Journal of Machine Learning Research 21 (2020) 1-61 Submitted 6/19; Revised 12/19; Published 3/20

Scalable Approximate MCMC Algorithms for the Horseshoe Prior

James Johndrow JOHNDROW@WHARTON.UPENN.EDU
Department of Statistics
Stanford University
Stanford, CA, 94305, USA

Paulo Orenstein PAULOO@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA, 94305, USA

Anirban Bhattacharya ANIRBANB@STAT.TAMU.EDU
Department of Statistics
Texas A& M University
College Station, TX 77843-3143, USA

Editor: Francois Caron

Abstract

The horseshoe prior is frequently employed in Bayesian analysis of high-dimensional models, and has been shown to achieve minimax optimal risk properties when the truth is sparse. While optimization-based algorithms for the extremely popular Lasso and elastic net procedures can scale to dimension in the hundreds of thousands, algorithms for the horseshoe that use Markov chain Monte Carlo (MCMC) for computation are limited to problems an order of magnitude smaller. This is due to high computational cost per step and growth of the variance of time-averaging estimators as a function of dimension. We propose two new MCMC algorithms for computation in these models that have significantly improved performance compared to existing alternatives. One of the algorithms also approximates an expensive matrix product to give orders of magnitude speedup in high-dimensional applications. We prove guarantees for the accuracy of the approximate algorithm, and show that gradually decreasing the approximation error as the chain extends results in an exact algorithm. The scalability of the algorithm is illustrated in simulations with problem size as large as $N = 5,000$ observations and $p = 50,000$ predictors, and an application to a genome-wide association study with $N = 2,267$ and $p = 98,385$. The empirical results also show that the new algorithm yields estimates with lower mean squared error, intervals with better coverage, and elucidates features of the posterior that were often missed by previous algorithms in high dimensions, including bimodality of posterior marginals indicating uncertainty about which covariates belong in the model.

Keywords: Bayesian; High dimensional; MCMC approximation; Perturbation theory; Shrinkage prior.

1. Introduction

Approximate Markov chain Monte Carlo (aMCMC) methods are increasingly popular in Bayesian analysis of big data problems. While MCMC algorithms remain one of the stan-

dard computational approaches in Bayesian statistics, their implementation can be prohibitively slow in large scale applications, contributed in part by expensive matrix calculations, likelihood evaluations and sampling operations at each iteration. The basic idea behind aMCMC is to create a computationally amenable approximation/perturbation $\mathcal{P}_\epsilon$ to an exact[1] Markov transition $\mathcal{P}$ so that extending the approximate chain $\mathcal{P}_\epsilon$ by one step requires substantially less computational effort than $\mathcal{P}$. Accordingly, perturbation theory has been a recent focus of the theoretical MCMC literature (Johndrow and Mattingly, 2017; Rudolf and Schweizer, 2018; Pillai and Smith, 2014), as well as algorithm development (Bardenet et al., 2017; Korattikara et al., 2014; Welling and Teh, 2011). Earlier examples of perturbation theory for Markov chains under stronger ergodicity conditions include Mitrophanov (2005) and Roberts et al. (1998). This theoretical literature provides conditions under which finite-length paths from the approximate kernel $\mathcal{P}_\epsilon$ give provably good approximations to the posterior. This approach is attractive from at least two perspectives: (1) it suggests the possibility of overcoming computational challenges for Bayesian inference in big data settings by replacing computational bottlenecks with faster numerical approximations, and (2) it allows practitioners to move beyond the setting of choosing a $\mathcal{P}$ that has exactly the "right" invariant measure from a set of alternatives that in practice is quite small.

The practical success of aMCMC has been mainly limited to applications involving very large sample sizes N and relatively modest number of parameters p. Recent activity has focused on using subsamples or "minibatches" of data in the large N setting to create an analogue of stochastic gradient methods for MCMC. As Bardenet et al. (2017) point out, achieving provably good approximations with significant computational advantage using subsampling typically requires the posterior to be well-approximated by a Gaussian, which is unlikely to be the case in large p applications. Accurate approximations using minibatches typically require the construction of control variates (Pollock et al., 2016; Baker et al., 2017; Bardenet et al., 2017), which in practice can be time-consuming, particularly when the target is high-dimensional and near-sparse in most directions, and the important directions are not known *a priori*. However, this is precisely the type of target distribution that one encounters in high-dimensional sparse regression problems, the object of interest in this paper.

Modern applications in genetics and other areas of biology have stimulated considerable interest in statistical inference in the high-dimensional setting where the number of predictors p is much larger than the number of observations N, and the truth is thought to be sparse or consist mostly of small signals. Regression models are frequently employed in this context. Consider a Gaussian linear model with likelihood

$$L(z \mid W\beta, \sigma^2) = (2\pi\sigma^2)^{-N/2} e^{-\frac{1}{2\sigma^2}(z-W\beta)'(z-W\beta)}, \tag{1}$$

where W is a $N \times p$ matrix of covariates, $\beta \in \mathbb{R}^p$ is assumed to be a sparse vector, and $z \in \mathbb{R}^N$ is an N-vector of response observations. A common hierarchical Bayesian approach employs a Gaussian scale-mixture prior on β of the form

$$\begin{aligned} \beta_j \mid \sigma^2, \eta, \xi &\overset{iid}{\sim} \mathrm{N}(0, \sigma^2\xi^{-1}\eta_j^{-1}), \quad \eta_j^{-1/2} \overset{iid}{\sim} \upsilon_L, \quad j = 1, \ldots, p, \\ \xi^{-1/2} &\sim \upsilon_G, \quad \sigma^2 \sim \mathrm{InvGamma}(\omega/2, \omega/2), \end{aligned} \tag{2}$$

1. That is, with the posterior distribution as its stationary distribution

where υ_L and υ_G are densities on $\mathbb{R}_+$. The prior structure (2) induces approximate sparsity in β by shrinking the null components aggressively toward zero while retaining the true signals (Polson and Scott, 2010). The global precision parameter ξ controls the number of signals, while the local precisions η_j dictate whether they are nulls. In this sense, the prior (2) approximates the properties of point-mass mixture priors (Johnson and Rossell, 2012; Scott and Berger, 2010; George and McCulloch, 1997), which allow some components of β to be exactly zero a posteriori.

While the core idea behind the algorithmic developments in this paper apply broadly to the class of global-local priors (2), see Polson and Scott (2010) and Bhattacharya et al. (2016) for a comprehensive list of such priors, for sake of concreteness we focus here on the popular horseshoe prior (Carvalho et al., 2010) which corresponds to both υ_L and υ_G being the half-Cauchy distribution. In the normal means setting, where $W = I_N$ (the $N \times N$ identity matrix), the horseshoe achieves the minimax adaptive rate of contraction when the true β is sparse (Van Der Pas et al., 2014; van der Pas et al., 2017a). Moreover, the marginal credible intervals have asymptotically correct frequentist coverage van der Pas et al. (2017b) for parameters that are either very close to zero or above the detection threshold, though signals in a certain "intermediate" range are shrunk too much toward zero for credible intervals to have correct coverage. Although early literature on the horseshoe prior justified it as a continuous approximation of the point-mass mixture prior, over time it has come to be recognized as a good default prior choice in high-dimensional settings in its own right, see Bhadra et al. (2017).

Despite the popularity of the horseshoe in the literature, there is a lack of MCMC algorithms that scale to large (N, p), owing to expensive linear algebra and slow mixing of the corresponding Markov chain. The current state-of-the-art algorithm for large p is Bhattacharya et al. (2016), which has only been employed successfully up to about $p = 10,000$, while the recently proposed algorithm of Hahn et al. (2018) scales very well in N but is less efficient than the exact algorithm we propose here when $p \gg N$ (see (Hahn et al., 2018, Section 3)). Another recent proposal is Makalic and Schmidt (2016), but it has only been compared to the implementation in the `monomvn` package for `R`, which is very slow relative to Bhattacharya et al. (2016). The lack of scalable algorithms has kept a useful model designed for high-dimensional regression out of many modern high-dimensional applications such as genome-wide association studies (GWAS), which often have N in the thousands and p in the hundreds of thousands or more.

In this paper, we develop an approximation scheme for the horseshoe posterior which is not based on subsampling and yet produces orders of magnitude speed-ups in large (N, p) settings. In addition to the issues with control variates mentioned earlier, another issue with subsampling in $p \gg N$ settings is that the posteriors within each sub-sample can be extremely noisy. Our approach is fundamentally different in that it relies on learning and exploiting the structural sparsity of the posterior to reduce the cost per step. Specifically, we make fast approximations to several matrix products exploiting the sparsity structure. These matrix approximations take place within a new MCMC algorithm for the horseshoe that exhibits faster mixing than existing algorithms and unearths subtle features of the posterior. The details of the exact and approximate algorithm are provided in Section 2. This is one of the first demonstrations we are aware of in which the perturbation strategy has resulted in a practically significant algorithmic advance in the high-dimensional set-

ting when the posterior is not remotely close to a Gaussian. In particular, the horseshoe posterior differs from a Gaussian not only in the tails, but also in its "center," since the posterior will often have many modes. Because the critical feature of our algorithm is exploitation of sparsity, a similar strategy could succeed in other canonical high-dimensional Bayesian models with an underlying lower-dimensional structure. One existing example of an approximate MCMC algorithm that exploits sparsity is the "skinny Gibbs" algorithm of Narisetty et al. (2018), which uses a thresholding-based approximate Gibbs sampler for spike-and-slab priors.

We prove bounds on the approximation error to the posterior both for invariant measures of the approximate algorithm and for finite-time Cesáro averages in Section 3. These results utilize and expand on recent work on perturbation theory for geometrically ergodic Markov chains. In the process, we prove a general lemma, Lemma 4, showing that one can typically work in unweighted metrics and nonetheless obtain an approximation error bound in the metric weighted by a Lyapunov function, which substantially reduces the effort needed to establish guarantees of approximation accuracy for large classes of unbounded functions. We further prove a new result that if one gradually reduces the approximation error as the chain extends, it is possible to construct exact algorithms that utilize only approximate transition kernels. These latter results are very general and apply to a wide array of algorithms constructed from approximate kernels beyond the immediate application considered here.

We complement our theoretical analysis with a detailed empirical study in $p \gg N$ settings which confirms that the approximation is empirically very accurate and has orders of magnitude lower computational cost per step than the exact algorithm. To analyze the approximation accuracy, we compute various metrics to compare the exact and approximate chain in Section 4. These include correlations between marginal means/variances of the β_js obtained from the exact and approximate chains as well as an average Kolmogorov–Smirnov distance between the marginal distributions of the β_js from the two chains. In addition to certifying the accuracy of the approximation, these empirical exercises provide a practical way of choosing the (only) threshold that appears in our approximation. To compare the overall computational complexity of the two algorithms, we consider the median effective sample size per second across an illustrative subset of the set of parameters. This metric combines the computational cost arising from extending the respective chains by one-step along with the amount of correlation in either chain. Based on this metric, the approximate algorithm is shown to be 50 times more efficient than the exact algorithm when $N = 2000$ and $p = 20,000$.

We conclude by utilizing the approximate algorithm to estimate the horseshoe on a GWAS dataset with $N = 2,267, p = 98,385$, which is an order of magnitude higher dimensional than the datasets considered by Bhattacharya et al. (2016). We compare these results to point estimates obtained using the Lasso, and show that while there is broad agreement in which variables are important, the horseshoe estimated using our approximate algorithm exhibits the expected behavior of shrinking the larger signals less and the smaller signals more than Lasso. We also show that the our approximate algorithm more accurately recovers nuanced features of the posterior compared to the exact algorithm of Bhattacharya et al. (2016), such as bimodality of marginals when the true signal is near the minimax threshold of detection. These bimodal marginals indicate uncertainty about which variables belong in

the model, which is an often-touted argument for the use of Bayesian procedures compared to frequentist methods such as the Lasso which return only a single selected model.

While this article was under preparation, we came across an interesting preprint by Nishimura and Suchard (2018) who use preconditioned conjugate gradient (pCG) methods to speed up otherwise expensive linear algebra calculations within an MCMC algorithm for high-dimensional logistic regression using the bridge prior of Polson et al. (2014). The usage of pCG algorithms is arguably underutilized in the MCMC literature, and using pCG in conjunction with our approximation scheme can potentially widen the scope of application for either algorithm. We leave this exploration for future work.

2. Algorithms

We begin by describing the update rules of an exact blocked Metropolis-within-Gibbs algorithm targeting the horseshoe posterior. This exact algorithm is new, though it is related to the algorithm of (Polson et al., 2014, Supplement) and that of Bhattacharya et al. (2016). The main motivation is to improve the mixing of the global parameter ξ, and we achieve that by making extensive use of block updating. For sake of brevity, we suppress dependence on z and W in the full conditionals of the state variables.

2.1. Exact Algorithm

We first define some quantities that will be used repeatedly. Let

$$\begin{aligned} D = \mathrm{diag}(\eta_j^{-1}), \quad M_\xi = I_N + \xi^{-1} W D W' \\ p(\xi \mid \eta) = |M_\xi|^{-1/2} \left(\frac{\omega}{2} + \frac{1}{2} z' M_\xi^{-1} z \right)^{-(N+\omega)/2} \frac{1}{\sqrt{\xi}(1+\xi)}. \end{aligned} \tag{3}$$

A blocked Metropolis-within-Gibbs algorithm that targets the exact horseshoe posterior is given by the update rule

$$\begin{aligned} &1.\ \text{sample } \eta \sim p(\eta \mid \xi, \beta, \sigma^2) \propto \prod_{j=1}^{p} \frac{1}{1+\eta_j} e^{-\frac{\beta_j^2 \xi \eta_j}{2\sigma^2}}. \\ &2.\ \text{propose } \log(\xi^*) \sim \mathrm{N}(\log(\xi), s), \text{ accept } \xi \text{ w.p. } \frac{p(\xi^* \mid \eta)\xi^*}{p(\xi \mid \eta)\xi}. \\ &3.\ \text{sample } \sigma^2 \mid \eta, \xi \sim \mathrm{InvGamma}\left(\frac{\omega + N}{2}, \frac{\omega + z' M_\xi^{-1} z}{2} \right). \\ &4.\ \text{sample } \beta \mid \eta, \xi, \sigma^2 \sim \mathrm{N}\left((W'W + (\xi^{-1}D)^{-1})^{-1} W' z, \sigma^2 (W'W + (\xi^{-1}D)^{-1})^{-1} \right). \end{aligned} \tag{4}$$

We refer generically to the Markov transition operator defined by this update rule as $\mathcal{P}$. The η_js in step 1 are independently sampled using the rejection sampler described in Section S1 of the Supplemental document.

The algorithm in (4) differs from Polson et al. (2014) in that the second step targets $p(\xi \mid \eta)$ rather than $p(\xi \mid \eta, \beta, \sigma^2)$ as in Polson et al. (2014), and thus blocks together (β, σ^2, ξ) instead of only (β, σ^2). It also differs from Bhattacharya et al. (2016), which did

not do any blocking of β, σ^2, ξ. Moreover, whereas Polson et al. (2014) and Bhattacharya et al. (2016) used slice sampling targeting $p(\eta \mid \xi, \beta, \sigma^2)$, we develop an exact rejection sampler to sample the η_js independently. The rejection sampler exploits that the full conditional density of η_j is log-convex to build a piecewise upper envelope which can be conveniently sampled from, with careful choices of the pieces ensuring very high acceptance rates. Being able to sample the η_js exactly is convenient as it avoids the introduction of additional p latent variables in the slice sampler, and also simplifies the convergence analysis of the Markov chain.

Like Bhattacharya et al. (2016), we use an efficient method for sampling from the Gaussian full conditional for β. The details of this method are relevant for understanding our approximate sampler, so we briefly summarize it here. To sample from $\beta \mid \eta, \xi, \sigma^2$, the following three steps suffice

$$\begin{aligned} &\text{sample } u \sim \mathrm{N}(0, \xi^{-1}D) \text{ and } f \sim \mathrm{N}(0, I_N) \text{ independently} \\ &\quad \text{set } v = Wu + f, \quad v^* = M_\xi^{-1}(z/\sigma - v), \\ &\quad \text{set } \beta = \sigma(u + \xi^{-1}DW'v^*). \end{aligned} \tag{5}$$

Notice that this algorithm – and indeed, all but one step of (4) – requires computing M_ξ defined in (3) and solving $M_\xi v^* = (z/\sigma - v)$ for v^*. When p is large, the computational bottleneck of the algorithm in (4) is, perhaps surprisingly, just computing the matrix WDW', which is needed to compute M_ξ. This has computational cost N^2p, which dominates every other calculation in the algorithm when $p > N$. In the next section, we propose an approximate sampler that has lower computational cost per step. Our approximate algorithm is designed for the case where $p > N$, and it is in these settings where it offers very large performance gains. As such, our sole focus in this paper is the $p > N$ setting. While the exact algorithm we propose could be modified to scale linearly in N rather than linearly in p to improve its performance in the $N > p$ case, the algorithm of Hahn et al. (2018) is a better choice than the exact algorithm presented here when $N > p$.

2.2. Approximate Algorithm

To reduce computational cost per step, we employ an approximation of the matrix product $\xi^{-1}WDW'$. The horseshoe prior is designed for the sparse setting, where most of the true β's are zero or very small. In this case, the horseshoe posterior will tend to concentrate strongly around zero for most of the true nulls, thus endowing it with its minimax adaptive properties. Of course, this means that the posterior has a great deal of structure, since we can typically expect it to be tightly concentrated around the origin in a subspace of dimension approximately $(p - s)$, where s is the unknown number of non-nulls.

We can exploit this structure to create very accurate approximations of $\xi^{-1}WDW'$. For entries of β_j to be shrunk to near zero, the precision $\xi\eta_j$ must be very large, as can be seen from (5). When this is the case, the jth column of W does not contribute much to the $N \times N$ matrix $\xi^{-1}WDW'$. An important practical consequence of this, hitherto unexplored, is that once the MCMC algorithm begins to converge, the matrix $\xi^{-1}WDW'$ will typically be well-approximated by hard-thresholding D in (3), resulting in

$$M_\xi \approx M_{\xi,\delta} := I_N + \xi^{-1}WD_\delta W', \quad D_\delta = \mathrm{diag}(\eta_j^{-1}\mathbf{1}(\xi_{\max}^{-1}\eta_j^{-1} > \delta)) \tag{6}$$

for "small" δ, where $\xi_{\max} = \max(\xi, \xi^*)$ in the first step of (4) (the choice of δ is considered in Section 4). This thresholding step reduces computational cost considerably, since the columns of W corresponding to the zero diagonal entries of D_δ can just be ignored. Letting

$$S = \{j : \xi_{\max}^{-1}\eta_j^{-1} > \delta\}, \tag{7}$$

we can also write the approximation as $M_{\xi,\delta} = I_N + \xi^{-1}W_S D_S W_S'$, where W_S consists of the columns of W with indices in the set S, and D_S consists of the rows and columns of D with indices in the set S. This makes clear the computational advantages of thresholding.

Using this strategy, we define an approximate algorithm that uses the same update rule as in (4), with only two changes:

1. M_ξ is replaced by $M_{\xi,\delta}$ everywhere that it appears in (4); and
2. In the final step of (5), the quantity DW' is replaced by $D_\delta W'$.

We denote the Markov transition operator corresponding to this variation of (4) by $\mathcal{P}_\epsilon$. The subscript ϵ is meant to indicate that $\mathcal{P}_\epsilon$ is "close" to $\mathcal{P}$ in some suitable metric on probability measures. The choice of metric and how close $\mathcal{P}$ is to $\mathcal{P}_\epsilon$ as a function of the current state and δ are the focus of Section 3.2.

The primary motivation behind the approximate algorithm is to improve per-iteration computational complexity without sacrificing accuracy. As discussed earlier, when the truth is sparse, we expect a large subset of $\{\xi^{-1}\eta_j^{-1}\}_{j\le p}$ to be small *a posteriori*, and hence thresholding the entries smaller than a small threshold δ should not affect the accuracy of the algorithm. Thresholding those small entries has significant computational advantages. The speedup from this approximation is best when p is large relative to N and the truth is sparse or close to sparse, so that most entries of β are shrunk to near zero. Critically, coordinates that are thresholded away at iteration k *need not be thresholded away at iteration* $(k+1)$, and in practice the set of variables that escapes the threshold does change considerably from one iteration to another. This can occur because the thresholded coordinates are never actually set to zero or omitted, but rather sampled from a Gaussian that closely approximates the exact full conditional. Thus, we are not sacrificing the primary benefit of Bayesian methods for sparse regression: estimates of uncertainty about the set of true signals are still valid.

Consider the computational cost of extending the Markov chain by a single step. The exact algorithm needs to calculate $|M_\xi|$, $z'M_\xi^{-1}z$ and solve a linear system in M_ξ in each iteration, each of which requires $O(N^3)$ operations. Further, formation of the matrix M_ξ itself requires computation of WDW', which has complexity $O(N^2p)$. The approximate algorithm on the other hand needs to calculate $WD_\delta W'$, with a subset of the diagonal entries of D_δ being zero. With S as in (7) denoting the *active set* of variables which escape the threshold, set

$$s_\delta = |S| = \sum_{j=1}^{p} \mathbf{1}(\xi_{\max}^{-1}\eta_j^{-1} > \delta).$$

Also, let D_S denote the $s_\delta \times s_\delta$ sub-matrix of D and W_S the $N \times s_\delta$ sub-matrix of W resulting from picking out the non-thresholded diagonal entries/columns indexed by S. When the truth is sparse, $s_\delta \ll p$ after a few iterations and $WD_\delta W' = W_S D_S W_S'$, which costs $N^2 s_\delta$,

providing significant savings. A second level of computational savings can be made when $s_\delta < N$, whence $WD_\delta W'$ is a reduced-rank approximation to WDW'. In such cases, our implementation altogether replaces the calculation of $WD_\delta W'$ and the formation of $M_{\xi,\delta}$ by directly calculating

$$M_{\xi,\delta}^{-1} = (I_N + \xi^{-1} W D_\delta W')^{-1} = I_N - W_S\big(\xi D_S^{-1} + W_S' W_S\big)^{-1} W_S',$$

using the Woodbury matrix identity. The calculation of $z' M_{\xi,\delta}^{-1} z$ and $M_{\xi,\delta}^{-1}(z/\sigma - v)$ are performed by substituting the above expression of $M_{\xi,\delta}^{-1}$, which only requires solving $s_\delta \times s_\delta$ systems, and has overall complexity $s_\delta^3 \vee s_\delta N$. The determinant of $I + \xi^{-1} W D_\delta W'$ is then computed by (a) performing a singular value decomposition of $W_S D_S^{1/2}$, which costs $\mathcal{O}(s_\delta^2 N)$, and then (b) calculating the eigenvalues as $1+\mathrm{s}^2$, where s is a vector of the singular values of $W_S D_S^{1/2}$, $(N - s_\delta)$ of which are identically zero. Accounting for the calculation of Wu performed when sampling β, which costs $\mathcal{O}(Np)$, the per step computational cost of the approximate algorithm when $s_\delta < N$ is order $(s_\delta^2 \vee p)N$. Thus by exploiting the sparse structure of the target, the algorithm achieves similar computational cost *per step* to coordinate descent algorithms for Lasso and Elastic Net (Friedman et al., 2010, Sections 2.1, 2.2).

Before we conclude this section, we provide some additional insight into the consequences of the approximation for β. The effects of the modified updates for ξ and σ^2 are relatively direct to see; however those for β modify multiple steps of the algorithm in Bhattacharya et al. (2016). Define $\Gamma := \xi^{-1} D$ and $\Gamma_\delta = \xi^{-1} D_\delta$. The approximate algorithm for β sets

$$\beta = \Gamma_\delta W' M_\delta^{-1} z + \sigma\,(u - \Gamma_\delta W' M_\delta^{-1} v).$$

Since (u, v) is jointly Gaussian, β obtained above continues to have a Gaussian distribution, $\beta \sim N(\mu_\delta, \sigma^2 \Sigma_\delta)$, with

$$\mu_\delta := \Gamma_\delta W' M_\delta^{-1} z, \quad \Sigma_\delta := \mathrm{cov}(u - \Gamma_\delta W' M_\delta^{-1} v).$$

Some further simplifications (see Appendix A.1 for details) yields,

$$\mu_\delta = (\mu_S; 0_{(p-s_\delta)\times 1}), \quad \mu_S = (W_S' W_S + \Gamma_S^{-1})^{-1} W_S' z, \tag{8}$$

and[2]

$$\Sigma_\delta = \begin{bmatrix} (W_S' W_S + \Gamma_S^{-1})^{-1} & -\Gamma_S W_S' M_S^{-1} W_{S^c} \Gamma_{S^c} \\ & \Gamma_{S^c} \end{bmatrix}. \tag{9}$$

Writing $\beta = (\beta_S; \beta_{S^c})$, we have $\mathbf{E}(\beta_{S^c}) = 0$, i.e, the entries of β outside the active set are drawn from a zero mean distribution. Second, the marginal distribution of β_S is $N\big((W_S' W_S + \Gamma_S^{-1})^{-1} W_S' z,\, \sigma^2 (W_S' W_S + \Gamma_S^{-1})^{-1}\big)$, which would exactly be the full conditional distribution of β if the model was fitted with only the current set of active variables.

Finally, we draw some distinctions between our approximate algorithm and the skinny Gibbs algorithm (Narisetty et al., 2018). The skinny Gibbs algorithm, designed for spike-and-slab priors, partitions the β vector into inactive and active coordinates, $\beta = (\beta_I; \beta_A)$,

2. When we write $\mu_\delta = (m_S; 0_{(p-s_\delta)\times 1})$, we simply mean that the sub-vector of μ_δ corresponding to the indices in S is m_S while the rest are zero. Similarly, blocks of Σ_δ are defined by S and S^c.

depending on whether a particular coordinate is assigned to the spike or slab component by the latent component indicator. Then, it proceeds to approximate the full-conditional $[\beta \mid -] \approx [\beta_I \mid -] \otimes [\beta_A \mid -]$ by breaking the dependence between the inactive and active coordinates. A diagonal approximation to the covariance of $[\beta_I \mid -]$ is subsequently made to sample the inactive coordinates independently. This leads to a per-iteration complexity of $n(p \vee |A|^2)$, where $|A|$ is the active model size.

