Paper Review: Simple, Scalable and Accurate Posterior Interval Estimation

Speaker: Jincheng Bai

Authors: Cheng Li, Sanvesh Srivastava, David B. Dunson

Duke University and The University of Iowa

June 14, 2016

Posterior Interval for Big Data?

- The typical MCMC algorithms face major problems in scaling up to big data. Turn to scalable sampling algorithms.
- Three of the most successful strategies are:
  - Approximating expensive MCMC transition kernels with easier to sample surrogates.
  - Running MCMC on a single machine but with different subsets of the data used as sampling proceeds (subsampling).
  - Running MCMC in parallel for different data subsets and then combining. (Embarrasingly parallel MCMC).
- The author's algorithm falls in the third category.

Example1: Subsamping [1]

- $$\frac{\pi(\theta^{'}|x)q(\theta|\theta^{'})}{\pi(\theta|x)q(\theta^{'}|\theta)} > u \Rightarrow log[\frac{\pi(x|\theta^{'})}{\pi(x|\theta)}] > log[u\frac{q(\theta^{'}|\theta)\pi_0(\theta))}{q(\theta|\theta)\pi_0(\theta^{'})}]$$

  .
- Let $\Lambda(\theta,\theta^{'}) = \frac{1}{N}\Sigma_{n=1}^{N} log[\frac{\pi(x_n|\theta^{'})}{\pi(x_n|\theta)}] = \frac{1}{N}\Sigma_{n=1}^{N} l_n$, and $\psi(u,\theta,\theta^{'}) = \frac{1}{N} log[u\frac{q(\theta^{'}|\theta)\pi_0(\theta))}{q(\theta|\theta)\pi_0(\theta^{'})}]$, the acceptance condition is then $\Lambda(\theta,\theta^{'}) > \psi(u,\theta,\theta^{'})$.
- Let $\{l_n^{\star}\}_{n=1}^{m}$ be a subsample of size $m < N$, without replacement from $\{l_n\}_{n=1}^{N}$, then the acceptance condition came to be

$$\hat{\Lambda}_m(\theta,\theta^{'}) > \psi(u,\theta,\theta^{'})$$

  where $\hat{\Lambda}_m(\theta,\theta^{'}) = \frac{1}{m}\Sigma_{n=1}^{m} l_n^{\star}$.

Example1: Subsamping [1]

- By concentration inequalty, for $\delta_m \in (0,1)$ and some constant c_m,

$$P(|\hat{\Lambda}_m(\theta,\theta') - \Lambda(\theta,\theta')| \leq c_m) \geq 1 - \delta_m$$

.

- if $|\hat{\Lambda}_m(\theta,\theta') - \psi(u,\theta,\theta')| > c_m$, then the approximate MH test agrees with the exact MH test with probability $1-\delta_m$. So the number of data points in the subset evaluated using this criterion is

$$M = min(N, inf_{m\geq 1}|\hat{\Lambda}_m(\theta,\theta') - \psi(u,\theta,\theta')| > c_m)$$

.

- In the implementation, the user need to set δ_m.

Example2: Sub-posterior density estimation [2]

- Suppose we divide the data into J partition elements, $x^{(1)}, \ldots, x^{(J)}$, then

$$\pi(\theta|x) = \prod_{j=1}^{J} \pi^{(j)}(\theta|x^{(j)})$$

where $\pi^{(j)}(\theta|x^{(j)}) = \pi_0(\theta)^{1/J} \prod_{x \in x^{(j)}} \pi(x|\theta)$, $j = 1, \ldots, J$. (product equation representation)

- Use a nonparametric kernel density estimate for each subposterior:

$$\tilde{\pi}^{(j)}(\theta|x^{(j)}) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{h^d} K(\frac{||\theta - \theta_{j,t}||}{h})$$

- The resulting product estimate is a mixture of T^m Gaussians.

Invariants to Example 2 and Problems

- Weierstrass samplers, Multi-scale histograms etc.
- Rely heavily on the accuracy of density estimators for the subset posteriors.
  - Curse of Dimensionality.
  - Suffering badly when subset posteriors have even slightly non-overlapping supports.
- The author's combining strategy falls in the category which use data subsamples to define noisy approximations to the full data posteriors, and then take an appropriate notion of geometric center, such as geometric median [3] or mean [4].

