Supplementary Materials: Appendices

The Generalized Multiset Sampler

Hang J. Kim and Steven N. MacEachern

1 PROOFS OF PROPOSITIONS

Proof of Proposition 1

The right hand side of expression (2) in the main text is the sum of the product of non-negative components, so $\pi^*(x, \mathbf{s}) \geq 0$ for all $(x, \mathbf{s})$. Working with the conditional density of $\mathbf{s}|x$,

$$\begin{aligned} \int_{\mathcal{S}} \pi^*(\mathbf{s}|x) d\mathbf{s} &= \int_{\mathcal{Y}} \cdots \int_{\mathcal{Y}} \sum_{k=1}^K \alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l) dy_1 \cdots dy_K \\ &= \sum_{k=1}^K \alpha_k \left[\int_{\mathcal{Y}} \cdots \int_{\mathcal{Y}} \pi(y_k|x) \prod_{l \neq k} f_l(y_l) dy_1 \cdots dy_K \right] \\ &= \sum_{k=1}^K \alpha_k \left[\int_{\mathcal{Y}} \pi(y_k|x) dy_k \right] \\ &= 1. \end{aligned}$$

Consequently,

$$\begin{aligned} \int_{\mathcal{S}} \pi^*(\mathbf{s}) d\mathbf{s} &= \int_{\mathcal{S}} \int_{\mathcal{X}} \pi^*(x, \mathbf{s}) dx d\mathbf{s} \\ &= \int_{\mathcal{S}} \int_{\mathcal{X}} \pi^*(x) \pi^*(\mathbf{s}|x) dx d\mathbf{s} \\ &= \int_{\mathcal{X}} \pi^*(x) dx \\ &= 1. \end{aligned}$$

Proof of Proposition 2

Without loss of generality, we take $k = 1$. We begin with the conditional density of $y_1|x$.

$$\begin{aligned}
\pi^*(y_1|x) &= \int_{y_1 \in \mathcal{S}} \pi^*(\mathbf{s}|x) d\mathbf{s} \\
&= \int_{y_1 \in \mathcal{S}} \sum_{k=1}^K \alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l) d\mathbf{s} \\
&= \int_{\mathcal{Y}} \cdots \int_{\mathcal{Y}} \alpha_1 \pi(y_1|x) \prod_{l \neq 1} f_l(y_l) dy_2 \cdots dy_K \\
&\quad + \sum_{k=2}^K \int_{\mathcal{Y}} \cdots \int_{\mathcal{Y}} \alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l) dy_2 \cdots dy_K \\
&= \alpha_1 \pi(y_1|x) + \sum_{k=2}^K \alpha_k f_1(y_1) \\
&= \alpha_1 \pi(y_1|x) + (1 - \alpha_1) f_1(y_1).
\end{aligned}$$

Passing to the marginal density for Y_1 ,

$$\begin{aligned}
\pi^*(y_1) &= \int_{\mathcal{X}} \pi^*(x) \pi^*(y_1|x) dx \\
&= \int_{\mathcal{X}} \pi(x) [\alpha_1 \pi(y_1|x) + (1 - \alpha_1) f_1(y_1)] dx \\
&= \alpha_1 \pi(y_1) + (1 - \alpha_1) f_1(y_1).
\end{aligned}$$

Proof of Theorem 3

$$\begin{aligned}
E_{\pi^*}[g^*(X, S)] &= \int_{\mathcal{X}} \int_{\mathcal{S}} \sum_{k=1}^K w_k g(x, y_k) \pi^*(\mathbf{s}|x) d\mathbf{s} \pi(x) dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{S}} \sum_{k=1}^K \frac{\alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l)}{\pi^*(\mathbf{s}|x)} g(x, y_k) \pi^*(\mathbf{s}|x) d\mathbf{s} \pi(x) dx \\
&= \int_{\mathcal{X}} \sum_{k=1}^K \alpha_k \left[\int_{\mathcal{Y}} \cdots \int_{\mathcal{Y}} \pi(y_k|x) \prod_{l \neq k} f_l(y_l) g(x, y_k) dy_1 \cdots dy_K \right] \pi(x) dx \\
&= \sum_{k=1}^K \alpha_k \left[\int_{\mathcal{X}} \int_{\mathcal{Y}} g(x, y_k) \pi(x, y_k) dy_k dx \right] \\
&= \sum_{k=1}^K \alpha_k E_{\pi}[g(X, Y)] \\
&= E_{\pi}[g(X, Y)].
\end{aligned}$$

