

$$\begin{aligned}\boldsymbol{v}^\top\left(\boldsymbol{A} + \boldsymbol{B}\right)^{-1}\boldsymbol{v} &= \boldsymbol{v}^\top\boldsymbol{A}^{-1/2}\left(\boldsymbol{I} + \boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}\right)^{-1}\boldsymbol{A}^{-1/2}\boldsymbol{v}^\top \\ &< \boldsymbol{v}^\top\boldsymbol{A}^{-1/2}\boldsymbol{A}^{-1/2}\boldsymbol{v}^\top\end{aligned}$$

for $\boldsymbol{v} \neq 0$. Hence we have $\left(\boldsymbol{A} + \boldsymbol{B}\right)^{-1} < \boldsymbol{A}^{-1}$.

To prove Property 2, we first show $(\boldsymbol{I} + \boldsymbol{B})^{-1} \succ \boldsymbol{I} - \boldsymbol{B}$. By applying a change of basis if necessary, we can assume that $\boldsymbol{B}$ is diagonal. Since $(1 + B_{ii})^{-1} > 1 - B_{ii}$, we have

$$\boldsymbol{v}^\top\left(\boldsymbol{I} + \boldsymbol{B}\right)^{-1}\boldsymbol{v} = \sum_i (1 + B_{ii})^{-1} v_i^2 > \sum_i (1 - B_{ii}) v_i^2 = \boldsymbol{v}^\top(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{v}\,.$$

Since the result $(\boldsymbol{I} + \boldsymbol{B})^{-1} \succ \boldsymbol{I} - \boldsymbol{B}$ holds when $\boldsymbol{B}$ is replaced by $\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}$, we obtain

$$\begin{aligned}(\boldsymbol{A}+\boldsymbol{B})^{-1} &= \boldsymbol{A}^{-1/2}\left(\boldsymbol{I}+\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}\right)^{-1}\boldsymbol{A}^{-1/2}\\ &\succ \boldsymbol{A}^{-1/2}\left(\boldsymbol{I}-\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}\right)\boldsymbol{A}^{-1/2}\\ &= \boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}.\end{aligned}$$

Property 3 is Theorem 8.1.5 of Golub and Van Loan (2012) and immediately implies Property 4.

For Property 5, observe that

$$|\boldsymbol{A}+\boldsymbol{B}| = \prod_i \nu_i(\boldsymbol{A}+\boldsymbol{B}) \le \prod_i \{\nu_i(\boldsymbol{A})+\nu_{\max}(\boldsymbol{B})\}.$$

Taking the logarithm and applying the inequality $\log(1+x)\le x$, we have

$$\begin{aligned}\log|\boldsymbol{A}+\boldsymbol{B}|-\log|\boldsymbol{A}| &\le \sum_i \log\left(1+\frac{\nu_{\max}(\boldsymbol{B})}{\nu_i(\boldsymbol{A})}\right)\\ &\le \sum_i \frac{\nu_{\max}(\boldsymbol{B})}{\nu_i(\boldsymbol{A})}\\ &= \nu_{\max}(\boldsymbol{B})\mathrm{tr}\left(\boldsymbol{A}^{-1}\right).\end{aligned}$$

□

Proof of Proposition C.2. Throughout the proof, we use the notation $\boldsymbol{\Phi}_\infty = \frac{1}{2}\boldsymbol{X}^\top\boldsymbol{X}\zeta^{-2}\boldsymbol{I}$ so that $\boldsymbol{\Phi}_{\tau\lambda} = \boldsymbol{\Phi}_\infty + \tau^{-2}\boldsymbol{\Lambda}^{-2}$. By Proposition C.4, we have

$$\begin{aligned}\left|\zeta^{-2}\boldsymbol{I}+\tau^{-2}\boldsymbol{\Lambda}^{-2}\right| &\ge \left|\zeta^{-2}\boldsymbol{I}\right|\\ \left|\boldsymbol{\Phi}_\infty+\tau^{-2}\boldsymbol{\Lambda}^{-2}\right| &\le |\boldsymbol{\Phi}_\infty|\exp\left\{\left(\max_j\tau^{-2}\lambda_j^{-2}\right)\mathrm{tr}\left(\boldsymbol{\Phi}_\infty^{-1}\right)\right\}.\end{aligned}$$

The above inequalities imply that

$$\frac{\left|\zeta^{-2}\boldsymbol{I}+\tau^{-2}\boldsymbol{\Lambda}^{-2}\right|^{1/2}}{|\boldsymbol{\Phi}|^{1/2}} \ge \frac{\left|\zeta^{-2}\boldsymbol{I}\right|}{|\boldsymbol{\Phi}_\infty|}\exp\left\{-\frac{1}{\min_j\tau^2\lambda_j^2}\mathrm{tr}\left(\boldsymbol{\Phi}_\infty^{-1}\right)\right\}. \tag{C.5}$$

Also by Proposition C.4, we have

$$\begin{aligned}\left(\zeta^{-2}\boldsymbol{I}+\tau^{-2}\boldsymbol{\Lambda}^{-2}\right)^{-1} &\prec \zeta^2\boldsymbol{I}\\ \left(\boldsymbol{\Phi}_\infty+\tau^{-2}\boldsymbol{\Lambda}^{-2}\right)^{-1} &\succ \boldsymbol{\Phi}_\infty^{-1}-\boldsymbol{\Phi}_\infty^{-1}\tau^{-2}\boldsymbol{\Lambda}^{-2}\boldsymbol{\Phi}_\infty^{-1}.\end{aligned}$$