Unlike spike-and-slab priors which naturally group the predictors into signal and noise variables, one-group shrinkage priors such as the horseshoe do not automatically select variables. Thus, the partitioning of $\beta = (\beta_S, \beta_{S^c})$ is neither naturally available nor central to our approximate algorithm, and is merely a by-product of approximating the matrix $\xi^{-1}WDW'$ in algorithm (5) with its thresholded version in equation (6). We also note that our approximate algorithm retains the covariance between β_S and β_{S^c} unlike the skinny Gibbs algorithm. To compensate for the loss of information due to breaking the dependence structure, the skinny Gibbs algorithm carefully handles the update of the latent indicators which doesn't have a parallel for our approximate algorithm. Thus, despite surface level similarities, the two algorithms have rather different motivation and seem appropriately suited for their respective targets.

2.3. Empirical performance: some highlights

Our main motivation for pursuing a variety of theoretical results about this algorithm is that it *performs very well empirically.* In the final section, we apply the approximate algorithm to a GWAS dataset with $N = 2267$ and $p = 98,385$. This is about an order of magnitude larger in p than any other application of horseshoe for linear models that we are aware of. Our exact algorithm has per-step linear complexity in p and quadratic complexity in N, while the approximate algorithm actually has per-step linear complexity in N and p in many cases. It therefore competes with coordinate descent for the Lasso in terms of per-step computational cost. While the dependence of the mixing properties of the Markov chain on dimension is not considered theoretically, empirically we find that both algorithms are insensitive to N and p on typical metrics like autocorrelations and effective sample sizes.

We give a brief empirical comparisons of our two algorithms to the algorithm of Polson et al. (2014) combined with the faster updates of β from Bhattacharya et al. (2016)[3] based on a simulation with $N = 2,000$ and $p = 20,000$, where the true β consists of a sparse sequence of signals of varying sizes. We use $\delta = 10^{-4}$ for the approximate algorithm; choosing δ is considered in detail in Section 4, along with a complete description of the simulation setup. The left panel of Figure 1 compares autocorrelations for $\log(\xi)$ for the "old" algorithm to our exact algorithm ("new") and our approximate algorithm. We focus on ξ since this parameter is known to mix poorly in MCMC algorithms for the horseshoe (Polson et al., 2014). Both of our algorithms improve mixing considerably. The right panel of Figure 1 shows the distribution of effective samples per second, a measure of overall computational efficiency, over a number of parameters for the three algorithms. The new algorithm has slightly worse performance at the median than the old algorithm because of the slightly higher per-step cost of the blocked sampler, but performs much better for the

3. We refer to this combined algorithm as the "old" algorithm throughout

slowest-mixing parameters. The approximate algorithm is about 50 times more efficient by this metric than the exact algorithm, and this gap widens with increasing p.

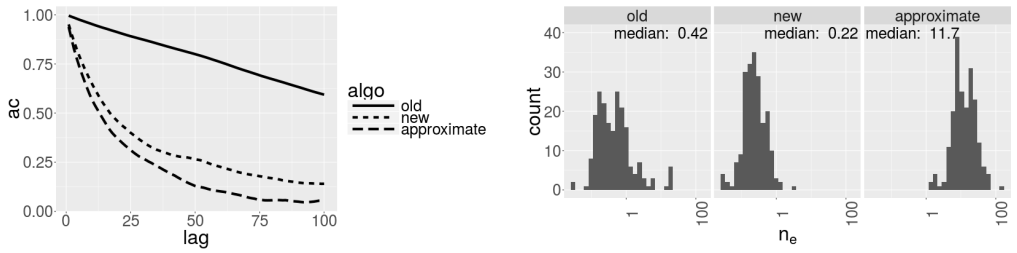

Figure 1: Left: Estimated autocorrelations for $\log(\xi)$ *for the three algorithms. Right: Effective samples per second for the three algorithms*

Figure 2 shows trace plots and density estimates for a single entry of β for the three algorithms. This particular β_j corresponds to a true signal of moderate size, and the resulting posterior marginal is bimodal, reflecting uncertainty about whether it is a signal or a null. Our exact and approximate algorithms both apparently mix well and result in visually similar estimates of the posterior marginal, while the old algorithm appears to become stuck at zero after a few thousand iterations, and the higher mode is lost after discarding a burn-in. Although this is a single entry of β, we later perform a more complete empirical comparison and find that the new algorithm outperforms the old algorithm on every metric we consider, while there is little discernible difference between the exact and approximate algorithms when $\delta = 10^{-4}$. Intuitively, the choice of δ should depend only weakly on dimension, since the matrix WDW' is always regularized by the identity. Thus a "small" value of δ is one that is small relative to the eigenvalues of the identity, which are all 1. This is discussed in more detail in Section 4.

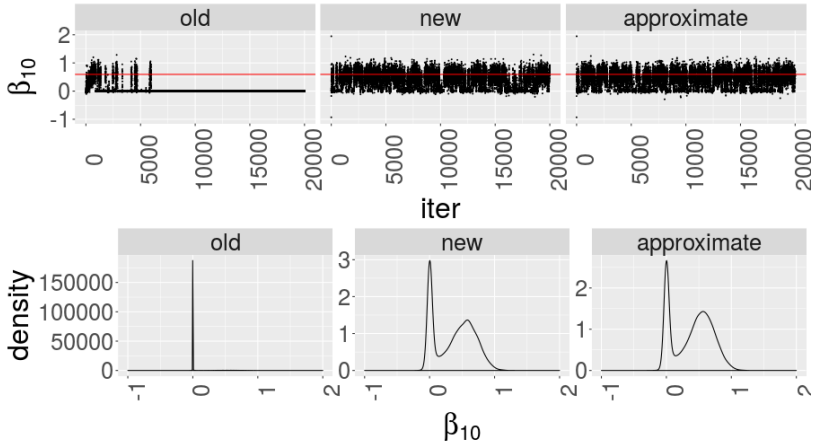

Figure 2: Trace plots (with true value indicated) and density estimates for one entry of β*.*

3. Theoretical results

We now give results connecting the convergence of the exact algorithm with the accuracy of the approximate algorithm. We also show that it is possible to construct exact algorithms using only approximate kernels. All proofs are deferred to the appendix.

3.1. Background

We first give some background on convergence rates of Markov chains. Our presentation follows closely that of Hairer and Mattingly (2011). Let $\mathcal{P}$ be a Markov transition operator on a measurable state space $\mathbf{X}$. Let x denote a generic element of the state space $\mathbf{X}$, so here $x = (\beta, \eta, \xi, \sigma^2)$. Denote by $P(\mathbf{X})$ the set of probability measures on $\mathbf{X}$. We shall follow the general convention of (Hairer and Mattingly, 2011) to denote the action of $\mathcal{P}$ on a measurable function $f : \mathbf{X} \to \mathbb{R}$ and a probability measure ν on $\mathbf{X}$ by

$$\mathcal{P}f(x) = \int f(y)\mathcal{P}(x, y)dy, \quad \nu\mathcal{P}(A) = \int_{\mathbf{X}} \mathcal{P}(x, A)\nu(dx).$$

We will state a form of geometric ergodicity of $\mathcal{P}$ that follows from two standard assumptions about $\mathcal{P}$, the first of which is existence of a Lyapunov function (Khasminskii, 1980; Meyn and Tweedie, 1993; Rosenthal, 1995).

Assumption 3.1 *There exists a function $V : \mathbf{X} \to [0, \infty)$ and constants $0 < \gamma < 1$ and $K > 0$ such that*

$$(\mathcal{P}V)(x) \equiv \int V(y)\mathcal{P}(x, dy) \leq \gamma V(x) + K.$$

The second condition is minorization on sublevel sets of V. We give the form of this condition used in Hairer and Mattingly (2011).

Assumption 3.2 *For every $R > 0$ there exists $\alpha \in (0, 1)$ (depending on R) such that, for $\mathcal{S}(R) = \{x : V(x) < R\}$,*

$$\sup_{x,y \in \mathcal{S}(R)} \|\delta_x\mathcal{P} - \delta_y\mathcal{P}\|_{\text{TV}} \leq 2(1 - \alpha). \tag{10}$$

Throughout, we will study convergence in a total variation norm weighted by the Lyapunov function; refer to Hairer and Mattingly (2011) for more details. For $\theta > 0$, define

$$d_\theta(\nu_1, \nu_2) = \int (1 + \theta V(x))|\nu_1 - \nu_2|(dx). \tag{11}$$

If $\theta = 0$, we recover the usual unweighted total variation metric.

We then have the following variation of Harris' theorem from Hairer and Mattingly (2011).

Theorem 1 (Hairer and Mattingly (2011), Theorem 3.1) *Suppose $\mathcal{P}$ satisfies assumptions 3.1 and 3.2. Then there exists $\bar{\alpha} \in (0, 1)$ and $\theta > 0$ such that*

$$d_\theta(\nu_1\mathcal{P}, \nu_2\mathcal{P}) \leq \bar{\alpha} d_\theta(\nu_1, \nu_2) \tag{12}$$

for any two probability measures $\nu_1, \nu_2 \in P(\mathbf{X})$. That is, $\mathcal{P}$ is geometrically ergodic (or V-uniformly ergodic).

Iterating the estimate in (12) and letting $\nu_1 = \nu$, and $\nu_2 = \nu^*$, the invariant measure, gives the result $d_\theta(\nu\mathcal{P}^n, \nu^*) \leq \bar{\alpha}^n d_\theta(\nu, \nu^*)$, so that convergence toward the target occurs at an exponential rate. If $\mathcal{P}$ satisfies Theorem 1, then there exists a $C < \infty$ such that

$$\sup_{|\varphi|<1+V} \int \varphi(y)(\delta_x \mathcal{P}^n - \nu^*)(dy) \leq C\bar{\alpha}^n V(x), \tag{13}$$

which is the more commonly used notion of geometric ergodicity in the MCMC literature.

In the high-dimensional linear model, the existence of a Lyapunov function and minorization on sub-level sets has been verified to establish geometric ergodicity of various global-local priors including the Bayes Lasso Gibbs sampler (Khare and Hobert, 2013) of Park and Casella (2008) and the Dirichlet-Laplace Gibbs sampler (Pal and Khare, 2014) of Bhattacharya et al. (2015). However, such a result hasn't been proved for any horseshoe sampler to best of our knowledge, possibly owing to the polynomial tails of the prior. Although not our main focus, we show in the supplemental document that the exact horseshoe sampler considered here is geometrically ergodic in $p > N$ settings provided υ_G is compactly supported and υ_L is truncated below. Truncation above and below of the prior υ_G on ξ has been recommended by van der Pas et al. (2017a), and an inspection of the proof of the prior concentration for the horseshoe in Chakraborty et al. (2016) reveals that the same concentration result goes through with a lower truncation on the η_js. Hence, the modified prior can be shown to be statistically optimal in a minimax sense. The geometric ergodicity result carries through when $p \leq N$ without any prior truncations.

3.2. Perturbation bounds

We now turn to proving error bounds for the approximate algorithm. Often in algorithm development, it is useful to identify computational bottlenecks, then design some computationally faster numerical approximation to alleviate the bottleneck. For example, the algorithm in section 2.2 substitutes an approximation of the matrix WDW' that is fast to compute when η^{-1} is near sparse. This defines some new Markov operator $\mathcal{P}_\epsilon$. One then typically wants to know that the long-time dynamics of $\mathcal{P}_\epsilon$ will approximate those of $\mathcal{P}$. For example, we might ask whether the invariant measure(s) of $\mathcal{P}_\epsilon$ (assuming they exist) are close to the invariant measure ν^* of $\mathcal{P}$, or whether the usual time-averaging estimator

$$n^{-1} \sum_{k=0}^{n-1} \varphi(X_k^\epsilon)$$

for $X_k^\epsilon \sim \nu\mathcal{P}_\epsilon^k$ gives a good approximation to expectations under ν^*. This is referred to as perturbation theory. This approach has significant advantages over studying $\mathcal{P}_\epsilon$ directly. For example, it is not necessary to show separately that $\mathcal{P}_\epsilon$ is geometrically ergodic (though in many cases it is), nor is it necessary that $\mathcal{P}_\epsilon$ has a unique invariant measure. Moreover, closeness of the invariant measure(s) of $\mathcal{P}_\epsilon$ to ν^* can be demonstrated as a corollary of bounds on the dynamics of the two chains.

Perturbation bounds for uniformly ergodic Markov operators date at least to Mitrophanov (2005), but more recent work Johndrow and Mattingly (2017); Rudolf and Schweizer (2018); Pillai and Smith (2014) focuses on the unbounded state space setting and the use

of Lyapunov functions. One effectively needs two conditions, the first of which is some pointwise control of the kernel approximation error, typically in the same metric used to study convergence. In our setting one such condition is

Assumption 3.3 *The approximate kernel $\mathcal{P}_\epsilon$ satisfies*

$$\sup_{x \in \mathbf{X}} \|\delta_x \mathcal{P} - \delta_x \mathcal{P}_\epsilon\|_{\mathrm{TV}} \leq \frac{\epsilon}{2}.$$

This differs from the basic error control assumption in both Johndrow and Mattingly (2017) and Rudolf and Schweizer (2018), which used variants of the condition $d_1(\delta_x \mathcal{P}, \delta_x \mathcal{P}_\epsilon) \leq \epsilon(1 + \kappa V(x))$. In Theorem 3, we show that the two conditions are essentially equivalent when one has control over stochastic stability of $\mathcal{P}_\epsilon$ via a Lyapunov function. It is often convenient if this is also a Lyapunov function of $\mathcal{P}$, so that the same weighted norms can be used to metrize convergence.

Assumption 3.4 *There exists $K_\epsilon > 0$ and $\gamma_\epsilon \in (0, 1)$ such that*

$$(\mathcal{P}_\epsilon V)(x) \leq \gamma_\epsilon V(x) + K_\epsilon.$$

Before stating our main results we point out an important general property of Lyapunov functions and weighted total variation metrics that allows us to use Assumption 3.3 instead of an approximation error condition in d_θ.

Remark 2 *If $\mathcal{P}$ has a Lyapunov function, then there must exist a Lyapunov function V of $\mathcal{P}$ for which V^2 is also a Lyapunov function. In particular, if $\tilde{V}$ is a Lyapunov function of $\mathcal{P}$, then $V = \tilde{V}^{1/2}$ is a Lyapunov function of $\mathcal{P}$ whose square is also a Lyapunov function. Moreover, if $\mathcal{P}$ satisfies Assumption 3.2 for V^2, then it also satisfies Assumption 3.2 for V. Thus, if Theorem 1 holds in the weighted total variation norm built on V^2, then it also holds in the weighted total variation norm built on V.*

Proof The first part is proved in Meyn and Tweedie (1993), but the argument is simple so we reproduce it here. Let $\tilde{V}$ be a Lyapunov function of $\mathcal{P}$ and put $V = \tilde{V}^{1/2}$. There exist $\tilde{\gamma}$ and $\tilde{K}$ so that

$$(\mathcal{P}\tilde{V})(x) \leq \tilde{\gamma}\tilde{V}(x) + \tilde{K}.$$

By Jensen's inequality

$$\begin{aligned}(\mathcal{P}\tilde{V}^{1/2})(x) \leq (\mathcal{P}\tilde{V}(x))^{1/2} \leq \sqrt{\tilde{\gamma}\tilde{V}(x) + \tilde{K}} &\leq \sqrt{\tilde{\gamma}}\sqrt{\tilde{V}(x)} + \sqrt{\tilde{K}} \\ &\equiv \gamma V(x) + K.\end{aligned}$$

and thus V is a Lyapunov function for which V^2 is also a Lyapunov function. For the second part, we only need to show minorization on sublevel sets of V. Since $\mathcal{P}$ satisfies Assumption 3.2 for $\tilde{V}$, for every $R^2 > 0$ there exists $\alpha_{R^2} \in (0, 1)$ (depending on R) so that

$$\sup_{x,y \in \tilde{\mathcal{S}}(R^2)} \|\delta_x \mathcal{P} - \delta_y \mathcal{P}\|_{TV} \leq 2(1 - \alpha_{R^2})$$

for $\tilde{\mathcal{S}}(R^2) = \{x : \tilde{V}(x) < R^2\}$. But then Assumption 3.2 is also satisfied for V, since letting $\mathcal{S}(R) = \{x : V(x) < R\}$, we have

$$\sup_{x,y \in \mathcal{S}(R)} \|\delta_x \mathcal{P} - \delta_y \mathcal{P}\|_{TV} \leq 2(1 - \alpha_R).$$

■

The next result shows that under Assumptions 3.4 and 3.3, we can obtain various bounds on the accuracy of $\mathcal{P}_\epsilon$.

Theorem 3 *Let V be a Lyapunov function of $\mathcal{P}$ and $\mathcal{P}_\epsilon$ for which V^2 is also a Lyapunov function of $\mathcal{P}$ and $\mathcal{P}_\epsilon$. Suppose $\mathcal{P}$ satisfies Assumptions 3.1 and 3.2, and $\mathcal{P}_\epsilon$ satisfies Assumptions 3.3 and 3.4, all defined with respect to Lyapunov function V. Then, with $\psi(\epsilon) = C^*\sqrt{\epsilon}$ for a constant $C^* > 0$, we have that for any probability measures ν_1, ν_2,*

$$\begin{aligned} d_\theta(\nu_1 \mathcal{P}_\epsilon^n, \nu_2 \mathcal{P}^n) \leq \frac{\psi(\epsilon)}{1 - \bar{\alpha}} \left(\frac{1 + K_\epsilon}{1 - \gamma_\epsilon} \right) + \psi(\epsilon)(\nu_1 V)(\bar{\alpha} \vee \gamma_\epsilon)^{n-1} n \\ + \bar{\alpha}^n d_\theta(\nu_1, \nu_2), \end{aligned} \tag{14}$$

which immediately implies that if ν_ϵ^* is any invariant measure of $\mathcal{P}_\epsilon$

$$d_\theta(\nu_\epsilon^*, \nu^*) \leq \frac{\psi(\epsilon)}{1 - \bar{\alpha}} \left(\frac{1 + K_\epsilon}{1 - \gamma_\epsilon} \right).$$

Furthermore, there exists $C, c_0, c_1 < \infty$ so that for any $|\varphi| < \sqrt{V}$

$$\begin{aligned} \mathbf{E} \left(\frac{1}{n} \sum_{k=0}^{n-1} \varphi(X_k^\epsilon) - \nu^* \varphi \right)^2 \leq 3C^2 \psi(\epsilon) c_0 \\ + \frac{3C^2}{n} \left(\frac{2(1 + K_\epsilon)}{1 - \gamma_\epsilon} + \frac{\psi(\epsilon) c_1 V(x_0)}{1 - \sqrt{\gamma_\epsilon}} \right) + \mathcal{O}\left(\frac{1}{n^2} \right), \end{aligned} \tag{15}$$

with $X_0^\epsilon = x_0$ and $X_k^\epsilon \sim \delta_{x_0} \mathcal{P}_\epsilon^{k-1}$. Moreover, the constants C, c_0, c_1 satisfy

$$C \leq \frac{1 \wedge \nu^* V}{1 - \bar{\alpha}_{(1/2)}}, \quad c_0 \leq 2 + 5 \frac{K_\epsilon \vee \sqrt{K_\epsilon}}{(1 - \sqrt{\gamma_\epsilon})^2} \quad c_1 = \left(2 + \frac{\sqrt{K_\epsilon}}{1 - \sqrt{\gamma_\epsilon}} \right),$$

where ν^* is the unique invariant measure of $\mathcal{P}$ and $1 - \bar{\alpha}_{(1/2)}$ is the spectral gap in the weighted total variation norm built on $V^{1/2}$ with an appropriate $\theta_{(1/2)} > \theta$.

Proof The key to the result is the following Lemma. This result improves upon the result in (Johndrow and Mattingly, 2017, Section 4.1), which showed that control in unweighted total variation was sufficient if the approximation error was tuned to the current state. This result requires only uniform control in the total variation or Hellinger distance over the entire state space.

Lemma 4 *Suppose V is a Lyapunov function of both $\mathcal{P}$ and $\mathcal{P}_\epsilon$ for which V^2 is also a Lyapunov function, and that Assumption 3.3 holds. Then with $\psi(\epsilon) = C^*\sqrt{\epsilon}$ we have*

$$\int (1 + V(y)) |\delta_x \mathcal{P} - \delta_x \mathcal{P}_\epsilon|(dy) \leq \psi(\epsilon)(1 + V(x))$$

Proof Write $\delta_x\mathcal{P}, \delta_x\mathcal{P}_\epsilon$ as densities

$$p(x,y) = \frac{d\delta_x\mathcal{P}}{d\nu}(y), \quad p_\epsilon(x,y) = \frac{d\delta_x\mathcal{P}_\epsilon}{d\nu}(y)$$

with respect to an appropriate dominating measure ν, which in our applications is just Lebesgue measure, and put $\tilde{V} = V^2$. Then

$$\begin{aligned}
I(x) &= \int V(y)|p(x,y) - p_\epsilon(x,y)|dy \\
I(x)^2 &= \left(\int V(y)(p^{1/2}(x,y) + p_\epsilon^{1/2}(x,y))|p^{1/2}(x,y) - p_\epsilon^{1/2}(x,y)|dy\right)^2 \\
&\leq 2\left(\int \tilde{V}(y)(p(x,y) + p_\epsilon(x,y))dy\right)\left(\int (p^{1/2}(x,y) - p_\epsilon^{1/2}(x,y))^2 dy\right) \\
&\leq 2\left(\int \tilde{V}(y)(p(x,y) + p_\epsilon(x,y))dy\right)\left(\int |p(x,y) - p_\epsilon(x,y)|dy\right) \\
&\leq 2\left((\gamma_0 + \gamma_\epsilon)\tilde{V}(x) + K_0 + K_\epsilon\right)\|\delta_x\mathcal{P} - \delta_x\mathcal{P}_\epsilon\|_{\mathrm{TV}} \\
I(x) &\leq \sqrt{2}(\sqrt{\gamma_0 + \gamma_\epsilon}\tilde{V}^{1/2}(x) + \sqrt{K_0 + K_\epsilon})\|\delta_x\mathcal{P} - \delta_x\mathcal{P}_\epsilon\|_{\mathrm{TV}}^{1/2} \\
&\leq \sqrt{\epsilon}\left(\sqrt{K_0 + K_\epsilon} + \sqrt{\gamma_0 + \gamma_\epsilon}V(x)\right) \\
&\leq 2\sqrt{\epsilon}\sqrt{K_0 + K_\epsilon + 2}\left(\frac{1}{2} + V(x)\right)
\end{aligned}$$

where we used the fact that $\gamma_0 + \gamma_\epsilon < 2$. So finally we obtain

$$\begin{aligned}
\int (1 + \sqrt{V(y)})|\delta_x\mathcal{P} - \delta_x\mathcal{P}_\epsilon|(dy) &\leq 2\sqrt{\epsilon}\sqrt{K_0 + K_\epsilon + 2}\left(\frac{1}{2} + V(x)\right) + \frac{\epsilon}{2} \\
&\leq \psi(\epsilon)(1 + V(x))
\end{aligned} \tag{16}$$

for $\psi(\epsilon) = 2\sqrt{\frac{\epsilon}{2}}\sqrt{K_0 + K_\epsilon + 2}$, and we used the fact that $\epsilon < 2$ so $2\sqrt{\epsilon} > \epsilon/2$. ■

Now, (14) follows from (Johndrow and Mattingly, 2017, equation (10)) and (15) follows from (Johndrow and Mattingly, 2017, Theorem 1.11) after substituting $\psi(\epsilon)$ for ϵ. ■

The next result follows immediately from (11) and (16).

Corollary 5 *Suppose $\mathcal{P}$ satisfies Assumption 3.3 and V, V^2 are Lyapunov functions of $\mathcal{P}$. Then with $\psi(\epsilon) = C^*\sqrt{\epsilon}$*

$$d_1(\delta_x\mathcal{P}, \delta_x\mathcal{P}_\epsilon) \leq \psi(\epsilon)(1 + V(x)).$$

Although these bounds are fairly transparent, a few comments are in order. First, all of the error bounds decrease to zero at rate $\sqrt{\epsilon}$. Second, $\mathcal{P}_\epsilon$ has an asymptotic bias proportional to $\sqrt{\epsilon}(1-\bar{\alpha})^{-1}$, and all of the constants will be small when $\sqrt{\epsilon}$ is small relative to the spectral gap $1 - \bar{\alpha}$. The implication is that there is more "room" to use approximations when the

exact chain mixes rapidly, and the bias will be small when ϵ is small relative to the spectral gap. Moreover, it seems that if ϵ is gradually decreased to zero as the chain extends, one can achieve an exact algorithm using only approximate kernels; the conditions under which this occurs are made precise in the next section.

In finite time, the practical tradeoff is between using a longer path from $\mathcal{P}_\epsilon$ with larger ϵ, which results in larger bias but smaller variance, or a shorter path from $\mathcal{P}_\epsilon$ with smaller ϵ, which has smaller bias but much larger variance. These tradeoffs are evident from (15), which gives an estimate of the squared error risk for the time-averaging estimator. Morally this is no different from choosing between two Markov kernels with the same invariant measure, where one mixes slowly but has low computational cost per step, and one mixes rapidly but has high computational cost per step.

The next result shows that our approximate horseshoe algorithm satisfies Assumptions 3.3 and 3.4.

