Control Variates

for Reversible MCMC Samplers

Petros Dellaportas and Yiannis Kontoyiannis

Athens U of Economics and Business

Control variates in simple (i.i.d.) Monte Carlo

Goal: Compute the expected value of some function F evaluated on i.i.d. samples $X_1, X_2, \ldots$

Idea: Variance of the standard ergodic averages $\frac{1}{n}\sum_{i=1}^{n} F(X_i)$ can be reduced by exploiting available zero-mean statistics

Modified estimators: If there is one or more functions $U_1, U_2, \ldots, U_k$ – the ***control variates*** – for which it is known that $E[U(X_i)] = 0$, then subtracting any linear combination

$$\frac{1}{n}\sum_{i=1}^{n}\Big[F(X_i) - \theta_1 U_1(X_i) - \theta_2 U_2(X_i) - \cdots - \theta_k U_k(X_i)\Big]$$

does *not* change the asymptotic mean

Practice: For the optimal choice of $\{\theta_j\}$, the variance is no larger than before and often much smaller. The optimal $\{\theta_j^*\}$ are usually estimated adaptively, based on the same samples

Control Variates for Markov chains

- Extension of the above methodology to estimators based on MCMC samples is limited
- Extensions include: Green and Han (1992), Barone and Frigessi (1989), Andradottir *et al.*(1993), Brooks & Gelman (1998), Robert & Casella (2004), Philippe & Robert (2001, 2004), Fan *et al.*(2006), Atchade & Perron (2005), Mira *et al.*(2003), Hammer and Hakon (2008), Henderson (1997), Henderson *et al.*(2003), Kim and Henderson (2007), Meyn (2006)

Two fundamental difficulties:

$\leadsto$ **$\{U_j\}$?** hard to find (nontrivial and useful) functions with known expectation wrt the stationary distribution of the chain

$\leadsto$ **$\{\theta_j\}$?** even in cases where control variates are available, no effective way to obtain useful estimates for the optimal coefficients $\{\theta_j^*\}$

Reason: This is a fundamentally difficult problem, because the MCMC variance of ergodic averages is *intrinsically an infinite-dimensional object*: It cannot be written in closed form as a function of the transition kernel and the stationary distribution

What we do [1/2]

Starting point: For *any* real-valued function G_j defined on the state space of a Markov chain $\{X_n\}$, the functions

$$U_j(x) := G_j(x) - E[G_j(X_{n+1})|X_n = x]$$

have zero mean with respect to the stationary distribution of the chain (Henderson,1997)

Estimating $\{\theta_j\}$: We use control variates of this form conjunction with a new, efficiently implementable and provably optimal estimator for the coefficients $\{\theta_j^*\}$ for reversible chains

- Our estimator for $\{\theta_j^*\}$ is adaptive, in the sense that is based on the same MCMC output
- Unlike the case of independent sampling where control variates need to be found in an *ad hoc* manner depending on the specific problem at hand, ***here the control variates (as well as the estimates of the corresponding optimal coefficients) come for free!***

What we do [2/2]

Choice of G: Identifying particular choices for the functions $\{G_j\}$ that lead to effective control variates $\{U_j\}$ in specific MCMC scenarios that arise from some of the most common families of Bayesian inference problems.

- Basic methodology: For an MCMC algorithm which simulates from $\pi(x) = \pi(x^{(1)}, x^{(2)}, \ldots, x^{(k)})$), use $G_j = x^{(j)}, j = 1, \ldots, k$; Control variates are constructed without any cost for nearly ALL random scan Gibbs samplers
- Extension 1: Use of a subset of $\{G_j\}$ functions
- Extension 2: General classes of basis functions G
- Extension 3: general statistics F

The setting [1/2]

- $\{X_n\}$ is a discrete-time Markov chain with initial state $X_0 = x$, and transition kernel P:

$$P(x, A) := \Pr\{X_{k+1} \in A \mid X_k = x\}, \quad \text{all } x,\ A$$

Typical application: Construct an easy-to-simulate Markov chain $\{X_n\}$ which has a target distribution π as its unique invariant measure

Ergodicity: If we write $PF(x) := E[F(X_1) \mid X_0 = x]$, then for appropriate F's:

$$P^n F(x) := E[F(X_n) \mid X_0 = x] \to \pi(F) := E_\pi[F(X)], \quad \text{as } n \to \infty$$

