Markov Interacting Importance Samplers

Eduardo F. Mendes Marcel Scharth Robert Kohn

ACEMS (ARC Center of Excellence in Mathematical and Statistical Frontiers) and UNSW

JULY 2015

Organization

- Background

Organization

- Background
- Motivation–Our contribution

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.
- MIIS algorithm.

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.
- MIIS algorithm.
- Augmented Space.

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.
- MIIS algorithm.
- Augmented Space.
- Fundamental MIIS Theorem.

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.
- MIIS algorithm.
- Augmented Space.
- Fundamental MIIS Theorem.
- Some examples of method.

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.
- MIIS algorithm.
- Augmented Space.
- Fundamental MIIS Theorem.
- Some examples of method.
- Variance reduction methods.

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.
- MIIS algorithm.
- Augmented Space.
- Fundamental MIIS Theorem.
- Some examples of method.
- Variance reduction methods.
- Conditional importance sampling within Gibbs.

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.
- MIIS algorithm.
- Augmented Space.
- Fundamental MIIS Theorem.
- Some examples of method.
- Variance reduction methods.
- Conditional importance sampling within Gibbs.
- Empirical Examples.

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.
- MIIS algorithm.
- Augmented Space.
- Fundamental MIIS Theorem.
- Some examples of method.
- Variance reduction methods.
- Conditional importance sampling within Gibbs.
- Empirical Examples.
- Some theory

Organization

- Background
- Motivation–Our contribution
- Conditional Importance Sampler and Comparison with importance sampling.
- MIIS algorithm.
- Augmented Space.
- Fundamental MIIS Theorem.
- Some examples of method.
- Variance reduction methods.
- Conditional importance sampling within Gibbs.
- Empirical Examples.
- Some theory
- Summary.

Background: Monte Carlo Integration

Estimator

1. Method: Given N draws $x_{1:N} = (x_1, \ldots, x_N)$ from $\pi(dx)$,

$$\widehat{E}^N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \to \int f(x)\pi(dx) = E_\pi(f),$$

with probability one as $N \to \infty$.

Background: Monte Carlo Integration

Estimator

1. Method: Given N draws $x_{1:N} = (x_1, \ldots, x_N)$ from $\pi(dx)$,

$$\widehat{E}^N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \to \int f(x)\pi(dx) = E_\pi(f),$$

with probability one as $N \to \infty$.

2. Performance: Central limit theorem ensures that, under regularity conditions,

$$\sqrt{N}[\widehat{E}^N(f) - E_\pi(f)] \to^d N(0, V_{MC}(f)),$$

where $V_{MC}(f)$ is the variance of $E^N(f)$.
Smaller variance $\to$ *better estimator.*

Monte Carlo Integration

Most used Monte Carlo methods:

- Markov chain Monte Carlo (MCMC)

Monte Carlo Integration

Most used Monte Carlo methods:

- Markov chain Monte Carlo (MCMC)
- Importance Sampling and variations

Importance Sampling

Basic Idea

1. Draw N samples $x_{1:n}$ from some proposal distribution $q(dx)$ and estimate $E_\pi(f) = E_q((\pi/q)f) = \int f(x)\frac{\pi(x)}{q(x)}q(dx)$, using

$$\widehat{E}_{IS}^N(f) = \sum_{i=1}^N \bar{w}(i, x_{1:N}) f(x_i),$$

where $w(x_i) = (m/q)(x_i)$, $\bar{w}(i; x_{1:N}) = w(x_i)/\sum_j w(x_j)$ with $\pi(x) \propto m(x)$.

Importance Sampling

Basic Idea

1. Draw N samples $x_{1:n}$ from some proposal distribution $q(dx)$ and estimate $E_\pi(f) = E_q((\pi/q)f) = \int f(x)\frac{\pi(x)}{q(x)}q(dx)$, using

$$\widehat{E}_{IS}^N(f) = \sum_{i=1}^N \bar{w}(i, x_{1:N}) f(x_i),$$

where $w(x_i) = (m/q)(x_i)$, $\bar{w}(i; x_{1:N}) = w(x_i)/\sum_j w(x_j)$ with $\pi(x) \propto m(x)$.

2. Performance: Depends on the variance of the weights $(\pi/q)(x)$ and choice of estimator.
Variance reduction methods: antithetic sampling, systematic sampling, Rao-Blackwellization (partial integration), and control variates.

Markov Chain Monte Carlo (MCMC)

Draw $X^{(i+1)}|x^{(i)}$ from a Markov Kernel $K(x^{(i)}, dx)$ that has π as equilibrium distribution. After a warmup period, the sequence of states $x^{(1)}, \ldots, x^{(N)}$ is used to construct the estimator

$$\widehat{E}^{N}_{MH}(f) = \frac{1}{N} \sum_{i=1}^{N} f\left(x^{(i)}\right),$$

of $E_{\pi}(f)$.
Performance: The variance of the estimator depends on the Integrated Auto-Correlation Time (or inefficiency factor; IACT), that measures the autocorrelation of the chain.

Some Limitations

Importance Sampling

- Difficult to construct high dimensional proposals.

MCMC

Some Limitations

Importance Sampling

- Difficult to construct high dimensional proposals.
- Unbounded weights.

MCMC

Some Limitations

Importance Sampling

- Difficult to construct high dimensional proposals.
- Unbounded weights.
- Hierarchical models requires many particles.

MCMC

Some Limitations

Importance Sampling

- Difficult to construct high dimensional proposals.
- Unbounded weights.
- Hierarchical models requires many particles.

MCMC

- May be hard to choose good proposals for MH.

Some Limitations

Importance Sampling

- Difficult to construct high dimensional proposals.
- Unbounded weights.
- Hierarchical models requires many particles.

MCMC

- May be hard to choose good proposals for MH.
- Difficult to use some variance reduction techniques.

Some Limitations

Importance Sampling

- Difficult to construct high dimensional proposals.
- Unbounded weights.
- Hierarchical models requires many particles.

MCMC

- May be hard to choose good proposals for MH.
- Difficult to use some variance reduction techniques.
- Convergence of the chain.

Our Contribution

- New MCMC algorithm that samples from an IS-type approximation of the target distribution.

Our Contribution

- New MCMC algorithm that samples from an IS-type approximation of the target distribution.
- Alternative to Metropolis-within-Gibbs approaches.

Our Contribution

- New MCMC algorithm that samples from an IS-type approximation of the target distribution.
- Alternative to Metropolis-within-Gibbs approaches.
- Use local proposals in the Importance Sampling.

Our Contribution

- New MCMC algorithm that samples from an IS-type approximation of the target distribution.
- Alternative to Metropolis-within-Gibbs approaches.
- Use local proposals in the Importance Sampling.
- Use variance reduction techniques previously unavailable or difficult in MCMC samplers.

Our methodology

- Can be written as a Gibbs sampler, hence targeting the lower dimensional conditionals and/or levels in a hierarchical model.

Our methodology

- Can be written as a Gibbs sampler, hence targeting the lower dimensional conditionals and/or levels in a hierarchical model.
- Easier to construct proposals.

Our methodology

- Can be written as a Gibbs sampler, hence targeting the lower dimensional conditionals and/or levels in a hierarchical model.
- Easier to construct proposals.
- Natural control of unbounded weights.