Theorem 6 *Let $\mathcal{P}_\epsilon$ be the Markov transition operator that uses the same update rule as $\mathcal{P}$, but approximates WDW' and DW by $WD_\delta W'$ and $D_\delta W'$ as in Section 2.2 with a fixed value of δ. Then*

1. *The function*

$$V(x) \equiv V\big(\eta, \beta, \sigma^2, \xi\big) = \frac{\|W\beta\|^2}{\sigma^2} + \xi^2 + \sum_{j=1}^{p}\left[\frac{\sigma^{2c}}{|\beta_j|^{2c}} + \frac{\eta_j^c|\beta_j|^c}{\sigma^c} + \eta_j^c\right]$$

   is a Lyapunov function of $\mathcal{P}$ and $\mathcal{P}_\epsilon$.

2. *There exists a constant $C > 0$ depending on W, z such that for any $x \in \mathbf{X}$,*

$$\sup_{x\in\mathbf{X}} \|\delta_x\mathcal{P} - \delta_x\mathcal{P}_\epsilon\|_{\text{TV}} \leq C\sqrt{\delta} + \mathcal{O}(\delta), \tag{17}$$

   where δ is the threshold tuning parameter for the matrix approximation in (6)*.*

In the Supplement, we further show that Theorem 1 holds for $\mathcal{P}$. Together, this result and Theorem 6 imply that all of the error bounds in Theorem 3 hold for the approximate algorithm. This result gives both a guarantee that taking δ sufficiently small, one can achieve any desired level of approximation error, and the rate at which the approximation error goes to zero with δ. Of course, without knowing exactly the value of all of the constants, we cannot give exact estimates of the approximation error for any δ. Section 4 focuses on choosing δ in practice.

3.3. Exact algorithms using only approximate kernels

We now give results showing how to construct "exact" versions of algorithms that only use approximating kernels. These results hold under general conditions and are not specific to the algorithms in Section 2.

In the MCMC literature, an algorithm is typically considered exact if

$$\|\nu\mathcal{P}^k - \nu^*\|_{\text{TV}} \to 0$$

as $k \to \infty$ for any starting measure ν. Similarly, one might require that time averages converge to expectations under the target, e.g.

$$\lim_{n\to\infty} \mathbf{E}\left(n^{-1}\sum_{k=0}^{n-1}\varphi(X_k^\epsilon) - \nu^*\varphi\right)^2 = 0$$

for some large class of functions φ. Of course, in most cases any pathwise quantity from $\mathcal{P}$ will still have bias for any finite running time (though see the method of Jacob et al. (2017) on de-biasing using couplings), and since one only ever has access to finite-time pathwise quantities, there is little practical difference between this guarantee and that given by Theorem 3. Nonetheless, this property is often seen as desirable. The following result shows that we can achieve this by employing a sequence of approximating transition kernels $\mathcal{P}_{\epsilon_k}$ at step k, and taking $\epsilon_k \to 0$ as $k \to \infty$ at a slow rate. Notice that while this result uses the approximation condition $d_\theta(\delta_x\mathcal{P}, \delta_x\mathcal{P}_\epsilon) \leq \epsilon_k(1 + V(x))$, this is implied by Assumption 3.3 by (11) and (16) and the fact that the norms d_1, d_θ are equivalent (see Hairer and Mattingly (2011)).

Theorem 7 *Let $\{\epsilon_k\} \in [0,1]^\infty$. Consider a Markov chain $\{X_k\}$ defined by $X_0 \sim \nu$, $X_k \mid X_{k-1} \sim \mathcal{P}_{\epsilon_k}(X_{k-1}, \cdot)$, and denote $\mathcal{P}_{\epsilon_1}\mathcal{P}_{\epsilon_2}\cdots\mathcal{P}_{\epsilon_n} \equiv \prod_{k=1}^n \mathcal{P}_{\epsilon_k}$. Suppose that for every ϵ_k,*

$$d_\theta(\delta_x\mathcal{P}, \delta_x\mathcal{P}_{\epsilon_k}) \leq \epsilon_k(1 + V(x)),$$

and that for every $\epsilon \in [0,1)$, $(\mathcal{P}_\epsilon V)(x) \leq \gamma_\epsilon V(x) + K_\epsilon$. Suppose further that $\tilde{\gamma} = \sup_{\epsilon\leq 1}\gamma_\epsilon < 1$ and $\tilde{K} = \sup_{\epsilon\leq 1} K_\epsilon < \infty$. Then if

$$\lim_{n\to\infty}\sum_{k=0}^{n}\epsilon_{n-k}\bar{\alpha}^k = 0, \tag{18}$$

we have $\lim_{n\to\infty}\|\nu\prod_{k=0}^n\mathcal{P}_{\epsilon_k} - \nu^*\|_{\mathrm{TV}} = 0$, and if

$$\lim_{n\to\infty} n^{-2}\sum_{k=1}^{n}\sum_{j=1}^{n}\sqrt{\epsilon_j\epsilon_k} = 0, \tag{19}$$

then for any function $|\varphi| < \sqrt{V}$, $\lim_{n\to\infty}\mathbf{E}\left(n^{-1}\sum_{k=0}^{n-1}\varphi(X_k^\epsilon) - \nu^*\varphi\right)^2 = 0$.

Taken together, the results in this section indicate that if one takes ϵ_k to zero, an exact algorithm can be obtained, including guarantees that time averages converge to expectations under the posterior measure uniformly over a large class of functions. This guarantee is similar to the guarantee of exactness for overdamped Langevin taking step sizes to 0 (Durmus and Moulines, 2016). However, even one step of exact Langevin is always computationally infeasible, so there is no upper bound on the computational cost of the algorithm as the step sizes decrease to zero. In contrast, in our setting the exact algorithm is just a polynomial time rather than a linear time algorithm per step. The behavior of sequences of approximate kernels with decreasing approximation error can also be compared with that of pseudo-marginal MCMC, for which one usually can choose between a slowly mixing algorithm that has low per step cost and a faster mixing algorithm that has higher cost, both

of which are exact. Similarly, in our case there is clearly some sequence ϵ_k that optimally trades off bias and variance for a fixed computational budget, but we must typically rely on empirical analysis to assess this. As such, it is often more practical to choose a single ϵ empirically, which we consider in the next section.

4. Analysis of approximation error

The development of the approximate algorithm suggests a "small" value for the threshold parameter δ, which is also backed up by the theoretical results in the previous section showing that the pointwise approximation error in d_θ decreases at rate $\delta^{1/2}$. Because the prior on $\beta_j \sim N(0, \sigma^2\eta_j^{-1}\xi^{-1})$ is scaled by σ^2, the choice of the threshold δ for $\xi^{-1}\eta_j^{-1}$ is conveniently independent of the noise level. However, because we do not have a quantitative estimate of the spectral gap $1 - \bar{\alpha}$, it is difficult to know how small δ needs to be to make the bias terms in Theorem 3 small. Here we give both some heuristic arguments for what a "small enough" value of δ will typically be, as well as an empirical analysis of the bias induced by different fixed values of δ. The latter is done by comparing paths from the exact and approximate algorithms for different values of δ for problem sizes where it is feasible to run the exact algorithm. Of course, one can always achieve an asymptotically exact algorithm by using a decreasing sequence ϵ_k of approximation errors per the results of Section 3.3. However, in practice simplicity is often highly valued, so the analysis in this section is aimed at choosing a default value of δ to be used in cases where the approximation error is held fixed over time.

4.1. Default choice for δ

An inspection of the proof of Theorem 6 reveals that the approximation error $\sup_x \|\delta_x\mathcal{P} - \delta_x\mathcal{P}_\epsilon\|_{\mathrm{TV}}$ is bounded above by a constant multiple of $\|W\|\|z\|\,\delta^{1/2}$. Assume that $\|W\|$ grows like $\sqrt{p}$, a reasonable assumption at least for sub-Gaussian random matrices, and also that $\|z\| \sim \sqrt{n}$. This implies that one needs δ to be $o\big(1/(np)\big)$ for the approximation error to be small. For problems with $p \sim 10^4$ and $n \sim 10^3$ employed in the simulations, we found this to be overly conservative and $\delta \sim p^{-1} = 10^{-4}$ already seemed sufficiently small. Admittedly, the leading constants in n and p for the bound in Theorem 6 may not be optimal as we employ Pinsker's inequality to bound a total variation distance in terms of the stronger Kullback–Leibler divergence inside the proof. As an alternative distance measure for which a more accurate calculation is possible, we investigate below the L^2-Wasserstein distance between the full-conditional distributions of β from the exact and approximate algorithms respectively.

For probability measures P and Q on $\mathbb{R}^d$, the L^2-Wasserstein distance with respect to the Euclidean metric, denoted $\mathcal{W}_2(P, Q)$, is defined as

$$\mathcal{W}_2(P, Q) = \inf_{(U,V)\in\mathcal{C}(P,Q)} (E\|U - V\|^2)^{1/2}, \tag{20}$$

where $\mathcal{C}(P, Q)$ denotes the collection of all random variables $(U, V) \in \mathbb{R}^d \times \mathbb{R}^d$ such that $U \sim P$ and $V \sim Q$. Let $P_e \equiv N(\mu, \sigma^2\Sigma)$ and $P_a \equiv N(\mu_\delta, \sigma^2\Sigma_\delta)$ denote the full-conditionals of β from the exact and approximate algorithms respectively. Although an exact expression for the $\mathcal{W}_2$ distance between two multivariate Gaussians is available, it is rather cumbersome

for further analysis in the present setting. We instead use the coupling interpretation in equation (20); the idea is to use the same u and v to generate the β for the exact and approximate algorithm. Specifically, set

$$\beta_e = \Gamma W' M^{-1} z + \sigma(u - \Gamma W' M^{-1} v), \quad \beta_a = \Gamma_\delta W' M_\delta^{-1} z + \sigma(u - \Gamma_\delta W' M_\delta^{-1} v).$$

By definition, $(\beta_e, \beta_a) \in \mathcal{C}(P_e, P_a)$ and hence, by (20),

$$\mathcal{W}_2^2(P_e, P_a) \leq \|\Gamma W' M^{-1} z - \Gamma_\delta W' M_\delta^{-1} z\|^2 + \sigma^2 \, E\|\Gamma W' M^{-1} v - \Gamma_\delta W' M_\delta^{-1} v\|^2.$$

In the above display, the expectation is with respect to the distribution of v. Some further analysis (see Appendix F) shows that $\mathcal{W}_2(P_e, P_a) \lesssim \|W\|\|z\| \, \delta$; note the difference in the powers from the total variation bound. It is readily seen that for $\mathcal{W}_2(P_e, P_a)$ to be small, one needs $\delta = o\big(1/\sqrt{np}\big)$. In particular, for $p > n$, the choice $\delta = 1/p$ satisfies this.

Based on the above considerations and our subsequent empirical analysis, we propose $\delta = 1/p$ as a default choice. It is important to keep in mind though that this is a suggestion based on heuristics and may not serve all settings uniformly well. In practice, the overriding factor in choosing such thresholds δ is computational budget, so we suggest taking δ as small as possible while satisfying computational constraints. The default $1/p$ can be taken as a safe upper bound for the choices involved.

4.2. Empirical analysis

We now empirically assess the approximation error for time averages by running the approximate algorithm for different values of δ. The results in the following sections are based on a series of simulations in which the data are generated from

$$\begin{aligned} w_i &\overset{iid}{\sim} \mathrm{N}_p(0, \Sigma) \\ z_i &\sim \mathrm{N}(w_i \beta, 4) \\ \beta_j &= \begin{cases} 2^{-(j/4 - 9/4)} & j < 24 \\ 0 & j > 23 \end{cases}, \end{aligned} \tag{21}$$

In contrast to typical simulations studies for shrinkage priors, in which signals are typically either zero or large relative to the residual variance, we use a decreasing sequence of signals. The largest signal size is 4, while 18 out of the 23 signals are smaller than the residual variance. For all of the problem sizes that we consider, this results in bimodal marginal posterior for at least some of the β_j, increasing the difficulty of sampling from the target. We consider two cases for Σ: the identity and $\Sigma_{ij} = \phi^{|i-j|}$. The latter is the covariance matrix for an autoregressive model of order 1 with autoregressive coefficient ϕ and stationary variance $(1 - \phi^2)^{-1}$. Throughout, we put $\phi = 0.9$ when simulating a dependent design. Because all of the nonzero signals are in the first 23 elements of β, all of the β_j corresponding to true signals will be highly correlated a posteriori, again considerably increasing the difficulty of efficiently sampling from the target.

For analysis of the approximation error, we simulate from (21) with $N = 1,000$ and $p = 10,000$ for $\delta = 10^{-2}, 10^{-3}, 10^{-4}$, and 10^{-5}. We also run the exact algorithm twice with different random number seeds. We collect paths of length 20,000 from each simulation

after discarding a burn-in of 5,000. For the first 100 entries of β, which includes the 23 non-nulls and 77 nulls, we compute (1) correlation of pathwise means between the exact and approximate algorithm, (2) correlation of pathwise variances between the exact and approximate algorithms, and (3) Kolmogorov-Smirnov statistics for comparing the approximate algorithm to the exact algorithm. Each metric is also computed between the paths from the exact algorithm using a different random number seed. This last measurement gives some notion of how much variation one can expect in the estimates just due to MCMC error. A value of δ that performs similarly by these metrics to another copy of the exact algorithm initiated with a different random seed is thus one that achieves almost undetectable approximation error.

The Kolmogorov-Smirnov statistics are shown in Figure 3. While $\delta = 10^{-2}$ or 10^{-3} have significant bias for at least some of the marginals, when $\delta = 10^{-4}$, none of the Kolmogorov-Smirnov statistics are greater that 0.1, and most are less than $10^{-1.5} \approx 0.03$. A slight inprovement is seen in decreasing δ to 10^{-5}, for which the distribution of Kolmogorov-Smirnov statistics is hardly distinguishable from the distribution from a replicate simulation using the exact algorithm initiated using a different random number seed.

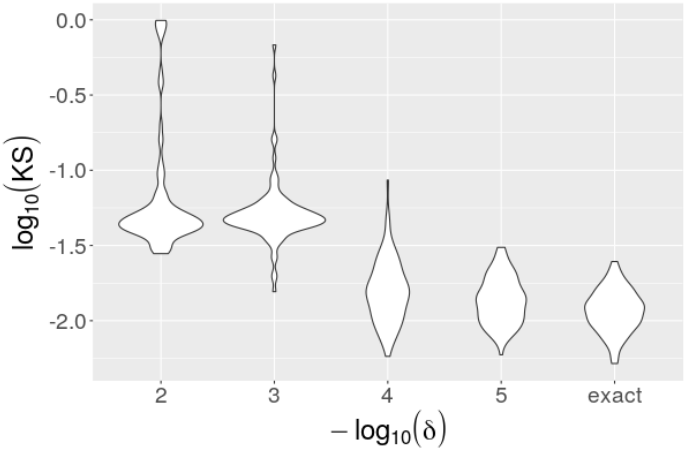

Figure 3: Distribution of Kolmogorov-Smirnov statistics comparing the marginals of 100 entries of β for different values of δ.

Table 1 shows correlations between the means and variances of 100 entries of β estimated using the exact and approximate algorithms. Similarly to the case of Kolmogorov-Smirnov statistics, significant disagreement is seen betweent the exact and approximate algorithms for $\delta = 10^{-2}$ or 10^{-3}, but they are virtually indistinguishable for $\delta = 10^{-4}$ or 10^{-5}. On the basis of these results, we typically choose $\delta = 10^{-4}$ in subsequent simulations.

5. Analysis of computational cost

5.1. Estimating dependence of constants on problem size

The results of Section 3 prove that the exact algorithm converges toward the posterior at an exponential rate, and give explicit bounds on the approximation error of time averages from $\mathcal{P}_\epsilon$ as a function of path length n. Moreover, we know the rate at which the computational

	$-\log_{10}(\delta)$	mean	variance
1	2	0.98	0.39
2	3	1.00	0.78
3	4	1.00	0.99
4	5	1.00	1.00
5	exact ($\delta = 0$)	1.00	1.00

Table 1: Correlations between estimates of means and varianes of β based on pathwise time averages for different values of δ

complexity of taking one step from $\mathcal{P}$ or $\mathcal{P}_\epsilon$ grows with N and p. However, rates are not always informative about the actual computational cost of an algorithm in finite dimensions, since one typically does not have sharp estimates of the constants. In particular, the spectral gap $1 - \bar{\alpha}$ that appears in the results of Section 3 often depends on N and p. It is typically very difficult to determine theoretically how $\bar{\alpha}$ depends on N, p in multistep Gibbs samplers like that in (4) (see e.g. Johndrow et al. (2018) for a very simple example that nonetheless required extensive calculation).

However, one can conduct an empirical analysis of computational cost in the following way. If we take $\epsilon = 0$ in (15), the bound becomes

$$\mathbf{E}\left(\frac{1}{n}\sum_{k=0}^{n-1}\varphi(X_k) - \nu^*\varphi\right)^2 \leq \frac{3}{n}\left(\frac{1 \vee \mu V}{1 - \bar{\alpha}_{(1/2)}}\right)^2\left(\frac{2(1+K_\epsilon)}{1-\gamma_\epsilon}\right) + \mathcal{O}\left(\frac{1}{n^2}\right).$$

It follows that the asymptotic (in n) variance of time averages of geometrically ergodic Markov chains is proportional to $(1 - \bar{\alpha}_{(1/2)})^{-1}$. This term can be thought of as an upper bound on the sum

$$\tau_\varphi^2 := \sum_{k=0}^{\infty}\mathrm{cov}(\varphi(X_0), \varphi(X_k)) \leq \frac{1}{1 - \bar{\alpha}_{(1/2)}} \tag{22}$$

for worst-case functions $|\varphi| < 1 + V$ with $X_0 \sim \nu$. A common approach to study how this constant varies as a function of N, p is thus to choose some collection of functions (usually coordinate projections) and compute an estimate of (22) via plugging in pathwise estimates of the covariances obtained after discarding a burn-in and truncating the sum. This is taken to be an estimate of the asymptotic variance of φ. Numerous other estimators are available; see Flegal and Jones (2010). Of course, there is no way to reliably find worst-case functions φ, but the empirical estimates at least give some sense of how this quantity behaves for statistically "important" functions like coordinate projections.

Estimates of τ_φ are referred to as MCMC standard error, and there is a significant literature on the properties of different estimators (see Flegal and Jones (2010) for a rigorous treatment). Several of these estimators are implemented in the `R` package `mcmcse`. We have consistently found the overlapping batch means estimator with the theoretically optimal $n^{1/3}$ batch size to perform the best, and we use this estimator throughout the paper. The asymptotic variance should be estimated after discarding the initial portion of the path; we discard 5,000 scans.

Using estimates of τ_φ^2 for coordinate projections, we empirically analyze the effect of problem size on the required path length as follows. Suppose that the relationship $\tau_\varphi^2 = BN^{a_1}p^{a_2}$ for constants B, a_1, a_2 dictates the growth rate of τ_φ^2 with N and p; that is to say, the asymptotic variance grows like a polynomial in N, p. Then,

$$\log(\tau_\varphi^2) = \log(B) + a_1 \log(N) + a_2 \log(p),$$

and thus one can obtain a rough estimate of the order of τ_φ^2 in p and N from a regression of $\log(\hat{\tau}_\varphi^2)$ on $\log(N)+\log(p)$. We propose to compare estimates $\widehat{a}_1, \widehat{a}_2$ of a_1, a_2 across different algorithms as a way to empirically evaluate the relative computational complexity arising from the growth of the asymptotic variance.

A related pathwise quantity is the effective sample size n_e, which is usually defined as

$$n_e = \frac{\mathrm{var}_{\nu^*}(\varphi)n}{\tau_\varphi^2}, \tag{23}$$

an adjustment to the path length n to reflect how much the asymptotic variance, τ_φ^2, is inflated by autocorrelation. Clearly, n_e is proportional to the reciprocal of the asymptotic variance, so larger n_e is better. To estimate n_e from paths of length n, we employ the procedure in `mcmcmse`, again using the overlapping batch means estimator with $n^{1/3}$ batch size and discarding 5,000 initial iterations.

5.2. Cost per step

Table 2 shows estimates of coefficients from a regression of $\log(t)$ on $\log(N) + \log(p)$ for the old, new, and approximate algorithms, where t is computation time in seconds. These estimates are based on 20 simulations from the model in (21) with N sampled uniformly at random from integers between 200 and $1,000$ and p sampled uniformly at random from integers between $1,000$ and $5,000$. The algorithm was run for 20,000 iterations and total wall clock time recorded. Computation was performed on multicore hardware with 12 threads, so matrix multiplications contribute less to the wall clock time than do matrix decompositions, resulting in the lower than expected exponents on N, p. Thus, these estimates are meant to reflect the actual performance on modern multicore hardware. Moreover, the computation time of the approximate algorithm is likely non-constant in N, p. For larger dimensions, the initial few iterations are likely to dominate the total computation time, since the benefits do not emerge until the algorithm locates most of the true nulls. This cost could be largely eliminated by "warm starting" the algorithm at, say, the cross-validated Lasso solution, which can be computed in nearly linear time in N, p. This approach could deliver a significant advantage in cases where lasso and horseshoe largely agree about the set of "important" variables, as in the application in Section 7.

Table 2: Estimates from regression of $\log(t)$ *on* $\log(N) + \log(p)$.

dimension	old	new	approximate
$\log(N)$	1.6478	1.6847	0.5449
$\log(p)$	0.7204	0.6065	0.3392

5.3. Cost related to variance of time averages

To assess the cost due to increased variance of the time-averaging estimator as a function of N, p, we conduct another set of simulations. We focus on the performance of the approximate algorithm, since its much lower computational cost per step allows a wider range of values of N, p in the simulation study, improving the reliability of the results (recall that the approximate algorithm is also geometrically ergodic by Theorem 6). The results that follow are based on two simulation studies from the setup in (21), each consisting of 20 independent simulations in which N was sampled uniformly at random from the integers between $1,000$ and $5,000$ and p was sampled uniformly at random from the integers between $10,000$ and $50,000$. In the first simulation study, we use an independent design. In the second simulation study, we use a correlated design with AR-1 structure and autocorrelation 0.9 as described above. The approximate algorithm was run for 20,000 iterations. Calculations of effective sample sizes n_e and standard errors were based on the final 15,000 iterations.

The left panels of Figure 4 shows the distribution of n_e based on the first 100 entries of β, the corresponding entries of η, $\log(\xi)$, and $-2\log(\sigma)$ as a function of p; each simulation also has a different value of N. No variation by p is evident in either the independent or correlated design case. The right panels of Figure 4 shows the analogous result, but as a function of N. A slight increase in effective sample size as N increases is possible. There is apparently little difference in n_e when the design matrix is correlated compared to independent design.

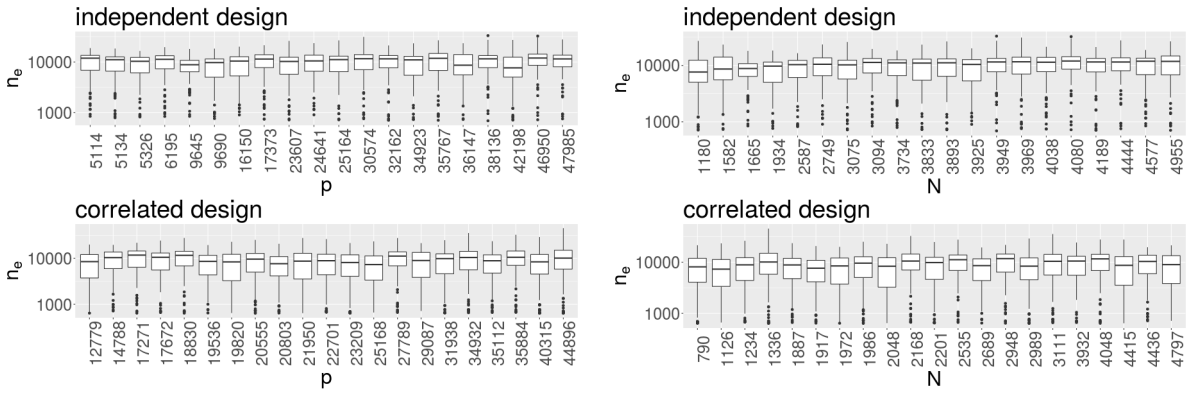

Figure 4: Left panels: Effective sample sizes n_e for 100 entries of β, 100 entries of η, $\log\xi$ and $-2\log\sigma$. The 100 entries of β, η include those corresponding to all of the true signals, and 76 null signals. The horizontal axis indicates the value of p used in each of the 20 simulations. Right panels: analogous to left panels, except that the horizontal axis indicates the value of N used in each of the 20 simulations rather than value of p.

Tables 3 and 4 show results of a linear model with specification

$$\log(\hat{\sigma}^2_{\varphi_j}(i)) = a_0 + a_1\log(N_i) + a_2\log(p_i) + b_j + \epsilon_{ij}$$

where $\varphi_j(x) = x_j$ is jth the coordinate projection of the partial state vector

$$x = (\beta_{1:100}, \eta_{1:100}, \log(\xi), -2\log(\sigma))$$

and $i = 1, \ldots, 20$ indexes the simulation scenario. Clearly, some coordinates tend to mix better than others, and the coordinate-specific intercepts allow for this variation. Results for independent design are shown in Table 3 and for dependent design in Table 4. The small, negative coefficient estimates suggest that if anything the Markov chain actually mixes slightly more rapidly as N, p increase. Thus, there is little evidence that a longer path is needed to achieve fixed Monte Carlo error as N, p grow.