Wasserstein Space and Distance

- Suppose $\Theta \in R^d$, $||\theta_1 - \theta_2||$ is the Euclidean distance between any $\theta_1, \theta_2 \in \Theta$.
- The Wasserstein-2 space is $P_2(\Theta) = \{v : \int_\Theta ||\theta||^2 dv(\theta) < \infty\}$.
- For any two measures, v_1, v_2 on Θ, their Wasserstein-2 distance is

$$W_2(v_1, v_2) = \{\inf_{\gamma \in \Gamma(v_1, v_2)} \int_{\Theta \times \Theta} ||\theta_1 - \theta_2||^2 d\gamma(v_1, v_2)\}^{1/2}$$

- $\Gamma(v_1, v_2)$ is the set of all probability measures on $\Theta \times \Theta$ with marginals v_1 and v_2, respectively.

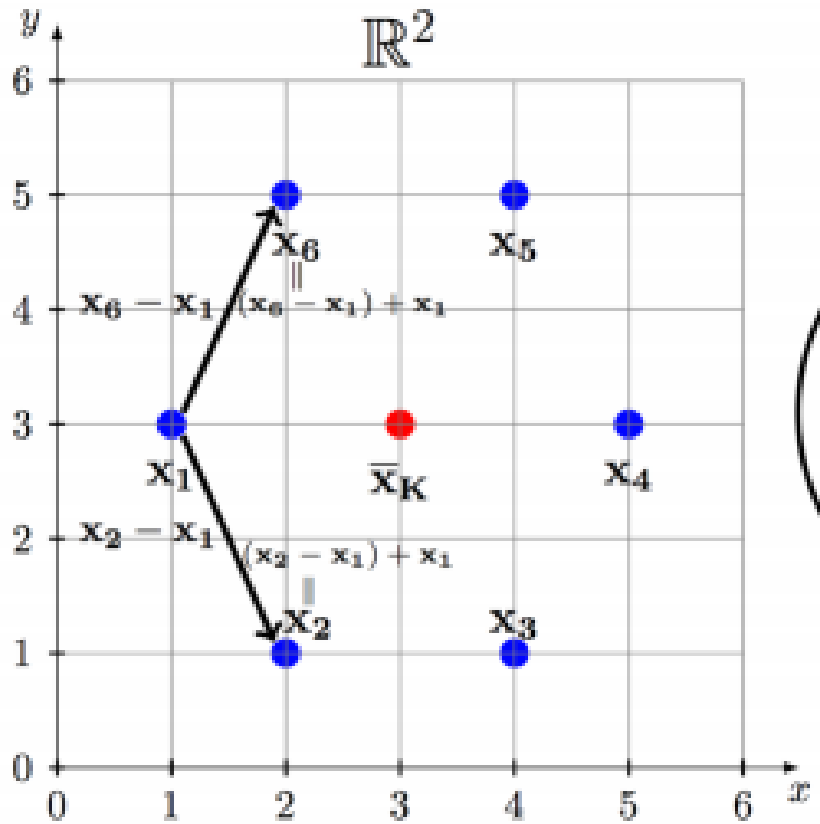
Wasserstein Space and Distance

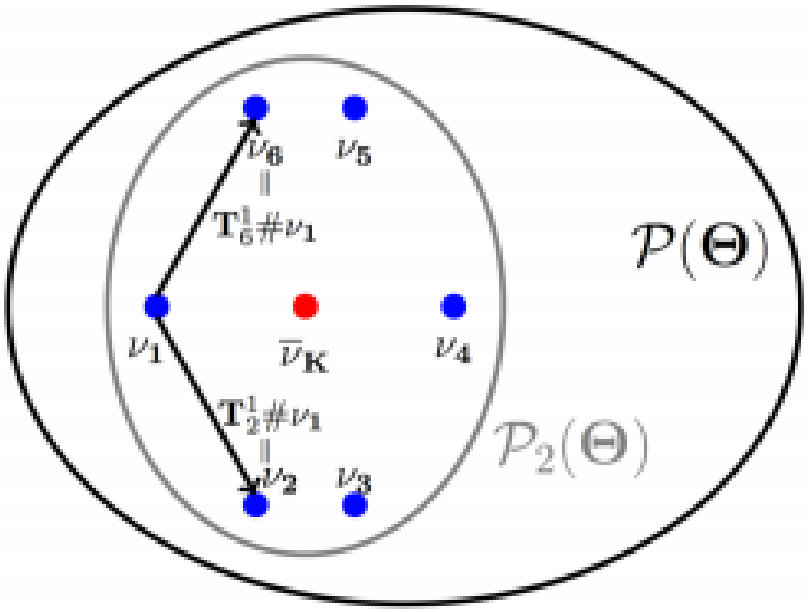
- W_2 distance is well defined for every pair of measures in $P_2(\theta)$.
- W_2 distance is the solution of an optimal transport problem: To minimize $\gamma \to \int_{X\times Y} c(x,y)d\gamma(x,y)$ in the set of all transport plans $\gamma \in P(X \times Y)$ from μ to v. i.e. the set of Borel Probability measures on $X \times Y$ s.t.

