Pseudo Marginal MCMC
Or
How to do Exact Inference with Approximate Methods and Playing Russian Roulette

Mark Girolami

Department of Statistics
University of Warwick

MLSS 2014

May, 2014

Joint Work

Talk Outline

- Motivation

Talk Outline

- Motivation
- Pseudo-Marginal Markov chain Monte Carlo

Talk Outline

- Motivation
- Pseudo-Marginal Markov chain Monte Carlo
- Hierarchic Gaussian Process Model working Example

Talk Outline

- Motivation
- Pseudo-Marginal Markov chain Monte Carlo
- Hierarchic Gaussian Process Model working Example
- Infinite Series Expansion of Likelihood

Talk Outline

- Motivation
- Pseudo-Marginal Markov chain Monte Carlo
- Hierarchic Gaussian Process Model working Example
- Infinite Series Expansion of Likelihood
- Targeting Absolute Measure via MCMC

Talk Outline

- Motivation
- Pseudo-Marginal Markov chain Monte Carlo
- Hierarchic Gaussian Process Model working Example
- Infinite Series Expansion of Likelihood
- Targeting Absolute Measure via MCMC
- Correction of Biased Estimators by Geometric & Exponential Tilting

Talk Outline

- Motivation
- Pseudo-Marginal Markov chain Monte Carlo
- Hierarchic Gaussian Process Model working Example
- Infinite Series Expansion of Likelihood
- Targeting Absolute Measure via MCMC
- Correction of Biased Estimators by Geometric & Exponential Tilting
- Russian Roulette Truncation and PM MCMC

Talk Outline

- Motivation
- Pseudo-Marginal Markov chain Monte Carlo
- Hierarchic Gaussian Process Model working Example
- Infinite Series Expansion of Likelihood
- Targeting Absolute Measure via MCMC
- Correction of Biased Estimators by Geometric & Exponential Tilting
- Russian Roulette Truncation and PM MCMC
- MCMC Sampling - Ising Lattice Model

Talk Outline

- Motivation
- Pseudo-Marginal Markov chain Monte Carlo
- Hierarchic Gaussian Process Model working Example
- Infinite Series Expansion of Likelihood
- Targeting Absolute Measure via MCMC
- Correction of Biased Estimators by Geometric & Exponential Tilting
- Russian Roulette Truncation and PM MCMC
- MCMC Sampling - Ising Lattice Model
- Large Scale GMRF Ozone Column Exact MCMC Posterior Sampling

Talk Outline

- Motivation
- Pseudo-Marginal Markov chain Monte Carlo
- Hierarchic Gaussian Process Model working Example
- Infinite Series Expansion of Likelihood
- Targeting Absolute Measure via MCMC
- Correction of Biased Estimators by Geometric & Exponential Tilting
- Russian Roulette Truncation and PM MCMC
- MCMC Sampling - Ising Lattice Model
- Large Scale GMRF Ozone Column Exact MCMC Posterior Sampling
- Conclusions and Discussion

Pseudo-Marginal Bayesian Inference for Gaussian Processes

- Challenge to carry out exact Bayesian inference and how to account for uncertainty on model parameters when making model-based predictions on out-of-sample data

Pseudo-Marginal Bayesian Inference for Gaussian Processes

- Challenge to carry out exact Bayesian inference and how to account for uncertainty on model parameters when making model-based predictions on out-of-sample data
- Exact Posterior Marginalisation is Hard

Pseudo-Marginal Bayesian Inference for Gaussian Processes

- Challenge to carry out exact Bayesian inference and how to account for uncertainty on model parameters when making model-based predictions on out-of-sample data
- Exact Posterior Marginalisation is Hard
- Using probit regression as an illustrative tutorial example, will present the pseudo-marginal approach to Markov chain Monte Carlo that efficiently addresses both of these issues.

Pseudo-Marginal Bayesian Inference for Gaussian Processes

- Challenge to carry out exact Bayesian inference and how to account for uncertainty on model parameters when making model-based predictions on out-of-sample data

- Exact Posterior Marginalisation is Hard

- Using probit regression as an illustrative tutorial example, will present the pseudo-marginal approach to Markov chain Monte Carlo that efficiently addresses both of these issues.

- This is particularly important as it offers a powerful tool to carry out full Bayesian inference of Gaussian Process based hierarchic statistical models in general.

Pseudo-Marginal Bayesian Inference for Gaussian Processes

- Challenge to carry out exact Bayesian inference and how to account for uncertainty on model parameters when making model-based predictions on out-of-sample data
- Exact Posterior Marginalisation is Hard
- Using probit regression as an illustrative tutorial example, will present the pseudo-marginal approach to Markov chain Monte Carlo that efficiently addresses both of these issues.
- This is particularly important as it offers a powerful tool to carry out full Bayesian inference of Gaussian Process based hierarchic statistical models in general.
- Empirically indicates Monte Carlo based integration of all model parameters is actually feasible in this class of models providing a superior quantification of uncertainty in predictions.

Pseudo-Marginal Bayesian Inference for Gaussian Processes

- Challenge to carry out exact Bayesian inference and how to account for uncertainty on model parameters when making model-based predictions on out-of-sample data
- Exact Posterior Marginalisation is Hard
- Using probit regression as an illustrative tutorial example, will present the pseudo-marginal approach to Markov chain Monte Carlo that efficiently addresses both of these issues.
- This is particularly important as it offers a powerful tool to carry out full Bayesian inference of Gaussian Process based hierarchic statistical models in general.
- Empirically indicates Monte Carlo based integration of all model parameters is actually feasible in this class of models providing a superior quantification of uncertainty in predictions.
- Extensive comparisons with respect to state-of-the-art probabilistic classifiers support this assertion.

Simple Gaussian Process Model

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$ with $y_i \in \{-1, +1\}$

Simple Gaussian Process Model

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$ with $y_i \in \{-1, +1\}$
- Let $\mathbf{f} = \{f_1, \ldots, f_n\}$ be a set of latent functions $\mathbf{f} \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, K)$ with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$ the function modeling the covariance between latent variables evaluated at the input vectors, parameterized by a vector of hyper-parameters $\boldsymbol{\theta}$.

Simple Gaussian Process Model

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$ with $y_i \in \{-1, +1\}$
- Let $\mathbf{f} = \{f_1, \ldots, f_n\}$ be a set of latent functions $\mathbf{f} \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, K)$ with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$ the function modeling the covariance between latent variables evaluated at the input vectors, parameterized by a vector of hyper-parameters $\boldsymbol{\theta}$.
- The data modelled as $p(y_i|f_i) = \Phi(y_i f_i)$ with $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i)$.

Simple Gaussian Process Model

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$ with $y_i \in \{-1, +1\}$
- Let $\mathbf{f} = \{f_1, \ldots, f_n\}$ be a set of latent functions $\mathbf{f} \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, K)$ with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$ the function modeling the covariance between latent variables evaluated at the input vectors, parameterized by a vector of hyper-parameters $\boldsymbol{\theta}$.
- The data modelled as $p(y_i|f_i) = \Phi(y_i f_i)$ with $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i)$.
- The GP classification model is hierarchical, as **y** is conditioned on **f**, and **f** is conditioned on $\boldsymbol{\theta}$ and the inputs X.

Simple Gaussian Process Model

- Let $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a set of n input vectors described by d covariates and associated with observed univariate responses $\mathbf{y} = \{y_1, \dots, y_n\}$ with $y_i \in \{-1, +1\}$
- Let $\mathbf{f} = \{f_1, \dots, f_n\}$ be a set of latent functions $\mathbf{f} \sim \mathcal{N}(\mathbf{f}|\mathbf{0}, K)$ with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta})$ the function modeling the covariance between latent variables evaluated at the input vectors, parameterized by a vector of hyper-parameters $\boldsymbol{\theta}$.
- The data modelled as $p(y_i|f_i) = \Phi(y_i f_i)$ with $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i)$.
- The GP classification model is hierarchical, as $\mathbf{y}$ is conditioned on $\mathbf{f}$, and $\mathbf{f}$ is conditioned on $\boldsymbol{\theta}$ and the inputs X.
- Require

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f}, \boldsymbol{\theta})p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})df_* d\mathbf{f} d\boldsymbol{\theta}.$$

Approximate Inference

- Object of interest

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f},\boldsymbol{\theta})p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})df_*d\mathbf{f}d\boldsymbol{\theta}.$$

Approximate Inference

- Object of interest

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f}, \boldsymbol{\theta})p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})df_*d\mathbf{f}d\boldsymbol{\theta}.$$

