Bayesian Dynamic Feature Partitioning in High-Dimensional Regression with Big Data

Rene Gutierrez*‡ Rajarshi Guhaniyogi*§

June 6, 2020

Abstract

Bayesian computation of high dimensional linear regression models using Markov Chain Monte Carlo (MCMC) or its variants can be extremely slow or completely prohibitive since these methods perform costly computations at each iteration of the sampling chain. Furthermore, this computational cost cannot usually be efficiently divided across a parallel architecture. These problems are aggravated if the data size is large or data arrive sequentially over time (streaming or online settings). This article proposes a novel dynamic feature partitioned regression (DFP) approach for efficient online inference for high dimensional linear regressions with large or streaming data. DFP constructs a *pseudo posterior density* of the parameters at every time point, followed by quickly updating the pseudo posterior when a new block of data (data shard) arrives. DFP updates the pseudo posterior at every time point suitably and partitions the parameter space to exploit parallelization for efficient posterior computation. The proposed approach is applied to high dimensional linear regression models with Gaussian scale mixture priors and spike and slab priors on large parameter spaces, along with large data, and is found to yield state-of-the-art inferential performance. The algorithm enjoys theoretical support with pseudo posterior densities over time being

‡Rene Gutierrez, Ph.D. Student, Department of Statistics, UC Santa Cruz (E-mail: rgutie17@ucsc.edu).
§Rajarshi Guhaniyogi, Assistant Professor, Department of Statistics, SOE2, UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 (E-mail: rguhaniy@ucsc.edu).

arbitrarily close to the full posterior as the data size grows, as shown in the supplementary material.

Key Words: Bayesian Statistics; Data Shards; High Dimensional Regression; Sufficient Statistics; Streaming Data; Shrinkage Prior.

1 Introduction

With recent technological progress, data containing a large number of predictors (a couple of thousand or more) are ubiquitous. In such settings, it is commonly of interest to consider the linear regression model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon, \;\; \epsilon \sim N(0, \sigma^2), \tag{1}$$

where $\boldsymbol{x}$ is a $p \times 1$ predictor, $\boldsymbol{\beta}$ is the corresponding $p \times 1$ coefficient, y is the continuous response and σ^2 is the error variance. Bayesian methods for estimating $\boldsymbol{\beta}$ provide a natural probabilistic characterization of uncertainty in the parameters and in predictions. Fitting Bayesian linear regression models in presence of very high dimensional predictors presents onerous computational burdens either due to decomposition of large matrices or due to poor convergence and inferential issues caused by the high correlations among the parameters. This article develops a dynamic approach, called Dynamic Feature Partitioning (DFP), for boosting the scalability of high dimensional Bayesian linear models for large/streaming data.

Broadly, two classes of prior distributions on $\boldsymbol{\beta}$ are typically employed in high dimensional regression literature. The traditional approach is to develop a discrete mixture of prior distributions (George and McCulloch, 1997; Scott and Berger, 2010). These methods enjoy the advantage of inducing exact sparsity for a subset of parameters and minimax rate of posterior contraction (Castillo et al., 2015) in high dimensional regression, but face computational challenges when the number of predictors is even moderately large. As an alternative to this approach, continuous shrinkage priors (Armagan et al., 2013; Carvalho et al., 2010; Caron and Doucet, 2008) have emerged which induce approximate sparsity in high-dimensional parameters. Such prior distributions can mostly be expressed as global-

∗**These authors contributed equally**

local scale mixtures of Gaussians (Polson and Scott, 2010) and offer an approximation to the operating characteristics of discrete mixture priors. Global-local priors allow parameters to be updated in blocks via a fairly automatic Gibbs sampler that leads to rapid mixing and convergence of the resulting Gibbs sampler. However, unless care is exercised, sampling can be expensive for large values of p. In fact, existing algorithms (Rue, 2001) to sample from the full conditional posterior of $\boldsymbol{\beta}$ require storing and computing the Cholesky decomposition of a $p \times p$ matrix, that necessitates p^3 floating point operations, which can be severely prohibitive for large p. There are available linear algebra artifacts such as the Sherman-Woodbury-Morrison matrix identity (Hager, 1989) to enable efficient computations in high dimensional regressions involving small n and large p, though it is less clear as to how these approaches can be adapted when the number of samples is massive to start with, or data is observed in a stream. Besides, having small sample size may limit the inferential accuracy for large p.

In fact, when the number of observations is massive, data processing and computational bottlenecks render all the above mentioned methods for high dimensional regression infeasible as they demand likelihood evaluations for updating model parameters at every sampling iteration, which can be costly. Matters are more complicated in the case of streaming data, where the posterior distribution changes once a new data shard arrives, so that the MCMC samples from the posterior distribution up to the last time point become useless.

We propose a novel online Bayesian sampling algorithm, referred to as Dynamic Feature Partitioning (DFP) that enables efficient computation of high dimensional regression in the presence of a large number of parameters and a large sample size. DFP splits a large dataset into a large number of moderately sized data shards and sequentially feeds data shards to the model. The DFP framework *dynamically* partitions the set of parameters with the onset of a new data shard, draws samples from the conditional posterior distribution in each partition, but instead of conditioning on all parameters, conditions on functions of sequential point estimates of parameters from other partitions along with the sufficient statistics from the observed data. This leads to an approximation of the conditional posterior distributions that enables rapid computation by parallelizing posterior updates of partitions in different processors. Additionally, it eliminates the need to store the entire data in time (process the

entire dataset at once), and leads to an approximation of the conditional distributions that produces samples from the correct target posterior asymptotically. The DFP algorithm is demonstrated to be highly versatile and efficient across a variety of high dimensional linear regression settings, enabling online sampling of parameters with dramatic reductions in the per-iteration computational requirement.

The rest of the article is organized as follows. Section 2 introduces a number of shrinkage priors and variable selection priors in high dimensional regression and describes the computational challenges with big n and p. Section 3 introduces the DFP algorithm and additionally provides a brief overview of the existing literature on high dimensional models with big data to highlight our contribution. Section 4 demonstrates the performance of DFP for high dimensional linear regression with (1) the Bayesian Lasso and (2) the Horseshoe shrinkage prior distributions and (3) the Spike and Lasso discrete mixture prior distribution for variable selection (described in Section 2.3). Further evidence on the empirical performance of DFP is provided in the analysis of a financial dataset consisting of the minute by minute average log-prices of the NASDAQ stock exchange from September 10 2018 to November 13 2018 during trading hours in Section 5. Finally, Section 6 concludes the article with discussions and possibilities of future directions. Theoretical insights into the convergence behavior of the DFP algorithm are provided in the supplementary material.

2 Computational Challenges in the High-Dimensional Regression Models

This section motivates the need for the dynamic feature partitioning algorithm by highlighting the issues with drawing online inference in Bayesian high dimensional linear models with big or streaming data. Let $\boldsymbol{D}_t = \{\boldsymbol{X}_t, \boldsymbol{y}_t\}$ be the data (responses and predictors) shard observed at time t and $\boldsymbol{D}^{(t)} = \{\boldsymbol{D}_s, s = 1, \ldots, t\}$ denote the data observed through time t, $t = 1, ..., T$. We assume that shards are of equal size, with each shard containing n samples, i.e., $\boldsymbol{X}_t$ is of dimension $n \times p$ and $\boldsymbol{y}_t$ is of dimension $n \times 1$. We emphasize that such an assumption is not required for the algorithmic development in the next section and is kept merely to simplify notations.

In the context of the linear regression model in (1), without the focus being on regularization or variable selection, a Bayesian hierarchical model is set up by assigning a prior $\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \sigma^2\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ and $\sigma^2 \sim IG(a, b)$. With data $\boldsymbol{D}^{(t)}$ observed through time t, the marginal posterior density of parameters σ^2 and $\boldsymbol{\beta}$ at time t appear in closed form and are given by $IG(a_t^*, b_t^*)$ and $Multivariate - t_{2a_t^*}(\boldsymbol{\mu}_t^*, (b_t^*/a_t^*)\boldsymbol{V}_t^*)$ respectively, where $a_t^* = a + \frac{nt}{2}$, $\boldsymbol{\mu}_t^* = (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \sum_{s=1}^t \boldsymbol{X}'_s\boldsymbol{X}_s)^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \sum_{s=1}^t \boldsymbol{X}'_s\boldsymbol{y}_s)$, $\boldsymbol{V}_t^* = (\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \sum_{s=1}^t \boldsymbol{X}'_s\boldsymbol{X}_s)^{-1}$, $b_t^* = b + \frac{\boldsymbol{\mu}'_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \sum_{s=1}^t \boldsymbol{y}'_s\boldsymbol{y}_s - \boldsymbol{\mu}_t^{*'}\boldsymbol{V}_t^{*-1}\boldsymbol{\mu}_t^*}{2}$. Notably, posterior distributions depend on the data only through the three sufficient statistics $\sum_{s=1}^t \boldsymbol{X}'_s\boldsymbol{X}_s$, $\sum_{s=1}^t \boldsymbol{X}'_s\boldsymbol{y}_s$ and $\sum_{s=1}^t \boldsymbol{y}'_s\boldsymbol{y}_s$. Hence posterior distribution at time t with the onset of data $\boldsymbol{D}_t$ can readily be constructed by storing and updating the sufficient statistics without having the need to store the entire data $\boldsymbol{D}^{(t)}$ through time t. When p is large, the major challenge in computing posterior distributions at time t comes from evaluating $\boldsymbol{V}_t^*$ which involves taking inverse of a $p \times p$ matrix. However, the marginal posterior distribution of $\boldsymbol{\beta}$ being in closed form, operating characteristics of the posteriors are available analytically, bypassing the need to follow an iterative sampling scheme to estimate these operating characteristics.

Such closed form expressions for the marginal posterior distributions of parameters are hard to come by when the focus is on Bayesian high dimensional regularization (shrinkage) or variable selection priors. This article considers the Bayesian Lasso and Horseshoe priors as two representative priors from the class of shrinkage priors and the Spike and Lasso prior from the class of variable selection priors. Below we briefly introduce online posterior computation with these priors with large or streaming data and describe computational challenges with large p. The computational challenges are similar in other Bayesian shrinkage or variable selection priors.

2.1 Bayesian Lasso Shrinkage Prior

The Bayesian Lasso shrinkage prior stands as an important example of the *global-local (GL) scale mixtures* (Polson and Scott, 2010) of normal prior distributions. The prior takes the specific form $p(\beta_j|\sigma^2, \lambda) = \frac{\lambda}{2\sigma}\exp(-\lambda|\beta_j|/\sigma)$, $j = 1, .., p$, $\lambda^2 \sim G(r, d)$, with the conditional posterior distribution of $\boldsymbol{\beta}$ given other parameters not available in closed form. However, conditional distributions can be obtained in closed form using a data augmenta-

tion approach. In fact, the hierarchical data augmented model with the Bayesian Lasso prior on $\boldsymbol{\beta}$ with data $\boldsymbol{D}^{(t)} = \{(\boldsymbol{y}_s, \boldsymbol{X}_s) : s = 1, ..., t\}$ upto time t is given by

$$\begin{gathered}\boldsymbol{y}_s|\boldsymbol{X}_s, \boldsymbol{\beta}, \sigma^2 \sim N_n\left(\boldsymbol{X}_s\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right), \;\; s = 1, ..., t \\ \boldsymbol{\beta}|\boldsymbol{\tau}^2, \sigma^2 \sim N_p\left(\boldsymbol{0}, \sigma^2\boldsymbol{M}_\tau\right), \;\; \tau_j^2 \sim Exp\left(\frac{\lambda^2}{2}\right), \;\; \pi\left(\sigma^2\right) \propto \frac{1}{\sigma^2}, \;\; \lambda^2 \sim G\left(r, d\right), \;\; j = 1, ..., p,\end{gathered}$$

where $\tau_1^2, ..., \tau_p^2$ are predictor specific latent variables employed for data augmentation, $\boldsymbol{\tau}^2 = (\tau_1^2, ..., \tau_p^2)'$ and $\boldsymbol{M}_\tau = diag(\boldsymbol{\tau}^2)$. The batch MCMC implemented using the customary Gibbs sampler alternates between the full conditional distributions of (1) $\boldsymbol{\beta}|\sigma^2, \lambda^2, \boldsymbol{\tau}^2, \boldsymbol{D}^{(t)}$; (2) $\sigma^2|\boldsymbol{\beta}, \lambda^2, \boldsymbol{\tau}^2, \boldsymbol{D}^{(t)}$; (3) $\lambda^2|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}^2, \boldsymbol{D}^{(t)}$ and (4) $\tau_j^2|\sigma^2, \lambda^2, \boldsymbol{\beta}, \boldsymbol{D}^{(t)}$, $j = 1, ..., p$, given by

$$\begin{gathered}\boldsymbol{\beta}|\sigma^2, \boldsymbol{\tau}^2, \lambda^2, \boldsymbol{D}^{(t)} \sim N_p\left(\left(\boldsymbol{S}_1^{(t)} + \boldsymbol{M}_\tau^{-1}\right)^{-1}\boldsymbol{S}_2^{(t)}, \sigma^2\left(\boldsymbol{S}_1^{(t)} + \boldsymbol{M}_\tau^{-1}\right)^{-1}\right) \\ \sigma^2|\boldsymbol{\beta}, \boldsymbol{\tau}^2, \lambda^2, \boldsymbol{D}^{(t)} \sim IG\left(\frac{nt + p}{2}, \frac{\left(\boldsymbol{S}_3^{(t)} + \boldsymbol{\beta}'\boldsymbol{S}_1^{(t)}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{S}_2^{(t)}\right) + \boldsymbol{\beta}'\boldsymbol{M}_\tau^{-1}\boldsymbol{\beta}}{2}\right) \\ \frac{1}{\tau_j^2}|\boldsymbol{\beta}, \sigma^2, \lambda^2, \boldsymbol{D}^{(t)} \sim Inv - Gaussian\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right), \; \lambda^2|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}^2, \boldsymbol{D}^{(t)} \sim IG\left(p + r, \frac{\sum_{j=1}^p \tau_j^2}{2} + d\right).\end{gathered} \tag{2}$$

The full conditional posterior distributions at time t depend on the data $\boldsymbol{D}^{(t)}$ only through a few sufficient statistics $\boldsymbol{S}_1^{(t)} = \boldsymbol{S}_1^{(t-1)} + \boldsymbol{X}_t'\boldsymbol{X}_t$, $\boldsymbol{S}_2^{(t)} = \boldsymbol{S}_2^{(t-1)} + \boldsymbol{X}_t'\boldsymbol{y}_t$ and $\boldsymbol{S}_3^{(t)} = \boldsymbol{S}_3^{(t-1)} + \boldsymbol{y}_t'\boldsymbol{y}_t$, which are updated at the onset of a new data shard. At each time $t = 1, ..., T$, the main computational issue lies in the Gibbs sampling step of $\boldsymbol{\beta}$ that requires decomposing a $p \times p$ covariance matrix costing $\sim p^3$ floating point operations (flops) and $\sim p^2$ storage units, and is rendered infeasible.

2.2 Horseshoe Shrinkage Prior

We also consider the popularly used Horseshoe (Carvalho et al., 2010) shrinkage prior on high dimensional predictor coefficients, which is well recognized in the Bayesian shrinkage literature for its ability to artfully shrink unimportant predictor coefficients while applying minimum shrinkage on important coefficients. Several recent articles theoretically prove its

ability to estimate true predictor coefficients a-posteriori in presence of both high and low sparsity (Armagan et al., 2013).

Similar to the Bayesian Lasso, the Horseshoe shrinkage prior also does not admit closed form full posterior of $\boldsymbol{\beta}$. Thus, Gibbs sampling is implemented by invoking a data augmentation approach similar to the Bayesian Lasso. The hierarchical data augmented model with the Horseshoe shrinkage prior is given by

$$\boldsymbol{y}_s|\boldsymbol{X}_s,\boldsymbol{\beta},\sigma^2 \sim N_n\left(\boldsymbol{X}_s\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n\right),\ s=1,..,t,\ \boldsymbol{\beta}|\sigma^2,\tau^2,\boldsymbol{\lambda} \sim N_p\left(\boldsymbol{0},\tau^2\sigma^2\boldsymbol{M}_{\boldsymbol{\lambda}}\right),\ \pi(\sigma^2)\propto\frac{1}{\sigma^2}$$
$$\lambda_j^2 \mid \nu_j \sim \mathcal{IG}\left(\frac{1}{2},\frac{1}{\nu_j}\right),\ \nu_j \sim \mathcal{IG}\left(\frac{1}{2},1\right),\ \tau^2 \mid \xi \sim \mathcal{IG}\left(\frac{1}{2},\frac{1}{\xi}\right),\ \xi \sim \mathcal{IG}\left(\frac{1}{2},1\right),\ j=1,..,p,$$

where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)'$, $\boldsymbol{M}_\lambda = diag\left(\lambda_1^2,\ldots,\lambda_p^2\right)$, $\boldsymbol{\lambda}=\left(\lambda_1^2,\ldots,\lambda_p^2\right)'$ and $\boldsymbol{\nu}=(\nu_1,\ldots,\nu_p)'$. The data augmentation allows the batch MCMC procedure to draw MCMC samples at time t from the following full conditional distributions,

$$\begin{aligned}
&\boldsymbol{\beta}|\sigma^2,\boldsymbol{\tau}^2,\lambda^2,\boldsymbol{D}^{(t)} \sim N_p\left(\left(\boldsymbol{S}_1^{(t)}+\frac{\boldsymbol{M}_\lambda^{-1}}{\tau^2}\right)^{-1}\boldsymbol{S}_2^{(t)},\sigma^2\left(\boldsymbol{S}_1^{(t)}+\frac{\boldsymbol{M}_\lambda^{-1}}{\tau^2}\right)^{-1}\right)\\
&\sigma^2|\boldsymbol{\beta},\boldsymbol{\tau}^2,\lambda^2,\boldsymbol{D}^{(t)} \sim IG\left(\frac{nt+p}{2},\frac{\boldsymbol{S}_3^{(t)}+\boldsymbol{\beta}'\boldsymbol{S}_1^{(t)}\boldsymbol{\beta}-2\boldsymbol{\beta}'\boldsymbol{S}_2^{(t)}}{2}+\frac{\boldsymbol{\beta}'\boldsymbol{M}_\lambda^{-1}\boldsymbol{\beta}}{2\tau^2}\right)\\
&\lambda_j^2|\beta_j,\nu_j,\tau^2,\sigma^2,\boldsymbol{D}^{(t)} \sim IG\left(1,\left[\frac{1}{\nu_j}+\frac{\beta_j^2}{2\tau^2\sigma^2}\right]\right),\ \nu_j|\lambda_j^2,\boldsymbol{D}^{(t)} \sim IG\left(1,\left(1+\frac{1}{\lambda_j^2}\right)\right)\\
&\xi|\boldsymbol{\beta},\sigma^2,\boldsymbol{\tau}^2,\boldsymbol{D}^{(t)} \sim IG\left(1,1+\frac{1}{\tau^2}\right),\tau^2|\boldsymbol{\beta},\boldsymbol{\lambda},\sigma^2,\boldsymbol{D}^{(t)} \sim IG\left(\frac{p+1}{2},\frac{1}{\xi}+\frac{\boldsymbol{\beta}'\boldsymbol{M}_\lambda^{-1}\boldsymbol{\beta}}{2\sigma^2}\right). \qquad (3)
\end{aligned}$$

The conditional distributions are dependent on the data $\boldsymbol{D}^{(t)}$ only through sufficient statistics $\boldsymbol{S}^{(t)}=\{\boldsymbol{S}_1^{(t)},\boldsymbol{S}_2^{(t)},\boldsymbol{S}_3^{(t)}\}$ which are updated using $\boldsymbol{S}_1^{(t)}=\boldsymbol{S}_1^{(t-1)}+\boldsymbol{X}_t'\boldsymbol{X}_t$, $\boldsymbol{S}_2^{(t)}=\boldsymbol{S}_2^{(t-1)}+\boldsymbol{X}_t'\boldsymbol{y}_t$ and $\boldsymbol{S}_3^{(t)}=\boldsymbol{S}_3^{(t-1)}+\boldsymbol{y}_t'\boldsymbol{y}_t$. Similar to the Bayesian Lasso, the Gibbs sampling step of $\boldsymbol{\beta}$ involves decomposing and storing a $p\times p$ matrix per iteration that becomes costly with big p.