- Moreover, $\hat{F}(x) = \sum_{n=0}^{\infty} \left[P^n F(x) - \pi(F) \right]$ where $\hat{F}$ satisfies **the Poisson equation for F**:

$$\boldsymbol{P\hat{F} - \hat{F} = -F + \pi(F)}$$

The setting [2/2]

Ergodic averages: Estimate $\pi(F)$ by $\mu_n(F) := \frac{1}{n}\sum_{i=0}^{n-1} F(X_i)$

Ergodic theorem: $\mu_n(F) \to \pi(F)$, a.s., as $n \to \infty$, for appropriate F's

Central limit theorem:

$$\sqrt{n}[\mu_n(F) - \pi(F)] = \frac{1}{\sqrt{n}}\sum_{i=0}^{n-1}[F(X_i) - \pi(F)] \xrightarrow{\mathcal{D}} N(0, \sigma_F^2), \quad \text{as } n \to \infty$$

where σ_F^2, the asymptotic variance of F, is given by

$$\sigma_F^2 := \lim_{n\to\infty} \mathrm{Var}_\pi(\sqrt{n}\mu_n(F)) = \lim_{n\to\infty} \mathrm{Var}_\pi\left(\frac{1}{\sqrt{n}}\sum_{i=0}^{n-1} F(X_i)\right) = \sum_{n=-\infty}^{\infty} \mathrm{Cov}_\pi(F(X_0), F(X_n))$$

Asymptotic variance: An alternative and more useful representation is in terms of the solution

$\hat{F}$ to Poisson's equation:

$$\boldsymbol{\sigma_F^2 = \pi\left(\hat{F}^2 - (P\hat{F})^2\right)}$$

Construction of control variates for Markov chains

- Suppose the chain $\{X_n\}$ takes values in some space S, typically $S \subset \mathbb{R}^d$

Construction of U: Given *any* $G : S \to \mathbb{R}$ with $\pi(|G|) < \infty$, if we let

$$\boldsymbol{U(x) := G(x) - PG(x)} = G(x) - E[G(X_1)|X_0 = x]$$

then $\pi(U) := E_\pi[U(X)] = 0$

Modified Estimators: Given such a function U with $\pi(U) = 0$ and $\theta \in \mathbb{R}$, define

$$\begin{aligned} \boldsymbol{F_\theta} &= \boldsymbol{F - \theta U} \\ \boldsymbol{\mu_n(F_\theta)} &= \boldsymbol{\mu_n(F) - \theta\mu_n(U)} \end{aligned}$$

Goals: Search for particular choices for: (i) G (with corresponding $U = G - PG$);

(ii) θ, so that the asymptotic variance $\sigma^2_{F_\theta}$ of the modified estimators is significantly smaller

than the variance σ^2_F of the standard ergodic averages $\mu_n(F)$

Ideal U? Zero Variance?

First suppose we have complete freedom in the choice of G. Set $\theta = 1$ without loss of generality.

We wish to make the asymptotic variance of

$$F - U = F - G + PG$$

as small as possible. But, in view of the Poisson equation

$$P\hat{F} - \hat{F} = -F + \pi(F)$$

the choice $G = \hat{F}$ yields

$$F - U = F - \hat{F} + P\hat{F} = \pi(F)$$

which has *zero* variance! Therefore, our first rule of thumb for choosing G is:

Choose a control variate $U = G - PG$ with $G \approx \hat{F}$

After choosing G

- With a choice G that (we hope) approximates $\hat{F}$, we form the modified estimators $\mu_n(F_\theta)$ with respect to the function $F_\theta = F - \theta U = F - \theta G + \theta PG$

Next task: Choose θ: Minimize the resulting variance

$$\sigma_\theta^2 := \sigma_{F_\theta}^2 = \pi\left(\hat{F}_\theta^2 - (P\hat{F}_\theta)^2\right)$$

From the definitions, $\hat{U} = G$ and $\hat{F}_\theta = \hat{F} - \theta G$. Therefore,

$$\sigma_\theta^2 = \pi\left((\hat{F} - \theta G)^2\right) - \pi\left((P\hat{F} - \theta PG)^2\right)$$