We therefore have

$$\begin{aligned}
&\boldsymbol{w}^{\mathsf{T}}\left[\boldsymbol{\Phi}_{\tau\lambda}^{-1}-\left(\zeta^{-2}\boldsymbol{I}+\tau^{-2}\boldsymbol{\Lambda}^{-2}\right)^{-1}\right]\boldsymbol{w}\\
&\geq \boldsymbol{w}^{\top}\boldsymbol{\Phi}_{\infty}^{-1}\boldsymbol{w}-\boldsymbol{w}^{\top}\boldsymbol{\Phi}_{\infty}^{-1}\tau^{-2}\boldsymbol{\Lambda}^{-2}\boldsymbol{\Phi}_{\infty}^{-1}\boldsymbol{w}-\zeta^{-2}\parallel\boldsymbol{w}\parallel^{2}\\
&\geq \boldsymbol{w}^{\top}\boldsymbol{\Phi}_{\infty}^{-1}\boldsymbol{w}-\frac{1}{\min_j\tau^{2}\lambda_j^{2}}\parallel\boldsymbol{\Phi}_{\infty}^{-1}\boldsymbol{w}\parallel^{2}-\zeta^{-2}\parallel\boldsymbol{w}\parallel^{2}.
\end{aligned} \tag{C.6}$$

From (C.5) and (C.6), we see that for all $\min_j\tau\lambda_j \geq R$

$$\delta_{\tau\lambda} \geq C_n\frac{\left|\zeta^{-2}\boldsymbol{I}\right|^{1/2}}{\left|\boldsymbol{\Phi}_{\infty}\right|^{1/2}}\exp\left\{\boldsymbol{w}^{\top}\boldsymbol{\Phi}_{\infty}^{-1}\boldsymbol{w}-\zeta^{-2}\parallel\boldsymbol{w}\parallel^{2}-\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_{\infty}^{-1}\right)+\parallel\boldsymbol{\Phi}_{\infty}^{-1}\boldsymbol{w}\parallel^{2}}{R^{2}}\right\}.$$

□

Proof of Proposition C.3. Note that

$$\inf_{\boldsymbol{\beta}}\frac{\mathcal{N}\left(\boldsymbol{\beta};\boldsymbol{\mu},\boldsymbol{\Phi}^{-1}\right)}{\mathcal{N}\left(\boldsymbol{\beta};\boldsymbol{\mu}',\boldsymbol{\Phi}'^{-1}\right)}=\frac{\left|\boldsymbol{\Phi}\right|^{1/2}}{\left|\boldsymbol{\Phi}'\right|^{1/2}}\exp\left\{\frac{1}{2}\inf_{\boldsymbol{\beta}}\Delta(\boldsymbol{\beta})\right\},$$

where

$$\Delta\left(\boldsymbol{\beta}\right)=(\boldsymbol{\beta}-\boldsymbol{\mu}')^{\top}\boldsymbol{\Phi}'(\boldsymbol{\beta}-\boldsymbol{\mu}')-\left(\boldsymbol{\beta}-\boldsymbol{\mu}\right)^{\top}\boldsymbol{\Phi}(\boldsymbol{\beta}-\boldsymbol{\mu}).$$

The quadratic function $\Delta(\boldsymbol{\beta})$ has a unique global minimum since the Hessian $\partial_{\boldsymbol{\beta}}^{2}\Delta=\boldsymbol{\Phi}'-\boldsymbol{\Phi}$ is positive definite by our assumption. Differentiating $\Delta(\boldsymbol{\beta})$, we see that the minimum occurs at $\hat{\boldsymbol{\beta}}$ such that

$$\boldsymbol{\Phi}'\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\mu}'\right)-\boldsymbol{\Phi}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\mu}\right)=0,\ \text{or equivalently}\ \hat{\boldsymbol{\beta}}=(\boldsymbol{\Phi}'-\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}'\boldsymbol{\mu}'-\boldsymbol{\Phi}\boldsymbol{\mu}).$$

The minimum $\widehat{\Delta}=\Delta\left(\hat{\boldsymbol{\beta}}\right)$ can be expressed as

$$\begin{aligned}
\widehat{\Delta}&=-(\boldsymbol{\mu}'-\boldsymbol{\mu})^{\top}\boldsymbol{\Phi}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\mu}\right)\\
&=-(\boldsymbol{\mu}'-\boldsymbol{\mu})^{\top}\boldsymbol{\Phi}\left[(\boldsymbol{\Phi}'-\boldsymbol{\Phi})^{-1}(\boldsymbol{\Phi}'\boldsymbol{\mu}'-\boldsymbol{\Phi}\boldsymbol{\mu})-\boldsymbol{\mu}\right].
\end{aligned}$$

In the special case $\boldsymbol{\mu}=\boldsymbol{\Phi}^{-1}\boldsymbol{w}$ and $\boldsymbol{\mu}'=\boldsymbol{\Phi}'^{-1}\boldsymbol{w}$, we have

$$\widehat{\Delta}=-(\boldsymbol{\mu}'-\boldsymbol{\mu})^{\top}\boldsymbol{\Phi}\boldsymbol{\mu}=-\left(\boldsymbol{\Phi}'^{-1}\boldsymbol{w}-\boldsymbol{\Phi}^{-1}\boldsymbol{w}\right)^{\top}\boldsymbol{w}=\boldsymbol{w}^{\top}\left(\boldsymbol{\Phi}^{-1}-\boldsymbol{\Phi}'^{-1}\right)\boldsymbol{w}\geq 0,$$

where the last inequality follows from $\boldsymbol{\Phi}^{-1}\succ\boldsymbol{\Phi}'^{-1}$. □

Appendix D: Proof of Proposition 3.10 and 3.11

Proof of Proposition 3.10. Winkelbauer (2012) tells us that a negative moment of Gaussian is given by

$$\mathbb{E}\left|\beta\right|^{-\alpha} = \frac{\Gamma\left(\frac{1-\alpha}{2}\right)}{\frac{\alpha}{2}\sqrt{\pi}}\sigma^{-\alpha}M\left(\frac{\alpha}{2},\frac{1}{2},-\frac{\mu^2}{2\sigma^2}\right),$$

where $M(\cdot,\cdot,\cdot)$ is Kummer's confluent hypergeometric function (see Proposition D.1). To complete the proof, therefore, it suffices to show that $M\left(\frac{\alpha}{2},\frac{1}{2},-\frac{\mu^2}{2\sigma^2}\right)$ is bounded by the smaller of 1 and the function $D(\mu/\sigma)$ as given in (3.14).