Our methodology

- Can be written as a Gibbs sampler, hence targeting the lower dimensional conditionals and/or levels in a hierarchical model.
- Easier to construct proposals.
- Natural control of unbounded weights.
- Use of variance reduction techniques.

Some Definitions

- We define the basic random variables on a set A that is a subset of Euclidean space.

Some Definitions

- We define the basic random variables on a set A that is a subset of Euclidean space.
- $\pi(x)$ is the target density.

Some Definitions

- We define the basic random variables on a set A that is a subset of Euclidean space.
- $\pi(x)$ is the target density.
- We often can evaluate $\pi(x)$ only up to a constant of proportionality $m(x)$, with $\pi(x) = m(x)/Z_m$, where $Z_m = \int_A m(x)dx$ is the normalizing constant.

Conditional Importance Sampler (CIS)

Discussion

- The *conditional importance sampler* (CIS) is the basic building block of the MCMC algorithms

Conditional Importance Sampler (CIS)

Discussion

- The *conditional importance sampler* (CIS) is the basic building block of the MCMC algorithms
- The CIS is motivated by the question: "*how to implement an importance sampler approximation to π that provides unbiased samples?*"

Conditional Importance Sampler (CIS)

Discussion

- The *conditional importance sampler* (CIS) is the basic building block of the MCMC algorithms
- The CIS is motivated by the question: "*how to implement an importance sampler approximation to π that provides unbiased samples?*"
- The CIS is our solution to this problem.

Conditional Importance Sampler (CIS)

Discussion ctd

- We go beyond a simple importance sampler and construct a general framework that not only covers the simple importance sampling approximation with variance reduction techniques, but also extends the basic importance sampling paradigm,
  allowing local exploration of the target inside an MCMC setting, for instance, by using a random-walk approach.

CIS Motivation (ctd)

- At each iterate of an MCMC algorithm, the CIS constructs an empirical approximation to the target density $\pi(\cdot)$.

CIS Motivation (ctd)

- At each iterate of an MCMC algorithm, the CIS constructs an empirical approximation to the target density $\pi(\cdot)$.
- It generates an auxiliary variable ξ and N particles $X_{1:N}$ conditional on the previous iterate y, in such a way that:

CIS Motivation (ctd)

- At each iterate of an MCMC algorithm, the CIS constructs an empirical approximation to the target density $\pi(\cdot)$.
- It generates an auxiliary variable ξ and N particles $X_{1:N}$ conditional on the previous iterate y, in such a way that:
- one particle X_k is generated through a Markov transition kernel

CIS Motivation (ctd)

- At each iterate of an MCMC algorithm, the CIS constructs an empirical approximation to the target density $\pi(\cdot)$.
- It generates an auxiliary variable ξ and N particles $X_{1:N}$ conditional on the previous iterate y, in such a way that:
- one particle X_k is generated through a Markov transition kernel
- and the other $N-1$ particles are generated conditional on X_k.

CIS preliminaries

- Let $\eta(\xi|y)$ be the conditional density of the auxiliary variable ξ, with $\xi, y \in A$,

CIS preliminaries

- Let $\eta(\xi|y)$ be the conditional density of the auxiliary variable ξ, with $\xi, y \in A$,
- Let $T(y, x; \xi)$ be the density of a Markov transition kernel from y to $x \in A$, conditional on ξ, that is reversible in the sense that

$$\pi(y)\eta(\xi|y)T(y, x; \xi) = \pi(x)\eta(\xi|x)T(x, y; \xi). \quad (1)$$

CIS preliminaries

- Let $\eta(\xi|y)$ be the conditional density of the auxiliary variable ξ, with $\xi, y \in A$,
- Let $T(y, x; \xi)$ be the density of a Markov transition kernel from y to $x \in A$, conditional on ξ, that is reversible in the sense that

$$\pi(y)\eta(\xi|y)T(y, x; \xi) = \pi(x)\eta(\xi|x)T(x, y; \xi). \tag{1}$$

- $q_i(x_i|\xi)$ $(i = 1, \ldots, N)$ are importance densities for x_i, $i = 1, \ldots, N$ conditional on ξ.

CIS algorithm

- The CIS generates $(X_{1:N}, \xi)$ using the following algorithm. Given (y, k),

CIS algorithm

- The CIS generates $(X_{1:N}, \xi)$ using the following algorithm. Given (y, k),
  1. sample $\xi \sim \eta(\xi|y)$;

CIS algorithm

- The CIS generates $(X_{1:N}, \xi)$ using the following algorithm. Given (y, k),
  1. sample $\xi \sim \eta(\xi|y)$;
  2. sample $X_k \sim T(y, x_k; \xi)$; i.e., generate the particle x_k using the Markov kernel.

CIS algorithm

- The CIS generates $(X_{1:N}, \xi)$ using the following algorithm. Given (y, k),
  1. sample $\xi \sim \eta(\xi|y)$;
  2. sample $X_k \sim T(y, x_k; \xi)$; i.e., generate the particle x_k using the Markov kernel.
  3. sample $X_i \sim q_i(x_i|x_k, \xi), i = 1, \ldots, N, i \neq k$; i.e., generate all the remaining particles conditional on ξ and the propagated particle x_k.

Using CIS to construct Empirical Distribution for π

- From the output of the Conditional Importance Sampler we define the weights for $i = 1, \ldots, N$

$$W_i(x_{1:N}; \xi) := \frac{w_i(x_i; \xi)}{\sum_{j=1}^{N} w_j(x_j; \xi)} \quad \text{where} \quad w_i(x; \xi) := \frac{m(x)}{q_i(x|\xi)} \eta(\xi|x) \tag{2}$$

Using CIS to construct Empirical Distribution for π

- From the output of the Conditional Importance Sampler we define the weights for $i = 1, \ldots, N$

$$W_i(x_{1:N}; \xi) := \frac{w_i(x_i; \xi)}{\sum_{j=1}^{N} w_j(x_j; \xi)} \quad \text{where} \quad w_i(x; \xi) := \frac{m(x)}{q_i(x|\xi)} \eta(\xi|x) \tag{2}$$

- Let $\widehat{\pi}^N_{CIS} := \{(x_1, W_1(x_{1:N}, \xi)), \ldots, (x_N, W_N(x_{1:N}, \xi))\}$ be the empirical approximation to π.

Using CIS to construct Empirical Distribution for π

- From the output of the Conditional Importance Sampler we define the weights for $i = 1, \ldots, N$

$$W_i(x_{1:N}; \xi) := \frac{w_i(x_i; \xi)}{\sum_{j=1}^{N} w_j(x_j; \xi)} \quad \text{where} \quad w_i(x; \xi) := \frac{m(x)}{q_i(x|\xi)} \eta(\xi|x) \tag{2}$$

- Let $\widehat{\pi}^N_{CIS} := \{(x_1, W_1(x_{1:N}, \xi)), \ldots, (x_N, W_N(x_{1:N}, \xi))\}$ be the empirical approximation to π.
- The weights depend on the marginals $q_i(\cdot|x_k, \xi)$ $(i = 1, \ldots, N)$, the auxiliary distribution $\eta(\xi|\cdot)$ and the target distribution $\pi(\cdot) \propto m(\cdot)$.