2 CONVERGENCE RESULTS

The strategies used to produce an ergodic Markov chain for π can also be used to produce an ergodic chain for π^* . To streamline the presentation, we will assume that the entire parameter vector has been multisetted. The two key conditions to establish convergence to a limiting distribution are ϕ -irreducibility and aperiodicity (see Robert and Casella, 2004, (7.4) and (7.5), page 273), although the latter condition is not needed to develop consistent estimators. Here, we follow Robert and Casella (2004) and Roberts and Rosenthal (2006) to prove that our estimators are consistent under the GMSS sampling algorithm in Section 3.1, which is, formally, a deterministic scan Metropolis-within-Gibbs Markov chain in the terminology of Roberts and Rosenthal (2006).

Theorem 4. (Theorem 12 of Roberts and Rosenthal, 2006) *Consider a ϕ -irreducible Metropolis-within-Gibbs Markov chain. Suppose that from any initial state S , with probability 1, the chain will eventually move at least once in each coordinate direction. Then the chain is Harris recurrent.*

Theorem 5. (Theorem 6.63 of Robert and Casella, 2004) *Suppose $\{S^{(t)}\}$ has a σ -finite invariance measure π^* . If $\{S^{(t)}\}$ is Harris recurrent, then*

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T g^*(S^{(t)}) = \int g^*(\mathbf{s}) \pi^*(\mathbf{s}) d\mathbf{s} \quad \text{for } g^* \in L^1(\pi^*).$$

Corollary 6. *The GMSS estimator $\sum_{t=1}^T g^*(S^{(t)})/T$ where $g^*(S) = \sum_{k=1}^K w_k g(Y_k)$ converges to $E_\pi[g(Y)]$.*

Proof. Theorem 4 holds for the GMSS algorithm which is ϕ -irreducible and eventually moves at least once for every y_k since π^* and $q(y_k^q | y_k^{(t-1)})$ for $k = 1, \dots, K$ are positive continuous functions. Then, by Theorem 5, $\lim_{T \rightarrow \infty} g^*(S^{(t)})/T = E_{\pi^*}[g^*(S)]$ for $g^* \in L^1(\pi^*)$. By Corollary 4 in the main text, it follows that the GMSS estimator converges to $E_\pi[g(Y)]$. Note that $g \in L^1(\pi)$ implies $g^* \in L^1(\pi^*)$. $\square$

3 INTEGRATED AUTOCORRELATION TIME

Let $\{X_n\}$ be a Markov chain with a stationary distribution π . Let h be a real-valued function defined on the state space. Then $\{h(X_n)\}$ is a stationary Markov process with mean $E[h(X_n)]$ and normalized autocorrelation function

$$\rho(t) = \frac{Cov[h(X_s), h(X_{s+t})]}{Var[h(X_s)]}.$$

Sokal (1989) defined the *integrated autocorrelation time* (IACT) by

$$\tau = \frac{1}{2} \sum_{t=-\infty}^{\infty} \rho(t).$$

Given N samples of a stationary Markov chain $\{h(x_1), \dots, h(x_N)\}$, the sample mean is $\bar{h} = \frac{1}{N} \sum_{n=1}^N h(x_n)$ and the variance of the sample mean is

$$Var[\bar{h}] \approx \frac{2\tau Var[h(X_n)]}{N} \text{ for } N \gg \tau.$$

If the samples of a Markov chain $\{h(X_n)\}$ are independent, the variance of its sample mean is $Var[h(X_n)]/N$. So we can interpret the factor, 2τ , as the cost of using dependent samples compared to independent samples. In other words, the effective sample size is roughly $N/(2\tau)$.

