Bayesian Inference in Common Microeconometric Models with Massive Datasets by Double Marginalized Subsampling

Hang Qian[*]

May 12, 2021

Abstract

Bayesian inference with a large dataset is computationally intensive, as Markov chain Monte Carlo simulation requires a complete scan of the dataset for each proposed parameter update. To reduce the number of data points evaluated at each iteration of posterior simulation, we develop a double marginalized subsampling method, which is applicable to a wide array of microeconometric models including Tobit, Probit, regressions with non-Gaussian errors, heteroscedasticity and stochastic volatility, hierarchical longitudinal models, time-varying-parameter regressions, Gaussian mixtures, etc. We also provide an extension to double pseudo-marginalized subsampling, which has more applications beyond conditionally conjugate models. With rank-one update of the cumulative statistics, both methods target the exact posterior distribution, from which a parameter draw can be obtained with every single observation. Simulation studies demonstrate the statistical and computational efficiency of the marginalized sampler. The methods are also applied to a real-world massive dataset on the incidentally truncated mortgage rates.

Keywords: Big data, Scalable computation, Latent variable, Gibbs sampler, Metropolis sampler

[*] 55 Centre Street, Natick, MA, U.S., 01760. Email: matlabist@gmail.com. The author would like to thank the editor, AE and reviewers for the comments and insights that greatly improve the manuscript.

1 Introduction

Markov chain Monte Carlo (MCMC) methods (Gelfand and Smith, 1990), such as the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampler with data augmentation (Tanner and Wong, 1987), are widely used in Bayesian microeconometric models. When the sample size is large, it is computationally expensive to implement the conventional MCMC methods, which require sweeping over the whole dataset at each iteration of posterior simulation.

Bayesian inference with massive data is an active field of research in several directions. First, the split-and-merge methods that partition a large dataset into multiple subsets and combine the subset estimation results. See Scott et al. (2016), Neiswanger et al. (2014), Qian (2018) and Nemeth and Sherlock (2018). Second, subsampling methods that reduce the number of data points for evaluating the likelihood function and/or gradients, with a small subset of the data randomly selected at each MCMC iteration. Various subsampling schemes have been proposed in the literature: firefly Monte Carlo by Maclaurin and Adams (2014), adaptive subsampling with MH tests by Korattikara et al. (2014), stochastic gradient Langevin dynamics by Teh et al. (2016), zig-zag process with subsampling by Bierkens et al. (2019), bias-corrected subsampling with control variates by Quiroz et al. (2019), MH sampling with delayed acceptance by Banterle et al. (2019), among others. See Bardenet et al. (2017) for an excellent review. Third, the variational Bayes methods that approximate complex densities through optimization by minimizing the Kullback-Leibler divergence. Massive data are handled by stochastic approximation of gradients for large-scale optimization. See Jordan et al. (1999) and Blei et al. (2017). Fourth, sequential Monte Carlo methods in which the approximated target distributions of increasing dimensions are constructed based on the previous approximation with importance weights. Balakrishnan and Madigan (2006) propose a one-pass sequential Monte Carlo method for Bayesian analysis of massive datasets. Also see Geweke and Durham (2019) for a recent development and Creal (2012) for a survey.

In the realm of classical inference, the stochastic gradient descent (SGD) traced back to Robbins and Monro (1951) is a well-known big data method for computing the M-estimator.

Traditionally, numerical optimization is implemented by Newton-type algorithms with the full-sample gradients. The batch algorithms can be cost prohibitive for large-scale problems, while SGD randomly samples one data point at each iteration of parameter updates.

We develop a double marginalized subsampling (DMS) method targeting the exact posterior distribution for scalable Bayesian inference, which is applicable to a variety of common microeconometric models, including Tobit, Probit, linear regressions with non-Gaussian errors, heteroscedasticity and stochastic volatility, hierarchical linear models, random coefficient regressions, Gaussian mixtures, etc. The Gibbs or MH algorithms for those models are described in popular textbooks (e.g., Koop et al., 2007). Compared to the conventional algorithms, our approach does not require a full pass of the entire dataset at each iteration. With rank-one update of the cumulative statistics, a posterior draw of the parameter vector can be obtained with every single observation, which is comparable to SGD that samples one data point at each iteration of numerical optimization of parameters. DMS is to a conventional MCMC sampler as SGD is to the full-sample gradient descent method. Also, compared to most subsampling methods that use a subset of n observations for the posterior evaluation, we subsample a latent vector of the length n for reconstructing the exact posterior density by online update of cumulative statistics.

The Gibbs sampler is convenient for models in which the posterior conditional distributions are of recognizable forms, so is DMS. However, there are situations where some parameters cannot be analytically integrated out. In the pseudo-marginal framework of Andrieu and Roberts (2009), we extend our method to double pseudo-marginalized subsampling for inference beyond conditionally conjugate models.

The remainder of the paper is organized as follows. Section 2 outlines the algorithm of DMS, of which use cases in microeconometrics are detailed in Section 3. Section 4 is devoted to simulation studies on the statistical and computational efficiency of DMS and alternative MCMC samplers. Section 5 provides extensions to pseudo-marginalization and Metropolis-within-Gibbs samplers that address non-conjugate distributions. In Section 6, DMS is applied to a real-world massive dataset for analyzing the incidentally truncated mortgage rates. Section 7 concludes the paper.

2 Double Marginalized Subsampling

Let θ be the model parameters and y be the data. In many microeconometric models, the posterior distribution $p\left(\theta\left|y\right.\right)$ is not analytically tractable. However, by data augmentation of the latent variable z, the posterior conditional distributions are of recognizable forms, which facilitate the Gibbs sampler that cycles through $p\left(\theta\left|y,z\right.\right)$ and $p\left(z\left|y,\theta\right.\right)$. Alternatively, the MH sampler that evaluates $p\left(\theta\right)p\left(y\left|\theta\right.\right)$ can be used, provided that the likelihood function $p\left(y\left|\theta\right.\right)$ is in a closed form. However, each iteration of the Gibbs or MH sampler requires at least one sweep over the whole dataset, which is computationally expensive if the sample size is large.

We propose DMS based on the marginal-conditional decomposition:

$$p\left(z,\theta\left|y\right.\right)=p\left(z\left|y\right.\right)p\left(\theta\left|y,z\right.\right). \tag{1}$$

Double marginalization refers to partition of parameters in two groups: $\theta=(\theta_1,\theta_2)$, which are marginalized separately:

$$p\left(z\left|y\right.\right)\propto p\left(z\right)p\left(y\left|z\right.\right)=\frac{p\left(\theta_1\right)p\left(z\left|\theta_1\right.\right)}{p\left(\theta_1\left|z\right.\right)}\frac{p\left(\theta_2\left|z\right.\right)p\left(y\left|z,\theta_2\right.\right)}{p\left(\theta_2\left|y,z\right.\right)}. \tag{2}$$

Equation (2) is useful for Bayesian models with the directed acyclic graph $\theta_1\rightarrow z\rightarrow y\leftarrow\theta_2$, which implies that θ_1 has no direct effect on y conditional on z, and θ_2 has an effect on y, but not on z. If the priors satisfy $p\left(\theta_1,\theta_2\right)=p\left(\theta_1\right)p\left(\theta_2\right)$, then equation (1) becomes

$$p\left(z,\theta_1,\theta_2\left|y\right.\right)=p\left(z\left|y\right.\right)p\left(\theta_1\left|z\right.\right)p\left(\theta_2\left|y,z\right.\right). \tag{3}$$

If $p\left(\theta_1\left|z\right.\right)$ and $p\left(\theta_2\left|y,z\right.\right)$ are analytically tractable, then equation (2) can be evaluated pointwise and equation (3) indicates the posterior sampling order: we first draw the latent variable from $p\left(z\left|y\right.\right)$, and then draw other parameters from $p\left(\theta_1\left|z\right.\right)$ and $p\left(\theta_2\left|y,z\right.\right)$. Table 1 summarizes the variable partition in common microeconometric models: some marginalize θ_1 or θ_2 (see Sections 3.1 to 3.3), and others marginalize both (see Sections 3.4 to 3.6).

It is well known that marginalization improves the convergence the MCMC chain and reduces sample autocorrelations (Chib and Carlin, 1999), but our main motivation is that marginalization concentrates the target distribution as $p\left(z\left|y\right.\right)$, and parameter updates

Sec	Model	y	z	θ_1	θ_2	Cumulative Statistics
3.1	Tobit	censored r.v.	continuous r.v.	β, σ^2	NA	$\sum x_i' x_i$, $\sum x_i' z_i$, $\sum z_i^2$
3.2	Probit	binary r.v.	continuous r.v.	β	NA	$\sum x_i' x_i$, $\sum x_i' z_i$, $\sum z_i^2$
3.3	Skew normal	r.v.	half-normal noise	NA	β, σ^2	$\sum x_i' x_i$, $\sum x_i' y_i$, $\sum y_i^2$, $\sum z_i^2$
3.4	Heteroscedasticity	r.v.	heteroscedasticity	α	β	$\sum e^{-z_i} x_i' x_i$, $\sum e^{-z_i} x_i' y_i$, $\sum e^{-z_i} y_i^2$ $\sum z_i$, $\sum w_i' w_i$, $\sum w_i' z_i$, $\sum z_i^2$
3.4	Stochastic volatility	r.v.	stochastic volatility	α, σ^2	β	same as above
3.5	Longitudinal	r.v.	random effect	σ_z^2	β, σ^2	$\sum x_{it}' x_{it}$, $\sum x_{it}' (y_{it} - z_i)$, $\sum (y_{it} - z_i)^2$, $\sum z_i^2$
3.5	Random coefficient	r.v.	time-varying param.	Φ, Ω_z	Ω	$\sum z_t' z_t$, $\sum z_t' z_{t-1}$, $\sum (y_t - x_t z_t)' (y_t - x_t z_t)$
3.6	Gaussian mixture	mixture data	class label	w	μ_j, Σ_j	$\sum I (z_i = j)$, $\sum y_i' I (z_i = j)$, $\sum y_i y_i' I (z_i = j)$
5.2	t-distributed noise	r.v.	scale of variance	v	β, σ^2	$\sum z_i^{-1} x_i' x_i$, $\sum z_i^{-1} x_i' y_i$, $\sum z_i^{-1} y_i^2$, $\sum \ln z_i$, $\sum z_i^{-1}$

Table 1: Variable partition and cumulative statistics in econometric models. Variable symbols (β, σ^2, etc.) are defined in the text of Section 3 and 5. For brevity, r.v. = response variable.

do not interfere with subsampling $z = (z_1, \ldots, z_n)$. A subsampling method reduces the number of data points evaluated in MCMC simulations. A single-move sampler updates one element at a time, but it is not necessarily a subsampling algorithm. If a complete scan of the whole dataset was required for updating the i^{th} element z_i, it would defeat the purpose of subsampling. It reads a single observation y_i for drawing z_i as well as θ, which is comparable to SGD that samples one data point at each iteration of numerical optimization of parameters.

