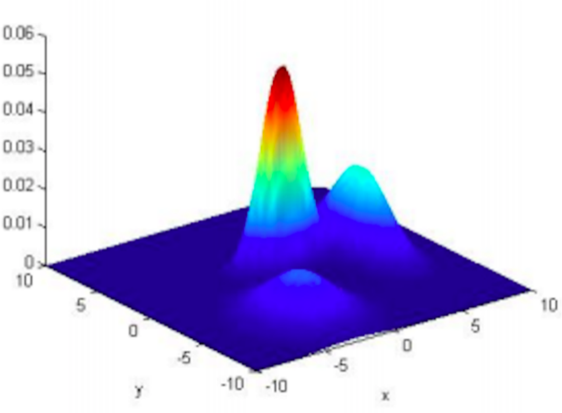
Inference from iterative simulation using multiple sequences

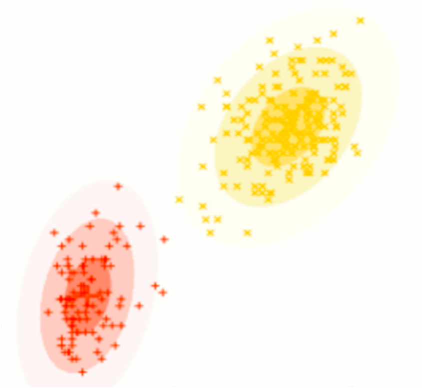
Yaxiong (Jason) Cai, Shiya Wang, Dustin Tran
Statistics Department, Harvard University

April 2, 2015

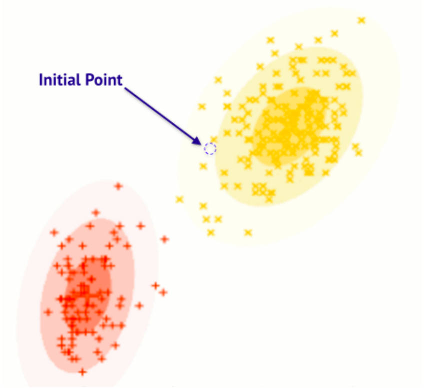
Motivation



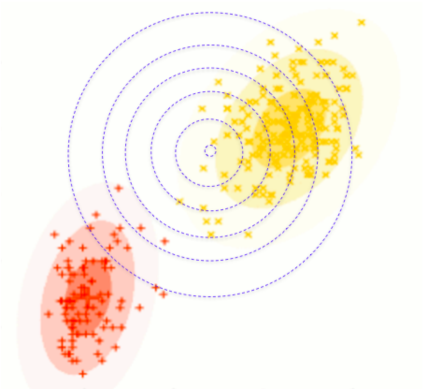
Motivation



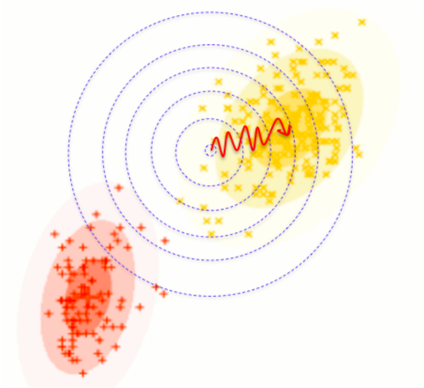
Motivation



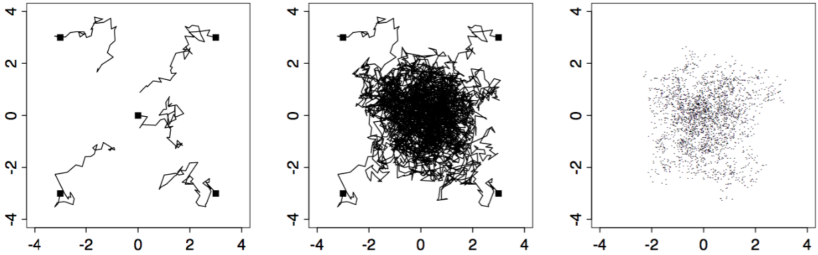
Motivation



Motivation



Iterative simulation using multiple sequences



Choosing initial points

1. Locate the modes using EM
2. Sample N points from a mixture of t-distributions centered at the modes
3. Sharpen the overdispersed approximation using sequential importance resampling (SIR; Rubin, 1987)
   - Gives m samples, one for starting each chain, from the set of N