Estimator of $E_\pi(f)$ using empirical distribution

- Based on $\widehat{\pi}^N_{CIS}$, we define the estimator of $E_\pi(f)$ as

$$\widehat{E}^N_{CIS}(f) := \sum_{i=1}^{N} W_i(x_{1:N}, \xi) f(x_i) = E_{\widehat{\pi}^N_{CIS}}(f). \tag{3}$$

Estimator of $E_\pi(f)$ using empirical distribution

- Based on $\widehat{\pi}^N_{CIS}$, we define the estimator of $E_\pi(f)$ as

$$\widehat{E}^N_{CIS}(f) := \sum_{i=1}^{N} W_i(x_{1:N}, \xi) f(x_i) = E_{\widehat{\pi}^N_{CIS}}(f). \tag{3}$$

- Note that this estimator uses all N particles in the empirical distribution.

Definition of an augmented target distribution

- Define the joint density

$$\widetilde{\pi}^N(k, y, x_{1:N}, \xi) := N^{-1}\pi(y)\Gamma^N(x_{1:N}, \xi|y, k). \tag{4}$$

Definition of an augmented target distribution

- Define the joint density

$$\widetilde{\pi}^N(k, y, x_{1:N}, \xi) := N^{-1}\pi(y)\Gamma^N(x_{1:N}, \xi|y, k). \tag{4}$$

- $$\Gamma^N(x_{1:N}, \xi|y, k) := \eta(\xi|y)\, T(y, x_k; \xi) \prod_{i=1, i\neq k}^{N} q_i(x_i|x_k, \xi) \tag{5}$$

Definition of an augmented target distribution

- Define the joint density

$$\widetilde{\pi}^N(k, y, x_{1:N}, \xi) := N^{-1}\pi(y)\Gamma^N(x_{1:N}, \xi|y, k). \tag{4}$$

- $$\Gamma^N(x_{1:N}, \xi|y, k) := \eta(\xi|y)\, T(y, x_k; \xi) \prod_{i=1, i\neq k}^{N} q_i(x_i|x_k, \xi) \tag{5}$$

- So, K is uniform from $1, \ldots, N$, y is from π, and then ξ and $X_{1:N}$ are generated by the CIS Γ^N.

Fundamental Result

Suppose that $E_\pi(|f|)$ is finite, (k, y) is a sample from $N^{-1}\pi(y)$, and that $(x_{1:N}, \xi)$ is generated from $\Gamma^N(x_{1:N}, \xi|y, k)$. Then,

(i) $\widetilde{\pi}^N(y) = \pi(y)$.

Fundamental Result

Suppose that $E_\pi(|f|)$ is finite, (k, y) is a sample from $N^{-1}\pi(y)$, and that $(x_{1:N}, \xi)$ is generated from $\Gamma^N(x_{1:N}, \xi|y, k)$. Then,

(i) $\widetilde{\pi}^N(y) = \pi(y)$.

(ii)

$$\widetilde{\pi}^N(k, y|x_{1:N}, \xi) = \sum_{i=1}^{N} W_i(x_{1:N}, \xi) I(k = i)\, T(x_i, y; \xi),$$

or equivalently,

$$\widetilde{\pi}^N(K = i|x_{1:N}, \xi) = W_i(x_{1:N}, \xi) \quad \text{and}$$

$$\widetilde{\pi}^N(y|x_{1:N}\xi, k) = T(x_k, y; \xi).$$

(6)

Fundamental Result

Suppose that $E_\pi(|f|)$ is finite, (k, y) is a sample from $N^{-1}\pi(y)$, and that $(x_{1:N}, \xi)$ is generated from $\Gamma^N(x_{1:N}, \xi|y, k)$. Then,

(i) $\widetilde{\pi}^N(y) = \pi(y)$.

(ii)

$$\widetilde{\pi}^N(k, y|x_{1:N}, \xi) = \sum_{i=1}^{N} W_i(x_{1:N}, \xi) I(k = i)\, T(x_i, y; \xi),$$

or equivalently,

$$\widetilde{\pi}^N(K = i|x_{1:N}, \xi) = W_i(x_{1:N}, \xi) \quad \text{and}$$
$$\widetilde{\pi}^N(y|x_{1:N}\xi, k) = T(x_k, y; \xi). \tag{6}$$

(iii) $E_{\widetilde{\pi}^N}\left(\widehat{E}^N_{CIS}(f)\right) = E_\pi(f)$.

Markov Interacting Importance Sampler Algorithm

These 3 steps use the CIS to draw an approximation to π.

Given $y^{(0)} \in A$ and $1 \leq k^{(0)} \leq N$, at step $t = 1, 2, \ldots$

1. Generate $\xi^{(t)} | y^{(t-1)} \sim \eta(\xi | y^{(t-1)})$.

Markov Interacting Importance Sampler Algorithm

These 3 steps use the CIS to draw an approximation to π.

Given $y^{(0)} \in A$ and $1 \leq k^{(0)} \leq N$, at step $t = 1, 2, \ldots$

1. Generate $\xi^{(t)} | y^{(t-1)} \sim \eta(\xi | y^{(t-1)})$.
2. Generate $X^{(t)}_{k^{(t-1)}} | (y^{(t-1)}, \xi^{(t)}) \sim T\left(y^{(t-1)}, x^{(t)}_{k^{(t-1)}}; \xi^{(t)}\right)$.

Markov Interacting Importance Sampler Algorithm

These 3 steps use the CIS to draw an approximation to π.

Given $y^{(0)} \in A$ and $1 \leq k^{(0)} \leq N$, at step $t = 1, 2, \ldots$

1. Generate $\xi^{(t)} | y^{(t-1)} \sim \eta(\xi | y^{(t-1)})$.
2. Generate $X^{(t)}_{k^{(t-1)}} | (y^{(t-1)}, \xi^{(t)}) \sim T\left(y^{(t-1)}, x^{(t)}_{k^{(t-1)}}; \xi^{(t)}\right)$.
3. Generate

$$X_i^{(t)} \Big| \left(x^{(t-1)}_{k^{(t-1)}}, \xi^{(t)}\right) \sim q_i\left(x_i^{(t)} \Big| x^{(t)}_{k^{(t-1)}}, \xi^{(t)}\right).$$

Markov Interacting Importance Sampler Algorithm (ctd)

Drawing from the approximation

(4) For $k = 1, \ldots, N$, calculate

$$w_k\left(x_k^{(t)}; \xi\right) = \frac{m(x_k^{(t)})}{q_k(x_k^{(t)}|\xi^{(t)})}\eta(\xi^{(t)}|x_k^{(t)}),$$

$$W_k(x_{1:N}^{(t)}, \xi^{(t)}) = \frac{w_k\left(x_k^{(t)}; \xi\right)}{\sum_{j=1}^{N} w_j\left(x_j^{(t)}; \xi\right)} \ .$$

Draw $K^{(t)} = k|(x_{1:N}^{(t)}, \xi^{(t)})$ with probability $W_k(x_{1:N}^{(t)}, \xi^{(t)})$.

Markov Interacting Importance Sampler Algorithm (ctd)

Drawing from the approximation

(4) For $k = 1, \ldots, N$, calculate

$$w_k\left(x_k^{(t)}; \xi\right) = \frac{m(x_k^{(t)})}{q_k(x_k^{(t)}|\xi^{(t)})}\eta(\xi^{(t)}|x_k^{(t)}),$$

$$W_k(x_{1:N}^{(t)}, \xi^{(t)}) = \frac{w_k\left(x_k^{(t)}; \xi\right)}{\sum_{j=1}^{N} w_j\left(x_j^{(t)}; \xi\right)} \ .$$

Draw $K^{(t)} = k|(x_{1:N}^{(t)}, \xi^{(t)})$ with probability $W_k(x_{1:N}^{(t)}, \xi^{(t)})$.