2.3 Spike and Lasso Variable Selection Prior

Although shrinkage priors are designed to shrink the posterior distributions of unimportant predictor coefficients close to zero, the shrinkage frameworks do not allow detection

of unimportant predictors. In contrast, the spike and slab discrete mixture of distributions are specifically designed for variable selection in high dimensional regressions (George and McCulloch, 1997). In this section, a variant of the spike and slab mixture prior is introduced as,

$$\beta_j|\sigma^2, \tau_j^2, \gamma_j \sim \gamma_j N\left(0, \sigma^2\tau_j^2\right) + (1-\gamma_j)N\left(0, c^2\right)$$
$$\tau_j^2 \sim Exp(\lambda^2/2),\ \gamma_j \sim Ber(\theta),\ \lambda^2 \sim Ga(r, d),\ \theta \sim Beta(a, b).$$

Integrating over the latent variables τ_j^2 and λ^2, we obtain $\beta_j|\sigma^2, \lambda^2, \gamma_j \sim \gamma_j DE(\lambda/\sigma) + (1-\gamma_j)N(0, c^2)$, for $j = 1, ..., p$, as a mixture of a double exponential and normal densities. We refer to this mixture distribution as the *Spike and Lasso* distribution. Choosing c^2 small, the prior performs simultaneous variable selection and parameter estimation, adaptively thresholding small effects with the concentrated normal spike while minimally shrinking the large effects with the heavy-tailed double exponential (DE) slab distribution. Allowing the prior inclusion probability θ to be random enables us to automatically adjust for multiple comparisons (Scott and Berger, 2010). Spike and slab discrete mixture priors enjoy attractive theoretical properties (Castillo et al., 2015) and a transformed spike and slab prior has recently been added as a penalty to the frequentist penalized optimization literature (Ročková and George, 2016).

With data upto time t, $\boldsymbol{D}^{(t)}$ and sufficient statistics $\boldsymbol{S}_1^{(t)}$, $\boldsymbol{S}_2^{(t)}$ and $\boldsymbol{S}_3^{(t)}$, the prior formulation and data model lead to the following closed form full conditional posteriors facilitating

implementation with Gibbs sampler

$$
\begin{aligned}
&\boldsymbol{\beta}|\sigma^2,\tau^2,\gamma,\boldsymbol{D}^{(t)} \sim N_p\left(\left(\boldsymbol{S}_1^{(t)}+\boldsymbol{M}^{-1}\right)^{-1}\boldsymbol{S}_2^{(t)},\sigma^2\left(\boldsymbol{S}_1^{(t)}+\boldsymbol{M}^{-1}\right)^{-1}\right)\\
&\sigma^2|\boldsymbol{\beta},\boldsymbol{\tau}^2,\lambda^2,\boldsymbol{D}^{(t)} \sim IG\left(\frac{nt+p}{2},\frac{\left(\boldsymbol{S}_3^{(t)}+\boldsymbol{\beta}'\boldsymbol{S}_1^{(t)}\boldsymbol{\beta}-2\boldsymbol{\beta}'\boldsymbol{S}_2^{(t)}\right)+\boldsymbol{\beta}'\boldsymbol{M}^{-1}\boldsymbol{\beta}}{2}\right)\\
&\lambda^2|\boldsymbol{\beta},\sigma^2,\tau^2,\boldsymbol{D}^{(t)} \sim IG\left(p+r,\frac{\sum_{j=1}^p\gamma_j\tau_j^2}{2}+d\right),\ \theta\sim Beta\left(a+\sum_{j=1}^p\gamma_j,b+p-\sum_{j=1}^p\gamma_j\right)\\
&\frac{1}{\tau_j^2}|\gamma_j=1,\boldsymbol{\beta},\sigma^2,\lambda^2,\boldsymbol{D}^{(t)} \sim Inv-Gaussian\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}},\lambda^2\right),\ \tau_j^2|\gamma_j=0,\boldsymbol{\beta},\sigma^2,\lambda^2,\boldsymbol{D}^{(t)}\sim Exp(\lambda^2/2)\\
&\gamma_j|\boldsymbol{\beta},\sigma^2,\tau^2,\theta,\boldsymbol{D}^{(t)} \sim Ber(\eta_j),\ \eta_j=\frac{\theta\left(\sigma^2\tau_j^2\right)^{-\frac{1}{2}}\exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right)}{\theta\left(\sigma^2\tau_j^2\right)^{-\frac{1}{2}}\exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right)+(1-\theta)\left(c^2\right)^{-\frac{1}{2}}\exp\left(-\frac{\beta_j^2}{2c^2}\right)}.
\end{aligned}
\tag{4}
$$

where $\boldsymbol{M}=diag(w_1,\ldots,w_p)$ with $w_j=\tau_j^2$ if $\gamma_j=1$; $=c^2$ otherwise. The computational issue arises from the Gibbs sampling step of $\boldsymbol{\beta}$ that incurs a complexity of $O(p^3)$, as well as due to updating γ_j's, $j=1,...,p$ resulting in high autocorrelation. Updating subsets of $\boldsymbol{\beta}$ parameters in smaller blocks may be an option. However, shrinkage or variable selection priors generally do not allow closed form marginal distributions for such blocks of regression parameters. Again, the sequential nature of Gibbs sampling prohibits updating blocks of parameters in $\boldsymbol{\beta}$ in parallel. The dynamic feature partitioning strategy developed in the next section will provide a solution to this computational challenge by parallelizing the approximate Bayesian computation of blocks of parameters into different processors.

3 Dynamic Feature Partition in High-Dimensional Regression

The dynamic feature partitioning (DFP) is a general online algorithm for streaming data (or massive data fed to a model in small batches or shards during model computation) that partitions the large parameter space and facilitates rapid Bayesian updating of different partitions of parameters in parallel. While the algorithm is applied to mitigate the

aforementioned computational issues in the Bayesian high dimensional linear regression, the algorithm per se is more general in nature and perhaps useful in other contexts.

3.1 Relevant Notations and Details of DFP

Let $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_q\}$ represent the parameter space with q parameters, which is bigger than p (the no. of predictors), since the parameter space includes the error variance σ^2 as well as latent variables from the data augmentation procedure described in Section 2. We further assume

(1) q is fixed over time, i.e., the parameter space does not change with the arrival of new data shards.

(2) At each time point, the posterior distribution of the parameters $\boldsymbol{\Theta}$ depends on the data only through lower dimensional functions of $\boldsymbol{D}^{(t)}$ which are referred to as sufficient statistics. More formally, $\boldsymbol{S}^{(t)}$ is a vector of sufficient statistics for $\boldsymbol{\Theta}$ if $\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}$ has the same distribution as $\boldsymbol{\Theta}|\boldsymbol{S}^{(t)}$. Denoting $f(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)})$ as the full posterior distribution of $\boldsymbol{\Theta}$, this assumption implies that $f(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}) = f(\boldsymbol{\Theta}|\boldsymbol{S}^{(t)})$.

Referring to Section 2, both (1) and (2) are valid for linear regression models with shrinkage prior distributions or discrete mixture variable selection priors on coefficients.

At time t, consider a partition of the parameter indices given by $\mathcal{G}^{(t)} = \{G_1^t, \ldots, G_{k_t}^t\}$, such that $G_l^t \cap G_{l'}^t = \emptyset, l \neq l'$ and $\bigcup_{l=1}^{k_t} G_l^t = \{1, \ldots, q\}$. Also let $\boldsymbol{\Theta}_{G_l^t} = \{\theta_i \mid i \in G_l^t\}$ and $\boldsymbol{\Theta}_{-G_l^t} = \boldsymbol{\Theta}_{\{1,\ldots,q\}\setminus G_l^t} = \{\theta_i \mid i \in \{1, \ldots, q\} \setminus G_l^t\} = \{\theta_i \mid i \notin G_l^t\}$ be parameters contained and not contained in the lth partition, respectively. We consider both the number of partitions k_t and the constitution of each partition to be adaptive and dynamically changing over time. The prior specifications and conditional independence assumptions often suggest natural parameter partitioning schemes. We provide an outline of the dynamic parameter partitioning schemes employed in this article in the context of high dimensional regressions with shrinkage and Spike and Lasso priors towards the end of this section.

Consider also a sequence of point estimates $\widehat{\boldsymbol{\Theta}}^{(t)}$ constructed dynamically over time for the parameter $\boldsymbol{\Theta}$. Given a partition of the parameter space at time t, the DFP approximation to the posterior full conditional distribution $f\left(\boldsymbol{\Theta}_{G_l^t}|\boldsymbol{\Theta}_{-G_l^t}, \boldsymbol{S}^{(t)}\right)$ of $\boldsymbol{\Theta}_{G_l^t}$ $(l = 1, ..., k_t)$, referred

to as the *DFP pseudo conditional posterior*, is given by $f\left(\boldsymbol{\Theta}_{G_l^t}|\widehat{\boldsymbol{\Theta}}^{(t-1)}_{-G_l^t}, \boldsymbol{S}^{(t)}\right)$, with $\boldsymbol{\Theta}_{-G_l^t}$ replaced by its point estimate $\widehat{\boldsymbol{\Theta}}^{(t-1)}_{-G_l^t}$ at time $(t-1)$. Since the conditioning set remains fixed throughout time t, conditional distributions $\boldsymbol{\Theta}_{G_l^t}$'s for $l = 1, ..., k_t$ are not dependent on each other at time t. This eliminates the need to sequentially update parameter blocks $\boldsymbol{\Theta}_{G_l^t}$'s, and samples can rather be drawn rapidly from k_t DFP pseudo conditional posteriors in parallel. All these concepts and notations will be used to describe the DFP algorithm below.

DFP Algorithm for online approximate MCMC inference:

The DFP algorithm provides an online approximate MCMC sampling based on dynamically adaptive parameter partitions and their point estimates constructed sequentially over time. The algorithm begins by initializing the point estimate of $\boldsymbol{\Theta}$ (call it $\widehat{\boldsymbol{\Theta}}^{(0)}$) at some default value and initializing sufficient statistics $\boldsymbol{S}^{(0)}$ at $\mathbf{0}$. When new data shard $\boldsymbol{D}_t$ arrives at time t $(t = 1, ..., T)$, sufficient statistics $\boldsymbol{S}^{(t)}$ are updated as a function of $\boldsymbol{S}^{(t-1)}$ and $\boldsymbol{D}_t$, denoted as $\boldsymbol{S}^{(t)} = g(\boldsymbol{S}^{(t-1)}, \boldsymbol{D}_t)$. In the examples of Section 2, $g(\cdot)$ is implicitly defined through the three equations, $\boldsymbol{S}_1^{(t)} = \boldsymbol{S}_1^{(t-1)} + \boldsymbol{X}_t'\boldsymbol{X}_t$, $\boldsymbol{S}_2^{(t)} = \boldsymbol{S}_2^{(t-1)} + \boldsymbol{X}_t'\boldsymbol{y}_t$ and $\boldsymbol{S}_3^{(t)} = \boldsymbol{S}_3^{(t-1)} + \boldsymbol{y}_t'\boldsymbol{y}_t$. The dynamic partitioning scheme (described later) then updates partitions of the set of parameters and creates new partitions $\mathcal{G}^{(t)}$ at time t. The DFP algorithm then proceeds by sampling from the DFP pseudo conditional posteriors at time t in parallel. If the DFP pseudo conditional posteriors are in closed form, one may consider block updating of $\boldsymbol{\Theta}_{G_l^t}$ from $f\left(\boldsymbol{\Theta}_{G_l^t}|\widehat{\boldsymbol{\Theta}}^{(t-1)}_{-G_l^t}, \boldsymbol{S}^{(t)}\right)$. Otherwise, the sampling in each partition proceeds by employing a Gibbs sampler with smaller blocks of parameters in the lth partition. More specifically, $\theta_j \in \boldsymbol{\Theta}_{G_l^t}$ is updated by drawing S (a moderately large number, taken to be 500 in Section 4) approximate MCMC samples $\tilde{\theta}_j^{(1,t)}, ..., \tilde{\theta}_j^{(S,t)}$ from $f\left(\theta_j|\boldsymbol{\Theta}_{G_l^t\setminus\{j\}}, \widehat{\boldsymbol{\Theta}}^{(t-1)}_{-G_l^t}, \boldsymbol{S}^{(t)}\right)$. Often this distribution depends on a lower dimensional function of $\boldsymbol{\Theta}_{G_l^t\setminus\{j\}}$, $\widehat{\boldsymbol{\Theta}}^{(t-1)}_{-G_l^t}$ and $\boldsymbol{S}^{(t)}$, given by $\boldsymbol{J}_{l,j}^{(t)} = h_j(\boldsymbol{\Theta}_{G_l^t\setminus\{j\}}, \widehat{\boldsymbol{\Theta}}^{(t-1)}_{-G_l^t}, \boldsymbol{S}^{(t)})$, i.e., $f\left(\theta_j|\boldsymbol{\Theta}_{G_l^t\setminus\{j\}}, \widehat{\boldsymbol{\Theta}}^{(t-1)}_{-G_l^t}, \boldsymbol{S}^{(t)}\right) = f\left(\theta_j|\boldsymbol{J}_{l,j}^{(t)}\right)$. Specific examples of $\boldsymbol{J}_{l,j}^{(t)}$ are provided in Sections 4.1, 4.2 and 4.3. Once S approximate MCMC samples are drawn from DFP pseudo conditional posteriors fairly rapidly, we use these samples to construct the point estimates of parameters at time t, given by $\widehat{\boldsymbol{\Theta}}^{(t)}$. In our exposition, we use mean of the S samples $\tilde{\theta}_j^{(1,t)}, ..., \tilde{\theta}_j^{(S,t)}$ to construct $\hat{\theta}_j^{(t)}$. The theoretical results in the supplementary material prove desirable performance of the proposed algorithm when the sequence of estimators $\widehat{\boldsymbol{\Theta}}^{(t)}$ is consistent in estimating the true parameters as $t \to \infty$.

Efficient updating of DFP pseudo conditional posteriors using the sufficient statistics and point estimates of parameters from the previous time point lead to scalable inference.

Partitioning schemes:

As discussed before, an efficient partitioning of parameter indices $\mathcal{G}^{(t)}$ at the tth time is achieved by heavily exploiting the nature of the model and prior distributions. We believe that a general partitioning scheme that is applicable to any model and/or any prior distribution is unappealing since it will not be able to fully exploit the specific features of the model and prior distributions. Since the main focus of this article is on Bayesian shrinkage and variable selection priors in high dimensional linear regression models, broadly two different partitioning schemes are proposed, one for the model (1) with shrinkage priors and the other for spike and slab priors.

(A) Partitioning algorithm for shrinkage priors: Referring to the discussion in Sections 2.1 and 2.2, the computational bottleneck mainly arises due to sampling from the posterior full conditional of $\boldsymbol{\beta}$. Therefore, in the course of developing a partitioning strategy for the parameters in (1) with shrinkage priors, the main focus rests on how to partition $\boldsymbol{\beta}$ into blocks of sub-vectors with a minimal loss of information due to separately updating these blocks residing in different partitions from their DFP full conditionals. To this end, we set the maximum size of each block of $\boldsymbol{\beta}$ residing in different partitions to be less than or equal to M at every time to keep a control on the computational complexity. M is user defined and its choice depends on the available computational resources. In high dimensional linear regression with Bayesian shrinkage priors, empirical investigations show $M = 100$ to be sufficient. Thereafter we envision the problem of partitioning $\boldsymbol{\beta}$ at time t as a graph partitioning problem. To elaborate, at time t, for $j, j' \in \{1, ..., p\}$, let the sample correlation between S iterates of β_j and $\beta_{j'}$ from time $(t-1)$ following the DFP algorithm, given by $\{\tilde{\beta}_j^{(s,t-1)}\}_{s=1}^S$ and $\{\tilde{\beta}_{j'}^{(s,t-1)}\}_{s=1}^S$, be denoted by $r_{j,j'}$. A graph is constructed with nodes as the predictor indices $\{1, ..., p\}$ and an edge between two nodes j, j' if $r_{j,j'} > c$ where $c \in (0, 1)$. Our proposed scheme constructs different graphs in this manner corresponding to different choices of the cut-off $c \in$ `seq(0.01,0.99,by=0.01)`. Thereafter we find connected components of all these constructed graphs and look for the smallest value of c (say c^*) for which the size of all connected components are less than M. Such an implementation

is readily achieved by the functionalities in the `igraph` package in `R`. The b_t connected components $\{\mathcal{P}_1^{(t)}, ..., \mathcal{P}_{b_t}^{(t)}\}$ corresponding to the cut-off value c^* at time t are recognized as partitions of the indices $\{1, ..., p\}$ and β_j's corresponding to different connected components go to different partitions at time t. Computational cost with Gibbs updating of the rest of the parameters is less substantial and thus there is more room to partition the other parameters. Since the data augmentation approaches in Sections 2.1 and 2.2 introduce latent vectors ($\boldsymbol{\tau}^2$ in Section 2.1, $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ in Section 2.2) of the same size as $\boldsymbol{\beta}$, we either keep all elements of a latent vector together in one partition or divide a latent vector into blocks with indices $\{\mathcal{P}_1^{(t)}, ..., \mathcal{P}_{b_t}^{(t)}\}$ and send the latent vector with indices $\mathcal{P}_k^{(t)}$ to the same partition where $\boldsymbol{\beta}_{\mathcal{P}_k^{(t)}}$ lies. Variance σ^2 and other hierarchical parameters are kept together in a separate partition. Since a partition involves blocks of $\boldsymbol{\beta}$ with size at most M, sampling them together from their DFP full conditionals incurs complexity at most of $O(M^3)$.

(B) Partitioning algorithm for Spike and Lasso priors: Since the Spike and Lasso example in Section 2.3 involves coefficients belonging to one of the two mixture components at every iteration of the posterior sampling, the parameter partitioning scheme adopted for shrinkage priors appears to be less efficient here. Instead, we propose a dynamic partitioning scheme of the parameter space by tacitly exploiting the natural partitioning of the $\boldsymbol{\beta}$ parameters and associated latent vector $\boldsymbol{\tau}$ into important and unimportant components. Define $\boldsymbol{\Theta}_{1t} = \{(\beta_j, \tau_j^2) : \widehat{\gamma}_j^{(t-1)} = 1\}$ and $\boldsymbol{\Theta}_{2t} = \{(\beta_j, \tau_j^2) : \widehat{\gamma}_j^{(t-1)} = 0\}$, where $\widehat{\gamma}_j^{(t-1)} \in \{0, 1\}$ corresponds to the point estimate of γ_j at time $(t-1)$. Thereafter our partitioning scheme suggests keeping the entire $\boldsymbol{\Theta}_{1t}$ in one partition and dividing $\boldsymbol{\Theta}_{2t}$ into partitions, with each partition of $\boldsymbol{\Theta}_{2t}$ containing (β_j, τ_j^2) for a single j. Additionally, all γ_j's are kept in the same partition and $\lambda^2, \sigma^2, \theta$ in another partition. Since spike and slab priors are typically employed to recover $\boldsymbol{\beta}$ parameters which are sparse in nature in the truth, $\boldsymbol{\Theta}_{1t}$ is expected to be of small to moderate size with cardinality much smaller than p as time progresses. Thus, updating $(\beta_j : \beta_j \in \boldsymbol{\Theta}_{1t})'$ together requires computation complexity of order $|\boldsymbol{\Theta}_{1t}|^3 << p^3$. On the other hand, β_j's for $j \in \boldsymbol{\Theta}_{2t}$ are updated individually without incurring any notable computational burden. A similar strategy is followed when the double exponential slab distribution in the Spike and Lasso prior is replaced by any other distribution.

3.2 Comparison of DFP with Other Approximate Bayesian Frameworks

Algorithm 1 presents a sketch of the Dynamic Feature Partition (DFP) in high dimensions. Although DFP is an approximate Bayesian algorithm, it has significant distinctions from the literature on frameworks involving approximate Bayesian inference as discussed below.

When the full posterior distribution is computationally prohibitive, methods like variational Bayes offer a computationally efficient alternative by optimizing the parameters in a class of analytic approximations to the posterior. Variational Bayes algorithms are extended to online variational Bayes algorithms (Sato, 2001; Hoffman et al., 2010; Campbell et al., 2015) for efficient online Bayesian learning for streaming or large data. Although the DFP framework proposes approximating the full posterior distribution, the approximation technique is fundamentally different from variational approximations. While variational Bayes approximates the full posterior distribution by a distribution with block independent marginals, the DFP framework invokes approximations by blocking independent posterior conditional distributions of parameters. More importantly, variational approximations often pre-decide parameter blocks which are to be considered independent in the posterior inference, while DFP dynamically adapts to ensure efficient partitioning of parameters. As a result, variational approximation may underestimate uncertainty from the variationally approximated posterior distribution of $\boldsymbol{\beta}$, while DFP is demonstrated to have close to nominal coverage in almost all high dimensional simulation examples.