$$\gamma(A \times Y) = \mu(A), \forall A \in B(X), \gamma(X \times B) = v(B), \forall B \in B(Y)$$

Equivalently: $\pi^X_\sharp \gamma = \mu, \pi^Y_\sharp \gamma = v$, where π^X, π^Y are the natural projections from $X \times Y$ onto X and Y respectively.

Wasserstein barycenter





Wasserstein barycenter

- Given N different measures $v_1, \ldots, v_N$ in $P_2(\Theta)$.
- The Wasserstein barycenter is defined as the solution to the following optimization problem:

$$\bar{v} = arg \min_{\mu \in P_2(\Theta)} \sum_{j=1}^{N} W_2^2(\mu, v_j)$$

- The geometric center of the N measures $v_1, \ldots, v_N$.

Wasserstein Posterior

- Consider n observations can be partitioned into K non-overlapping subsets, for simplicity, n=Km.
- $X = \cup_{j=1}^{K} X_j$.
- The overall posterior density of θ given X:

$$\pi_n(\theta|X) = \frac{\prod_{j=1}^{K} \prod_{i=1}^{m} p(X_{ij}|\theta)\pi(\theta)}{\int_\Theta \prod_{j=1}^{K} \prod_{i=1}^{m} p(X_{ij}|\theta)\pi(\theta)d\theta}$$

- The jth subset posterior density of θ given X_j, $j = 1, \ldots, K$:

$$\pi_m(\theta|X_j) = \frac{\{\prod_{i=1}^{m} p(X_{ij}|\theta)\}^K \pi(\theta)}{\int_\Theta \{\prod_{i=1}^{m} p(X_{ij}|\theta)\}^K \pi(\theta)d\theta}$$

Wasserstein Posterior

- Denote their corresponding distribution functions as $\Pi_n(\theta|X)$ and $\Pi_m(\theta|X_j)$.
- The variance of each subset posterior given X_j is roughly of the same order as the variance of the overall posterior $\Pi_n(\theta|X_j)$ after raising to the kth power.
- Run MCMC on the K subsets in parallel, the Wasserstein barycenter of the draws for all K subsets is used as an approximation of the overall posterior $\Pi_n(\theta|X)$.

Wasserstein Posterior

- Only interested in a scalar parameter $\xi = h(\theta) \in \Xi$ with $h : \Theta \to \Xi \subseteq R$
- $\Pi_n(\xi|X)$: overall posterior for ξ; $\Pi_m(\xi|X_j)$: jth subset posterior for ξ.
- For theory development, focus on the linear functional $\xi = h(\theta) = a^T\theta + b,\ a \in R^d, b \in R$
- The one-dimensional Wasserstein posterior $\overline{\Pi}_n(\xi|X)$ is defined as the Wasserstein barycenter of $\Pi_m(\xi|X_j)$:

$$\overline{\Pi}_n(\xi|X) = arg \min_{\mu \in P_2(\Xi)} \sum_{j=1}^{K} W_2^2(\mu, \Pi_m(\xi|X_j))$$

Wasserstein Posterior in the one-dimensional case

- Let F_1 and F_2 be two generic univariate continous distributions in $P_2(\Xi)$.
- $F_1^{-1}(u)$ and $F_2^{-1}(u)$ are the quantile functions, for any $u \in (0, 1)$
- The W_2 distance between F_1 and F_2 is:

$$W_2(F_1, F_2) = [\int_0^1 \{F_1^{-1}(u) - F_2^{-1}(u)\}^2 du]^{1/2}$$

- $\overline{\Pi}_n(\xi|X)$ is explicitly related to the subset posteriors $\Pi_m(\xi|X_j))$ by

$$\overline{\Pi}_n^{-1}(u|X) = \frac{1}{K} \sum_{j=1}^{K} \Pi_m^{-1}(u|X_j)$$

- The above results were derived in [5] from an optimal transport perspective via linear programming.