(5) Generate $Y^{(t)}|(x_{1:N}^{(t)}, k^{(t)}, \xi^{(t)}) \sim T\left(x_{k^{(t)}}^{(t)}, y^{(t)}; \xi^{(t)}\right)$.

Characterization of MIIS

The *Markov Interacting Importance Sampler* is a Gibbs sampler targeting the augmented density

$$\widetilde{\pi}^N(k, y, x_{1:N}, \xi) := N^{-1}\pi(y)\Gamma^N(x_{1:N}, \xi|y, k). \tag{7}$$

that has $\pi(y)$ as a marginal density.

Comparison with Importance Sampling

Description of IS

1. In importance sampling, we draw particles $x_{1:N}$ from *importance* or *proposal* densities densities $q_i(x_i)$ and calculate their *importance weights*

$$W_i(x_{1:N}) := \frac{w_i(x_i)}{\sum_{j=1}^{N} w_j(x_j)}, \quad \text{where} \quad w_i(x_i) := \frac{m(x_i)}{q_i(x_i)},$$

Comparison with Importance Sampling

Description of IS

1. In importance sampling, we draw particles $x_{1:N}$ from *importance* or *proposal* densities densities $q_i(x_i)$ and calculate their *importance weights*

$$W_i(x_{1:N}) := \frac{w_i(x_i)}{\sum_{j=1}^{N} w_j(x_j)}, \quad \text{where} \quad w_i(x_i) := \frac{m(x_i)}{q_i(x_i)},$$

2. We obtain the approximation $\widehat{\pi}_{IS}^{N} := \{W_{1:N}(x_{1:N}), x_{1:N}\}$ to π.

Comparison with Importance Sampling

Description of IS

1. In importance sampling, we draw particles $x_{1:N}$ from *importance* or *proposal* densities densities $q_i(x_i)$ and calculate their *importance weights*

$$W_i(x_{1:N}) := \frac{w_i(x_i)}{\sum_{j=1}^N w_j(x_j)}, \quad \text{where} \quad w_i(x_i) := \frac{m(x_i)}{q_i(x_i)},$$

2. We obtain the approximation $\widehat{\pi}_{IS}^N := \{W_{1:N}(x_{1:N}), x_{1:N}\}$ to π.
3. The IS sampling estimate of $E_\pi(f)$ is

$$\widehat{E}_{IS}^N(f) := \sum_{i=1}^N W_i(x_{1:N}) f(x_i) = E_{\widehat{\pi}_{IS}^N}(f) \tag{8}$$

In the simplest case, the *particles* $x_{1:N}$ are sampled independently from the same proposal distribution q, i.e., $q_1 = \ldots = q_N = q$

Comparison with Importance Sampling

Difference between CIS and IS

1. The marginal distribution of a sample X from $\widehat{\pi}^N_{IS}$ is not $\pi(X)$, while the distribution of Y from $\widehat{\pi}^N_{CIS}$ is $\pi(Y)$. Similarly,

$$E_q\left(\widehat{E}^{IS}_{\pi}(f)\right) \neq E_{\pi}(f), \tag{9}$$

whereas $E_{\widetilde{\pi}^N}\left(\widehat{E}^N_{CIS}(f)\right) = E_{\pi}(f)$.

Comparison with Importance Sampling

Difference between CIS and IS

1. The marginal distribution of a sample X from $\widehat{\pi}_{IS}^{N}$ is not $\pi(X)$, while the distribution of Y from $\widehat{\pi}_{CIS}^{N}$ is $\pi(Y)$. Similarly,

$$E_q\left(\widehat{E}_{\pi}^{IS}(f)\right) \neq E_{\pi}(f), \tag{9}$$

   whereas $E_{\widetilde{\pi}^N}\left(\widehat{E}_{CIS}^{N}(f)\right) = E_{\pi}(f)$.

2. The weights w_i in the CIS may depend on an auxiliary variable ξ, with density $\eta(\cdot|y)$, that incorporates past information in the proposal opening the possibility for using local proposals. Moreover, it can be used as a mechanism to bound the weights and provide more robust estimators.

Example

Random walk CIS

1. The random walk importance sampler draws particles from a symmetric proposal ϕ dependent on its past.

Example

Random walk CIS

1. The random walk importance sampler draws particles from a symmetric proposal ϕ dependent on its past.
2. The advantage is that the method bounds the weights by construction.

Example

Random walk CIS

1. The random walk importance sampler draws particles from a symmetric proposal ϕ dependent on its past.
2. The advantage is that the method bounds the weights by construction.
3. The random walk proposal performs local exploration around the auxiliary variable ξ, which we sample conditionally on the previous state.

Example

Random walk CIS

1. The random walk importance sampler draws particles from a symmetric proposal ϕ dependent on its past.
2. The advantage is that the method bounds the weights by construction.
3. The random walk proposal performs local exploration around the auxiliary variable ξ, which we sample conditionally on the previous state.
4. $\eta(\xi|y) = \phi(\xi - y)$ denote the proposal functions for ξ.

Example

Random walk CIS

1. The random walk importance sampler draws particles from a symmetric proposal ϕ dependent on its past.
2. The advantage is that the method bounds the weights by construction.
3. The random walk proposal performs local exploration around the auxiliary variable ξ, which we sample conditionally on the previous state.
4. $\eta(\xi|y) = \phi(\xi - y)$ denote the proposal functions for ξ.
5. $T(y, x; \xi) = \delta_y(x)$

Example

Random walk CIS

1. The random walk importance sampler draws particles from a symmetric proposal ϕ dependent on its past.
2. The advantage is that the method bounds the weights by construction.
3. The random walk proposal performs local exploration around the auxiliary variable ξ, which we sample conditionally on the previous state.
4. $\eta(\xi|y) = \phi(\xi - y)$ denote the proposal functions for ξ.
5. $T(y, x; \xi) = \delta_y(x)$
6. $q(x_i|\xi) = \phi(x_i - \xi)$ denote the proposal functions for the q_i, $i \neq k$

Example

Antithetic Variables

1. In the importance sampling literature, the method of antithetic variables consists of drawing perfectly negatively correlated particles to reduce the variance of the Monte Carlo estimate.

Example

Antithetic Variables

1. In the importance sampling literature, the method of antithetic variables consists of drawing perfectly negatively correlated particles to reduce the variance of the Monte Carlo estimate.
2. We can use this method within the MIIS framework.

Example

Antithetic Variables

1. In the importance sampling literature, the method of antithetic variables consists of drawing perfectly negatively correlated particles to reduce the variance of the Monte Carlo estimate.
2. We can use this method within the MIIS framework.
3. The importance sampler with antithetic variables draws the particles in pairs from a proposal distribution.