Table 3: estimated parameters from regression of $-\log(n_e)$ *on* $\log(N) + \log(p)$*, independent design*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.65	0.32	-23.65	0.00
log(N)	-0.17	0.03	-5.39	0.00
log(p)	-0.03	0.02	-1.83	0.07

Table 4: estimated parameters from regression of $-\log(n_e)$ *on* $\log(N) + \log(p)$*, dependent design*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.56	0.54	-14.03	0.00
log(N)	-0.15	0.03	-5.04	0.00
log(p)	-0.06	0.04	-1.47	0.14

6. Statistical performance

Because the old algorithm can sometimes become trapped in potential wells (see Figure 2), computational cost is not a complete measure of the difference between the old and new algorithms. In this section, we analyze the performance of the three algorithms in the estimation of β, which is typically the focus of inference. We again use the simulation setup in (21) with N sampled uniformly at random from the integers between 200 and $1,000$ and p sampled uniformly at random from the integers between $1,000$ and $5,000$. Mean squared error (MSE) for estimation of β by MCMC time averages is shown in the left panel of Figure 5. There is no discernible difference between the performance of the new and approximate algorithms, but the old algorithm has about double the MSE at the median over the 20 simulations. Similarly, median empirical coverage of 95 percent credible intervals is about 90 percent for the old algorithm, and in only one case did the empirical coverage achieve 95 percent. In contrast, the new and approximate algorithms have median empirical coverage of about 93 percent, and never exhibited empirical coverage below 90 percent. We know from van der Pas et al. (2017b) that credible intervals for intermediate-sized signals cannot achieve the nominal coverage, even asymptotically. Since our simulation involves a sequence of decreasing signals, undoubtedly some of them fall into this "intermediate" categorization. As such, the performance of the new and approximate algorithms with respect to empirical coverage is probably near optimal.

Figures S3, S4, and S2 in Supplementary Materials also evaluate statistical performance of the approximate algorithm. Figures S3 and S4 show posterior marginals for the first 25 entries of β for simulations with $N = 1,000$, $p = 5,000$ and $N = 5,000$, $p = 50,000$, respectively, along with the true values of β. In general, the marginals have single modes

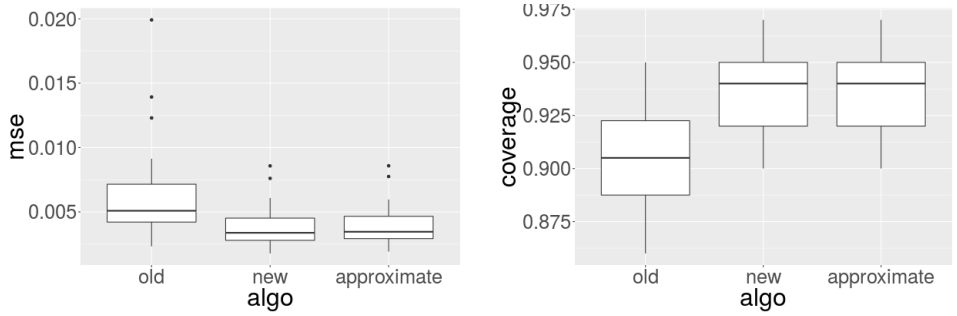

Figure 5: Mean squared error for estimation of β by time averages (left) and coverage of 95 percent equal-tailed credible intervals based on time averaged quantiles (right). Boxplot is over results of 20 simulations.

centered near the truth for larger true signals, two modes with one centered near the truth and one centered at zero for intermediate sized true signals, and single modes at zero when the true signal is small or identically zero. This is consistent with the expected behavior of the horseshoe. Figure S2 shows violin plots with indicated 95 percent credible intervals for σ^2 over 20 independent simulations each with $1,000 \leq N \leq 5,000$ and $5,000 \leq p \leq 50,000$. All but two of the intervals cover the true value of 2. Overall, the approximate algorithm has exhibited excellent statistical performance by every metric we have considered.

Finally, we conduct a series of replicate simulations to assess the ability of the approximate algorithm to concentrate around the true parameter β with increasing N. For each of $N = 200, 400, 600, \ldots, 2000$, we perform ten replicates of the simulation in (21) with $p = 20,000$ and $\delta = 2p^{-1} = 10^{-4}$. We run the approximate algorithm for $n = 21,000$ iterations, discarding $B = 1,000$ iterations and computing the pathwise average $\hat{\beta} = (n - B)^{-1} \sum_{t=B+1}^{n} \beta^{(t)}$, where $\beta^{(t)}$ is the state of β at time t. We then compute the mean squared error (MSE) $p^{-1}\|\beta - \hat{\beta}\|^2$ and provide boxplots across the ten replicates for each N in Figure 6. Clearly, pathwise averages from the approximate algorithm concentrate around the true value of β as N grows large.

7. GWAS Application

We use the horseshoe with computation by the approximate algorithm to analyze a genome-wide association study (GWAS) dataset. The data consist of $N = 2,267$ observations and $p = 98,385$ single nucleotide polymorphisms (SNPs) in the genome of maize. These data have been previously studied by Liu et al. (2016) and Zeng and Zhou (2017). Each observation corresponds to a different inbred maize line from the USDA Ames seed bank (Romay et al., 2013). As the response, we use growing degree days to silking, a measure of the average number of days exceeding a certain temperature that are necessary for the maize to "silk." Maize is typically ready to harvest about 60 days after silking, so this is a measure of the length of the growth cycle for a particular line of maize, crudely controlling for temperature. This response is also considered by Zeng and Zhou (2017).

We run the approximate algorithm for 30,000 iterations, discarding 5,000 iterations as burn-in. Figure 7 shows histograms of n_e and n_e/t for 200 entries of β, the corresponding

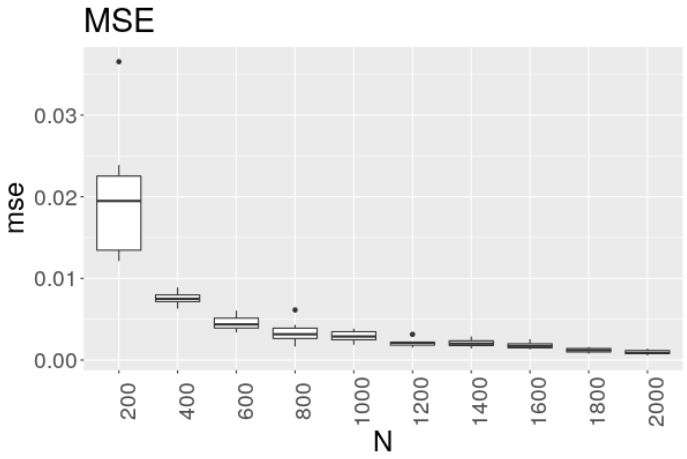

Figure 6: Mean squared error for β for the approximate algorithm as a function of N; boxplot shows variation across 10 replicate simulations.

200 entries of η, $\log(\xi)$, and $-2\log(\sigma)$. The 200 entries of β, η includes the 100 entries for which the posterior mean is largest in absolute value, as well as 100 other entries. The smallest value of n_e observed was 893, and the smallest value of n_e/t 0.05. The median values were 4531 and 0.24, respectively. Thus the algorithm remains quite efficient, even on a fairly large, real dataset.

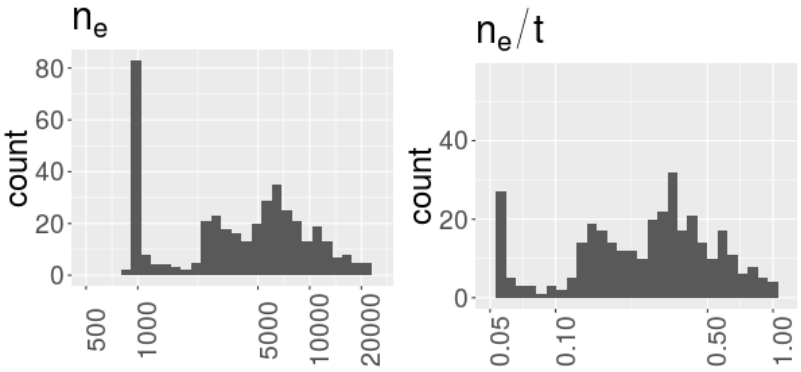

Figure 7: Effective sample size (left) and effective samples per second (right) for maize application.

Figure 8 shows density plots of samples for the nine entries of β with largest estimated absolute posterior mean, as well as the estimated posterior mean. The Lasso estimates for these parameters, with the penalty chosen by 10-fold cross-validation, are also indicated. It is clear that, even for the entries of β for which the signal strength is largest, the horseshoe marginals are typically bimodal, with the weight in the mode centered at zero increasing with decreasing signal strength. This suggests that the bimodal shape of the marginals may be quite common in applications, and gives some sense of the level of uncertainty about which entries correspond to true signals. The Lasso estimates for these relatively large

parameters are typically shrunken toward zero relative to the horseshoe posterior mean, a behavior that has been observed previously (see Bhadra et al. (2017)).

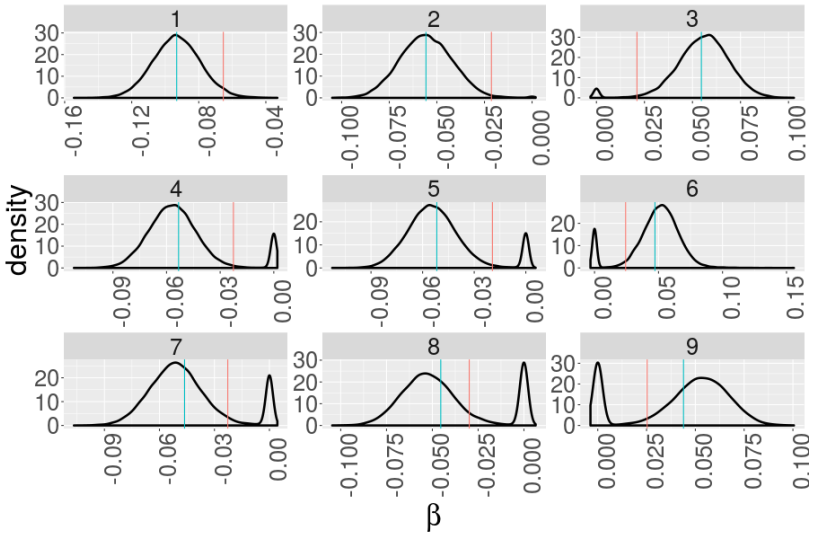

Figure 8: Density plots for the 9 entries of β with largest estimated posterior mean along with $\hat{\mathbf{E}}[\beta_j \mid y]$ from horseshoe (blue) and $\hat{\beta}_j$ from Lasso (red).

Figure 9 plots the number of entries of β for which the absolute value of the corresponding Lasso or horseshoe point estimates exceed a threshold between 0.0005 and 0.1. Also shown is the size of the intersection of these two sets. For larger thresholds, the number of horseshoe point estimates exceeding the threshold is typically larger than that for Lasso, while for smaller thresholds, this trend is reversed. This is again consistent with the tendency of Lasso to overshrink large signals and undershrink small signals (Bhadra et al., 2017). The size of the intersection closely tracks the minimum size of the two sets, suggesting that Lasso and horseshoe largely agree as to which coefficients represent signals, but disagree somewhat about their magnitude. Of course, Lasso provides no notion of uncertainty in the selected variables such as that conveyed by the posterior marginals of horseshoe.

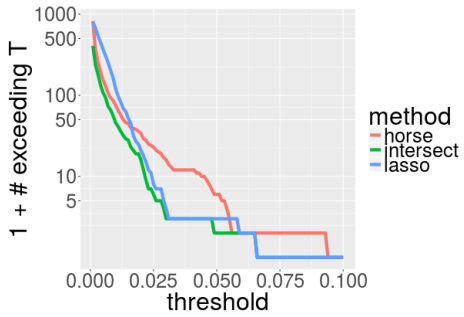

Figure 9: Plot of the number of variables for which $\hat{\mathbf{E}}[\beta_j \mid y] > T$ (horseshoe) or $\hat{\beta}_j > T$ (where $\hat{\beta}_j$ is the Lasso estimate) vs T (threshold) for $T \geq 0.0005$.

8. Discussion

It is now ten years since the Bayesian Lasso and the associated Gibbs sampling algorithm were proposed in Park and Casella (2008), eight years since the horseshoe prior appeared in Carvalho et al. (2010), and 22 years since the landmark Lasso paper of Tibshirani (1996). While the introduction of the least angle regression algorithm (Efron et al., 2004) and coordinate descent algorithms (Friedman et al., 2010) have made L_1 regularized regression with hundreds of thousands of predictors possible on standalone computing hardware, no existing implementation of Bayes Lasso, horseshoe, or any other Bayesian global-local shrinkage prior scales to this problem size. This has probably limited the adoption of these attractive Bayesian methods by practitioners, especially in the biological sciences where large p is common. Regardless of the virtues of a statistical procedure, it is of little practical use if it is not computable.

Here we have offered for the first time computational algorithms for horseshoe that can scale to hundreds of thousands of predictors. The algorithms have strong theoretical convergence and approximation error guarantees. Our approximate algorithm has the same computational cost per step as coordinate descent for elastic net and Lasso when the truth is sparse, though naturally more computation time is required to obtain a Markov chain of the requisite length than to obtain a single path of Lasso solutions. However, one gains more information from the horseshoe, perhaps most critically some measure of uncertainty regarding which β_j correspond to true signals. The Bayesian community has long recommended against selecting single models without reporting its uncertainty, but has often not provided algorithms that scale well to large p problems. This has perhaps contributed to the growing importance of selective inference over Bayesian methods, as practitioners have mostly adopted the strategy of selecting a single model. We hope that the computational strategies and results outlined here will contribute to the use of Bayesian methods in high-dimensional settings, and that exploiting sparsity and other special structure of the target will be more widely adopted as a means to develop efficient approximate MCMC algorithms for modern applications.

In designing approximate MCMC algorithms, our experience suggests that it is important for the corresponding exact algorithm to have good mixing behavior. For this reason, we introduced the new exact algorithm with superior mixing behavior over previous alternatives before considering approximations. Performing the same thresholding operation with the older algorithm that doesn't marginalize over β and σ^2 to update the global parameter ξ wasn't nearly as successful. Therefore, while the approximation scheme in the update of β can be readily applied to most global-local shrinkage priors as well as spike-and-slab priors, a careful study of the exact algorithm for the corresponding prior is recommended before exporting the approximation scheme.

Acknowledgements

The authors thank Professors Xiaolei Liu and Xiang Zhou for sharing the Maize GWAS data. Prof. Bhattacharya's research is supported by an NSF CAREER award (DMS 1653404).

Appendix A. Preliminaries

We introduce notation and make several observations that are used throughout the appendix. For a square matrix A, $\mathrm{tr}(A)$ denotes its trace. We use I_d to denote the $d \times d$ identity matrix. For an $m \times r$ matrix A (with $m > r$), $s_i(A) := s_i = \sqrt{\lambda_i}$ for $i = 1, \ldots, r$ denote the singular values of A, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$ are the eigenvalues of $A'A$. The largest and smallest (non-zero) singular values are $s_{\max}(A) = s_1(A)$ and $s_{\min}(A) = s_r(A)$. Unless otherwise stated, $\|A\| := s_{\max}(A)$ denotes the operator norm of a matrix. We often make use of the standard facts $\|AB\| = \|BA\| \leq \|A\|\,\|B\|$, $\|A + B\| \leq \|A\| + \|B\|$, and $\|A^{-1}\| = 1/s_{\min}(A)$. We use $\succeq$ to denote the partial order on the space of nonnegative definite (nnd) matrices, i.e., $A \succeq B$ implies $(A - B)$ is nnd. We also record the fact that if A, B are nnd matrices of the same size, then ABA is also nnd.

For probability measures P, Q on $(\mathcal{X}, \mathcal{B})$ having densities p and q with respect to some dominating measure ν, recall the following equivalent definitions of the total variation distance

$$\|P - Q\|_{\mathrm{TV}} = \sup_{B \in \mathcal{B}} \big|P(B) - Q(B)\big| = \frac{1}{2} \int_{\mathcal{X}} |p - q| d\nu = \sup_{|\phi| < 1} \int \phi(p - q) d\nu.$$

The Kullback–Leibler (KL) divergence $\mathrm{KL}(P \,||\, Q) = \int p \log(p/q) d\nu$. From Pinsker's inequality, $\mathrm{KL}(P \,||\, Q) \geq 2\|P - Q\|^2_{\mathrm{TV}}$.

Define the multivariate normal inverse-gamma (MNIG) distribution to be the joint distribution of $(\beta, \sigma^2) \in \mathbb{R}^p \otimes \mathbb{R}_+$ defined by the hierarchy

$$\beta \mid \sigma^2 \sim N(\mu, \sigma^2 \Sigma), \quad \sigma^2 \sim \mathrm{InvGamma}(a, a'), \tag{24}$$

where an $\mathrm{InvGamma}(a, a')$ distribution has density proportional to $x^{-(a+1)} e^{-a'/x} \mathbb{1}_{(0,\infty)}(x)$. We denote the above distribution by $\mathrm{MNIG}(\mu, \Sigma, a, a')$.

We record a lemma which calculates the KL divergence between two MNIG distributions with the same shape parameter; a proof is provided in Appendix E.3.

Lemma 8 *Suppose $p_i \sim \mathrm{MNIG}(\mu_i, \Sigma_i, a_i, a_i')$ for $i = 0, 1$, with $a_0 = a_1$. Then,*

$$\begin{aligned}\mathrm{KL}(p_0 \,||\, p_1) =& \frac{1}{2}\left[\mathrm{tr}(\Sigma_1^{-1}\Sigma_0 - I_p) - \log|\Sigma_1^{-1}\Sigma_0| + (\mu_1 - \mu_0)'\Sigma_1^{-1}(\mu_1 - \mu_0)\,\frac{a_0}{a_0'}\right] \\ &+ a_0 \log(a_0'/a_1') + \frac{(a_1' - a_0')a_0}{a_0'}.\end{aligned}$$

A.1. Derivation of mean and covariance for β in the approximate chain

Let us first derive μ_δ in (8). Recalling the definition of D_δ, we have $\Gamma_\delta W' = (\Gamma_S W_S'; 0_{p - s_\delta \times N})$ and $W \Gamma_\delta W' = W_S \Gamma_S W_S'$. Thus, $\mu_\delta = (\mu_S; 0_{p - s_\delta \times 1})$, with

$$\mu_S = \Gamma_S W_S'(I_N + W_S \Gamma_S W_S')^{-1} z = (W_S' W_S + \Gamma_S^{-1})^{-1} W_S' z.$$

A proof of the second equality can be found in the proof of Proposition 1 in Bhattacharya et al. (2016).

We now derive Σ_δ in (9). Again, using the definition of D_δ, we have $u - \Gamma W' M_\delta^{-1} v = (u_S - \Gamma_S W_S' M_S^{-1} v\,;\, u_{S^c})$, where $M_S = (I_N + W_S \Gamma_S W_S')$. Also, recall that $v = Wu + f = W_S u_S + W_{S^c} u_{S^c} + f$, and $u_S \perp\!\!\!\perp u_{S^c}$ since Γ is diagonal, which together imply $\text{cov}(u_S, v) = \Gamma_S W_S'$ and $\text{cov}(u_{S^c}, v) = \Gamma_{S^c} W_{S^c}'$. We now derive the blocks of Σ_δ.
1. We have,

$$\text{cov}(u_S - \Gamma_S W_S' M_S^{-1} v) = \Gamma_S - \Gamma_S W_S' M_S^{-1} W_S \Gamma_S = (W_S' W_S + \Gamma_S^{-1})^{-1},$$

where the proof of the second equality can be found in the proof of Proposition 1 in Bhattacharya et al. (2016).
2. Next, using $u_S \perp\!\!\!\perp u_{S^c}$,

$$\text{cov}(u_S - \Gamma_S W_S' M_S^{-1} v, u_{S^c}) = -\Gamma_S W_S' M_S^{-1} W_{S^c} \Gamma_{S^c}.$$

3. Finally, $\text{cov}(u_{S^c}) = \Gamma_{S^c}$.

Appendix B. Transition densities of the exact and approximate chain

We lay down the transition densities of the exact and approximate chains. Recall $D = \text{diag}(\eta_j^{-1})$ and $\Gamma = \xi^{-1} D$.

Exact chain. First, a comment on the state space for the Markov chain(s) of interest. $\mathcal{P}(x, \cdot)$ is not defined for x in the set

$$\mathcal{X}_0 = \{x = (\beta, \sigma^2, \xi, \eta) : \beta_j = 0 \text{ for one or more } j\}$$

Thus, we exclude $\mathcal{X}_0$ from the state space, and construct a Lyapunov function that is infinite on this set, so that points in this set are not on the boundary of sublevel sets of the Lyapunov function. By (Hairer and Mattingly, 2011, Remark 1.1), the Lyapunov condition and minorization on its sublevel sets are sufficient to establish exponential convergence toward a unique invariant measure. In particular, (Hairer and Mattingly, 2011, Theorem 3.1) requires only that the state space $\mathbf{X}$ be a measurable space. Since $\mathcal{X}_0$ has $\nu\mathcal{P}^k$-measure zero for any $k > 0$ whenever it has ν-measure zero, computational problems are avoided by simply not initializing the Markov chain at a point in $\mathcal{X}_0$.

Define $\mathbb{R}_{\setminus 0} = \mathbb{R} \setminus \{0\}$. Consider the Markov transition kernel $\mathcal{P}$ for our exact algorithm on state space $\mathbf{X} = \mathbf{X}_1 \times \mathbf{X}_2 = \mathbb{R}_+^p \times (\mathbb{R}_{\setminus 0}^p \times \mathbb{R}_+ \times \mathbb{R}_+)$ with state variable $x = (\eta, x_{\setminus \eta})$, where $x_{\setminus \eta} = (\beta, \sigma^2, \xi)$. Letting $x = (\tilde\eta, \tilde\beta, \tilde\sigma^2, \tilde\xi)$ denote the current state and $y = (\eta, \beta, \sigma^2, \xi)$ the new state, the transition kernel $\mathcal{P}(x, \cdot)$ has density with respect to Lebesgue measure

$$p(x, y) = p\big((\tilde\eta, x_{\setminus\eta}), (\eta, y_{\setminus\eta})\big) = p_1(\eta \mid x_{\setminus\eta})\, p_2(y_{\setminus\eta} \mid \eta, \tilde\xi), \tag{25}$$

where

$$\begin{aligned} p_1(\eta \mid x_{\setminus\eta}) &= \prod_{j=1}^p p_1(\eta_j \mid x_{\setminus\eta}), \\ p_1(\eta_j \mid x_{\setminus\eta}) &= \frac{e^{-\tilde m_j}}{\Gamma(0, \tilde m_j + b\tilde m_j)} \frac{e^{-\tilde m_j \eta_j}}{1 + \eta_j} \mathbb{1}_{(b,\infty)}(\eta_j), \\ p_2(y_{\setminus\eta} \mid \eta, \tilde\xi) &= p_2(\beta \mid \sigma^2, \xi, \eta)\, p_2(\sigma^2 \mid \xi, \eta)\, p_2(\xi \mid \eta, \tilde\xi), \end{aligned} \tag{26}$$

where $\tilde{m}_j = \tilde{\xi}\tilde{\beta}_j^2/(2\tilde{\sigma}^2)$ and $\Gamma(\cdot,\cdot)$ is the incomplete Gamma function defined in (50). We describe the various components of p_2 below.

The transition kernel of the MH-within-Gibbs update for ξ can be written as

$$p_2(\xi \mid \eta, \tilde{\xi}) = \alpha_\eta(\tilde{\xi}, \xi)\, h(\xi \mid \tilde{\xi}) + r_\eta(\tilde{\xi})\, \delta_{\tilde{\xi}}(\xi), \tag{27}$$

where $h(\cdot \mid \cdot)$ is the log-normal proposal kernel for ξ. Here,

$$\alpha_\eta(\tilde{\xi}, \xi) = \min\left\{1, q_\eta(\tilde{\xi}, \xi) = \frac{p(\xi \mid \eta)\xi}{p(\tilde{\xi} \mid \eta)\tilde{\xi}}\right\}$$

is the probability of accepting a move to ξ from $\tilde{\xi}$, with $p(\xi \mid \eta)$ as defined in (3), $\delta_{\tilde{\xi}}(\cdot)$ denotes a point-mass at $\tilde{\xi}$, and

$$r_\eta(\tilde{\xi}) = 1 - \int \alpha_\eta(\tilde{\xi}, \xi)\, h(\xi \mid \tilde{\xi})\, d\xi$$

is the probability of staying at $\tilde{\xi}$.

The full conditional $p_2(\beta \mid \sigma^2, \xi, \eta)$ is $N(\mu, \sigma^2\Sigma)$ with $\mu = \Sigma W'z$ and $\Sigma = (W'W + (\xi^{-1}D)^{-1})^{-1}$, while $p_2(\sigma^2 \mid \xi, \eta)$ has an InvGamma distribution. Thus, using the definition of MNIG above, the full conditional for (β, σ^2) from the exact algorithm, $p_2(\beta, \sigma^2 \mid \xi, \eta)$, is distributed as MNIG$(\mu, \Sigma, a, a')$, with

$$\begin{aligned} &\mu = \Sigma W'z = \Gamma W' M^{-1} z,\ \Sigma = (W'W + \Gamma^{-1})^{-1} = \Gamma - \Gamma W' M^{-1} W \Gamma, \\ &a = (N+\omega)/2,\ a' = (z' M^{-1} z + \omega)/2. \end{aligned} \tag{28}$$

In the fist line of the above display, we used various equivalent representations of μ and Σ which follow from the Woodbury matrix identity and are encapsulated in the algorithm (5).

