Application of Rao-Blackwell Theorem to Variance Reduction in Markov Chain Monte Carlo Simulation

Di Wang

EN 660.664

Rao-Blackwell Theorem

- **Definition.** Averaging with respect to a sufficient statistic does not cause increase of risk of an estimator with respect to an arbitrary continuous convex loss function (V. G. Voinov, M. S. Nikulin, 1989).
- **Intuition.**

$$E\big[E[h(X)|T]\big] = E[h(X)]$$

$$var(h(X)) = var[E\{h(X)|\underbrace{T}_{\text{sufficient}}\}] + \underbrace{E\{var[h(X)|T]\}}_{>0}$$

Sufficiency

- A statistic $T = T(\mathbb{X})$ is sufficient for a parameter θ if the conditional distribution for any fixed value $T(\mathbb{X}) = t$ of a random vector $\mathbb{X}$ does not depend on θ.

Metropolis-Hastings models - Independence Sampler

- Original

$$\tau = \frac{1}{n+1}\sum_{i=0}^{n} h(Z_i)$$

- Rao-Blackwellised:

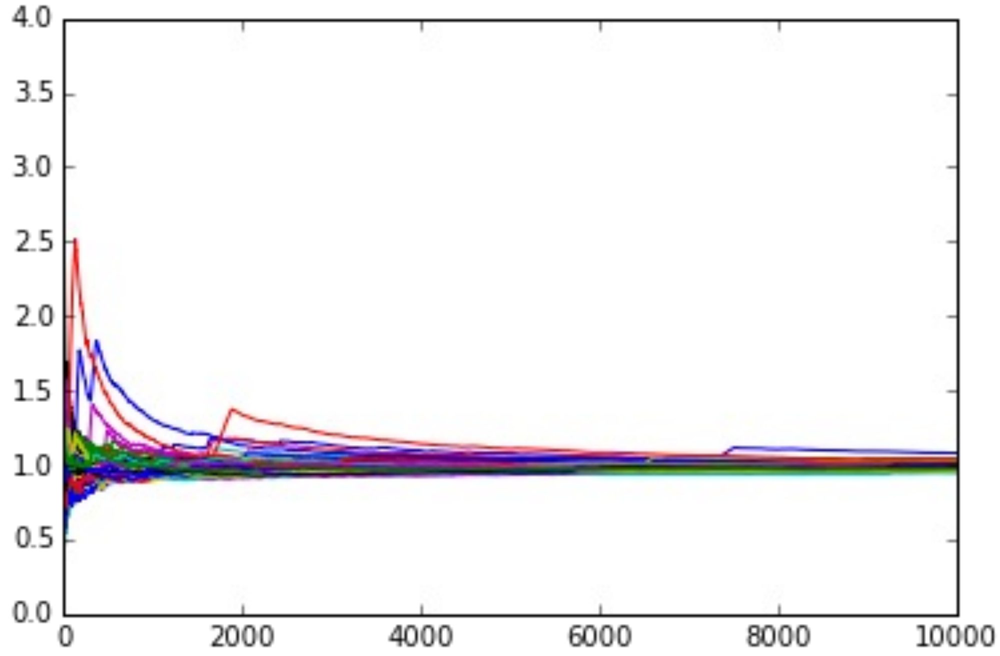
$$\tau' = \frac{1}{n+1}\sum_{i=0}^{n} \varphi_i h(Y_i)\,,$$

With

$$\varphi_i = \delta_i \textstyle\sum_{j=i}^{n} \xi_{ij}$$

Metropolis-Hastings models - Independence Sampler

- Implementation:
  - Target: $X \sim Gamma(2.0, 0.5)$
  - Proposal: $N(1.0, 0.5)$
  - $h(X_i) = E[X] = 1.0$



M-H sampling to $Gamma(2.0, 0.5)$ with normal proposal distribution

Metropolis-Hastings models - Independence Sampler

	Sample Mean (RB Metro)	Sample Mean(Metrop)
N=10	1.05074222067	1.00505328912
N=15	1.11333643423	1.03013167617