Since $\alpha/2 < 1/2$, Proposition D.1 tells us that $M\left(\frac{\alpha}{2},\frac{1}{2},-\frac{\mu^2}{2\sigma^2}\right)$ is bounded by 1 and admits the integral representation

$$M\left(\frac{\alpha}{2},\frac{1}{2},-\frac{\mu^2}{2\sigma^2}\right) = \frac{1}{B\left(\frac{\alpha}{2},\frac{1-\alpha}{2}\right)}\int_0^1\left(1-u\right)^{\frac{1-\alpha}{2}-1}u^{\frac{\alpha}{2}-1}\exp\left(-\frac{\mu^2}{2\sigma^2}u\right)\mathrm{d}u\,. \tag{D.1}$$

To bound the integral, we break up the domain of integration into [0, 1/2] and [1/2, 1] and observe that

$$\begin{aligned}\int_{1/2}^1(1-u)^{\frac{1-\alpha}{2}-1}u^{\frac{\alpha}{2}-1}\exp\left(-\frac{\mu^2}{2\sigma^2}u\right)\mathrm{d}u &\le 2^{1-\frac{\alpha}{2}}\exp\left(-\frac{\mu^2}{4\sigma^2}\right)\int_{1/2}^1(1-u)^{\frac{1-\alpha}{2}-1}\,\mathrm{d}u\\ &= \frac{2^{\frac{5}{2}-\alpha}}{1-\alpha}\exp\left(-\frac{\mu^2}{4\sigma^2}\right),\end{aligned} \tag{D.2}$$

and that

$$\begin{aligned}\int_0^{1/2}(1-u)^{\frac{1-\alpha}{2}-1}u^{\frac{\alpha}{2}-1}\exp\left(-\frac{\mu^2}{2\sigma^2}u\right)\mathrm{d}u &\le 2^{1-\frac{1-\alpha}{2}}\int_0^{1/2}u^{\frac{\alpha}{2}-1}\exp\left(-\frac{\mu^2}{2\sigma^2}u\right)\mathrm{d}u\\ &= 2^{\frac{1+\alpha}{2}}\left(\frac{\mu^2}{2\sigma^2}\right)^{-\frac{\alpha}{2}}\int_0^{\frac{\mu^2}{4\sigma^2}}v^{\frac{\alpha}{2}-1}\exp(-v)\mathrm{d}v\\ &\le 2^{\frac{1+\alpha}{2}}\left(\frac{\mu^2}{2\sigma^2}\right)^{-\frac{\alpha}{2}}\int_0^{\infty}v^{\frac{\alpha}{2}-1}\exp(-v)\mathrm{d}v\\ &= 2^{\frac{1}{2}+\alpha}\left|\frac{\mu}{\sigma}\right|^{-\alpha}\Gamma\left(\frac{\alpha}{2}\right).\end{aligned} \tag{D.3}$$

By (D.1), (D.2), and (D.3), we obtain

$$M\left(\frac{\alpha}{2}, \frac{1}{2}, -\frac{\mu^2}{2\sigma^2}\right) \leq \frac{1}{B\left(\frac{\alpha}{2}, \frac{1-\alpha}{2}\right)}\left[\frac{2^{\frac{5}{2}-\alpha}}{1-\alpha}\exp\left(-\frac{\mu^2}{4\sigma^2}\right) + 2^{\frac{1}{2}+\alpha}\Gamma\left(\frac{\alpha}{2}\right)\left|\frac{\mu}{\sigma}\right|^{-\alpha}\right]$$

□

Proposition D.1. *For* $b > a > 0$, *Kummer's confluent hypergeometric function 1) satisfies the inequality* $M(a, b, z) \leq \max\{1, \exp(z)\}$ *and* 2) *admits the integral representations*

$$M(a, b, z) = \frac{2^{1-b}e^{z/2}}{B(a, b-a)}\int_{-1}^{1}(1-u)^{b-a-1}(1+u)^{a-1}e^{zu/2}\,\mathrm{d}u \tag{D.4}$$

$$= \frac{1}{B(a, b-a)}\int_{0}^{1}(1-u)^{b-a-1}u^{a-1}e^{zu}\,\mathrm{d}u\,. \tag{D.5}$$

Proof. Kummer's function can be represented as the following infinite series (Gradshteyn and Ryzhik 2014, Section 9.210):

$$M\left(a, b, z\right) = 1 + \frac{a}{b}\frac{z}{1!} + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{z^3}{3!} + \ldots\,.$$

Since $b > a > 0$, the series representation immediately implies

$$M(a, b, z) \leq 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots = \exp(z)\,. \tag{D.6}$$

for $z \geq 0$. For $z \leq 0$, we first note that

$$M(a, b, z) = \exp(z)M(b-a, a, -z) \tag{D.7}$$

by the identity (9.212.1) in Gradshteyn and Ryzhik (2014). Since $b > b - a > 0$ and $-z \geq 0$, we can apply our previous bound (D.6) to conclude that $M(b-a, a, -z) \leq \exp(-z)$. Combined with (D.7), this yields $M(a, b, z) \leq 1$ for $z \leq 0$.

The integral representation (D.4) is given in Section 9.211 of Gradshteyn and Ryzhik (2014). To obtain (D.5), we apply the change of variable $v = (1+u)/2$:

$$M\left(a,b,z\right)=\frac{2^{1-b}e^{z/2}}{B(a,b-a)}\int_0^1\left[2\left(1-v\right)\right]^{b-a-1}\left(2v\right)^{a-1}e^{z(2v-1)/2}2\,\mathrm{d}v$$

$$=\frac{1}{B(a,b-a)}\int_0^1\left(1-v\right)^{b-a-1}v^{a-1}e^{zv}\,\mathrm{d}v$$

□

Proof of Proposition 3.11. A conditional precision (in expectation) is always larger than the marginal one, so we have

$$\sigma_j^{-2}\leq\left(\mathbf{\Sigma}^{-1}\right)_{jj}=\zeta^{-2}+\tau^{-2}\lambda_j^{-2}+\sum_{i=1}^n\omega_i x_{ij}^2.$$