Approximate chain. We now describe the transition density of the approximate chain. Noting that the update for η remains the same in the approximate algorithm, the approximate Markov kernel $\mathcal{P}_\epsilon(x, \cdot)$ has transition density

$$p_\epsilon(x, y) = p_1(\eta \mid x_{\setminus\eta})\, p_{2,\epsilon}(\beta, \sigma^2 \mid \xi, \eta)\, p_{2,\epsilon}(\xi \mid \eta, \tilde{\xi}) \quad x \in \mathbf{X}, \tag{29}$$

where $p_{2,\epsilon}(\beta, \sigma^2 \mid \xi, \eta)$ denotes the approximate full conditional of (β, σ^2) resulting from the approximations of DW and WDW' by $D_\delta W$ and $WD_\delta W'$ described in Section 2.2, which is distributed as MNIG$(\mu_\delta, \Sigma_\delta, a_\delta, a'_\delta)$, with

$$\begin{aligned} &\mu_\delta = \Gamma_\delta W' M_\delta^{-1} z, \quad \Sigma_\delta = \Gamma - \big(2\Gamma W' M_\delta^{-1} W \Gamma_\delta - \Gamma_\delta W' M_\delta^{-1} M M_\delta^{-1} W \Gamma_\delta\big), \\ &a_\delta = (N+\omega)/2, \quad a'_\delta = (z' M_\delta^{-1} z + \omega)/2. \end{aligned} \tag{30}$$

Also,

$$p_{2,\epsilon}(\xi \mid \tilde{\xi}, \eta) = \alpha_{\eta,\epsilon}(\tilde{\xi}, \xi)\, h(\xi \mid \tilde{\xi}) + r_{\eta,\epsilon}(\tilde{\xi})\, \delta_{\tilde{\xi}}(\xi), \tag{31}$$

is the approximate MH-within-Gibbs transition density obtained by the approximation of the acceptance probability $\alpha_{\eta,\epsilon}(\xi, \xi') = \min\big\{1, q_{\eta,\delta}(\xi, \xi')\big\}$ where

$$q_{\eta,\delta} = \frac{p_\epsilon(\xi \mid \eta)\xi}{p_\epsilon(\tilde{\xi} \mid \eta)\tilde{\xi}}, \tag{32}$$

and $p_\epsilon(\xi \mid \eta)$ is obtained by replacing M_ξ by $M_{\xi,\delta}$ in (3).

Appendix C. Proof of Theorem 6

We first show that V continues to define a Lyapunov function for the approximate chain $\mathcal{P}_\epsilon$, and then bound the total variation distance between the exact and approximate transition densities.

C.1. Lyapunov condition for approximate chain

The proof of this part to a large extent closely resembles the first part of the proof of Theorem 14, and we only point out the key features. The only place where one requires more work is to bound the trace term of $E(\|W\beta\|^2 \mid \sigma^2, \xi, \eta) = \mu_\delta' W' W \mu_\delta + \sigma^2 \operatorname{tr}(W\Sigma_\delta W')$ under $\mathcal{P}_\epsilon$. Proceeding as before, we can show $\mu_\delta' W' W \mu_\delta = \|(I_N - M_\delta^{-1})z\|^2 \leq \|z\|^2$. Write $\operatorname{tr}(W\Sigma_\delta W') = \operatorname{tr}(W\Sigma W') + \operatorname{tr}(W\Delta W')$, where $\Delta = \Sigma_\delta - \Sigma$. We have already showed in the proof of Theorem 14 that $\operatorname{tr}(W\Sigma W')$ is small. For the other term, write $\operatorname{tr}(W\Delta W') = \operatorname{tr}(W\Sigma^{1/2}\Sigma^{-1/2}\Delta\Sigma^{-1/2}\Sigma^{1/2}W')$. We prove in the next subsection (see equation 46) that

$$\|\Sigma^{-1}\Delta\|_2 \leq 8\|W\|^2\delta + O(\delta^2).$$

Since $\Sigma^{-1/2}\Delta\Sigma^{-1/2}$ is similar to $\Sigma^{-1}\Delta$, this means $\Sigma^{-1/2}\Delta\Sigma^{-1/2} \preceq CI_p$ for some constant $C > 0$. This means $\operatorname{tr}(W\Sigma^{1/2}\Sigma^{-1/2}\Delta\Sigma^{-1/2}\Sigma^{1/2}W') \leq C \operatorname{tr}(W\Sigma W')$, which we already know is bounded above by a constant.

The other fact used to complete the proof is that the same two-sided bound for σ_j^2 continues to hold as before. To see this, for $j \notin S$, $\sigma_j^2 = \xi^{-1}\eta_j^{-1}$, while for $j \in S$, $\sigma_j^2 \geq 1/(s_{\max}^2(W_S) + \xi\eta_j) \geq 1/(s_{\max}^2(W) + \xi\eta_j)$, and the upper bound $\sigma_j^2 \leq (\xi\eta_j)^{-1}$ holds for all j.

C.2. Proof of uniform total variation bound in (17)

Recall we denote $x = (\beta, \sigma^2, \xi, \eta)$ for the entire state vector. We shall also call $\theta = (\beta, \sigma^2)$. The various pieces of the transition density for the exact algorithm is given in (25) – (28), while the same for the approximate algorithm is given in (29) – (32).

We now proceed to bound the total variation distance between $\mathcal{P}(x,\cdot)$ and $\mathcal{P}_\epsilon(x,\cdot)$. We have, for a fixed $x \in \mathbf{X}$,

$$
\begin{aligned}
&2\|\mathcal{P}(x,\cdot) - \mathcal{P}_\epsilon(x,\cdot)\|_{\text{TV}} = \int |p(x,y) - p_\epsilon(x,y)| dy \\
&= \int p_1(\eta \mid x_{\setminus\eta}) \left\{ \int |p_2(y_{\setminus\eta} \mid \eta, \tilde{\xi}) - p_{2,\epsilon}(y_{\setminus\eta} \mid \eta, \tilde{\xi})| dy_{\setminus\eta} \right\} d\eta \\
&\leq \sup_\eta \int |p_2(y_{\setminus\eta} \mid \eta, \tilde{\xi}) - p_{2,\epsilon}(y_{\setminus\eta} \mid \eta, \tilde{\xi})| dy_{\setminus\eta} \\
&= \sup_\eta \int \left| \; p_2(\theta \mid \xi, \eta)\, p_2(\xi \mid \eta, \tilde{\xi}) - p_{2,\epsilon}(\theta \mid \xi, \eta)\, p_{2,\epsilon}(\xi \mid \eta, \tilde{\xi}) \; \right| dy'_{\setminus\eta} \\
&\overset{(i)}{\leq} \sup_\eta \left[\int \left\{ \int |p_2(\theta \mid \xi, \eta) - p_{2,\epsilon}(\theta \mid \xi, \eta)| \; d\theta \right\} p_2(\xi \mid \eta, \tilde{\xi})\, d\xi \right. \\
&\left. + \int \left| p_2(\xi \mid \eta, \tilde{\xi}) - p_{2,\epsilon}(\xi \mid \eta, \tilde{\xi}) \right| d\xi \right] \\
&\overset{(ii)}{\leq} 2 \sup_{\xi,\eta} \|p_2(\theta \mid \xi, \eta) - p_{2,\epsilon}(\theta \mid \xi, \eta)\|_{\text{TV}} + 2 \sup_{\xi,\tilde{\xi},\eta} \left| \alpha_\eta(\tilde{\xi}, \xi) - \alpha_{\eta,\epsilon}(\tilde{\xi}, \xi) \right|.
\end{aligned}
$$

For (i), we used triangle inequality and that $\int p_{2,\epsilon}(\theta \mid \xi, \eta)\, d\theta = 1$. For (ii), we used that

$$
\begin{aligned}
&\int \left| p_2(\xi \mid \eta, \tilde{\xi}) - p_{2,\epsilon}(\xi \mid \eta, \tilde{\xi}) \right| d\xi \\
&\leq \int \left| \alpha_\eta(\tilde{\xi}, \xi) - \alpha_{\eta,\epsilon}(\tilde{\xi}, \xi) \right| h(\xi \mid \tilde{\xi})\, d\xi + \left| r_\eta(\tilde{\xi}) - r_{\eta,\epsilon}(\tilde{\xi}) \right| \\
&\leq 2 \sup_{\xi,\tilde{\xi},\eta} \left| \alpha_\eta(\tilde{\xi}, \xi) - \alpha_{\eta,\epsilon}(\tilde{\xi}, \xi) \right|.
\end{aligned}
$$

Since the bound in (ii) is independent of x, we conclude that

$$
\begin{aligned}
\sup_{x \in \mathbf{X}} \|\delta_x \mathcal{P} - \delta_x \mathcal{P}_\epsilon\|_{\text{TV}} \leq & \underbrace{\sup_{\xi,\eta} \|p_2(\theta \mid \xi, \eta) - p_{2,\epsilon}(\theta \mid \xi, \eta)\|_{\text{TV}}}_{\text{TV}_1} \\
& + \underbrace{\sup_{\xi,\tilde{\xi},\eta} \left| \alpha_\eta(\tilde{\xi}, \xi) - \alpha_{\eta,\epsilon}(\tilde{\xi}, \xi) \right|}_{\text{TV}_2}.
\end{aligned} \tag{33}
$$

We now separately bound TV_1 and TV_2. We show that

$$
\begin{aligned}
\text{TV}_1^2 &= 4\|W\|^2\delta + \frac{N+\omega}{\omega}\|W\|^2\delta + \frac{N}{2}\frac{\|z\|^2}{\omega}\|W\|^2\delta + \mathcal{O}(\delta^2), \\
\text{TV}_2 &= N\,\|W\|^2\,(1 + \|z\|^2/\omega)\delta + \mathcal{O}(\delta^2),
\end{aligned}
$$

for sufficiently small δ, which produce the desired bound. Since the derivations to obtain these bounds are somewhat lengthy, we split them into two different sections below.

C.3. Bounding TV_2: MH ratio approximations for ξ

We first record a couple of useful auxiliary results. The first result is a well-known eigenvalue perturbation bound due to Weyl.

Lemma 9 (Weyl) *Let A, E be $n \times n$ Hermitian matrices. Then, for $i = 1, \ldots, n$,*

$$|\nu_i(A + E) - \nu_i(A)| \leq \|E\|,$$

where $\nu_i(A)$ denotes the ith eigenvalue of A, and $\|\cdot\|$ denotes the operator norm of a matrix.

Next, we present a simple yet useful result to bound the difference between MH acceptance probabilities.

Lemma 10 *For any $a, b > 0$,*

$$|\min(a, 1) - \min(b, 1)| \leq \max\left\{|(a/b) - 1|, |(b/a) - 1|\right\} \leq e^{|\Delta|} - 1,$$

where $\Delta = \log(a/b)$.

Proof First observe that $|\min(a, 1) - \min(b, 1)| \leq |a - b|$, which can be verified by enumerating the 4 different cases (i) $a, b < 1$, (ii) $a < 1 < b$, (iii) $b < 1 < a$, and (iv) $a, b > 1$. In case (iv), the left hand side is 0 and the claimed bound is trivially satisfied. In the remaining cases, bound

$$\begin{aligned} |a - b| &= |\{\max(a, b)/\min(a, b)\} - 1| \min(a, b) \leq |\{\max(a, b)/\min(a, b)\} - 1| \\ &\leq \max\left\{|(a/b) - 1|, |(b/a) - 1|\right\}. \end{aligned}$$

This proves the first part. The second part simply follows from the monotonicity of $x \mapsto e^x$. ■

As noted in Appendix B, we have that

$$\alpha_\eta(x, y) = \min\{1, q_\eta(x, y)\}, \quad \alpha_{\eta,\epsilon}(x, y) = \min\{1, q_{\eta,\delta}(x, y)\}$$

with

$$q_\eta(x, y) = \frac{|M_y|^{-1/2} (\omega + z' M_y^{-1} z)^{-(N+\omega)/2}}{|M_x|^{-1/2} (\omega + z' M_x^{-1} z)^{-(N+\omega)/2}} \frac{y\sqrt{x}\,(1 + x)}{x\sqrt{y}\,(1 + y)},$$

and $q_{\eta,\delta}(x, y)$ is obtained by replacing M_t by $M_{t,\delta}$, where, recall that

$$M_t = I_N + t^{-1} W D W', \; M_{t,\delta} = I_N + t^{-1} W D_\delta W', \quad t \in \{x, y\}.$$

It then follows from Lemma 10 that

$$\left|\alpha_\eta(x, y) - \alpha_{\eta,\epsilon}(x, y)\right| \leq \exp(|\Delta|) - 1,$$

where

$$\Delta = \log \frac{q_{\eta,\delta}(x,y)}{q_\eta(x,y)} = \Delta_1 + \Delta_2,$$
$$\Delta_1 = \Delta_{1,y} - \Delta_{1,x}, \quad \Delta_{1,t} = -\frac{1}{2}\left[\log|M_{t,\delta}| - \log|M_t|\right],\ t \in \{x,y\},$$
$$\Delta_2 = \Delta_{2,y} - \Delta_{2,x}, \quad \Delta_{2,t} = -\frac{n+\omega}{2}\left[\log(1 + z' M_{t,\delta}^{-1} z/\omega) - \log(1 + z' M_t^{-1} z/\omega)\right],\ t \in \{x,y\}.$$

We shall prove below that

$$|\Delta| \le N \|W\|^2 (1 + \|z\|^2/\omega)\delta. \tag{34}$$

Observe the right hand side is independent of ξ and η.

To establish (34), we bound

$$|\Delta| \le \sum_{t\in\{x,y\}} [|\Delta_{1,t}| + |\Delta_{2,t}|]. \tag{35}$$

We now proceed to individually bound $|\Delta_{1,t}|$ and $|\Delta_{2,t}|$ for $t \in \{x,y\}$.

For $t \in \{x,y\}$, we have

$$\begin{aligned}
\left|\log|M_t| - \log|M_{t,\delta}|\right| &= \left|\sum_{i=1}^N \left[\log\{1 + t^{-1}\nu_i(WDW')\} - \log\{1 + t^{-1}\nu_i(WD_\delta W')\}\right]\right| \\
&\le \sum_{i=1}^N \left|\log\{1 + t^{-1}\nu_i(WDW')\} - \log\{1 + t^{-1}\nu_i(WD_\delta W')\}\right| \\
&\le \sum_{i=1}^N \left|t^{-1}\nu_i(WDW') - t^{-1}\nu_i(WD_\delta W')\right|,
\end{aligned}$$

where the last step uses the fact that the map $u \mapsto \log(1+u)$ for $u > 0$ is Lipschitz. Write

$$t^{-1}WDW' = t^{-1}WD_\delta W' + t^{-1}WD_{<\delta}W',$$

where $D_{<\delta} = \mathrm{diag}\big((\eta_j^{-1})\,\mathbf{1}(j \in \mathcal{I}^c)\big)$ retains the entries of D which are thresholded. By Weyl's perturbation bound (see Lemma 9), for any $i = 1, \ldots, N$,

$$\left|t^{-1}\nu_i(WDW') - t^{-1}\nu_i(WD_\delta W')\right| \le t^{-1}\|WD_{<\delta}W'\| \le \delta\|W\|^2,$$

where we use the fact that, given our thresholding rule, all non-zero diagonal entries of the matrix $t^{-1}D_{<\delta}$ is bounded by δ for $t \in \{x,y\}$. Substituting the bound, we obtain,

$$\sum_{t\in\{x,y\}} |\Delta_{1,t}| \le N\|W\|^2\,\delta. \tag{36}$$

Next, we bound $|\Delta_{2,t}|$ for $t \in \{x, y\}$. To that end, once again using that the map $u \mapsto \log(1+u)$ is Lipschitz, bound

$$\begin{aligned}\left|\log(1+z'M_t^{-1}z/\omega) - \log(1+z'M_{t,\delta}^{-1}z/\omega)\right| &\leq \left|z'(M_t^{-1} - M_{t,\delta}^{-1})z/\omega\right| \\ &\leq (\|z\|^2/\omega)\,\|M_t^{-1} - M_{t,\delta}^{-1}\| \\ &\leq (\|z\|^2/\omega)\,\|M_t^{-1}(M_{t,\delta} - M_t)M_{t,\delta}^{-1}\| \\ &\leq (\|z\|^2/\omega)\,\|M_{t,\delta} - M_t\|,\end{aligned}$$

where we have used the identity $A^{-1}-B^{-1} = A^{-1}(B-A)B^{-1}$, the bound $\|AB\| \leq \|A\|\|B\|$, and the fact that both $\|M_t^{-1}\|$ and $\|M_{t,\delta}^{-1}\|$ are bounded above by 1. Continuing from the last line of the display, $\|M_t - M_{t,\delta}\| = \|t^{-1}WD_{<\delta}W'\| \leq \delta\|W\|^2$ using the same argument as in the bound for $\Delta_{1,t}$. Substituting this bound, we obtain,

$$\sum_{t\in\{x,y\}} |\Delta_{2,t}| \leq (N+\omega)\,(\|z\|^2/\omega)\,\|W\|^2\,\delta. \tag{37}$$

Substituting (36) and (37) in (35), we obtain (34). Now, making a Taylor expansion of $e^x - 1$ about zero, we obtain for $0 < x < 1$

$$e^x - 1 = (1 + x + \mathcal{O}(x^2)) - 1 = x + \mathcal{O}(x^2),$$

which gives

$$\mathrm{TV}_2 = N\,\|W\|^2\,(1 + \|z\|^2/\omega)\delta + \mathcal{O}(\delta^2) \tag{38}$$

for sufficiently small δ.

C.4. Bounding TV_1

To bound the total variation distance between $p_2(\cdot \mid \xi, \eta)$ and $p_{2,\epsilon}(\cdot \mid \xi, \eta)$, we use Pinsker's inequality,

$$\|p_2(\theta \mid \xi, \eta) - p_{2,\epsilon}(\theta \mid \xi, \eta)\|_{\mathrm{TV}}^2 \leq \frac{1}{2}\,\mathrm{KL}\big(p_{2,\epsilon}(\theta \mid \xi, \eta)\,||\,p_2(\theta \mid \xi, \eta)\big), \tag{39}$$

and subsequently use the expression for KL between two MNIGs derived in Lemma 8; note that the shape parameters $a_\delta = a = (N+\omega)/2$ and hence the Lemma applies.

Let us define

$$\begin{aligned}\mathrm{KL}_1 &= \mathrm{tr}(\Sigma^{-1}\Sigma_\delta - I_p) - \log|\Sigma^{-1}\Sigma_\delta|, \\ \mathrm{KL}_2 &= (\mu - \mu_\delta)'\Sigma^{-1}(\mu - \mu_\delta)\,\frac{a_\delta}{a'_\delta}, \\ \mathrm{KL}_3 &= a_\delta \log(a'_\delta/a') + (a' - a'_\delta)\,\frac{a_\delta}{a'_\delta}\end{aligned} \tag{40}$$

so that

$$\mathrm{KL}\big(p_{2,\epsilon}(\cdot \mid \xi, \eta)\,||\,p_2(\cdot \mid \xi, \eta)\big) = 0.5(\mathrm{KL}_1 + \mathrm{KL}_2) + \mathrm{KL}_3.$$

We now proceed to bound each of the terms subsequently.

C.4.1. Bounds for KL_1

The matrix $\Sigma^{-1}\Sigma_\delta$ is similar to the positive definite matrix $\Sigma^{-1/2}\Sigma_\delta\Sigma^{-1/2}$, and hence its eigenvalues $\{\zeta_j\}_{j=1}^p$ are all positive. Expressing the trace and determinant in terms of the eigenvalues, we obtain,

$$\mathrm{KL}_1 = \sum_{j=1}^{p}(\zeta_j - 1 - \log\zeta_j). \tag{41}$$

Now, write

$$\Sigma_\delta = \Sigma + \Delta, \quad \Delta = \Gamma W' M^{-1} W\Gamma - \left(2\Gamma W' - \Gamma_\delta W' M_\delta^{-1} M\right) M_\delta^{-1} W\Gamma_\delta,$$

and

$$\Sigma^{-1}\Sigma_\delta = I_p + \Sigma^{-1}\Delta.$$

Using $\mathrm{rank}(B_1B_2) \le \min\{\mathrm{rank}(B_1), \mathrm{rank}(B_2)\}$, Δ is the difference of two matrices with rank at most N each, and using $\mathrm{rank}(B_1+B_2) \le \mathrm{rank}(B_1)+\mathrm{rank}(B_2)$, we can bound $\mathrm{rank}(\Delta) \le 2N$, which then implies $\mathrm{rank}(\Sigma^{-1}\Delta) \le 2N$. Letting $\{\bar{\zeta}_j\}_{j=1}^p$ denote the eigenvalues of $\Sigma^{-1}\Delta$, it then follows that $\bar{\zeta}_j = 0$ for $j \ge 2N$. Since $\zeta_j = 1 + \bar{\zeta}_j$, we conclude that $\zeta_j = 1$ for $j \ge 2N$, and

$$\mathrm{KL}_1 = \sum_{j=1}^{2N}\left(\zeta_j - 1 - \log\zeta_j\right) = \sum_{j=1}^{2N}\left[\bar{\zeta}_j - \log(1+\bar{\zeta}_j)\right]. \tag{42}$$

Observe that the right hand side is a positive quantity, since $\log(1+x) \le x$ for $x > -1$ and $\bar{\zeta}_j > -1$ for all j (since $\zeta_j > 0$ for all j). Using Taylor expansion, it can be further shown that $x - \log(1+x) < x^2$ whenever $|x| \le 1/2$. Using that the magnitude of the eigenvalues of a matrix are bounded by its operator norm, we have $|\bar{\zeta}_j| \le \|\Sigma^{-1}\Delta\|$ for all $j = 1, \ldots, 2N$. Hence, if we can show that $\|\Sigma^{-1}\Delta\|$ is small, we can bound

$$\mathrm{KL}_1 \le \sum_{j=1}^{2N}|\bar{\zeta}_j|^2 \le 2N\,\|\Sigma^{-1}\Delta\|^2. \tag{43}$$

With this motivation, we now proceed to bound $\|\Sigma^{-1}\Delta\|$. To facilitate our bounds, we decompose

$$(\Sigma_\delta - \Sigma) = (\Sigma_\delta - \Sigma_*) + (\Sigma_* - \Sigma),$$

where

$$\Sigma_* = \Gamma - \Gamma W'(2M_\delta^{-1} - M_\delta^{-1} M M_\delta^{-1})W\Gamma.$$

Σ_* itself is a covariance matrix, although this isn't used in the subsequent analysis. Letting $A = M_\delta^{-1} M M_\delta^{-1}$,

$$\Gamma_\delta W' A W\Gamma_\delta - \Gamma W' A W\Gamma = \Gamma_\delta W' A W(\Gamma_\delta - \Gamma) + (\Gamma_\delta - \Gamma)W' A W\Gamma.$$

Hence,

$$\Sigma_\delta - \Sigma_* = \underbrace{2\Gamma W' M_\delta^{-1} W(\Gamma - \Gamma_\delta)}_{T_1} + \underbrace{\Gamma_\delta W' A W(\Gamma_\delta - \Gamma)}_{T_2} + \underbrace{(\Gamma_\delta - \Gamma) W' A W \Gamma}_{T_3}.$$

Recall that $\Sigma^{-1} = (W'W + \Gamma^{-1})$. Let us now calculate

$$\Sigma^{-1}(\Sigma_\delta - \Sigma_*) = (W'W + \Gamma^{-1})(T_1 + T_2 + T_3) = H_1 + H_2 + H_3,$$

with

$$\begin{aligned} H_1 &= 2\big[W' W\Gamma W' M_\delta^{-1} W(\Gamma - \Gamma_\delta) + W' M_\delta^{-1} W(\Gamma - \Gamma_\delta)\big] \\ H_2 &= \big[W' W\Gamma_\delta W' A W(\Gamma - \Gamma_\delta) + \Gamma^{-1}\Gamma_\delta W' A W(\Gamma_\delta - \Gamma)\big] \\ H_3 &= \big[W'W (\Gamma_\delta - \Gamma) W' A W\Gamma + \Gamma^{-1} (\Gamma_\delta - \Gamma) W' A W\Gamma\big]. \end{aligned} \tag{44}$$

Next,

$$\Sigma_* - \Sigma = \Gamma W' \underbrace{\left[M^{-1} + M_\delta^{-1} M M_\delta^{-1} - 2M_\delta^{-1}\right]}_{E} W\Gamma.$$

Hence,

$$H_4 := \Sigma^{-1}(\Sigma_* - \Sigma) = (W'W + \Gamma^{-1})\Gamma W' E W\Gamma. \tag{45}$$

Combining (44) and (45) and using the triangle inequality for the operator norm,

$$\|\Sigma^{-1}\Delta\| = \|\Sigma^{-1}(\Sigma_\delta - \Sigma)\| = \|H_1 + H_2 + H_3 + H_4\| \le \sum_{i=1}^{4} \|H_i\|.$$

We now record a Lemma which collects various results required to bound the operator norms of the H_is; a proof is provided in Appendix E.4.

Lemma 11 *The following inequalities hold:*
(i) $\max\big\{\|M^{-1}\|, \|M_\delta^{-1}\|\big\} \le 1$.
(ii) $\max\big\{\|M - M_\delta\|, \|M^{-1}M_\delta - I_N\|, \|M_\delta M^{-1} - I_N\|, \|M_\delta^{-1}M - I_N\|, \|MM_\delta^{-1} - I_N\|\big\} \le \|W\|^2\delta$.
(iii) $\max\big\{\|W\Gamma W' M^{-1}\|, \|W\Gamma_\delta W' M_\delta^{-1}\|\big\} \le 1$.
(iv) $\|W\Gamma W' M_\delta^{-1}\| \le 1 + \|W\|^2\delta$.
(v) *Recalling that* $A = M_\delta^{-1} M M_\delta^{-1}$, *we have* $\|A\| \le (1 + \|W\|^2\delta)$. *Further,* $\|W\Gamma_\delta W' A\| \le (1 + \|W\|^2\delta)$ *and* $\|A W\Gamma W'\| \le (1 + \|W\|^2\delta)^2$.

Using Lemma 11, we now proceed to bound the $\|H_i\|$s; that $\|\Gamma - \Gamma_\delta\| < \delta$ is used throughout, along with the facts $\|B_1B_2\| = \|B_2B_1\|$ and $\|B_1 + B_2\| \le \|B_1\| + \|B_2\|$.