- Laplace Approximation

Approximate Inference

- Object of interest

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f},\boldsymbol{\theta})p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})df_*d\mathbf{f}d\boldsymbol{\theta}.$$

- Laplace Approximation
- Variational Approximation

Approximate Inference

- Object of interest

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f},\boldsymbol{\theta})p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})df_*d\mathbf{f}d\boldsymbol{\theta}.$$

- Laplace Approximation
- Variational Approximation
- Expectation Propagation

Approximate Inference

- Object of interest

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f},\boldsymbol{\theta})p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})df_*d\mathbf{f}d\boldsymbol{\theta}.$$

- Laplace Approximation
- Variational Approximation
- Expectation Propagation
- Maximum Approximate Marginal Likelihood

Approximate Inference

- Object of interest

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f},\boldsymbol{\theta})p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})df_*d\mathbf{f}d\boldsymbol{\theta}.$$

- Laplace Approximation
- Variational Approximation
- Expectation Propagation
- Maximum Approximate Marginal Likelihood
- Monte Carlo to tackle intractability in characterizing $p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})$

Approximate Inference

- Object of interest

$$p(y_*|\mathbf{y}) = \int p(y_*|f_*)p(f_*|\mathbf{f}, \boldsymbol{\theta})p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})df_* d\mathbf{f} d\boldsymbol{\theta}.$$

- Laplace Approximation
- Variational Approximation
- Expectation Propagation
- Maximum Approximate Marginal Likelihood
- Monte Carlo to tackle intractability in characterizing $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$
- Draw samples from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$ using MCMC methods, so that a Monte Carlo estimate of the predictive distribution can be used

$$p(y_*|\mathbf{y}) \simeq \frac{1}{N}\sum_{i=1}^{N}\int p(y_*|f_*)p(f_*|\mathbf{f}^{(i)}, \boldsymbol{\theta}^{(i)})df_*,$$

where $\mathbf{f}^{(i)}, \boldsymbol{\theta}^{(i)}$ denotes the ith sample from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$.

MCMC Posterior Sampling from $p(\mathbf{f}, \theta | \mathbf{y})$

- Sampling from the posterior over $\mathbf{f}$ and θ by joint proposals is not feasible; it is extremely unlikely to propose a set of latent variables and hyper-parameters that are compatible with each other and observed data.

MCMC Posterior Sampling from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$

- Sampling from the posterior over $\mathbf{f}$ and $\boldsymbol{\theta}$ by joint proposals is not feasible; it is extremely unlikely to propose a set of latent variables and hyper-parameters that are compatible with each other and observed data.
- In order to draw samples from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$, it is therefore necessary to resort to a Gibbs sampler, whereby $\mathbf{f}$ and $\boldsymbol{\theta}$ are updated in turn.

MCMC Posterior Sampling from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$

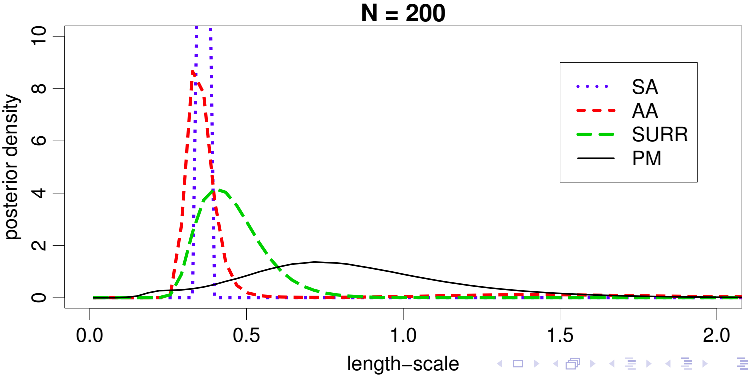
- Sampling from the posterior over $\mathbf{f}$ and $\boldsymbol{\theta}$ by joint proposals is not feasible; it is extremely unlikely to propose a set of latent variables and hyper-parameters that are compatible with each other and observed data.
- In order to draw samples from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$, it is therefore necessary to resort to a Gibbs sampler, whereby $\mathbf{f}$ and $\boldsymbol{\theta}$ are updated in turn.
- Sampling from $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$, Elliptic Slice Sampling, HMC

MCMC Posterior Sampling from $p(\mathbf{f}, \boldsymbol{\theta} | \mathbf{y})$

- Sampling from the posterior over $\mathbf{f}$ and $\boldsymbol{\theta}$ by joint proposals is not feasible; it is extremely unlikely to propose a set of latent variables and hyper-parameters that are compatible with each other and observed data.
- In order to draw samples from $p(\mathbf{f}, \boldsymbol{\theta} | \mathbf{y})$, it is therefore necessary to resort to a Gibbs sampler, whereby $\mathbf{f}$ and $\boldsymbol{\theta}$ are updated in turn.
- Sampling from $p(\mathbf{f} | \mathbf{y}, \boldsymbol{\theta})$, Elliptic Slice Sampling, HMC
- Sampling from $p(\boldsymbol{\theta} | \mathbf{f}, \mathbf{y})$, problematic requiring reparametrisation

MCMC Posterior Sampling from $p(\mathbf{f}, \boldsymbol{\theta} | \mathbf{y})$

- Sampling from the posterior over $\mathbf{f}$ and $\boldsymbol{\theta}$ by joint proposals is not feasible; it is extremely unlikely to propose a set of latent variables and hyper-parameters that are compatible with each other and observed data.
- In order to draw samples from $p(\mathbf{f}, \boldsymbol{\theta} | \mathbf{y})$, it is therefore necessary to resort to a Gibbs sampler, whereby $\mathbf{f}$ and $\boldsymbol{\theta}$ are updated in turn.
- Sampling from $p(\mathbf{f} | \mathbf{y}, \boldsymbol{\theta})$, Elliptic Slice Sampling, HMC
- Sampling from $p(\boldsymbol{\theta} | \mathbf{f}, \mathbf{y})$, problematic requiring reparametrisation



MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$

- The use of reparameterization techniques mitigates the problems due to the coupling of latent variables and hyper-parameters, but sampling efficiency for GP models is still an issue

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$

- The use of reparameterization techniques mitigates the problems due to the coupling of latent variables and hyper-parameters, but sampling efficiency for GP models is still an issue
- Intuitively, the best strategy to break the correlation between latent variables and hyper-parameters in sampling from the posterior over the hyper-parameters would be to integrate out the latent variables altogether.

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$

- The use of reparameterization techniques mitigates the problems due to the coupling of latent variables and hyper-parameters, but sampling efficiency for GP models is still an issue

- Intuitively, the best strategy to break the correlation between latent variables and hyper-parameters in sampling from the posterior over the hyper-parameters would be to integrate out the latent variables altogether.

- This is not possible, but here we present a strategy that uses an unbiased estimate of the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ to devise an MCMC strategy that produces samples from the correct posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$.

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$, the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$, the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

  $$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

  In order to do that, we would need to integrate out the latent variables:

  $$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$$

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$, the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

  $$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

  In order to do that, we would need to integrate out the latent variables:

  $$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$$

  and use this along with the prior $p(\boldsymbol{\theta})$ in the Hastings ratio:

  $$z = \frac{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})}$$

  As already discussed, analytically integrating out $\mathbf{f}$ is not possible.

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$, the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

  $$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

  In order to do that, we would need to integrate out the latent variables:

  $$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$$

  and use this along with the prior $p(\boldsymbol{\theta})$ in the Hastings ratio:

  $$z = \frac{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})}$$

  As already discussed, analytically integrating out **f** is not possible.
- Resort to approximations

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$, the Pseudo-Marginal Approach

- We are interested in sampling from the posterior distribution

  $$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

  In order to do that, we would need to integrate out the latent variables:

  $$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$$

  and use this along with the prior $p(\boldsymbol{\theta})$ in the Hastings ratio:

  $$z = \frac{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})}$$

  As already discussed, analytically integrating out **f** is not possible.
- Resort to approximations and still retain exactness of MCMC

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$, the Pseudo-Marginal Approach

- We could just plug into the Hastings ratio an estimate $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ of the marginal $p(\mathbf{y}|\boldsymbol{\theta})$.

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$, the Pseudo-Marginal Approach

- We could just plug into the Hastings ratio an estimate $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ of the marginal $p(\mathbf{y}|\boldsymbol{\theta})$.
- If the estimate of the margin is unbiased and positive, then the sampler will draw samples from the correct exact posterior $p(\boldsymbol{\theta}|\mathbf{y})$.