In the general Bayesian literature of streaming data, Sequential Monte Carlo (SMC) (Chopin, 2002; Arulampalam et al., 2002; Lopes and Tsay, 2011; Doucet et al., 2001; Zhou and Jasra, 2015) is one of the most popular online methods that relies on resampling particles sequentially as data shards arrive over time. A naive implementation of SMC might be less efficient and less accurate involving large n and p due to the need to employ very large numbers of particles to obtain adequate approximations and prevent particle degeneracy. The latter is addressed through rejuvenation steps using all the data (or sufficient statistics), which may become expensive in an online setting (Snyder et al., 2008). There

are approaches in recent years to overcome the dimensionality issue in the SMC algorithm mainly in the context of fitting state-space models. To this end, carefully constructed SMC algorithms (Chopin et al., 2004; Beskos et al., 2014; Schweizer, 2012; Carvalho et al., 2010) show promise in terms of scaling in a polynomial complexity with the number of parameters, though the complexity as a function of the size of the dataset is either growing with time (e.g., for Chopin et al. (2004)) or is not apparent from the context. Rebeschini et al. (2015) develop a blocking strategy for high dimensional particle learning (PL) where the error of approximation is free of the dimension of the parameter space. Unfortunately, the numerical examples for high dimensions provided by Rebeschini et al. (2015) do not demonstrate satisfactory performance with large state-space models. Furthermore, the results rely on the decay of correlations for state-space varying parameters in the fitted model, which is suitable in the context of state-space models, but less satisfactory for our problem of interest. Wigren et al. (2018) propose another approach for high dimensional particle learning in state-space models, though the numerical illustration of the approach may struggle to comfortably scale beyond a few dozen dimensional state-space models. While most of these developments have taken place in the high dimensional problem of particle filtering in state-space models, we are concerned with estimation of high dimensional parameters, which has been given far less attention. To this end, Lindsten et al. (2017) propose a new SMC algorithm based on parameter partitioning, though difficulties may arise when joining the partitions, which requires a careful resampling. In the same vein, Gunawan et al. (2018) propose an approach that employs a sub-sampling technique to combat the problem of large data in the realm of high dimensional problems. Arguably, there is a general lack of extensive empirical investigations of SMC or PL algorithms proposed for high dimensional problems, and most of them do not come with any open source code for implementation. Perhaps a static parameter estimation presents a bigger challenge than state filtering in high dimensions. One plausible reason can be the fact that new data points add more information for the state in a state-space setting. On a separate note, Hamiltonian Monte Carlo (HMC) methods with stochastic gradient descent can also leverage the online nature of the data (Betancourt, 2017) while exploring the distribution efficiently. However, HMC may not be suitable for computing high dimensional regression with a discrete mixture of prior distributions involving a large number of binary

variables. Additionally, the performance of HMC with a high dimensional parameter space is yet to be fully reckoned with.

In a recent article, Wang et al. (2016) introduce a two stage predictor partitioned high dimensional linear regression for fast computation. In the first stage, predictors are de-correlated, which is followed by partitioning the predictor space into subsets. In the second stage, Lasso is fit on each subset of predictors. This approach yields only point estimates of $\boldsymbol{\beta}$ with no straightforward Bayesian extension. Moreover, it somewhat loses its appeal when dealing with streaming data, as the first stage of de-correlation would have to be done repeatedly at the onset of a new data shard.

4 Illustrations of DFP with Shrinkage and Discrete Mixture Priors in High Dimensional Regressions

This section illustrates parametric and predictive performances of the online DFP algorithm for (i) Bayesian Lasso, (ii) Horseshoe and (iii) Spike and Lasso discrete mixture priors. For the simulation examples in (i)-(iii), shards of size $n = 1000$ observations arrive sequentially over $T = 500$ time horizons. Data shard $\boldsymbol{D}_t$ at time t consists of an $n \times 1$ response vector $\boldsymbol{y}_t$ and an $n \times p$ predictor matrix $\boldsymbol{X}_t = (\boldsymbol{x}_{1t}, ..., \boldsymbol{x}_{nt})'$, $t = 1, ..., T$. At each time, $S = 500$ approximate MCMC samples of $\boldsymbol{\Theta}_{G_1^t}, ..., \boldsymbol{\Theta}_{G_{k_t}^t}$ are drawn from their respective DFP pseudo conditional posteriors to approximate the full posterior distribution $f(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)})$.

The $p \times 1$ predictor vector $\boldsymbol{x}_{jt}$ $(j = 1, ..., n)$ at time t is generated as $\boldsymbol{x}_{jt} \sim N(\boldsymbol{0}, \boldsymbol{H})$, where $\boldsymbol{H} = \text{Block-diag}(\boldsymbol{H}_1, ..., \boldsymbol{H}_{100})$, with each $\boldsymbol{H}_l$ being a 50×50 Toeplitz structured matrix having the (m, m')th element as $\rho^{|m-m'|}$, $\rho \in (0, 1)$. This is to mimic the scenario where there are blocks of predictors such that predictors within a block are correlated and predictors across blocks are uncorrelated. All simulation examples consider high correlations among predictors in a block with $\rho = 0.9$. This presumably induces strong associations among parameters, which is often challenging for any high dimensional regression framework to estimate. The inferential challenge appears to be more critical for the DFP framework as it relies on parameter partitioning, which might naturally weaken correlations a-posteriori among parameters. To simulate the true predictor coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, the following

Algorithm 1: Dynamic Feature Partition

1 1

Input: (1) Data shard $\boldsymbol{D}_t$ at time t; (2) Parameter partition $\mathcal{G}^{(t-1)}$; (3) Sufficient Statistics $\boldsymbol{S}^{(t-1)}$ (4) Approximate posterior draws $\tilde{\boldsymbol{\Theta}}^{(1,t-1)}, \ldots, \tilde{\boldsymbol{\Theta}}^{(S,t-1)}$ at time $(t-1)$; (5) Parameter Estimates $\widehat{\boldsymbol{\Theta}}^{(t-1)}$

Output: (1) Approximate posterior draws $\tilde{\boldsymbol{\Theta}}^{(1,t)}, \ldots, \tilde{\boldsymbol{\Theta}}^{(S,t)}$ at time t; (2) Sufficient Statistics $\boldsymbol{S}^{(t)}$; (3) Parameter Estimates $\widehat{\boldsymbol{\Theta}}^{(t)}$

2 **DFP**$(\boldsymbol{D}_t, \mathcal{G}^{(t)}, \boldsymbol{S}^{(t-1)}, \widehat{\boldsymbol{\Theta}}^{(t-1)})$

3 **begin**

/* Step 1: Update the Partitioning of the set of parameters at time t: the partitioning schemes should ideally exploit the nature of the model and prior distributions. We propose partitioning schemes specific to the high dimensional linear regression models with shrinkage priors and spike and slab priors in Section 3, page 12 and 13. */

4 $\mathcal{G}^{(t)} = \textbf{PartitionUpdate}\left(\tilde{\boldsymbol{\Theta}}^{(1,t-1)}, \ldots, \tilde{\boldsymbol{\Theta}}^{(S,t-1)}\right)$

/* step 2: Update Sufficient Statistics */

5 Update $\boldsymbol{S}^{(t)} = g(\boldsymbol{D}_t, \boldsymbol{S}^{(t-1)})$

6 **for** $G_l^t \in \mathcal{G}^{(t)}$ **do**

7 **for** $\theta_j \in \Theta_{G_l^t}$ **do**

8 set $\boldsymbol{J}_{l,j}^{(t)} \leftarrow h_j\left(\boldsymbol{\Theta}_{G_l^t \setminus \{j\}}, \boldsymbol{S}^{(t-1)}, \widehat{\boldsymbol{\Theta}}_{-G_l^t}\right)$

9 **end**

10 **end**

/* step 3: Approximate Sampling for Parameter Blocks in Parallel */

11 **for** $G_l^t \in \mathcal{G}^{(t)}$ **do**

12 **for** $\theta_j \in \boldsymbol{\Theta}_{G_l^t}$ **do**

13 **for** *s=1:S* **do**

14 sample $\tilde{\theta}_j^{(s,t)} \sim f\left(\theta_j | \boldsymbol{J}_{l,j}^{(t)}\right)$

15 **end**

16 **end**

17 **end**

/* step 4: Update Estimates */

18 **for** $G_l^t \in \mathcal{G}^{(t)}$ **do**

19 **for** $\theta_j \in \boldsymbol{\Theta}_{G_l^t}$ **do**

/* Compute relevant point estimates for the parameters from approximate MCMC samples. We consider the mean of the samples as the point estimate for each parameter */

20 set $\hat{\theta}_j^{(t)} \leftarrow \text{stat}\left(\tilde{\theta}_j^{(1,t)}, \ldots, \tilde{\theta}_j^{(S,t)}\right)$

21 **end**

22 **end**

23 **return** $\{\tilde{\boldsymbol{\Theta}}^{(1,t)}, \ldots, \tilde{\boldsymbol{\Theta}}^{(S,t)}\}, \boldsymbol{S}^{(t)}, \widehat{\boldsymbol{\Theta}}^{(t)}$ 17

24 **end**

scenarios are considered:

Simulation 1: 50 randomly selected β_j's are drawn i.i.d. from N(3,1), 50 randomly selected β_j's are drawn i.i.d. from N(1,1), rest are all set to 0.

Simulation 2: 50 randomly selected β_j's are drawn i.i.d. from N(3,1), rest are all set to 0.

Simulation 3: All β_j's are drawn i.i.d. from $U(-1, 1)$.

Simulation 1 focuses on a sparse case with varying magnitudes of nonzero coefficients. We will refer to it as the *low and high sparse case*. Simulation 2 corresponds to a *sparse case* with similar magnitudes of nonzero coefficients, while Simulation 3 corresponds to a *dense case* which is motivated by practical applications where each of the covariates has a small effect on the outcome. The responses $\boldsymbol{y}_t$ for $t = 1, ..., T$ are generated from $\boldsymbol{X}_t$ and the true predictor coefficients using (1), with σ^2 chosen so as to keep a signal to noise ratio of 1 for the generated data.

The performance of DFP is compared with a set of competitors suitable for high dimensional linear regressions models. We specifically compare with (a) batch MCMC that draws S MCMC samples from the full conditional distributions at every time point with the full data $\boldsymbol{D}^{(t)}$ through time t at disposal; and (b) Conditional Density Filtering (C-DF) (Guhaniyogi et al., 2018). Batch MCMC offers the "gold standard" for ordinary Gibbs sampling that uses the full data $\boldsymbol{D}^{(t)}$ at time t. At time t, batch MCMC initializes the MCMC chain at the last iterate in time $(t-1)$. In examples (i)-(iii), the conditional posterior distributions depend on the data through lower dimensional sufficient statistics, and hence batch MCMC only stores and propagates the sufficient statistics to update the conditional distributions in successive time points. Conditional density filtering is proposed in the same vein as DFP with an important difference. While DFP proposes dynamic partitioning of the parameter space, C-DF works with parameter partitions fixed over time. We find that the naive implementation of C-DF demonstrates considerably inferior performance than DFP. To make C-DF more competitive, we employ a version of C-DF that draws samples from parameter partitions sequentially rather than in parallel, to be able to use samples from one partition to construct more accurate point estimates for the other partitions at every time. Such an implementation of C-DF considerably improves its performance, though at the expense of added computational burden. Overall, comparison with this improved version of C-DF

will demonstrate the advantages of dynamic partitioning over fixed partitioning as a tool to provide a better approximation to the full posterior distribution of parameters. Online variational inference provides an alternate strategy to draw approximate inference in presence of big data and a large number of parameters. However, in absence of any open source code for online variational inference in high dimensional linear regression, we refrain from employing it as a competitor. Finally, we compare our approach with a variant of the Sequential Monte Carlo (SMC) approach. As discussed in Section 3.2, most of the developments in SMC and PL algorithms have taken place in the high-dimensional state-space models and they do not assume seamless extensions to very high dimensional static parametric models. There are only a handful of approaches using SMC and its variants in static parametric models, mostly for moderately large dimensional problems. However, to the best of our knowledge, SMC or any of its variants have not been empirically investigated in static parametric models with dimensions as high as we consider ($p = 5000$). Therefore, we adapt the recent sub-sampled SMC approach outlined in Gunawan et al. (2018) to our setting. Note that the approach in Gunawan et al. (2018) is designed for the scenario when the entire dataset is available to the user. To adapt it to the streaming data context, we employ a data annealing approach instead of the temperature annealing approach used by the authors. Our data annealing approach performs data sub-sampling from the entire data $\boldsymbol{D}^{(t)}$ when a new batch arrives at time t and uses the sub-sampling density approximation as well as the Hamiltonian Monte-Carlo technique for efficient drawing of high dimensional Monte Carlo samples. This approach uses the entire data set (upto time t) $\boldsymbol{D}^{(t)}$ in drawing SMC samples at time t, and strictly speaking is not an online Bayesian competitor. Nevertheless, it can demonstrate the state-of-the-art performance from SMC which will be helpful in assessing the performance of DFP. We refer to this approach as sub-sampled SMC (SSMC).

Plots of kernel density estimates for marginal approximate DFP posterior densities on representative model parameters are shown at various time points with the true value of the respective parameters overlaid to assess the posterior inference from DFP. Additionally, to measure the predictive performance of competitors, we report: (a1) mean squared prediction error (MSPE); (a2) Interval score (Gneiting and Raftery, 2007) of the 95% predictive interval; (a3) coverage of the 95% predictive interval and (a4) average run time for each batch or shard.

Note that (a1) demonstrates the performance in terms of point prediction, while (a2) and (a3) show how well calibrated the predictions turn out to be. Finally, (a4) helps readers gauge the computation time vis-a-vis accuracy of the competitors. At time $(t-1)$, evaluations of predictive performance metrics (a1)-(a3) are based on the data shard observed at time t. All results are based on averages over 10 independent replications. All computation times are based on an R implementation in a cluster computing environment with three interactive analysis servers, 32 cores each with the Dell PE R820: 4x Intel Xeon Sandy Bridge E5-4640 processor, 16GB RAM and 1TB SATA hard drive.

4.1 DFP with Bayesian Lasso

We consider the first application of DFP with the popular Bayesian Lasso (Park and Casella, 2008) shrinkage prior on high dimensional predictor coefficients. Details of the Bayesian Lasso prior and challenges regarding posterior computation with the Bayesian Lasso prior has already been presented in Section 2.1.

The DFP algorithm applied to this setting proposes dynamic partitioning of the parameter space over $k_t = b_t + 1$ subsets at time t. Let the partition of the parameter space at time t be defined by

$$\boldsymbol{\Theta}_{G_l^{(t)}} = \left\{\beta_{i^{(t)}_{m_1+\cdots+m_{l-1}+1}}, \tau^2_{i^{(t)}_{m_1+\cdots+m_{l-1}+1}}, ..., \beta_{i^{(t)}_{m_1+\cdots+m_l}}, \tau^2_{i^{(t)}_{m_1+\cdots+m_l}}\right\},\ l = 1, .., b_t, \quad \boldsymbol{\Theta}_{G^{(t)}_{b_t+1}} = \left\{\sigma^2, \lambda^2\right\},$$

where the lth partition, $l = 1, .., b_t$ consists of $2m_l$ parameters (m_l is also a function of t) and $i^{(t)}_{m_1+\cdots+m_{l-1}+1}, ..., i^{(t)}_{m_1+\cdots+m_l} \in \{1, ..., p\}$ correspond to the indices of predictor coefficients and latent variables belonging to the lth partition at time t. Let at time t, $\boldsymbol{\beta}_l = \left(\beta_{i^{(t)}_{m_1+\cdots+m_{l-1}+1}}, ..., \beta_{i^{(t)}_{m_1+\cdots+m_l}}\right)'$, $\boldsymbol{\tau}^2_l = \left(\tau^2_{i^{(t)}_{m_1+\cdots+m_{l-1}+1}}, ..., \tau^2_{i^{(t)}_{m_1+\cdots+m_l}}\right)'$, $\boldsymbol{M}_{\tau,l} = diag(\boldsymbol{\tau}^2_l)$ and $\boldsymbol{\beta}_{-l}$ be the vector of all β_j's except those included in $\boldsymbol{\beta}_l$. $\widehat{\boldsymbol{\beta}}^{(t-1)}_l$, $\widehat{\boldsymbol{\beta}}^{(t-1)}_{-l}$, $\widehat{\boldsymbol{\tau}}^{2(t-1)}_l$ are the point estimates of $\boldsymbol{\beta}_l, \boldsymbol{\beta}_{-l}, \boldsymbol{\tau}^2_l$ respectively at time $(t-1)$. $\boldsymbol{S}^{(t)}_{1,l}$ and $\boldsymbol{S}^{(t)}_{2,l}$ are analogously defined. Also assume $\boldsymbol{S}^{(t)}_{1,l,-l} = \boldsymbol{S}^{(t-1)}_{1,l,-l} + \boldsymbol{X}'_{t,l}\boldsymbol{X}_{t,-l}$, where $\boldsymbol{X}_{t,l}$ and $\boldsymbol{X}_{t,-l}$ are the sub-matrices of $\boldsymbol{X}_t$ corresponding to $\boldsymbol{\beta}_l$ and $\boldsymbol{\beta}_{-l}$, respectively. Following Algorithm 1, sampling proceeds using DFP as follows:

1. *Initialize:* Initialize variables $\boldsymbol{\beta}$, $\boldsymbol{\tau}^2$, σ^2 and λ. Set $\widehat{\boldsymbol{\beta}}^{(0)}, \widehat{\sigma}^{2(0)}, \widehat{\boldsymbol{\tau}}^{2(0)}, \widehat{\lambda}^{2(0)}$ at their initial

values.

2. *Observe data and partition parameter space at time t:* Observe data $\boldsymbol{D}_t = \{\boldsymbol{y}_t, \boldsymbol{X}_t\}$ at time t. Update the partitions of the parameters based on the iterates of the parameters at time $(t-1)$. The parameter partitioning algorithm at time t for the shrinkage priors is given in Section 3.

3. *Update sufficient statistics:* Update sufficient statistics $\boldsymbol{S}_1^{(t)}, \boldsymbol{S}_2^{(t)}, \boldsymbol{S}_3^{(t)}$ based on $\boldsymbol{S}_1^{(t-1)}, \boldsymbol{S}_2^{(t-1)}, \boldsymbol{S}_3^{(t-1)}$ and $\boldsymbol{D}_t$ with the equations given in Section 2.1.

4. *Drawing approximate posterior samples:* Draw S samples from the DFP full conditional posterior distributions of $\boldsymbol{\beta}_l$ and $\boldsymbol{\tau}_l^2$ given by

$$\frac{1}{\tau_j^2}|\cdot \sim Inv-Gaussian\left(\sqrt{\frac{\widehat{\lambda}^{2(t-1)}\widehat{\sigma}^{2(t-1)}}{\beta_j^2}}, \widehat{\lambda}^{2(t-1)}\right) \forall \tau_j^2 \in \boldsymbol{\tau}_l^2, \quad \boldsymbol{\beta}_l \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}_l^{(t)}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_l^{(t)}}\right)$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_l^{(t)}} = \left(\boldsymbol{S}_{1,l}^{(t)} + \boldsymbol{M}_{\tau,l}^{-1}\right)^{-1}\left(\boldsymbol{S}_{2,l}^{(t)} - \boldsymbol{S}_{1,l,-l}^{(t)}\widehat{\boldsymbol{\beta}_{-l}}^{(t-1)}\right), \; \boldsymbol{\Sigma}_{\boldsymbol{\beta}_l^{(t)}} = \widehat{\sigma}^{2(t-1)}\left(\boldsymbol{S}_{1,l}^{(t)} + \boldsymbol{M}_{\tau,l}^{-1}\right)^{-1}.$$

   The conditional distributions of the parameters in the lth server depends on the lower dimensional functions of sufficient statistics, point estimates from time $(t-1)$ and the other parameters from the same partition. This is conceptualized in the notation $\boldsymbol{J}_{l,j}^{(t)}$ in Section 3. Sampling from the DFP full conditionals of $\{\boldsymbol{\beta}_l, \boldsymbol{\tau}_l^2\}$ $(l = 1, ..., b_t)$ is performed on b_t servers in parallel. In the (b_t+1)-th server, draw S samples from the DFP conditional distributions of λ^2 and σ^2 given by $\lambda^2 \sim Gamma\left(p+r, \frac{\sum_{j=1}^{p}\widehat{\tau}_j^{2(t-1)}}{2} + d\right)$, $\sigma^2 \sim IG\left(\frac{nt+p}{2}, \frac{\left(\boldsymbol{S}_3^{(t)} + \widehat{\boldsymbol{\beta}}^{(t-1)'}\boldsymbol{S}_1^{(t)}\widehat{\boldsymbol{\beta}}^{(t-1)} - 2\widehat{\boldsymbol{\beta}}^{(t-1)'}\boldsymbol{S}_2^{(t)}\right) + \widehat{\boldsymbol{\beta}}^{(t-1)'}(\widehat{\boldsymbol{M}}_{\tau}^{(t-1)})^{-1}\widehat{\boldsymbol{\beta}}^{(t-1)}}{2}\right)$.