Bound for $\|H_1\|$. We obtain, using (i) and (iv) in Lemma 11,

$$\|H_1\| \le 2\|W\|^2\delta\left[\|W\Gamma W' M_\delta^{-1}\| + \|M_\delta^{-1}\|\right] \le 2\|W\|^2\delta\big[2 + \|W\|^2\delta\big].$$

Bound for $\|H_2\|$. We obtain, using (v) in Lemma 11 and the fact that $\|\Gamma^{-1}\Gamma_\delta\| \leq 1$,

$$\|H_2\| \leq \|W\|^2\delta \left[\|W\Gamma_\delta W'\, A\| + \|A\|\right] \leq 2\|W\|^2\delta \left[1 + \|W\|^2\delta\right].$$

Bound for $\|H_3\|$. We obtain, using (v) in Lemma 11 and the fact that $\|\Gamma^{-1}\Gamma_\delta\| \leq 1$,

$$\|H_3\| \leq \|W\|^2\delta \left[\|A\, W\Gamma W'\| + \|A\|\right] \leq \|W\|^2\delta \left[(1 + \|W\|^2\delta)^2 + (1 + \|W\|^2\delta)\right].$$

Bound for $\|H_4\|$. We have,

$$\begin{aligned}\|H_4\| &= \|W\Gamma(W'W + \Gamma^{-1})\Gamma W'\, E\| \\ &= \|W\Gamma W'\,(I_N + W\Gamma W')\, E\| \\ &= \|W\Gamma W'\, ME\|.\end{aligned}$$

Now, $ME = I_N + MM_\delta^{-1}MM_\delta^{-1} - 2MM_\delta^{-1} = (MM_\delta^{-1} - I_N)^2$. Substituting in the above display, and once again invoking Lemma 11,

$$\begin{aligned}\|H_4\| &= \|M_\delta^{-1}W\Gamma W'\,(M - M_\delta)M_\delta^{-1}(M - M_\delta)\| \\ &\leq \|W\Gamma W'M_\delta^{-1}\|\, \|M - M_\delta\|^2 \\ &\leq (1 + \|W\|^2\delta)\,(\|W\|^2\delta)^2.\end{aligned}$$

Bound for $\|\Sigma^{-1}\Delta\|$. Collecting the bounds for $\|H_i\|$ and substituting in the display before Lemma 11 plus some simplifying algebra yields,

$$\|\Sigma^{-1}\Delta\| \leq (\|W\|^2\delta)\left[3 + 3(1 + \|W\|^2\delta) + 2(1 + \|W\|^2\delta)^2\right] = 8\|W\|^2\delta + \mathcal{O}(\delta^2) \tag{46}$$

for sufficiently small δ.

C.4.2. Bound for KL_2

Focus first on $(\mu - \mu_\delta)'\Sigma^{-1}(\mu - \mu_\delta)$. We have $\mu - \mu_\delta = (\Gamma W'M^{-1} - \Gamma_\delta W'M_\delta^{-1})z$. Write

$$\Gamma W'M^{-1} - \gamma_\delta W'M_\delta^{-1} = \underbrace{(\Gamma - \Gamma_\delta)W'M^{-1}}_{U} + \underbrace{\Gamma_\delta W'M^{-1}(I_N - MM_\delta^{-1})}_{V}.$$

We can now write

$$\begin{aligned}(\mu - \mu_\delta)'\Sigma^{-1}(\mu - \mu_\delta) &= z'(U + V)'\Sigma^{-1}(U + V)z \\ &\leq \|(U + V)'\Sigma^{-1}(U + V)\|\, \|z\|^2 \\ &\leq \|\Sigma^{-1/2}(U + V)\|^2\, \|z\|^2 \\ &\leq 2(\|\Sigma^{-1/2}U\|^2 + \|\Sigma^{-1/2}V\|^2)\, \|z\|^2 \\ &= 2(\|U'\Sigma^{-1}U\| + \|V'\Sigma^{-1}V\|)\, \|z\|^2,\end{aligned}$$

where we used the inequality $\|B_1 + B_2\|^2 \leq 2(\|B_1\|^2 + \|B_2\|^2)$. Next,

$$\begin{aligned}\|U'\Sigma^{-1}U\| &= \|M^{-1}W(\Gamma - \Gamma_\delta)(W'W + \Gamma^{-1})(\Gamma - \Gamma_\delta)W'M^{-1}\| \\ &= \|(\Gamma - \Gamma_\delta)(W'W + \Gamma^{-1})\,(\Gamma - \Gamma_\delta)W'M^{-2}W\| \\ &\leq \|(\Gamma - \Gamma_\delta)(W'W + \Gamma^{-1})\|\,\|W\|^2\delta \\ &\leq \|W\|^2\delta(1 + \|W\|^2\delta) = \|W\|^2\delta + \mathcal{O}(\delta^2),\end{aligned}$$

for sufficiently small δ, where we have used conclusions of Lemma 11 in multiple places and in the last step, we used $\|(\Gamma - \Gamma_\delta)\Gamma^{-1}\| \leq 1$ since it is a diagonal matrix with zeros and ones on the diagonal. Similarly, it can be verified that $\|V'\Sigma^{-1}V\| \leq \|W\|^2\delta(1 + \|W\|^2\delta)$. So then it follows

$$\mathrm{KL}_2 \leq \left(2\|W\|^2\delta + \mathcal{O}(\delta^2)\right)\frac{N+\omega}{\omega} = 2\frac{N+\omega}{\omega}\|W\|^2\delta + \mathcal{O}(\delta^2).$$

where the last factor is an upper bound on a_δ/a'_δ which originates from bounding a'_δ below by $\omega/2$.

C.4.3. Bound for KL_3

Using $\log(x) \leq (x-1)$ for $x > 0$, we have,

$$\begin{aligned}\mathrm{KL}_3 &\leq a_\delta\left\{(a'_\delta/a' - 1) + (a'/a'_\delta - 1)\right\} \\ &\leq 2a_\delta|a' - a'_\delta|,\end{aligned}$$

since $a', a'_\delta > 1$. Since $|a' - a'_\delta| = |z'(M^{-1} - M_\delta^{-1})z|/\omega \leq (\|z\|^2/\omega)\,\|W\|^2\delta$, we have

$$\mathrm{KL}_3 \leq N\,(\|z\|^2/\omega)\,\|W\|^2\delta.$$

C.4.4. Summing up

We now combine the bounds to obtain the final result. We have

$$\mathrm{KL}_1 + \mathrm{KL}_2 + \mathrm{KL}_3 \leq 8\|W\|^2\delta + 2\frac{N+\omega}{\omega}\|W\|^2\delta + N\,(\|z\|^2/\omega)\,\|W\|^2\delta + \mathcal{O}(\delta^2)$$

so by Pinsker's inequality

$$\mathrm{TV}_1^2 \leq 4\|W\|^2\delta + \frac{N+\omega}{\omega}\|W\|^2\delta + \frac{N}{2}\,\frac{\|z\|^2}{\omega}\,\|W\|^2\delta + \mathcal{O}(\delta^2)$$

and so finally, combining with (38) – which contributes only factors of order δ^2 or smaller after squaring – via (33), we obtain

$$\begin{aligned}\sup_x \|\delta_x\mathcal{P} - \delta_x\mathcal{P}_\epsilon\|_{\mathrm{TV}} &= \sqrt{4\|W\|^2\delta + \frac{N+\omega}{\omega}\|W\|^2\delta + \frac{N}{2}\,\frac{\|z\|^2}{\omega}\,\|W\|^2\delta} + \mathcal{O}(\delta) \\ &= \sqrt{\delta}\|W\|\sqrt{4 + \frac{N+\omega}{\omega} + \frac{N}{2}\,\frac{\|z\|^2}{\omega}} + \mathcal{O}(\delta),\end{aligned}$$

since none of the bounds depend upon the remaining state variable η, giving the result.

Appendix D. Proof of Theorem 7

Let $\tilde{\mathcal{P}}_k$ be the k-step transition kernel $\tilde{\mathcal{P}}_k = \prod_{j=1}^{k} \mathcal{P}_{\epsilon_j}$. By (Johndrow and Mattingly, 2017, Corollary 1.6)

$$d_\theta(\nu_1\tilde{\mathcal{P}}_n, \nu_2\mathcal{P}^n) \leq \bar{\alpha} d_\theta(\nu_1\tilde{\mathcal{P}}_{n-1}, \nu_2\mathcal{P}^{n-1}) + \epsilon_n(1 + \nu_1\tilde{\mathcal{P}}_{n-1}V).$$

Iterating this estimate we obtain

$$d_\theta(\nu_1\tilde{\mathcal{P}}_n, \nu_2\mathcal{P}^n) \leq \bar{\alpha}^n d_\theta(\nu_1, \nu_2) + \sum_{k=0}^{n-1} \bar{\alpha}^k \epsilon_{n-k}(1 + \nu_1\tilde{\mathcal{P}}_{n-k-1}V).$$

Since V is a Lyapunov function of $\mathcal{P}_{\epsilon_k}$ for every ϵ_k, we obtain using the uniform bound on the constants

$$\begin{aligned}
\nu_1\tilde{\mathcal{P}}_k V &\leq \nu_1\tilde{\mathcal{P}}_{k-1}(\gamma_{\epsilon_k}V + K_{\epsilon_k}) \\
&\leq \prod_{j=1}^{k} \gamma_{\epsilon_j}\nu_1 V + \sum_{j=0}^{k-1} \gamma_{\epsilon_j} K_{\epsilon_j} \\
&\leq \tilde{\gamma}^k(\nu_1 V) + \frac{\tilde{K}}{1-\tilde{\gamma}} \leq \frac{\nu_1 V + \tilde{K}}{1-\tilde{\gamma}}
\end{aligned}$$

so

$$\begin{aligned}
d_\theta(\nu_1\tilde{\mathcal{P}}_n, \nu_2\mathcal{P}^n) &\leq \bar{\alpha}^n d_\theta(\nu_1, \nu_2) + \sum_{k=0}^{n-1} \bar{\alpha}^k \epsilon_{n-k}\left(1 + \frac{\nu_1 V + \tilde{K}}{1-\tilde{\gamma}}\right) \\
&= \bar{\alpha}^n d_\theta(\nu_1, \nu_2) + \left(1 + \frac{\nu_1 V + \tilde{K}}{1-\tilde{\gamma}}\right) \sum_{k=0}^{n-1} \bar{\alpha}^k \epsilon_{n-k}.
\end{aligned}$$

Substituting $\nu_2 = \nu^*$ we obtain

$$d_\theta(\nu_1\tilde{\mathcal{P}}_n, \nu^*) \leq \bar{\alpha}^n d_\theta(\nu_1, \nu^*) + \left(1 + \frac{\nu_1 V + \tilde{K}}{1-\tilde{\gamma}}\right) \sum_{k=0}^{n-1} \bar{\alpha}^k \epsilon_{n-k}.$$

Since d_θ bounds total variation from above, the result follows if $\lim_{n\to\infty} \sum_{k=0}^{n-1} \bar{\alpha}^k \epsilon_{n-k} = 0$.

To show the second part, we apply (Johndrow and Mattingly, 2017, equation (45)) to obtain

$$\frac{1}{n}\sum_{k=0}^{n-1} \varphi(X_k^\epsilon) - \nu^*\varphi = \frac{U(X_0^\epsilon) - U(X_n^\epsilon)}{n} + \frac{1}{n}M_n^\epsilon + \frac{1}{n}\sum_{k=0}^{n-1} (\mathcal{P}_{\epsilon_{k+1}} - \mathcal{P})U(X_k^\epsilon)$$

where $\tilde{\varphi} = \varphi - \nu^*\varphi$,

$$U = \sum_{k=0}^{\infty} \mathcal{P}\tilde{\varphi}, \quad M_n^\epsilon = \sum_{k=1}^{n} m_k^\epsilon, \quad m_{k+1}^\epsilon = U(X_k^\epsilon) - \mathcal{P}_{\epsilon_{k+1}}U(X_k^\epsilon).$$

Put $C = \frac{(1 \vee \mu V^{1/2})}{\theta_{(1/2)}(1-\bar{\alpha}_{(1/2)})}$, where $1-\bar{\alpha}_{(1/2)}$ is the spectral gap of $\mathcal{P}$ in the weighted total variation norm built on $V^{1/2}$ with an appropriate $\theta_{(1/2)}$. Simple modifications to the calculations in Johndrow and Mattingly (2017) using the uniform bounds on K_ϵ and γ_ϵ give

$$\mathbf{E}\left[\left(\frac{1}{n}M_n^\epsilon\right)^2\right] \leq 2C^2\left(\frac{1}{n} + \frac{\tilde{K}}{n\{1-\tilde{\gamma}\}} + \frac{1-\tilde{\gamma}^n}{n^2\{1-\tilde{\gamma}\}}V(x_0)\right), \tag{47}$$

and

$$\mathbf{E}[(\mathcal{P}_{\epsilon_k} - \mathcal{P})U(X_k^\epsilon)(\mathcal{P}_{\epsilon_j} - \mathcal{P})U(X_j^\epsilon)] \leq C^2\sqrt{\epsilon_k\epsilon_j}\left[c_0 + c_1\tilde{\gamma}^{(j\wedge k)/2}V(x_0)\right]$$

so

$$\begin{aligned}
\sum_{k=0}^{n-1}\sum_{j=0}^{n-1}\mathbf{E}\left[(\mathcal{P}_{\epsilon_{k+1}} - \mathcal{P})U(X_k^\epsilon)(\mathcal{P}_{\epsilon_{j+1}} - \mathcal{P})U(X_j^\epsilon)\right] &\leq \sum_{k=0}^{n-1}\sum_{j=0}^{n-1} C^2\sqrt{\epsilon_{k+1}\epsilon_{j+1}}\left[c_0 + c_1\tilde{\gamma}^{(j\wedge k)/2}V(x_0)\right] \\
&\leq C^2\left[c_0 + c_1V(x_0)\right]\sum_{k=1}^{n}\sum_{j=1}^{n}\sqrt{\epsilon_k\epsilon_j} \\
n^{-2}\sum_{k=0}^{n-1}\sum_{j=0}^{n-1}\mathbf{E}\left[(\mathcal{P}_{\epsilon_{k+1}} - \mathcal{P})U(X_k^\epsilon)(\mathcal{P}_{\epsilon_{j+1}} - \mathcal{P})U(X_j^\epsilon)\right] &\leq \frac{1}{n^2}C^2\left[c_0 + c_1V(x_0)\right]\sum_{k=1}^{n}\sum_{j=1}^{n}\sqrt{\epsilon_k\epsilon_j} \\
&\leq \frac{1}{n^2}\tilde{C}\sum_{k=1}^{n}\sum_{j=1}^{n}\sqrt{\epsilon_k\epsilon_j}
\end{aligned} \tag{48}$$

Finally

$$\frac{(U(X_0^\epsilon) - U(X_n^\epsilon))^2}{n^2} \leq \frac{4C^2}{n^2}\left(1 + (1+\tilde{\gamma}^n)V(x_0) + \frac{\tilde{K}}{1-\tilde{\gamma}}\right). \tag{49}$$

The quantities (47) and (49) converge to zero at rate n^{-1} and n^{-2}, respectively. The quantity in (48) will converge to zero whenever

$$\lim_{n\to\infty} n^{-2}\sum_{k=1}^{n}\sum_{j=1}^{n}\sqrt{\epsilon_k\epsilon_j} = 0;$$

a sufficient condition is that $\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\sqrt{\epsilon_k\epsilon_j}$ is finite, in which case convergence to zero occurs at rate n^{-2}. This completes the proof of the second part.

Appendix E. Some integrals, inequalities, & proofs of auxiliary lemmas

E.1. Incomplete Gamma function

The incomplete Gamma function $\Gamma(a,x)$ for $x > 0$ is defined as

$$\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}dt. \tag{50}$$

When $a = 0$, this reduces to the exponential integral function $\mathrm{En}_1(x) = \int_x^\infty t^{-1}e^{-t}dt$.

We record an integral from Gradstheyn and Ryzhik (see 3.383.10 of Gradshteyn and Ryzhik (1996)),

$$\int_0^\infty \frac{x^{\nu-1}e^{-\mu x}}{x+\beta}\,dx = \beta^{\nu-1}e^{\beta\mu}\,\Gamma(\nu)\,\Gamma(1-\nu,\beta\mu), \tag{51}$$

for $\nu, \mu, \beta > 0$.

We record a result relating the ratio of certain incomplete gamma functions.

Lemma 12 *Fix $c \in (0, 1/2]$ and $b \in (0, 1)$. For any small $\varepsilon > 0$, there exists a positive constant C_ε such that*

$$r_{b,c}(x) := \frac{e^{-x}}{x^c}\,\frac{\Gamma(c, bx)}{\Gamma(0, x+bx)} \le \varepsilon x^{-c} + C_\varepsilon, \quad \forall\, x \in (0, \infty).$$

Proof For $x \ge 1/2$, bound $\Gamma(0, x) \ge \int_x^{2x} e^{-t}/t\,dt \ge e^{-x}(1-e^{-x})/(2x) \ge e^{-x}/(8x)$, where we used that for $x \ge 1/2$, $1 - e^{-x} \ge 1/4$. Also bound $\Gamma(c, x) = \int_x^\infty t^{c-1}e^{-t}dt \le x^{c-1}\int_x^\infty e^{-t}dt = x^{c-1}e^{-x}$. Substituting these bounds, we have for $x \ge 1/2$ that

$$\frac{e^{-x}}{x^c}\,\frac{\Gamma(c, bx)}{\Gamma(0, x+bx)} \le 8b^{c-1}(1+b). \tag{52}$$

We also have that $\lim_{x\to 0}\frac{\Gamma(c,bx)}{\Gamma(0,x+bx)} = 0$. Pick $\delta > 0$ such that $\frac{\Gamma(c,bx)}{\Gamma(0,x+bx)} < \varepsilon$ for all $x < \delta$. We can then bound

$$r_{b,c}(x) \le \varepsilon x^{-c} + \max\{r_{b,c}(\delta), 8b^{c-1}(1+b)\}, \quad \forall\, x \in (0, \infty).$$

■

E.2. Normal inverse moments & Hypergeometric function

We state a formula for inverse absolute moments of a normal distribution from Winkelbauer (2012); specifically, see equations (17) and (6) therein. Let $X \sim \mathrm{N}(\mu, \sigma^2)$. One has, for $\nu > -1$,

$$E(|X|^\nu) = \sigma^\nu 2^{\nu/2}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}}\,e^{-\frac{\mu^2}{2\sigma^2}}\,M\left(\frac{\nu+1}{2}, \frac{1}{2}; \frac{\mu^2}{2\sigma^2}\right), \tag{53}$$

where

$$M(\alpha, \gamma; z) = {}_1F_1(\alpha; \gamma; z) := \sum_{n=0}^\infty \frac{(\alpha)_n}{(\gamma)_n n!}z^n$$

is the confluent hypergeometric function of the first kind, with

$$(x)_n \equiv \frac{\Gamma(x+n)}{\Gamma(x)},$$

the ascending factorial. Letting

$$\mathbf{M}(\alpha, \gamma; z) = \frac{M(\alpha, \gamma; z)}{\Gamma(\gamma)},$$

one has, if $\gamma > \alpha > 0$ then

$$\mathbf{M}(\alpha, \gamma; z) = \frac{1}{\Gamma(\alpha)\Gamma(\gamma - \alpha)} \int_0^1 e^{zt} t^{\alpha-1}(1-t)^{\gamma-\alpha-1} dt.$$

We state a useful result below.

Lemma 13 *Fix $\gamma > \alpha > 0$. The function*

$$z \mapsto e^{-z} M(\alpha, \gamma; z)$$

is a non-increasing function of z for $z \geq 0$. In particular, $e^{-z}M(\alpha, \gamma; z) \leq 1$ for any $z > 0$.

Proof We can write, based on the above integral representation,

$$\begin{aligned} e^{-z} M(\alpha, \gamma; z) &= \frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\gamma - \alpha)} \int_0^1 e^{-z(1-t)} t^{\alpha-1}(1-t)^{\gamma-\alpha-1} dt \\ &= \frac{\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\gamma - \alpha)} \int_0^1 e^{-zt} t^{\gamma-\alpha-1}(1-t)^{\alpha-1} dt, \end{aligned}$$

which is a non-increasing function of z on $[0, \infty)$.

The second part follows from evaluating the last line of the above display at $z = 0$, which reduces to the integral of a beta pdf at 0, so that $e^{-z}M(\alpha, \gamma; z)|_{z=0} = 1$. The bound then follows. ■

E.3. Proof of Lemma 8

We have

$$\begin{aligned} &\int p_0(\beta, \sigma^2) \log \frac{p_0(\beta, \sigma^2)}{p_1(\beta, \sigma^2)} \, d\beta d\sigma^2 \\ &= \int p_0(\beta \mid \sigma^2) \, p_0(\sigma^2) \left[\log \frac{p_0(\beta \mid \sigma^2)}{p_1(\beta \mid \sigma^2)} + \log \frac{p_0(\sigma^2)}{p_1(\sigma^2)} \right] d\beta d\sigma^2 \\ &= \int \mathrm{KL}\Big(p_0(\cdot \mid \sigma^2) \,||\, p_1(\cdot \mid \sigma^2) \Big) \, p_0(\sigma^2) \, d\sigma^2 + \int p_0(\sigma^2) \log \frac{p_0(\sigma^2)}{p_1(\sigma^2)} \, d\sigma^2. \end{aligned}$$

First, using normality of $p_i(\cdot \mid \sigma^2)$ and the standard expression for the KL divergence between two multivariate normals,

$$\mathrm{KL}\Big(p_0(\cdot \mid \sigma^2) \,||\, p_1(\cdot \mid \sigma^2) \Big) = \frac{1}{2} \left[\mathrm{tr}(\Sigma_1^{-1}\Sigma_0 - I_p) - \log |\Sigma_1^{-1}\Sigma_0| + \frac{(\mu_1 - \mu_0)' \Sigma_1^{-1} (\mu_1 - \mu_0)}{\sigma^2} \right].$$

Thus,

$$\int \mathrm{KL}\Big(p_0(\cdot \mid \sigma^2) \,||\, p_1(\cdot \mid \sigma^2)\Big)\, p_0(\sigma^2)\, d\sigma^2$$
$$= \frac{1}{2}\bigg[\operatorname{tr}(\Sigma_1^{-1}\Sigma_0 - I_p) - \log|\Sigma_1^{-1}\Sigma_0| + (\mu_1-\mu_0)'\Sigma_1^{-1}(\mu_1-\mu_0)\,\frac{a_0}{a_0'}\bigg].$$

Next, using that $a_0 = a_1$,

$$\int p_0(\sigma^2)\log\frac{p_0(\sigma^2)}{p_1(\sigma^2)}\,d\sigma^2$$
$$= \int p_0(\sigma^2)\left[(a_0\log a_0' - a_0\log a_1') + (a_1'-a_0')(\sigma^2)^{-1}\right]d\sigma^2$$
$$= a_0\log(a_0'/a_1') + \frac{(a_1'-a_0')a_0}{a_0'}.$$

E.4. Proof of Lemma 11

We make multiple usage of the following facts. For matrices A and B of compatible size, $\|AB\| = \|BA\| \le \|A\|\,\|B\|$ and $\|A+B\| \le \|A\|+\|B\|$. For invertible A, $\|A^{-1}\| = 1/s_{\min}(A)$. For a symmetric p.d. matrix A, its eigenvalues and singular values are identical.

(i) Follows since $s_{\min}(M)$ and $s_{\min}(M_\delta)$ are both bounded below by 1.

(ii) First, $\|M - M_\delta\| = \|W(\Gamma-\Gamma_\delta)W'\| \le \|W\|^2\delta$ since $\Gamma-\Gamma_\delta$ is a diagonal matrix with the non-zero entries bounded by δ. The remaining 4 inequalities have near identical proofs so we only prove one of them. We have $\|M_\delta^{-1}M - I_N\| = \|M_\delta^{-1}(M_\delta - M)\| \le \|M - M_\delta\|$ by (i).

(iii) Writing $W\Gamma W'\,M^{-1} = I_N - M^{-1}$, all its eigenvalues are bounded above by 1. Similarly for the second part.

(iv) Bound $\|W\Gamma W'\,M_\delta^{-1}\| \le \|W\Gamma W'\,M^{-1}\| + \|W\Gamma W'\,M^{-1}\,(I_N - MM_\delta^{-1})\| \le \|W\Gamma W'\,M^{-1}\|\,\big(1 + \|I_N - MM_\delta^{-1}\|\big)$. Conclude from (ii) and (iii).

(v) First, bound $\|A\| \le \|M_\delta^{-1}M\|\,\|M_\delta^{-1} - M^{-1}\| + \|M_\delta^{-1}\|$. The bound then follows from (i) and (ii) and noting that $(M_\delta^{-1} - M^{-1}) = M_\delta^{-1}(M - M_\delta)M^{-1}$. For the third bound, write $\|A\,W\Gamma W'\| = \|M_\delta^{-1}M\,M_\delta^{-1}W\Gamma W'\| \le \|M_\delta^{-1}M\|\,\|M_\delta^{-1}W\Gamma W'\|$. The bound then follows from (ii) and (iii). The bound for $\|W\Gamma_\delta W'\,A\|$ follows similarly.