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$, the Pseudo-Marginal Approach

- We could just plug into the Hastings ratio an estimate $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ of the marginal $p(\mathbf{y}|\boldsymbol{\theta})$.
- If the estimate of the margin is unbiased and positive, then the sampler will draw samples from the correct exact posterior $p(\boldsymbol{\theta}|\mathbf{y})$.

$$\tilde{z} = \frac{\tilde{p}(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{\tilde{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})}$$

MCMC Posterior Sampling from $p(\boldsymbol{\theta}|\mathbf{y})$, the Pseudo-Marginal Approach

- We could just plug into the Hastings ratio an estimate $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ of the marginal $p(\mathbf{y}|\boldsymbol{\theta})$.
- If the estimate of the margin is unbiased and positive, then the sampler will draw samples from the correct exact posterior $p(\boldsymbol{\theta}|\mathbf{y})$.

$$\tilde{z} = \frac{\tilde{p}(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{\tilde{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})}$$

- This result is remarkable as it gives a simple recipe to be used in hierarchical models to tackle the problem of strong coupling between groups of variables when using MCMC algorithms.

Exploiting Approximate Posteriors in the Pseudo-Marginal Approach

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ for the marginal $p(\mathbf{y}|\boldsymbol{\theta})$, we propose to employ importance sampling.

Exploiting Approximate Posteriors in the Pseudo-Marginal Approach

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ for the marginal $p(\mathbf{y}|\boldsymbol{\theta})$, we propose to employ importance sampling.
- We draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$, so that we can approximate the marginal $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$ by:

Exploiting Approximate Posteriors in the Pseudo-Marginal Approach

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ for the marginal $p(\mathbf{y}|\boldsymbol{\theta})$, we propose to employ importance sampling.
- We draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$, so that we can approximate the marginal $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$ by:

$$\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y}|\mathbf{f}_i)p(\mathbf{f}_i|\boldsymbol{\theta})}{q(\mathbf{f}_i|\mathbf{y}, \boldsymbol{\theta})}$$

Exploiting Approximate Posteriors in the Pseudo-Marginal Approach

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ for the marginal $p(\mathbf{y}|\boldsymbol{\theta})$, we propose to employ importance sampling.

- We draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$, so that we can approximate the marginal $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$ by:

$$\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \simeq \frac{1}{N_{\mathrm{imp}}} \sum_{i=1}^{N_{\mathrm{imp}}} \frac{p(\mathbf{y}|\mathbf{f}_i)p(\mathbf{f}_i|\boldsymbol{\theta})}{q(\mathbf{f}_i|\mathbf{y}, \boldsymbol{\theta})}$$

- It is easy to verify that the approximation yields an unbiased estimate of $p(\mathbf{y}|\boldsymbol{\theta})$, as its expectation is the exact marginal $p(\mathbf{y}|\boldsymbol{\theta})$.

Exploiting Approximate Posteriors in the Pseudo-Marginal Approach

- In order to obtain an unbiased estimator $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ for the marginal $p(\mathbf{y}|\boldsymbol{\theta})$, we propose to employ importance sampling.

- We draw N_{imp} samples $\mathbf{f}_i$ from the approximating distribution $q(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$, so that we can approximate the marginal $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$ by:

$$\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \simeq \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \frac{p(\mathbf{y}|\mathbf{f}_i)p(\mathbf{f}_i|\boldsymbol{\theta})}{q(\mathbf{f}_i|\mathbf{y}, \boldsymbol{\theta})}$$

- It is easy to verify that the approximation yields an unbiased estimate of $p(\mathbf{y}|\boldsymbol{\theta})$, as its expectation is the exact marginal $p(\mathbf{y}|\boldsymbol{\theta})$.

- Therefore, this estimate can be used in the Hastings ratio to construct an MCMC approach that samples from the correct invariant distribution $p(\boldsymbol{\theta}|\mathbf{y})$.

Exploiting Approximate Posteriors in the Pseudo-Marginal Approach

Algorithm 1 Pseudo-marginal MH transition operator to sample $\boldsymbol{\theta}$.

Input: The current pair $(\theta, \tilde{p}(\mathbf{y}|\boldsymbol{\theta}))$, a routine to approximate $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$ by $q(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$, and number of importance samples N_{imp}

Output: A new pair $(\theta, \tilde{p}(\mathbf{y}|\boldsymbol{\theta}))$

```
Draw $\boldsymbol{\theta}'$ from the proposal distribution $\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})$
Approximate $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}')$ by $q(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}')$
Draw $N_{\mathrm{imp}}$ samples from $q(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}')$
Compute $\tilde{p}(\mathbf{y}|\boldsymbol{\theta}')$ using IMPORTANCE SAMPLER
Compute $A = \min\left\{1, \frac{\tilde{p}(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{\tilde{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$
Draw $u$ from $U_{[0,1]}$
if $A > u$ then
    return $(\boldsymbol{\theta}', \tilde{p}(\mathbf{y}|\boldsymbol{\theta}'))$
else
    return $(\boldsymbol{\theta}, \tilde{p}(\mathbf{y}|\boldsymbol{\theta}))$
end if
```

Impact of Approximating distribution

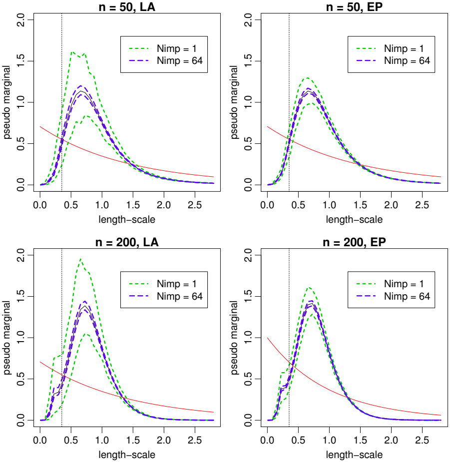


Figure: Plot of the PM as a function of the length-scale τ; black solid lines represent the average over 500 repetitions and dashed lines represent 2.5th and 97.5th quantiles for $N_{\text{imp}} = 1$ and $N_{\text{imp}} = 64$. The solid red line is the prior density.

Sampling Efficiency

Breast $n = 682$

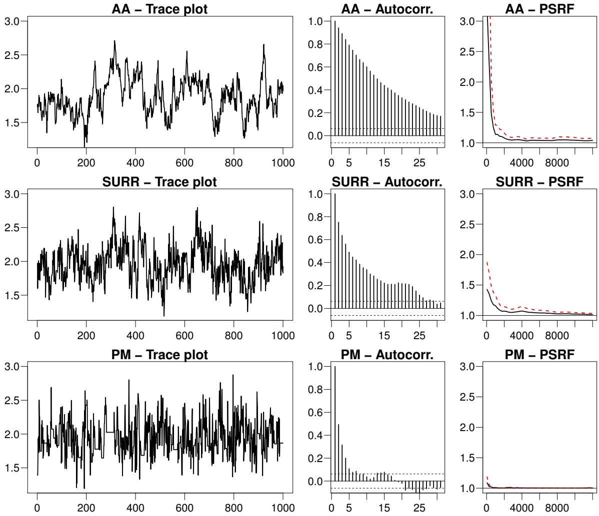
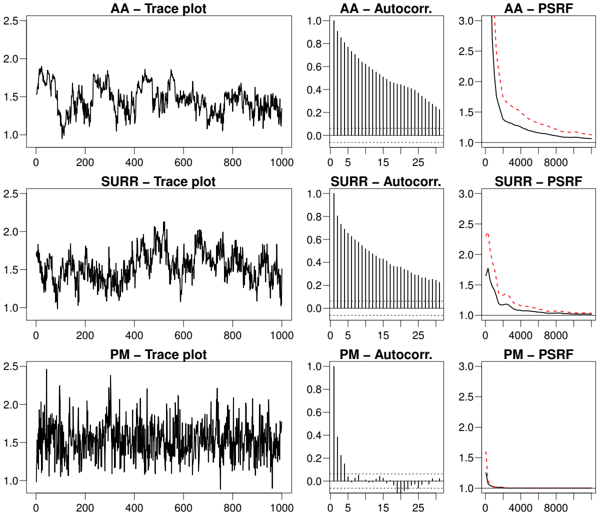


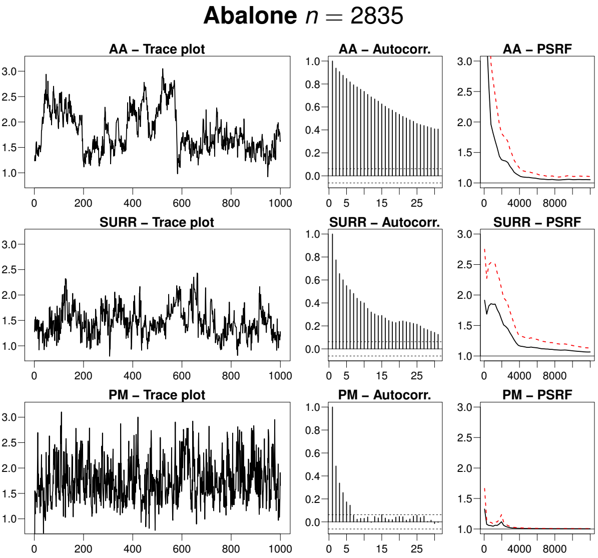
Figure: Summary of efficiency and convergence speed on Breast data set. All plots show the sampling of the logarithm of the length-scale parameter τ. The right panel reports the evolution of the PSRF after burn-in; in this plot the solid line and the red dashed line represent the median and the 97.5% percentile respectively.