5. *Compute the sequence of estimators at time t:* Set $\widehat{\boldsymbol{\beta}}^{(t)}$, $\widehat{\boldsymbol{\tau}}^{2(t)}$, $\widehat{\sigma}^{2(t)}$, $\widehat{\lambda}^{2(t)}$ from their respective sample averages from S MCMC samples.

Figure 1 presents MSPE, coverage, interval score for the 95% predictive intervals and computation time in seconds per batch of the competing methods for Simulation 1. Figures 2 and 3 highlight the same quantities for Simulations 2 and 3 respectively, except the computation time which is similar for competitors across the three simulations. Batch MCMC,

being a batch method, is expected to converge faster. The predictive inference of DFP improves rapidly and becomes indistinguishable from batch MCMC within $t \approx 100 - 150$ for all three simulations. In contrast, the predictive performance of C-DF appears to be inferior to batch MCMC even at $t = 150$. The substantial gain in predictive inference of DFP over C-DF can perhaps be attributed to the dynamic partitioning of the parameter space, thereby learning posterior correlations among parameters more accurately over time which yields a better approximation of the full posterior. Additionally, DFP approximation enjoys significant reduction in per batch run time over its competitors.

The average MSPE, run time, coverage and interval scores of 95% predictive intervals over the last 100 time points for all the competitors are presented in Table 1. Table 1 shows that in all three simulations, DFP emerges as a computationally efficient replacement for batch MCMC, both in terms of point prediction as well as characterizing predictive uncertainties. As mentioned earlier, naive implementation of C-DF demonstrates inferior predictive inference. An improved implementation of C-DF presented here, in contrast, loses appeal with minimal gain in computation time over batch MCMC. The SSMC approach also demonstrates similar inferential performance with DFP with a higher computation time.

Due to space constraint, density estimates for a few selected predictor coefficients are displayed at $t = 250, 500$. Since Simulation 1 is the most interesting scenario, posterior densities of a randomly chosen zero coefficient, a nonzero coefficient with a lower magnitude and a nonzero coefficient with a higher magnitude are presented in Figure 1. Posterior densities of the selected β_j's in the batch MCMC and DFP tend to show discrepancies in the earlier time points. These discrepancies diminish at $t = 500$, empirically validating the fact that approximate DFP draws converge to the full posterior distribution in time. This conclusion remains valid for Simulations 2 and 3.

4.2 DFP with Horseshoe

Our second application considers implementing DFP on the Horseshoe shrinkage prior (Carvalho et al., 2010). The full conditional distributions of parameters along with computational issues in implementing Gibbs sampling with the Horseshoe shrinkage prior are given in Section 2.2. The DFP algorithm is employed to incur computational benefits in

Table 1: Bayesian Lasso performance statistics for MCMC, CDF, DFP and SSMC. Coverage and length are based on the average of the 95% credible predictive intervals in the last 100 batches. The subscript provides standard errors calculated over 10 replications.

Low & High Sparse

Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
MCMC	$0.914_{0.019}$	$0.002_{0.000}$	$3.827_{0.345}$	$339.578_{66.343}$
DFP	$0.897_{0.021}$	$0.002_{0.000}$	$3.925_{0.370}$	$148.292_{43.878}$
CDF	$0.902_{0.021}$	$0.002_{0.000}$	$3.897_{0.370}$	$303.215_{73.600}$
SSMC	$0.903_{0.018}$	$0.002_{0.000}$	$3.811_{0.355}$	$234.198_{57.627}$

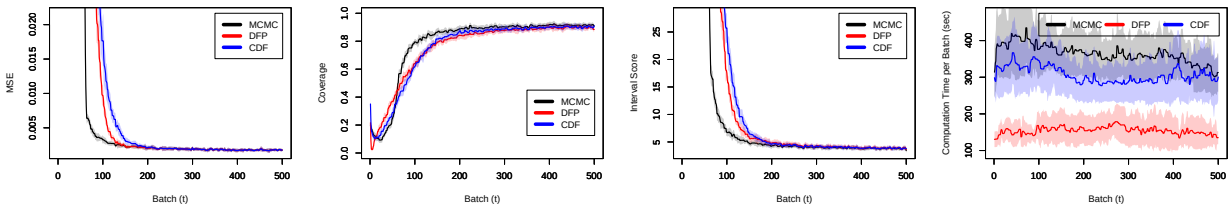
Sparse

Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
MCMC	$0.915_{0.021}$	$0.002_{0.000}$	$3.502_{0.345}$	$400.203_{88.666}$
DFP	$0.898_{0.023}$	$0.002_{0.000}$	$3.592_{0.393}$	$162.788_{58.104}$
CDF	$0.903_{0.023}$	$0.002_{0.000}$	$3.556_{0.380}$	$365.983_{71.200}$
SSMC	$0.912_{0.021}$	$0.002_{0.000}$	$3.512_{0.346}$	$289.179_{66.265}$

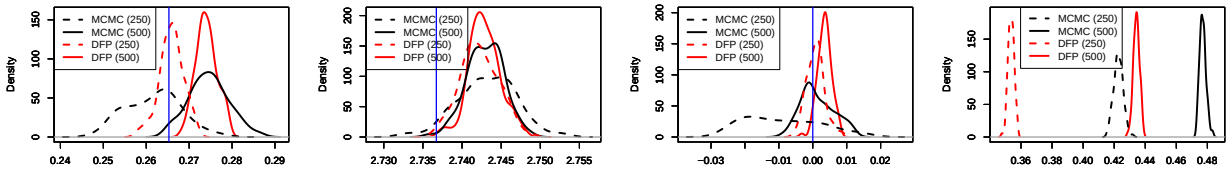
Dense

Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
MCMC	$0.940_{0.017}$	$4e-05_{1e-05}$	$1.629_{0.121}$	$377.822_{128.891}$
DFP	$0.917_{0.019}$	$4e-05_{1e-05}$	$1.662_{0.148}$	$145.340_{48.056}$
CDF	$0.919_{0.018}$	$4e-05_{1e-05}$	$1.654_{0.143}$	$352.099_{105.388}$
SSMC	$0.943_{0.016}$	$4e-05_{1e-05}$	$1.628_{0.121}$	$278.354_{65.505}$

Figure 1: Performance measures for MCMC, DFP and CDF in the case of Bayesian Lasso under the high and low sparse case are presented in the first row. Coverage and Interval scores are based on the average of the 95% predictive intervals. The second row shows estimated densities of selected parameters at $t = 250$ and $t = 500$ for DFP and batch MCMC. Confidence bands are based on repeating the analysis over 10 replications.



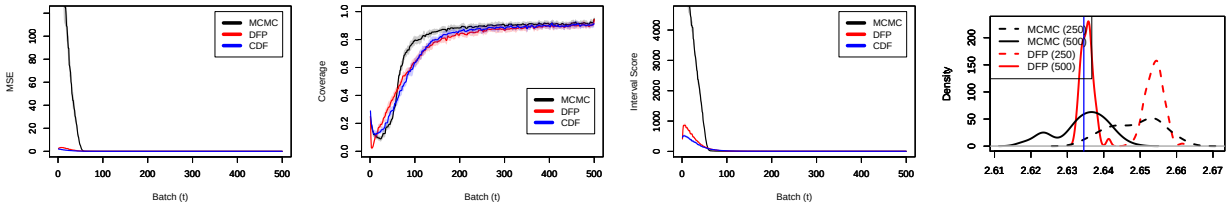
(a) MSPE (b) Coverage (c) Interval Score (d) Time



(e) Low β (f) High β (g) Zero β (h) σ^2

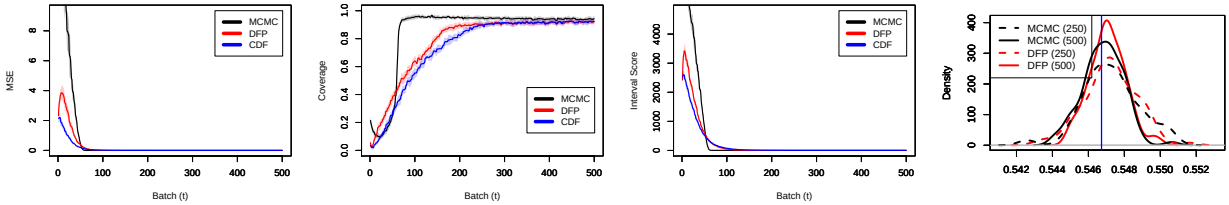
situations with large p. The DFP algorithm applied to this problem considers partitioning

Figure 2: Performance measures for MCMC, DFP and CDF for Bayesian Lasso under the sparse case are presented. Coverage and Interval scores are based on the average of the 95% predictive intervals. We also show estimated densities for a selected β_j at $t = 250$ and $t = 500$ for both batch MCMC and DFP.



(a) MSPE (b) Coverage (c) Interval Score (d) Non zero β

Figure 3: Performance measures for MCMC, DFP and CDF for Bayesian Lasso under the dense case. Coverage and Interval scores are based on the average of the 95% predictive intervals. Estimated densities of selected parameters at $t = 250$ and $t = 500$ for both batch MCMC and DFP are also added.



(a) MSPE (b) Coverage (c) Interval Score (d) β

the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \sigma^2, \tau^2, \xi\}$ into $k_t = b_t + 2$ subsets at time t given by

$$\boldsymbol{\Theta}_{G_l^{(t)}} = \left\{ \beta_{i^{(t)}_{m_1+\cdots+m_{l-1}+1}}, \lambda^2_{i^{(t)}_{m_1+\cdots+m_{l-1}+1}}, ..., \beta_{i^{(t)}_{m_1+\cdots+m_l}}, \lambda^2_{i^{(t)}_{m_1+\cdots+m_l}} \right\}, \;\; l = 1, .., b_t,$$

$$\boldsymbol{\Theta}_{G^{(t)}_{b_t+1}} = \Big\{ \boldsymbol{\nu} \Big\}, \;\; \boldsymbol{\Theta}_{G^{(t)}_{b_t+2}} = \Big\{ \sigma^2, \tau^2, \xi \Big\}.$$

Let $\boldsymbol{\beta}_l$ and $\boldsymbol{\lambda}_l$ be the vector of β_js and λ_j^2s, respectively, corresponding to the lth partition. Define $\boldsymbol{S}^{(t)}_{1,l}$, $\boldsymbol{S}^{(t)}_{2,l}$ and $\boldsymbol{S}^{(t)}_{1,l,-l}$ as in Section 4.1. Let $\boldsymbol{M}_{\lambda,l} = diag(\boldsymbol{\lambda}_l)$ and $\boldsymbol{\beta}_{-l}$ be the β_js not contained in $\boldsymbol{\beta}_l$. The DFP algorithm proceeds as follows.

1. Set $\widehat{\boldsymbol{\beta}}^{(0)}, \widehat{\sigma}^{2(0)}, \widehat{\boldsymbol{\lambda}}^{2(0)}, \widehat{\boldsymbol{\nu}}^{2(0)}, \widehat{\tau}^{2(0)}$ and $\widehat{\xi}^{(0)}$ at their initial values.

2. Observe data $\boldsymbol{D}_t = \{\boldsymbol{y}_t, \boldsymbol{X}_t\}$ at time t. Update the partitions of the parameters based on the iterates of the parameters at time $(t-1)$. The dynamic partitioning scheme for

parameters for shrinkage priors described in Section 3 is employed. Throughout, the partitions $G_{b_t+1}^{(t)}$ and $G_{b_t+2}^{(t)}$ are kept fixed.

3. Update sufficient statistics $\boldsymbol{S}_1^{(t)}, \boldsymbol{S}_2^{(t)}, \boldsymbol{S}_3^{(t)}$ based on $\boldsymbol{S}_1^{(t-1)}, \boldsymbol{S}_2^{(t-1)}, \boldsymbol{S}_3^{(t-1)}$ and $\boldsymbol{D}_t$ with the equations given in Section 2.2.

4. Draw S samples from the DFP conditional distributions of $\boldsymbol{\beta}_l$ and $\boldsymbol{\lambda}_l$ given by

$$\lambda_j^2 \sim IG\left(1, \left[\frac{1}{\widehat{\nu}_j^{(t-1)}} + \frac{\beta_j^2}{2\widehat{\tau}^{2(t-1)}\widehat{\sigma}^{2(t-1)}}\right]\right),\ \lambda_j^2 \in \boldsymbol{\lambda}_l^2,\quad \boldsymbol{\beta}_l \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}_l^{(t)}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_l^{(t)}}\right)$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_l^{(t)}} = \left(\boldsymbol{S}_{1,l}^{(t)} + \frac{\boldsymbol{M}_{\lambda,l}^{-1}}{\tau^2}\right)^{-1}\left(\boldsymbol{S}_{2,l}^{(t)} - \boldsymbol{S}_{1,l,-l}^{(t)}\widehat{\boldsymbol{\beta}_{-l}}^{(t-1)}\right),\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_l^{(t)}} = \widehat{\sigma}^{2(t-1)}\left(\boldsymbol{S}_{1,l}^{(t)} + \frac{\boldsymbol{M}_{\lambda,l}^{-1}}{\tau^2}\right)^{-1},$$

Sampling from the DFP full conditionals of $\{\boldsymbol{\beta}_l, \boldsymbol{\lambda}_l\}$ $(l = 1, .., b_t)$ are performed on b_t servers in parallel with the number of flops at most M^3 at every server. Draw S samples from the DFP full conditionals of $\boldsymbol{\nu}$ given by $\nu_j \sim IG\left(1, \left(1 + \frac{1}{\widehat{\lambda}_j^{2(t-1)}}\right)\right)$, $j = 1, ..., p$, in the $(b_t + 1)$-th server. Finally, in the $(b_t + 2)$-th server, draw S samples from the DFP full conditional posterior distributions of τ^2, ξ, σ^2 given by $\xi \sim IG\left(1, 1 + \frac{1}{\tau^2}\right)$,
$\tau^2 \sim IG\left(\frac{p+1}{2}, \frac{1}{\xi} + \frac{\hat{\boldsymbol{\beta}}^{(t-1)'}(\widehat{\boldsymbol{M}}_\lambda^{(t-1)})^{-1}\hat{\boldsymbol{\beta}}^{(t-1)}}{2\sigma^2}\right)$,
$\sigma^2 \sim IG\left(\frac{nt+p}{2}, \frac{\boldsymbol{S}_3^{(t)} + \hat{\boldsymbol{\beta}}^{(t-1)'}\boldsymbol{S}_1^{(t)}\hat{\boldsymbol{\beta}}^{(t-1)} - 2\hat{\boldsymbol{\beta}}^{(t-1)'}\boldsymbol{S}_2^{(t)}}{2} + \frac{\hat{\boldsymbol{\beta}}^{(t-1)'}(\widehat{\boldsymbol{M}}_\lambda^{(t-1)})^{-1}\hat{\boldsymbol{\beta}}^{(t-1)}}{2\tau^2}\right)$.

5. Set $\widehat{\boldsymbol{\beta}}^{(t)}$, $\widehat{\boldsymbol{\lambda}}^{2(t)}$, $\widehat{\boldsymbol{\nu}}^{(t)}$, $\widehat{\tau}^{2(t)}$, $\widehat{\sigma}^{2(t)}$ and $\widehat{\xi}^{(t)}$ as their respective sample averages from S MCMC samples.

Figure 4 presents dynamically evolving MSPE, coverage, interval score for the 95% predictive interval and computation time in seconds per batch of the competing methods for Simulation 1. As observed in Section 4.1, MSPE for DFP falls sharply as time progresses and becomes indistinguishable with the MSPE of batch MCMC after $t \approx 200 - 250$. While accurate point prediction is one of our primary objectives, characterizing uncertainty is of paramount importance given the recent development in the frequentist literature on characterizing uncertainties in high dimensional regressions (Javanmard and Montanari, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014). Although Bayesian procedures provide an automatic characterization of uncertainty, the resulting credible intervals may not possess

Table 2: Horseshoe performance statistics for MCMC, C-DF, SSMC and DFP. Coverage and interval scores are based on the average of the 95% credible predictive intervals of the last 100 batches. Subscripts provide standard errors over 10 simulations.

Low & High Sparse

Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
MCMC	$0.924_{0.019}$	$0.002_{0.001}$	$3.725_{1.006}$	$298.126_{52.808}$
DFP	$0.905_{0.020}$	$0.002_{0.000}$	$3.715_{0.341}$	$143.587_{30.989}$
CDF	$0.909_{0.020}$	$0.002_{0.000}$	$3.704_{0.338}$	$289.120_{58.688}$
SSMC	$0.922_{0.021}$	$0.002_{0.001}$	$3.722_{1.006}$	$288.783_{83.226}$

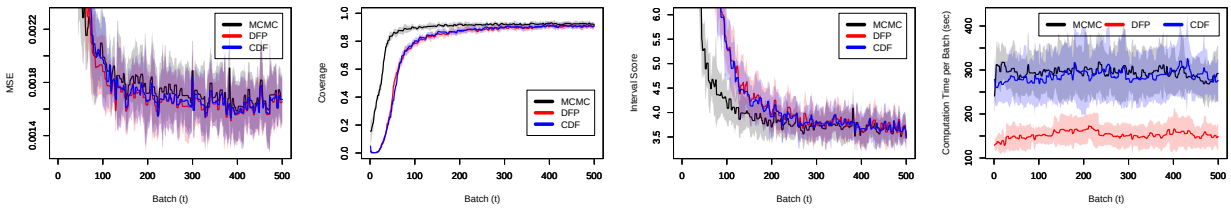
Sparse

Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
MCMC	$0.925_{0.021}$	$0.002_{0.001}$	$3.375_{1.004}$	$357.010_{64.220}$
DFP	$0.906_{0.021}$	$0.002_{0.000}$	$3.386_{0.343}$	$164.555_{42.560}$
CDF	$0.910_{0.022}$	$0.002_{0.000}$	$3.372_{0.349}$	$329.129_{83.201}$
SSMC	$0.923_{0.022}$	$0.002_{0.001}$	$3.377_{1.026}$	$338.996_{66.246}$

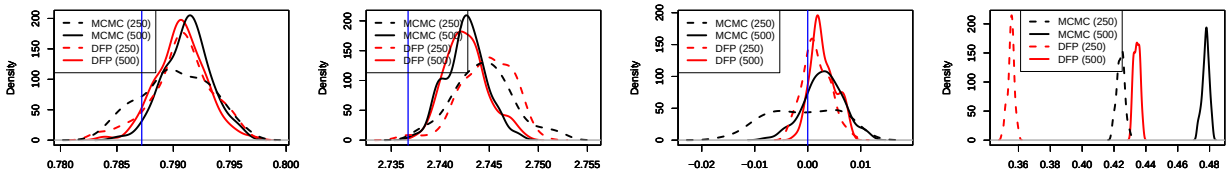
Dense

Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
MCMC	$0.931_{0.018}$	$0.001_{0.000}$	$2.383_{21.448}$	$262.594_{34.915}$
DFP	$0.891_{0.022}$	$4e-05_{1e-05}$	$1.749_{0.180}$	$117.416_{14.589}$
CDF	$0.903_{0.021}$	$3e-05_{1e-05}$	$1.696_{0.162}$	$261.798_{68.321}$
SSMC	$0.932_{0.017}$	$0.001_{0.001}$	$2.221_{3.996}$	$311.438_{70.867}$

Figure 4: Performance measures for MCMC, DFP and C-DF in the case of Horseshoe under the high and low sparse case are presented in the first row. Coverage and Interval scores are based on the average of the 95% predictive intervals. The second row shows estimated densities of selected parameters at $t = 250$ and $t = 500$ for both batch MCMC and DFP. Confidence bands are based on the analysis over 10 replications.