Posterior Interval Estimation

Algorithm 1 Posterior Interval Estimation

Input: K subsets of data $X_1, \ldots, X_K$, each with sample size m.

Output: Posterior credible intervals $[\overline{q}_{\alpha/2}, \overline{q}_{1-\alpha/2}]$, for $\alpha \in (0,1)$.

For $j = 1$ to K # (Parallel in K subsets)

  For $i = 1$ to N

    Draw θ_{ij} from $\Pi_m(\theta \mid X_j)$, using appropriate posterior sampler.

    Calculate $\xi_{ij} = h(\theta_{ij})$.

  End for

  Sort $\{\xi_{1j}, \ldots, \xi_{Nj}\}$; Obtain the empirical $\alpha/2$ and $1-\alpha/2$ quantiles $q_{\alpha/2,j}$ and $q_{1-\alpha/2,j}$.

End for

Set $\overline{q}_{\alpha/2} = \frac{1}{K}\sum_{j=1}^{K} q_{\alpha/2,j}$ and $\overline{q}_{1-\alpha/2} = \frac{1}{K}\sum_{j=1}^{K} q_{1-\alpha/2,j}$.

Return: $[\overline{q}_{\alpha/2}, \overline{q}_{1-\alpha/2}]$.

Main Results

Assumptons:

- 1. θ_0 is an interior point of $\Theta \in R^d$. $P_\theta = P_{\theta_0}$ almost everywhere if and only if $\theta = \theta_0$. X contains i.i.d observations generated from P_{θ_0}.
- 2. The support of $p(x|\theta)$ is the same for all $\theta \in \Theta$.
- 3. $logp(x|\theta)$ is three times differentiable w.r.t. θ in a neighborhood $B_{\delta_0}(\theta_0) = \{\theta \in \Theta : ||\theta - \theta_0|| \leq \delta_0\}$ of θ_0.
  $E_{P_{\theta_0}}\{\dot{p}(X|\theta_0)/p(X|\theta_0)\} = 0$. Furthermore, there exists an envelope function s.t. $\sup_{\theta \in B_{\delta_0}(\theta_0)} |\partial logp(x|\theta)/\partial\theta_{l_1}| \leq M(x)$,
  $\sup_{\theta \in B_{\delta_0}(\theta_0)} |\partial^2 logp(x|\theta)/\partial\theta_{l_1}\partial\theta_{l_2}| \leq M(x)$,
  $\sup_{\theta \in B_{\delta_0}(\theta_0)} |\partial^3 logp(x|\theta)/\partial\theta_{l_1}\partial\theta_{l_2}\partial\theta_{l_3}| \leq M(x)$ for all
  $l_1, l_2, l_3 = 1, \ldots, d$ and $E_{P_{\theta_0}} M(X)^4 < \infty$

Main Results

- 4. $I(\theta) = E_{P_{\theta_0}}\{-\partial^2 p(X|\theta)/\partial\theta\partial\theta^T\} = -l_1^{''}/m$ is positive definite with eigenvalues bounded from below and above by constants.
- 5. For any $\delta > 0$, there exists an $\epsilon > 0$ s.t.

$$\lim_{m\to\infty} P_{\theta_0}[sup_{|\theta-\theta_0|\geq\delta}\{l_1(\theta) - l_1(\theta_0)/m \leq -\epsilon\}] = 1$$

- 6. The prior density $\pi(\theta)$ is continuous, bounded from above in Θ and bounded below at θ_0. The prior has finite second moment $\int_\Theta ||\theta||^2\pi(\theta)d\theta < \infty$.
- 7. Let $\psi(X_1) = E_{\Pi_m(\theta|X_1)}Km||\theta - \hat{\theta}_1||^2$, where $E_{\Pi_m(\theta|X_1)}$ is the expectation w.r.t. θ under posterior $\Pi_m(\theta|X_1)$. Then there exists an interger $m_0 \geq 1$, s.t. $\psi(X_1) : m \geq m_0, K \geq 1$ is uniformly integrable under P_{θ_0}.