Expanding the above quadratic in θ, the optimal value is

$$\theta^* = \frac{\pi(\hat{F}G - (P\hat{F})(PG))}{\pi(G^2 - (PG)^2)}$$

- Hard to estimate θ^* – it depends on $\hat{F}$

Interpretation of θ^*

$$\sigma_\theta^2 = \lim_{n\to\infty} \mathrm{Var}_\pi\left(\frac{1}{\sqrt{n}}\sum_{i=0}^{n-1}[F(X_i) - \theta U(X_i)]\right),$$

$$\sigma_\theta^2 = \sigma_F^2 + \theta^2\sigma_U^2 - 2\theta \sum_{n=-\infty}^{\infty} \mathrm{Cov}_\pi(F(X_0), U(X_n)),$$

so that θ^* can also be expressed as

$$\theta^* = \frac{1}{\sigma_U^2}\sum_{n=-\infty}^{\infty} \mathrm{Cov}_\pi(F(X_0), U(X_n))$$

leading to the optimal asymptotic variance

$$\sigma_{\theta^*}^2 = \sigma_F^2 - \frac{1}{\sigma_U^2}\left[\sum_{n=-\infty}^{\infty} \mathrm{Cov}_\pi(F(X_0), U(X_n))\right]^2$$

This leads to our second rule of thumb for selecting control variates:

Choose a control variate $U = G - PG$ so that U and F are highly correlated

A different representation of θ^*

$$\theta^* = \frac{\pi(\hat{F}G - (P\hat{F})(PG))}{\pi(G^2 - (PG)^2)}.$$

Since $\hat{U} = G$, the denominator is simply σ_U^2, and the fact that σ_U^2 is always nonnegative suggests that there should be a way to rewrite the expression $\pi(G^2 - (PG)^2)$ in the denominator of θ^* in a way which makes this nonnegativity obvious. Indeed:

Proposition.

$$\sigma_U^2 = \pi(G^2 - (PG)^2) = E_\pi\left[\Big(G(X_1) - PG(X_0)\Big)^2\right]$$

and

$$\theta^* = \frac{\pi(\hat{F}G - (P\hat{F})(PG))}{E_\pi\left[\Big(G(X_1) - PG(X_0)\Big)^2\right]}$$

Optimal empirical estimates

Theorem. If the chain $\{X_n\}$ is *reversible*, then the optimal coefficient θ^* for the control variate $U = G - PG$ can be expressed as

$$\theta^* = \theta^*_{\text{rev}} := \frac{\pi\big((F - \pi(F))(G + PG)\big)}{E_\pi\Big[\Big(G(X_1) - PG(X_0)\Big)^2\Big]}$$

Therefore, we can estimate:

$$\theta^* \quad \text{as} \quad \hat{\theta}_{n,\text{rev}} = \frac{\mu_n(F(G + PG)) - \mu_n(F)\mu_n(G + PG)}{\frac{1}{n}\sum_{i=0}^{n-1}(G(X_i) - PG(X_{i-1}))^2}$$

$$\pi(F) \quad \text{as} \quad \mu_{n,\text{rev}}(F) := \mu_n(F_{\hat{\theta}_{n,\text{rev}}}) = \mu_n(F - \hat{\theta}_{n,\text{rev}}U)$$

Key: Expressions do *not* involve the solution $\hat{F}$ to Poisson's equation

Proof

Let $\Delta = P - I$ denote the generator of a discrete time Markov chain $\{X_n\}$ with transition kernel P.

Reversibility $\iff$ Δ is a self-adjoint linear operator on the space $L_2(\pi)$:

$$\pi(F\,\Delta G) = \pi(\Delta F\, G), \qquad \text{for any two functions } \; F, G \in L_2(\pi)$$

Let $\bar{F} = F - \pi(F)$ denote the centered version of F, and recall that $\hat{F}$ solves Poisson's equation for F, so $P\hat{F} = \hat{F} - \bar{F}$. Therefore, the numerator in the expression for θ^* can be expressed as

$$\begin{aligned}
\pi\big(\hat{F}G - (P\hat{F})(PG)\big) &= \pi\big(\hat{F}G - (\hat{F} - \bar{F})(PG)\big) \\
&= \pi\big(\bar{F}PG - \hat{F}\Delta G\big) \\
&= \pi\big(\bar{F}PG - \Delta\hat{F}G\big) \\
&= \pi\big(\bar{F}PG + \bar{F}G\big) \\
&= \pi\big(\bar{F}(G + PG)\big)
\end{aligned}$$