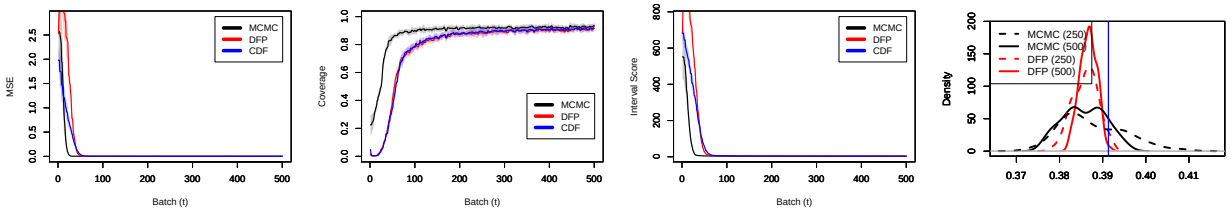
(a) MSPE (b) Coverage (c) Interval Score (d) Time



(e) Low β (f) High β (g) Zero β (h) σ^2

the correct frequentist coverage in nonparametric/high-dimensional problems (Szabó et al., 2015). An attractive adaptive property of the shrinkage priors, including Horseshoe, is that

Figure 5: Performance measures for MCMC, DFP and C-DF for Horseshoe under the sparse case are presented. Coverage and Interval scores are based on the average of the 95% predictive intervals. We also show estimated densities of a selected β_j at $t = 250$ and $t = 500$ for both batch MCMC and DFP.



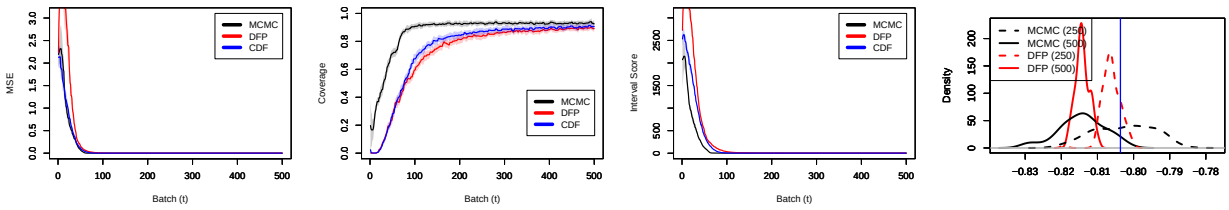
(a) MSPE (b) Coverage (c) Interval Score (d) Selected β

the lengths of the intervals automatically adapt between the signal and noise variables, maintaining close to nominal coverage. Approximate Bayesian inference with the DFP algorithm is found to preserve this desirable property of the Horseshoe prior. In fact, Figures 4, 5 and 6 show similar coverage and interval scores for DFP and batch MCMC as time progresses. This observation is further reinforced from Table 2 which demonstrates practically identical performances of batch MCMC, CDF, SSMC and DFP, with DFP having notably reduced computation time.

Density estimates for a few selected predictor coefficients are displayed at $t = 250, 500$. Since Simulation 1 is the most interesting scenario, posterior densities of a randomly chosen zero coefficient, a nonzero coefficient with a lower magnitude and a nonzero coefficient with a higher magnitude are presented in Figure 4. For nonzero coefficients, the density estimates seem to be similar in DFP and in batch MCMC, though DFP yields marginally narrower credible intervals than batch MCMC corresponding to zero coefficients. One fundamental advantage of the Horseshoe shrinkage prior over frequentist penalized optimization is its ability to accurately characterize parametric and predictive uncertainties without any user dependent choice of tuning parameters. However, it might lose this appeal due to its high computation time and inability to provide rapid inference with big n and p. DFP applied to the Horseshoe prior solves the computational bottleneck for big n and p, perhaps offering wider applicability to the Horseshoe prior in regression problems at a much larger scale. We expect similar conclusions to hold for other state-of-the-art shrinkage priors such as, the

Generalized Double Pareto (Armagan et al., 2013) and the normal gamma (Griffin et al., 2010) prior distributions.

Figure 6: Performance measures for MCMC, DFP and C-DF for Horseshoe under the dense case are presented. Coverage and Interval scores are based on the average of the 95% predictive intervals. We also show estimated densities of a selected β_j at $t = 250$ and $t = 500$ for both batch MCMC and DFP.



(a) MSPE (b) Coverage (c) Interval Score (d) β

4.3 Spike and Lasso

Without being too repetitive, we briefly sketch the main steps of implementing the DFP algorithm with the Spike and Lasso prior as follows.

1. *Initialize:* Set $\widehat{\boldsymbol{\beta}}^{(0)}, \widehat{\sigma}^{2(0)}, \widehat{\lambda}^{2(0)}, \widehat{\boldsymbol{\gamma}}^{2(0)}, \widehat{\boldsymbol{\tau}}^{2(0)}$ and $\widehat{\theta}^{(0)}$ at their initial values.

2. *Parameter space partitioning at time t:* Observe data $\boldsymbol{D}_t = \{\boldsymbol{y}_t, \boldsymbol{X}_t\}$ at time t. Update the partitions of the parameters based on the iterates of the parameters at time $(t-1)$. As discussed in the partitioning scheme for the Spike and Lasso prior in Section 3, the number of partitions is $k_t = 1 + |\{j : (\beta_j, \tau_j^2) \in \boldsymbol{\Theta}_{2t}\}|$, where $|\cdot|$ denotes the cardinality of the set.

3. *Update sufficient statistics:* Update sufficient statistics $\boldsymbol{S}_1^{(t)}, \boldsymbol{S}_2^{(t)}$ and $\boldsymbol{S}_3^{(t)}$ based on $\boldsymbol{S}_1^{(t)} = \boldsymbol{S}_1^{(t-1)} + \boldsymbol{X}_t'\boldsymbol{X}_t$, $\boldsymbol{S}_2^{(t)} = \boldsymbol{S}_2^{(t-1)} + \boldsymbol{X}_t'\boldsymbol{y}_t$ and $\boldsymbol{S}_3^{(t)} = \boldsymbol{S}_3^{(t-1)} + \boldsymbol{y}_t'\boldsymbol{y}_t$.

4. *Draw approximate posterior samples at time t:* Define $\mathcal{I}_{1t} = \{j : (\boldsymbol{\beta}_j, \tau_j^2) \in \boldsymbol{\Theta}_{1t}\}$, where $\boldsymbol{\Theta}_{1t} = \{(\beta_j, \tau_j^2) : \hat{\gamma}_j^{(t-1)} = 1\}$. In a server, draw S samples from the DFP full conditional posterior distributions of $\boldsymbol{\beta}_{\mathcal{I}_{1t}} = (\beta_j : j \in \mathcal{I}_{1t})'$ and $\boldsymbol{\tau}^2_{\mathcal{I}_{1t}} = (\tau_j^2 : j \in \mathcal{I}_{1t})'$

Table 3: Spike and Lasso performance statistics for MCMC, CDF, SSMC and DFP. MSPE, Coverage and interval scores are based on the average of the 95% credible predictive intervals for the last 100 batches.

Sparse				
Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
MCMC	$0.921_{0.021}$	$0.002_{0.000}$	$3.479_{0.335}$	$396.730_{97.681}$
DFP	$0.898_{0.023}$	$0.002_{0.000}$	$3.587_{0.388}$	$9.262_{3.476}$
CDF	$0.894_{0.023}$	$0.002_{0.000}$	$3.595_{0.385}$	$395.402_{136.833}$
SSMC	$0.922_{0.02}$	$0.002_{0.001}$	$3.483_{0.379}$	$311.897_{52.019}$
Low & High Sparse				
Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
MCMC	$0.922_{0.019}$	$0.002_{0.000}$	$3.795_{0.324}$	$393.422_{55.556}$
DFP	$0.897_{0.021}$	$0.002_{0.000}$	$3.929_{0.385}$	$9.406_{2.886}$
CDF	$0.892_{0.021}$	$0.002_{0.000}$	$3.982_{0.380}$	$407.424_{50.365}$
SSMC	$0.925_{0.017}$	$0.002_{0.001}$	$3.802_{0.333}$	$314.783_{45.451}$

given by

$$\frac{1}{\tau_j^2}|\cdot \sim Inv-Gaussian\left(\sqrt{\frac{\widehat{\lambda}^{2(t-1)}\widehat{\sigma}^{2(t-1)}}{\beta_j^2}}, \widehat{\lambda}^{2(t-1)}\right)\ \forall\, j\in\mathcal{I}_{1t},\ \boldsymbol{\beta}_{\mathcal{I}_{1t}}\sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}^{(t)}_{\mathcal{I}_{1t}}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(t)}_{\mathcal{I}_{1t}}}\right)$$

$$\boldsymbol{\mu}_{\boldsymbol{\beta}^{(t)}_{\mathcal{I}_{1t}}}=\left(\boldsymbol{S}^{(t)}_{1,\mathcal{I}_{1t}}+\boldsymbol{M}^{-1}_{\mathcal{I}_{1t}}\right)^{-1}\left(\boldsymbol{S}^{(t)}_{2,\mathcal{I}_{1t}}-\boldsymbol{S}^{(t)}_{1,\mathcal{I}_{1t},-\mathcal{I}_{1t}}\widehat{\boldsymbol{\beta}_{-\mathcal{I}_{1t}}}^{(t-1)}\right),\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}^{(t)}_{\mathcal{I}_{1t}}}=\widehat{\sigma}^{2(t-1)}\left(\boldsymbol{S}^{(t)}_{1,\mathcal{I}_{1t}}+\boldsymbol{M}^{-1}_{\mathcal{I}_{1t}}\right)^{-1},$$

where $\boldsymbol{M}_{\mathcal{I}_{1t}}$ is a sub-matrix of $\boldsymbol{M}$ corresponding to the indices of $\mathcal{I}_{1t}$, $\boldsymbol{S}^{(t)}_{1,\mathcal{I}_{1t}}, \boldsymbol{S}^{(t)}_{1,\mathcal{I}_{1t},-\mathcal{I}_{1t}}$ and $\boldsymbol{S}^{(t)}_{2,\mathcal{I}_{1t}}$ are defined analogous to the last section. Similarly draw (β_j,τ_j^2) for $j\in\mathcal{I}_{2t}=\{j:(\beta_j,\tau_j^2)\in\boldsymbol{\Theta}_{2t}\}$, $\boldsymbol{\Theta}_{2t}=\{(\beta_j,\tau_j^2):\hat{\gamma}_j^{(t-1)}=0\}$ in different processors from their DFP full conditional distributions. Draw S samples from the DFP full conditional posterior distributions of $\boldsymbol{\gamma}$ given in (4) with $\sigma^2,\boldsymbol{\beta},\boldsymbol{\tau}$ replaced by their point estimates from time $(t-1)$. Finally, draw S samples from the DFP full conditional posterior distributions of λ^2,σ^2 and θ in a server.

5. *Compute the sequence of estimators at time t:* Set $\widehat{\boldsymbol{\beta}}^{(t)}$, $\widehat{\boldsymbol{\tau}}^{2(t)}$, $\widehat{\lambda}^{2(t)}$, $\widehat{\sigma}^{2(t)}$ and $\widehat{\theta}^{(t)}$ as their respective sample averages from S MCMC samples. Set $\widehat{\gamma}_j^{(t)}=1$ if out of S approximate posterior samples of γ_j at time $(t-1)$, at least $S/2$ have resulted in $\gamma_j=1$.

Since spike and slab prior distributions are primarily designed to identify important variables in sparse high dimensional regressions, we investigate DFP with the Spike and Lasso prior for Simulations 1 and 2. Figure 7 presents the dynamic progression of various

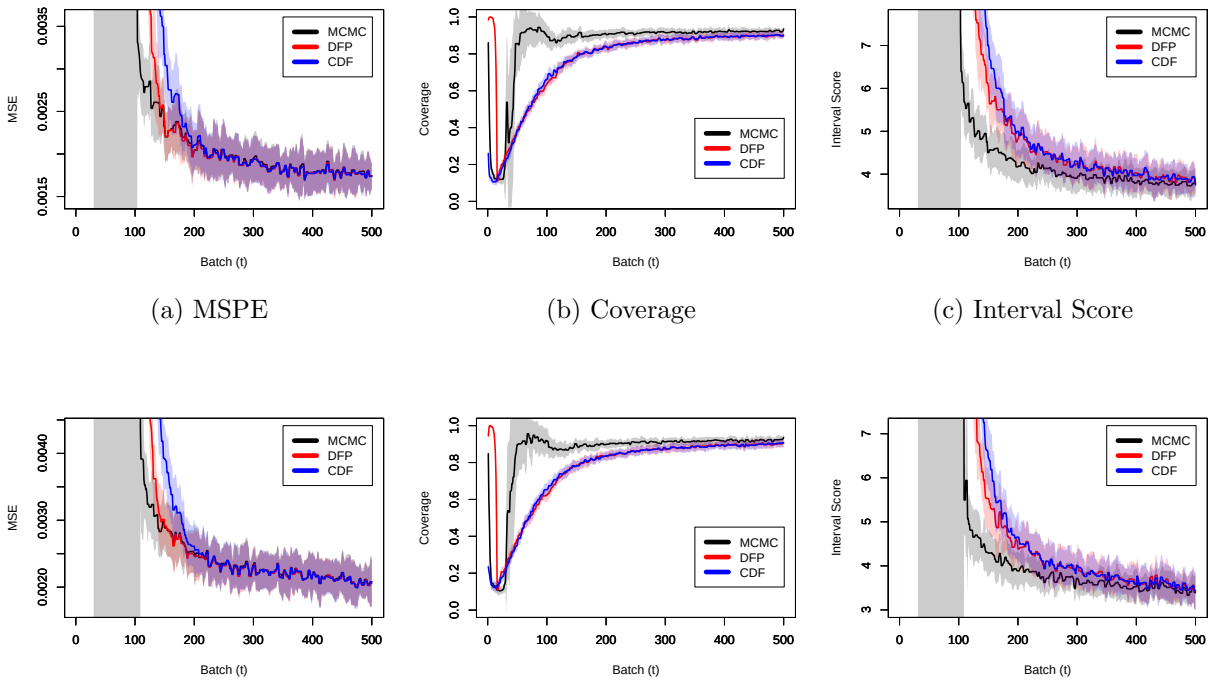
performance metrics for DFP, batch MCMC and C-DF over $T = 500$ time points. Unlike Sections 4.1 and 4.2, the operating characteristics of the Spike and Lasso applied to all three competitors take longer time to stabilize. This is not surprising, given that batch MCMC with spike and slab mixture priors is known to offer less accurate performance with a smaller sample size due to the high correlation between various γ_j's. As before, DFP approximates batch MCMC accurately in terms of the operating characteristics. In fact, Table 3 shows practically indistinguishable performance of DFP and batch MCMC, while C-DF yields marginally larger interval scores even at latter time points. SSMC continues to show competitive performance with a much higher computation time compared to DFP. DFP dynamically learns the partition based on $\mathbf{\Theta}_{1t}$ and $\mathbf{\Theta}_{2t}$. Since we consider sparse examples, the cardinality of the set $\mathbf{\Theta}_{1t}$ is never large, and hence the parameters therein can be updated quickly. Our detailed investigation also reveals that even a large number of partitions of $\mathbf{\Theta}_{2t}$ does not compromise the accuracy of the inference and prediction. This helps to accrue substantial gains in computation time for DFP compared to its competitors, as demonstrated in Table 3. In contrast, C-DF fixes the partitions in the beginning and is unable to leverage the information of the zero and nonzero β_j's as the approximate posterior sampling progresses.

Representative posterior densities of β_j's from DFP and batch MCMC are presented in Figure 8. Both in Simulations 1 and 2, the posterior densities of β_j's for DFP and batch MCMC are centered around the truth and have similar tails. Both Simulations 1 and 2 involve high sparsity, resulting in the posterior density of θ centered at a small value. Again there is a considerable agreement in the posterior densities of θ from DFP and batch MCMC. Finally, posterior densities of σ^2 for DFP and batch MCMC are found to differ by a small margin from the truth.

4.4 Sensitivity to the choice of S

One of the important ingredients in the development of DFP is the choice of the number of Monte Carlo samples S at every time and it is instructive to see the effect on inference with different choices of S. The simulation section presents results of DFP with $S = 500$. To assess the sensitivity to the choice of S in our simulations, we compute DFP after moderately

Figure 7: Performance measures for MCMC, DFP and C-DF with the Spike and Lasso prior under Simulations 1 (1st row) and 2 (second row). Coverage and interval scores are based on the average of the 95% predictive intervals.



(a) MSPE (b) Coverage (c) Interval Score

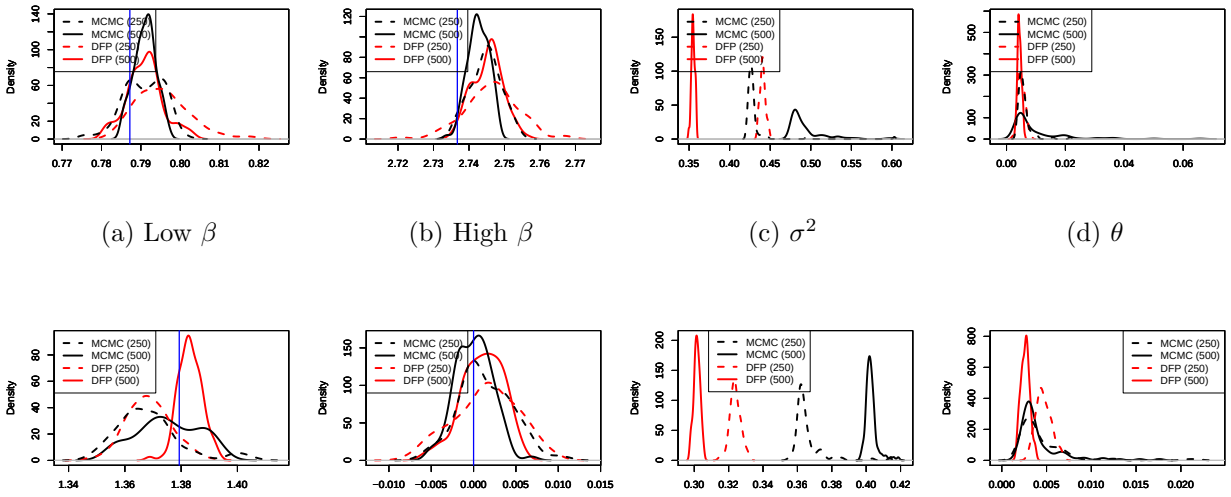
(d) MSPE (e) Coverage (f) Interval Score

perturbing S. Table 4 presents the predictive inference with DFP for $S = 500, 750, 1000$ in the different simulation cases with the Bayesian Lasso prior. The results show practically indistinguishable inference with different choices of S, with $S = 750$ and $S = 1000$ naturally incurring much more computational cost. In our experience, the inference can be marginally improved with much larger choices of S, though such choicees practically diminish any computational advantage of DFP.

5 Application to Financial Stock Database

To illustrate the performance of DFP, we implement DFP for a financial data set consisting of minute by minute average log-prices of the NASDAQ stock exchange from September 10, 2018 to November 13, 2018 during trading hours. The data consists of log-prices of Apple stocks along with 3430 assets, and the aim of the data analysis is to evaluate the elasticity of the price of Apple stocks with respect to the prices of the remaining assets. This is of

Figure 8: Estimated densities for a few selected β_js, σ^2 and θ at $t = 250$ and $t = 500$. The first row presents results for Simulation 1 while the second row demonstrates performance of DFP in Simulation 2.



(a) Low β (b) High β (c) σ^2 (d) θ

(e) Non zero β (f) Zero β (g) σ^2 (h) θ

particular interest, since Apple, one of the biggest publicly traded companies in the world, is ubiquitous in portfolios ranging from retirement funds to small portfolios managed by individuals in the financial market. Thus accurate inference on the relationship between Apple and other financial stocks allows better portfolio diversification. We envision it as a high dimensional linear regression problem with the log-price of the Apple stock as the response and log-prices of other assets as predictors. Along with prediction, the inferential interest lies mainly in identifying important predictors significantly associated with the response. Hence the *Spike & Lasso* prior on regression coefficients are employed.

The data includes several assets, such as ETFs, Trust Funds, stock tracker indexes, and banks, which as expected, present a very high degree of collinearity. To avoid less desirable inference due to high collinearity, a few financial assets are removed along with assets which have very few transactions (less than 40), yielding 2015 predictors for the analysis. The data set consists of 18330 observations collected over two months.