Appendix F. Wasserstein bounds

Here, we provide details for the claim in Section 4.1 regarding the Wasserstein bound. Write

$$\Gamma W'M^{-1} - \Gamma_\delta W'M_\delta^{-1} = (\Gamma-\Gamma_\delta)W'M^{-1} + \Gamma_\delta W'M_\delta^{-1}(I - M_\delta^{-1}M).$$

Since $\|M^{-1}\| \le 1$, we can bound $\|(\Gamma-\Gamma_\delta)W'M^{-1}z\|^2 \le \|W\|^2\|z\|^2\delta^2$. Note also that $\|I - MM_\delta^{-1}\| < \delta$ and $\Gamma_\delta W'M_\delta^{-1} = \Gamma_S W_S'(I + W_S\Gamma_S W_S')^{-1} = (W_S'W_S + \Gamma_S^{-1})^{-1}W_S' \precsim (W_S'W_S)^{-1}W_S'$. In the expression for $\mathcal{W}_2^2(P_e, P_a)$, this is the dominating term, which implies the order of $\mathcal{W}_2(P_e, P_a)$ is $\|W\|\|z\|\delta$.

References

Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *arXiv preprint arXiv:1706.05439*, 2017.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon T Willard. Lasso meets Horseshoe. *arXiv preprint arXiv:1706.10179*, 2017.

Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

Anirban Bhattacharya, Antik Chakraborty, and Bani K Mallick. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991, 2016.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Antik Chakraborty, Anirban Bhattacharya, and Bani K Mallick. Bayesian sparse multiple regression for simultaneous rank reduction and variable selection. *arXiv preprint arXiv:1612.00877*, 2016.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

James M Flegal and Galin L Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070, 2010.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL `https://www.jstatsoft.org/v033/i01`.

Edward I George and Robert E McCulloch. Approaches for Bayesian variable selection. *Statistica sinica*, 7:339–373, 1997.

Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products*. Academic press, 1996.

P Richard Hahn, Jingyu He, and Hedibert F Lopes. Efficient sampling for Gaussian linear regression with arbitrary priors. *Journal of Computational and Graphical Statistics*, (to appear), 2018.

Martin Hairer and Jonathan C Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pages 109–117. Springer, 2011.

Pierre E Jacob, John O'Leary, and Yves F Atchadé. Unbiased Markov chain Monte Carlo with couplings. *arXiv preprint arXiv:1708.03625*, 2017.

James E Johndrow and Jonathan C Mattingly. Error bounds for approximations of Markov chains. *arXiv preprint arXiv:1711.05382*, 2017.

James E Johndrow, Aaron Smith, Natesh Pillai, and David B Dunson. MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, (to appear), 2018.

Valen E Johnson and David Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

Kshitij Khare and James P Hobert. Geometric ergodicity of the Bayesian lasso. *Electronic Journal of Statistics*, 7:2150–2163, 2013.

Rafail Khasminskii. *Stochastic stability of differential equations*. Springer, 1980.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning*, pages 181–189, 2014.

Xiaolei Liu, Meng Huang, Bin Fan, Edward S Buckler, and Zhiwu Zhang. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS genetics*, 12(2):e1005767, 2016.

Enes Makalic and Daniel F Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2016.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer, 1993.

Alexander Y. Mitrophanov. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Naveen N Narisetty, Juan Shen, and Xuming He. Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *Journal of the American Statistical Association*, pages 1–13, 2018.

Akihiko Nishimura and Marc A Suchard. Prior-preconditioned conjugate gradient for accelerated gibbs sampling in" large n & large p" sparse bayesian logistic regression models. *arXiv preprint arXiv:1810.12437*, 2018.

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

Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Natesh S Pillai and Aaron Smith. Ergodicity of approximate mcmc chains with applications to large data sets. *arXiv preprint arXiv:1405.0182*, 2014.

Murray Pollock, Paul Fearnhead, Adam M Johansen, and Gareth O Roberts. The scalable Langevin exact algorithm: Bayesian inference for big data. *arXiv preprint arXiv:1609.03436*, 2016.

Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. In J.M. Bernardo, M.J. Bayarri, and J.O. Berger, editors, *Bayesian Statistics*, volume 9, pages 501–538. Oxford, 2010.

Nicholas G Polson, James G Scott, and Jesse Windle. The Bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):713–733, 2014.

Gareth O. Roberts, Jeffrey S. Rosenthal, and Peter O. Schwartz. Convergence properties of perturbed Markov chains. *Journal of Applied Probability*, 35(1):1–11, 1998.

Maria C Romay, Mark J Millard, Jeffrey C Glaubitz, Jason A Peiffer, Kelly L Swarts, Terry M Casstevens, Robert J Elshire, Charlotte B Acharya, Sharon E Mitchell, Sherry A Flint-Garcia, et al. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, 14(6):R55, 2013.

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

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 24(4):2610–2639, 2018.

James G Scott and James O Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

S.L. Van Der Pas, B.J.K. Kleijn, and A.W. Van Der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8 (2):2585–2618, 2014.

Stéphanie van der Pas, Botond Szabó, and Aad van der Vaart. Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics*, 11(2):3196–3225, 2017a.

Stéphanie van der Pas, Botond Szabó, and Aad van der Vaart. Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274, 2017b.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

Andreas Winkelbauer. Moments and absolute moments of the normal distribution. *arXiv preprint arXiv:1209.4340*, 2012.

Ping Zeng and Xiang Zhou. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nature communications*, 8(1):456, 2017.

Supplementary Materials

This supplement contains derivation of the rejection sampler and some additional figures as described in the main text.

Appendix S1. Rejection sampler for local scales

Fix $\varepsilon \in (0,1)$ and consider sampling from the density

$$h_\varepsilon(t) = C_\varepsilon \frac{e^{-\varepsilon t}}{1+t}, \quad t > 0,$$

where the normalizing constant $C_\varepsilon = e^{-\varepsilon}/\mathrm{Ei}(\varepsilon)$, with $\mathrm{Ei}(x) = \int_x^\infty e^{-t}/t\, dt = \Gamma(0,x)$ the exponential integral function. The constant C_ε is a decreasing function of ε, with $C_1 \approx 1.6$ and $C_\varepsilon < 1$ for $\varepsilon < 0.40$.

First we record useful fact about the density h_ε. If $X \sim \mathrm{Expo}(\varepsilon)$ with $E(X) = 1/\varepsilon$, then $P(X > b/\varepsilon) = e^{-b}$ for any $b > 0$. We show a similar upper bound for h_ε. Bound

$$C_\varepsilon \int_{b/\varepsilon}^\infty \frac{e^{-\varepsilon t}}{1+t}\, dt \le \frac{C_\varepsilon}{1+b/\varepsilon} \int_{b/\varepsilon}^\infty e^{-\varepsilon t}\, dt = \frac{C_\varepsilon}{1+b/\varepsilon} \frac{1}{\varepsilon} e^{-b} \le C_\varepsilon \frac{e^{-b}}{b}.$$

Let

$$f(x) := f_\varepsilon(x) = \varepsilon x + \log(1+x), \quad x > 0,$$

be the negative log-density up to constants. It is easily verified that f is an increasing concave function on $(0,\infty)$. We now develop a lower bound to f.

For any real-valued function g and an interval $[\underline{v}, \bar{v}] \subset \mathrm{dom}(g)$, recall that the line segment on the interval $[\underline{v}, \bar{v}]$ joining $g(\underline{v})$ and $g(\bar{v})$ is given by

$$x \mapsto g(\underline{v}) + \frac{g(\bar{v}) - g(\underline{v})}{\bar{v} - \underline{v}} (x - \underline{v}), \quad x \in [\underline{v}, \bar{v}].$$

Fix $0 < a < 1 < b$, and set

$$A = f(a/\varepsilon),\ I = f(1/\varepsilon),\ B = f(b/\varepsilon).$$

Also, set

$$\lambda_2 = \frac{I - A}{(1-a)/\varepsilon},\ \lambda_3 = \frac{B - I}{(b-1)/\varepsilon}.$$

With these notations, set

$$f_{L,\varepsilon}(x) := f_L(x) = \begin{cases} \log(1+x) & x \in [0, a/\varepsilon), \\ A + \lambda_2(x - a/\varepsilon) & x \in [a/\varepsilon, 1/\varepsilon), \\ I + \lambda_3(x - b/\varepsilon) & x \in [1/\varepsilon, b/\varepsilon), \\ B + \varepsilon(x - b/\varepsilon) & x \ge b/\varepsilon. \end{cases}$$

Some comments about the approximation f_L. First, f_L is an increasing function and is piecewise linear on $[a/\varepsilon, \infty)$. It has a jump discontinuity at a/ε and is continuous everywhere

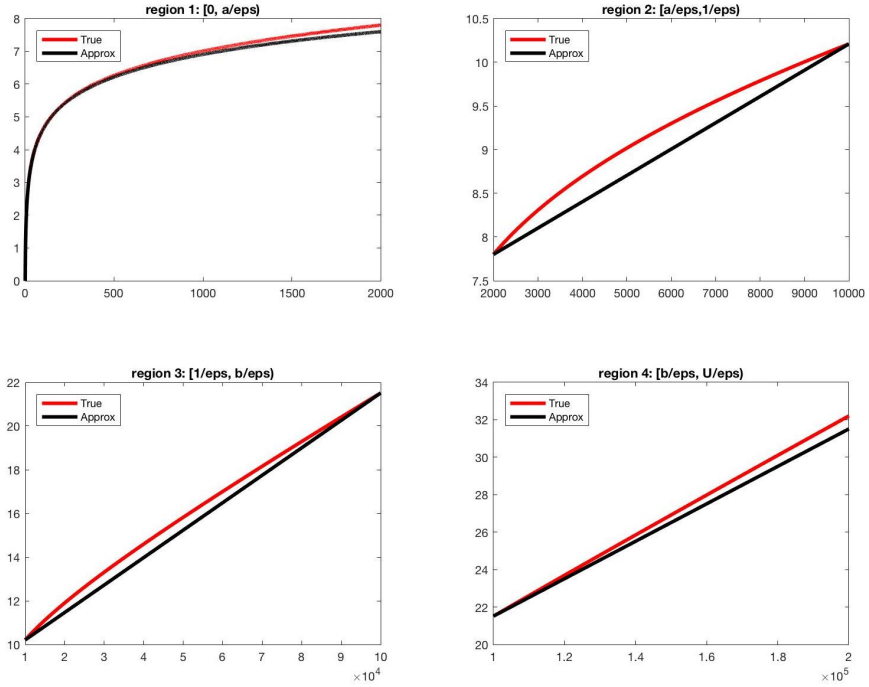

Figure S1: Comparison of f and f_L with $\varepsilon = 10^{-4}$, $a = 1/5$, and $b = 10$.

else. f_L is identical to $\log(1+x)$ on $[0, a/\varepsilon)$, linearly interpolates between (i) $f(a/\varepsilon)$ and $f(1/\varepsilon)$ on $[a/\varepsilon, 1/\varepsilon)$ and (ii) $f(1/\varepsilon)$ and $f(b/\varepsilon)$ on $[1/\varepsilon, b/\varepsilon)$, and equals $\varepsilon x + \log(1 + b/\varepsilon)$ on $[b/\varepsilon, \infty)$. By construction, $f_L \leq f$ on $[0, a/\varepsilon)$, and the concavity of f implies $f_L \leq f$ on $[a/\varepsilon, \infty)$, implying that f_L is globally bounded from above by f.

Let $h_L(x) = e^{-f_L(x)}/\nu$ for $x \in (0, \infty)$, with $\nu = \int_0^\infty e^{-f_L(x)}\, dx$. A rejection sampling algorithm to sample from h proceeds as follows:

(i) draw $z \sim h_L$ and $u \sim U(0,1)$ independently.
(ii) Accept z as a sample from h if $u < e^{-(f - f_L)(z)}$. Otherwise, back to step (i).

We now describe sampling from h_L. To that end, let us first calculate the normalizing constant ν. We have,

$$
\begin{aligned}
\nu &= \nu_1 + \nu_2 + \nu_3 + \nu_4 \\
\nu_1 &= \int_0^{a/\varepsilon} \frac{dx}{1+x} = \log(1 + a/\varepsilon), \\
\nu_2 &= e^{-A} \int_{a/\varepsilon}^{1/\varepsilon} e^{-\lambda_2 (x - a/\varepsilon)}\, dx = \lambda_2^{-1}\, e^{-A} \left[1 - e^{-(I-A)}\right], \\
\nu_3 &= e^{-I} \int_{1/\varepsilon}^{b/\varepsilon} e^{-\lambda_3 (x - 1/\varepsilon)}\, dx = \lambda_3^{-1}\, e^{-I} \left[1 - e^{-(B-I)}\right], \\
\nu_4 &= e^{-B} \int_{b/\varepsilon}^{\infty} e^{-\varepsilon (x - b/\varepsilon)}\, dx = \varepsilon^{-1}\, e^{-B}.
\end{aligned}
$$

We can thus write h_L as a mixture of four densities,

$$h_L = (\nu_1/\nu)\, h_1 + (\nu_2/\nu)\, h_2 + (\nu_3/\nu)\, h_3 + (\nu_4/\nu)\, h_4,$$

where

$$\begin{aligned}
h_1(x) &= \frac{1}{\nu_1}\, \frac{\mathbb{1}_{[0,a/\varepsilon)}(x)}{1+x}, \\
h_2(x) &= \frac{1}{\nu_2}\, e^{-A}\, e^{-\lambda_2(x-a/\varepsilon)}\, \mathbb{1}_{[a/\varepsilon,1/\varepsilon)}(x), \\
h_3(x) &= \frac{1}{\nu_3}\, e^{-I}\, e^{-\lambda_3(x-1/\varepsilon)}\, \mathbb{1}_{[1/\varepsilon,1b/\varepsilon)}(x), \\
h_4(x) &= \frac{1}{\nu_4}\, e^{-B}\, e^{-\varepsilon(x-b/\varepsilon)}\, \mathbb{1}_{[b/\varepsilon,\infty)}(x).
\end{aligned}$$

Observe that h_2, h_3 and h_4 are truncated exponential densities. We now describe the inverse cdf method to sample from a truncated exponential.

Sampling from truncated exponential. Let $\text{Expo}(\lambda, \underline{v}, \bar{v})$ denote the distribution with density

$$\gamma(x) = \frac{\lambda e^{-\lambda(x-\underline{v})}}{H}, \quad x \in [\underline{v}, \bar{v}],$$

where $\lambda > 0$, $0 \le \underline{v} < \bar{v} \le \infty$, and $H = 1 - e^{-\lambda(\bar{v}-\underline{v})}$ (Note:when $\bar{v} = \infty$, this means $H = 1$). The cdf

$$F_\gamma(x) = \frac{1 - e^{-\lambda(x-\underline{v})}}{H}, \quad x \in [\underline{v}, \bar{v}].$$

The inverse-cdf method to sample from $\text{Expo}(\lambda, \underline{v}, \bar{v})$ is then given by:
Sample $u \sim U(0,1)$ and set

$$x = \underline{v} + \frac{-\log(1-uH)}{\lambda} = \underline{v} + \frac{-\log(1-u+u\, e^{-\lambda(\bar{v}-\underline{v})})}{\lambda}.$$

After some simplification, the value of H corresponding to h_2 and h_3 is respectively,

$$H_2 = 1 - e^{-(I-A)}, \quad H_3 = 1 - e^{-(B-I)}.$$

The density h_1 can also be sampled using inverse cdf method. We have

Sampling from h_1: The cdf of h_1 is

$$F_1(x) = \frac{\log(1+x)}{\nu_1}, \quad x \in [0, a/\varepsilon).$$

The inverse cdf sampler sets:
draw $u \sim U(0,1)$ and set $x = e^{u\nu_1} - 1 = (1 + a/\varepsilon)^u - 1$.

Appendix S2. Geometric ergodicity of the exact algorithm

Consider the following slight modification to the horseshoe prior from (2)

$$\begin{aligned} \beta_j \mid \sigma^2, \eta, \xi \overset{iid}{\sim} \mathrm{N}(0, \sigma^2 \xi^{-1} \eta_j^{-1}), \quad \eta_j^{-1/2} \overset{iid}{\sim} \mathrm{Cauchy}_{[0, b^{-1/2}]}(0,1), \quad j = 1, \ldots, p, \\ \xi^{-1/2} \sim \mathrm{Cauchy}_{[a_\xi, b_\xi]}(0,1), \quad \sigma^2 \sim \mathrm{InvGamma}(\omega/2, \omega/2), \end{aligned} \tag{54}$$

where $0 \leq a_\xi < b_\xi \leq \infty$, $b \geq 0$, and $\mathrm{Cauchy}_{[\underline{a}, \overline{a}]}(0,1)$ is the standard Cauchy distribution restricted to the interval $[\underline{a}, \overline{a}]$. The original horseshoe prior Carvalho et al. (2010) corresponds to $a_\xi = b = 0, b_\xi = \infty$. The truncation of the prior on ξ was introduced in van der Pas et al. (2017a) for theoretical tractability, while we additionally truncate η_j for our convergence analysis of the MCMC algorithms developed below. As noted in the main text, this truncation retains the statistical accuracy.

Posterior sampling with this modified horseshoe prior proceeds exactly as in the blocked Gibbs sampler in (4), with the minor difference being the η_js and ξ now need to be sampled from truncated densities. Let us continue to use $\mathcal{P}$ to denote the Markov transition operator corresponding to the modified update rules.

Theorem 14 *For any $c < 1/2$*

1. *The function*

$$V(\eta, \beta, \sigma^2, \xi) = \frac{\|W\beta\|^2}{\sigma^2} + \xi^2 + \sum_{j=1}^{p} \left[\frac{\sigma^{2c}}{|\beta_j|^{2c}} + \frac{\eta_j^c |\beta_j|^c}{\sigma^c} + \eta_j^c \right] \tag{55}$$

   is a Lyapunov function of $\mathcal{P}$, even if no truncation of the prior on η is used (i.e. $b = 0$ in (2)).

2. *If $b > 0$ in (2), $\mathcal{P}$ is geometrically ergodic in the sense of (12).*

Remark 15 *If $p \leq N$ and W is full-rank, then $\mathcal{P}$ is geometrically ergodic without any truncation of the prior on η. That is, one can take the constant $b = 0$ in (2).*

The Lyapunov function in (55) is somewhat unusual in that – as a function of $\beta_j^2 \sigma^{-2}$ – it both grows at infinity and has a pole at zero, whereas most commonly encountered Lyapunov functions simply grow at infinity and are bounded on compact sets containing the origin. This is necessary because showing the minorization condition in Assumption 3.2 requires a uniform bound on the total variation distance between any two densities in the set

$$\left\{ p_1(\eta_j \mid \beta_j, \sigma^2, \xi) = \frac{e^{-m_j}}{\Gamma\left(0, m_j(1+b)\right)} \frac{1}{1+\eta_j} e^{-m_j \eta_j} \mathbb{1}\{\eta_j > b\} : (\beta, \sigma^2, \xi) \in \mathcal{S}(R) \right\}$$

for any $0 < R < \infty$ and every j, where $\Gamma(0, x)$ is the upper incomplete gamma function defined in (50), $\mathcal{S}(R)$ is as defined in Assumption 3.2 and $m_j = \beta_j^2 \xi / (2\sigma^2)$. Clearly, minorization requires bounding $\beta_j^2 \sigma^{-2}$ away from infinity inside sublevel sets $\mathcal{S}(R)$ of V, since $p_1(\eta_j \mid \beta_j, \sigma^2, \xi) \to \delta_0(\eta_j)$ as $\beta_j^2 \sigma^{-2} \to \infty$, and $\delta_0(\eta_j)$ has total variation distance 1 from

every measure with a continuous density. We must also bound $\beta_j^2\sigma^{-2}$ away from zero for every j, since the limiting distribution is improper there. So the Lyapunov function must have sublevel sets in which $\beta_j^2\sigma^{-2}$ is bounded away from both zero and infinity. The former is accomplished by the term $\sum_j \sigma^{2c}|\beta_j|^{-2c}$ regardless of the values of N and p. However, truncation of the prior on η is needed to achieve the latter. If $p > N$, then the function $\|W\beta\|^2\sigma^{-2}$ is constant in the kernel of the linear function $W : \mathbb{R}^p \to \mathbb{R}^N$, and the only other appearance of positive powers of β in (55) is in the term $\eta_j^c|\beta_j|^c\sigma^{-c}$. Thus, in order to ensure that β cannot go to infinity in the kernel of W, we must have η_j bounded away from zero in sublevel sets. In contrast, when $p \leq N$ and W is full rank, the term $\|W\beta\|^2\sigma^{-2}$ is enough to keep $\beta_j^2\sigma^{-2}$ bounded away from infinity in sublevel sets.

Proof We prove the second assertion, that is, prove geometric ergodicity of $\mathcal{P}$ for $b > 0$ by verifying the Lyapunov condition in Assumption 3.1 and the minorization condition in Assumption 3.2. An inspection of the verification of the Lyapunov condition will show that the same proof continues to work for $b = 0$ with some minor modifications.

S2.1. Lyapunov condition for the exact chain

We first show that

$$V\big(\eta, \beta, \sigma^2, \xi\big) = \frac{\|W\beta\|^2}{\sigma^2} + \sum_{j=1}^{p} \big[\sigma^{2c}|\beta_j|^{-2c} + \sigma^{-c}\eta_j^c|\beta_j|^c + \eta_j^c\big] + \xi^2 \tag{56}$$

is a Lyapunov function for $\mathcal{P}$ for any $c \in (0, 1/2)$. We have,

$$\begin{aligned}
(\mathcal{P}V)(\tilde{\eta}, x_{\setminus\eta}) &= \int V(\eta, y_{\setminus\eta})\, p\big((\tilde{\eta}, x_{\setminus\eta}), (\eta, y_{\setminus\eta})\big) d\eta dy_{\setminus\eta} \\
&= \int_\eta \bigg[\int_{y_{\setminus\eta}} V(\eta, y_{\setminus\eta}) p_2(y_{\setminus\eta} \mid \eta, \tilde{\xi})\, dy_{\setminus\eta} \bigg] p_1(\eta \mid x_{\setminus\eta})\, d\eta \\
&= \int V_1(\eta, \tilde{\xi}) p_1(\eta \mid x_{\setminus\eta})\, d\eta,
\end{aligned} \tag{57}$$

where

$$\begin{aligned}
V_1(\eta, \tilde{\xi}) = \mathbf{E}\left(\frac{\|W\beta\|^2}{\sigma^2} \,\Big|\, \eta, \tilde{\xi} \right) + \mathbf{E}(\xi^2 \mid \eta, \tilde{\xi}) \\
+ \sum_{j=1}^{p} \big[\mathbf{E}(\sigma^{2c}|\beta_j|^{-2c} \mid \eta, \tilde{\xi}) + \mathbf{E}(\sigma^{-c}|\beta_j|^c \mid \eta, \tilde{\xi}) + \eta_j^c \big].
\end{aligned} \tag{58}$$

We now aim to bound $V_1(\eta, \tilde{\xi})$. We shall make repeated use of the fact that

$$\mathbf{E}[g(\beta)h(\sigma^2) \mid \eta] = \mathbf{E}\Big[h(\sigma^2)\, \mathbf{E}[g(\beta) \mid \sigma^2, \xi, \eta] \,\Big|\, \eta \Big]$$

for integrable functions g and h, using the tower property of conditional expectations.

We begin by integrating over β to bound $\mathbf{E}(\|W\beta\|^2 \mid \sigma^2, \xi, \eta)$, $\mathbf{E}(|\beta_j|^{-2c} \mid \sigma^2, \xi, \eta)$, and $\mathbf{E}(|\beta_j|^c \mid \sigma^2, \xi, \eta)$, respectively. We first show that

$$\mathbf{E}(\|W\beta\|^2 \mid \sigma^2, \xi, \eta) \leq \|z\|^2 + N\sigma^2. \tag{59}$$

To that end, we have, from (28),

$$\mathbf{E}(\|W\beta\|^2 \mid \eta) = \mu' W'W\mu + \sigma^2 \operatorname{tr}[(W'W)\Sigma] = \|W\Sigma W'z\|^2 + \sigma^2 \operatorname{tr}[W\Sigma W'].$$

Let us now calculate $W\Sigma W'$. Recall,

$$\Gamma = \xi^{-1}D, \quad M = I_N + W\Gamma W', \quad \Sigma = \Gamma - \Gamma W' M^{-1} W\Gamma,$$

where the last equality follows from the Woodbury matrix identity. Then,

$$\begin{aligned} W\Sigma W' &= W\Gamma W' - W\Gamma W' M^{-1} W\Gamma W' = W\Gamma W'\big[I_N - M^{-1}W\Gamma W'\big] \\ &= W\Gamma W' M^{-1} = I_N - M^{-1}, \end{aligned}$$

where we have used that $M^{-1}W\Gamma W' = W\Gamma W' M^{-1} = I_N - M^{-1}$. We then have $\|W\Sigma W'z\|^2 = \|(I_N - M^{-1})z\|^2 \leq \|z\|^2$, and $\operatorname{tr}(W\Sigma W') \leq N$, delivering (59).

We next focus on $\mathbf{E}(|\beta_j|^{-2c} \mid \sigma^2, \xi, \eta)$ and show that for universal constants $0 < C_1, C_2 < \infty$,

$$\mathbf{E}\big(|\beta_j|^{-2c} \mid \sigma^2, \xi, \eta\big) \leq \sigma^{-2c}(C_1 \eta_j^c + C_2). \tag{60}$$

A formula for negative absolute moments of Gaussians is available from Gradshteyn and Ryzhik (1996) and recorded in equation (53) in Section E.2. Specifically, let μ_j and σ_j^2 respectively denote the jth entry of μ and the jth diagonal entry of Σ in (28). We then have, from (53), that

$$\mathbf{E}\big(|\beta_j|^{-2c} \mid \sigma^2, \xi, \eta\big) = \sigma^{-2c}\sigma_j^{-2c}\,\frac{2^{-c}\,\Gamma\left(\frac{1-2c}{2}\right)}{\sqrt{\pi}}\, e^{-\frac{\mu_j^2}{2\sigma^2\sigma_j^2}}\, M\left(\frac{1-2c}{2}, \frac{1}{2}; \frac{\mu_j^2}{2\sigma^2\sigma_j^2}\right),$$

where $M(\cdot,\cdot;\cdot)$ is the confluent hypergeometric function of the first kind; see Section E.2 for definition and properties. Since $c \in (0, 1/2)$, the condition of Lemma 13 is satisfied, so that we can bound

$$e^{-\frac{\mu_j^2}{2\sigma^2\sigma_j^2}}\, M\left(\frac{1-2c}{2}, \frac{1}{2}; \frac{\mu_j^2}{2\sigma^2\sigma_j^2}\right) \leq 1.$$

This implies

$$\mathbf{E}\big(|\beta_j|^{-2c} \mid \sigma^2, \xi, \eta\big) \leq \tilde{C}_1 \sigma^{-2c}\sigma_j^{-2c}, \tag{61}$$

where $\tilde{C}_1 = \pi^{-1/2}2^{-c}\Gamma(1/2 - c)$.