$\square$

Generalisation

Let $\mathrm{K}(G)$ denote the covariance matrix of the random variables

$$Y_j := G_j(X_1) - PG_j(X_0), \quad j = 1, 2, \ldots, k,$$

where $X_0 \sim \pi$. Then the optimal coefficient vector θ^* can also be expressed as,

$$\theta^* = \mathrm{K}(G)^{-1}\pi(\hat{F}G - (P\hat{F})(PG)).$$

and assuming that the chain $\{X_n\}$ is reversible,

$$\theta^* = \theta^*_{\text{rev}} := \mathrm{K}(G)^{-1}\pi\big((F - \pi(F))(G + PG)\big),$$

$$\hat{\theta}_{n,\mathrm{K}} = \mathrm{K}_n(G)^{-1}[\mu_n(F(G + PG)) - \mu_n(F)\mu_n(G + PG)],$$

where the $k \times k$ matrix $\mathrm{K}_n(G)$ is defined by

$$(\mathrm{K}_n(G))_{ij} = \frac{1}{n}\sum_{t=0}^{n-1}(G_i(X_t) - PG_i(X_{t-1}))(G_j(X_t) - PG_j(X_{t-1})).$$

Normal posterior, random scan Gibbs

Theorem: Let $\{X_n\}$ denote the Markov chain constructed from the random-scan Gibbs sampler used to simulate from an arbitrary multivariate normal distribution $\pi \sim N(\mu, \Sigma)$ in $\mathbb{R}^k$. If the goal is to estimate the mean of the first component of π, then letting $F(x) = x^{(1)}$ for each $x = (x^{(1)}, x^{(2)}, \ldots, x^{(k)})^t \in \mathbb{R}^k$, the solution $\hat{F}$ of the Poisson equation for F can be expressed as linear combination of the basis functions $G_j(x) := x^{(j)}$, $x \in \mathbb{R}^k$, $1 \leq j \leq k$,

$$\hat{F} = \sum_{j=1}^{k} \theta_j G_j.$$

Moreover, writing $Q = \Sigma^{-1}$, the coefficient vector θ is given by the first row of the matrix $k(I - A)^{-1}$ where A has entries $A_{ij} = -Q_{ij}/Q_{ii}$, $1 \leq i \neq j \leq k$, $A_{ii} = 0$ for all i, and $(I - A)$ is always invertible.

Outline of the Basic Methodology

(i) Given:

- A multivariate posterior distribution $\pi(x) = \pi(x^{(1)}, x^{(2)}, \ldots, x^{(d)})$
- A reversible Markov chain $\{X_n\}$ with stationary distribution π
- A sample of length n from the chain $\{X_n\}$

(ii) Goal:

- Estimate the posterior mean $\mu^{(i)}$ of $x^{(i)}$

(iii) Define:

- $F(x) = x^{(i)}$
- Basis functions $G_j(x) = x^{(j)}$ for all components j

  for which $PG_j(x) = E[X_{n+1}^{(j)} \mid X_n = x]$ is computable in closed form
- The corresponding control variates $U_j = G_j - PG_j$

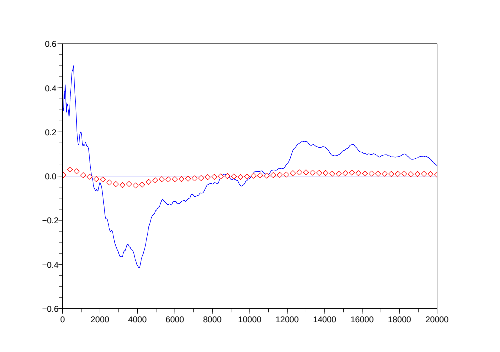
(iv) Estimate:

- The optimal coefficient vector θ^* by $\hat{\theta}_{n,\mathrm{K}}$
- The quantity of interest $\mu^{(i)}$ by the adaptive estimators $\mu_{n,\mathrm{K}}(F)$

Example: bivariate Gaussian [1/2]

- Let $(X, Y) \sim \pi(x, y)$ be an arbitrary bivariate normal distribution, with $E(X) = E(Y) = 0$, $Var(X) = 1$, $\mathrm{Var}(Y) = 10$ and $Corr(X, Y) = .99$.
- Random-scan Gibbs sampler, initial values $x_0 = y_0 = 0.5$
- $F(x, y) = x$, $G_1(x, y) = x$ and $G_2(x, y) = y$.
- $PG_1(x, y) = \frac{1}{2}\left[x + \frac{.99y}{10}\right]$ and $PG_2(x, y) = \frac{1}{2}(y + .99 \times 10x)$,