To compare the predictive inference of DFP with respect to the gold standard "batch MCMC," the dataset is divided into 183 approximately equal shards to implement DFP and the "gold standard" batch MCMC. Both are implemented 10 times with 10 different

Table 4: Bayesian Lasso performance statistics for DFP with $S = 500, 750, 1000$. Coverage and length are based on the average of the 95% predictive intervals on the last 100 batches. The subscript provides standard errors calculated over 10 replications.

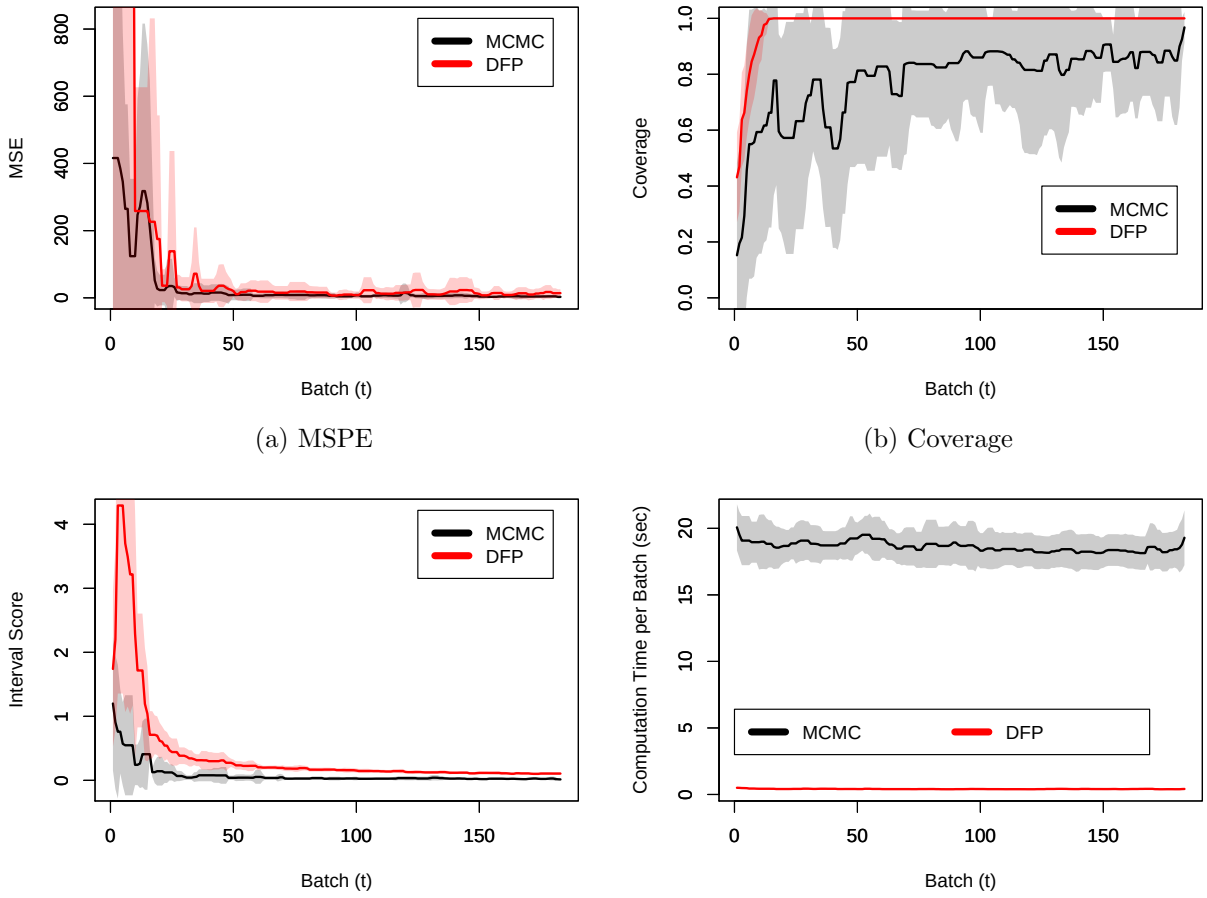
Low & High Sparse				
Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
DFP($S = 500$)	$0.897_{0.021}$	$0.002_{0.000}$	$3.925_{0.370}$	$148.292_{43.878}$
DFP($S = 750$)	$0.906_{0.024}$	$0.002_{0.000}$	$3.957_{0.344}$	$243.176_{48.245}$
DFP($S = 1000$)	$0.912_{0.015}$	$0.002_{0.000}$	$3.954_{0.358}$	$309.542_{44.268}$
Sparse				
Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
DFP($S = 500$)	$0.898_{0.023}$	$0.002_{0.000}$	$3.592_{0.393}$	$162.788_{58.104}$
DFP($S = 750$)	$0.903_{0.028}$	$0.002_{0.000}$	$3.578_{0.369}$	$248.927_{54.200}$
DFP($S = 1000$)	$0.911_{0.022}$	$0.002_{0.000}$	$3.589_{0.327}$	$316.178_{59.264}$
Dense				
Method	*Predictive Coverage*	*MSPE*	*Int. Score*	*Runtime (sec)*
DFP($S = 500$)	$0.917_{0.019}$	$4e-05_{1e-05}$	$1.662_{0.148}$	$145.340_{48.056}$
DFP($S = 750$)	$0.919_{0.017}$	$4e-05_{1e-05}$	$1.684_{0.143}$	$234.099_{46.498}$
DFP($S = 750$)	$0.919_{0.016}$	$4e-05_{1e-05}$	$1.678_{0.141}$	$305.354_{46.491}$

permutations of the dataset to minimize the effect of sample ordering on the identification of influential variables. Furthermore, this allows us to examine if the predictive inferential mechanism in DFP is sufficiently robust to the inaccurate posterior approximations at earlier time points.

Figure 9 tracks the progression of MSPE, interval score and coverage of 95% predictive intervals for both DFP and batch MCMC as more batches are processed. At time t, the predictive inference is assessed with the data shard obtained at time $t + 1$. Similar to simulation studies, the behavior of DFP in the early batches is somewhat erratic due to the inaccurate posterior approximation in the initial phase of the algorithm, though it stabilizes as more data shards arrive. Furthermore, the performances of the competitors become closer as time progresses, with batch MCMC demonstrating marginally superior performance at higher time points. The dramatic improvement of DFP over batch MCMC is mainly observed in terms of computation time. While batch MCMC runs 500 iterations per batch in 18.35 seconds, DFP finishes 500 iterations per batch in 0.40 seconds. Such a dramatic improvement in computation time is consistent with our findings in Section 3.3 and can be attributed to efficient partitioning of the parameter space as well as parallel inference on parameter partitions at each time.

While fitting the data using the high dimensional regression model with the Spike and

Figure 9: Performance measures for MCMC and DFP. MSPE, coverage and interval scores for 95% predictive intervals are presented. Confidence bands (in a lighter color) are calculated by observing the variations of these metrics over 10 permutations.



(a) MSPE

(b) Coverage

(c) Interval Score

(d) Time

Lasso prior, we observe a high degree of multi-modality in the posterior distribution. Although a high degree of multi-modality in the high dimensional regression is known to have minimal effects on the predictive inference, it may provide somewhat unreliable inference in terms of variable selection. This is observed and noted in the earlier literature on high dimensional regression (see e.g., Guha and Rodriguez (2018)). In such cases, it is customary to run the posterior computation multiple times, record the set of variables being identified in each of these runs, and finally declare those variables as influential which have appeared as influential in more than half of the runs. Due to the multi-modality in the posterior

distribution, we observe that 10 runs of both DFP and batch MCMC do not lead to the same set of variables identified. In fact, we find a difference in the conclusion between DFP and MCMC in terms of identified variables.

To ensure more reliable inference from DFP and the "gold standard" batch MCMC for variable selection, we run both these competitors 10 more times on the dataset of interest. In these 10 runs, the data is divided into 163 shards with the first shard having 20% observations, and the rest 162 shards all approximately equal. We observe that feeding more data early on leads to reliable variable selection with minimal variation between different runs. To provide concrete evidence on this observation, we refer to Table 5 which presents all predictors identified by either DFP or batch MCMC in any of the 10 runs. The table also records the number of times among the 10 runs they are identified as influential. It shows that the number of times a predictor is selected by either batch MCMC or DFP is very close to 0 or 10, indicating quite reliable variable selection. Importantly, much less discrepancy is observed between DFP and batch MCMC, with them identifying 17 and 21 variables as influential respectively, with 14 identified by both.

6 Conclusion and Future Work

The emergence of large volumes of high dimensional data mandates that model fitting tools evolve quickly to keep pace with the rapidly growing dimension and size of data. Although the literature in high dimensional Bayesian inference has witnessed recent upsurge, there are limited number of Bayesian methods, online in nature, which enable efficient Bayesian model fitting in high dimensional linear regression in presence of large or streaming data. The DFP algorithm proposed in this article dynamically partitions the parameter space after observing every data shard and employs fast and approximate Bayesian inference at each partition in parallel. The detailed simulation studies of DFP with popular Bayesian shrinkage priors (Bayesian Lasso, Horseshoe and Spike and Lasso) show indistinguishable inference from batch MCMC with a considerable reduction of per batch computation time. The supplementary material contains the proof of convergence of the DFP algorithm for high dimensional linear regression as time $t \to \infty$.

The scope of DFP extends well beyond the realm of high dimensional linear regression

Table 5: Number of times a stock is selected under DFP and MCMC out of 10 runs of both methods.

Company	*DFP*	*MCMC*
Allscripts Healthcare Solutions, Inc.	10	10
Alphabet Inc.	10	10
Century Aluminum Company	10	10
Ferroglobe PLC	10	10
Skyworks Solutions, Inc.	10	10
Red Robin Gourmet Burgers, Inc.	9	10
Viavi Solutions Inc.	9	10
The Kraft Heinz Company	8	10
Amazon.com, Inc.	7	10
Popular, Inc.	7	9
Caesarstone Ltd.	7	9
Microsoft Corporation	8	9
SeaSpine Holdings Corporation	6	10
Qorvo, Inc.	7	10
Costco Wholesale Corporation	7	0
iQIYI, Inc.	8	0
The Ultimate Software Group, Inc.	7	0
Global Water Resources, Inc.	0	10
Kala Pharmaceuticals, Inc.	0	10
National General Holdings Corp	0	10
Applied Optoelectronics, Inc.	0	9
Atlas Air Worldwide Holdings	0	9
Baozun Inc.	0	9
Genprex, Inc.	0	9

with Gaussian errors. For example, as part of our future work, we will employ DFP for high dimensional logistic and probit regressions. While data augmentation schemes (Albert and Chib, 1993; Polson et al., 2013) in high dimensional binary regression allow Gibbs sampling for parameter blocks, making the DFP formulation natural, they also violate assumptions (1) and (2) in the formulation of DFP in Section 3. Thus, one needs to develop a modification of DFP to account for a growing number of latent variables as time progresses. To this end, instead of propagating the sufficient statistics over time, we intend to propagate a quantity known as the surrogate conditional sufficient statistics (Guhaniyogi et al., 2018) that eliminates the need for these assumptions. In the same vein, we propose to extend the DFP formulation for high dimensional linear regression with heavy tailed error distributions. Notably, a heavy tailed error distribution can often be expressed as a scale mixture of Gaussian errors. Thus, upon using a data augmentation scheme, developing DFP under this model will require extending the DFP framework when the number of parameters increases with the onset of a new data shard. Another important research direction is to develop the DFP algorithm for non-local priors in high dimensions. We would also like to extend our theoretical results on the convergence of the DFP kernel to the full posterior from a fixed partitioning set up to an adaptive dynamic partitioning set up. The DFP approach can presumably provide competitive inference with Sequential Monte Carlo in high dimensional factor models and dynamic factor models where shrinkage priors have been employed recently. Some of these constitute our current area of research.

7 Acknowledgement

The research of Rajarshi Guhaniyogi is partially supported by grants from the Office of Naval Research (ONR-BAA N000141812741) and the National Science Foundation (DMS-1854662).

References

Albert, J. H. and S. Chib (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American statistical Association 88*(422), 669–679.

Armagan, A., D. B. Dunson, and J. Lee (2013). Generalized Double Pareto Shrinkage.

Statistica Sinica 23(1), 119.

Armagan, A., D. B. Dunson, J. Lee, W. U. Bajwa, and N. Strawn (2013). Posterior Consistency in Linear Models under Shrinkage Priors. *Biometrika 100*(4), 1011–1018.

Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002). A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. *IEEE Transactions on signal processing 50*(2), 174–188.

Beskos, A., D. O. Crisan, A. Jasra, and N. Whiteley (2014). Error Bounds and Normalising Constants for Sequential Monte Carlo Samplers in High Dimensions. *Advances in Applied Probability 46*(1), 279–306.

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Campbell, T., J. Straub, J. W. Fisher III, and J. P. How (2015). Streaming, Distributed Variational Inference for Bayesian Nonparametrics. In *Advances in Neural Information Processing Systems*, pp. 280–288.

Caron, F. and A. Doucet (2008). Sparse Bayesian Nonparametric Regression. In *Proceedings of the 25th international conference on Machine learning*, pp. 88–95. ACM.

Carvalho, C. M., H. F. Lopes, N. G. Polson, M. A. Taddy, et al. (2010). Particle Learning for General Mixtures. *Bayesian Analysis 5*(4), 709–740.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The Horseshoe Estimator for Sparse Signals. *Biometrika 97*(2), 465–480.

Castillo, I., J. Schmidt-Hieber, A. Van der Vaart, et al. (2015). Bayesian Linear Regression with Sparse Priors. *The Annals of Statistics 43*(5), 1986–2018.

Chopin, N. (2002). A Sequential Particle Filter Method for Static Models. *Biometrika 89*(3), 539–552.

Chopin, N. et al. (2004). Central Limit Theorem for Sequential Monte Carlo Methods and Its Application to Bayesian Inference. *The Annals of Statistics 32*(6), 2385–2411.

Doucet, A., N. De Freitas, and N. Gordon (2001). An Introduction to Sequential Monte Carlo Methods. In *Sequential Monte Carlo methods in practice*, pp. 3–14. Springer.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian Variable Selection. *Statistica sinica*, 339–373.

Gneiting, T. and A. E. Raftery (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Griffin, J. E., P. J. Brown, et al. (2010). Inference with Normal-Gamma Prior Distributions in Regression Problems. *Bayesian Analysis 5*(1), 171–188.

Guha, S. and A. Rodriguez (2018). Bayesian regression with undirected network predictors with an application to brain connectome data. *arXiv preprint arXiv:1803.10655*.

Guhaniyogi, R., S. Qamar, and D. B. Dunson (2018). Bayesian Conditional Density Filtering. *Journal of Computational and Graphical Statistics* (just-accepted).

Gunawan, D., K.-D. Dang, M. Quiroz, R. Kohn, and M.-N. Tran (2018). Subsampling sequential monte carlo for static bayesian models. *arXiv preprint arXiv:1805.03317*.

Hager, W. W. (1989). Updating The Inverse of a Matrix. *SIAM review 31*(2), 221–239.

Hoffman, M., F. R. Bach, and D. M. Blei (2010). Online Learning for Latent Dirichlet Allocation. In *advances in neural information processing systems*, pp. 856–864.

Javanmard, A. and A. Montanari (2014). Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *The Journal of Machine Learning Research 15*(1), 2869–2909.

Lindsten, F., A. M. Johansen, C. A. Naesseth, B. Kirkpatrick, T. B. Schön, J. Aston, and A. Bouchard-Côté (2017). Divide-and-conquer With Sequential Monte Carlo. *Journal of Computational and Graphical Statistics 26*(2), 445–458.

Lopes, H. F. and R. S. Tsay (2011). Particle Filters and Bayesian Inference in Financial Econometrics. *Journal of Forecasting 30*(1), 168–209.

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

Polson, N. G. and J. G. Scott (2010). Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. *Bayesian statistics 9*, 501–538.

Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association 108*(504), 1339–1349.

Rebeschini, P., R. Van Handel, et al. (2015). Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability 25*(5), 2809–2866.

Ročková, V. and E. I. George (2016). The Spike-and-Slab Lasso. *Journal of the American Statistical Association* (just-accepted).

Rue, H. (2001). Fast Sampling of Gaussian Markov Random Fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 325–338.

Sato, M.-A. (2001). Online Model Selection Based on The Variational Bayes. *Neural computation 13*(7), 1649–1681.

Schweizer, N. (2012). Non-asymptotic Error Bounds for Sequential MCMC Methods in Multimodal Settings. *arXiv preprint arXiv:1205.6733*.

Scott, J. G. and J. O. Berger (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *The Annals of Statistics*, 2587–2619.

Snyder, C., T. Bengtsson, P. Bickel, and J. Anderson (2008). Obstacles to High-Dimensional Particle Filtering. *Monthly Weather Review 136*(12), 4629–4640.

Szabó, B., A. van der Vaart, J. van Zanten, et al. (2015). Frequentist Coverage of Adaptive Nonparametric Bayesian Credible Sets. *The Annals of Statistics 43*(4), 1391–1428.

Van de Geer, S., P. Bühlmann, Y. Ritov, R. Dezeure, et al. (2014). On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models. *The Annals of Statistics 42*(3), 1166–1202.

Wang, X., D. B. Dunson, and C. Leng (2016). Decorrelated Feature Space Partitioning for Distributed Sparse Regression. In *Advances in Neural Information Processing Systems*, pp. 802–810.

Wigren, A., L. Murray, and F. Lindsten (2018). Improving the particle filter in high dimensions using conjugate artificial process noise. *IFAC-PapersOnLine 51*(15), 670–675.

Zhang, C.-H. and S. S. Zhang (2014). Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zhou, Y. and A. Jasra (2015). Biased Online Parameter Inference for State-Space Models. *arXiv preprint arXiv:1503.00266*.

Supplementary Material: Bayesian Dynamic Feature Partitioning in High-Dimensional Regression with Big Data

February 23, 2020

Convergence Behavior of Approximate Samplers

We study convergence behavior for the DFP algorithm provided in Section 3 of the main article. Since developing results with dynamic partitioning is challenging given that any partitioning scheme exploits specifics of model and prior distributions, the results developed here establish convergence of the DFP algorithm with the assumption that the partitioning of the parameter set is fixed over time. Although this is a restrictive assumption, DFP seems to enjoy desirable asymptotic behavior even under this assumption. With dynamic partitioning, we expect to witness stronger theoretical results for DFP, which needs separate attention in a future work.

The theoretical development proceeds in a few steps. DFP algorithm being a Markov chain framework assumes a transition kernel (denoted by $T_t(\cdot, \cdot)$ at time t) and a stationary distribution of the transition kernel at each time t (referred to as the DFP stationary distribution and denoted by π_t). At first, we establish the general form of the DFP stationary distribution π_t at each time t. Next, we develop sufficient conditions on the transition kernel ($T_t(\cdot, \cdot)$), no. of samples (S) drawn from the transition kernel at each time t, dynamic evolution of the DFP stationary distributions over time and conditions on the point estimates ($\widehat{\boldsymbol{\Theta}}^{(t)}$) to ensure convergence of the DFP transition kernel to the full posterior distribution

asymptotically. Some of these conditions are verified for the specific cases of high dimensional linear regression with shrinkage priors and spike and slab priors. To begin with, we define a few quantities.

Notation and Framework

For the sake of simplicity denote $\boldsymbol{\Theta}_{G_l^t} = \boldsymbol{\Theta}_{l,t} \in \mathbb{R}^{q_l}$ for $l = 1, ..., k_t$. Since our theoretical exposition fixes partitions over time t, $k_t = k$ and q_l's are not functions of time t and $\sum_{l=1}^{k_t} q_l = q = dim(\boldsymbol{\Theta})$. Assume $\boldsymbol{\Theta}_{l,t} = (\theta_{l,t,1}, ..., \theta_{l,t,q_l})'$. The full posterior distribution of $\boldsymbol{\Theta}$ at time t, denoted by $f(\boldsymbol{\Theta}|\boldsymbol{S}^{(t)})$ in Section 3, is also shortened as $f_t(\boldsymbol{\Theta})$. Assume that the density $f_t(\boldsymbol{\Theta})$ is admitted with respect to the Lebesgue measure ν. $T_t : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}^+$ is a transition kernel at time t having the property that $T_t(\boldsymbol{z}, \cdot)$ is a probability measure for all $\boldsymbol{z} \in \mathbb{R}^q$ and $T_t(\cdot, \boldsymbol{A})$ is a measurable function for all $\boldsymbol{A}$ in the Borel sigma algebra of $\mathbb{R}^q$. Finally, we denote $\widehat{\boldsymbol{\Theta}}_{-G_l^t}^{(t)}$ and $\widehat{\boldsymbol{\Theta}}_{G_l^t}^{(t)}$ as $\widehat{\boldsymbol{\Theta}}_{-l}^{(t)}$ and $\widehat{\boldsymbol{\Theta}}_{l}^{(t)}$ respectively, $l = 1, ..., k$, for the simplicity of notation.

