Output Analysis for High-Dimensional Markov Chain Monte Carlo

Dootika Vats (vatsx007@umn.edu), Galin Jones
University of Minnesota
School of Statistics

James Flegal
University of California, Riverside
Department of Statistics

Markov Chain Monte Carlo

Let F be a probability density and g be a function on $\mathbb{R}^p$, so that $\theta = \mathrm{E}_F g$ is of interest. We construct a Harris ergodic Markov chain $\{X_1, X_2, X_3, \dots\}$ with invariant distribution F and θ is estimated by

$$\theta_n = \frac{1}{n}\sum_{t=1}^{n} g_t,$$

where $g_t = g(X_t)$. Under certain conditions, a Markov chain CLT of the following form exists

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{d} N_p(0, \Sigma). \tag{1}$$

Goal: Estimate Σ consistently.

Multivariate Batch Means

Let $n = a_n b_n$, where a_n = number of batches, and b_n = batch size. We find the mean of batch k, $\bar{Y}_k$.

$$\underbrace{g_1, \dots, g_{b_n}}_{\bar{Y}_1},\ \underbrace{g_{b_n+1}, \dots, g_{2b_n}}_{\bar{Y}_2},\ \dots,\ \underbrace{g_{n-b_n+1}, \dots, g_n}_{\bar{Y}_{a_n}}$$

The ***multivariate batch means*** **(mBM)** estimator is defined as

$$\Sigma_n = \frac{b_n}{a_n - 1}\sum_{k=1}^{a_n}(\bar{Y}_k - \theta_n)(\bar{Y}_k - \theta_n)^T.$$

In [1], we show that Σ_n is strongly consistent under conditions on the Markov chain, b_n and g.

Stopping Rules

Using Σ_n we obtain the $100(1-\alpha)\%$ confidence region for θ

$$R_{\alpha,n} = \left\{\theta \in \mathbb{R}^p : n(\theta_n - \theta)^T \Sigma_n^{-1}(\theta_n - \theta) \le T^2_{\alpha,p,a_n-1}\right\}.$$

The ***relative standard deviation fixed volume sequential stopping rule*** stops simulation the first time

$$\mathrm{Vol}(R_{\alpha,n})^{1/p} + \epsilon|\Lambda_n|^{1/2p} I(n < n^*) + n^{-1} \le \epsilon|\Lambda_n|^{1/2p}$$

where n^* is the minimum simulation effort, Λ_n is the sample covariance matrix of g_t, ϵ is the tolerance level.

The resulting confidence regions are asymptotically valid. We show that this termination rule is asymptotically equivalent to terminating using effective sample size.

Multivariate Effective Sample Size (ESS)

ESS is the number of iid samples with the same standard error as this sample. Only a univariate definition of ESS exists, and thus ESS is calculated for *each* component separately. Let $|\cdot|$ be the determinant, $\Lambda = \mathrm{Var}_F g(X_1)$, and λ_i^2, σ_i^2 be the ith diagonal of Λ and Σ

Univariate	Multivariate Extension
$\mathrm{ESS}_i = n\dfrac{\lambda_i^2}{\sigma_i^2}$	$\mathrm{mESS} = n\left(\dfrac{\lvert\Lambda\rvert}{\lvert\Sigma\rvert}\right)^{1/p}$

Equivalence

Stopping the first time that

$$\mathrm{Vol}(R_{\alpha,n})^{1/p} + \epsilon|\Lambda_n|^{1/2p} I(n < n^*) + n^{-1} \le \epsilon|\Lambda_n|^{1/2p}$$

is asymptotically equivalent to stopping when

$$\widehat{\mathrm{mESS}} \ge \frac{2^{2/p}\pi}{(p\Gamma(p/2))^{2/p}}\frac{\chi^2_{1-\alpha,p}}{\epsilon^2}.$$

This lower bound is only a function of the dimension of the problem p, the tolerance level ϵ and the desired confidence level.

Example: To achieve Monte Carlo error of at most 5% of the posterior standard deviation with 95% confidence for $p = 5$ we need $\widehat{\mathrm{mESS}} \ge 8605$.

