Quantile estimation via Markov chain Monte Carlo

James M. Flegal

University of California, Riverside, CA

joint work with Galin L. Jones

August 19, 2010

General Setting

- Let π be a probability distribution we want to explore.
- Summarize π with expectations and quantiles.
- When i.i.d. observations unavailable, Markov chain is often used.

Expectations

- Consider estimating

$$E_\pi g := \int_{\mathcal{X}} g(x)\, \pi(dx)\ ,$$

  which is intractable.
- $E_\pi g$ is an unknown quantity I would like to estimate using simulated data.

- Let $X = \{X_1, X_2, \ldots, X_n\}$ be a Markov chain.
- Usually, $X_i \sim F_i \neq \pi$ and $\mathrm{Cov}(X_i, X_{i+1}) > 0$, but

$$\bar{g}_n := \frac{1}{n} \sum_{i=1}^{n} g(X_i) \overset{a.s.}{\to} E_\pi g \quad \text{as } n \to \infty \ (SLLN).$$

- Want $\bar{g}_n - E_\pi g$, the Monte Carlo error, to be small.

- Sampling distribution via a Markov chain CLT, that is

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} \mathrm{N}(0, \sigma_g^2) \quad \text{as } n \to \infty \,,$$

$$\sigma_g^2 = \mathrm{Var}_\pi\{g(X_1)\} + 2\sum_{s=1}^{\infty} \mathrm{Cov}_\pi\{g(X_1), g(X_{1+s})\}.$$

MCMC Settings

- Let $X = \{X_1, X_2, \ldots, X_n\}$ be a Markov chain with state space $\mathcal{X}$ and invariant distribution π.
- Define the n-step Markov transition kernel as

$$P^n(x, A) = \Pr(X_{n+i} \in A | X_i = x) \, .$$

- As $n \to \infty$, a Markov chain CLT requires

$$\| P^n(x, \cdot) - \pi(\cdot) \| := \sup_{A \in \mathcal{B}} | P^n(x, A) - \pi(A) | \downarrow 0 \, ,$$

  where the rate of convergence is important.

X is *geometrically ergodic* if there exists a constant $0 < t < 1$ and a function $M : \mathcal{X} \mapsto [0, \infty)$ such that for any $x \in \mathcal{X}$,

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq M(x)t^n \ .$$

Markov Chain Assumptions

Suppose $X = \{X_0, X_1, X_2, X_3, \dots\}$ *is a Markov chain such that* X

- *has state space* $\mathcal{X}$ *and invariant distribution* π
- *and is geometrically ergodic.*

Now, we have one set of regularity conditions for a MC CLT.

- If our assumption holds and $E_\pi |g|^{2+\epsilon} < \infty$, then for any initial distribution, as $n \to \infty$,

$$\sqrt{n}(\bar{g}_n - E_\pi g) \xrightarrow{d} \mathrm{N}(0, \sigma_g^2) \ ,$$

where $\sigma_g^2 = \mathrm{Var}_\pi\{g(X_1)\} + 2\sum_{s=1}^{\infty} \mathrm{Cov}_\pi\{g(X_1), g(X_{1+s})\}$.

How can we estimate σ_g^2?

- Batch Means
- Overlapping Batch Means
- Spectral Variance Estimators
- Regeneration

Overlapping Batch Means Estimator

- Let b be the batch size resulting in $n - b + 1$ overlapping batches.
- Let $\bar{Y}_j(b) = b^{-1} \sum_{i=1}^{b} g(X_{j+i})$, then the OBM estimator of σ_g^2 is

$$\hat{\sigma}^2_{OBM} = \frac{nb}{(n-b)(n-b+1)} \sum_{j=0}^{n-b} (\bar{Y}_j(b) - \bar{Y}_n)^2 \ .$$

Theorem

Suppose our assumptions hold and $E_\pi |g|^{2+\delta+\epsilon} < \infty$. *Further assume* $b = \lfloor n^\nu \rfloor$ *for* $(1 + \delta/2)^{-1} < \nu < 3/4$; *Then for any initial distribution* $\hat{\sigma}^2_{OBM}$ *is strongly consistent for* σ_g^2.