Sampling Efficiency

Pima $n = 768$



Sampling Efficiency



Predictive Performance

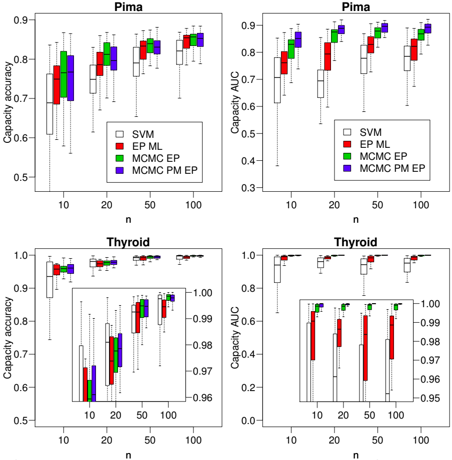


Figure: Plots of performance scores with respect to size of training set for the Pima (first row) and the Thyroid (second row) data sets. The legend is reported in the first row only and it applies to all the plots. In the remaining plots, a closeup is reported to make it easier to compare the results.

Predictive Performance

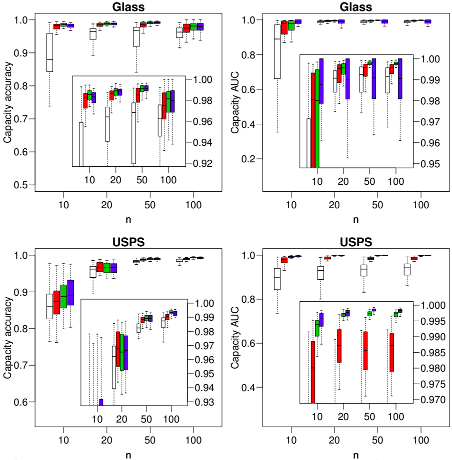
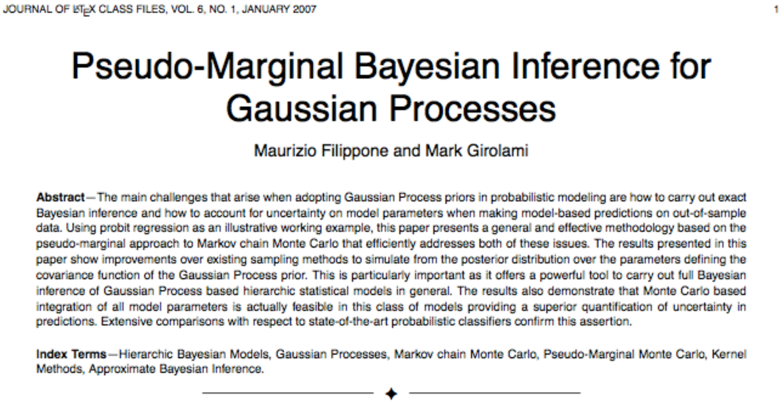


Figure: Plots of performance scores with respect to size of training set for the Glass (first row) and the USPS (second row) data sets. The legend is reported in the first row only and it applies to all the plots. In the remaining plots, a closeup is reported to make it easier to compare the results.

IEEE Trans PAMI



Pseudo-Marginal Bayesian Inference for Gaussian Processes

Maurizio Filippone and Mark Girolami

Abstract—The main challenges that arise when adopting Gaussian Process priors in probabilistic modeling are how to carry out exact Bayesian inference and how to account for uncertainty on model parameters when making model-based predictions on out-of-sample data. Using probit regression as an illustrative working example, this paper presents a general and effective methodology based on the pseudo-marginal approach to Markov chain Monte Carlo that efficiently addresses both of these issues. The results presented in this paper show improvements over existing sampling methods to simulate from the posterior distribution over the parameters defining the covariance function of the Gaussian Process prior. This is particularly important as it offers a powerful tool to carry out full Bayesian inference of Gaussian Process based hierarchic statistical models in general. The results also demonstrate that Monte Carlo based integration of all model parameters is actually feasible in this class of models providing a superior quantification of uncertainty in predictions. Extensive comparisons with respect to state-of-the-art probabilistic classifiers confirm this assertion.

Index Terms—Hierarchic Bayesian Models, Gaussian Processes, Markov chain Monte Carlo, Pseudo-Marginal Monte Carlo, Kernel Methods, Approximate Bayesian Inference.

◆

Motivation

- Bayesian inference data $\mathbf{y} \in \mathcal{Y}$, posterior inference for variables $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

Motivation

- Bayesian inference data $\mathbf{y} \in \mathcal{Y}$, posterior inference for variables $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
- Prior $\pi(\boldsymbol{\theta})$, data density $p(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}; \boldsymbol{\theta})/\mathcal{Z}(\boldsymbol{\theta})$ with $\mathcal{Z}(\boldsymbol{\theta}) = \int f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$

Motivation

- Bayesian inference data $\mathbf{y} \in \mathcal{Y}$, posterior inference for variables $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
- Prior $\pi(\boldsymbol{\theta})$, data density $p(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta})/\mathcal{Z}(\boldsymbol{\theta})$ with $\mathcal{Z}(\boldsymbol{\theta}) = \int f(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}$
- *Doubly-Intractable* Posterior follows as

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})} = \frac{f(\mathbf{y};\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})}$$

where $\mathcal{Z}(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$

Motivation

- Bayesian inference data $\mathbf{y} \in \mathcal{Y}$, posterior inference for variables $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
- Prior $\pi(\boldsymbol{\theta})$, data density $p(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}; \boldsymbol{\theta})/\mathcal{Z}(\boldsymbol{\theta})$ with $\mathcal{Z}(\boldsymbol{\theta}) = \int f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x}$
- *Doubly-Intractable* Posterior follows as

  $$\pi(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})} = \frac{f(\mathbf{y}; \boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})}$$

  where $\mathcal{Z}(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$
- Bayesian inference proceeds by taking posterior expectations of functions of interest i.e.

  $$E_{\pi(\boldsymbol{\theta}|\mathbf{y})}\left\{\varphi(\boldsymbol{\theta})\right\} = \int \varphi(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

Motivation

- Construct Markov chain whose invariant distribution has density $\pi(\boldsymbol{\theta}|\mathbf{y})$ via transition kernel constructed by employing $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ and acceptance probability

$$\alpha(\boldsymbol{\theta}',\boldsymbol{\theta}) = \min\left\{1, \frac{f(\mathbf{y};\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')}{f(\mathbf{y};\boldsymbol{\theta})\pi(\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \times \frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')}\right\}$$

Motivation

- Construct Markov chain whose invariant distribution has density $\pi(\boldsymbol{\theta}|\mathbf{y})$ via transition kernel constructed by employing $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ and acceptance probability

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{f(\mathbf{y}; \boldsymbol{\theta}')\pi(\boldsymbol{\theta}')}{f(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \times \frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')}\right\}$$

- If $\mathcal{Z}(\boldsymbol{\theta}')$ is non-analytic or non-computable kernel infeasible

Motivation

- Construct Markov chain whose invariant distribution has density $\pi(\boldsymbol{\theta}|\mathbf{y})$ via transition kernel constructed by employing $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ and acceptance probability

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{f(\mathbf{y}; \boldsymbol{\theta}')\pi(\boldsymbol{\theta}')}{f(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \times \frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')}\right\}$$

- If $\mathcal{Z}(\boldsymbol{\theta}')$ is non-analytic or non-computable kernel infeasible
- Biased approximations e.g. pseudo-likelihoods, plugin $\hat{\mathcal{Z}}(\boldsymbol{\theta}')$ estimates

Motivation

- Construct Markov chain whose invariant distribution has density $\pi(\boldsymbol{\theta}|\mathbf{y})$ via transition kernel constructed by employing $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ and acceptance probability