Marginalization is closely related to subsampling in that evaluating equation (2) requires conjugate distributions (conditional on z), which usually come from exponential-family distributions with sufficient statistics whose dimension does not increase as the sample size increases. For all models listed in Table 1, equation (2) is fully determined by a set of the cumulative statistics in the form of a sum $\sum_{i=1}^{n} \ldots$, which are reminiscent of the sufficient statistics, though they depend on both latent variables and data (including covariates x_i, without conditioning on covariates explicitly denoted). For example, the model in Section 3.4 has the cumulative statistics $S_1 (z) = \sum_{i=1}^{n} e^{-z_i} x_i' x_i$, $S_2 (z) = \sum_{i=1}^{n} e^{-z_i} x_i' y_i$, among others. Subsampling yields z^* that differs from z by only one element (or a mini-batch of data points), say $z_i \neq z_i^*$ and $z_{-i} = z_{-i}^*$, where $z_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$. Instead of recalculating $S_1(z^*)$ and $S_2(z^*)$ by sweeping over the whole dataset, we only read the i^{th} data point and perform the rank-one update of the cumulative statistics:

$$S_1(z^*) = S_1(z) - e^{-z_i} x_i' x_i + e^{-z_i^*} x_i' x_i,$$

$$S_2(z^*) = S_2(z) - e^{-z_i} x_i' y_i + e^{-z_i^*} x_i' y_i.$$

We propose Algorithm 1, a scalable DMS method, following the MH sampling paradigm. The proposal transition kernel is strictly positive on the support of the target distribution, hence irreducibility. The chain is aperiodic since the acceptance probability in step 2.6 is not always equal one. Also, the detailed balance condition is satisfied, and the chain has the stationary distribution $p(z, \theta \,|y)$. See Tierney (1994) and Robert and Casella (2004) for the convergence theorems.

Step 2.7 is optional and flexible, as we may sample θ with every update of z_i, or after each pass of the data, or at an arbitrary frequency. Algorithm 1 without step 2.7 corresponds to the MH sampler that targets $p(z\,|y)$.

The idea of DMS is sampling from $p(z\,|y)$ by marginalizing over θ and rank-one update of cumulative statistics, and Algorithm 1 implements DMS by the random-walk MH sampler. In some models where $p(z_i\,|y, z_{-i})$ is of a recognizable form (e.g., the categorical distribution in the Gaussian mixture model, and the truncated normal distribution in the Probit model), it is possible to update z_i directly.

Implementation of Algorithm 1 involves a tuning parameter λ, which is a positive scalar. Our experience is that a smaller λ leads to a larger acceptance rate in step 2.6, and a reasonable value for adequate mixing of the MCMC chain can be found by trial and error.

Algorithm 1 can be generalized to mini-batch processing, in which the proposed value z^* randomly changes multiple elements of z simultaneously. Compared to SGD that one update relies only on a mini-batch of data points, the acceptance/reject MH step depends on the whole dataset (so as to target the exact posterior distribution). However, the full-sample dependency does not imply a complete scan of the dataset, because the full-sample information has been incorporated in the cumulative statistics. At each iteration, we only read a mini-batch of observations associated with the changed elements of z for the low-rank update of the cumulative statistics.

Algorithm 1:

```
Input  : data $y$, an initial value of the latent variable $z$, and a tuning parameter $\lambda$.
Output: MCMC draws from $p(z, \theta \,|y)$.
1. compute cumulative statistics $S_j(z), j = 1, \ldots, J$, by one sweep over the data.
2. iterate the following steps:
    2.1 randomly draw an index $i \in \{1, \ldots, n\}$.
    2.2 propose a new value $z^*$ such that $z^*_{-i} = z_{-i}$, $z^*_i = z_i + \lambda u_i$, $u_i \sim N(0,1)$.
    2.3 read the $i^{th}$ data point only.
    2.4 compute cumulative statistics $S_j(z^*)$ by rank-one update of $S_j(z)$.
    2.5 evaluate $p(z^* \,|y)$ by equation (2) with cumulative statistics.
    2.6 accept the proposal (set $z = z^*$) with the probability $\min\left(1, \frac{p(z^*|y)}{p(z|y)}\right)$.
    2.7 sample $\theta$ from $p(\theta \,|y, z)$ with cumulative statistics (without original data).
```

3 Use Cases

In this section, we revisit the common microeconometric models and compare DMS (Algorithm 1) to the conventional MCMC methods routinely used in those models.

3.1 Tobit Model

Consider a Tobit model (Tobin, 1958), in which the observations are lower bounded by a known censoring point c.

$$z_i = x_i\beta + \sigma\varepsilon_i,$$

$$y_i = \max(c, z_i),$$

where $\{\varepsilon_i\}$ are independent and identically distributed (IID) standard normal disturbances. We specify the d-dimensional Normal-Inverse-Gamma (NIG) conjugate prior $(\beta, \sigma^2) \sim NIG(\mu, \Lambda, a, b)$ such that

$$p\left(\beta, \sigma^2\right) \propto \left(\sigma^2\right)^{-\left(a+\frac{d}{2}+1\right)} e^{-\sigma^{-2}\left[b+\frac{1}{2}(\beta-\mu)'\Lambda(\beta-\mu)\right]}.$$

For notational convenience, we parameterize the NIG distribution by the precision matrix Λ. That is, $NIG(\mu, \Lambda, a, b)$ indicates that $p(\beta \,|\sigma^2)$ is the multivariate normal density with the mean μ and the precision $\sigma^{-2}\Lambda$. The covariance matrix is $\sigma^2\Lambda^{-1}$.

The Gibbs sampler of Chib (1992) proceeds with the sampling of the full conditional distributions $p\left(\beta,\sigma^2\left|y,z\right.\right)$ and $p\left(z\left|y,\beta,\sigma^2\right.\right)$, while DMS resorts to the decomposition: $p\left(z,\beta,\sigma^2\left|y\right.\right)=p\left(z\left|y\right.\right)p\left(\beta,\sigma^2\left|y,z\right.\right)$.

On the one hand, $p\left(\beta,\sigma^2\left|y,z\right.\right)$ follows $NIG\left(\overline{\mu},\overline{\Lambda},\overline{a},\overline{b}\right)$, where

$$\overline{\mu}=\left(\Lambda+\sum_{i=1}^{n}x_i'x_i\right)^{-1}\left(\Lambda\mu+\sum_{i=1}^{n}x_i'z_i\right), \tag{4}$$

$$\overline{\Lambda}=\Lambda+\sum_{i=1}^{n}x_i'x_i, \tag{5}$$

$$\overline{a}=a+\frac{n}{2}, \tag{6}$$

$$\overline{b}=b+\frac{1}{2}\sum_{i=1}^{n}z_i^2+\frac{1}{2}\mu'\Lambda\mu-\frac{1}{2}\overline{\mu}'\overline{\Lambda}\overline{\mu}. \tag{7}$$

On the other hand, $p\left(z\left|y\right.\right)\propto p\left(z\right)p\left(y\left|z\right.\right)$ can be evaluated pointwise:

$$p\left(z\right)\propto\left(\overline{b}\right)^{-\overline{a}},$$

$$p\left(y\left|z\right.\right)=\prod_{i=1}^{n}\left[I\left(y_i=c,z_i\leq c\right)+I\left(y_i>c,z_i=y_i\right)\right],$$

where $I\left(\cdot\right)$ denotes an indicator function that equals one if $\left(\cdot\right)$ is true, and zero otherwise.

To implement Algorithm 1, we construct a set of cumulative statistics: $S_1\left(z\right)=\sum_{i=1}^{n}x_i'x_i$, $S_2\left(z\right)=\sum_{i=1}^{n}x_i'z_i$, $S_3\left(z\right)=\sum_{i=1}^{n}z_i^2$, which are updated by (step 2.4 of Algorithm 1)

$S_1\left(z^*\right)=S_1\left(z\right)$,

$S_2\left(z^*\right)=S_2\left(z\right)-x_i'z_i+x_i'z_i^*$,

$S_3\left(z^*\right)=S_3\left(z\right)-z_i^2+z_i^{*2}$.

We can streamline the sampling process described in Algorithm 1. If $y_i>c$, then we skip step 2.2 - 2.6, because the proposal will be rejected at step 2.6. Otherwise, we proceed to step 2.2 and propose z_i^*. We skip step 2.3 - 2.6 if $z_i^*>c$, as z_i^* will be rejected in that case. We always keep $p\left(y\left|z\right.\right)=1$ throughout the MCMC simulation.

A difference between the conventional Gibbs sampler and DMS is the dependency on the parameters. Whenever the parameters β and σ^2 are updated by the Gibbs sampler, the posterior conditional distribution $p\left(z_i \left| y, \beta, \sigma^2 \right.\right)$ changes for all z_i, $i = 1, \ldots, n$. In contrast, the scalable DMS proceeds without dependency on the parameters. That is, parameter updates at an arbitrary frequency do not change the target distribution $p\left(z \left| y \right.\right)$, and thus do not interfere with subsampling z.

3.2 Probit Model

The latent variable representation of the Probit model is

$$z_i = x_i \beta + \varepsilon_i,$$

$$y_i = I\left(z_i > 0\right),$$

where $\{\varepsilon_i\}$ are IID standard normal disturbances. We specify a normal prior $\beta \sim N\left(\mu, \Lambda\right)$ such that $p\left(\beta\right) \propto e^{-\frac{1}{2}\left(\beta-\mu\right)' \Lambda \left(\beta-\mu\right)}$. Then we have $\beta \left| y, z \right. \sim N\left(\overline{\mu}, \overline{\Lambda}\right)$, where $\overline{\mu}, \overline{\Lambda}$ are given by equations (4) and (5).

The Gibbs sampler of Albert and Chib (1993) cycles through the full posterior conditionals $p\left(\beta \left| y, z \right.\right)$ and $p\left(z \left| y, \beta \right.\right)$, while DMS evaluates $p\left(z\right) p\left(y \left| z \right.\right)$, where

$$p\left(z\right) \propto e^{-\frac{1}{2}\sum_{i=1}^{n} z_i^2 + \frac{1}{2}\overline{\mu}' \overline{\Lambda} \overline{\mu}},$$

$$p\left(y \left| z \right.\right) = \prod_{i=1}^{n} \left[I\left(y_i = 1, z_i > 0\right) + I\left(y_i = 0, z_i \leq 0\right)\right],$$

which are determined by the same set of cumulative statistics as those in the Tobit model.