Monitoring convergence

1. Gelman-Rubin diagnostic ($\hat{R}$)
2. Effective number of simulation draws ($\hat{n}_{eff}$)

Other diagnostics:

- Visuals: traceplots, autocorrelation plots, empirical density plots
- Acceptance rate
- Heidelberg and Welch diagnostic (1983)
- Geweke diagnostic (1992)
- Raftery and Lewis diagnostic (1992)

Procedure

- Run $m \geq 2$ sequences of length $2n$ from the starting values and discard the first n draws in each sequence

For each parameter of interest:

1. Calculate the between- and within-sequence variances
2. Estimate the total variance using a combining rule
3. Compute a ratio of variances $\sqrt{\hat{R}}$, known as the *potential scale reduction factor* (Gelman and Rubin, 1992)

Continue running chains until PSRF is sufficiently close to 1 for all (or "enough") estimands.

Within-sequence variance

For each scalar estimand ψ, let ψ_{ij} be the i^{th} estimand (p estimands) from the j^{th} chain (m chains)

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2$$

where

$$s_j^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\psi_{ij} - \bar{\psi}_{.j})^2, \ \bar{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij}$$

Between-sequence variance

$$B/n = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{.j} - \bar{\psi}_{..})^2$$

where

$$\bar{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{ij},\ \ \bar{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\psi_{.j}$$

Total variance

Estimate $\mathrm{var}(\psi|y)$, the marginal posterior variance of the estimator, by a weighted average of W and B

$$\hat{var}^{+}(\psi|y) = \frac{n-1}{n} W + B/n$$

- Consistency guarantee!

Total variance

Estimate $\mathrm{var}(\psi|y)$, the marginal posterior variance of the estimator, by a weighted average of W and B

$$\hat{var}^+(\psi|y) = \frac{n-1}{n}W + B/n$$

- Consistency guarantee!
- For finite n, $\hat{var}^+(\psi|y)$ will underestimate the true variance.

Using

$$\widehat{V} = \hat{var}^+(\psi|y) + (1/m)B/n$$

adjusts for the uncertainty in estimating $\bar{\psi}_{..}$ (Rubin, 1987; Ch. 4)

Potential scale reduction factor

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{V}}{W}}$$

- $\widehat{V}$ *overestimates* the true variance of the target distribution
- W *underestimates* the true variance of the target distribution
- If $\sqrt{\hat{R}}$ is high (say, greater than 1.1), then we have reason to believe that the chains have not converged

Evolution of $\hat{R}$

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{V}}{W}}$$ Gelman and Rubin (1992)

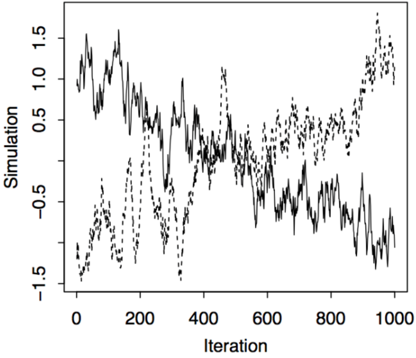
$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{var}^{+}(\psi|y)}{W}}$$ BDA1 (1995)

$$\hat{R}^{p} = \frac{n-1}{n} + \frac{m+1}{m} \max_{a} \frac{a^{T}\widehat{V}_{cov}a}{a^{T}W_{cov}a}$$ Brooks and Gelman (1998)

$$\hat{R}_{split} = \sqrt{\frac{\hat{var}^{+}(\psi|y)}{W}}$$ on *split chains*, BDA3 (2013)

Unsplit-$\hat{R}$ and split-$\hat{R}$

The unsplit-$\hat{R}$ would not correctly diagnose the poor convergence in nonstationarity as below:



Effective number of simulation draws

1. Account for ρ_t, autocorrelation of the sequence ψ at lag t
2. Define the effective sample size as:

$$n_{eff} = \frac{mn}{1+2\sum_{t=1}^{\infty} \rho_t}$$

3. Estimate autocorrelations:

$$\hat{\rho}_t = 1 - \frac{V_t}{2\widehat{var}^+(\psi|y)}$$

4. Compute a positive partial sum for autocorrection:

$$\hat{n}_{eff} = \frac{mn}{1+2\sum_{t=1}^{T} \hat{\rho}_t}$$

Where the Gelman-Rubin diagnostic fails

- Full coverage: If important areas of the target distribution were not captured by the starting distribution and are not easily reachable by the proposals...
- Computational efficiency: Given finite runtime constraints, how to assess the time it takes to get mn desired samples?
- Precognition: Can we assess the the rate of convergence before the actual convergence of the chains?