Exponentiating both sides of the inequality, we obtain

$$\sigma_j^{-\alpha}\leq\left(\zeta^{-2}+\tau^{-2}\lambda_j^{-2}+\sum_{i=1}^n\omega_i x_{ij}^2\right)^{\alpha/2}$$

$$\leq\zeta^{-\alpha}+\tau^{-\alpha}\lambda_j^{-\alpha}+\left(\sum_{i=1}^n\omega_i x_{ij}^2\right)^{\alpha/2}\tag{D.8}$$

$$\leq\zeta^{-\alpha}+\tau^{-\alpha}\lambda_j^{-\alpha}+1+\frac{\alpha}{2}\left(\sum_{i=1}^n\omega_i x_{ij}^2-1\right),\tag{D.9}$$

where (D.8) follows from the property of L^α-norm $(|a|+|b|)^\alpha\leq|a|^\alpha+|b|^\alpha$ and (D.9) from the Taylor expansion of the concave function $x\to x^\alpha$ at $x=1$. □

Appendix E: Properties of Bayesian bridge prior

Bayesian bridge is characterized by the density of $\beta_j\mid\tau$ given as

$$\pi\left(\beta\mid\tau\right)\propto\tau^{-1}\exp\left(-|\beta/\tau|^a\right).\tag{E.1}$$

We obtain the global-local representation of (E.1) with the conditional $\beta \mid \tau, \lambda \sim \mathcal{N}\left(0, \tau^2\lambda^2\right)$ when

$$\pi_{\text{loc}}\left(\lambda\right) \propto \lambda^{-2}\pi_{\text{st}}\left(\lambda^{-2}/2\right),$$

where $\pi_{\text{st}}(\cdot)$ denote the density of the one-sided stable distribution, characterized by location $\mu = 0$, skewness $\beta = 1$, characteristic exponent $a/2$, and scale $c = \cos(a\pi/4)^{2/a}$ (Hofert, 2011). This follows from the Laplace transform identity for the stable distribution:

$$\begin{aligned}\exp\left(-\left|\beta/\tau\right|^a\right) &= \frac{1}{2}\int_0^\infty \exp\left(-\frac{\phi\beta^2}{2\tau^2}\right)\pi_{\text{st}}\left(\frac{\phi}{2}\right)\mathrm{d}\phi \\ &\propto \int_0^\infty \mathcal{N}\left(\beta; 0, \tau^2\phi^{-1}\right)\pi(\phi)\mathrm{d}\phi,\end{aligned}$$

for $\pi\left(\phi\right) \propto \phi^{-1/2}\pi_{\text{st}}\left(\phi/2\right)$, the density of $\phi = \lambda^{-2}$.

We can characterize the behavior of $\pi_{\text{loc}}(\lambda)$ at $\lambda \approx 0$ from the following asymptotic behavior of the stable distribution as $x \to 0$ (Nolan, 2018).

$$\pi_{\text{st}}\left(x\right) \sim \frac{1}{x^{(1+a)}}\sin\left(\varpi a\right)\frac{\Gamma(a+1)}{\varpi}$$

where $\varpi \approx 3.14159$ is Archimedes' constant. In particular, we have

$$\pi_{\text{loc}}\left(\lambda\right) = O\left(\lambda^{2a}\right) \text{ as } \lambda \to 0\,.$$

The availability of the marginal $\pi(\beta_j \mid \tau) = \int \mathcal{N}\left(\beta_j; 0, \tau^2\lambda_j^2\right)\pi_{\text{loc}}(\lambda_j)\mathrm{d}\lambda_j$ allows for a Gibbs update of τ from the posterior with the local scale parameters λ_j's marginalized out. More precisely, instead of drawing from $\tau \mid \boldsymbol{\beta}, \lambda$, the Bayesian bridge Gibbs sampler can directly target the conditional

$$\pi(\tau \mid \boldsymbol{\beta}) \propto \left(\tau^{-p}\prod_{j=1}^{p}\exp\left(-\left|\frac{\beta_j}{\tau}\right|^a\right)\right)\pi_{\text{glo}}(\tau)\,.$$

Since $\boldsymbol{\beta} \mid \tau$ belongs to the location-scale family, the reference prior is $\pi_{\text{glo}}\left(\tau\right) \propto \tau^{-1}$ (Berger et al., 2015), which also happens to be a conjugate prior. More generally, in terms of the parametrization $\phi = \tau^{-\alpha}$, a prior $\phi \sim \text{Gamma}(\text{shape} = s, \text{rate} = r)$ belongs to a conjugate family, yielding the posterior conditional

$$\pi(\phi \mid \boldsymbol{\beta}) \sim \text{Gamma}\left(\text{shape} = s + p, \text{rate} = r + \textstyle\sum_{j=1}^{p}\left|\beta_j\right|\right).$$

In the limit $s, r \to 0$, the gamma prior on ϕ recovers the reference prior $\pi_{\text{glo}}(\tau) \propto \tau^{-1}$ which is invariant under reparametrization,

Appendix F: Sampler for local scale posterior under horseshoe prior

Our theoretical results on convergence rate assume the ability to sample independently from the conditionals $\lambda_j \mid \beta_j, \tau$ for $j = 1, \ldots, p$. While not necessarily trivial, this task is typically quite manageable given the wide range of algorithms available to deal with univariate distributions (Devroye, 2006; Ripley, 2009).

As an illustration, we present a simple rejection sampler for the conditional $\lambda_j \mid \beta_j, \tau$ under the prior $\pi_{\text{loc}}(\lambda_j) \propto 1/(1+\lambda_j^2)$ — corresponding to the horseshoe prior, arguably the most popular of the existing shrinkage priors (Bhadra et al., 2017). The rejection sampler, as we will show, has uniformly high acceptance probability for all β_j and τ with the minimum acceptance probability ≈ 0.6975 (Figure F.3). On the precision scale $\eta_j = \lambda_j^{-2}$, the prior is given by

$$\pi_{\text{loc}}(\eta_j) = \pi_{\text{loc}}(\lambda_j)\left|\mathrm{d}\lambda/\mathrm{d}\eta_j\right| \propto \frac{1}{1+\eta_j^{-1}}\eta_j^{-3/2} = \frac{1}{\eta_j^{1/2}(1+\eta_j)}.$$

The full conditional $\eta_j \mid \beta_j, \tau$ has the density

$$\pi(\eta_j \mid \beta_j, \tau) \propto \pi_{\text{loc}}(\eta_j)\pi(\beta_j \mid \tau, \eta_j) \propto \frac{1}{1+\eta_j}\exp\left(-\eta_j\frac{\beta_j^2}{2\tau^2}\right).$$

The task of sampling from the local scale posterior, therefore, boils down to that of sampling from the family of univariate densities

$$\pi(\eta) \propto \frac{1}{1+\eta}\exp(-b\eta) \text{for } b > 0\,. \tag{F.1}$$

To sample from (F.1), the online supplement of Polson et al. (2014) describes a slice sampling approach and Makalic and Schmidt (2015) a data augmentation method. However, we find that both approaches suffer from slow-mixing as $b \to 0$ and the slow-decaying term $(1+\eta)^{-1}$ becomes significant (Figure F.1 and F.2).