Example

Antithetic Variables

1. In the importance sampling literature, the method of antithetic variables consists of drawing perfectly negatively correlated particles to reduce the variance of the Monte Carlo estimate.
2. We can use this method within the MIIS framework.
3. The importance sampler with antithetic variables draws the particles in pairs from a proposal distribution.
4. Suppose that N is even. For $k \leq N/2$, let $q_k(x_k)$ be the density of x_k with corresponding cumulative distribution function $Q_k(\cdot)$

Example

Antithetic Variables

1. In the importance sampling literature, the method of antithetic variables consists of drawing perfectly negatively correlated particles to reduce the variance of the Monte Carlo estimate.
2. We can use this method within the MIIS framework.
3. The importance sampler with antithetic variables draws the particles in pairs from a proposal distribution.
4. Suppose that N is even. For $k \leq N/2$, let $q_k(x_k)$ be the density of x_k with corresponding cumulative distribution function $Q_k(\cdot)$
5. let $x_{N/2+k} = Q_k^{-1}(1 - Q_k(x_k))$, where Q_k^{-1} is the inverse of Q_k.

Conditional Importance Sampling within Gibbs

1. Divide y into d blocks, $y(1), ..., y(d)$.

Conditional Importance Sampling within Gibbs

1. Divide y into d blocks, $y(1), ..., y(d)$.
2. Apply MIIS within each block.

Estimation of expectations

Basic Monte Carlo Estimate

1. Suppose that the MIIS algorithm targeting π runs for M iterations after reaching the stationary distribution.

Estimation of expectations

Basic Monte Carlo Estimate

1. Suppose that the MIIS algorithm targeting π runs for M iterations after reaching the stationary distribution.
2. The simplest estimator of $E_\pi(f)$, which uses only the output $\{x^{(t)}\}$ from the Markov Chain, is

$$\widehat{E}^M_{MC}(f) := \frac{1}{M}\sum_{t=1}^{M} f\left(x^{(t)}\right) = E_{\widehat{\pi}^M_{MC}}(f) \tag{10}$$

Estimation of expectations

Basic Monte Carlo Estimate

1. Suppose that the MIIS algorithm targeting π runs for M iterations after reaching the stationary distribution.
2. The simplest estimator of $E_\pi(f)$, which uses only the output $\{x^{(t)}\}$ from the Markov Chain, is

$$\widehat{E}^M_{MC}(f) := \frac{1}{M}\sum_{t=1}^{M} f\left(x^{(t)}\right) = E_{\widehat{\pi}^M_{MC}}(f) \tag{10}$$

3. where $\widehat{\pi}^M_{MC} = \{(1/N, x^{(1)}), \ldots, (1/N, x^{(M)})\}$.

1. The MIIS algorithm constructs an unbiased approximation

$$\widehat{E}^{N}_{CIS,t}(f) := \sum_{i=1}^{N} W_i \left(x^{(t)}_{1:N}; \xi^{(t)} \right) f \left(x^{(t)}_i \right) \tag{11}$$

to $E_\pi(f)$ at each iteration t of the Markov chain, after the chain has converged.

Estimation of expectations

Reusing all the particles

1. The MIIS algorithm constructs an unbiased approximation

$$\widehat{E}^{N}_{CIS,t}(f) := \sum_{i=1}^{N} W_i \left(x^{(t)}_{1:N}; \xi^{(t)} \right) f \left(x^{(t)}_{i} \right) \tag{11}$$

   to $E_\pi(f)$ at each iteration t of the Markov chain, after the chain has converged.

2. The MIIS estimator that averages over the terms $\widehat{E}^{N}_{CIS,t}(f)$ is

$$\widehat{E}^{M,N}_{MIIS}(f) := \frac{1}{M} \sum_{t=1}^{M} \widehat{E}^{N}_{CIS,t}(f) = \widehat{E}^{M}_{MC} \left(\widehat{E}^{N}_{CIS,t}(f) \right) \tag{12}$$

Estimation of expectations

Variance Reduction

The power of the MIIS relies on reusing the approximations to the target constructed at each step. We show how to reuse all generated particles to reduce the variance of the estimators.

1. Show how to construct a Rao-Blackwellized estimator for the Gibbs sampler

Estimation of expectations

Variance Reduction

The power of the MIIS relies on reusing the approximations to the target constructed at each step. We show how to reuse all generated particles to reduce the variance of the estimators.

1. Show how to construct a Rao-Blackwellized estimator for the Gibbs sampler
2. Show how to implement control variates in the MIIS

Estimation of expectations

Rao Blackwellization

1. The motivation for Rao-Blackwellized estimators is that the variance of $f(x(s))$ is larger than the variance of $E_{\pi_s(\cdot|x(\setminus s)}(f)$.

Estimation of expectations

Rao Blackwellization

1. The motivation for Rao-Blackwellized estimators is that the variance of $f(x(s))$ is larger than the variance of $E_{\pi_s(\cdot|x(\setminus s)}(f)$.
2. However, the latter requires knowledge of the conditional expectation in closed form.

Estimation of expectations

Rao Blackwellization

1. The motivation for Rao-Blackwellized estimators is that the variance of $f(x(s))$ is larger than the variance of $E_{\pi_s(\cdot|x(\setminus s)}(f)$.
2. However, the latter requires knowledge of the conditional expectation in closed form.
3. The MIIS for the Gibbs sampler overcomes this limitation by using an asymptotically (i.e., as the MCMC converges to π) unbiased approximation of the unknown integral.

Estimation of expectations

Rao Blackwellization

1. The motivation for Rao-Blackwellized estimators is that the variance of $f(x(s))$ is larger than the variance of $E_{\pi_s(\cdot|x(\setminus s)}(f)$.
2. However, the latter requires knowledge of the conditional expectation in closed form.
3. The MIIS for the Gibbs sampler overcomes this limitation by using an asymptotically (i.e., as the MCMC converges to π) unbiased approximation of the unknown integral.
4. Thus, for two components, we would get an unbiased estimator of $E_{\pi_1(\cdot|x(2))}(f(x_1)))$ and $E_{\pi_1(\cdot|x(1))}(f(x_2)))$

Estimation of expectations

Control Variates

1. For $j = 1, \ldots, p$, suppose that $g_j(x)$ is an integrable function with respect to the density π and $U(g_j)$ a real function such that $E_{\widetilde{\pi}^N}\big(U(g_j)\big) = 0$.

Estimation of expectations

Control Variates

1. For $j = 1, \ldots, p$, suppose that $g_j(x)$ is an integrable function with respect to the density π and $U(g_j)$ a real function such that $E_{\widetilde{\pi}^N}\big(U(g_j)\big) = 0$.
2. Let $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p)$ be a $p \times 1$ vector of parameters and let $F = f - \sum_{j=1}^{p} \kappa_j U(g_j)$.

Estimation of expectations

Control Variates

1. For $j = 1, \ldots, p$, suppose that $g_j(x)$ is an integrable function with respect to the density π and $U(g_j)$ a real function such that $E_{\widetilde{\pi}^N}\big(U(g_j)\big) = 0$.
2. Let $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p)$ be a $p \times 1$ vector of parameters and let $F = f - \sum_{j=1}^{p} \kappa_j U(g_j)$.
3. For an optimal choice of $\boldsymbol{\kappa}$, we would like the variance of the estimate of the posterior mean of F to be smaller than that of f.

Estimation of expectations

Control Variates

1. For $j = 1, \ldots, p$, suppose that $g_j(x)$ is an integrable function with respect to the density π and $U(g_j)$ a real function such that $E_{\widetilde{\pi}^N}\big(U(g_j)\big) = 0$.
2. Let $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p)$ be a $p \times 1$ vector of parameters and let $F = f - \sum_{j=1}^{p} \kappa_j U(g_j)$.
3. For an optimal choice of $\boldsymbol{\kappa}$, we would like the variance of the estimate of the posterior mean of F to be smaller than that of f.
4. The variables $U(g_i)$ are the *control variates*.