Algorithm 1 provides a general-purpose implementation of DMS, as it only requires pointwise evaluation of $p\left(z \left| y \right.\right)$, which is not necessarily a common distribution. Holmes and Held (2006) note that it is the n-dimensional truncated normal density in the Probit model, and therefore step 2.2 - 2.6 of Algorithm 1 can be replaced by a truncated normal sampler for $p\left(z_i \left| y, z_{-i} \right.\right)$.

3.3 Skew Normal Regression

When the regression error departs from normality, a parsimonious model for asymmetry is the skew normal distribution (Azzalini and Dalla Valle, 1996):

$$y_i = \tilde{x}_i \tilde{\beta} + \delta z_i + \sigma \varepsilon_i,$$

where $\{\varepsilon_i\}$ are IID standard normal disturbances and z_i is a half-normal latent variable: $p(z_i) \propto e^{-\frac{1}{2}z_i^2} I(z_i > 0)$. Let $x_i = (\tilde{x}_i, z_i)$, $\beta = \left(\tilde{\beta}', \delta\right)'$. The model specification is completed by adding the prior $(\beta, \sigma^2) \sim NIG(\mu, \Lambda, a, b)$.

The Gibbs sampler of Koop et al. (2007, p.276) cycles through the full posterior conditionals $p(\beta, \sigma^2 \,|y, z\,)$ and $p(z\,|y, \beta, \sigma^2\,)$, while DMS evaluates $p(z)\, p(y\,|z\,)$. Unlike Tobit and Probit models where parameters are marginalized in $p(z)$, parameters in this model are marginalized in $p(y\,|z\,)$. That is,

$$p(z) \propto e^{-\frac{1}{2}\sum_{i=1}^n z_i^2} \prod_{i=1}^n I(z_i > 0),$$

$$p(y\,|z\,) \propto \left(\overline{b}\right)^{-\overline{a}} \left|\overline{\Lambda}\right|^{-\frac{1}{2}},$$

where $\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b}$ are defined in equations (4) - (7), with z_i replaced by y_i (since the response variable is y_i). They are fully determined by the cumulative statistics: $S_1(z) = \sum_{i=1}^n x_i' x_i$, $S_2(z) = \sum_{i=1}^n x_i' y_i$, $S_3(z) = \sum_{i=1}^n y_i^2$, $S_4(z) = \sum_{i=1}^n z_i^2$. Although the functional form of S_1 is the same as that in the Tobit model, the difference is that x_i depends on z_i, hence an extra term $\left|\overline{\Lambda}\right|^{-\frac{1}{2}}$ in evaluating $p(y\,|z\,)$.

3.4 Heteroscedasticity and Stochastic Volatility

Li and Tobias (2009) consider a linear regression with a multiplicative, parametric form of heteroscedasticity

$$y_i = x_i\beta + e^{\frac{1}{2}w_i\alpha}\varepsilon_i,$$

where $\{\varepsilon_i\}$ are IID standard normal disturbances. We specify the normal priors $\beta \sim N(\mu_\beta, \Lambda_\beta)$ and $\alpha \sim N(\mu_\alpha, \Lambda_\alpha)$.

Li and Tobias (2009) propose a Metropolis-within-Gibbs sampler that iteratively draws from $p(\beta\,|y, \alpha\,)$ and $p(\alpha\,|y, \beta\,)$. The former follows the conjugate normal distribution, while the latter is not of a known form and a Metropolis component is added.

We address the problem by randomizing the heteroscedasticity. The original model is approximated by

$$y_i = x_i\beta + e^{\frac{1}{2}z_i}\varepsilon_i,$$

$$z_i = w_i \alpha + \sigma u_i,$$

where $\{u_i\}$ are IID standard normal disturbances and σ is treated as a tuning parameter. As $\sigma \to 0$, $e^{\frac{1}{2}z_i}$ converges to the original specification of heteroscedasticity.

DMS evaluates $p(z)\, p(y\,|z)$ such that

$$p(z) = \frac{p(\alpha)\, p(z\,|\alpha)}{p(\alpha\,|z)},$$

$$p(y\,|z) = \frac{p(\beta\,|z)\, p(y\,|z,\beta)}{p(\beta\,|y,z)}.$$

On the one hand, $p(\alpha)$ and $p(\alpha\,|z)$ are densities of $N(\mu_\alpha, \Lambda_\alpha)$ and $N\left(\bar{\mu}_\alpha, \bar{\Lambda}_\alpha\right)$, where $\bar{\mu}_\alpha = \bar{\Lambda}_\alpha^{-1}\left(\Lambda_\alpha \mu_\alpha + \sigma^{-2}\sum_{i=1}^n w_i' z_i\right)$, $\bar{\Lambda}_\alpha = \Lambda_\alpha + \sigma^{-2}\sum_{i=1}^n w_i' w_i$. It follows that

$$p(z) \propto e^{-\frac{1}{2}\sigma^{-2}\sum_{i=1}^n z_i^2 + \frac{1}{2}\bar{\mu}_\alpha' \bar{\Lambda}_\alpha \bar{\mu}_\alpha}. \tag{8}$$

On the other hand, $p(\beta\,|z)$ and $p(\beta\,|y,z)$ are densities of $N(\mu_\beta, \Lambda_\beta)$ and $N\left(\bar{\mu}_\beta, \bar{\Lambda}_\beta\right)$, where $\bar{\mu}_\beta = \bar{\Lambda}_\beta^{-1}\left(\Lambda_\beta \mu_\beta + \sum_{i=1}^n e^{-z_i} x_i' y_i\right)$, $\bar{\Lambda}_\beta = \Lambda_\beta + \sum_{i=1}^n e^{-z_i} x_i' x_i$. It follows that

$$p(y\,|z) \propto \left|\bar{\Lambda}_\beta\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\sum_{i=1}^n z_i - \frac{1}{2}\sum_{i=1}^n e^{-z_i} y_i^2 + \frac{1}{2}\bar{\mu}_\beta' \bar{\Lambda}_\beta \bar{\mu}_\beta}. \tag{9}$$

Algorithm 1 can be implemented by rank-one update of the cumulative statistics: $S_1 = \sum_{i=1}^n e^{-z_i} x_i' x_i$, $S_2 = \sum_{i=1}^n e^{-z_i} x_i' y_i$, $S_3 = \sum_{i=1}^n e^{-z_i} y_i^2$, $S_4 = \sum_{i=1}^n z_i$, $S_5 = \sum_{i=1}^n w_i' w_i$, $S_6 = \sum_{i=1}^n w_i' z_i$, $S_7 = \sum_{i=1}^n z_i^2$.

The randomized heteroscedasticity, or more commonly called the stochastic volatility model, is notoriously difficult to estimate, because the likelihood function is not in a closed form. DMS addresses the problem by noting that the marginal likelihood conditional on z (marginalized over other parameters) is tractable. We consider a stylized stochastic volatility model:

$$y_i = x_i \beta + e^{\frac{1}{2}z_i} \varepsilon_i,$$

$$z_i = c + \phi z_{i-1} + \sigma u_i,$$

where $\{\varepsilon_i, u_i\}$ are IID standard normal disturbances, and the subscript i denotes the time index. Assume that the initial state z_0 is known. Let $w_i = (1, z_{i-1})$, $\alpha = (c, \phi)'$. The priors are specified as $\beta \sim N(\mu_\beta, \Lambda_\beta)$ and $(\alpha, \sigma^2) \sim NIG(\mu_\alpha, \Lambda_\alpha, a, b)$.

Jacquier et al. (1994) propose a local neighbor sampler. The main step is to sample from $p\left(z_i\left|z_{i-1}, z_{i+1}, y, \alpha, \beta, \sigma^2\right.\right)$, which does not have an analytic form, so an independence MH algorithm with the Gamma proposal density is used.

DMS partials out α, β, σ^2 and evaluates $p\left(z\right)p\left(y\left|z\right.\right)$, where $p\left(y\left|z\right.\right)$ is given by equation (9), and $p\left(z\right) \propto \left(\bar{b}\right)^{-\bar{a}}\left|\bar{\Lambda}_\alpha\right|^{-\frac{1}{2}}$, where $\bar{\Lambda}_\alpha = \Lambda_\alpha + \sum_{i=1}^n w_i' w_i$ and $\bar{a}, \bar{b}$ are essentially given by equations (6) and (7).

The cumulative statistics $S_1, \ldots, S_7$ are the same as those in the heteroscedasticity model. To accommodate lagged values in the regressor $w_i = (1, z_{i-1})$, we adapt step 2.3 and 2.4 of Algorithm 1 by reading adjacent latent variables z_{i-1}, z_i, z_{i+1} to update S_5, S_6.

3.5 Hierarchical Longitudinal Model and Random Coefficients

We consider a hierarchical linear model with longitudinal data

$$y_{it} = z_i + x_{it}\beta + \sigma\varepsilon_{it}, \tag{10}$$

where $\{\varepsilon_{it}\}$ are IID standard normal disturbances, and the random effect z_i is normally distributed with the mean zero and the variance σ_z^2. The hierarchical specification is completed by adding the prior distributions $(\beta, \sigma^2) \sim NIG\left(\mu, \Lambda, a, b\right)$ and $\sigma_z^2 \sim IG\left(a_z, b_z\right)$.

The Gibbs sampler of Gelfand and Smith (1990) cycles through the full posterior conditionals, while Chib and Carlin (1999) propose a block sampler based on the marginal-conditional decomposition

$$p\left(z, \beta\left|y, \sigma^2, \sigma_z^2\right.\right) = p\left(\beta\left|y, \sigma^2, \sigma_z^2\right.\right)p\left(z\left|y, \beta, \sigma^2, \sigma_z^2\right.\right).$$

DMS resorts to a different decomposition scheme

$$p\left(z, \beta, \sigma^2, \sigma_z^2\left|y\right.\right) = p\left(z\left|y\right.\right)p\left(\beta, \sigma^2, \sigma_z^2\left|y, z\right.\right).$$

The main difference is that Chib and Carlin (1999) integrate out z, while we integrate out all parameters other than z, and evaluate

$$p\left(z\right) = \frac{p\left(\sigma_z^2\right)p\left(z\left|\sigma_z^2\right.\right)}{p\left(\sigma_z^2\left|z\right.\right)},$$

$$p\left(y\left|z\right.\right)=\frac{p\left(\beta,\sigma^{2}\left|z\right.\right)p\left(y\left|z,\beta,\sigma^{2}\right.\right)}{p\left(\beta,\sigma^{2}\left|y,z\right.\right)}.$$

On the one hand, $p\left(\sigma_z^2\right)$ and $p\left(\sigma_z^2\left|z\right.\right)$ are densities of $IG\left(a_z,b_z\right)$ and $IG\left(\bar{a}_z,\bar{b}_z\right)$, respectively, where $\bar{a}_z=a_z+\frac{n}{2}$, $\bar{b}_z=b_z+\frac{1}{2}\sum_i z_i^2$, It follows that $p\left(z\right)\propto\left(\bar{b}_z\right)^{-\bar{a}_z}$.