Example: bivariate Gaussian [2/2]



Variance reduction factors						
	Simulation steps					
Estimator	$n = 10^3$	$n = 10^4$	$n = 5 \times 10^4$	$n = 10^5$	$n = 2 \times 10^5$	$n = 5 \times 10^5$
$\mu_{n,\mathrm{K}}(F)$	4.13	27.91	122.4	262.5	445.0	1196.6

Example: hierarchical normal [1/2]

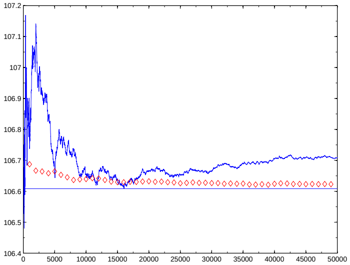
- $N = 5$ weekly weight measurements of $k = 30$ young rats whose weight is assumed to increase linearly in time (Gelfand, Smith and Hills, 1990, JASA)
- $Y_{ij} \sim N\left(\alpha_i + \beta_i x_{ij}, \sigma_c^2\right), \qquad 1 \le i \le k,\ 1 \le j \le N,$

$$\begin{aligned}
\phi_i &= \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N(\mu_c, \Sigma_c) \\
\mu_c &= \begin{pmatrix} \alpha_c \\ \beta_c \end{pmatrix} \sim N(\eta, C) \\
\Sigma_c^{-1} &\sim \mathrm{W}((\rho R)^{-1}, \rho) \\
\sigma_c^2 &\sim \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \tau_0^2}{2}\right),
\end{aligned}$$

with known values for η, C, ν_0, ρ, R and τ_0.

- The posterior has $2k + 2 + 3 + 1 = 66$ parameters
- random scan Gibbs samples from $((\phi_i), \mu_c, \Sigma_c, \sigma_c^2)$

Examples



Variance reduction factors						
Parameter	$n = 1000$	$n = 10000$	$n = 20000$	$n = 50000$	$n = 100000$	$n = 200000$
(ϕ_i)	1.59-3.58	9.12-31.02	11.73-61.08	10.04-81.36	12.44-85.99	9.38-109.2
α_c	2.99	15.49	32.28	31.14	28.82	36.48
β_c	3.05	19.96	34.05	39.22	32.33	36.04
Σ_c	1.15-1.38	4.92-5.74	5.36-7.60	3.88-5.12	4.91-5.34	3.65-6.50
σ_c^2	2.01	5.06	5.23	5.17	4.75	5.79

Example: Metropolis-within-Gibbs, heavy-tailed posterior

- Roberts and Rosenthal (2006, Can. J. of Stats, with discussion)
- N i.i.d. observations $x = (x_1, x_2, \ldots, x_N)$ are drawn from a $N(\phi, V)$
- $\phi \sim$ Cauchy$(0, 1)$, $V \sim$ IG$(1, 1)$

$$\pi(\phi|V, x) \quad \propto \quad \left(\frac{1}{1+\phi^2}\right) \exp\left\{ -\frac{1}{2V} \sum_i (\phi - x_i)^2 \right\},$$

$$\text{and} \quad \pi(V|\phi, x) \quad \sim \quad \mathrm{IG}\left(1 + \frac{N}{2}, 1 + \frac{1}{2} \sum_i (\phi - x_i)^2\right).$$

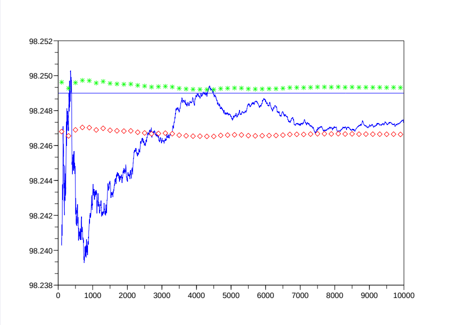
- Random scan: update V from its conditional (Gibbs step), or update ϕ in a random walk-Metropolis step with a $\phi' \sim N(\phi, 1)$ proposal, each case chosen with probability $1/2$.
- Simulate data of $N = 100$ i.i.d. $N(2, 4)$ observations, initial values $\phi_0 = 0$ and $V_0 = 1$.
- $F(\phi, V) = V$, $G(\phi, V) = V$.
- Variance reduction factors, estimated from $T = 100$ repetitions of the same experiment, are **7.89**, **7.48**, **10.46** and **8.54**, after $n = 10000$, 50000, 100000 and 200000 MCMC steps.