To reduce noise present in the sample autocorrelation estimator $\hat{\rho}(t)$, Sokal proposed the use of a “window” $\lambda(t)$ in the estimator for IACT

$$\hat{\tau} = \frac{1}{2} \sum_{t=-(N-1)}^{N-1} \lambda(t) \hat{\rho}(t)$$

where $\lambda(t) = 1$ if $|t| \leq M$ and 0 if $|t| > M$ where M is a user-defined cutoff. Then the IACT estimator has a bias, $E[\hat{\tau} - \tau] = -\frac{1}{2} \sum_{|t| > M} \rho(t) + o(1/N)$, and a variance, $Var(\hat{\tau}) \approx \tau^2(4M+2)/N$. We can see the tradeoff between bias and variance as taking a large M making the variance large, but the bias small. Sokal suggested an “automatic windowing” algorithm which chooses M to be the smallest number satisfying $M \geq c \hat{\tau}$ with the pre-specified value c around $4 \sim 10$.

4 SIMULATION EXAMPLE WITH BAYESIAN REGRESSION ANALYSIS

This section presents a simulation study to demonstrate GMSS application to a continuous distribution with unbounded support, and to show the relative importance of instrumental densities derived from regional information for local modes.

4.1 Bayesian Regression Analysis for Data With Outliers

Consider a regression model $z_i = \beta_0 + \beta_1 d_i + \varepsilon_i$ on a simulate dataset with 50 observations, similar to the example used by Rousseeuw (1984), where 40 points are randomly generated from $d_i \sim \text{Unif}(1, 4)$, $z_i|d_i \sim N(2 + d_i, 0.6^2)$, and the other 10 points are generated from $d_i \sim N(7, 0.1^2)$, $z_i \sim N(2, 0.1^2)$. In Figure 1, we can see two regression fits, the dotted line representing the ordinary least square (OLS) fit on all 50 points, $z_i = 5.40 - 0.38d_i$, and the solid line representing the OLS fit on only 40 points of the top-left, $z_i = 2.32 + 0.89d_i$. One can interpret the 10 points of the bottom-right as outliers which severely affect on the regression analysis, resulting in the negative slope. Justel and Peña (1996) call this phenomenon *strong masking* because the relatively small number of outliers mask the positive linear relation among the “inliers”.

As an alternative to the classical linear regression model, Bayesian regression analysis adopts a model to explain the influence of outliers. A simple model is the scale contaminated model used by Justel and Peña (1996) where

$$\begin{aligned} z_i &= \beta_0 + \beta_1 d_i + \varepsilon_i, \\ p(\varepsilon_i) &= (1 - w)N(0, \sigma^2) + wN(0, c^2 \sigma^2), \\ p(\boldsymbol{\theta}) &\propto \sigma^{-2} \end{aligned} \tag{1}$$

where $\boldsymbol{\theta}$ denotes $(\beta_0, \beta_1, \sigma^2)$. For demonstration purpose, the contamination w and the scale parameter c are fixed as $w = 0.2$ and $c = 10.0$ in this simulation example. For this kind of data, it has been known that Gibbs sampling has problems of convergence in fitting the Bayesian regression model (Justel and Peña, 1996).

Figure 1: Simulated data for regression analysis on the two groups of data, 40 inliers randomly generated from $d_i \sim \text{Unif}(1, 4)$, $z_i|d_i \sim N(2 + d_i, 0.6^2)$ (black dots) and 10 outliers generated from $d_i \sim N(7, 0.1^2)$, $z_i \sim N(2, 0.1^2)$ (red dots). The dashed line is the ordinary least square (OLS) fit on all observations, $z_i = 5.40 - 0.38d_i$, and the solid line is the OLS fit on the 40 inliers only, $z_i = 2.32 + 0.89d_i$.