On the other hand, $p\left(\beta,\sigma^{2}\left|z\right.\right)$ and $p\left(\beta,\sigma^{2}\left|y,z\right.\right)$ are densities of $NIG\left(\mu,\Lambda,a,b\right)$ and $NIG\left(\overline{\mu},\overline{\Lambda},\overline{a},\overline{b}\right)$, respectively, where

$$\overline{\mu}=\overline{\Lambda}^{-1}\left[\Lambda\mu+\sum_{i,t}x_{it}'\left(y_{it}-z_i\right)\right],$$

$$\overline{\Lambda}=\Lambda+\sum_{i,t}x_{it}'x_{it},$$

$$\overline{a}=a+\frac{nT}{2},$$

$$\overline{b}=b+\frac{1}{2}\sum_{i,t}\left(y_{it}-z_i\right)^2+\frac{1}{2}\mu'\Lambda\mu-\frac{1}{2}\overline{\mu}'\overline{\Lambda}\overline{\mu}.$$

It follows that $p\left(y\left|z\right.\right)\propto\left(\overline{b}\right)^{-\overline{a}}$.

We implement Algorithm 1 with the cumulative statistics:

$S_1=\sum_{i,t}x_{it}'x_{it}$, $S_2=\sum_{i,t}x_{it}'\left(y_{it}-z_i\right)$, $S_3=\sum_{i,t}\left(y_{it}-z_i\right)^2$, $S_4=\sum_i z_i^2$.

This model is a special case of the random coefficient regression, especially a time-varying-parameter vector-autoregression (Cogley and Sargent, 2005), which can be formulated in a general form:

$$y_t=x_tz_t+\varepsilon_t,$$

$$z_t=\Phi z_{t-1}+u_t.$$

The latent vector z_t captures the time-varying coefficients and the predictors x_t may contain lagged variables like $y_{t-1},\ldots,y_{t-p}$. The disturbances ε_t and u_t are assumed to be zero-mean Gaussian variates with the covariance matrices Ω and Ω_z, respectively. If we specify an inverse-Wishart prior for Ω and a normal-inverse-Wishart (NIW) prior for Φ,Ω_z, the conditionally conjugate specification facilities double marginalization over Φ,Ω_z and Ω:

$$p\left(z\right)=\frac{p\left(\Phi,\Omega_z\right)p\left(z\left|\Phi,\Omega_z\right.\right)}{p\left(\Phi,\Omega_z\left|z\right.\right)},$$

$$p\left(y\left|z\right.\right)=\frac{p\left(\Omega\left|z\right.\right)p\left(y\left|z,\Omega\right.\right)}{p\left(\Omega\left|y,z\right.\right)}.$$

For brevity, we omit the lengthy expressions of $p\left(z\right)$ and $p\left(y\left|z\right.\right)$, but it is clear that $p\left(z\right)$ is analytically tractable because $p\left(\Phi,\Omega_z\right)$ and $p\left(\Phi,\Omega_z\left|z\right.\right)$ are the conjugate NIW densities. Similarly, $p\left(y\left|z\right.\right)$ can be evaluated by the conjugate inverse-Wishart densities of $p\left(\Omega\left|z\right.\right)$ and $p\left(\Omega\left|y,z\right.\right)$.

DMS of the random coefficient model is similar to that of the stochastic volatility model (see Section 3.4), as the simulation smoothing (i.e., posterior sampling from $p\left(z\left|y\right.\right)$) of the state space model does not rely on the Kalman filter/smoother (Carter and Kohn, 1994; Durbin and Koopman, 2002).

3.6 Gaussian Mixtures

In a recent review of the computational solutions for Bayesian mixture models, Celeux et al. (2019) comment that "there nonetheless remain major difficulties in running Bayesian inference on mixtures of moderate to large dimensions."

Consider a d-dimensional multivariate Gaussian mixture with k classes:

$$p\left(y,z\left|w,\mu,\Sigma\right.\right)=\prod_{i=1}^{n}\sum_{j=1}^{k}w_j p\left(y_i;\mu_j,\Sigma_j\right)I\left(z_i=j\right), \tag{11}$$

where $z_i\in\{1,\ldots,k\}$ indicates the latent class label of the observation y_i. For the j^{th} class, μ_j, Σ_j and w_j are the class-specific mean, covariance matrix and weight, with the priors: $w\sim Dirichlet\left(a_1,\ldots,a_k\right)$ and $\left(\mu_j,\Sigma_j\right)\sim NIW\left(U_j,\Lambda_j,\Omega_j,v_j\right)$ such that $p\left(w\right)\propto\prod_{j=1}^{k}w_j^{a_j}$,

$$p\left(\mu_j,\Sigma_j\right)\propto\left|\Sigma_j\right|^{-\frac{v_j+d+2}{2}}\exp\left\{-\frac{1}{2}tr\left[\left(\mu_j-U_j\right)'\Lambda_j\left(\mu_j-U_j\right)\Sigma_j^{-1}+\Omega_j\Sigma_j^{-1}\right]\right\}.$$

The model contains $dk+\frac{d(d+1)}{2}k+k-1$ unknown parameters. If the dimension d (i.e., the number of variables in y_i) is large, multitudinous parameters make MCMC samplers delicate to tune. Celeux et al. (2019) note that the Gibbs sampling is the only algorithm that scales to high dimensions. However, the Gibbs sampler may lead to trapping states that require an enormous number of iterations to escape from (Marin et al., 2005). In many implementations, it fails to converge (Celeux et al., 2000). In Jasra et al. (2005, p.53-54), "the Gibbs sampler is not always appropriate for sampling from a mixture posterior." See also Stephens (2000), Geweke (2007) and Yao and Lindsay (2009) for more discussions on label switching and convergence issues.

DMS addresses the curse of dimensionality by integrating out the high dimensional parameters:

$$p\left(z\left|y\right.\right)\propto p\left(z\right)p\left(y\left|z\right.\right)=\frac{p\left(w\right)p\left(z\left|w\right.\right)}{p\left(w\left|z\right.\right)}\frac{p\left(\mu,\Sigma\left|z\right.\right)p\left(y\left|z,\mu,\Sigma\right.\right)}{p\left(\mu,\Sigma\left|y,z\right.\right)}. \tag{12}$$

Equation (12) can be evaluated by a three-step procedure.

First, given the class labels z, we calculate the cumulative statistics for each class: $S_{1j}=\sum_{i=1}^{n}I\left(z_i=j\right)$, $S_{2j}=\sum_{i=1}^{n}y_i'\cdot I\left(z_i=j\right)$, $S_{3j}=\sum_{i=1}^{n}y_iy_i'\cdot I\left(z_i=j\right)$, $j=1,\ldots,k$.

Second, we write the posterior conditional distribution as a function of the cumulative statistics in the conjugate form: $p\left(w\left|z\right.\right)$ follows $Dirichlet\left(\bar{\alpha}_1,\ldots,\bar{\alpha}_k\right)$ and $p\left(\mu_j,\Sigma_j\left|y,z\right.\right)$ follows $NIW\left(\bar{U}_j,\bar{\Lambda}_j,\bar{\Omega}_j,\bar{v}_j\right)$, where

$\bar{\alpha}_j=\alpha_j+S_{1j}$,

$\bar{U}_j=\left(\Lambda_j+S_{1j}\right)^{-1}\left(\Lambda_jU_j+S_{2j}\right)$,

$\bar{\Lambda}_j=\Lambda_j+S_{1j}$,

$\bar{\Omega}_j=\Omega_j+S_{3j}+U_j'\Lambda_jU_j-\bar{U}_j'\bar{\Lambda}_j\bar{U}_j$,

$\bar{v}_j=v_j+S_{1j}$.

Third, we evaluate $p\left(z\left|y\right.\right)$ by either an arbitrary value of w,μ,Σ or simplifying equation (12) by cancelling out terms involving w,μ,Σ. The simplified expression reads

$$p\left(z\left|y\right.\right)\propto\prod_{j=1}^{k}\Gamma\left(\bar{\alpha}_j\right)\Gamma_d\left(\frac{\bar{v}_j}{2}\right)\left|\bar{\Omega}_j\right|^{-\bar{v}_j/2}\left(\bar{\Lambda}_j\right)^{-d/2},$$

where Γ and Γ_d are univariate and multivariate gamma functions, respectively.

Because the class label z_i is discrete-valued, DMS can be implemented by sampling from the categorical distribution $p\left(z_i\left|y,z_{-i}\right.\right)$, which is proportional to $p\left(z\left|y\right.\right)$, as an alternative to the random-walk MH proposal (i.e., random change of a label for z_i).

DMS of the Gaussian mixture model has three features: 1) it is scalable to mixtures of large dimensions, as it integrates out high dimensional parameters; 2) it is scalable to massive datasets, as a single observation is used for rank-one update of the cumulative statistics that fully determine the exact posterior distribution; and 3) marginalization alleviates the label switching problem and accelerates convergence of the MCMC chain.

4 Simulation Studies

The Gibbs sampler updates the whole vector z conditional on θ, while DMS updates each element of z by integrating out θ. The block sampler of Chib and Carlin (1999) updates θ by integrating out z. It is of interest of compare the convergence properties, statistical and computational efficiencies of different samplers.

4.1 Evidence from Gaussian Mixture Model

We consider the multivariate Gaussian mixture model (11) in three scenarios of the increasing sample size (n) and the number of variables (d). The number of classes is 2.

The first is a small data scenario with $n = 4$ and $d = 4$. As the sample size is small and z is discrete-valued, it is feasible to calculate the closed-form posterior distribution $p\left(z\left|y\right.\right)$ by enumerating the probabilities of 2^4 combinations of class labels, shown in the first two columns of Table 2. Given such ground truth, we compare the asymptotic and transitional behavior of the MCMC samplers.

The data are extracted from Fisher Iris data: (5.1, 3.5, 1.4, 0.2) and (4.9, 3.0, 1.4, 0.2) of setosa, as well as (6.2, 3.4, 5.4, 2.3) and (5.9, 3.0, 5.1, 1.8) of virginica. We specify an exchangeable prior: $w \sim Dirichlet\left(1,1\right)$ and $\left(\mu_j, \Sigma_j\right) \sim NIW\left(\left(5,4,3,2\right), 1, 10I_d, 10\right)$.

Table 2 shows the empirical distributions of z draws after 100, 500 and 5000 passes of data. The results of DMS and Gibbs samplers closely match the true probabilities after 5000 passes, which validate the convergence of the MCMC chains.