F. 1 Rejection sampler algorithm

Our rejection sampler acts on a transformed parameter $\psi = \log(1+\eta)$ that maps back as $\eta = e^{\psi} - 1$. The density of ψ is given by

$$\pi(\psi) \propto \pi(\eta)\left|\mathrm{d}\eta/\mathrm{d}\psi\right| = \frac{1}{e^{\psi}}\exp(-be^{\psi})e^{\psi} = \exp(-be^{\psi}) \text{ on } \psi \geq 0\,.$$

We now define a function g_b that upper bounds the unnormalized target density

$$f_b(\psi) \coloneqq \exp(-be^{\psi}).$$

For $b \geq 1$, we set

$$g_b(\psi) = \exp\{-b(1+\psi)\},$$

which coincides with an unnormalized density of the distribution exp(rate $= b$). For $b < 1$, we set

$$g_b(\psi) = \begin{cases} \exp(-b) & \text{for } \psi \leq \log(1/b) \\ \exp\{-1-(\psi-\log(1/b))\} & \text{for } \psi \geq \log(1/b) \end{cases},$$

which coincides with an unnormalized density of a mixture of Uniform $(0, \log(1/b))$ and Exp(1) shifted by $\log(1/b)$. To draw a random variable X from this mixture, we set $X \sim$Uniform $(0, \log(1/b))$ with probability $\log(1/b)/(\log(1/b) + e^{b-1})$ and $X - \log(1/b) \sim$exp(1) otherwise. R and Python code of the rejection sampler are available at https://github.com/aki-nishimura/horseshoe-scale-sampler.

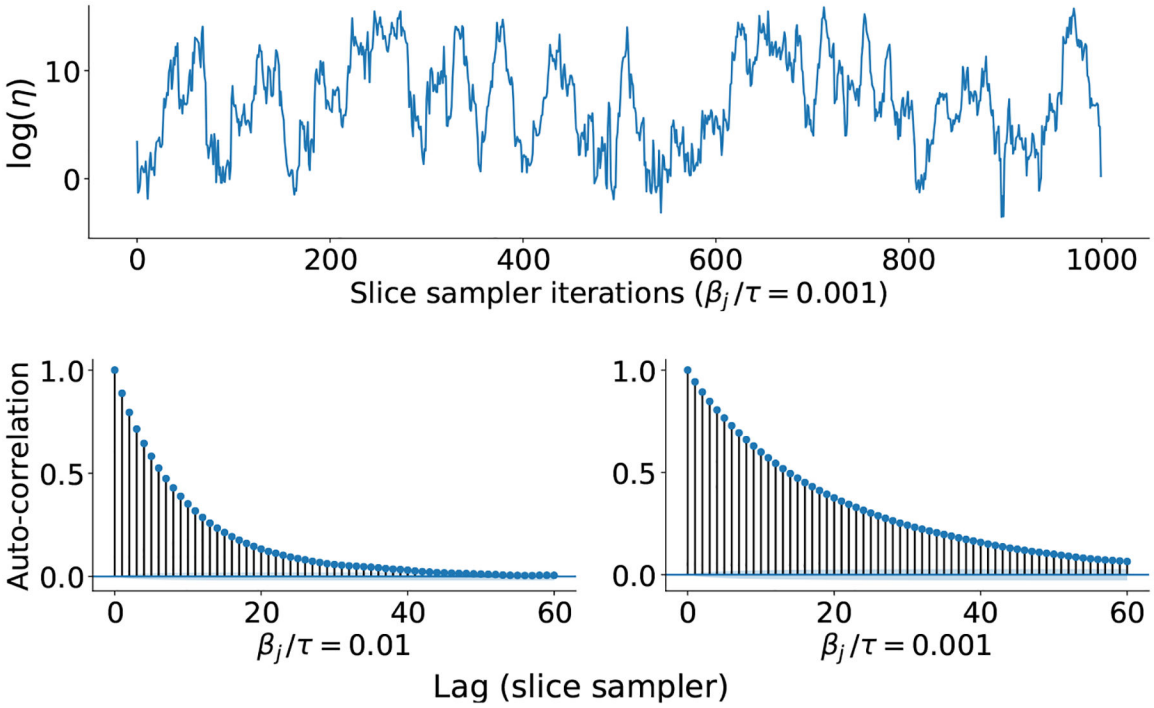


Figure F.1:
Trace and auto-correlation plots when slice sampling η from (F.1) as proposed in Polson et al. (2014). For the two different values of $b = \beta_j^2/2\tau^2$, the auto-correlations at stationarity are computed from 10,000 iterations of the sampler to demonstrate how the mixing rate degrades as $b \to 0$.