4.2 GMSS Application Using Informative Instrumental Densities

Instead of Gibbs sampling, we run the GMSS with $K = 2$, $\alpha_k = 1/2$, and a permutation step, to draw posterior samples of $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$. We divide the parameters into $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and σ^2 , and define the instrumental density by

$$f_k(\boldsymbol{\theta}_k) = f_{\beta,k}(\boldsymbol{\beta}_k) \times f_{\sigma,k}(\sigma_k^2), \quad k = 1, 2.$$

For σ^2 , we put a relatively flat instrumental density, $f_{\sigma,k}(\sigma_k^2) \sim \text{Gamma}(2, 2)$ for $k = 1, 2$.

For this specific problem, we can utilize the information about the possible modes of $\boldsymbol{\beta}$ to set the “informative” form of $f_{\beta,k}(\boldsymbol{\beta}_k)$. Motivated by the different regression fits in Figure 1, we adopt the robust regression with the Tukey’s biweight estimator to set $f_{\beta,k}(\boldsymbol{\beta}_k)$ with the following steps:

1. Choose the initial values for the robust regression from the median regression fits.
 - (a) Make a sub-dataset by randomly drawing 25 points from the original data.
 - (b) Fit the quantile regression on the sub-datasets and set the resulting 50th regression quantile as the initial values for the robust regression in Step 2.
2. Fit robust regression with the Tukey’s biweight function on the original data (with all 50 points), using the initial points obtained in Step 1-(b).
3. Repeat R times.

We repeat the process $R = 100$ times and cluster the results so that $\hat{\beta}_1$ of replications within each group differ at most 0.01. Out of the 100 replications, the first group consists of two replications whose mean values of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})$ are $\bar{\boldsymbol{\theta}}_a = (2.31, 0.89, 0.75)$ while the second group consists of 98 replications whose mean values are $\bar{\boldsymbol{\theta}}_b = (5.45, -0.41, 1.03)$. The standard deviations of all regression estimates within each group are smaller than 10^{-4} . Now we set

$$f_{\beta,1}(\boldsymbol{\beta}_1) \sim N((2.31, 0.89)^T, \nu^2 I), \quad f_{\beta,2}(\boldsymbol{\beta}_2) \sim N((5.45, -0.41)^T, \nu^2 I). \quad (2)$$

Note that $\bar{\boldsymbol{\theta}}_a$ and $\bar{\boldsymbol{\theta}}_b$ are similar to the two fitted lines in Figure 1. However, it is not always easy to find the cluster of outliers by seeing the scatter plot. By adopting our procedure, one can find the possible location of modes of $\boldsymbol{\beta}$, systematically and easily. We use `rq` function

in R-Package ‘quantreg’ with default setting for Step 1-(b) and `rlm` function in R-package ‘MASS’ for Step 2.

4.3 Simulation Results

For fitting the model in Equation (1), we run the random-walk Metropolis-Hastings (MH), the GMSS with the informative instrumental densities (GMSSi) using Equation (2), and the GMSS with a normal instrumental density (GMSSn) using

$$f_{\beta,k}(\beta_k) \sim N \left(\left(\frac{2.31 + 5.45}{2}, \frac{0.89 - 0.41}{2} \right)^T, v^2 I \right), \quad k = 1, 2. \quad (3)$$

For all methods, the random walk proposals are used as follows:

$$\beta'_0 \sim N(\beta_0^{(t-1)}, \tau_{\beta_0}^2), \quad \beta'_1 \sim N(\beta_1^{(t-1)}, \tau_{\beta_1}^2), \quad \sigma^{2'} \sim \text{logNormal}(\log \sigma^{2(t-1)}, \tau_{\sigma^2}^2)$$

where $(\tau_{\beta_0}, \tau_{\beta_1}, \tau_{\sigma^2}) = (0.2, 0.1, 1.5)$. MH has two versions with different starting points, $\theta^{(0)} = \bar{\theta}_a$ (MH-a) and $\theta^{(0)} = \bar{\theta}_b$ (MH-b), while GMSSi and GMSSn start the chains with the two multiset elements by setting $\theta_1^{(0)} = \bar{\theta}_a$ and $\theta_2^{(0)} = \bar{\theta}_b$.