However, DMS converges substantially faster than the Gibbs sampler. With 100 passes of data, the empirical and theoretical distributions of $p\left(z\left|y\right.\right)$ are reasonably close for DMS, while the results of the Gibbs sampler depart remotely from the true probabilities. Not until the Gibbs sampler runs for 500 passes of data, does the empirical distribution approach the stationary distribution.

The second example is a tall data scenario with $n = 50000$ and $d = 20$. The data generating process is the model (11) with $\mu_1 = -\iota_d, \mu_2 = \iota_d, \Sigma_1 = \Sigma_2 = \frac{1}{2}I_d + \frac{1}{2}\iota_d\iota_d', w_1 = \frac{1}{3}, w_2 = \frac{2}{3}$, where ι_d is the $d \times 1$ vector of ones, and I_d is the $d \times d$ identity matrix.

We evaluate the (statistical) speed of convergence of MCMC samplers by the root

Ground Truth		5000 Passes		500 Passes		100 Passes	
z	Prob.	DMS	Gibbs	DMS	Gibbs	DMS	Gibbs
1111	0.186	0.184	0.190	0.192	0.214	0.180	0.370
1112	0.015	0.011	0.020	0.012	0.018	0.010	0.050
1121	0.022	0.023	0.023	0.020	0.024	0.020	0.020
1122	0.235	0.254	0.233	0.236	0.224	0.230	0.070
1211	0.019	0.019	0.022	0.016	0.026	0.030	0.010
1212	0.002	0.003	0.001	0.006	0.000	0.010	0.000
1221	0.002	0.001	0.002	0.000	0.000	0.000	0.000
1222	0.019	0.019	0.018	0.020	0.012	0.020	0.000
2111	0.019	0.017	0.016	0.018	0.018	0.020	0.000
2112	0.002	0.004	0.002	0.002	0.000	0.000	0.000
2121	0.002	0.002	0.002	0.002	0.002	0.000	0.000
2122	0.019	0.020	0.018	0.022	0.018	0.020	0.000
2211	0.235	0.232	0.221	0.220	0.188	0.190	0.070
2212	0.022	0.021	0.021	0.020	0.016	0.030	0.010
2221	0.015	0.015	0.017	0.018	0.022	0.030	0.030
2222	0.186	0.176	0.195	0.196	0.218	0.210	0.370

Table 2: Analytical distribution of $p\left(z\,|y\right)$ and empirical distributions produced by DMS and Gibbs samplers running for 5000, 500 and 100 passes of data, in the Gaussian mixture model. The analytical distribution enumerates $p\left(z_1 = i, z_2 = j, z_3 = k, z_4 = l\,|y\right)$, $i, j, k, l \in \{1, 2\}$.

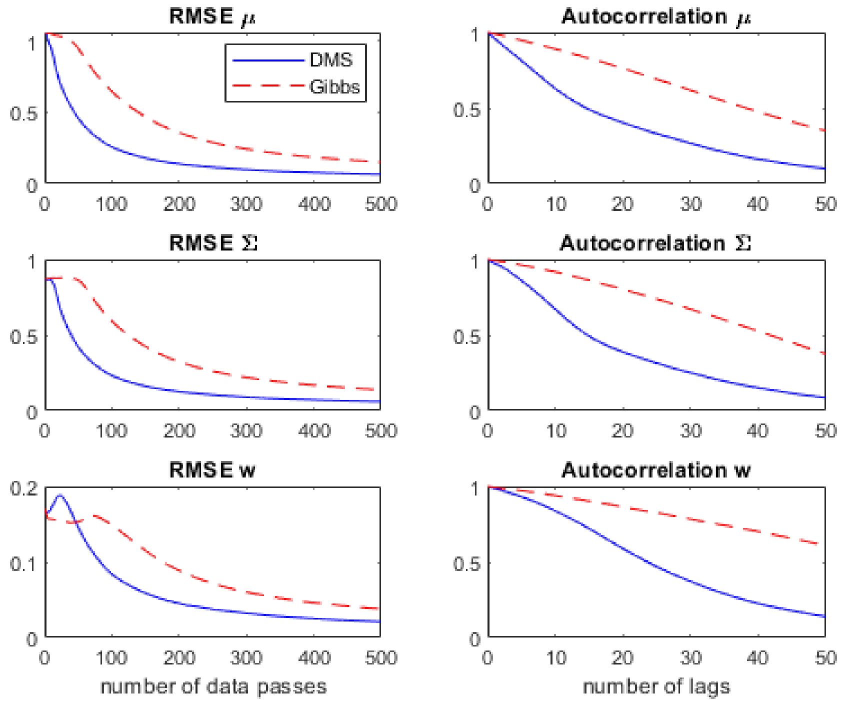

Figure 1: RMSE and autocorrelations of DMS (solid lines) and the Gibbs sampler (dashed lines) in the Gaussian mixture model. RMSE indicates the discrepancy between the estimated and true parameters, averaged across elements of μ, Σ and w.

mean squared error (RMSE) between estimated and true parameters, after each pass of data. Figure 1 plots RMSE as well as autocorrelations of the MCMC draws. The RMSE of DMS (solid line) is smaller than that of the Gibbs sampler (dashed line) in all cases, and the autocorrelations of DMS draws are lower as well.

The third example is a tall and distributed data scenario with $n = 10^7$ and $d = 50$. The artificial data are saved in 100 files separately on the hard disk. As we do not load the whole dataset into the computer memory, it is cost prohibitive to run the conventional Gibbs sampler. However, it is still feasible to implement DMS. We initialize z by random assignment of class labels, and scan the dataset once to obtain the initial cumulative statistics. Then, the MCMC sampler sequentially loads one of the data files in the memory, reads one data point y_i at a time, and generates a draw of z_i by rank-one update of the cumulative statistics. Updating parameters μ, Σ, w by cumulative statistics is an optional step and does not interfere with subsampling z. For the purpose of monitoring convergence of the MCMC chain, we generate a parameter draw every 10^4 updates of z_i (that is, 10^3 parameter draws after a pass of data).

We illustrate the performance of DMS by the in-sample fit and the out-of-sample prediction of class labels. The former is measured by RMSE between the estimated and true parameters (averaged across elements of parameter matrices), and the latter is characterized by the correct classification rate. A thousand data points are reserved for forecast evaluation. The class with the highest posterior predictive probability is considered as the most likely class. Because of random assignment of initial class labels, the correct prediction rate is about 0.5 at the beginning. As is shown in Table 3, the initial RMSE of parameter estimation is high, but drops drastically after a few passes of data. Meanwhile, the correct prediction rate increases steadily: 0.55, 0.74, 0.92, 0.99, and levels off at 0.9998 after 10 passes of data. Also, DMS achieves good in-sample fit, as the estimation error of μ, Σ, w decreases to 0.003, 0.052 and 0.001, respectively.

Passes	1	2	3	4	5	6	7	8	9	10
RMSE μ	1.053	1.049	1.009	0.787	0.409	0.124	0.012	0.002	0.003	0.003
RMSE Σ	0.396	0.396	0.395	0.366	0.222	0.057	0.049	0.052	0.052	0.052
RMSE w	0.167	0.168	0.174	0.198	0.143	0.045	0.005	0.001	0.001	0.001
Prediction	0.501	0.506	0.548	0.740	0.918	0.987	1.000	1.000	1.000	1.000

Table 3: In-sample fit by RMSE and out-of-sample prediction by correct classification rate, in the Gaussian mixture model with tall and distributed data. Results are reported after each pass of data by DMS.

4.2 Evidence from Hierarchical Longitudinal Model

Artificial data are generated by equation (10), in which $z_i, x_{it}, \varepsilon_{it}$, $i = 1, \ldots, 10000$, $t = 1, 2, 3$ are normally distributed. The true parameters are specified as $\beta = (1, 1)'$, $\sigma^2 = 1$, $\sigma_z^2 = 10$ with the prior $(\beta, \sigma^2) \sim NIG\,(0, I_2, 3, 3)$, $\sigma_z^2 \sim IG\,(3, 10)$.

Compared to the Gibbs sampler of Gelfand and Smith (1990) that cycles through the full posterior conditionals, marginalization over $\beta, \sigma^2, \sigma_z^2$ by DMS, or alternatively over z as in Chib and Carlin (1999), may reduce the sample autocorrelations measured by the inefficiency factor (IF):

$$IF = 1 + 2\sum_{j=1}^{\infty} \rho_j,$$

where ρ_j is the j^{th} order autocorrelations of MCMC draws.

As is shown in Table 4, it comes to a consensus of the posterior means and standard deviations among the samplers, suggesting that they have converged to the same stationary distribution. Because the sample size is large and the prior is not tight, the posterior means are close to the true parameters.

Autocorrelations of the Gibbs sampler are highest for all variables. The block sampler of Chib and Carlin (1999) produces near perfect outputs for β, making them essentially independent draws. However, sampling β and z in the same block does not reduce autocorrelations of σ^2 and σ_z^2. The magnitude of IF is similar to that of the Gibbs sampler.

Compared to the Gibbs sampler, DMS effectively reduces autocorrelations of σ^2, σ_z^2 and β, because they are analytically integrated out. Although the reduction in β is not as drastic as the block sampler, the drop in σ^2 and σ_z^2 appears substantial, which provides evidence that marginalization can improve statistical efficiency of MCMC sampling.

	Gibbs Sampler			Block Sampler			DMS		
	Mean	Std	IF	Mean	Std	IF	Mean	Std	IF
β_1	0.990	0.007	2.031	0.990	0.007	1.023	0.990	0.007	1.353
β_2	1.001	0.007	2.067	1.001	0.007	0.965	1.001	0.007	1.340
σ^2	0.979	0.010	2.152	0.979	0.010	2.214	0.979	0.010	1.310
σ_z^2	10.007	0.146	1.114	10.009	0.146	1.172	10.009	0.146	1.022

Table 4: Posterior mean, standard deviation and inefficiency factor of draws generated by the Gibbs sampler, the block sampler of Chib and Carlin (1999) and DMS, in the hierarchical longitudinal model.

5 Extensions

In Algorithm 1, all parameters other than the latent variables are marginalized. However, there are situations where some parameters cannot be analytically integrated out, for lack of conjugate distributions. The problem can be addressed by two methods: double pseudo-marginalized subsampling and Metropolis within Algorithm 1, both of which slightly adapt Algorithm 1 while maintain the true target distribution.