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{f(\mathbf{y}; \boldsymbol{\theta}')\pi(\boldsymbol{\theta}')}{f(\mathbf{y}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \times \frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')}\right\}$$

- If $\mathcal{Z}(\boldsymbol{\theta}')$ is non-analytic or non-computable kernel infeasible
- Biased approximations e.g. pseudo-likelihoods, plugin $\hat{\mathcal{Z}}(\boldsymbol{\theta}')$ estimates
- Do not wish to sacrifice exactness of MCMC (simulation or expectation)

Motivation

- Directional Statistics and distributions on manifolds

Motivation

- Directional Statistics and distributions on manifolds
- Machine Learning - Boltzman Machines, Deep Learning

Motivation

- Directional Statistics and distributions on manifolds
- Machine Learning - Boltzman Machines, Deep Learning
- Diffusion Processes

Motivation

- Directional Statistics and distributions on manifolds
- Machine Learning - Boltzman Machines, Deep Learning
- Diffusion Processes
- Markov Random Fields - Ising, Potts Colouring, Autologistic, Spatial Point Processes

Motivation

- Directional Statistics and distributions on manifolds
- Machine Learning - Boltzman Machines, Deep Learning
- Diffusion Processes
- Markov Random Fields - Ising, Potts Colouring, Autologistic, Spatial Point Processes
- Large Scale Gaussian Markov Random Fields

Motivation

- Directional Statistics and distributions on manifolds
- Machine Learning - Boltzman Machines, Deep Learning
- Diffusion Processes
- Markov Random Fields - Ising, Potts Colouring, Autologistic, Spatial Point Processes
- Large Scale Gaussian Markov Random Fields
- Statistical Models of Network Connectivity

Existing Approaches to Solution

- Unbiased plugin estimate Møller *et al*, 2006 and Murray *et al* 2006

Existing Approaches to Solution

- Unbiased plugin estimate Møller *et al*, 2006 and Murray *et al* 2006

$$\frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')} \approx \frac{f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x};\boldsymbol{\theta}')} \quad \text{where} \quad \mathbf{x} \sim \frac{f(\mathbf{x};\boldsymbol{\theta}')}{\mathcal{Z}(\boldsymbol{\theta}')}$$

Existing Approaches to Solution

- Unbiased plugin estimate Møller *et al*, 2006 and Murray *et al* 2006

$$\frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')} \approx \frac{f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x};\boldsymbol{\theta}')} \quad \text{where} \quad \mathbf{x} \sim \frac{f(\mathbf{x};\boldsymbol{\theta}')}{\mathcal{Z}(\boldsymbol{\theta}')}$$

- Major methodological step forward in addressing *Doubly-Intractable* problem

Existing Approaches to Solution

- Unbiased plugin estimate Møller *et al*, 2006 and Murray *et al* 2006

$$\frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')} \approx \frac{f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x};\boldsymbol{\theta}')} \quad \text{where} \quad \mathbf{x} \sim \frac{f(\mathbf{x};\boldsymbol{\theta}')}{\mathcal{Z}(\boldsymbol{\theta}')}$$

- Major methodological step forward in addressing *Doubly-Intractable* problem
- Require to simulate from model - exploit Perfect Sampling where possible

Exact Approximate Methods

- Pseudo-Marginal construction -

Exact Approximate Methods

- Pseudo-Marginal construction - Simply a miraculous result

Exact Approximate Methods

- Pseudo-Marginal construction - Simply a miraculous result
- Beaumont (2003); Andrieu and Roberts (2009); Doucet *et al* (2012)

Exact Approximate Methods

- Pseudo-Marginal construction - Simply a miraculous result
- Beaumont (2003); Andrieu and Roberts (2009); Doucet *et al* (2012)
- Obtain unbiased, positive estimate of target posterior and use in acceptance expression

Exact Approximate Methods

- Pseudo-Marginal construction - Simply a miraculous result
- Beaumont (2003); Andrieu and Roberts (2009); Doucet *et al* (2012)
- Obtain unbiased, positive estimate of target posterior and use in acceptance expression

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{\hat{\pi}(\boldsymbol{\theta}'|\mathbf{y})}{\hat{\pi}(\boldsymbol{\theta}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$$

Exact Approximate Methods

- Pseudo-Marginal construction - Simply a miraculous result
- Beaumont (2003); Andrieu and Roberts (2009); Doucet *et al* (2012)
- Obtain unbiased, positive estimate of target posterior and use in acceptance expression

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{\hat{\pi}(\boldsymbol{\theta}'|\mathbf{y})}{\hat{\pi}(\boldsymbol{\theta}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$$

- Transition kernel has invariant distribution with target density $\pi(\boldsymbol{\theta}|\mathbf{y})$

Exact Approximate Methods

- Pseudo-Marginal construction - Simply a miraculous result
- Beaumont (2003); Andrieu and Roberts (2009); Doucet *et al* (2012)
- Obtain unbiased, positive estimate of target posterior and use in acceptance expression

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min \left\{ 1, \ \frac{\hat{\pi}(\boldsymbol{\theta}'|\mathbf{y})}{\hat{\pi}(\boldsymbol{\theta}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})} \right\}$$

- Transition kernel has invariant distribution with target density $\pi(\boldsymbol{\theta}|\mathbf{y})$
- Historical Note - Pseudo-Marginal Result exploited in Bosonic Gauge Theory literature almost 30 years ago e.g. Bhanot and Kennedy (1985) - predating Beaumont (2003)

Exact Approximate Methods

- Consequence of Monte Carlo error appearing in estimate of target

Exact Approximate Methods

- Consequence of Monte Carlo error appearing in estimate of target
- Represent Monte Carlo error with r.v. $\boldsymbol{\xi} \sim P_\theta$ and $\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$

Exact Approximate Methods

- Consequence of Monte Carlo error appearing in estimate of target
- Represent Monte Carlo error with r.v. $\boldsymbol{\xi} \sim P_\theta$ and $\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$
- MCMC with target

$$\Pi(d\boldsymbol{\theta}, d\boldsymbol{\xi}|\mathbf{y}) := \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) d\boldsymbol{\theta} P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})$$

Exact Approximate Methods

- Consequence of Monte Carlo error appearing in estimate of target
- Represent Monte Carlo error with r.v. $\boldsymbol{\xi} \sim P_\theta$ and $\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$
- MCMC with target

$$\Pi(d\boldsymbol{\theta}, d\boldsymbol{\xi}|\mathbf{y}) := \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})d\boldsymbol{\theta}P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})$$

- and overall proposal

$$Q(\boldsymbol{\theta}, \boldsymbol{\xi}; d\boldsymbol{\theta}', d\boldsymbol{\xi}') := q(\boldsymbol{\theta}'|\boldsymbol{\theta})d\boldsymbol{\theta}'P_{\boldsymbol{\theta}'}(d\boldsymbol{\xi}')$$

Exact Approximate Methods

- Consequence of Monte Carlo error appearing in estimate of target
- Represent Monte Carlo error with r.v. $\boldsymbol{\xi} \sim P_\theta$ and $\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$
- MCMC with target

$$\Pi(d\boldsymbol{\theta}, d\boldsymbol{\xi}|\mathbf{y}) := \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) d\boldsymbol{\theta} P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})$$

- and overall proposal

$$Q(\boldsymbol{\theta}, \boldsymbol{\xi}; d\boldsymbol{\theta}', d\boldsymbol{\xi}') := q(\boldsymbol{\theta}'|\boldsymbol{\theta}) d\boldsymbol{\theta}' P_{\boldsymbol{\theta}'}(d\boldsymbol{\xi}')$$

- has acceptance probability

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \ \frac{\hat{\pi}(\boldsymbol{\theta}'|\mathbf{y})}{\hat{\pi}(\boldsymbol{\theta}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$$

Exact Approximate Methods

- Consequence of Monte Carlo error appearing in estimate of target
- Represent Monte Carlo error with r.v. $\boldsymbol{\xi} \sim P_\theta$ and $\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$
- MCMC with target

$$\Pi(d\boldsymbol{\theta}, d\boldsymbol{\xi}|\mathbf{y}) := \pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) d\boldsymbol{\theta} P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})$$

- and overall proposal

$$Q(\boldsymbol{\theta}, \boldsymbol{\xi}; d\boldsymbol{\theta}', d\boldsymbol{\xi}') := q(\boldsymbol{\theta}'|\boldsymbol{\theta}) d\boldsymbol{\theta}' P_{\boldsymbol{\theta}'}(d\boldsymbol{\xi}')$$

- has acceptance probability

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{\hat{\pi}(\boldsymbol{\theta}'|\mathbf{y})}{\hat{\pi}(\boldsymbol{\theta}|\mathbf{y})} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$$

- Given unbiasedness of $\pi(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$, $\Pi(d\theta, d\xi|\mathbf{y})$ admits the required target $\pi(\boldsymbol{\theta}|\mathbf{y})$ as its marginal distribution.