	Sample Var (RB Metro)	Sample Var (Metrop)
N=10	0.220772540984	0.235231103452
N=15	0.179379932536	0.22260909921

Metropolis-Hastings models – Random Walk

- $Y_t = X_t + \varepsilon_t$ where $\varepsilon_t \sim g(\sigma)$ -> spherically symmetric
- $X_{t+1} = \begin{cases} Y_t & if\ U \leq f(Y_t)/f(X_t) \\ X_t & other\ wise \end{cases}$

Or

- $Y_t \sim N(X_t, \sigma\mathbb{I})$ or $Y_t \sim U(X_t - \delta\mathbb{I}, X_t + \delta\mathbb{I})$

Metropolis-Hastings models – Random Walk

- Original:

$$\tau = \frac{1}{n+1}\sum_{i=0}^{n} X_i$$

- Rao-Blackwellised:

$$\tau' = X_0 + \sum_{i=0}^{n}(n+1-i)\varepsilon_i * pr\{D_i = \varepsilon_i \mid \varepsilon_0, \varepsilon_1, .. \varepsilon_i\}$$

where $pr\{D_i = \varepsilon_i \mid \varepsilon_0, \varepsilon_1, .. \varepsilon_i\} = \{\frac{f(X_i+\varepsilon_i)}{f(X_i)} \mid \varepsilon_0, \varepsilon_1, .. \varepsilon_i\}$

- Independent case: $\gamma(\varepsilon_i)$

Metropolis-Hastings models – Random Walk

- General case: complex!
- $pr\{D_i = \varepsilon_i | \varepsilon_0, \varepsilon_1, .. \varepsilon_i, K\} = \frac{f(X_0 + \sum_K \varepsilon_k + \varepsilon_i)}{f(X_0)} \prod_{k \in K} \prod_{k_m \le j \le k_{m+1},} \{1 - \frac{f(X_0 + \sum_{k \le k_m} \varepsilon_k + \varepsilon_j)}{f(X_0 + \sum_{k \le k_m} \varepsilon_k)}\}$
- $\binom{i}{0} + \binom{i}{1} + \cdots + \binom{i}{i}$ terms like this in $pr\{D_i = \varepsilon_i | \varepsilon_0, \varepsilon_1, .. \varepsilon_i\}$!
- i ranges from 0 to n!

Gibbs Sampling

- Standard Gibbs Sampling

  $f(X_1|X_2{}^t, X_3{}^t)$ ->$f(X_2|X_1{}^{t+1}, X_3{}^t)$->$f(X_3|X_1{}^{t+1}, X_2{}^{t+1})$
- Collapsed Gibbs Sampling
  - $f(X_1|X_3{}^t)$ ->$f(X_2|X_1{}^{t+1})$ ->$f(X_3|X_2{}^{t+1})$
- Block Gibbs Sampling
  - $f\left(X_1\middle|X_2{}^t, X_3{}^t\right)$ -> $f(X_2, X_2|X_1{}^{t+1})$
- Partially Collapsed Gibbs Sampling
  - $f(X_1|X_2{}^t)$ ->$f(X_2|X_1{}^{t+1}, X_3{}^t)$->$f(X_3|X_1{}^{t+1}, X_2{}^{t+1})$