5.1 Double Pseudo-marginalized Subsampling

Suppose that equation (2) cannot be evaluated in a closed form due to the intractable $p(z)$ and/or $p(y\,|z)$. Following Andrieu and Roberts (2009), we construct a non-negative and unbiased estimator such that

$$\hat{p}(z) = \frac{1}{M}\sum_{m=1}^{M} p(z\,|\theta_{1m}),$$

$$\hat{p}(y\,|z) = \frac{1}{M}\sum_{m=1}^{M} p(y\,|z, \theta_{2m}),$$

where the auxiliary variables $\theta_{1m}, \theta_{2m}, m = 1, \ldots, M$ follow the prior distributions $p(\theta_1), p(\theta_2\,|z)$, respectively. Unbiasness refers to the fact that

$$E[\hat{p}(z)] = \frac{1}{M}\sum_{m=1}^{M}\int p(z\,|\theta_{1m})\, p(\theta_{1m})\, d\theta_{1m} = p(z),$$

$$E[\hat{p}(y\,|z)] = \frac{1}{M}\sum_{m=1}^{M}\int p(y\,|z, \theta_{2m})\, p(\theta_{2m}\,|z)\, d\theta_{2m} = p(y\,|z).$$

Double pseudo-marginalized subsampling follows all the steps of Algorithm 1, except that the intractable $p(z)$ and $p(y\,|z)$ are replaced by the unbiased estimators $\hat{p}(z)$ and $\hat{p}(y\,|z)$. It is an exact approximation in that the extended Markov chain targets the joint density (up to a proportionality constant $p(y)$):

$$\pi\left(z, \overrightarrow{\theta}_1, \overrightarrow{\theta}_2\right) = \hat{p}(z)\,\hat{p}(y\,|z)\prod_{m=1}^{M} p(\theta_{1m})\, p(\theta_{2m}\,|z),$$

where $\overrightarrow{\theta}_1 = (\theta_{11}, \ldots, \theta_{1M})$, $\overrightarrow{\theta}_2 = (\theta_{21}, \ldots, \theta_{2M})$. The marginal distribution is the original target of the MH sampler. That is,

$$\int\int \pi\left(z, \overrightarrow{\theta}_1, \overrightarrow{\theta}_2\right) d\overrightarrow{\theta}_1 d\overrightarrow{\theta}_2 = p\left(z\right) p\left(y\left|z\right.\right).$$

If the auxiliary variables are drawn independently from the prior distribution, the naive unbiased estimator may have a prohibitively large variance, which leads to a sticky Markov chain, and the problem looms larger as the sample size increases. Following Deligiannidis et al. (2018), we generate correlated auxiliary variables by a reversible transition kernel: $u^* = \rho u + \sqrt{1-\rho^2}e$, where e is a standard normal variate and $\rho < 1$ determines correlation. The correlated normal variates u and u^* are mapped to θ_{1m} and θ_{1m}^* by the inverse cumulative distribution function method, so that the marginal distribution remains $p\left(\theta_1\right)$. The correlations between θ_{2m} and θ_{2m}^* are generated in the same manner. The correlated pseudo-marginal version of Algorithm 1 proceeds with evaluating the MH acceptance ratio $\frac{\hat{p}(z^*)\hat{p}(y|z^*)}{\hat{p}(z)\hat{p}(y|z)}$, where the auxiliary variables θ_{1m}, θ_{2m} are used for calculating $\hat{p}\left(z\right)\hat{p}\left(y\left|z\right.\right)$, while $\theta_{1m}^*, \theta_{2m}^*$ for $\hat{p}\left(z^*\right)\hat{p}\left(y\left|z^*\right.\right)$. The extended Markov chain still targets the exact posterior distribution by marginalizing over the auxiliary variables.

We illustrate the correlated pseudo-marginalized subsampling by a stochastic volatility model:

$$y_i = \beta_0 + \beta_1 e^{z_i} + e^{\frac{1}{2}z_i}\varepsilon_i,$$

$$z_i = c + \phi z_{i-1} + \gamma y_{i-1} + \sigma u_i,$$

where the latent variable z_i is the stochastic volatility associated with the data point y_i, the daily return of S&P 500 indices from January 2012 to January 2020. The model accommodates the stochastic volatility in mean (Koopman and Uspensky, 2002) if $\beta_1 \neq 0$ and the leverage effect (Yu, 2005) if $\gamma < 0$. The disturbances $\{\varepsilon_i, u_i\}$ are IID standard normal variates. The priors are specified as: $(c, \phi, \gamma, \sigma^2) \sim NIG\left(0, I_3, 3, 1\right)$, β_0 and β_1 are normal with the mean zero and the standard deviation 0.2.

We consider an unbiased estimator of $p\left(y\left|z\right.\right)$ such that

$$\hat{p}\left(y\left|z\right.\right) = \frac{1}{M}\sum_{m=1}^{M} p\left(y\left|z, \beta_{0m}, \beta_{1m}\right.\right),$$

where the auxiliary variables β_{0m}, $m = 1, \ldots, M, M = 100$ are generated by

$$\beta_{0m}^{*} = \rho \beta_{0m} + \sqrt{1-\rho^2} e_0,$$

and e_0 follows the normal distribution with the mean zero and the standard deviation 0.2. We implement the correlated pseudo-marginal (CPM) method with $\rho = 0.999$, and the independent pseudo-marginal (IPM) method with $\rho = 0$. The auxiliary variables β_{1m} and β_{1m}^{*} are handled in the same way. In step 2.6 of Algorithm 1, we compute the MH acceptance ratio $\frac{p(z^*)\hat{p}(y|z^*)}{p(z)\hat{p}(y|z)}$, where β_{0m}, β_{1m} are used for calculating $\hat{p}\left(y\left|z\right.\right)$, while $\beta_{0m}^{*}, \beta_{1m}^{*}$ for $\hat{p}\left(y\left|z^*\right.\right)$. The marginal likelihood $p\left(z\right)$ and $p\left(z^*\right)$ are calculated by equation (8).

Equation (9) is still applicable, and it is feasible to evaluate the absolute and relative error defined as $\ln \hat{p}\left(y\left|z\right.\right) - \ln p\left(y\left|z\right.\right)$ and $\ln \frac{\hat{p}(y|z^*)}{\hat{p}(y|z)} - \ln \frac{p(y|z^*)}{p(y|z)}$, respectively. Because $\hat{p}\left(y\left|z\right.\right)$ is a noisy estimator of $p\left(y\left|z\right.\right)$, the absolute error is inevitable unless we increase M. However, the correlation of the auxiliary variables may reduce the relative error. That is, the Monte Carlo noise may be partially cancelled in the MH acceptance ratio.

The stochastic volatility model is a standard tester of MCMC samplers. Typically, a draw of the parameters $\theta = (\beta_0, \beta_1, c, \phi, \gamma, \sigma^2)$ is generated after a pass of data. We follow this convention and run the samplers for 5000 passes of data. Samplers are initialized by setting $z_i = \ln \left(y_i - \frac{1}{n}\sum y_i\right)^2$, without the need of initializing θ. The trace plots of the first 700 draws of ϕ are presented in Figure 2. DMS (the top panel) that computes the analytical $p\left(y\left|z\right.\right)$ converges after 200 passes of data. It takes about 300 passes for CPM (the middle panel) to converge. IPM (the bottom panel) does not converge until 600 passes. They eventually reach an invariant distribution with the posterior mean 0.9 and the standard deviation 0.01. In our implementation of Algorithm 1, we set the tuning parameter $\lambda = 0.7$, which leads to an acceptance rate of 40.5% by DMS. With pseudo-marginalization, the acceptance rate drops: 36.5% for CPM and 27.7% for IPM. Experimenting with fewer auxiliary variables and more diffused prior distributions, we note that IPM may lead to a sticky chain or even get stuck, because of an excessively noisy estimator. The performance of CPM is good even if M is smaller.

The size of the Monte Carlo error depends on the number of auxiliary variables (M), the sample size (n) and the tightness of the prior. As is shown in Table 5, although both

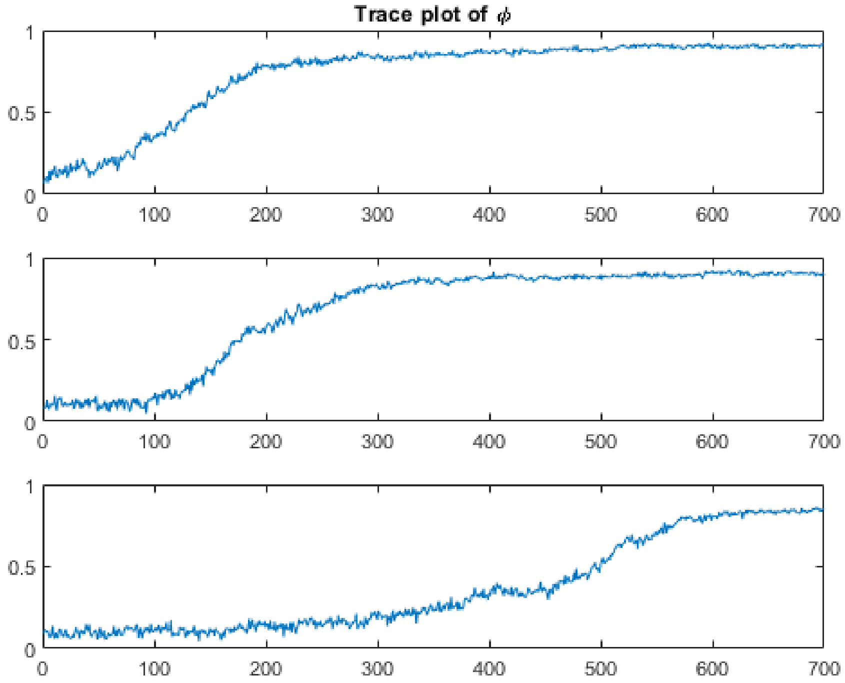

Figure 2: Trace plot of the parameter ϕ in the stochastic volatility model by DMS (top panel), CPM (middle panel) and IPM (bottom panel).

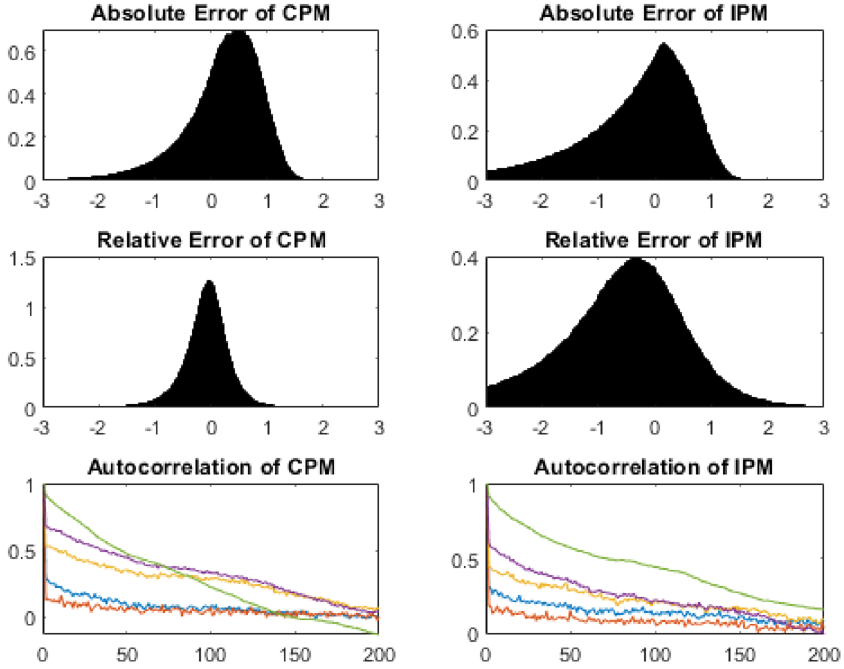

Figure 3: Histograms of the absolute and relative Monte Carlo error in evaluating $\ln \hat{p}\,(y\,|z\,)$ by CPM and IPM, as well as autocorrelations of parameter draws $(\beta_0, \beta_1, c, \phi, \gamma)$ in the stochastic volatility model.