Estimation of expectations

Control Variates

1. For $j = 1, \ldots, p$, suppose that $g_j(x)$ is an integrable function with respect to the density π and $U(g_j)$ a real function such that $E_{\widetilde{\pi}^N}\big(U(g_j)\big) = 0$.
2. Let $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p)$ be a $p \times 1$ vector of parameters and let $F = f - \sum_{j=1}^{p} \kappa_j U(g_j)$.
3. For an optimal choice of $\boldsymbol{\kappa}$, we would like the variance of the estimate of the posterior mean of F to be smaller than that of f.
4. The variables $U(g_i)$ are the *control variates*.
5. Control variates are not commonly used in an MCMC setting because their use makes it more difficult to ensure that the sampling scheme is Markov.

Simulations and Applications

We show the performance of the method in three examples:

1. MIIS-Gibbs sampler on Bivariate Gaussian

Simulations and Applications

We show the performance of the method in three examples:

1. MIIS-Gibbs sampler on Bivariate Gaussian
2. Random-Walk MIIS on a Markov Modulated Poisson Process

Simulations and Applications

We show the performance of the method in three examples:

1. MIIS-Gibbs sampler on Bivariate Gaussian
2. Random-Walk MIIS on a Markov Modulated Poisson Process
3. Gibbs sampler on Mixed Logit Model

Bivariate Gaussian

- Target distribution

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

for $\rho \in \{.5, .99\}$.

The *ideal* Gibbs sampler

Bivariate Gaussian

- Target distribution

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

  for $\rho \in \{.5, .99\}$.
- Algorithms: Gibbs sampler , Metropolis within Gibbs (MwG), and MIIS (N=50).

The *ideal* Gibbs sampler

Bivariate Gaussian

- Target distribution

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

  for $\rho \in \{.5, .99\}$.
- Algorithms: Gibbs sampler , Metropolis within Gibbs (MwG), and MIIS (N=50).
- Proposal non-central t_5 with same mean and variance as the target.

The *ideal* Gibbs sampler

Bivariate Gaussian

- Target distribution

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

  for $\rho \in \{.5, .99\}$.
- Algorithms: Gibbs sampler , Metropolis within Gibbs (MwG), and MIIS (N=50).
- Proposal non-central t_5 with same mean and variance as the target.
- Goal: estimate mean and variance of X_1, and $P(X_1 < -2.32)$.

The *ideal* Gibbs sampler

Bivariate Gaussian

- Target distribution

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

  for $\rho \in \{.5, .99\}$.
- Algorithms: Gibbs sampler , Metropolis within Gibbs (MwG), and MIIS (N=50).
- Proposal non-central t_5 with same mean and variance as the target.
- Goal: estimate mean and variance of X_1, and $P(X_1 < -2.32)$.

The *ideal* Gibbs sampler

1. Sample $X_1|x_2 \sim N(\rho x_2, 1 - \rho^2)$;

Bivariate Gaussian

- Target distribution

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

  for $\rho \in \{.5, .99\}$.
- Algorithms: Gibbs sampler , Metropolis within Gibbs (MwG), and MIIS (N=50).
- Proposal non-central t_5 with same mean and variance as the target.
- Goal: estimate mean and variance of X_1, and $P(X_1 < -2.32)$.

The *ideal* Gibbs sampler

1. Sample $X_1|x_2 \sim N(\rho x_2, 1 - \rho^2)$;
2. Sample $X_2|x_1 \sim N(\rho x_1, 1 - \rho^2)$.

Table: Bivariate Gaussian simulation – Monte Carlo MSE of target density expectation estimates relative to MwG.

	$\rho = 0.99$				
	Gibbs	Gibbs-CV	MwG	MIIS-CV	MIIS/A-CV
Mean	1.087	0.002	1.000	0.011	0.002
Variance	0.805	0.001	1.000	0.011	0.001
Covariance	0.789	0.001	1.000	0.022	0.002
$P(X(1) < -2.32)$	0.942	0.746	1.000	0.966	0.874

	$\rho = 0.5$				
	Gibbs	Gibbs-CV	MwG	MIIS-CV	MIIS/A-CV
Mean	0.931	0.000	1.000	0.025	0.000
Variance	0.974	0.000	1.000	0.177	0.225
Covariance	0.988	0.000	1.000	0.066	0.022
$P(X(1) < -2.32)$	0.906	0.148	1.000	0.270	0.240

Markov Modulated Poisson Process

- $Y_t|X_t \sim Poisson(\psi_{X_t})$.

Markov Modulated Poisson Process

- $Y_t|X_t \sim Poisson(\psi_{X_t})$.
- X_t is a continuous time Markov Process on a discrete state space $\{1, \ldots, d\}$, with generator matrix Q.

Markov Modulated Poisson Process

- $Y_t|X_t \sim Poisson(\psi_{X_t})$.
- X_t is a continuous time Markov Process on a discrete state space $\{1, \ldots, d\}$, with generator matrix Q.
- $\psi = (\psi_1, \ldots, \psi_d)'$ is a (non-negative) vector of intensities.

Markov Modulated Poisson Process

- $Y_t|X_t \sim Poisson(\psi_{X_t})$.
- X_t is a continuous time Markov Process on a discrete state space $\{1, \ldots, d\}$, with generator matrix Q.
- $\psi = (\psi_1, \ldots, \psi_d)'$ is a (non-negative) vector of intensities.
- We observe the n event times $t_1, \ldots, t_{n+1}$, where t_i is the time elapsed since the last event (t_1 is the time for the first event).

Markov Modulated Poisson Process

- $Y_t|X_t \sim Poisson(\psi_{X_t})$.
- X_t is a continuous time Markov Process on a discrete state space $\{1, \ldots, d\}$, with generator matrix Q.
- $\psi = (\psi_1, \ldots, \psi_d)'$ is a (non-negative) vector of intensities.
- We observe the n event times $t_1, \ldots, t_{n+1}$, where t_i is the time elapsed since the last event (t_1 is the time for the first event).
- The Likelihood of this model is

$$L(Q, \Psi, \mathbf{t}) = \nu' e^{(Q-\Psi)t_1}\Psi \ldots e^{(Q-\Psi)t_n}\Psi e^{(Q-\Psi)t_{n+1}}\mathbf{1},$$

where $\Psi = \text{diag}(\psi)$ and ν is the initial distribution of X_t.

Markov Modulated Poisson Process

Applications:

- Occurrence of rare DNA motif in bacteria genome. (*)

Markov Modulated Poisson Process

Applications:

- Occurrence of rare DNA motif in bacteria genome. (*)
- Frequency of bank transactions.

Markov Modulated Poisson Process

Applications:

- Occurrence of rare DNA motif in bacteria genome. (*)
- Frequency of bank transactions.
- Requests for web pages from the WWW.

Markov Modulated Poisson Process

Applications:

- Occurrence of rare DNA motif in bacteria genome. (*)
- Frequency of bank transactions.
- Requests for web pages from the WWW.
- Overflow in telecommunications networks.

Markov Modulated Poisson Process

Applications:

- Occurrence of rare DNA motif in bacteria genome. (*)
- Frequency of bank transactions.
- Requests for web pages from the WWW.
- Overflow in telecommunications networks.
- Packeted voice and data streams.