Main Results

Theorem (1)

Suppose Assumptions 1-7 hold and $\xi = a^T\theta + b$ *for some fixed* $a \in R^d$ *and* $b \in R$. *Let* $I_\xi = \{a^T I^{-1}(\theta_0)a\}^{-1}$. *Let* $\bar{\xi} = a^T\bar{\theta} + b$, $\hat{\xi} = a^T\hat{\theta} + b$. *Let* $\Phi(.;\mu,\Sigma)$ *be the normal distribution with mean* μ *and variance* Σ.
(i) As $m \to \infty$,

$$n^{1/2} W_2(\overline{\Pi}_n(\xi|X), \Phi[\xi;\bar{\xi},\{nI_\xi(\theta_0)\}^{-1}]) \to 0,$$
$$n^{1/2} W_2(\Pi_n(\xi|X), \Phi[\xi;\hat{\xi},\{nI_\xi(\theta_0)\}^{-1}]) \to 0,$$
$$m^{1/2} W_2\{\overline{\Pi}_n(\xi|X), \Pi_n(\xi|X)\} \to 0,$$

where the convergence is in $P_{\theta_0} -$ *probability*, $\hat{\theta}$ *is the mle of* θ *based on the full dataset* X.

Main Results

Theorem (1 cont.)

(ii) If $\hat{\theta}_1$ is an unbiased estimator for θ, then as $m \to \infty$,

$$n^{1/2} W_2\{\overline{\Pi}_n(\xi|X), \Pi_n(\xi|X)\} \to 0$$

in P_{θ_0} – probability

Main Results

Theorem (2)

Suppose Assumptions 1-7 hold. Let $\xi_0 = a^T\theta_0) + b$ *and* $\hat{\xi}$ *be the same as defined in Theorem 1. For a generic distribution F on* Ξ*, let* $bias(F) = E_F(\xi) - \xi_0$ *and var(F) be the variance of F. Let u be any fixed number in (0,1). Then the following relation hold:*

$$(i) bias\{\overline{\Pi}_n(\xi|X)\} = \overline{\xi} - \xi_0 + o_p(n^{-\frac{1}{2}}), bias\{\Pi_n(\xi|X)\} = \hat{\xi} - \xi_0 + o_p(n^{-\frac{1}{2}})$$

$$(ii) var\{\overline{\Pi}_n(\xi|X)\} = \frac{1}{n} I_\xi^{-1}(\theta_0) + o_p(n^{-1}), var\{\Pi_n(\xi|X)\} = \frac{1}{n} I_\xi^{-1}(\theta_0) + o_p(n^{-$$

$$(iii) \overline{\Pi}_n^{-1} - \Pi_n^{-1} = \overline{\xi} - \hat{\xi} + o_p(n^{-\frac{1}{2}}) = o_p(m^{-\frac{1}{2}})$$

where o_p *is in* $P_{\theta_0} - probability$*.*

Main Results

Theorem (2 conts)

Furthermore, if $\hat{\theta}_1$ is an unbiased estimator of θ_0), then

$$bias\{\overline{\Pi}_n(\xi|X)\} - bias\{\Pi_n(\xi|X)\} = o_p(n^{-\frac{1}{2}})$$
$$\overline{\Pi}_n^{-1} - \Pi_n^{-1} = o_p(n^{-\frac{1}{2}})$$

Experiments

- Use posterior summaries from MCMC applied to the full data as the benchmark for comparisons.
- All sampling algorithms were run for 10,000 iterations. After 5000 burn-in, retain every fifth sample in all the chains.
- Example 1: Linear model with varying dimension

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I_n), \beta \sim gdP, \sigma \sim Half - t$$

- Apply their approach for inference on β compared with an asymptotic normal approximation.

Experiments

- Use posterior summaries from MCMC applied to the full data as the benchmark for comparisons.
- All sampling algorithms were run for 10,000 iterations. After 5000 burn-in, retain every fifth sample in all the chains.
- Example 1: Linear model with varying dimension

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I_n), \beta \sim gdP, \sigma \sim Half - t$$

- Apply their approach for inference on β compared with an asymptotic normal approximation.

Experiments

Table 1: Accuracy of approximate posteriors for the non-zero and zero elements of β in (5). The accuracies are averaged over 10 simulation replications. Normal, the asymptotic normal approximation based on the full data; PIE, our posterior interval estimation algorithm; NB, the W_2 barycenter of K asymptotic normal approximations of subset posteriors.