Non-conjugate Normal-Gamma

In the iid case, it is well known that the use of many control variates may be problematic since the variance increases due to the use of estimated coefficients; (see the notion of loss factors).

- Body temperature data, Mackowiak *et al.*(1992), JASA.
- $x = (x_1, x_2, \ldots, x_N) \sim$ i.i.d. $N(\mu, \sigma^2)$, $N = 130$.
- Priors $\mu \sim N(0, 100)$ and $\sigma^2 \sim \mathrm{IG}(0.001, 0.001)$
- $F(\mu, \gamma) = \mu$, random-scan Gibbs sampler

Example: Gaussian-Gamma posterior



Variance reduction factors						
	Simulation steps					
$\mu_{n,\mathrm{K}}(F)$	$n = 10^3$	$n = 10^4$	$n = 5 \times 10^4$	$n = 10^5$	$n = 2 \times 10^5$	$n = 5 \times 10^5$
G_1	1.40	18.15	49.06	1578.6	4474.6	69659
G_1, G_2	0.02	0.05	0.06	0.33	0.47	0.68

Model space search with Metropolis

- Two-threshold AR model, Data U.S. 3-month treasury bill rates 1962 – 1999

$$\Delta r_t = \left\{ \begin{array}{ll} \alpha_{10} + \alpha_{11} r_{t-1} & r_{t-1} < c_1 \\ \alpha_{20} + \alpha_{21} r_{t-1} & r_{t-1} \geq c_1 \end{array} \right\} + \left\{ \begin{array}{ll} \sigma \epsilon_t & r_{t-1} < c_2 \\ \sigma (1+\gamma)^{1/2} \epsilon_t & r_{t-1} \geq c_2 \end{array} \right\}, \quad (1)$$

  where $\gamma \geq -1$ characterizes the jump in σ^2 between the two volatility regimes.

- *Sampling*: 6-dim'al integration, and a discrete Metropolis-Hastings algorithm over (c_1, c_2) (we replace the 8-dimensional Gibbs sampler of Pfann *et al.*(1996, J of Econometrics) by a five-dimensional analytical integration over α and σ, a numerical integration over γ, and a Metropolis-Hastings algorithm over (c_1, c_2)).

- *Control variates*: Indicator functions of the three most likely models (c_1, c_2)
  Variance reduction factors: In estimating the posterior prob of MAP model, around **30-120**

A log-linear model

Data: $2 \times 3 \times 4$ table of Knuiman and Speed (1988): 491 subjects classified according to hypertension (yes, no), obesity (low, average, high) and alcohol consumption (0, 1-2, 3-5, or $6+$ drinks/day)

"Best" (main effects) model: $y_i \sim \text{Poisson}(\mu_i), \quad \log(\mu_i) = x_i^t \beta, \quad i = 1, 2 \ldots, 24$

Prior: Flat improper prior on $\beta \in \mathbb{R}^7$

Sampling: Standard Bayesian inference via MCMC performed either by a Gibbs sampler (full conditional densities are log-concave) or by a multivariate random walk Metropolis-Hastings sampler

Coplex G_j: A log-linear model

- *Sampling*: Here we use a simple random-scan Gibbs sampler, noting that a sample from the full conditional density of each β_j can be obtained directly as the logarithm of a

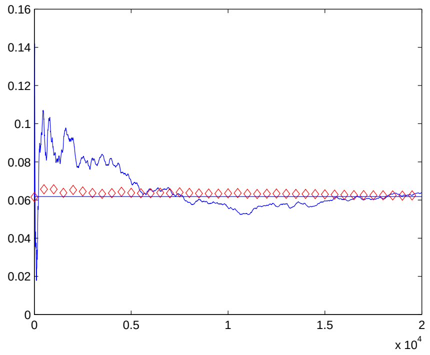
$$\text{Gamma}\left(\sum_i y_i x_{ij}, \sum_{i:x_{ij}=1} \exp\left\{\sum_{\ell \neq j} \beta_\ell x_{i\ell}\right\}\right) \quad \text{random variable}$$