Bayesian Probit Regression

- Suppose $Y_1, \ldots, Y_m$ are independent Bernoulli with $Pr(Y_i = 1) = \Phi(x_i^T \beta)$.
  - x_i is a $p \times 1$ vector of known covariates associated with Y_i.
  - β is a $p \times 1$ vector of unknown regression coefficients.
- Then for $y_i \in \{0, 1\}$

$$Pr(Y_1 = y_1, \ldots, Y_m = y_m | \beta) = \prod_{i=1}^{m} \Phi(x_i^T \beta)^{y_i} \left[1 - \Phi(x_i^T \beta)\right]^{1-y_i}.$$

- Bayesian inference on β with a flat prior (p-dimensional Lebesgue measure) is common resulting in

$$\pi(\beta | y) \propto \prod_{i=1}^{m} \Phi(x_i^T \beta)^{y_i} \left[1 - \Phi(x_i^T \beta)\right]^{1-y_i},$$

which under regularity conditions is proper.

- We will sample from $\pi(\beta|y)$ using the PX-DA algorithm of Liu and Wu (1999).
  1. Draw $z_1, \ldots, z_m$ independently with $z_i \sim TN(x_i^T\beta, 1, y_i)$.
  2. Draw $g^2 \sim \Gamma\left(\frac{m}{2}, \frac{1}{2}\sum_{i=1}^{m}\left[z_i - x_i^T(X^TX)^{-1}X^Tz\right]^2\right)$ and set $z' = (gz_1, \ldots, gz_m)^T$.
  3. Draw $\beta' \sim N\left((X^TX)^{-1}X^Tz', (X^TX)^{-1}\right)$.
- Roy and Hobert (2007) show conditions to ensure the chain is geometrically ergodic.
- Chen and Shao (2000) show conditions to ensure the appropriate moment conditions for estimating β.

van Dyk and Meng's Lupus Data (2001)

- Trying to predict the occurrence of latent membranous lupus nephritis.
  - x_{i1} is the difference between IgG3 and IgG4 (immunoglobulin G).
  - x_{i2} is IgA (immunoglobulin A).
  - y_i in an indicator for latent membranous lupus nephritis (1 for present).
- We consider the following Bayesian analysis using a flat prior

$$Pr(Y_i = 1) = \Phi\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right).$$

Simulation

Compare coverage probabilities of OLB and Regeneration based on 1000 independent replications.

- Regeneration
  - Starting from new regeneration, i.e. burn-in is not an issue.
  - Stopped the simulation after 50 regenerations, resulting in a mean simulation effort of 7.12e5 (3.2e3).
- Overlapping Batch Means
  - Starting from the maximum likelihood estimate of β given by $\hat{\beta} = (-1.778, 4.374, 2.482)$.
  - Used 7e5 iterations and $b = \lfloor n^{1/2} \rfloor$.

	β_0	β_1	β_2
$\hat{\beta}_i$	-3.0166 (1.18E-3)	6.9107 (2.26E-3)	3.9792 (1.47E-3)

Table: "Truth" based on 1E8 iterations.

	β_0	β_1	β_2
$\hat{\beta}_i$	-3.0166 (1.18E-3)	6.9107 (2.26E-3)	3.9792 (1.47E-3)

Table: "Truth" based on 1E8 iterations.

Method	β_0	β_1	β_2
OBM	0.942	0.942	0.948
RS	0.938	0.938	0.938

Table: Coverage probabilities for OBM and RS, MCSEs $\leq$ 7.6E-3.

Where Are We Now?

What about quantiles?

1. Quantile Estimator
2. Central Limit Theorem
3. Estimate Asymptotic Variance

Quantiles

Consider a univariate marginal distributions of π with c.d.f. F. The quantile function of F is the generalized inverse, i.e.

$$F^{-1}(q) = \inf\{y : F(y) \geq q\}.$$

- Let $Y := \{Y_1, \ldots, Y_n\}$ be the observed sample from F.
- Estimate F^{-1} with the empirical quantile function,

$$\mathbb{F}_n^{-1} = Y_{n(j)} \text{ for } q \in \left(\frac{j-1}{n}, \frac{j}{n}\right] ,$$

where $Y_{n(1)}, \ldots, Y_{n(n)}$ are the order statistics of the sample.