	$p = 10$				$p = 100$				$p = 200$			
	$n = 10^4$		$n = 10^5$		$n = 10^4$		$n = 10^5$		$n = 10^4$		$n = 10^5$	
	0s	non-0s	0s	non-0s	0s	non-0s	0s	non-0s	0s	non-0s	0s	non-0s
Normal	0.95	0.89	0.96	0.96	0.95	0.90	0.96	0.95	0.95	0.89	0.96	0.95
NB (K=10)	0.94	0.91	0.96	0.95	0.89	0.87	0.95	0.94	0.84	0.83	0.94	0.94
PIE (K=10)	0.95	0.97	0.97	0.97	0.90	0.92	0.96	0.96	0.85	0.85	0.95	0.95
NB (K=20)	0.93	0.92	0.96	0.95	0.84	0.84	0.94	0.93	0.75	0.76	0.92	0.92
PIE (K=20)	0.94	0.97	0.97	0.97	0.85	0.87	0.95	0.95	0.77	0.78	0.93	0.93
	$p = 300$				$p = 400$				$p = 500$			
	$n = 10^4$		$n = 10^5$		$n = 10^4$		$n = 10^5$		$n = 10^4$		$n = 10^5$	
	0s	non-0s	0s	non-0s	0s	non-0s	0s	non-0s	0s	non-0s	0s	non-0s
Normal	0.95	0.89	0.96	0.95	0.94	0.89	0.96	0.95	0.94	0.89	0.96	0.95
NB (K=10)	0.80	0.79	0.93	0.93	0.75	0.75	0.93	0.92	0.71	0.71	0.92	0.91
PIE (K=10)	0.82	0.81	0.94	0.94	0.77	0.78	0.93	0.93	0.73	0.74	0.93	0.93
NB (K=20)	0.65	0.67	0.91	0.91	0.51	0.52	0.90	0.90	-	-	0.89	0.88
PIE (K=20)	0.67	0.68	0.92	0.91	0.52	0.53	0.91	0.91	0.31	0.31	0.90	0.89

Experiments

- Example 2: Linear mixed effects model

$$y_i \sim N(X_i\beta + Z_i\mu_i, \sigma^2 I_{n_i}), \mu_i \sim N(0, \Sigma), i = 1, \ldots, s.$$

Table 2: 90% credible intervals for covariance matrix of random effects in simulation for linear mixed effects model. The upper and lower bounds are averaged over 10 replications. MLE, maximum likelihood estimator; MCMC, Markov chain Monte Carlo based on the full data; VB, variational Bayes; WASP, the algorithm in [19]; PIE, our posterior interval estimation algorithm.

	Σ_{11}	Σ_{22}	Σ_{33}	Σ_{12}	Σ_{13}	Σ_{23}
MLE	0.99	2	3	-0.57	0.52	0
MCMC	(0.96, 1.03)	(1.94, 2.07)	(2.9, 3.1)	(-0.61, -0.53)	(0.48, 0.56)	(-0.06, 0.05)
VB	(0.9, 0.96)	(1.88, 2)	(2.84, 3.04)	(-0.61, -0.54)	(0.48, 0.56)	(-0.06, 0.05)
WASP	(0.96, 1.03)	(1.94, 2.07)	(2.9, 3.1)	(-0.61, -0.53)	(0.48, 0.56)	(-0.06, 0.06)
PIE	(0.96, 1.03)	(1.94, 2.07)	(2.9, 3.1)	(-0.61, -0.53)	(0.48, 0.56)	(-0.06, 0.06)

Experiments

Table 3: Accuracy of approximate posteriors for covariances of random effects in simulation for linear mixed effects model. The standard deviation of accuracy across 10 folds of cross-validation is in parentheses. VB, variational Bayes; WASP, the algorithm in [19]; PIE, our posterior interval estimation algorithm.

	Σ_{11}	Σ_{22}	Σ_{33}	Σ_{12}	Σ_{13}	Σ_{23}
VB	0.11 (0.01)	0.45 (0.02)	0.62 (0.02)	0.94 (0.02)	0.94 (0.01)	0.96 (0.01)
WASP	0.94 (0.03)	0.95 (0.02)	0.95 (0.01)	0.95 (0.02)	0.94 (0.02)	0.95 (0.02)
PIE	0.94 (0.02)	0.95 (0.02)	0.95 (0.01)	0.95 (0.02)	0.95 (0.02)	0.95 (0.01)