- *Estimation*: To estimate the posterior means of the β_j, set $F_j(\beta) = \beta_j$ for each $j = 1, 2, \ldots, 7$ and use the same seven control variates $U_1, U_2, \ldots, U_7$ for each F_j, where each $U_\ell = G_\ell - PG_\ell$ is defined in terms of $G_\ell(\beta) = \exp(\beta_\ell)$
- *Computing PG*: The computation of PG_ℓ is straightforward, since the mean of $\exp(\beta_j)$ under the full conditional density of β_j is

$$\frac{\sum_i y_i x_{ij}}{\sum_{i:x_{ij}=1} \exp\left(\sum_{\ell \neq j} \beta_\ell x_{i\ell}\right)}$$

A log-linear model

The variance reduction factors obtained by our estimator $\mu_{n,\text{rev}}(F)$ for different parameters β_j are in the range **3.55–5.57, 38.2–57.69, 66.20–135.51, 57.16–170.34** and **85.41–179.11**, after $n = 1000$, 10000, 50000, 100000 and 200000 simulation steps, respectively



Coplex G_j: Gaussian mixtures

- Still numerous unresolved issues in inference for finite mixtures. Such models are often ill-posed or non-identifiable. Difficulties reflect important problems in prior specifications and label switching
- Improper priors are hard to use, and proper mixing over all (many!) posterior modes may require enforcing label-switching moves through Metropolis steps
- We begin with $N = 500$ data points $x = (x_1, x_2, \ldots, x_N)$ generated from the mixture $\frac{7}{10}N(0, \frac{1}{4}) + \frac{3}{10}(0.1, 9)$
- Assume the means, variances and mixing proportions are all unknown. Usual conjugate prior setting with non-informative priors based on Richardson and Green (1997)
- Impose *a priori* restriction $\mu_1 < \mu_2$
- To facilitate sampling from the posterior, introduce latent indicator variables $Z_1, Z_2, \ldots, Z_N$
- Problem: Estimate the two means μ_1, μ_2

Gaussian mixtures: Sampling

- Standard random-scan Gibbs sampler that selects one of the four parameter blocks (μ_1, μ_2), (σ_1, σ_2), Z or p, each with probability $1/4$
- Preferable to first obtain draws from the unconstrained posterior distribution and then to impose the identifiability (ordering) constraint at the post-processing stage
- The data x have been generated so that the two means are very close, which results in frequent label switching throughout the MCMC run and in near-identical (unordered) marginal densities of μ_1 and μ_2
- We perform a post-processing relabelling of the sampled values according to the above restriction, and we denote the ordered sampled vector by $(\mu_1^o, \mu_2^o, \sigma_1^o, \sigma_2^o, Z^o, p^o)$

Gaussian mixtures: Estimation

- In order to estimate the posterior mean of the smaller of the two means, we let,

$$F(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p) := \mu_1^o = \min\{\mu_1, \mu_2\}$$

- To reduce the variance of $\mu_n(F)$ we use a bivariate control variate $U = G - PG$, where $G = (G_1, G_2) = (\mu_1^o, \sigma_1^o)$
- $PG_1(\mu_1, \mu_2, \sigma_1, \sigma_2, Z, p)$ is the one-step expected value of $\min\{\mu_1, \mu_2\}$

$$\frac{3}{4}\mu_1^o + \frac{\nu_1}{4}\Phi\left(\frac{\nu_2 - \nu_1}{\sqrt{\tau_1^2 + \tau_2^2}}\right) + \frac{\nu_2}{4}\Phi\left(\frac{\nu_1 - \nu_2}{\sqrt{\tau_1^2 + \tau_2^2}}\right) - \frac{1}{4}\sqrt{\tau_1^2 + \tau_2^2}\,\phi\left(\frac{\nu_2 - \nu_1}{\sqrt{\tau_1^2 + \tau_2^2}}\right)$$

where ν_j and τ_j^2 are the means and variances of μ_j, respectively, for $j = 1, 2$, under the corresponding full conditional densities