Define $g_n(y) = I\{Y_n \leq y\} - F(y)$ and covariance function

$$\Gamma(y) = E\left[(g_1(y))^2\right] + 2\sum_{n=2}^{\infty} E\left[g_1(y)g_n(y)\right].$$

Theorem

Suppose X is a geometrically ergodic Markov chain. Fix $0 < q < 1$ and further suppose F differentiable at $F^{-1}(q)$ with positive derivative $f\left(F^{-1}(q)\right)$, then

$$\sqrt{n}\left(\mathbb{F}_n^{-1}(q) - F^{-1}(q)\right) \xrightarrow{d} N(0, \gamma_q^2),$$

where $\gamma_q^2 = \Gamma\left(F^{-1}(q)\right) / f^2\left(F^{-1}(q)\right)$.

Subsampling Methods

- Let b be the batch size resulting in $n - b + 1$ overlapping batches.
- The ith batch is $\{Y_i, \ldots, Y_{i+b-1}\}$ with $\{Y^*_{b(1)}, \ldots, Y^*_{b(b)}\}$.
- Define the quantile based on the ith batch as

$$\phi^*_i = Y^*_{b(j)} \text{ where } \frac{j-1}{n} < q \leq \frac{j}{n} \text{ for } i = 1, \ldots, n - b + 1 \, .$$

- Estimate γ^2_q using

$$\hat{\gamma}^2_q = \frac{b}{n - b + 1} \sum_{i=1}^{n-b+1} (\phi^*_i - \bar{\phi}^*)^2 \, .$$

- Requires $b/n \to 0$ and $b \to \infty$ as $n \to \infty$.

Bayesian Probit Regression

- Recall van Dyk and Meng's Lupus Data (2001) model

$$Pr(Y_i = 1) = \Phi\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right).$$

- Goal is estimating median and 90% credible region β,

$$\Phi = \left(\phi^{(0)}_{.05}, \phi^{(0)}_{.5}, \phi^{(0)}_{.95}, \phi^{(1)}_{.05}, \phi^{(1)}_{.5}, \phi^{(1)}_{.95}, \phi^{(2)}_{.05}, \phi^{(2)}_{.5}, \phi^{(2)}_{.95}\right).$$

- 1000 independent replications.
- 25 regenerations resulting in 3.56E5 (2300).

q	0.05	0.5	0.95
β_0	-6.301 (8.38E-03)	-2.692 (4.01E-03)	-0.848 (2.18E-03)
β_1	2.809 (4.40E-03)	6.294 (7.69E-03)	13.136 (1.57E-02)
β_2	1.277 (2.82E-03)	3.575 (5.03E-03)	8.060 (1.02E-02)

Table: "Truth" based on 9E6 iterations.

q	0.05	0.5	0.95
β_0	-6.301 (8.38E-03)	-2.692 (4.01E-03)	-0.848 (2.18E-03)
β_1	2.809 (4.40E-03)	6.294 (7.69E-03)	13.136 (1.57E-02)
β_2	1.277 (2.82E-03)	3.575 (5.03E-03)	8.060 (1.02E-02)

Table: "Truth" based on 9E6 iterations.

	q	0.05	0.5	0.95
β_0	SBM	0.940	0.944	0.955
	RS	0.941	0.939	0.932
β_1	SBM	0.947	0.947	0.939
	RS	0.921	0.944	0.949
β_2	SBM	0.958	0.947	0.949
	RS	0.934	0.940	0.951

Table: Coverage probabilities, MCSEs between 6.3E-3 and 8.5E-3.

Summary

- Monte Carlo Standard Errors
  - Easy to calculate.
  - Useful for interpretation.
- Compare well to Regeneration