Experiments

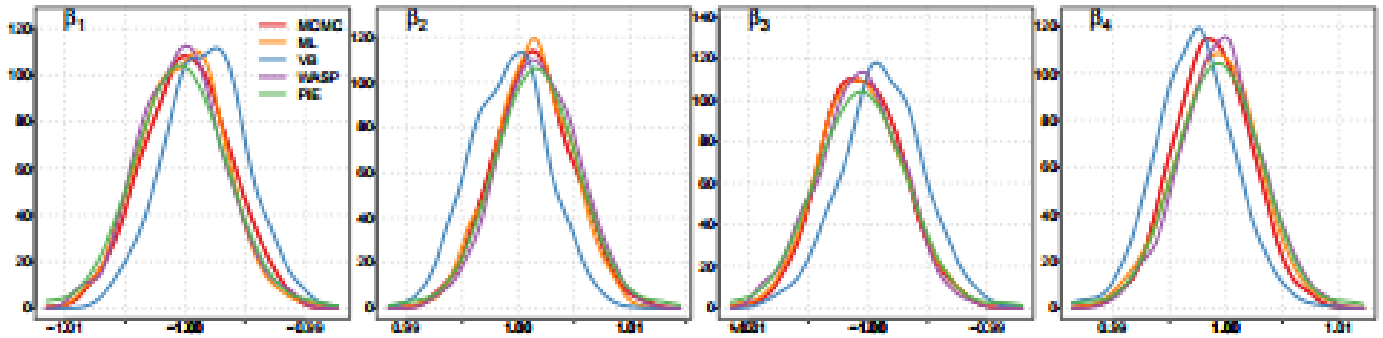


Figure 1: Posterior density plots for fixed effects in simulation for linear mixed effects model. MCMC, Markov chain Monte Carlo based on the full data; ML, maximum likelihood estimator; VB, variational Bayes; WASP, the algorithm in [19]; PIE, our posterior interval estimation algorithm.

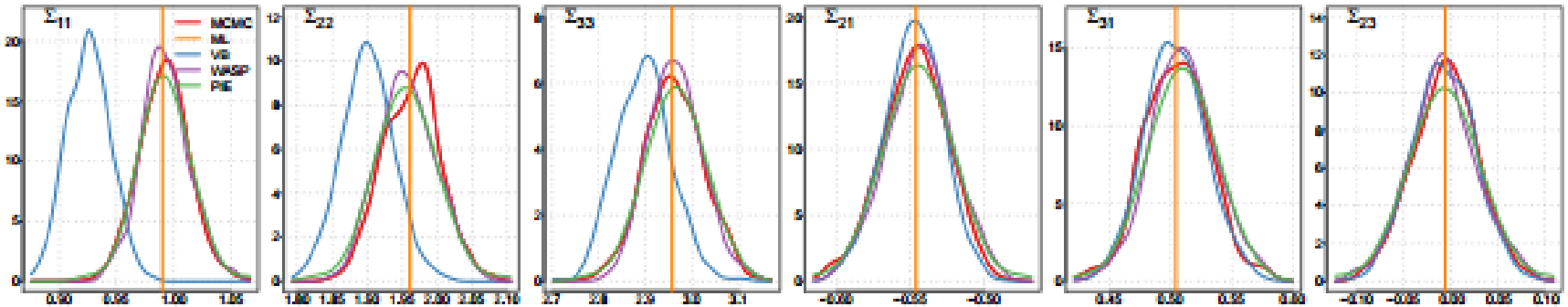


Figure 2: Posterior density plots for covariance matrix of random effects in simulation for linear mixed effects model. MCMC, Markov chain Monte Carlo based on the full data; ML, maximum likelihood estimator; VB, variational Bayes; WASP, the algorithm in [19]; PIE, our posterior interval estimation

Selected References

[1] Rmi Bardenet and Odalric-Ambrym Maillard (2015)

Concentration inequalities for sampling without replacement.

Bernoulli Volume 21, Number 3 (2015), 1361-1385.

[2] W. Neiswanger et al. (2014)

Asymptotically Exact, Embarrassingly Parallel MCMC.

[3] Minker et al.(2014)

Scalable and robust Bayesian inference via the median posterior.

Proceedings of the 31st International Conference on Machine Learning(ICML) 32, 1656-1664.

[4] Srivastava et al. (2015)

WASP: Scalable Bayes via barycenters of subset posteriors.

Proceedings of the 18th International Conference on Artificial Intelligence and Statistics(ATSTATS) 38, 912-920.

[5] Martial Agueh and Guillaume Carlier (2011)

Barycenters in the Wasserstein Space.

SIAM J. Math. Anal. 43(2), 904924.