Infinite Series Expansion Construction

- For each $\boldsymbol{\theta}$ and $\boldsymbol{y}$, construct random variable $\{V_{\boldsymbol{\theta}}^{(j)},\ j \geq 0\}$ such that

$$\hat{\pi}(\boldsymbol{\theta}, \{V_{\boldsymbol{\theta}}^{(j)}\}|\mathbf{y}) := \sum_{j=0}^{\infty} V_{\boldsymbol{\theta}}^{(j)}$$

Infinite Series Expansion Construction

- For each $\boldsymbol{\theta}$ and $\boldsymbol{y}$, construct random variable $\{V_{\boldsymbol{\theta}}^{(j)},\ j \geq 0\}$ such that

$$\hat{\pi}(\boldsymbol{\theta}, \{V_{\boldsymbol{\theta}}^{(j)}\}|\mathbf{y}) := \sum_{j=0}^{\infty} V_{\boldsymbol{\theta}}^{(j)}$$

  is finite almost surely, having finite expectation where

$$\mathbb{E}\left(\hat{\pi}(\boldsymbol{\theta}, \{V_{\boldsymbol{\theta}}^{(j)}\}\,|\mathbf{y})\right) = \pi(\boldsymbol{\theta}|\mathbf{y})$$

Infinite Series Expansion Construction

- For each $\boldsymbol{\theta}$ and $\boldsymbol{y}$, construct random variable $\{V_{\boldsymbol{\theta}}^{(j)},\ j \geq 0\}$ such that

$$\hat{\pi}(\boldsymbol{\theta}, \{V_{\boldsymbol{\theta}}^{(j)}\}|\mathbf{y}) := \sum_{j=0}^{\infty} V_{\boldsymbol{\theta}}^{(j)}$$

  is finite almost surely, having finite expectation where

$$\mathbb{E}\left(\hat{\pi}(\boldsymbol{\theta}, \{V_{\boldsymbol{\theta}}^{(j)}\}\,|\mathbf{y})\right) = \pi(\boldsymbol{\theta}|\mathbf{y})$$

- Introduce a random time $\tau_{\boldsymbol{\theta}}$, such that with $\boldsymbol{\xi} := (\tau_{\boldsymbol{\theta}}, \{V_{\boldsymbol{\theta}}^{(j)}, 0 \leq j \leq \tau_{\boldsymbol{\theta}}\})$ the estimate

$$\hat{\pi}(\theta, \boldsymbol{\xi}|\mathbf{y}) := \sum_{j=0}^{\tau_\theta} V_{\theta}^{(j)}$$

Infinite Series Expansion Construction

- For each $\boldsymbol{\theta}$ and $\boldsymbol{y}$, construct random variable $\{V_{\boldsymbol{\theta}}^{(j)},\ j \geq 0\}$ such that

  $$\hat{\pi}(\boldsymbol{\theta}, \{V_{\boldsymbol{\theta}}^{(j)}\}|\mathbf{y}) := \sum_{j=0}^{\infty} V_{\boldsymbol{\theta}}^{(j)}$$

  is finite almost surely, having finite expectation where

  $$\mathbb{E}\left(\hat{\pi}(\boldsymbol{\theta}, \{V_{\boldsymbol{\theta}}^{(j)}\}\,|\mathbf{y})\right) = \pi(\boldsymbol{\theta}|\mathbf{y})$$

- Introduce a random time $\tau_{\boldsymbol{\theta}}$, such that with $\boldsymbol{\xi} := (\tau_{\boldsymbol{\theta}}, \{V_{\boldsymbol{\theta}}^{(j)},\ 0 \leq j \leq \tau_{\boldsymbol{\theta}}\})$ the estimate

  $$\hat{\pi}(\theta, \boldsymbol{\xi}|\mathbf{y}) := \sum_{j=0}^{\tau_\theta} V_\theta^{(j)}$$

  satisfies

  $$\mathbb{E}\left(\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})|\{V_\theta^{(j)},\ j \geq 0\}\right) = \sum_{j=0}^{\infty} V_{\boldsymbol{\theta}}^{(j)}.$$

Targeting Absolute Measure

- Unbiased estimate $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ using series construction no general guarantee of positivity

Targeting Absolute Measure

- Unbiased estimate $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ using series construction no general guarantee of positivity
- Well studied problem in Solid State and QCD literature with conference devoted to *Sign Problem*

Targeting Absolute Measure

- Unbiased estimate $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ using series construction no general guarantee of positivity
- Well studied problem in Solid State and QCD literature with conference devoted to *Sign Problem*
- Own feeble attempts at resolving Sign problem unsuccessful ...... to date

Targeting Absolute Measure

- Unbiased estimate $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ using series construction no general guarantee of positivity
- Well studied problem in Solid State and QCD literature with conference devoted to *Sign Problem*
- Own feeble attempts at resolving Sign problem unsuccessful ...... to date
- Inspiration from QCD literature, exploit result in Lin, Lui, Sloan, (2000)

Targeting Absolute Measure

- Unbiased estimate $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ using series construction no general guarantee of positivity
- Well studied problem in Solid State and QCD literature with conference devoted to *Sign Problem*
- Own feeble attempts at resolving Sign problem unsuccessful ...... to date
- Inspiration from QCD literature, exploit result in Lin, Lui, Sloan, (2000)
- Despite sign problem

Targeting Absolute Measure

- Unbiased estimate $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ using series construction no general guarantee of positivity
- Well studied problem in Solid State and QCD literature with conference devoted to *Sign Problem*
- Own feeble attempts at resolving Sign problem unsuccessful ...... to date
- Inspiration from QCD literature, exploit result in Lin, Lui, Sloan, (2000)
- Despite sign problem
- Retain exactness of Monte Carlo estimates of expectations w.r.t. $\pi(\boldsymbol{\theta}|\mathbf{y})$
- Details in paper

Targeting Absolute Measure

- W.L.O.G, write
  $$\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})}\hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$$
  where $Z(\mathbf{y})$ is some intractable normalizing constant.

Targeting Absolute Measure

- W.L.O.G, write

  $$\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})}\hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$$

  where $Z(\mathbf{y})$ is some intractable normalizing constant.

- By unbiasedness of $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$, $Z(\mathbf{y}) = \int\int \hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})d\boldsymbol{\theta}$

Targeting Absolute Measure

- W.L.O.G, write
$$\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})}\hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$$
where $Z(\mathbf{y})$ is some intractable normalizing constant.
- By unbiasedness of $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$, $Z(\mathbf{y}) = \int\int \hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})d\boldsymbol{\theta}$
- Although measure $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ integrates to 1, it is not a probability measure because of the positivity issue.

Targeting Absolute Measure

- W.L.O.G, write
$$\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})}\hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$$
where $Z(\mathbf{y})$ is some intractable normalizing constant.
- By unbiasedness of $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$, $Z(\mathbf{y}) = \int\int \hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})d\boldsymbol{\theta}$
- Although measure $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ integrates to 1, it is not a probability measure because of the positivity issue.
- Write
$$\hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) = \sigma(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})|\hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})|$$

Targeting Absolute Measure

- W.L.O.G, write
$$\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) = \frac{1}{Z(\mathbf{y})}\hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$$
where $Z(\mathbf{y})$ is some intractable normalizing constant.
- By unbiasedness of $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$, $Z(\mathbf{y}) = \int\int \hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})d\boldsymbol{\theta}$
- Although measure $\hat{\pi}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})$ integrates to 1, it is not a probability measure because of the positivity issue.
- Write
$$\hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y}) = \sigma(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})|\hat{p}(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathbf{y})|$$
- Require to obtain expectation
$$\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

Targeting Absolute Measure

- We can write integral as

$$\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \quad = \quad \int\int h(\boldsymbol{\theta})\hat{\pi}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})d\boldsymbol{\theta}$$

Targeting Absolute Measure

- We can write integral as

$$\begin{aligned}\int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} &= \int\int h(\boldsymbol{\theta})\hat{\pi}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})d\boldsymbol{\theta} \\ &= \frac{\int\int h(\boldsymbol{\theta})\sigma(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y})}{\int\int \sigma(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y})}\end{aligned}$$