Gaussian mixtures: PG_2

First calculate the probability $p(\text{order})$ that $\mu_1 < \mu_2$:

$$p(\text{order}) = \frac{\Phi(E(\mu_2|\cdot) - E(\mu_1|\cdot))}{\sqrt{E(\sigma_1^2|\cdot) + E(\sigma_2^2|\cdot)}},$$

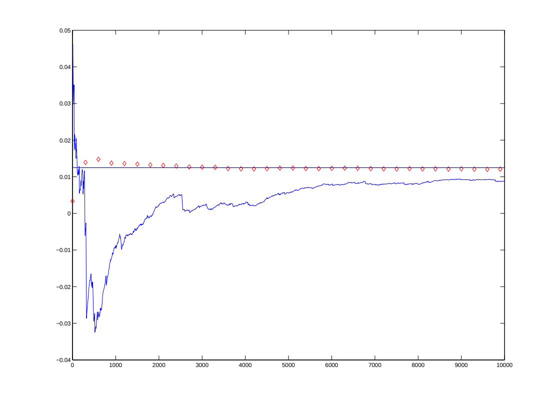
where all four expectations above are taken under the corresponding full conditional densities, and, since the full conditional of each σ_j^{-2} is a Gamma density, the expectations of σ_1, σ_2, σ_1^2, and σ_2^2, are all available in closed form. Therefore, $p(\text{order})$ can be computed explicitly, and, PG_2 is:

$$\frac{\sigma_1^o}{2} + \frac{1}{4}\left[\mathbb{I}_{\{\mu_1<\mu_2\}}E(\sigma_1|\cdot) + \mathbb{I}_{\{\mu_1>\mu_2\}}E(\sigma_2|\cdot)\right] + \frac{1}{4}\left[p(\text{order})\sigma_1 + (1 - p(\text{order}))\sigma_2\right]$$

where all expectations are taken under the corresponding full conditional densities

Gaussian mixtures: Variance reduction

With this choice for G_1, G_2 and corresponding control variates U_1, U_2, the variance reduction factors obtained by $\mu_{n,\mathrm{rev}}(F)$ are **16.17, 25.36, 38.99, 44.5** and **36.16**, after $n = 1000, 10000, 50000, 100000$ and 200000 simulation steps, respectively



Discussion: Applicability

1. The methodology presented applies immediately to *any* reversible MCMC sampler, as long as it is possible to compute the one-step expectation of *some* function G of the parameters, in closed form
2. These estimators can be used in a "black-box" fashion to various state-of-the-art samplers used in Bayesian inference via MCMC:
   - $\leadsto$ all conjugate Gibbs samplers
   - $\leadsto$ all random-walk Metropolis-Hastings samplers with a *discrete proposal*
   - $\leadsto$ many hybrid, Metropolis-within-Gibbs samplers
3. As in the iid case, blind use of all available control variates is not a good idea -standard hypothesis testing for zero-mean θ_j can be used
4. Beyond black-box: Rules of thumb should be used to derive good control variates in broad families of models as demonstrated in log-linear and finite mixture models
5. See next talk for some interesting ongoing research with many open problems

Theorem: "Under minimal assumptions, it all works"

Suppose $\{X_n\}$ is ψ-irreducible, aperiodic, reversible and satisfies the Lyapunov drift condition (V3), $PV \leq V - W + b\mathbb{I}_C$. If $F, G \in L_\infty^W$ and they are non-degenerate, then:

(i) [Ergodicity] The chain is positive Harris recurrent, it has a unique invariant measure π, and it converges in distribution to π in a strong sense

(ii) [LLN] The ergodic averages $\mu_n(F)$, as well as the adaptive averages $\mu_{n,\text{rev}}(F)$, both converge to $\pi(F)$ a.s., as $n \to \infty$.

(iii) [Poisson Equation] There is an essentially unique solution $\hat{F} \in L_\infty^{V+1}$ to the Poisson eqn

(iv) [CLT for $\mu_n(F)$] The normalized ergodic averages $\sqrt{n}[\mu_n(F) - \pi(F)]$ converge in distribution to $N(0, \sigma_F^2)$

(v) [CLT for $\mu_{n,\text{rev}}(F)$] The normalized adaptive averages $\sqrt{n}[\mu_{n,\text{rev}}(F) - \pi(F)]$ converge in distribution to $N(0, \sigma^2_{F_{\theta*}})$, where the variance $\sigma^2_{F_{\theta*}}$ is minimal among all estimators based on the control variate $U = G - PG$