Markov Modulated Poisson Process

Applications:

- Occurrence of rare DNA motif in bacteria genome. (*)
- Frequency of bank transactions.
- Requests for web pages from the WWW.
- Overflow in telecommunications networks.
- Packeted voice and data streams.
- Wet deposition of a radionuclide emitted from a point source.

Markov Modulated Poisson Process

Applications:

- Occurrence of rare DNA motif in bacteria genome. (*)
- Frequency of bank transactions.
- Requests for web pages from the WWW.
- Overflow in telecommunications networks.
- Packeted voice and data streams.
- Wet deposition of a radionuclide emitted from a point source.
- Arrival of photons from a single-molecule fluorescent experiments where the arrival rate is modulated by the state of a molecule which (in the simplest setting) alternates between ground-state and excited state.

Table: MMPP DGP with $\psi_1 = 10$ and $\psi_2 = 17$. Monte Carlo MSE of posterior estimates relative to MH (average across parameters). We define the time adjusted MSE as $(\text{time}_{alg} \times \text{mse}_{alg})/(time_{MH} \times mse_{MH})$, which approximates the MSE relative to the MH algorithm for the same amount of computing time.

			MIIS		MIIS-CV	
	MH	MTM	N=8	N=16	N=8	N=16
Acceptance	0.30	0.46	0.32	0.47	0.32	0.47
IACT	15.0	9.3	7.5	5.6	7.5	5.6
Time	8	15	9	14	9	14
	Initializing at the true parameters					
Mean	1.00	0.47	0.24	0.16	0.15	0.12
Variance	1.00	0.67	0.25	0.14	0.20	0.15
	Time adjusted MSE					
Mean	1.00	0.88	0.27	0.29	0.17	0.21
Variance	1.00	1.26	0.29	0.25	0.22	0.27

Table: MMPP DGP with $\psi_1 = 10$ and $\psi_2 = 17$. Monte Carlo MSE of posterior estimates relative to MH (average across parameters). We define the time adjusted MSE as $(\text{time}_{alg} \times \text{mse}_{alg})/(time_{MH} \times mse_{MH})$, which approximates the MSE relative to the MH algorithm for the same amount of computing time.

	Initializing from the prior					
Mean	1.00	0.07	0.06	0.02	0.01	0.01
Variance	1.00	0.24	0.19	0.06	0.05	0.04
	Time adjusted MSE					
Mean	1.00	0.13	0.07	0.03	0.01	0.01
Variance	1.00	0.46	0.21	0.10	0.06	0.06

MMPP : Chi-site data for *Escherichia coli*

- Understanding different features of the genome and what affect biological processes.
- Genome segmentation: partition genome in regions that are homogeneous w.r.t. some characteristics and then looking for correlations between other characteristics or biological process of interest.
- Segmentation of a bacterial genome based on rate of occurrence of a particular DNA motif.
- The Chi site is a motif of eight base pairs: GCTGGTGG.

MMPP : Chi-site data for *Escherichia coli*

- Estimate MMPP for the lagging part of the outer ring of the E. coli genome strand.
- Total of 117 observations and $d = 2$ states.
- Exponential priors for ψ_1 and ψ_2, and uniform priors for q_{12} and q_{21}.
- Algorithms: Random Walk Metropolis, Multiple-try-Metropolis (N=8), MIIS and MIIS-CV, both with N=8 and N=16 particles.
- Inverse of information matrix for the shape of the RW proposal.
- Run 10,000 iterations after a warmup of 1,000.
- Chain is initialized at maximum likelihood estimator.
- Monte Carlo MSE calculated over 500 replications.

Table: Empirical example for the MMPP model – Monte Carlo MSE of posterior estimates relative to MH.

We define the time adjusted MSE as $(time_{alg} \times mse_{alg})/(time_{MH} \times mse_{MH})$, which approximates the MSE relative to the MH algorithm for the same amount of computing time.

				MIIS		MIIS-CV	
		MH	MTM	N=8	N=16	N=8	N=16
	Acceptance	0.24	0.54	0.42	0.57	0.42	0.57
	IACT	55.2	20.2	13.3	9.4	13.3	9.4
	Time	20	62	55	59	58	63
		Posterior mean					
	ψ_1	1.00	0.45	0.25	0.17	0.18	0.13
	ψ_2	1.00	0.50	0.21	0.15	0.15	0.10
	q_{12}	1.00	0.39	0.20	0.14	0.16	0.10
	q_{21}	1.00	0.52	0.27	0.15	0.15	0.12
		Time adjusted MSE					
	ψ_1	1.00	1.39	0.68	0.50	0.51	0.43
	ψ_2	1.00	1.55	0.59	0.45	0.45	0.33
	q_{12}	1.00	1.21	0.54	0.42	0.46	0.33
	q_{21}	1.00	1.60	0.73	0.45	0.44	0.37

Table: Empirical example for the MMPP model – Monte Carlo MSE of posterior estimates relative to MH.

We define the time adjusted MSE as $(time_{alg} \times mse_{alg})/(time_{MH} \times mse_{MH})$, which approximates the MSE relative to the MH algorithm for the same amount of computing time.

	Posterior variance					
ψ_1	1.00	0.53	0.34	0.19	0.22	0.17
ψ_2	1.00	0.65	0.29	0.21	0.18	0.11
q_{12}	1.00	0.45	0.20	0.12	0.17	0.12
q_{21}	1.00	0.50	0.23	0.14	0.15	0.10
	Time adjusted MSE					
ψ_1	1.00	1.66	0.93	0.56	0.64	0.56
ψ_2	1.00	2.02	0.81	0.63	0.51	0.36
q_{12}	1.00	1.39	0.55	0.35	0.50	0.38
q_{21}	1.00	1.55	0.63	0.41	0.45	0.33

Binary Mixed Logit Model

The Model

1.

$$p(i \text{ chooses } j = 1 \text{at } t | Z_{it}, \delta_i) = \frac{\exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})}{1 + \exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})},$$

where $\delta_i = (\beta_{0i}, ..., \beta_{Li})$ are the utility weights for individual i, and $Z_{it} = (z_{1it}, ..., z_{Lit})'$ is the corresponding vector of attributes.

Binary Mixed Logit Model

The Model

1. $$p(i \text{ chooses } j = 1 \text{at } t | Z_{it}, \delta_i) = \frac{\exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})}{1 + \exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})},$$

   where $\delta_i = (\beta_{0i}, ..., \beta_{Li})$ are the utility weights for individual i, and $Z_{it} = (z_{1it}, ..., z_{Lit})'$ is the corresponding vector of attributes.

2. Individual specific constants are

   $$\beta_{li} = \beta_l + \epsilon_{li}, \quad l = 1, ..., L,$$

   with $\epsilon_{li} \sim N(0, \sigma_l^2)$.

Binary Mixed Logit Model

The Model

1. $$p(i \text{ chooses } j = 1 \text{at } t | Z_{it}, \delta_i) = \frac{\exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})}{1 + \exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})},$$

   where $\delta_i = (\beta_{0i}, ..., \beta_{Li})$ are the utility weights for individual i, and $Z_{it} = (z_{1it}, ..., z_{Lit})'$ is the corresponding vector of attributes.