0.1 The DFP transition kernel

It follows from the DFP algorithm that the DFP transition kernel $T_t : \mathbb{R}^{q_1} \times \cdots \times \mathbb{R}^{q_k} \to \mathbb{R}^+$ at time t is given by:

$$T_t(\boldsymbol{\Theta}, \boldsymbol{\Theta}') = \prod_{l=1}^{k} \prod_{i=1}^{q_l} f_t\big(\theta'_{l,t,i} | \widehat{\boldsymbol{\Theta}}_{-l}^{(t-1)}, \theta'_{l,t,j}, j < i, \theta_{l,t,j}, j > i\big) \tag{1}$$

The unique stationary distribution $\pi_t : \mathbb{R}^q \to \mathbb{R}^+$ of the transition kernel T_t at time t is given in the following lemma.

Lemma 0.1 *DFP approximate kernel T_t has a unique stationary distribution $\pi_t(\boldsymbol{\Theta}) = \prod_{l=1}^{k_t} f_t(\boldsymbol{\Theta}_l | \widehat{\boldsymbol{\Theta}}_{-l}^{(t-1)})$.*

Proof In order to prove the lemma, we will simply show that π_t given by the equation above

satisfies $\int T_t(\mathbf{\Theta}, \mathbf{\Theta}')\pi_t(\mathbf{\Theta})\, d\mathbf{\Theta} = \pi_t(\mathbf{\Theta}')$. Note that

$$\int T_t(\mathbf{\Theta}, \mathbf{\Theta}')\pi_t(\mathbf{\Theta}) d\mathbf{\Theta} = \int \prod_{l=1}^{k} \Big[\prod_{i=1}^{q_l} f_t(\theta'_{l,t,i} | \widehat{\mathbf{\Theta}}_{-l}^{(t-1)}, \theta'_{l,t,j}, j < i, \theta_{l,t,j}, j > i) f_t(\mathbf{\Theta}_l | \widehat{\mathbf{\Theta}}_{-l}^{(t-1)}) \Big] d\mathbf{\Theta}$$

$$= \prod_{l=1}^{k} \int \Big[\prod_{i=1}^{q_l} f_t(\theta'_{l,t,i} | \widehat{\mathbf{\Theta}}_{-l}^{(t-1)}, \theta'_{l,t,j}, j < i, \theta_{l,t,j}, j > i) f_t(\mathbf{\Theta}_l | \widehat{\mathbf{\Theta}}_{-l}^{(t-1)}) \Big] d\mathbf{\Theta} = \prod_{l=1}^{k} f_t(\mathbf{\Theta}'_l | \widehat{\mathbf{\Theta}}_{-l}^{(t-1)}).$$

Here the last step follows by recognizing that the kernel T_t is a product of various independent Gibbs sampler (or Metropolis Hastings) kernels in different parameter partitions.

0.2 Main convergence results

We will now state a theorem and a corollary. The theorem states reasonable assumptions to ensure decay of the total variation distance between DFP transition kernel and its stationary distribution as t increases. The corollary then adds a few more sufficient conditions to ensure that DFP kernel becomes close to the full posterior distribution as t increases. Let π_0 denote the initial distribution from which parameters are drawn. Suppose T_t^S denotes the kernel corresponding to S draws from the DFP kernel T_t. We use $||\cdot||_{TV}$ to denote the total variation distance and $d_H(\cdot, \cdot)$ to denote the Hellinger distance between two densities. The statement of the theorem is given below.

Theorem 0.2 *Let $\epsilon \in (0, 1)$. Assume $\exists$ a constant $C > 0$, a positive integer S and a function $V : \mathbb{R}^q \to [1, \infty)$ s.t. for all large t,*

(i) $E_{\pi_t}(V^2) \leq C$

(ii) $||T_t(\mathbf{\Theta}, \cdot)^S - \pi_t||_{TV} \leq V(\mathbf{\Theta})\alpha_t^S < 1 - \epsilon \ \forall \ \mathbf{\Theta}$ *and for some* $\alpha_t \in (0, 1)$.

Then,

$$||T_t^S \cdots T_1^S - \pi_t||_{TV} \leq \sum_{s=1}^{t} \epsilon^{t+1-s} \rho_s, \tag{2}$$

where $\rho_t = 2\sqrt{C} d_H(\pi_t, \pi_{t-1})$.

The proof of Theorem 0.2 follows along the same line of the proof of Theorem 3.6 in (Yang and Dunson, 2013) and is thus omitted.

Corollary 0.3 *If conditions (i) and (ii) of Theorem 0.2 are satisfied and additionally we assume (iii)* $\rho_t \to 0$ *and (iv)* $||\pi_t - f_t||_{TV} \to 0$*, as* $t \to \infty$*. Then* $||T_t^S \cdots T_1^S - f_t||_{TV} \to 0$*.*

Poof: Using conditions (i), (ii) and (iii), as $t \to \infty$, $||T_t^S \cdots T_1^S - \pi_t||_{TV} \to 0$, following Theorem 0.2. Now we use (iv) to deduce that $||T_t^S \cdots T_1^S - f_t||_{TV} \leq ||T_t^S \cdots T_1^S - \pi_t||_{TV} + ||\pi_t - f_t||_{TV} \to 0$, as $t \to \infty$.

Remark Corollary 0.3 shows that the DFP transition kernel after S draws each at time $1, ..., t$ becomes close to the full posterior distribution f_t at time t. This implies that as time t increases, samples drawn from the DFP full conditional distributions can be taken as the draws from the un-approximated full posterior distribution f_t.

Next, we argue that the assumptions in Theorem 0.2 and Corollary 0.3 are reasonable. Note that conditions (i) and (ii) refer to the assumption that the DFP transition kernel at time t converges to the DFP stationary distribution at time t at a geometric rate. This assumption is also referred to as the *Geometric Ergodicity* assumption. We first prove that this assumption holds for shrinkage and spike and lasso priors used in this article. Condition (iii) ensures that the stationary distribution of the approximating kernel changes slowly as time proceeds. This is a mild condition satisfied by any regular parametric model by applying the Bernstein-Von Mises theorem. Finally, we prove condition (iv) under mild assumptions.

We will now proceed to verify Geometric Ergodicty for the DFP kernel with some of the Gaussian scale mixture priors and spike and lasso prior. The theorem below shows conditions for geometric ergodicity under Bayesian lasso prior. The proof uses some of the techniques outlined in Pal and Khare (2014).

Theorem 0.4 *Assume there exists* $m_0 > 0$ *s.t.* $e_{min}(\boldsymbol{S}_{1,\nabla}^{(t)}) \geq m_0$*, for any set* $\nabla \subseteq \{1, ..., p\}$ *and any* $t = 1, ..., T$*, where* $\boldsymbol{S}_{1,\nabla}^{(t)}$ *is a submatrix of* $\boldsymbol{S}_1^{(t)}$ *with columns corresponding to the indices* ∇*. Then the DFP Bayesian lasso transition kernel is geometrically ergodic.*

Proof If $T_t((\boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2, \lambda^2), (\boldsymbol{\beta}', (\boldsymbol{\tau}^2)', (\sigma^2)', (\lambda^2)'))$ is the transition kernel of the DFP and $\pi_t(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2, \lambda^2)$ is the stationary distribution of the transition kernel, then $T_t(\cdot, \cdot)$ and $\pi_t(\cdot)$

for the Bayesian lasso are given by

$$
\begin{aligned}
&T_t((\boldsymbol{\beta},\boldsymbol{\tau}^2,\sigma^2,\lambda^2),((\boldsymbol{\beta})',(\boldsymbol{\tau}^2)',(\sigma^2)',(\lambda^2)') \\
&= \prod_l \prod_{j\in G_l^t} \left\{ f_t((\beta_j)'|(\tau_j^2),\hat{\boldsymbol{\beta}}_{-l}^{(t-1)},\hat{\boldsymbol{\tau}}_{-l}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)}) f_t((\tau_j^2)'|(\beta_j)',\hat{\boldsymbol{\beta}}_{-l}^{(t-1)},\hat{\boldsymbol{\tau}}_{-l}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)}) \right\} \\
&\quad f_t((\sigma^2)'|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)}) f_t((\lambda^2)'|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)}) \qquad (3)
\end{aligned}
$$

$$
\begin{aligned}
&\pi_t(\boldsymbol{\beta},\boldsymbol{\tau}^2,\sigma^2,\lambda^2) \\
&= \prod_{l=1}^{k} \left\{ f_t(\boldsymbol{\beta}_l,\boldsymbol{\tau}_l^2|\hat{\boldsymbol{\beta}}_{-l}^{(t-1)},\hat{\boldsymbol{\tau}}_{-l}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)}) \right\} f_t(\sigma^2|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)}) f_t(\lambda^2|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)}). \\
&\qquad (4)
\end{aligned}
$$

Hence, $||T_t - \pi_t||_{TV} = ||\tilde{T}_{t,1} - \tilde{\pi}_{t,1}||_{TV}$, where

$$
\begin{aligned}
\tilde{T}_{t,1} &= \prod_{j\in G_l^t} \left\{ f_t((\beta_j)'|(\tau_j^2),\hat{\boldsymbol{\beta}}_{-l}^{(t-1)},\hat{\boldsymbol{\tau}}_{-l}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)}) f_t((\tau_j^2)'|(\beta_j)',\hat{\boldsymbol{\beta}}_{-l}^{(t-1)},\hat{\boldsymbol{\tau}}_{-l}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)}) \right\} \\
\tilde{\pi}_{t,1} &= \prod_{l=1}^{k} \left\{ f_t(\boldsymbol{\beta}_l,\boldsymbol{\tau}_l^2|\hat{\boldsymbol{\beta}}_{-l}^{(t-1)},\hat{\boldsymbol{\tau}}_{-l}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)}) \right\}.
\end{aligned}
$$

Thus it is enough to show the geometric ergodicity of the chain by establishing a geometric drift condition and a geometric minorization condition for the $(\boldsymbol{\beta},\boldsymbol{\tau}^2)$ chain.

Minorization condition.

Define, $\tilde{V}_t(\boldsymbol{\beta},\boldsymbol{\tau}^2) = \frac{\sum_{j=1}^p \beta_j^2}{\frac{1}{m_0}\sum_{l=1}^k \boldsymbol{H}_l'\boldsymbol{H}_l+1} + \sum_{j=1}^p \tau_j^2$, where $\boldsymbol{H}_l = \boldsymbol{S}_{2,l}^{(t)} - \boldsymbol{S}_{1,l,-l}^{(t)}\hat{\boldsymbol{\beta}}_{-l}^{(t-1)}$. Let $\mathcal{S}_{\tilde{V}_t,d} = \{(\boldsymbol{\beta},\boldsymbol{\tau}^2) : \tilde{V}_t(\boldsymbol{\beta},\boldsymbol{\tau}^2) \le d\}$. While showing minorization condition, we will establish that there exists a constant $0 < c(\tilde{V}_t,d) < 1$ depending on $\tilde{V}_t$ and d such that $\tilde{T}_{t,1}((\boldsymbol{\beta}_0,\boldsymbol{\tau}_0^2),(\boldsymbol{\beta},\boldsymbol{\tau}^2)) \ge c(\tilde{V}_t,d)g(\boldsymbol{\beta},\boldsymbol{\tau}^2)$ for some density function $g(\cdot)$ for every $(\boldsymbol{\beta}_0,\boldsymbol{\tau}_0^2) \in \mathcal{S}_{\tilde{V}_t,d}$. Denote $\tilde{\lambda} = \hat{\lambda}^{2(t-1)}$

and $\tilde{\mu}_j = \sqrt{\frac{\hat{\sigma}^{2(t-1)}\hat{\lambda}^{2(t-1)}}{\beta_{0j}^2}}$. Then

$$\begin{aligned} f_t(\tau_j^2|\boldsymbol{\beta}_0) &= \sqrt{\frac{\tilde{\lambda}}{2\pi}}(\tau_j^2)^{-1/2}\exp\left\{-\tilde{\lambda}\frac{(1-\tau_j^2\tilde{\mu}_j)^2}{2\tilde{\mu}_j^2\tau_j^2}\right\} \\ &= \sqrt{\frac{\tilde{\lambda}}{2\pi}}(\tau_j^2)^{-1/2}\exp\left\{-\frac{\tilde{\lambda}\tau_j^2}{2} - \frac{\tilde{\lambda}}{2\tilde{\mu}_j^2\tau_j^2} + \frac{\tilde{\lambda}}{\tau_j}\right\} \\ &\geq \sqrt{\frac{\tilde{\lambda}}{2\pi}}(\tau_j^2)^{-1/2}\exp\left\{-\frac{\tilde{\lambda}\tau_j^2}{2} - \frac{\tilde{\lambda}}{2\tilde{\mu}_j^2\tau_j^2}\right\}. \end{aligned}$$

Note that $\tilde{V}_t(\boldsymbol{\beta}_0, \tau_0^2) \leq d$ implies $\frac{\boldsymbol{\beta}_0'\boldsymbol{\beta}_0}{\frac{1}{m_0}\sum_{l=1}^k \boldsymbol{H}_l'\boldsymbol{H}_l+1} + \sum_{j=1}^p \tau_j^2 \leq d$ when $(\boldsymbol{\beta}_0, \boldsymbol{\tau}_0^2) \in \mathcal{S}_{\tilde{V}_t,d}$. Thus $\boldsymbol{\beta}_0'\boldsymbol{\beta}_0 \leq d\left[\frac{1}{m_0}\sum_{l=1}^k \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]$ when $(\boldsymbol{\beta}_0, \boldsymbol{\tau}_0^2) \in \mathcal{S}_{\tilde{V}_t,d}$.

$$\begin{aligned} f_t(\tau_j^2|\boldsymbol{\beta}_0) &\geq \sqrt{\frac{\hat{\lambda}^{2(t-1)}}{2\pi}}(\tau_j^2)^{-1/2}\exp\left\{-\frac{\hat{\lambda}^{2(t-1)}\tau_j^2}{2} - \frac{d\left[\frac{1}{m_0}\sum_{l=1}^k \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]}{2\tau_j^2\hat{\sigma}^{2(t-1)}}\right\} \\ &\geq \sqrt{\frac{\hat{\lambda}^{2(t-1)}}{2\pi}}(\tau_j^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\sqrt{\hat{\lambda}^{2(t-1)}\tau_j^2} - \sqrt{\frac{d\left[\frac{1}{m_0}\sum_{l=1}^k \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]}{\tau_j^2\hat{\sigma}^{2(t-1)}}}\right)^2\right\} \\ &\quad \exp\left\{-\sqrt{\frac{\hat{\lambda}^{2(t-1)}d\left[\frac{1}{m_0}\sum_{l=1}^k \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]}{\hat{\sigma}^{2(t-1)}}}\right\} \end{aligned}$$

Let $c(\tilde{V}_t, d) = \exp\left\{-\sqrt{\frac{\hat{\lambda}^{2(t-1)}d\left[\frac{1}{m_0}\sum_{l=1}^k \boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}{\hat{\sigma}^{2(t-1)}}}\right\}$. Thus

$$\begin{aligned} \tilde{T}_{t,1}((\boldsymbol{\beta}_0, \boldsymbol{\tau}_0^2), (\boldsymbol{\beta}, \boldsymbol{\tau}^2)) &= \prod_{l=1}^k \prod_{j\in G_l^t} f_t(\beta_j|\tau_j^2, \hat{\sigma}^{2(t-1)}, \hat{\lambda}^{2(t-1)}) f_t(\tau_j^2|\beta_{0j}, \hat{\sigma}^{2(t-1)}, \hat{\lambda}^{2(t-1)}) \\ &\geq c(\tilde{V}_t, d)\prod_{l=1}^k \prod_{j\in G_l^t}\prod_{j=1}^p f_t(\beta_j|\tau_j^2, \hat{\sigma}^{2(t-1)}, \hat{\lambda}^{2(t-1)}) g_t(\tau_j^2|\hat{\sigma}^{2(t-1)}, \hat{\lambda}^{2(t-1)}), \end{aligned}$$

where $g_t(\tau_j^2|\hat{\sigma}^{2(t-1)}, \hat{\lambda}^{2(t-1)}) = \sqrt{\frac{\hat{\lambda}^{2(t-1)}}{2\pi}}(\tau_j^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\sqrt{\hat{\lambda}^{2(t-1)}\tau_j^2} - \sqrt{\frac{d\left[\frac{1}{m_0}\sum_{l=1}^k \boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}{\tau_j^2\hat{\sigma}^{2(t-1)}}}\right)^2\right\}$

is a density function. Hence the minorization condition is established.

Geometric drift condition.

$$E[\sum_{l=1}^{k}\tilde{V}_t(\boldsymbol{\beta}_l,\boldsymbol{\tau}_l^2)|\boldsymbol{\beta}_0,\boldsymbol{\tau}_0^2]=E_2[E_1[\sum_{l=1}^{k}\tilde{V}_t(\boldsymbol{\beta}_l,\boldsymbol{\tau}_l^2)|\boldsymbol{\beta}_0,\boldsymbol{\tau}_0^2]],$$

where the inner expectation is w.r.t conditional distribution of $\boldsymbol{\beta}|\boldsymbol{\tau}_0^2$ and the outer expectation is w.r.t. $\boldsymbol{\tau}^2|\boldsymbol{\beta}_0$.

$$\begin{aligned}
&E_1[\sum_{l=1}^{k}\tilde{V}_t(\boldsymbol{\beta}_l,\boldsymbol{\tau}_l^2)|\boldsymbol{\beta}_0,\boldsymbol{\tau}_0^2]\\
&=\frac{\sum_{l=1}^{k}\boldsymbol{H}_l'(\boldsymbol{S}_{1,l}^{(t)}+\boldsymbol{M}_{\boldsymbol{\tau}_{0,l}}^{-1})^{-1}(\boldsymbol{S}_{1,l}^{(t)}+\boldsymbol{M}_{\boldsymbol{\tau}_{0,l}}^{-1})^{-1}\boldsymbol{H}_l+tr(\hat{\sigma}^{2(t-1)}(\boldsymbol{S}_{1,l}^{(t)}+\boldsymbol{M}_{\boldsymbol{\tau}_{0,l}}^{-1})^{-1})}{\left[\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}+\sum_{j=1}^{p}\tau_j^2\\
&\leq\frac{(\sum_{j=1}^{p}\tau_{0j}^2)\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l+\hat{\sigma}^{2(t-1)}\frac{p}{m_0}}{\left[\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}+\sum_{j=1}^{p}\tau_j^2\\
&\leq(\sum_{j=1}^{p}\tau_{0j}^2)\frac{\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l}{\left[\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}+\frac{\hat{\sigma}^{2(t-1)}\frac{p}{m_0}}{\left[\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}+\sum_{j=1}^{p}\tau_j^2, \qquad (5)
\end{aligned}$$

where the second step follows

$$(\boldsymbol{S}_{1,l}^{(t)}+\boldsymbol{M}_{\boldsymbol{\tau}_{0,l}}^{-1})^{-1}\leq\frac{1}{m_0}\boldsymbol{I},\ \ (\boldsymbol{S}_{1,l}^{(t)}+\boldsymbol{M}_{\boldsymbol{\tau}_{0,l}}^{-1})^{-1}\leq\sum_{j=1}^{p}\tau_j^2. \qquad (6)$$

$$\begin{aligned}
E_2[\sum_{j=1}^{p}\tau_j^2]&=\sum_{j=1}^{p}\left[\sqrt{\frac{\beta_{0j}^2}{\hat{\sigma}^{2(t-1)}\hat{\lambda}^{2(t-1)}}}+\frac{1}{\hat{\lambda}^{2(t-1)}}\right]\\
&=\sum_{j=1}^{p}\sqrt{\frac{\beta_{0j}^2}{\left[\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}\frac{\left[\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}{\hat{\sigma}^{2(t-1)}\hat{\lambda}^{2(t-1)}}}+\frac{p}{\hat{\lambda}^{2(t-1)}}\\
&\leq\frac{\boldsymbol{\beta}_0'\boldsymbol{\beta}_0}{2\left[\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}+\frac{p\left[\frac{1}{m_0}\sum_{l=1}^{k}\boldsymbol{H}_l'\boldsymbol{H}_l+1\right]}{2\hat{\sigma}^{2(t-1)}\hat{\lambda}^{2(t-1)}}+\frac{p}{\hat{\lambda}^{2(t-1)}}, \qquad (7)
\end{aligned}$$

where the last inequality follows by the Cauchy-Schwartz inequality. Using (5) and (7),

$$
\begin{aligned}
E[\sum_{l=1}^{k} V(\boldsymbol{\beta}_l, \boldsymbol{\tau}_l^2)|\boldsymbol{\beta}_0, \boldsymbol{\tau}_0^2] &\leq (\sum_{j=1}^{p} \tau_{0j}^2) \frac{\frac{1}{m_0}\sum_{l=1}^{k} \boldsymbol{H}_l'\boldsymbol{H}_l}{\left[\frac{1}{m_0}\sum_{l=1}^{k} \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]} + \frac{\hat{\sigma}^{2(t-1)}\frac{p}{m_0}}{\left[\frac{1}{m_0}\sum_{l=1}^{k} \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]} \\
&\quad + \frac{\boldsymbol{\beta}_0'\boldsymbol{\beta}_0}{2\left[\frac{1}{m_0}\sum_{l=1}^{k} \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]} + \frac{p\left[\frac{1}{m_0}\sum_{l=1}^{k} \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]}{2\hat{\sigma}^{2(t-1)}\hat{\lambda}^{2(t-1)}} + \frac{p}{\hat{\lambda}^{2(t-1)}} \\
&\leq \gamma V(\boldsymbol{\beta}_0, \boldsymbol{\tau}_0^2) + b,
\end{aligned}
$$

where $0 < \gamma = \max\left\{\frac{1}{2}, \frac{\frac{1}{m_0}\sum_{l=1}^{k} \boldsymbol{H}_l'\boldsymbol{H}_l}{\left[\frac{1}{m_0}\sum_{l=1}^{k} \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]}\right\} < 1$ and $b = \frac{\hat{\sigma}^{2(t-1)}\frac{p}{m_0}}{\left[\frac{1}{m_0}\sum_{l=1}^{k} \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]} + \frac{p}{\hat{\lambda}^{2(t-1)}} + \frac{p\left[\frac{1}{m_0}\sum_{l=1}^{k} \boldsymbol{H}_l'\boldsymbol{H}_l + 1\right]}{2\hat{\sigma}^{2(t-1)}\hat{\lambda}^{2(t-1)}} > 0$. Hence the geometric drift condition is satisfied. Geometric drift and minorization condition together implies geometric ergodicity of the chain.