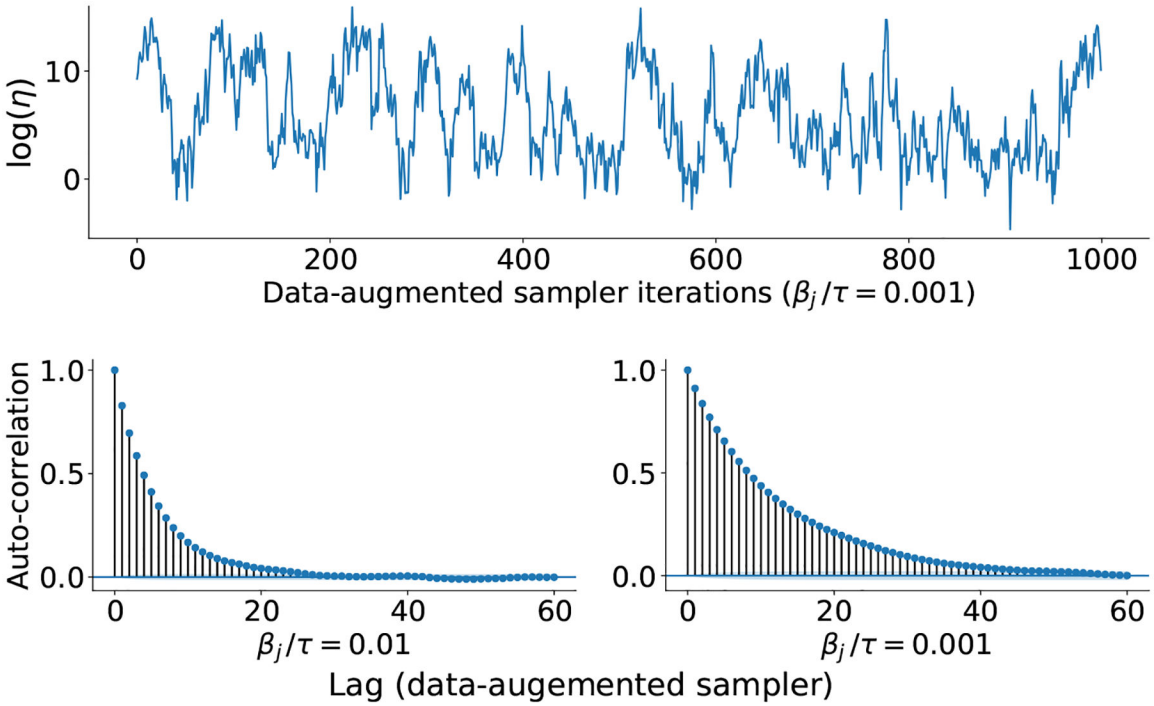


Figure F.2:

Trace and auto-correlation plots when sampling η from (F.1) with the data-augmentation scheme of Makalic and Schmidt (2015). The auto-correlations at stationarity are computed from 10,000 iterations of the sampler.

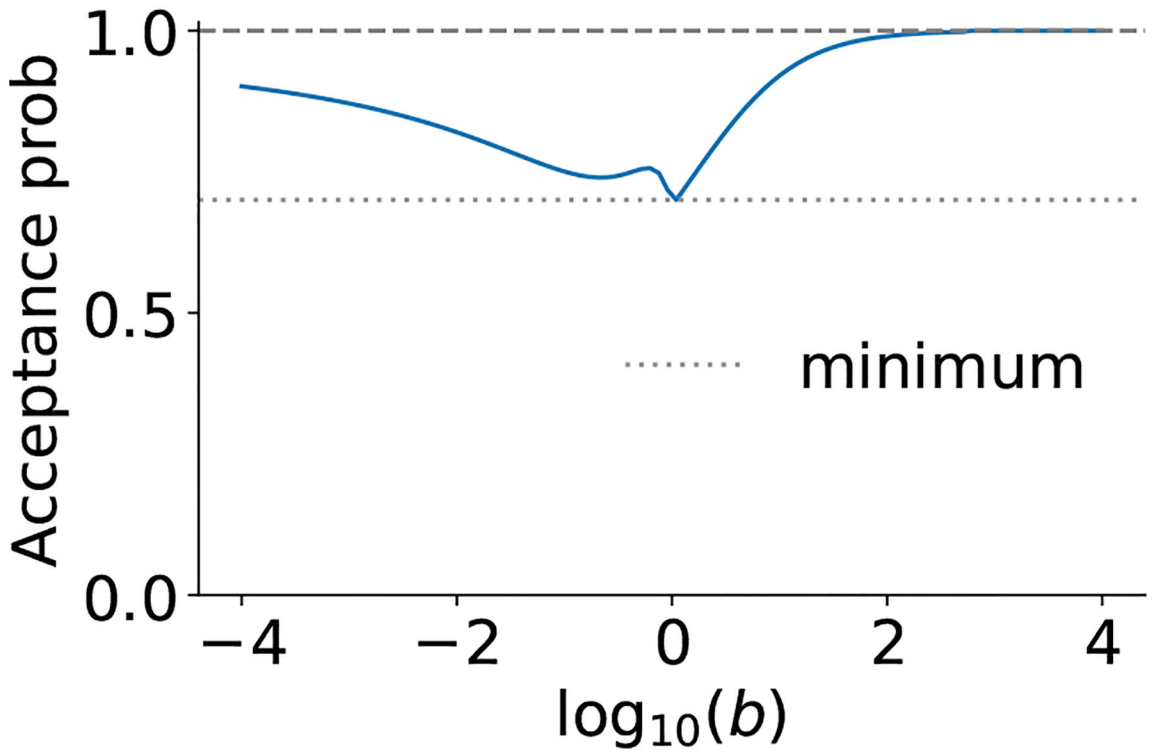


Figure F.3:
Acceptance probability of the proposed rejection sampler as a function of $b = \beta_j^2/2\tau^2$. The probability is uniformly lower-bounded and increases to 1 as $b \to 0$ and $b \to \infty$ (see Theorem F.1). The minimum probability is ≈ 0.6975.

F. 2 Analysis of acceptance probability

The acceptance probability of a rejection sampler is given by the ratio of the integrals of the target to the bounding density (Ripley, 2009). In particular, the rejection sampler described in Section F.1 has the acceptance probability

$$A\left(b\right) = \frac{\int_0^\infty f_b(\eta)\mathrm{d}\eta}{\int_0^\infty g_b(\eta)\mathrm{d}\eta}. \tag{F.2}$$

Figure F.2 plots the acceptance probability $A(b)$, evaluated to high accuracy via numerical integration of the integrals in (F.2), and supports the theoretical results below.

Theorem F.1. *The acceptance probability* $A(b)$ *is uniformly lower bounded over* $b > 0$ *by a positive constant. Moreover,* $A(b)$ converges to 1 *as* $b \to 0$ *and* $b \to \infty$.