Targeting Absolute Measure

- We can write integral as

$$\begin{aligned} \int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} &= \int\int h(\boldsymbol{\theta})\hat{\pi}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})d\boldsymbol{\theta} \\ &= \frac{\int\int h(\boldsymbol{\theta})\sigma(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y})}{\int\int \sigma(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y})} \end{aligned}$$

- where $\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y})$ is the distribution

$$\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y}) := \frac{|\hat{p}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})|d\boldsymbol{\theta}P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})}{\int\int|\hat{p}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})|d\boldsymbol{\theta}P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})}.$$

Targeting Absolute Measure

- We can write integral as

$$\begin{aligned} \int h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} &= \int\int h(\boldsymbol{\theta})\hat{\pi}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})d\boldsymbol{\theta} \\ &= \frac{\int\int h(\boldsymbol{\theta})\sigma(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y})}{\int\int \sigma(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y})} \end{aligned}$$

- where $\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y})$ is the distribution

$$\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y}) := \frac{|\hat{p}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})|d\boldsymbol{\theta}P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})}{\int\int |\hat{p}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})|d\boldsymbol{\theta}P_{\boldsymbol{\theta}}(d\boldsymbol{\xi})}.$$

- Exact-approximate MH algorithm with target $\check{\pi}(d\boldsymbol{\theta},d\boldsymbol{\xi}|\mathbf{y})$ and proposal $Q(\boldsymbol{\theta},\boldsymbol{\xi};d\boldsymbol{\theta}',d\boldsymbol{\xi}') = q(\boldsymbol{\theta}'|\boldsymbol{\theta})d\boldsymbol{\theta}'P_{\boldsymbol{\theta}'}(d\boldsymbol{\xi}')$ has acceptance probability given by

$$\min\left\{1,\ \frac{|\hat{p}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})|}{|\hat{p}(\boldsymbol{\theta},\boldsymbol{\xi}|\mathbf{y})|} \times \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}.$$

Targeting Absolute Measure

- Estimate

$$I = \frac{\int h(x)\pi(x)dx}{\int \pi(x)dx},$$

  using

$$\hat{I}_n = \frac{\sum_{k=1}^{n} \sigma(X_k)h(X_k)}{\sum_{k=1}^{n} \sigma(X_k)}.$$

Targeting Absolute Measure

- Estimate
$$I = \frac{\int h(x)\pi(x)dx}{\int \pi(x)dx},$$
using
$$\hat{I}_n = \frac{\sum_{k=1}^{n} \sigma(X_k)h(X_k)}{\sum_{k=1}^{n} \sigma(X_k)}.$$
- If $\{X_n,\ n \geq 0\}$ is irreducible & aperiodic, $\hat{I}_n \to I$ a.s. as $n \to \infty$.

Targeting Absolute Measure

- Estimate
  $$I = \frac{\int h(x)\pi(x)dx}{\int \pi(x)dx},$$
  using
  $$\hat{I}_n = \frac{\sum_{k=1}^{n} \sigma(X_k)h(X_k)}{\sum_{k=1}^{n} \sigma(X_k)}.$$
- If $\{X_n,\ n \geq 0\}$ is irreducible & aperiodic, $\hat{I}_n \to I$ a.s. as $n \to \infty$.
- Approximation of the Monte Carlo variance of $\hat{I}_n$ is given by
  $$\frac{1}{n} \times \left\{ \frac{\sum_{k=1}^{n} h^2(X_k)\sigma(X_k)}{\sum_{k=1}^{n} \sigma(X_k)} - \left(\frac{\sum_{k=1}^{n} h(X_k)\sigma(X_k)}{\sum_{k=1}^{n} \sigma(X_k)} \right)^2 \right\} \times \frac{\hat{V}}{\left\{ \frac{1}{n} \sum_{k=1}^{n} \sigma(X_k) \right\}^2},$$
  where $\hat{V}$ is an estimate of the common autocorrelation sum.

Targeting Absolute Measure

- Estimate

  $$I = \frac{\int h(x)\pi(x)dx}{\int \pi(x)dx},$$

  using

  $$\hat{I}_n = \frac{\sum_{k=1}^n \sigma(X_k)h(X_k)}{\sum_{k=1}^n \sigma(X_k)}.$$

- If $\{X_n,\ n \geq 0\}$ is irreducible & aperiodic, $\hat{I}_n \to I$ a.s. as $n \to \infty$.
- Approximation of the Monte Carlo variance of $\hat{I}_n$ is given by

  $$\frac{1}{n} \times \left\{ \frac{\sum_{k=1}^n h^2(X_k)\sigma(X_k)}{\sum_{k=1}^n \sigma(X_k)} - \left(\frac{\sum_{k=1}^n h(X_k)\sigma(X_k)}{\sum_{k=1}^n \sigma(X_k)} \right)^2 \right\} \times \frac{\hat{V}}{\left\{\frac{1}{n}\sum_{k=1}^n \sigma(X_k)\right\}^2},$$

  where $\hat{V}$ is an estimate of the common autocorrelation sum.
- Quantity $\sum_{k=1}^n \sigma(X_k)$ indicates severity of sign problem, the smaller the harder it is to estimate I accurately.

Unbiased Estimators via Geometric Tilting

- The approximation $\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta})/\tilde{\mathcal{Z}}(\boldsymbol{\theta})$, where $\tilde{\mathcal{Z}}(\boldsymbol{\theta})$ is an estimate, approximation, an upper-bound, or a deterministic approximation

Unbiased Estimators via Geometric Tilting

- The approximation $\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta})/\tilde{\mathcal{Z}}(\boldsymbol{\theta})$, where $\tilde{\mathcal{Z}}(\boldsymbol{\theta})$ is an estimate, approximation, an upper-bound, or a deterministic approximation
- A multiplicative correction can take form of an infinite expansion such as

$$p(\mathbf{y}|\boldsymbol{\theta}) = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta}) \left[1 + \sum_{n=1}^{\infty} \kappa(\boldsymbol{\theta})^n \right]$$

Unbiased Estimators via Geometric Tilting

- The approximation $\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta})/\tilde{\mathcal{Z}}(\boldsymbol{\theta})$, where $\tilde{\mathcal{Z}}(\boldsymbol{\theta})$ is an estimate, approximation, an upper-bound, or a deterministic approximation
- A multiplicative correction can take form of an infinite expansion such as

$$p(\mathbf{y}|\boldsymbol{\theta}) = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta}) \left[1 + \sum_{n=1}^{\infty} \kappa(\boldsymbol{\theta})^n\right]$$

- Note that if $\kappa(\boldsymbol{\theta}) = 1 - c(\boldsymbol{\theta})\mathcal{Z}(\boldsymbol{\theta})/\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$

Unbiased Estimators via Geometric Tilting

- The approximation $\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta})/\tilde{\mathcal{Z}}(\boldsymbol{\theta})$, where $\tilde{\mathcal{Z}}(\boldsymbol{\theta})$ is an estimate, approximation, an upper-bound, or a deterministic approximation
- A multiplicative correction can take form of an infinite expansion such as

$$p(\mathbf{y}|\boldsymbol{\theta}) = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta}) \left[1 + \sum_{n=1}^{\infty} \kappa(\boldsymbol{\theta})^n\right]$$

- Note that if $\kappa(\boldsymbol{\theta}) = 1 - c(\boldsymbol{\theta})\mathcal{Z}(\boldsymbol{\theta})/\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$then for a choice of the constant $c(\boldsymbol{\theta})$ that ensures the region of convergence of a geometric series i.e. $|\kappa(\boldsymbol{\theta})| < 1$, by convergence of a geometric series it follows trivially that

$$p(\mathbf{y}|\boldsymbol{\theta}) = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta}) \left[1 + \sum_{n=1}^{\infty} \kappa(\boldsymbol{\theta})^n\right] = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times \frac{c(\boldsymbol{\theta})}{1 - \kappa(\boldsymbol{\theta})} = p(\mathbf{y}|\boldsymbol{\theta})$$

Unbiased Estimators via Geometric Tilting

- The approximation $\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y};\boldsymbol{\theta})/\tilde{\mathcal{Z}}(\boldsymbol{\theta})$, where $\tilde{\mathcal{Z}}(\boldsymbol{\theta})$ is an estimate, approximation, an upper-bound, or a deterministic approximation
- A multiplicative correction can take form of an infinite expansion such as

$$p(\mathbf{y}|\boldsymbol{\theta}) = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta}) \left[1 + \sum_{n=1}^{\infty} \kappa(\boldsymbol{\theta})^n \right]$$