Other ideas: Markov-normal based analysis

(Liu and Rubin, 1996; Liu and Rubin, 2002)

Assumption: Target distribution of the simulated Markov chains is approximately normal.

1. Obtain ML estimates of the target distribution from multiple sequences *before their convergence*
2. Use pre-convergence ML estimates to *define a restarting distribution for the simulations*, *assess convergence rate*, and/or *analyze problematic areas in the simulation*

Other ideas: Markov-normal based analysis

Suppose all transition distributions for chains are

$$X_j^{(t)}|(X_j^{(t-1)}, \ldots, X_j^{(0)} \sim N_d(\beta X_j^{(t-1)} + \gamma, \Delta), t = 1, 2, \ldots$$

This is the $AR(1)$ multivariate time-series model, and the target distribution is

$$N_d(\beta\mu + \gamma, \beta\Psi\beta^T + \Delta)$$

Other ideas: Markov-normal based analysis

We have

$$(X^{(t)} - \mu) = \beta(X^{(t-1)} - \mu) + \epsilon^{(t)},\ \ \epsilon^{(t)} \sim N_d(0, \Delta)$$

- $X^{(t)}$'s are independent of $\epsilon^{(t)}$
- $\mathrm{Cov}(X^{(t)}, X^{(t-1)}) = \beta\Psi$
- Correlation of normalized values $\mathrm{Cor}(\Psi^{-1/2}X^{(t)}, \Psi^{-1/2}X^{(t-1)}) = \Psi^{-1/2}\beta\Psi^{1/2}$

Other ideas: Markov-normal based analysis

The rate of convergence of $P^{(t)}(X)$ to $P(X)$ is the *spectral radius* of the matrix β: $\max_{|\lambda|}\{|\lambda_i(\beta)|\}$. This can be difficult to compute (possibly complex number).

Alternatively consider the square root of the spectral radius ρ of the PSD matrix

$$M = (\mathrm{Cor}(\Psi^{-1/2}X^{(t)}, \Psi^{-1/2}X^{(t-1)}))(\mathrm{Cor}(\Psi^{-1/2}X^{(t)}, \Psi^{-1/2}X^{(t-1)}))^T$$
$$= \Psi^{-1/2}\beta\Psi\beta^T\Psi^{-1/2}$$

Other ideas: Markov-normal based analysis

Consider

$$S = I - M = \Psi^{-1/2} \Delta \Psi^{-1/2}$$

the *squared speed matrix* of the underlying MCMC scheme.

- S is related to the fraction of observed information: $\operatorname{Var}(X) / \operatorname{Var}(X | X^{(t)})$
- Intuition: the larger the within-variance covariance matrix Δ with respect to the target covariance matrix Ψ, the faster the algorithm converges because the fraction of missing information is less

Other ideas: Markov-normal based analysis

Let $Z^{(t)} = \Psi^{-1/2}(X^{(t)} - \mu)$. Then $Z^{(t)}|Z^{(t-1)} \sim N_d(0, S)$.

- If the eigenvalues of S are all 1, converges in one step; if zero, then MCMC scheme never converges
- The eigenvalues govern the rate of convergence in each dimension, and the minimal eigenvalue is the slowest

$$\rho = (1 - \lambda_1)^{1/2}$$

References I

C. Liu and D. B. Rubin, "Model-based analysis to improve the performance of iterative simulations," *Statistica Sinica*, vol. 12, no. 3, pp. 751–767, 2002.

C. Liu and D. B. Rubin., "Markov-normal analysis of iterative simulations before their convergence," *Journal of Econometrics*, vol. 75, no. 1, pp. 69–78, 1996.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, 2013.

References II

S. P. Brooks and A. Gelman, “General methods for monitoring convergence of iterative simulations,” *Journal of Computational and Graphical Statistics*, vol. 7, no. 4, pp. 434–455, 1998.

D. B. Rubin, “A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm,” *Journal of American Statistical Association*, no. 82, pp. 543–546, 1987.

D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

A. Gelman and D. Rubin, “Inference from iterative simulation using multiple sequences,” *Statistical Science*, vol. 7, pp. 457–511, 1992.