We will now prove a similar result for the spike and lasso model. Indeed,

Theorem 0.5 *Assume there exists $m_0 > 0$ s.t. $e_{min}(\boldsymbol{S}_{1,\nabla}^{(t)}) \geq m_0$, for any set $\nabla \subseteq \{1, ..., p\}$ and any $t = 1, ..., T$, where $\boldsymbol{S}_{1,\nabla}^{(t)}$ is a submatrix of $\boldsymbol{S}_1^{(t)}$ with columns corresponding to the indices ∇. Then the DFP Bayesian spike and lasso transition kernel is geometrically ergodic.*

Proof If $T_t((\boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2, \lambda^2, \theta, \boldsymbol{\gamma}), (\boldsymbol{\beta}', (\boldsymbol{\tau}^2)', (\sigma^2)', (\lambda^2)', (\theta)', (\boldsymbol{\gamma})'))$ is the transition kernel of the DFP and $\pi_t(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2, \lambda^2, \theta, \boldsymbol{\gamma})$ is the stationary distribution of the transition kernel, then

$T_t(\cdot,\cdot)$ and $\pi_t(\cdot)$ for the Bayesian spike and lasso model are given by

$$
\begin{aligned}
&T_t((\boldsymbol{\beta},\boldsymbol{\tau}^2,\sigma^2,\lambda^2,\theta,\boldsymbol{\gamma}),((\boldsymbol{\beta})',(\boldsymbol{\tau}^2)',(\sigma^2)',(\lambda^2)',(\theta)',(\boldsymbol{\gamma})'))\\
&=\Big\{f_t((\boldsymbol{\beta}_1)'|(\boldsymbol{\tau}_1^2),\hat{\boldsymbol{\beta}}_{-1}^{(t-1)},\hat{\boldsymbol{\tau}}_{-1}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)})f_t((\boldsymbol{\tau}_1^2)'|(\boldsymbol{\beta}_1)',\hat{\boldsymbol{\beta}}_{-1}^{(t-1)},\hat{\boldsymbol{\tau}}_{-1}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)})\Big\}\\
&\prod_l\prod_{j\in G_l^t}\Big\{f_t((\beta_j)'|(\tau_j^2),\hat{\boldsymbol{\beta}}_{-l}^{(t-1)},\hat{\boldsymbol{\tau}}_{-l}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)})f_t((\tau_j^2)'|(\beta_j)',\hat{\beta}_{-l}^{(t-1)},\hat{\tau}_{-l}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)})\Big\}\\
&\qquad f_t((\sigma^2)'|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)})f_t((\lambda^2)'|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)})f_t((\theta)'|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)})\prod_{j=1}^{p}f_t((\boldsymbol{\gamma})'|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)})
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
&\pi_t(\boldsymbol{\beta},\boldsymbol{\tau}^2,\sigma^2,\lambda^2,\theta,\boldsymbol{\gamma})\\
&=\prod_{l=1}^{k}\Big\{f_t(\boldsymbol{\beta}_l,\boldsymbol{\tau}_l^2|\hat{\boldsymbol{\beta}}_{-l}^{(t-1)},\hat{\boldsymbol{\tau}}_{-l}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)})\Big\}\,f_t(\sigma^2|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)})f_t(\lambda^2|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)})\\
&f_t(\theta|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)})f_t(\boldsymbol{\gamma}|\hat{\boldsymbol{\beta}}^{(t-1)},\hat{\boldsymbol{\tau}}^{2(t-1)}).
\end{aligned}
\tag{9}
$$

where $\boldsymbol{\beta}_1=\{\beta_j:\beta_j\in\boldsymbol{\Theta}_{1,t}\},\boldsymbol{\tau}_1^2=\{\tau_j^2:\tau_j^2\in\boldsymbol{\Theta}_{1,t}\}$. Hence, $||T_t-\pi_t||_{TV}=||\tilde{T}_{t,1}-\tilde{\pi}_{t,1}||_{TV}$, where

$$
\begin{aligned}
\tilde{T}_{t,1}&=\Big\{f_t((\boldsymbol{\beta}_1)'|(\boldsymbol{\tau}^2),\hat{\boldsymbol{\beta}}_{-1}^{(t-1)},\hat{\boldsymbol{\tau}}_{-1}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)},\hat{\theta}^{(t-1)},\hat{\boldsymbol{\gamma}}^{(t-1)})\\
&\qquad f_t((\boldsymbol{\tau}_1^2)'|(\boldsymbol{\beta}_1)',\hat{\boldsymbol{\beta}}_{-1}^{(t-1)},\hat{\boldsymbol{\tau}}_{-1}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)},\hat{\theta}^{(t-1)},\hat{\boldsymbol{\gamma}}^{(t-1)})\Big\}\\
\tilde{\pi}_{t,1}&=\Big\{f_t(\boldsymbol{\beta}_1,\boldsymbol{\tau}_1^2|\hat{\boldsymbol{\beta}}_{-1}^{(t-1)},\hat{\boldsymbol{\tau}}_{-1}^{2(t-1)},\hat{\sigma}^{2(t-1)},\hat{\lambda}^{2(t-1)},\hat{\theta}^{(t-1)},\hat{\boldsymbol{\gamma}}^{(t-1)})\Big\}.
\end{aligned}
$$

Thus it is enough to show the geometric ergodicity of the chain by establishing a geometric drift condition and a geometric minorization condition for the $(\boldsymbol{\beta}_1,\boldsymbol{\tau}_1^2)$ chain. Define, $\tilde{V}_t(\boldsymbol{\beta}_1,\boldsymbol{\tau}_1^2)=\frac{\boldsymbol{\beta}_1'\boldsymbol{\beta}_1}{\frac{1}{m_0}\boldsymbol{H}_1'\boldsymbol{H}_1+1}+\mathbf{1}'\boldsymbol{\tau}_1^2\mathbf{1}$, where $\boldsymbol{H}_l=\boldsymbol{S}_{2,l}^{(t)}-\boldsymbol{S}_{1,l,-l}^{(t)}\hat{\boldsymbol{\beta}}_{-l}^{(t-1)}$. Using similar calculations as in Theorem 0.4, the proof of minorization and geometric drift conditions follow.

It remains to show (iv) in Corollary 0.3. The lemma presented below develops sufficient conditions to derive (iv). The lemma is presented for a general likelihood function $p_{\boldsymbol{\Theta}}(\cdot)$.

Lemma 0.6 *Assume that the likelihood function $p_{\boldsymbol{\Theta}}(\cdot)$ is continuous as a function of $\boldsymbol{\Theta}$ at $\boldsymbol{\Theta}^0=(\boldsymbol{\Theta}_1^0,..,\boldsymbol{\Theta}_k^0)$ and $\sqrt{t}p_{\boldsymbol{\Theta}^0}(\boldsymbol{D}^{(t)})$ in limit is bounded away from 0 and ∞. Suppose $\boldsymbol{\Theta}^0$*

is an interior point in the domain and prior distribution $\pi_0(\boldsymbol{\Theta})$ *is positive and continuous at* $\boldsymbol{\Theta}^0$. *Further, assume* $\widehat{\boldsymbol{\Theta}}^{(t)} \to \boldsymbol{\Theta}^0$ *a.s. under the data generating law at* $\boldsymbol{\Theta}^0$, *and* f_t *and* π_t *both converge to* $\boldsymbol{\Theta}^0$ *at a rate* Δ_t. *Then* $\exists\ \kappa_t$ *depending on* Δ_t, *s.t.* $\kappa_t \to 0$ *and* $||f_t - \pi_t||_{TV} = 2\int \left|\pi_t(\boldsymbol{\Theta}) - f_t(\boldsymbol{\Theta})\right| d\boldsymbol{\Theta} \leq 2\kappa_t$ *for large* t, *a.s. under the data generating model at* $\boldsymbol{\Theta}^0$.

proof of lemma 0.6

Stationary distribution π_t of the C-DF transition kernel T_t is the approximate posterior distribution to π_t obtained at time t, and by Lemma 0.1 is given by

$$\pi_t(\boldsymbol{\Theta}_1, ..., \boldsymbol{\Theta}_k) = \prod_{s=1}^{k} f_t(\boldsymbol{\Theta}_s | \widehat{\boldsymbol{\Theta}}_{-s}^{(t)}) = \frac{\prod_{s=1}^{k}\prod_{l=1}^{t}\left\{p_{\boldsymbol{\Theta}_s, \widehat{\boldsymbol{\Theta}}_{-s}^{(t)}}(\boldsymbol{D}_l)\pi_0(\widehat{\boldsymbol{\Theta}}_{-s}^{(t)}, \boldsymbol{\Theta}_s)\right\}}{\int \prod_{s=1}^{k}\prod_{l=1}^{t}\left\{p_{\boldsymbol{\Theta}_s, \widehat{\boldsymbol{\Theta}}_{-s}^{(t)}}(\boldsymbol{D}_l)\pi_0(\widehat{\boldsymbol{\Theta}}_{-s}^{(t)}, \boldsymbol{\Theta}_s)\right\}}.$$

By assumption, $\widehat{\boldsymbol{\Theta}}_s^{(t)} \to \boldsymbol{\Theta}_s^0$ a.s. under $\boldsymbol{\Theta}^0$, there exists Ω_0 which has probability 1 under the data generating law s.t. for all $\omega \in \Omega_0$, $\widehat{\boldsymbol{\Theta}}_s^{(t)}(\omega)$ is in an arbitrarily small neighborhood of $\boldsymbol{\Theta}_s^0$, $s = 1, ..., k$. Also by assumption, prior π_0 is continuous at $\boldsymbol{\Theta}^0$. That is, given $\epsilon_t > 0$ and $\eta_{1,t}, \eta_{2,t} > 0$, there exists a neighborhood $N_{\epsilon_t,\eta_{1,t},\eta_{2,t}} = \{\boldsymbol{\Theta} : ||\boldsymbol{\Theta} - \boldsymbol{\Theta}^0|| \leq M\Delta_t\}$ s.t. for all $\boldsymbol{\Theta} \in N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}$ one has $|\pi_0(\boldsymbol{\Theta}_1, ..., \boldsymbol{\Theta}_k) - \pi_0(\boldsymbol{\Theta}_1^0, ..., \boldsymbol{\Theta}_k^0)| < \epsilon_t$. Using this and the consistency of $\widehat{\boldsymbol{\Theta}}_s^{(t)}$, $s = 1, ..., k$ as above, one obtains for all $t > t_0$ and $\omega \in \Omega_0$

$$|\pi_0(\boldsymbol{\Theta}_s, \widehat{\boldsymbol{\Theta}}_{-s}^{(t)}) - \pi_0(\boldsymbol{\Theta}^0)| < \epsilon_t, \tag{10}$$

Similarly, continuity of $p_{\boldsymbol{\Theta}}(\cdot)$ at $\boldsymbol{\Theta}^0$ leads to the condition that for all $t > t_0$,

$$|p_{\boldsymbol{\Theta}_1,...,\boldsymbol{\Theta}_k}(\boldsymbol{D}_t) - p_{\boldsymbol{\Theta}_1^0,...,\boldsymbol{\Theta}_k^0}(\boldsymbol{D}_t)| < \epsilon_t. \tag{11}$$

Further, convergence assumptions on f_t and π_t yield that for all $t > t_1$ and $\omega \in \Omega_1$ $f_t(N_{\epsilon_t,\eta_{1,t},\eta_{2,t}} | \boldsymbol{D}^{(t)}(\omega)) > 1 - \eta_{1,t}, \pi_t(N_{\epsilon_t,\eta_{1,t},\eta_{2,t}} | \boldsymbol{D}^{(t)}(\omega)) > 1 - \eta_{2,t}$, where Ω_1 has probability 1 under the data generating law. Considering $\Omega = \Omega_0 \cap \Omega_1$ and $t_2 = \max\{t_1, t_0\}$ it is evident that Ω has probability 1 under the true data generating law and all of the above conditions hold for $t > t_2$ and $\omega \in \Omega$. Simple algebraic manipulations yield $\frac{\pi_t(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}(\omega))}{f_t(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}(\omega))} =$

$\frac{\pi_t(N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}|\boldsymbol{D}^{(t)}(\omega))}{f_t(N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}|\boldsymbol{D}^{(t)}(\omega))}\frac{\int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)\pi_0(\boldsymbol{\Theta})}{\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)\pi_0(\boldsymbol{\Theta})}\frac{\left[\prod_{l=1}^t\prod_{s=1}^k p_{\boldsymbol{\Theta}_s,\widehat{\boldsymbol{\Theta}}^{(t)}_{-s}}(\boldsymbol{D}_l)\pi_0(\widehat{\boldsymbol{\Theta}}^{(t)}_{-s},\boldsymbol{\Theta}_s)\right]}{\int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\left[\prod_{l=1}^t\prod_{s=1}^k p_{\boldsymbol{\Theta}_s,\widehat{\boldsymbol{\Theta}}^{(t)}_{-s}}(\boldsymbol{D}_l)\pi_0(\widehat{\boldsymbol{\Theta}}^{(t)}_{-s},\boldsymbol{\Theta}_s)\right]}$ Using (10)
we have $(\pi_0(\boldsymbol{\Theta}^0)-\epsilon)\int\limits_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l) \leq \int\limits_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\left[\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)\right]\pi_0(\boldsymbol{\Theta})$
$\leq (\pi_0(\boldsymbol{\Theta}^0)+\epsilon)\int\limits_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)$. Thus,

$$\frac{\pi_t(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}(\omega))}{f_t(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}(\omega))} \leq \frac{(1-\eta_{1,t})^{-1}\prod_{l=1}^t\prod_{s=1}^k p_{\boldsymbol{\Theta}_s,\widehat{\boldsymbol{\Theta}}^{(t)}_{-s}}(\boldsymbol{D}_l)}{\int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\prod_{l=1}^t\prod_{s=1}^k p_{\boldsymbol{\Theta}_s,\widehat{\boldsymbol{\Theta}}^{(t)}_{-s}}(\boldsymbol{D}_l)}\frac{\int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)}{\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)}\frac{(\pi_0(\boldsymbol{\Theta}^0)+\epsilon)^3}{(\pi_0(\boldsymbol{\Theta}^0)-\epsilon)^3}.$$

Using similar calculations we have

$$\frac{\pi_t(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}(\omega))}{f_t(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}(\omega))} \geq \frac{(1-\eta_{2,t})\prod_{l=1}^t\prod_{s=1}^k p_{\boldsymbol{\Theta}_s,\widehat{\boldsymbol{\Theta}}^{(t)}_{-s}}(\boldsymbol{D}_l)}{\int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\prod_{l=1}^t\prod_{s=1}^k p_{\boldsymbol{\Theta}_s,\hat{\boldsymbol{\Theta}}^{(t)}_{-s}}(\boldsymbol{D}_l)}\frac{\int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)}{\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)}\frac{(\pi_0(\boldsymbol{\Theta}^0)-\epsilon)^3}{(\pi_0(\boldsymbol{\Theta}^0)+\epsilon)^3}.$$

Condition (11) now gives us

$$\frac{\prod_{l=1}^t(p_{\boldsymbol{\Theta}^0}(\boldsymbol{D}_l)-\epsilon)^3}{\prod_{l=1}^t(p_{\boldsymbol{\Theta}^0}(\boldsymbol{D}_l)+\epsilon)^3} \leq \frac{\prod_{l=1}^t\prod_{s=1}^k p_{\boldsymbol{\Theta}_s,\hat{\boldsymbol{\Theta}}^t_{-s}}(\boldsymbol{D}_l)}{\int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\prod_{l=1}^t\prod_{s=1}^k p_{\boldsymbol{\Theta}_s,\hat{\boldsymbol{\Theta}}^{(t)}_{-s}}(\boldsymbol{D}_l)}\frac{\int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)}{\prod_{l=1}^t p_{\boldsymbol{\Theta}}(\boldsymbol{D}_l)}$$
$$\leq \frac{\prod_{l=1}^t(p_{\boldsymbol{\Theta}^0}(\boldsymbol{D}_l)+\epsilon)^3}{\prod_{l=1}^t(p_{\boldsymbol{\Theta}^0}(\boldsymbol{D}_l)-\epsilon)^3}.$$

Using the condition that $\lim_{t\to\infty}\sqrt{t}p_{\boldsymbol{\Theta}^0}(\boldsymbol{D}^{(t)})$ is bounded away from 0 and ∞ and choosing ϵ,η sufficiently small, we have $\left|\frac{\pi_t(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}(\omega))}{f_t(\boldsymbol{\Theta}|\boldsymbol{D}^{(t)}(\omega))}-1\right| < \upsilon_t \quad$ for all $t > t_2$ and $\omega \in \Omega$. Finally,

$$\int|\pi_t(\boldsymbol{\Theta})-f_t(\boldsymbol{\Theta})| \leq \int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}|\pi_t(\boldsymbol{\Theta})-f_t(\boldsymbol{\Theta})| + \int_{N^c_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}|\pi_t(\boldsymbol{\Theta})-f_t(\boldsymbol{\Theta})|$$
$$\leq \int_{N_{\epsilon_t,\eta_{1,t},\eta_{2,t}}}|\pi_t(\boldsymbol{\Theta})-f_t(\boldsymbol{\Theta})| + \eta_{1,t}+\eta_{2,t} \leq f_t(N_{\epsilon_t,\eta_{1,t},\eta_{2,t}})\upsilon_t + \eta_{1,t}+\eta_{2,t} < \upsilon_t+\eta_{1,t}+\eta_{2,t} = \kappa_t.$$

Remark: Lemma outlines sufficient conditions for the DFP stationary distribution to be close to the full posterior distribution as t increases. One of the important sufficient conditions presented is the consistency of the sequence of estimators $\widehat{\boldsymbol{\Theta}}^{(t)}$.

References

Pal, S. and K. Khare (2014). Geometric Ergodicity for Bayesian Shrinkage Models. *Electronic Journal of Statistics 8*(1), 604–645.

Yang, Y. and D. B. Dunson (2013). Sequential Markov Chain Monte Carlo. *arXiv preprint arXiv:1308.3861*.