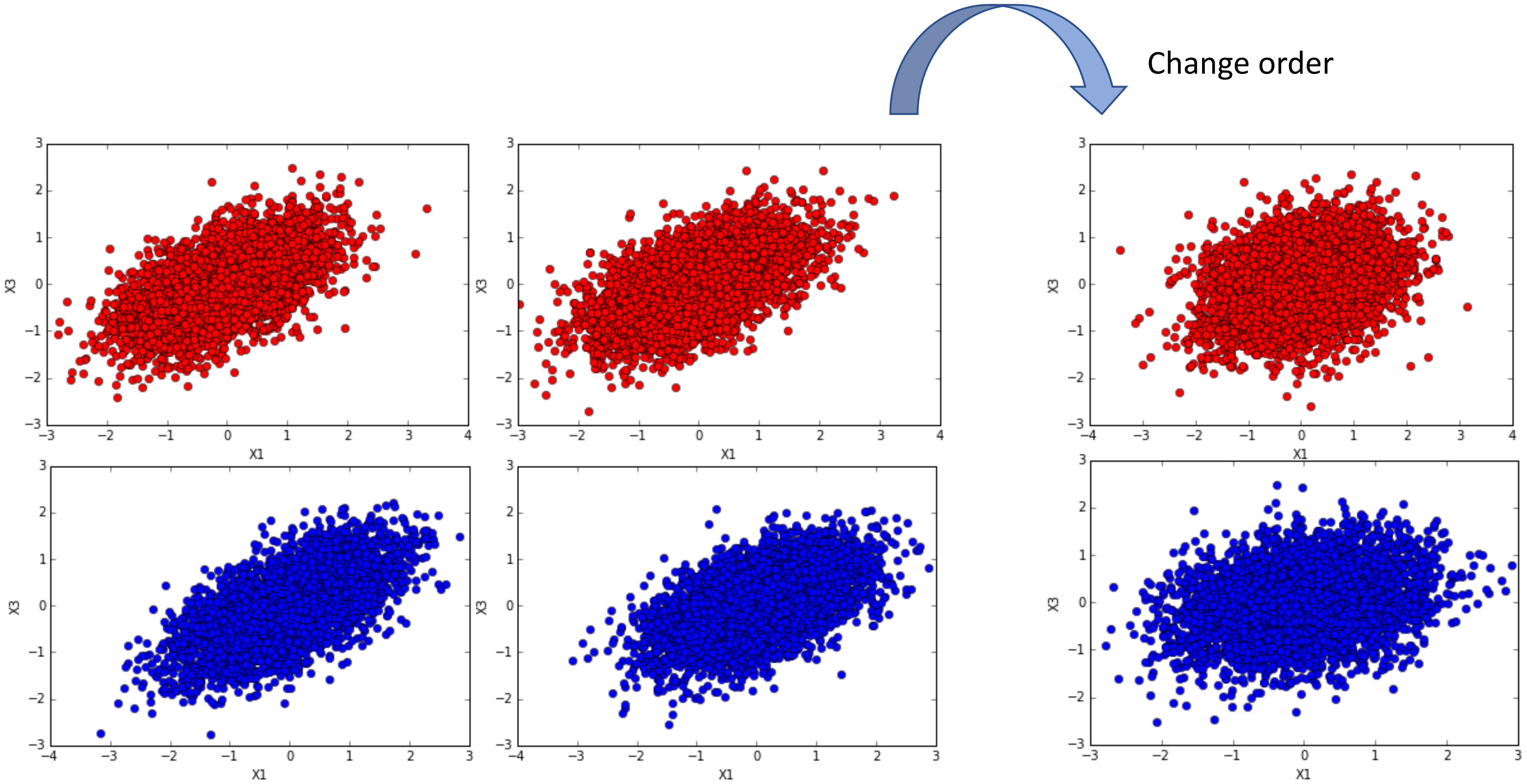
Rao-Blackwellised!

Standard vs Block vs Collapsed

Implementation: Multivariate Normal Dist.

	Standard	Block	Collapsed
Σ_1 Case	1.0920e-05	9.4915e-06	1.2331e-05
Σ_2 Case	4.2727e-05	5.4143e-05	2.4865e-05
Σ_3 Case	7.9752e-05	7.7695e-05	3.5266e-05

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1.0 & 0.5 \\ 0.5 & 0.5 & 1.0 \end{bmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1 & 0.5 & 0.9 \\ 0.5 & 1 & 0.5 \\ 0.9 & 0.5 & 1 \end{bmatrix}$$



standard

Partially collapsed
Order 1

Partially collapsed
Order 2

Discussion

- Advantage: reduce variance
- Disadvantage: computational complexity
- Future Improvements
  - Use Dynamic Programming or Data Structures to avoid repeated calculations
  - Combine with other variance reduction methods