Proof. We can show that both the denominator and numerator of (F.2) depend continuously on b, and so does $A(b)$, by a simple application of the dominated convergence theorem. The

continuity of $A(b)$ implies a uniform lower bound on $b \in (0, \infty)$ as soon as we establish $A(b) \to 1$ towards the boundary $b \to 0$ and $b \to \infty$.

We establish a lower bound on the acceptance probability (F.2) by explicitly computing the denominator and then lower bounding the numerator. We first consider the case $b \geq 1$, when the denominator is given by

$$\int_0^\infty g_b(\eta)\mathrm{d}\psi = \int_0^\infty \exp\{-b(1+\psi)\}\mathrm{d}\psi = b^{-1}e^{-b}. \tag{F.3}$$

Then, using Taylor's theorem and the fact $\frac{\mathrm{d}^2}{\mathrm{d}\psi^2}e^\psi = e^\psi$, we have

$$0 \leq e^\psi - \left(1+\psi\right) \leq \psi^2 \max_{\psi' \in [0,\psi]} e^{\psi'} = \psi^2 e^\psi.$$

The above inequality in particular implies that

$$f_b(\psi) = \exp(-be^\psi) \geq \exp\{-b(1+\psi)\}\exp\left(-b\psi^2 e^\psi\right). \tag{F.4}$$

We now apply (F.4) to lower bound the numerator of (F.2); for any $L > 0$,

$$\begin{aligned}\int_0^\infty \exp(-be^\psi)\mathrm{d}\psi &\geq \int_0^L \exp\{-b(1+\psi)\}\exp\left(-b\psi^2 e^\psi\right)\mathrm{d}\psi \\ &\geq \exp\left(-bL^2 e^L\right)\int_0^L \exp\{-b(1+\psi)\}\mathrm{d}\psi \\ &= b^{-1}e^{-b}\exp\left(-bL^2 e^L\right)\left(1-e^{-bL}\right).\end{aligned} \tag{F.5}$$

From (F.3) and (F.5), we obtain the following lower bound on the acceptance probability, which holds for any $L > 0$:

$$A(b) \geq \exp\left(-bL^2 e^L\right)\left(1-e^{-bL}\right).$$

Choosing $L = \log(\kappa b)/b$ with $\kappa > 1$, for example, we obtain the lower bound

$$A\left(b\right) \geq \exp\left(-\frac{(\log\kappa b)^2}{b}\kappa^{1/b}b^{1/b}\right)\left(1-\frac{1}{\kappa b}\right). \tag{F.6}$$

It is straightforward to show that, for example by the derivative test, the function $b \to b^{1/b}$ has the global maximum $\exp\left(e^{-1}\right)$ on $b > 0$. We can therefore simplify the lower bound (F.6) to

$$A(b) \geq \exp\left(-\exp\left(e^{-1}\right)\kappa\frac{1}{b}\frac{(\log\kappa b)^2}{b}\right)\left(1-\frac{1}{\kappa b}\right). \tag{F.7}$$

The lower bound in (F.7), and hence $A(b)$, converges to 1 as $b \to \infty$.

We now turn to establishing a lower bound on the acceptance probability in the case $b < 1$. We have

$$\begin{aligned}\int_0^\infty g_b(\psi)\mathrm{d}\psi &= \int_0^{\log(1/b)} e^{-b}\,\mathrm{d}\psi + \int_{\log(1/b)}^\infty \exp\left\{-1-\left(\psi+\log b\right)\right\}\mathrm{d}\psi \\ &= e^{-b}\log\left(1/b\right) + e^{-1}.\end{aligned} \tag{F.8}$$

To lower bound $\int f_b(\psi)\mathrm{d}\psi$, we first observe that, by the change of variable $\psi' = \psi/\log(1/b)$

$$\int_0^{\log(1/b)} \exp(-be^{\psi})\mathrm{d}\psi = \log\left(\frac{1}{b}\right)C(b) \text{ where } C(b) = \int_0^1 \exp\left(-b^{1-\psi'}\right)\mathrm{d}\psi'. \tag{F.9}$$

On the interval $\psi' \in [0,1)$, the integrand converges to 1 as $b \to 0$ and hence the dominated convergence theorem implies $C(b) \to 1$ as $b \to 0$. On the interval $\psi \in [\log(1/b), \infty)$, we have

$$\begin{aligned}&\int_{\log(1/b)}^\infty \exp(-be^{\psi})\mathrm{d}\psi = \int_{\log(1/b)}^\infty \exp\left\{-be^{\log(1/b)}e^{\psi}-\log(1/b)\right\}\mathrm{d}\psi = \int_0^\infty \exp(-e^{\psi})\mathrm{d}\psi \geq e^{-1}C'(\kappa) \text{ for } C'(\kappa) \\ &= \exp\left(-(\log\kappa)^2\kappa\right)\left(1-\frac{1}{\kappa}\right),\end{aligned} \tag{F.10}$$

where the last inequality follows from (F.5) with $b = 1$ and $L = \log(\kappa)$ for $\kappa > 1$. It follows from (F.8), (F.9), and (F.10) that for $b < 1$

$$A\left(b\right) \geq \frac{\log\left(1/b\right)C\left(b\right)+e^{-1}C'\left(\kappa\right)}{e^{-b}\log\left(1/b\right)+e^{-1}}, \tag{F.11}$$

where $\lim_{b\to 0}C(b) = 1$ and $C'(\kappa) \approx 0.264$ for $\kappa = 1.57$. The lower bound in (F.11), and hence $A(b)$, converges to 1 as $b \to 0$. □

References

Abramowitz M and Stegun I (1965). Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables. Applied mathematics series. Dover Publications.

Albert JH and Chib S (1993). "Bayesian analysis of binary and polychotomous response data." Journal of the American statistical Association, 88(422): 669–679.

Berger JO, Bernardo JM, and Sun D (2015). "Overall objective priors." Bayesian Analysis, 10(1): 189–221.