Example: Bayesian Logistics Regression

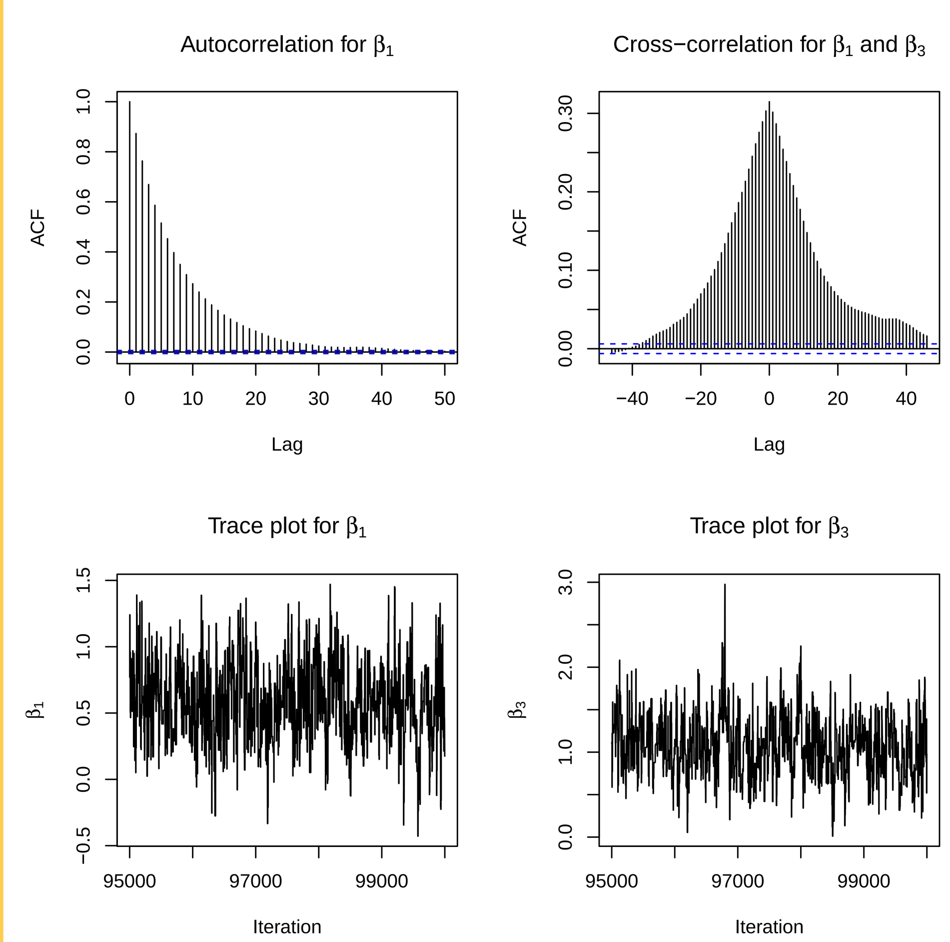
Consider the `logit` dataset in the `R` package `mcmc`. The response is binary with four covariates and 100 observations. We fit the following Bayesian logistic regression model. For $i = 1, \dots, 100$

$$Y_i | x_i, \beta \overset{ind}{\sim} \text{Bernoulli}\left(\frac{1}{1+e^{-x_i^T\beta}}\right), \quad \text{and} \quad \beta \sim N_5(0, I_5) \qquad \text{Posterior: } f(\beta|y) \propto \left(\prod_{i=1}^{100}\frac{e^{-y_i x_i^T\beta}}{1+e^{-x_i^T\beta}}\right)e^{-\frac{1}{2}\beta^T\beta} \quad \text{is intractable}$$

Goal is to estimate posterior mean $\theta = \mathrm{E}_F\beta$. We implement a random walk Metropolis-Hastings algorithm with proposal distribution $N_5(0, .35I_5)$.

Example: Confidence Regions and Termination

The autocorrelation plot shows significant autocorrelation for β_1. This autocorrelation is captured by both univariate and multivariate methods. The cross-correlation plot of β_1 and β_3 shows the correlation ignored by univariate methods and captured by multivariate methods.