To bound the right hand side of (61), we need a lower bound on σ_j^2. To that end, we have $s_{\max}^2(W)\, I_p + (\xi^{-1}D)^{-1} \succeq W'W + (\xi^{-1}D)^{-1}$, where $A \succeq B$ denotes $(A - B)$

is nonnegative definite (nnd). Using the fact that $A \succeq B$ implies $B^{-1} \succeq A^{-1}$, we have $\Sigma \succeq (s^2_{\max}(W)\, I_p + (\xi^{-1}D)^{-1})^{-1}$. Next, use the fact that if $A \succeq B$, then $a_{jj} \geq b_{jj}$, since $(a_{jj} - b_{jj}) = e_j'(A-B)e_j \geq 0$ with e_j the jth unit vector. This implies

$$\sigma_j^2 \geq \frac{1}{s^2_{\max}(W) + \xi\eta_j}, \quad \sigma_j^{-2c} \leq (s^2_{\max}(W) + \xi\eta_j)^c \leq (s^{2c}_{\max}(W) + b_\xi^c \eta_j^c),$$

where in the last step, we used that for $a, b > 0$ and $c \in (0, 1/2)$, $(a+b)^c \leq (a^c + b^c)$, and that $\xi \leq b_\xi$. Substitute the bound in (61) to obtain (60).

Next, we consider $\mathbf{E}(|\beta_j|^c \mid \sigma^2, \xi, \eta)$, and show that there exist universal constants $0 < C_3, C_4 < \infty$ such that

$$\sum_{j=1}^p \eta_j^c \mathbf{E}(|\beta_j|^c \mid \sigma^2, \xi, \eta) \leq C_3 \sigma^c \sum_{j=1}^p (\eta_j^c + 1) + C_4. \tag{62}$$

We have, using that $x \mapsto x^{2/c}$ is convex for $x > 0$, that $\big[\mathbf{E}(|\beta_j|^c \mid \sigma^2, \xi, \eta)\big]^{2/c} \leq \mathbf{E}(\beta_j^2 \mid \sigma^2, \xi, \eta) = \mu_j^2 + \sigma^2\sigma_j^2$, and hence, $\mathbf{E}(|\beta_j|^c \mid \sigma^2, \xi, \eta) \leq (\mu_j^2 + \sigma^2\sigma_j^2)^{c/2} \leq |\mu_j|^c + \sigma^c(\sigma_j^2)^{c/2}$. Following a similar argument as in the paragraph after (61), $(W'W + (\xi^{-1}D)^{-1})^{-1} \preceq \xi^{-1}D$, implying $\sigma_j^2 \leq (\xi\eta_j)^{-1}$. These together imply,

$$\sum_{j=1}^p \eta_j^c \mathbf{E}(|\beta_j|^c \mid \sigma^2, \xi, \eta) \leq \sum_{j=1}^p \eta_j^c \{|\mu|_j^c + \sigma^c\, (\sigma_j^2)^{c/2}\} \leq \sum_{j=1}^p \eta_j^c |\mu_j|^c + C\sigma^c \sum_{j=1}^p \eta_j^{c/2}.$$

for a universal constant C. Next, using Hölder's inequality,

$$\sum_{j=1}^p \eta_j^c |\mu_j|^c \leq \Big[\sum_{j=1}^p (\eta_j^c |\mu_j|^c)^{2/c}\Big]^{c/2} p^{1-c/2} = p^{1-c/2}\, (\|D^{-1}\mu\|^2)^{c/2}.$$

We have $\|D^{-1}\mu\|^2 = w' D^{-1}\Sigma^2 D^{-1} w$, with $w = W'z$. Since $\Sigma^2 = (W'W + (\xi^{-1}D)^{-1})^{-2} \preceq \xi^2 D^2$, we have $D^{-1}\Sigma^2 D^{-1} \preceq \xi^{-2} I_p$, where we have used that if $B_1 \preceq B_2$, and B_1, B_2, A are positive definite (pd), then $AB_1A \preceq AB_2A$. This implies $\|D^{-1}\mu\|^2 \leq \xi^{-2}\|w\|^2 \leq a_\xi^{-2}\|w\|^2$. Cascading this bound through the previous two displays, and using the inequality $x^{c/2} \leq (x^c + 1)$ for $x > 0$, (62) is obtained.

Combining the bounds (62), (60), and (59), we obtain for universal constants $0 < C_5, C_6, C_7, C_8 < \infty$ not depending on η

$$\mathbf{E}[V \mid \sigma^2, \xi, \eta] \leq C_5 \sum_{j=1}^p \eta_j^c + \frac{C_6}{\sigma^2} + \frac{C_7}{\sigma^c} + \xi^2 + \tilde{C}_7$$

Next, take an expectation w.r.t. $\sigma^2 \mid \xi, \eta$. Note that $\mathbf{E}(1/\sigma^2 \mid \xi, \eta) = (N+a)/(z'M^{-1}z + b) \leq (N+a)/b$, and similarly, $\mathbf{E}(1/\sigma^c \mid \xi, \eta)$ is also bounded above by a constant not depending on ξ, η. This leads to

$$V_1(\eta, \tilde{\xi}) = \mathbf{E}[V \mid \eta, \tilde{\xi}] \leq C_5 \sum_{j=1}^p \eta_j^c + C_8, \tag{63}$$

for a universal constant C_8 not depending on $\tilde{\xi}$, where we additionally used that ξ is compactly supported. Notice that while V_1 is a function of $\tilde{\xi}$, the upper bound on the right side is not, a consequence of the fact that $\xi \in [a_\xi, b_\xi]$ for $0 < a_\xi < b_\xi < \infty$.

We now proceed to bound $\mathbf{E}\big(V_1(\eta, \tilde{\xi}) \mid x_{\setminus\eta}\big) = \int V_1(\eta, \tilde{\xi})\, p_2(\eta \mid x_{\setminus\eta})$. For a small $\varepsilon > 0$ to be chosen later, bound

$$\begin{aligned}
\int \eta_j^c\, p(\eta_j \mid x_{\setminus\eta})\, d\eta_j &= \frac{e^{-\tilde{m}_j}}{\Gamma(0, \tilde{m}_j + b\tilde{m}_j)} \int_b^\infty \eta_j^c \frac{e^{-\tilde{m}_j \eta_j}}{1+\eta_j}\, d\eta_j \\
&\le \frac{e^{-\tilde{m}_j}}{\Gamma(0, \tilde{m}_j + b\tilde{m}_j)} \int_b^\infty \eta_j^{c-1} e^{-\tilde{m}_j \eta_j}\, d\eta_j \\
&= \frac{e^{-\tilde{m}_j}}{m_j^c} \frac{\Gamma(c, b\tilde{m}_j)}{\Gamma(0, \tilde{m}_j + b\tilde{m}_j)} \\
&\le \varepsilon \tilde{m}_j^{-c} + C_\varepsilon \\
&\le (a_\xi/2)^{-c} \varepsilon\, \tilde{\sigma}^{2c} |\tilde{\beta}_j|^{-2c} + C_\varepsilon
\end{aligned}$$

In the first inequality, we used the bound $\eta_j/(1+\eta_j) < 1$ for $\eta_j \in (b, \infty)$, while the penultimate inequality follows from Lemma 12. From (63), we then obtain

$$\mathbf{E}\big(V_1(\eta, \tilde{\xi}) \mid x_{\setminus\eta}\big) \le \sum_{j=1}^p (a_\xi/2)^{-c} C_5\, \varepsilon\, \sigma^{2c} |\beta_j|^{-2c} + C_9.$$

Now pick ε such that $(a_\xi/2)^{-c} C_5\, \varepsilon < 1$, and we have proved that V is Lyapunov.

S2.2. Minorization condition on sublevel sets

Consider sublevel sets of the Lyapunov function in (55):

$$\mathcal{S}(R) := \left\{ x : \frac{\|W\beta\|^2}{\sigma^2} + \sum_{j=1}^p \left[\sigma^{2c}|\beta_j|^{-2c} + \sigma^{-c}\eta_j^c |\beta_j|^c + \eta_j^c\right] + \xi^2 < R \right\}$$

where $D = \mathrm{diag}(\eta^{-1})$. Consider two points $x, y \in \mathbf{X}$. Observe that

$$\begin{aligned}
\|\delta_x \mathcal{P} - \delta_y \mathcal{P}\|_{\mathrm{TV}} &= \int |p_1(\eta \mid x_{\setminus\eta}) p_2(z_{\setminus\eta} \mid \eta, \tilde{\xi}) - p_1(\eta \mid z_{\setminus\eta}) p_2(z_{\setminus\eta} \mid \eta, \tilde{\xi})| dz_{\setminus\eta} d\eta \\
&= \int p_2(z_{\setminus\eta} \mid \eta, \tilde{\xi}) |p_1(\eta \mid x_{\setminus\eta}) - p_1(\eta \mid y_{\setminus\eta})| dz_{\setminus\eta} d\eta \\
&= \int |p_1(\eta \mid x_{\setminus\eta}) - p_1(\eta \mid y_{\setminus\eta})| d\eta \\
&= \|\delta_{x_{\setminus\eta}} \mathcal{P}_1 - \delta_{y_{\setminus\eta}} \mathcal{P}_1\|_{\mathrm{TV}},
\end{aligned}$$

the total variation distance just between the η conditionals started at two points $x_{\setminus\eta}, y_{\setminus\eta}$.

Let

$$S_{\setminus\eta}(R) = \{x_{\setminus\eta} \in \mathbf{X}_2 : (x_{\setminus\eta}, \eta) \in \mathcal{S}(R) \text{ for some } \eta \in \mathbf{X}_1\}$$

Consider a point $x_{\setminus\eta} \in \mathcal{S}_{\setminus\eta}(R)$. Any such point must satisfy

$$\begin{aligned}
\frac{|\beta_j|^c}{\sigma^c}\eta^c < R \Rightarrow \frac{\beta_j^2\xi}{\sigma^2} < b_\xi R^{2/c}\eta^{-2} \quad j = 1, \ldots, p \\
\frac{\sigma^{2c}}{|\beta_j|^{2c}} < R \Rightarrow \frac{\beta_j^2\xi}{\sigma^2} > a_\xi R^{-1/c} \quad j = 1, \ldots, p \\
\text{if } p \leq N \text{ then } \frac{\|W\beta\|^2}{\sigma^2} < R \Rightarrow \frac{\beta_j^2\xi}{\sigma^2} < b_\xi R \quad j = 1, \ldots, p
\end{aligned}$$

It follows that when $p \leq N$, $a_\xi R^{-1/c} < \beta_j^2\xi\sigma^{-2} < b_\xi R$ for every $x_{\setminus\eta} \in \mathcal{S}_{\setminus\eta}(R)$. Moreover, if $p > N$ and the prior on η is truncated below by b, then we have $a_\xi R^{-1/c} < \beta_j^2\xi\sigma^{-2} < b_\xi R^{2/c}b^{-2}$ for every $x_{\setminus\eta} \in \mathcal{S}_{\setminus\eta}(R)$.

The remainder of the proof uses the upper bound $b_\xi R^{2/c}b^{-2}$ from the $p > N$ case; the proof for the $p \leq N$ case is virtually identical and omitted. Define the interval

$$I(R) = \left[\frac{1}{2}a_\xi R^{-1/c}, \frac{1}{2}b_\xi R^{2/c}b^{-2}\right]$$

and collection of densities corresponding to the full conditional of $\eta_j \mid x_{\setminus\eta}$ for a generic η_j

$$\mathcal{F}(R) = \left\{ f_{m_j}(\eta_j) = \frac{e^{-m_j}}{\Gamma(0, m_j(1+b))} \frac{e^{-m_j\eta_j}}{1+\eta_j}\mathbb{1}\{\eta_j > b\}, m_j \in I(R) \right\},$$

and recall that $m_j = \beta_j^2\xi\sigma^{-2}$. We have that for any $m_j \in I(R)$

$$\frac{e^{-m_j}}{\Gamma(0, m_j(1+b))} \frac{e^{-m_j\eta_j}}{1+\eta_j}\mathbb{1}\{\eta_j > b\} \geq \frac{e^{-\frac{1}{2}b_\xi R^{2/c}b^{-2}}}{\Gamma(0, \frac{1}{2}a_\xi R^{-1/c}(1+b))} \frac{e^{-\frac{1}{2}b_\xi R^{2/c}b^{-2}\eta_j}}{1+\eta_j}\mathbb{1}\{\eta_j > b\}$$

and since the function $e^{-cx}/(1+x)$ for $c > 0$ is monotone decreasing in x, it follows

$$\inf_{\substack{\eta_j \in (b, b+1] \\ m_j \in I(R)}} f_{m_j}(\eta_j) \geq \frac{e^{-\frac{1}{2}b_\xi R^{2/c}b^{-2}}}{\Gamma(0, \frac{1}{2}a_\xi R^{-1/c}(1+b))} \frac{e^{-\frac{1}{2}b_\xi R^{2/c}b^{-2}(b+1)}}{2+b}\mathbb{1}\{\eta_j > b\}.$$

Now since the transition density corresponding to $\mathcal{P}_1$ can be written $p_1(x_{\setminus\eta}, \eta) = \prod_{j=1}^p f_{m_j}(\eta_j)$, we have, with $m = (m_1, \ldots, m_p)$,

$$\inf_{\substack{\eta_j \in (b, b+1]^p \\ m_j \in I(R)^p}} p_1(x_{\setminus\eta}, \eta) \geq \frac{e^{-\frac{p}{2}b_\xi R^{2/c}b^{-2}}}{\Gamma^p(0, \frac{1}{2}a_\xi R^{-1/c}(1+b))} \frac{e^{-\frac{p}{2}b_\xi R^{2/c}b^{-2}(b+1)}}{(2+b)^p} \equiv C(R) > 0.$$

Define

$$\mathcal{M}(R) = \{m : m_j = \beta_j^2\xi\sigma^{-2} \text{ for some } (\beta, \xi, \sigma^2) \in \mathcal{S}_{\setminus\eta}(R)\},$$

and observe that $\mathcal{M}(R) \subset I(R)^p$. It follows that

$$\inf_{x_{\setminus\eta}, y_{\setminus\eta} \in \mathcal{S}_{\setminus\eta}(R)} \int_{(b,b+1]^p} (p_1(x_{\setminus\eta}, \eta) \wedge p_1(y_{\setminus\eta}, \eta))d\eta \geq C(R) \int_{(b,b+1]^p} d\eta = C(R),$$

so

$$\begin{aligned}\sup_{x_{\setminus\eta}, y_{\setminus\eta} \in \mathcal{S}_{\setminus\eta}(R)} \|\delta_{x_{\setminus\eta}} \mathcal{P}_1 - \delta_{y_{\setminus\eta}} \mathcal{P}_1\|_{\text{TV}} &= 1 - \inf_{x_{\setminus\eta}, y_{\setminus\eta} \in \mathcal{S}_{\setminus\eta}(R)} \int_{(b,\infty)^p} (p_1(x_{\setminus\eta}, \eta) \wedge p_1(y_{\setminus\eta}, \eta)) d\eta \\ &\leq 1 - \inf_{x_{\setminus\eta}, y_{\setminus\eta} \in \mathcal{S}_{\setminus\eta}(R)} \int_{(b,b+1]^p} (p_1(x_{\setminus\eta}, \eta) \wedge p_1(y_{\setminus\eta}, \eta)) d\eta \\ &\leq 1 - C(R) < 1,\end{aligned}$$

completing the proof. ∎

Appendix S3. Extra Figures

Here we provide additional figures relevant to the statistical performance of time-averaging estimators from the Approximate algorithm.

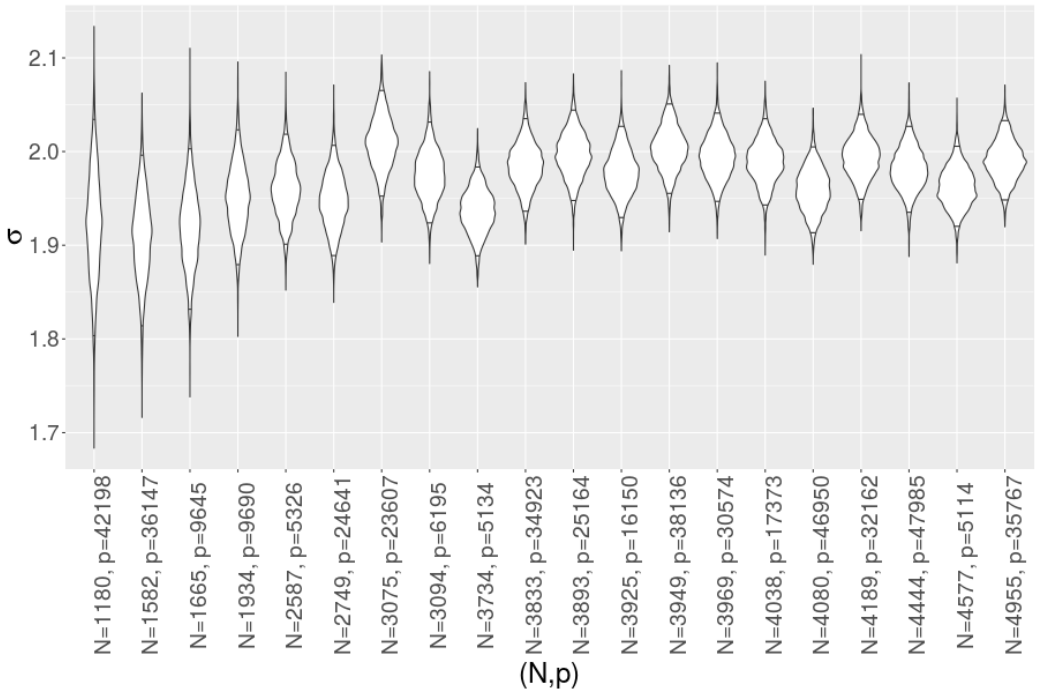

Figure S2: Marginals for the residual standard deviation σ over 20 values of N, p using the approximate algorithm. The small horizontal lines indicate the 0.025 and 0.975 approximate posterior quantiles. The true value is 2 in all cases.

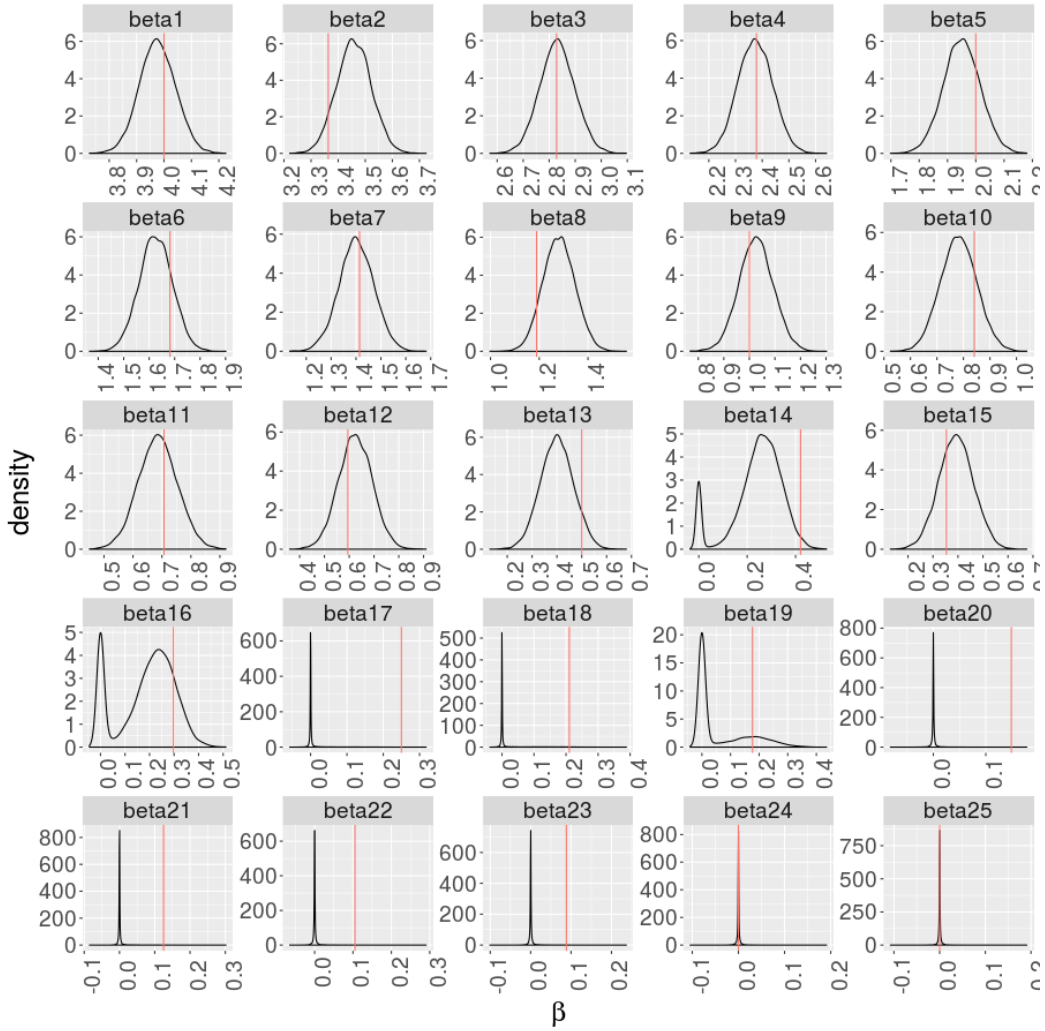

Figure S3: Marginals for the first 25 entries of β for $N = 1000, p = 5000$, true value indicated with red line. Approximate algorithm.

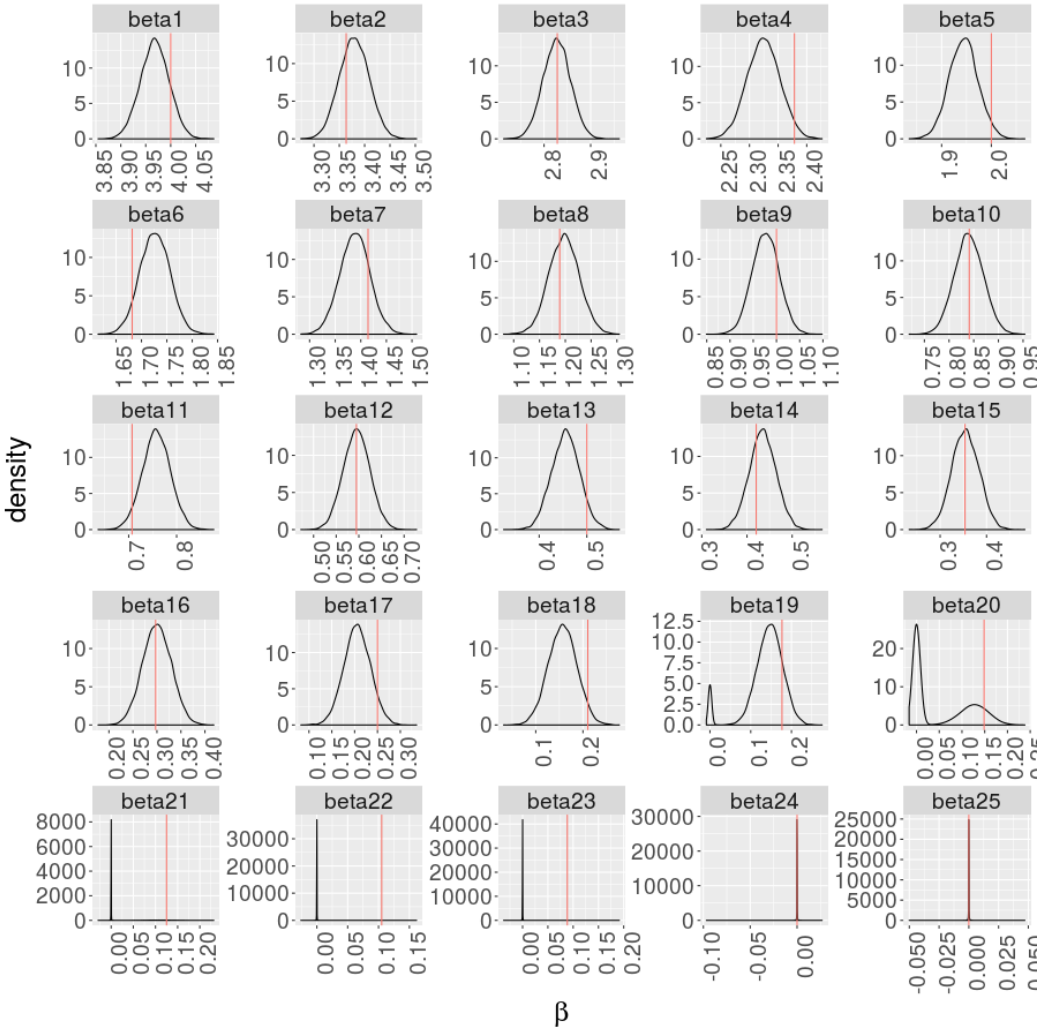

Figure S4: Marginals for the first 25 entries of β for $N = 5000, p = 50,000$, true value indicated with red line. Approximate algorithm.