Bhadra A, Datta J, Polson NG, and Willard BT (2017). "Lasso Meets Horseshoe." arXiv:1706.10179

Carvalho CM, Polson NG, and Scott JG (2009). "Handling sparsity via the horseshoe." In Artificial Intelligence and Statistics, 73–80.

Carvalho CM, Polson NG, and Scott JG (2010). "The horseshoe estimator for sparse signals." Biometrika, 97(2): 465–480.

Choi HM and Hobert JP (2013). "The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic." Electronic Journal of Statistics, 7: 2054–2064.

Devroye L (2006). "Nonuniform random variate generation." In Handbooks in Operations Research and Management Science, volume 13, 83–121. Elsevier.

Durante D (2019). "Conjugate Bayes for probit regression via unified skew-normal distributions." Biometrika, 106(4): 765–779.

Flegal JM and Jones GL (2011). "Implementing MCMC: estimating with confidence." In Brooks S, Gelman A, Jones G, and Meng X-L (eds.), Handbook of Markov chain Monte Carlo, 175–197. CRC Press.

Gautschi W (1959). "Some elementary inequalities relating to the gamma and incomplete gamma function." Journal of Mathematics and Physics, 38(1–4): 77–81.

Ghosh J, Li Y, and Mitra R (2018). "On the use of Cauchy prior distributions for Bayesian logistic regression." Bayesian Analysis, 13(2): 359–383.

Ghosh P and Chakrabarti A (2017). "Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems." Bayesian Analysis, 12(4): 1133–1161.

Golub GH and Van Loan CF (2012). Matrix Computations, volume 3. Johns Hopkins University Press.

Gradshteyn IS and Ryzhik IM (2014). Table of integrals, series, and products. Academic press.

Greenland S, Mansournia MA, and Altman DG (2016). "Sparse data bias: a problem hiding in plain sight." bmj, 352.

Griffin JE and Brown PJ (2010). "Inference with normal-gamma prior distributions in regression problems." Bayesian Analysis, 5(1): 171–188.

Hastie T, Tibshirani R, and Friedman J (2009). The Elements of Statistical Learning. Springer Series in Statistics. Springer.

Hofert M (2011). "Sampling exponentially tilted stable distributions." ACM Transactions on Modeling and Computer Simulation, 22(1): 3.

Johndrow JE, Orenstein P, and Bhattacharya A (2018). "Bayes Shrinkage at GWAS scale: Convergence and Approximation Theory of a Scalable MCMC Algorithm for the Horseshoe Prior." arXiv:1705.00841

Jones GL and Hobert JP (2001). "Honest exploration of intractable probability distributions via Markov chain Monte Carlo." Statistical Science, 312–334.

Kastner G (2019). "Sparse Bayesian time-varying covariance estimation in many dimensions." Journal of Econometrics, 210(1): 98–115.

Kowal DR, Matteson DS, and Ruppert D (2019). "Dynamic shrinkage processes." Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(4): 781–804.

Li K-C (1989). "Honest confidence regions for nonparametric regression." The Annals of Statistics, 17(3): 1001–1008.

Li Y, Craig BA, and Bhadra A (2019). "The graphical horseshoe estimator for inverse covariance matrices." Journal of Computational and Graphical Statistics, 28(3): 747–757.

Louizos C, Ullrich K, and Welling M (2017). "Bayesian compression for deep learning." In Advances in neural information processing systems, 3288–3298.

Makalic E and Schmidt DF (2015). "A simple sampler for the horseshoe estimator." IEEE Signal Processing Letters, 23(1): 179–182.

Meyn S and Tweedie RL (2009). Markov Chains and Stochastic Stability. New York, NY, USA: Cambridge University Press.

Nishimura A and Suchard MA (2018). "Prior-preconditioned conjugate gradient for accelerated Gibbs sampling in" large n & large p" sparse Bayesian logistic regression models." arXiv:1810.12437

Nolan JP (2018). Stable Distributions - Models for Heavy Tailed Data. Boston: Birkhauser.

Pal S and Khare K (2014). "Geometric ergodicity for Bayesian shrinkage models." Electronic Journal of Statistics, 8(1): 604–645.

Park T and Casella G (2008). "The Bayesian lasso." Journal of the American Statistical Association, 103(482): 681–686.

Piironen J and Vehtari A (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors." Electronic Journal of Statistics, 11(2): 5018–5051.

Polson NG and Scott JG (2010). "Shrink globally, act locally: Sparse Bayesian regularization and prediction." Bayesian Statistics, 9: 501–538.

Polson NG, Scott JG, and Windle J (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." Journal of the American Statistical Association, 108(504): 1339–1349.

Polson NG, Scott JG, and Windle J (2014). "The Bayesian bridge." Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(4): 713–733.

Ripley BD (2009). Stochastic simulation, volume 316. John Wiley & Sons.

Roberts GO and Rosenthal JS (2001). "Markov chains and de-initializing processes." Scandinavian Journal of Statistics, 28(3): 489–504.

Roberts GO and Rosenthal JS (2004). "General state space Markov chains and MCMC algorithms." Probability Surveys, 1: 20–71.

Rosenthal JS (1995). "Minorization conditions and convergence rates for Markov chain Monte Carlo." Journal of the American Statistical Association, 90(430): 558–566.

Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, and Suchard MA (2018). "Improving reproducibility by using high-throughput observational studies with empirical calibration." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2128): 20170356.

Tian Y, Schuemie MJ, and Suchard MA (2018). "Evaluating large-scale propensity score performance through real-world and synthetic data experiments." International Journal of Epidemiology.

van der Pas S, Salomond J-B, and Schmidt-Hieber J (2016). "Conditions for posterior contraction in the sparse normal means problem." Electronic journal of statistics, 10(1): 976–1000.

van der Pas S, Szabó B, and van der Vaart A (2017). "Adaptive posterior contraction rates for the horseshoe." Electronic Journal of Statistics, 11(2): 3196–3225.

Wang X and Roy V (2018). "Geometric ergodicity of Pólya-Gamma Gibbs sampler for Bayesian logistic regression with a flat prior." Electronic Journal of Statistics, 12(2): 3295–3311.

Winkelbauer A (2012). "Moments and absolute moments of the normal distribution." arXiv:1209.4340