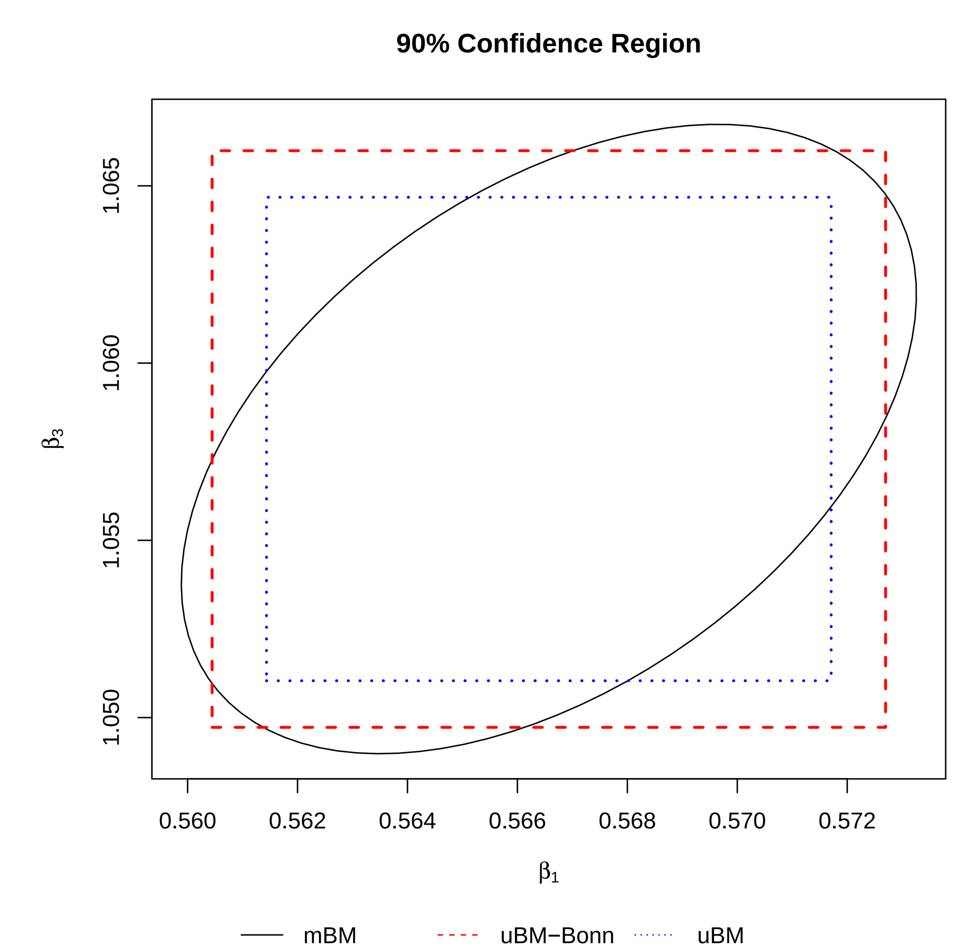


Method	Cov Prob.	Volume$^{1/5}$
mBM	0.880 (0.0103)	0.020 (1.20e-05)
uBM-Bonn	0.910 (0.0090)	0.021 (1.42e-05)
uBM	0.578 (0.0156)	0.015 (1.00e-05)

90% confidence regions for β when $n = 10^5$ over 1000 replications.

ϵ	mBM	uBM-Bonn
$\epsilon = 0.05$	133005 (196)	201497 (391)
$\epsilon = 0.02$	844082 (1158)	1262194 (1880)

Iterations to termination over 1000 replications.



We compare 90% confidence ellipsoids obtained through the mBM with the Bonferroni corrected and uncorrected confidence cuboids obtained through univariate batch means. Termination is earlier using mBM because of

- accounting for the multivariate nature of the samples,
- conservative behavior of the Bonferroni correction,
- the additional conservative step of choosing the smallest univariate effective sample size.

Sample Size	mESS	ESS_1	ESS_2	ESS_3	ESS_4	ESS_5
10^5	6209	6796	5164	6152	6323	4781

Conclusion

- MCMC samples are inherently multivariate, and thus their treatment should be multivariate whenever possible.
- Uncorrected confidence regions have bad coverage probabilities.
- Similar/better results were seen in more complicated models including Bayesian Lasso, Bayesian dynamic spatial temporal models.
- R package `mcmcse` includes mBM estimator, multivariate spectral variance estimator, and mESS estimator.

References

[1] Dootika Vats, James M Flegal, and Galin L Jones. Multivariate output analysis for Markov chain Monte Carlo. *arXiv:1512.07713*, 2015.