CPM and IPM have acceptable error sizes under the current specification, the relative error of IPM is larger than that of CPM. Also, the histograms in Figure 3 suggest that the distribution of the relative error of CPM is bell-shaped, like a zero-mean normal distribution, while that of the absolute error appears biased and skewed, possibly due to the Jansen's inequality.

Summary statistics of the estimated parameters are reported in Table 5. The posterior mean of γ is -0.31 with the standard deviation 0.04, which indicates significant leverage effect: the volatility increases with the debt-to-equity ratio as the asset return falls. The posterior mean of β_1 is -0.04, with a large standard deviation 0.03. The negative but weak relationship between returns and contemporaneous volatility is in line with the finding of Koopman and Uspensky (2002).

	DMS	CPM	IPM
β_0	0.073	0.067	0.071
	0.017	0.017	0.018
	11.169	11.825	12.083
β_1	-0.042	-0.031	-0.039
	0.034	0.035	0.036
	4.661	4.122	5.008
c	-0.081	-0.071	-0.075
	0.014	0.015	0.014
	19.663	46.34	26.519
ϕ	0.895	0.904	0.902
	0.012	0.013	0.012
	38.023	87.689	63.457
γ	-0.314	-0.319	-0.311
	0.04	0.038	0.037
	149.597	109.834	187.451
abs err	0	0.739	1.285
rel err	0	0.435	1.546

Table 5: Stochastic volatility model estimated by DMS, CPM and IPM. For each parameter, we report the posterior mean (the first row), standard deviation (the second row) and the inefficiency factor (the third row). The last two rows show the RMSE of the absolute and relative Monte Carlo error in evaluating $\ln \hat{p}\,(y\,|z\,)$.

5.2 Metropolis within Algorithm 1

Just as a Metropolis element can be inserted in a Gibbs sampler (i.e., Metropolis-within-Gibbs sampler), an additional Metropolis step can also be included in Algorithm 1, which provides an alternative solution to the problem of the intractable $p\left(z\right)$ or $p\left(y\left|z\right.\right)$.

As an illustration, we consider the linear regression with t-distributed disturbances, which have a scale-mixture representation:

$$y_i = x_i\beta + \sigma\sqrt{z_i}\varepsilon_i,$$

where $\{\varepsilon_i\}$ are IID standard normal disturbances and the latent variable z_i follows the inverse gamma distribution $IG\left(\frac{v}{2}, \frac{v}{2}\right)$. Also, we specify the prior $(\beta, \sigma^2) \sim NIG\left(\mu, \Lambda, a, b\right)$.

If the degree of freedom parameter v were known, marginalized subsampling would proceed with $p\left(z\left|y, v\right.\right) \propto p\left(z\left|v\right.\right) p\left(y\left|z\right.\right)$ such that

$$p\left(z\left|v\right.\right) = \exp\left[-n\ln\mathit{\Gamma}\left(\frac{v}{2}\right) + \frac{nv}{2}\ln\frac{v}{2} - \left(\frac{v}{2}+1\right)\sum_{i=1}^{n}\ln z_i - \frac{v}{2}\sum_{i=1}^{n}z_i^{-1}\right],$$

$$p\left(y\left|z\right.\right) \propto \left(\overline{b}\right)^{-\overline{a}}\left|\overline{\Lambda}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\ln z_i},$$

where $\overline{\mu}, \overline{\Lambda}, \overline{a}, \overline{b}$ are essentially given by equations (4) - (7) with the rescaled data.

As the degree of freedom parameter v cannot be marginalized analytically, we alternately sample from $p\left(z\left|y, v\right.\right)$ and $p\left(v\left|y, z\right.\right)$. The sampler of the former is given by step 2.1 - 2.6 of Algorithm 1, while the latter is proportional to $p\left(z\left|v\right.\right)$ under the uniform prior of v, so we adapt step 2.7 of Algorithm 1 by inserting a Metropolis element that evaluates $IG\left(\frac{v}{2}, \frac{v}{2}\right)$ densities.

The Metropolis-within-Gibbs sampler cycles through $p\left(\beta, \sigma^2\left|y, z\right.\right)$, $p\left(z\left|y, v, \beta, \sigma^2\right.\right)$ and $p\left(v\left|z\right.\right)$. The timing of simulating v is after the full sample of z has been drawn, so it is updated once in a sweep over the whole dataset. However, for the Metropolis step within Algorithm 1, the best timing of simulating v is after the i^{th} element of z has been drawn, because the cumulative statistics has been refreshed at that time and it is computationally cheap to recycle them for generating a draw.

6 An Application of Mortgage Rates

Advances in technology make larger datasets available from various sources, one of which is the administrative data collected by the state or local authorities. The Home Mortgage Disclosure Act (HMDA) requires financial institutions to disclose loan-level data to the public. In the year 2018 alone, over five thousand institutions reported 15 million loans to HMDA database, with various variables on the characteristics of loans, properties, applicants and lenders. The loan purposes included home purchase (51%), refinancing (32%), home improvement (8%) and others. About 56% of the loans for home purchase were originated, 9% denied, and the remaining were withdrawn or purchased on the secondary market.

Our objective is to study factors that determine the mortgage rates for home buyers, so we consider the loan records for single-family home purchases and the sample size is about 4.7 million. If the mortgage rates of all applicants were available, we would run a linear regression. However, missing data account for 15% of the loan applications, but they are not missing at random: among the denied applications, none of them reports an interest rate; among the 4.2 million originated loans, 96% of them have interest rate data, which implies that the mortgage rate sample is selected by the loan action, and we observe mortgage rates only if the loans are originated. Heuristically, mortgage rates and loan actions are correlated, as a good applicant with a high propensity of loan approval might get a low mortgage rate. Such negative correlation induces incidental truncation of mortgage rates, as described by a sample selection model:

loan action: $z_{1i} = x_i \beta_1 + \varepsilon_{1i}$,

mortgage rate: $z_{2i} = x_i \beta_2 + \varepsilon_2$,

observation: $y_i = z_{2i} I\left(z_{1i} > 0\right)$,

where $\varepsilon_{1i}, \varepsilon_{2i}$ are zero-mean bivariate normal disturbances with the covariance matrix Σ.

The Heckman two-step procedure (Heckman, 1979) addresses the incidental truncation problem at a low computational cost. The first step is a Probit regression, in which the dependent variable is the action taken (loan originated or denied). The regressors include debt-to-income ratio (DIR), loan-to-value ratio (LVR), loan size, popular loan terms (15 or 30 years), dummy variables for the fixed rate, subordinate lien, loans insured by Federal Housing Administration (FHA) and Veterans Affairs (VA), the number of units, applicant age, gender, race, ethnicity, joint filing status and regional income. The second step is the mortgage rate regression for the originated loans. The regressors are augmented by the inverse Mills ratio constructed from the previous Probit estimator to correct the sample selection bias.

As there are millions of observations, we resort to the marginalized subsampling algorithm presented in Section 3.2 for the Probit regression. The tuning parameter in Algorithm 1 is chosen as $\lambda = 1$, and the MH sampler accepts proposal draws with the probability around 65%. The Probit regression results are presented in the second column of Table 6. The posterior means have the expected signs, and posterior standard deviations are small. The empirical results suggest various strategies for the home buyers to increase the chance of loan approval: lower DIR and LVR, choose the fixed rate 15 or 30 years loan term, consider a jumbo loan, avoid subordinate lien, take advantage of FHA and VA loans, buy a single-unit house, apply mortgage earlier in life, find a co-applicant for joint filing, etc.

The mortgage rate regression results are reported in the third column of Table 6, from which we find that 1) most factors that increase the probability of loan approval decease the mortgage rate as well. 2) If a home buyer decreases her DIR and LVR, she is expected to get a lower mortgage rate. 3) Traditionally the jumbo loan interest rate was higher than standard loan rate, but in recent years the tide has been reversed. Data suggest that the loan size is negatively correlated with the interest rate. 4) Currently the fixed rate is higher than the adjustable rate mortgage. 5) There is no evidence of gender discrimination. A female applicant tends to get a slightly better rate. 6) Young couples are considered favorably in mortgage rates. 7) Mortgage rates are lower in affluent regions.

With the two-step estimator as our starting value, we consider the full-fledged Bayesian analysis on the joint posterior distribution $p\left(\beta_1, \beta_2, \Sigma, z_1, z_2 \,|y\right)$ by marginalized subsampling. A technical issue is that the first diagonal element of the covariance matrix Σ is normalized to one. Following McCulloch and Rossi (1994), the sampler ignores the constraint and posterior draws are post-processed to focus on the identifiable quantities of