Figure 2 shows result from MH with two starting points, GMSSi with $\nu = 0.25$ (GMSSi-M), and GMSSn with $\nu = 1.0$ (GMSSn-M). The upper panels show scatter plots of drawn samples from the M-H’s while the lower panels show the weighted samples (β_0^*, β_1^*) of the GMSS calculated by $\beta_p^* = \sum_{k=1}^K w_k \beta_{p,k}$, $p = 0, 1$. From the figure, we can see that the MH’s have convergence problems as the results are dependent to its starting points, while the GMSS’s can travel between the two posterior modes.

Figure 3 shows the result of simulation study with 50 replications, each representing the Markov chain with 10^5 iterations. For comparisons, we change the spread of the instrumental densities for the GMSS. Specifically, we add GMSSi with $\nu = 0.1$ (GMSSi-S) and $\nu = 0.5$ (GMSSi-L), and GMSSn with $\nu = 0.5$ (GMSSn-S) and with $\nu = 2.0$ (GMSSn-L). The results are summarized as follows. First, the relatively small standard deviations of the results from GMSSi-S and GMSSi-M imply that finding good instrumental densities results in more efficient estimators and accurate inference. Second, the success of GMSSn in finding two modes for all settings also implies that one can use a normal instrumental distribution for continuous, unbounded case, when no enough information about the location of modes exists.

Figure 2: Scatter plot of drawn samples for the regression coefficients (β_0, β_1) . The upper panels show the results of MH with two starting points, $\boldsymbol{\theta}^{(0)} = \bar{\boldsymbol{\theta}}_a$ (MH-a) and $\boldsymbol{\theta}^{(0)} = \bar{\boldsymbol{\theta}}_b$ (MH-b). The bottom panel shows the weighted samples (β_0^*, β_1^*) calculated by $\beta_p^* = \sum_{k=1}^K w_k \beta_{p,k}$, $p = 0, 1$ from the GMSS with the informative instrumental densities with $\nu = 0.25$ (GMSSi-M), and the GMSS with normal instrumental densities with $\nu = 1.0$ (GMSSn-M). The M-H's find only one of either modes while the GMSS's find the two modes.

Figure 3: The box plots of point estimators for β_1 over 50 replications of the MCMC methods. The M-H are implemented with two starting points, $\theta^{(0)} = \bar{\theta}_a$ (MH-a) and $\theta^{(0)} = \bar{\theta}_b$ (MH-b). The GMSS with the informative instrumental densities is implemented with the different values of ν : 0.1 (GMSSi-S), 0.25 (GMSSi-M) and 0.5 (GMSSi-L). The GMSS with normal instrumental densities is also implemented with the different values of ν : 0.5 (GMSSn-S), 1.0 (GMSSn-M) and 2.0 (GMSSn-L). The red dotted line is the result from a long-run GMSS implementation with 5×10^7 iterations.

References

- Justel, A., and Peña, D. (1996), “Gibbs Sampling Will Fail in Outlier Problems With Strong Masking,” *Journal of Computational and Graphical Statistics*, 5, 176–189.
- Robert, C. P., and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd ed.), New York: Springer-Verlag.
- Roberts, G. O., and Rosenthal, J. S. (2006), “Harris Recurrence of Metropolis-Within-Gibbs and Trans-Dimensional Markov Chains,” *The Annals of Applied Probability*, 16, 2123–2139.
- Rousseeuw, P. J. (1984), “Least Median of Squares Regression,” *Journal of the American Statistical Association*, 79, 871–880.
- Sokal, A. D. (1989), “Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms,” Troisième Cycle de la Physique en Suisse Romande lecture notes.