2. Individual specific constants are

   $$\beta_{li} = \beta_l + \epsilon_{li}, \quad l = 1, ..., L,$$

   with $\epsilon_{li} \sim N(0, \sigma_l^2)$.

1. Parameter vector: $\theta = (\beta_0, \sigma_0, \beta_1, ..., \beta_L, \sigma_1^2, \sigma_L^2)$.

Binary Mixed Logit Model

The Model

1. $$p(i \text{ chooses } j = 1 \text{at } t | Z_{it}, \delta_i) = \frac{\exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})}{1 + \exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})},$$

   where $\delta_i = (\beta_{0i}, ..., \beta_{Li})$ are the utility weights for individual i, and $Z_{it} = (z_{1it}, ..., z_{Lit})'$ is the corresponding vector of attributes.

2. Individual specific constants are

$$\beta_{li} = \beta_l + \epsilon_{li}, \quad l = 1, ..., L,$$

   with $\epsilon_{li} \sim N(0, \sigma_l^2)$.

1. Parameter vector: $\theta = (\beta_0, \sigma_0, \beta_1, ..., \beta_L, \sigma_1^2, \sigma_L^2)$.
2. Latent variables for each individual: $\zeta_i = (\epsilon_{0i}, ..., \epsilon_{Li}, \lambda_i)$

Binary Mixed Logit Model

The Model

1. $$p(i \text{ chooses } j = 1 \text{at } t | Z_{it}, \delta_i) = \frac{\exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})}{1 + \exp(\beta_{0i} + \sum_{l=1}^{L} \beta_{li} z_{lit})},$$

   where $\delta_i = (\beta_{0i}, ..., \beta_{Li})$ are the utility weights for individual i, and $Z_{it} = (z_{1it}, ..., z_{Lit})'$ is the corresponding vector of attributes.

2. Individual specific constants are

   $$\beta_{li} = \beta_l + \epsilon_{li}, \quad l = 1, ..., L,$$

   with $\epsilon_{li} \sim N(0, \sigma_l^2)$.

1. Parameter vector: $\theta = (\beta_0, \sigma_0, \beta_1, ..., \beta_L, \sigma_1^2, \sigma_L^2)$.
2. Latent variables for each individual: $\zeta_i = (\epsilon_{0i}, ..., \epsilon_{Li}, \lambda_i)$
3. βs have Gaussian priors and σs have half-Cauchy priors.

MixLogit: Pap Smear data

1. $I = 79$ women choose whether or not to have Pap smear test on $T = 32$ choice scenarios.
2. We observe the choice of individual i in scenario t
3. $y_{it} = 1$ if the woman chooses to take the test and $y_{it} = 0$ otherwise.
4. We fix β_5 as the MLE for computational simplicity

Table: Choice attributes for the pap smear data set

Choice attributes	Values	Param
Alternative specific constant for test	1	β_0, σ_0
Whether patient knows doctor	0 (no), 1 (yes)	β_1, σ_1
Whether doctor is male	0 (no), 1 (yes)	β_2, σ_2
Whether test is due	0 (no), 1 (yes)	β_3, σ_3
Whether doctor recommends test	0 (no), 1 (yes)	β_4, σ_4
Test cost	$\{0, 10, 20, 30\}/10$	β_5

MixLogit: Results

1. Sampling algorithm: Metropolis-within-Gibbs sampler and MIIS wihting Gibbs sampler with N = 50

Parameter	MwG	MIIS	MIIS-CV
β_0	1.000	0.906	0.067
β_1	1.000	1.227	0.062
β_2	1.000	0.917	0.052
β_3	1.000	0.983	0.062
β_4	1.000	0.662	0.084
σ_0	1.000	0.947	0.068
σ_1	1.000	1.019	0.162
σ_2	1.000	0.943	0.080
σ_3	1.000	1.174	0.076
σ_4	1.000	0.535	0.252

Table: Mean squared errors for the posterior mean estimates relative to MwG.

MixLogit: Results

1. Sampling algorithm: Metropolis-within-Gibbs sampler and MIIS wihting Gibbs sampler with N = 50
2. results are based on 500 independent MCMC and 20,000 iterations after warmup

Parameter	MwG	MIIS	MIIS-CV
β_0	1.000	0.906	0.067
β_1	1.000	1.227	0.062
β_2	1.000	0.917	0.052
β_3	1.000	0.983	0.062
β_4	1.000	0.662	0.084
σ_0	1.000	0.947	0.068
σ_1	1.000	1.019	0.162
σ_2	1.000	0.943	0.080
σ_3	1.000	1.174	0.076
σ_4	1.000	0.535	0.252

Table: Mean squared errors for the posterior mean estimates relative to MwG.

MIIS Theory

1. The marginal chain $\{(y^{(t)}, k^{(t)})\}$ is Markov and ergodic when sampled using MIIS, i.e.,

$$\lim_{t \to \infty} \mid P^t(I, y; \cdot) - N^{-1}\pi(\cdot)|_{TV} = 0,$$

where $P(y, I; B \times \{k\})$ is the Markov transition kernel from (y, I) to $B \times \{k\}$, $B \in \Omega$, $k \in \{1{:}N\}$.

MIIS Theory

1. The marginal chain $\{(y^{(t)}, k^{(t)})\}$ is Markov and ergodic when sampled using MIIS, i.e.,

$$\lim_{t \to \infty} \mid P^t(I, y; \cdot) - N^{-1}\pi(\cdot)|_{TV} = 0,$$

   where $P(y, I; B \times \{k\})$ is the Markov transition kernel from (y, I) to $B \times \{k\}$, $B \in \Omega$, $k \in \{1{:}N\}$.

2. This implies that sample means converge to their corresponding expectations.

MIIS Theory

1. The marginal chain $\{(y^{(t)}, k^{(t)})\}$ is Markov and ergodic when sampled using MIIS, i.e.,

$$\lim_{t \to \infty} \mid P^t(I, y; \cdot) - N^{-1}\pi(\cdot)|_{TV} = 0,$$

   where $P(y, I; B \times \{k\})$ is the Markov transition kernel from (y, I) to $B \times \{k\}$, $B \in \Omega$, $k \in \{1{:}N\}$.
2. This implies that sample means converge to their corresponding expectations.
3. In addition, under some more regularity conditions, the marginal chain is uniformly ergodic. This implies a central limit theorem.

Highlights and Extensions

- Novel MCMC method that allows for use of variance reduction techniques and has good convergence properties.

Highlights and Extensions

- Novel MCMC method that allows for use of variance reduction techniques and has good convergence properties.
- Extends importance sampler methods for a Gibbs-type approach.

Highlights and Extensions

- Novel MCMC method that allows for use of variance reduction techniques and has good convergence properties.
- Extends importance sampler methods for a Gibbs-type approach.
- Significant improvement over traditional methods.

Highlights and Extensions

- Novel MCMC method that allows for use of variance reduction techniques and has good convergence properties.
- Extends importance sampler methods for a Gibbs-type approach.
- Significant improvement over traditional methods.
- Can use more general IS such as IS^2, Annealed importance samplers, and particle filters.

Highlights and Extensions

- Novel MCMC method that allows for use of variance reduction techniques and has good convergence properties.
- Extends importance sampler methods for a Gibbs-type approach.
- Significant improvement over traditional methods.
- Can use more general IS such as IS^2, Annealed importance samplers, and particle filters.
- Adaptive proposals and parallel chains.