	Two-step		Unconstrained		Constrained	
	Action	Rate	Action	Rate	Action	Rate
Intercept	-1.479	7.022	-1.454	6.631	-1.505	6.628
	(0.047)	(0.020)	(0.051)	(0.017)	(0.029)	(0.015)
debt/Income	-2.815	0.822	-2.816	0.536	-2.821	0.530
	(0.007)	(0.007)	(0.006)	(0.008)	(0.004)	(0.002)
loan/Value	-0.407	0.506	-0.411	0.473	-0.400	0.473
	(0.006)	(0.003)	(0.007)	(0.002)	(0.004)	(0.002)
loanSize	1.702	-1.915	1.698	-1.742	1.686	-1.739
	(0.014)	(0.007)	(0.015)	(0.006)	(0.009)	(0.005)
popularTerm	0.224	0.001	0.225	0.029	0.223	0.029
	(0.004)	(0.002)	(0.004)	(0.002)	(0.003)	(0.001)
fixedRate	0.227	0.244	0.226	0.269	0.227	0.270
	(0.003)	(0.001)	(0.003)	(0.001)	(0.002)	(0.001)
subordinateLien	-0.668	1.133	-0.670	1.038	-0.666	1.038
	(0.007)	(0.004)	(0.007)	(0.003)	(0.005)	(0.003)
FHA	0.025	-0.072	0.025	-0.073	0.023	-0.073
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
VA	0.239	-0.296	0.239	-0.273	0.236	-0.273
	(0.003)	(0.001)	(0.003)	(0.001)	(0.002)	(0.001)
units	-0.194	0.179	-0.194	0.153	-0.191	0.152
	(0.003)	(0.002)	(0.003)	(0.001)	(0.002)	(0.001)
age	-0.460	0.185	-0.462	0.135	-0.457	0.134
	(0.006)	(0.003)	(0.006)	(0.002)	(0.004)	(0.002)
female	0.106	-0.025	0.105	-0.013	0.108	-0.013
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
jointFilers	0.142	-0.007	0.141	0.007	0.144	0.007
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
asian	-0.185	-0.014	-0.185	-0.033	-0.180	-0.034
	(0.004)	(0.001)	(0.004)	(0.001)	(0.002)	(0.001)
african	-0.275	0.137	-0.275	0.100	-0.273	0.100
	(0.003)	(0.002)	(0.003)	(0.001)	(0.002)	(0.001)
hispanic	-0.144	0.149	-0.144	0.131	-0.142	0.131
	(0.003)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
regionIncome	1.913	-0.953	1.900	-0.744	1.948	-0.743
	(0.043)	(0.018)	(0.043)	(0.016)	(0.028)	(0.014)

Table 6: Incidentally truncated mortgage rate regressions by Heckman two-step approach (columns 2 and 3), full Bayesian analysis by marginalized subsampling with the unconstrained covariance matrix (columns 4 and 5), and pseudo-marginalized subsampling with the constrained covariance matrix (columns 6 and 7). Posterior means are reported with standard deviations in parentheses.

interest. Also see Li and Tobias (2009, p.268) for a description.

Let $B = (\beta_1, \beta_2)$, $z_i = (z_1, z_2)'$, We specify a d-dimensional $NIW\left(U, \Lambda, \mathit{\Omega}, v\right)$ prior:

$$p\left(B, \Sigma\right) \propto \left|\Sigma\right|^{-\frac{v+d+3}{2}} \exp\left\{-\frac{1}{2}tr\left[\left(B-U\right)'\Lambda\left(B-U\right)\Sigma^{-1} + \mathit{\Omega}\Sigma^{-1}\right]\right\}.$$

Then $p\left(B, \Sigma\left|z\right.\right)$ takes the conjugate form: $NIW\left(\bar{U}, \bar{\Lambda}, \bar{\mathit{\Omega}}, \bar{v}\right)$ where

$$\bar{U} = \left(\Lambda + \sum_{i=1}^{n} x_i'x_i\right)^{-1}\left(\Lambda U + \sum_{i=1}^{n} x_i'z_i'\right),$$

$$\overline{\Lambda} = \Lambda + \sum_{i=1}^{n} x_i'x_i,$$

$$\bar{\mathit{\Omega}} = \mathit{\Omega} + \sum_{i=1}^{n} z_iz_i' + U'\Lambda U - \overline{U}'\overline{\Lambda U},$$

$$\overline{v} = v + n,$$

which are multivariate extensions to equations (4) - (7). It follows that $p\left(z\left|y\right.\right) \propto p\left(z\right)p\left(y\left|z\right.\right)$, where

$$p\left(z\right) \propto \left|\bar{\mathit{\Omega}}\right|^{-\overline{v}/2},$$

$$p\left(y\left|z\right.\right) = \prod_{i=1}^{n}\left[I\left(y_i \neq 0, z_{1i} > 0, z_{2i} = y_i\right) + I\left(y_i = 0, z_{1i} < 0\right)\right].$$

With the cumulative statistics $S_1 = \sum_{i=1}^{n} x_i'x_i$, $S_2 = \sum_{i=1}^{n} x_i'z_i'$, $S_3 = \sum_{i=1}^{n} z_iz_i'$, we implement Algorithm 1 and the posterior means and standard deviations are presented in the middle columns of Table 6. Compared to the two-step approach, the numerical results are similar, but the computing cost increases.

Lastly, we consider imposing the constraint $\Sigma = \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}$ and working directly with the identified parameters ρ and σ by the pseudo-marginalization technique presented in Section 5.1. For the priors, we assume that ρ and σ are normally distributed with the mean -0.1 and 0.5, respectively, with the standard deviation 0.1. Although $p\left(z\right)$ does not have a closed-form expression, we construct a non-negative and unbiased estimator: $\hat{p}\left(z\right) = \frac{1}{M}\sum_{m=1}^{M} p\left(z\left|\Sigma_m\right.\right)$, where Σ_m is the Monte Carlo counterpart of Σ, with ρ, σ replaced by ρ_m, σ_m, $m = 1, \ldots, M$, which are correlated draws from the prior distributions. The last two columns of Table 6 report the pseudo-marginalized subsampling results, which are similar to those produced by previous regressions.

7 Conclusion

As the fast growing data outpace the available computational resources, it is challenging to scale up the MCMC methods to handle the big data. DMS mainly addresses the problem of massive observations, while it also tackles the high dimensional problem by integrating out parameters, leaving only a vector of latent variables for subsampling. The multivariate Gaussian mixture model illustrates the scalability of our approach in both directions.

Subsampling methods can be exact or approximate. DMS and its pseudo-marginal extension target the exact posterior density. Subsampling increases computational efficiency without loss of statistical efficiency. Monte Carlo evidence suggests that marginalization reduces sample autocorrelations and improves convergence.

DMS is applicable to various econometric models, among which the stochastic volatility model is a specialty. The challenge is that the likelihood function $p(y|\theta)$ does not have a closed form. We note that $p(z|y)$ remains tractable by marginalizing over θ under the conjugate prior specification, and DMS proceeds without distributional approximations.

References

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics 37*(2), 697–725.

Azzalini, A. and A. Dalla Valle (1996). The multivariate skew-normal distribution. *Biometrika 4*, 715–726.

Balakrishnan, S. and D. Madigan (2006). A one-pass sequential Monte Carlo method for Bayesian analysis of massive datasets. *Bayesian Analysis 1*, 345–362.

Banterle, M., C. Crazian, A. Lee, and C. Robert (2019). Accelerating Metropolis-Hastings algorithms by delayed acceptance. *Foundations of Data Science 1*, 103–128.

Bardenet, R., A. Doucet, and C. Holmes (2017). On Markov Chain Monte Carlo methods for tall data. *Journal of Machine Learning Research 18*, 1–43.

Bierkens, J., P. Fearnhead, and G. Roberts (2019). The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *The Annals of Statistics 47*, 1288–1320.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*(3), 541–553.

Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*(451), 957–970.

Celeux, G., K. Kamary, G. Malsiner-Walli, J.-M. Marin, and C. P. Robert (2019). Computational solutions for Bayesian inference in mixture models. In S. Fruhwirth-Schnatter, G. Celeux, and C. P. Robert (Eds.), *Handbook of Mixture Analysis*. New York: Chapman and Hall.

Chib, S. (1992). Bayesian inference in the Tobit censored regression mode. *Journal of Econometrics 51*, 79–99.

Chib, S. and B. P. Carlin (1999). On mcmc sampling in hierarchical longitudinal models. *Statistics and Computing 9*, 17–26.

Cogley, T. and T. J. Sargent (2005). Drift and volatilities: Monetary policies and outcomes in the post WWII U.S. *Review of Economic Dynamics 8*(2), 262–302.

Creal, D. (2012). A survey of sequential Monte Carlo methods for economics and finance. *Econometric Reviews 31*(3), 245–296.

Deligiannidis, G., A. Doucet, and M. Pitt (2018). The correlated pseudo-marginal method. *Journal of the Royal Statistical Society: Series B 80*(5), 839–870.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89*(3), 603–615.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works. *Computational Statistics & Data Analysis 51*(7), 3529–3550.

Geweke, J. and G. Durham (2019). Sequentially adaptive Bayesian learning algorithms for inference and optimization. *Journal of Econometrics 210*(1), 4–25.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Holmes, C. C. and L. Held (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis 1*, 146–168.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics 20*, 69–87.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science 20*(1), 50–67.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. *Machine Learning 37*, 183–233.

Koop, G., D. J. Poirier, and J. L. Tobias (2007). *Bayesian Econometric Methods.* Cambridge: Cambridge University Press.

Koopman, S. J. and E. H. Uspensky (2002). The stochastic volatility in mean model: Empirical evidence from international stock market. *Journal of Applied Econometrics 17*, 667–689.

Korattikara, A., Y. Chen, and M. Welling (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. *Proceedings of the 31st International Conference on Machine Learning 1*, 181–189.

Li, M. and J. L. Tobias (2009). Bayesian methods in microeconometrics. In J. Geweke, G. Koop, and H. Van Dijk (Eds.), *The Oxford Handbook of Bayesian Econometrics*. Oxford: Oxford University Press.

Maclaurin, D. and R. Adams (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. *Uncertainty in Artificial Intelligence - Proceedings of the 30th Conference, UAI 2014*, 4289–4295.

Marin, J.-M., K. Mengersen, and C. P. Robert (2005). Bayesian modelling and inference on mixtures of distributions. In D. Dey and C. Rao (Eds.), *Bayesian Thinking*, Volume 25 of *Handbook of Statistics*, pp. 459–507. New York: Elsevier Science Publishers.

McCulloch, R. E. and P. Rossi (1994). An exact likelihood analysis of the multinomial Probit model. *Journal of Econometrics 64*, 207–240.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*, 1087–1092.

Neiswanger, W., C. Wang, and E. P. Xing (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pp. 623–632. Arlington: AUAI Press.

Nemeth, C. and Sherlock (2018). Merging MCMC subposteriors through Gaussian-process approximations. *Bayesian Analysis 13*(2), 507–530.

Qian, H. (2018). Big data Bayesian linear regression and variable selection by normal-inverse-gamma summation. *Bayesian Analysis 13*(4), 1011–1035.

Quiroz, M., R. Kohn, M. Villani, and M. Tran (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association 114*, 831–843.

Robbins, H. and S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics 22*(3), 400–407.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods.* New York: Springer Science Business Media.

Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management 11*, 78–88.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 62*(4), 795–809.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 528–550.

Teh, Y. W., H. A. Thiery, and S. J. Vollmer (2016). Consistency and fluctuations for stochastic gradient langevin dynamics. *Journal of Machine Learning Research 17*, 1–33.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics 22*, 1701–1728.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*, 24–36.

Yao, W. and B. G. Lindsay (2009). Bayesian mixture labeling by highest posterior density. *Journal of the American Statistical Association 104*(486), 758–767.

Yu, J. (2005). On leverage in a stochastic volatility model. *Journal of Econometrics 127*(2), 165–178.