- Note that if $\kappa(\boldsymbol{\theta}) = 1 - c(\boldsymbol{\theta})\mathcal{Z}(\boldsymbol{\theta})/\widetilde{\mathcal{Z}}(\boldsymbol{\theta})$then for a choice of the constant $c(\boldsymbol{\theta})$ that ensures the region of convergence of a geometric series i.e. $|\kappa(\boldsymbol{\theta})| < 1$, by convergence of a geometric series it follows trivially that

$$p(\mathbf{y}|\boldsymbol{\theta}) = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta}) \left[1 + \sum_{n=1}^{\infty} \kappa(\boldsymbol{\theta})^n \right] = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times \frac{c(\boldsymbol{\theta})}{1 - \kappa(\boldsymbol{\theta})} = p(\mathbf{y}|\boldsymbol{\theta})$$

- An infinite independent number of unbiased estimates of $\mathcal{Z}(\boldsymbol{\theta})$ each denoted as $\hat{\mathcal{Z}}_i(\boldsymbol{\theta})$ yields an unbiased estimate of the target density

$$\hat{p}(\mathbf{y}|\boldsymbol{\theta}) = \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta}) \left[1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \left(1 - c(\boldsymbol{\theta}) \frac{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})} \right) \right]$$

Unbiased Estimators via Geometric Tilting

- Notice that the series is finite almost surely and has finite expectation if

$$E\left(\left|1 - c(\boldsymbol{\theta})\frac{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}\right|\right) < 1$$

Unbiased Estimators via Geometric Tilting

- Notice that the series is finite almost surely and has finite expectation if

$$E\left(\left|1 - c(\boldsymbol{\theta})\frac{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}\right|\right) < 1$$

- As $E(|X|) \leq E^{1/2}(|X|^2)$, sufficient that $c(\theta) < 2\widetilde{\mathcal{Z}}(\boldsymbol{\theta})\hat{\mathcal{Z}}(\boldsymbol{\theta})/E\left(\hat{\mathcal{Z}}_1^2(\boldsymbol{\theta})\right)$

Unbiased Estimators via Geometric Tilting

- Notice that the series is finite almost surely and has finite expectation if

$$E\left(\left|1 - c(\boldsymbol{\theta})\frac{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}\right|\right) < 1$$

- As $E(|X|) \leq E^{1/2}(|X|^2)$, sufficient that $c(\theta) < 2\widetilde{\mathcal{Z}}(\boldsymbol{\theta})\hat{\mathcal{Z}}(\boldsymbol{\theta})/E\left(\hat{\mathcal{Z}}_1^2(\boldsymbol{\theta})\right)$
- Under this assumption expectation of $\hat{p}(\mathbf{y}|\boldsymbol{\theta})$ can be computed as

$$\begin{aligned} E\left\{\hat{p}(\mathbf{y}|\boldsymbol{\theta})\right\} &= \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\left(1 - c(\boldsymbol{\theta})\frac{E\left\{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})\right\}}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}\right)\right] \\ &= \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty}\kappa(\boldsymbol{\theta})^n\right] = p(\mathbf{y}|\boldsymbol{\theta}) \end{aligned}$$

Unbiased Estimators via Geometric Tilting

- Notice that the series is finite almost surely and has finite expectation if

$$E\left(\left|1 - c(\boldsymbol{\theta})\frac{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}\right|\right) < 1$$

- As $E(|X|) \leq E^{1/2}(|X|^2)$, sufficient that $c(\theta) < 2\widetilde{\mathcal{Z}}(\boldsymbol{\theta})\hat{\mathcal{Z}}(\boldsymbol{\theta})/E\left(\hat{\mathcal{Z}}_1^2(\boldsymbol{\theta})\right)$
- Under this assumption expectation of $\hat{p}(\mathbf{y}|\boldsymbol{\theta})$ can be computed as

$$\begin{aligned} E\left\{\hat{p}(\mathbf{y}|\boldsymbol{\theta})\right\} &= \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty}\prod_{i=1}^{n}\left(1 - c(\boldsymbol{\theta})\frac{E\left\{\hat{\mathcal{Z}}_i(\boldsymbol{\theta})\right\}}{\widetilde{\mathcal{Z}}(\boldsymbol{\theta})}\right)\right] \\ &= \tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \times c(\boldsymbol{\theta})\left[1 + \sum_{n=1}^{\infty}\kappa(\boldsymbol{\theta})^n\right] = p(\mathbf{y}|\boldsymbol{\theta}) \end{aligned}$$

- Therefore essential property, $E\left\{\hat{p}(\mathbf{y}|\boldsymbol{\theta})\right\} = p(\mathbf{y}|\boldsymbol{\theta})$, required of a plugin estimator for exact-approximate MCMC is satisfied

Unbiased Estimators via Exponential Tilting

- Introduction auxiliary variable $\nu \sim \text{Expon}(\mathcal{Z}(\boldsymbol{\theta}))$ defines joint distribution

$$\begin{aligned}\pi(\boldsymbol{\theta}, \nu|\mathbf{y}) &= \mathcal{Z}(\boldsymbol{\theta}) \exp(-\nu\mathcal{Z}(\boldsymbol{\theta})) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})} \\ &= \exp\left(-\nu\mathcal{Z}(\boldsymbol{\theta})\right) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})}\end{aligned}$$

Unbiased Estimators via Exponential Tilting

- Introduction auxiliary variable $\nu \sim \text{Expon}(\mathcal{Z}(\boldsymbol{\theta}))$ defines joint distribution

$$\begin{aligned}\pi(\boldsymbol{\theta}, \nu | \mathbf{y}) &= \mathcal{Z}(\boldsymbol{\theta}) \exp(-\nu \mathcal{Z}(\boldsymbol{\theta})) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})} \\ &= \exp\left(-\nu \mathcal{Z}(\boldsymbol{\theta})\right) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})}\end{aligned}$$

- Exact-approximate scheme constructed by estimating $\exp\left(-\nu \mathcal{Z}(\boldsymbol{\theta})\right)$

Unbiased Estimators via Exponential Tilting

- Introduction auxiliary variable $\nu \sim \text{Expon}(\mathcal{Z}(\boldsymbol{\theta}))$ defines joint distribution

$$\begin{aligned}\pi(\boldsymbol{\theta}, \nu | \mathbf{y}) &= \mathcal{Z}(\boldsymbol{\theta}) \exp(-\nu \mathcal{Z}(\boldsymbol{\theta})) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})} \\ &= \exp\left(-\nu \mathcal{Z}(\boldsymbol{\theta})\right) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})}\end{aligned}$$

- Exact-approximate scheme constructed by estimating $\exp\left(-\nu \mathcal{Z}(\boldsymbol{\theta})\right)$
- The MacLaurin series expansion is

$$\exp(-\nu \mathcal{Z}(\boldsymbol{\theta})) = 1 + \sum_{n=1}^{\infty} \frac{(-\nu)^n}{n!} \mathcal{Z}(\boldsymbol{\theta})^n$$

Unbiased Estimators via Exponential Tilting

- Introduction auxiliary variable $\nu \sim \text{Expon}(\mathcal{Z}(\boldsymbol{\theta}))$ defines joint distribution

$$\begin{aligned}\pi(\boldsymbol{\theta}, \nu|\mathbf{y}) &= \mathcal{Z}(\boldsymbol{\theta}) \exp(-\nu \mathcal{Z}(\boldsymbol{\theta})) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})} \\ &= \exp\left(-\nu \mathcal{Z}(\boldsymbol{\theta})\right) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})}\end{aligned}$$

- Exact-approximate scheme constructed by estimating $\exp\left(-\nu \mathcal{Z}(\boldsymbol{\theta})\right)$
- The MacLaurin series expansion is

$$\exp(-\nu \mathcal{Z}(\boldsymbol{\theta})) = 1 + \sum_{n=1}^{\infty} \frac{(-\nu)^n}{n!} \mathcal{Z}(\boldsymbol{\theta})^n$$

- Suggesting an unbiased estimator of the form

$$\exp(\widehat{-\nu \mathcal{Z}}(\boldsymbol{\theta})) = 1 + \sum_{n=1}^{\infty} \frac{(-\nu)^n}{n!} \prod_{i=1}^{n} \hat{\mathcal{Z}}_i(\boldsymbol{\theta}),$$

Unbiased Estimators via Exponential Tilting

- Introduction auxiliary variable $\nu \sim \text{Expon}(\mathcal{Z}(\boldsymbol{\theta}))$ defines joint distribution

$$\begin{aligned}\pi(\boldsymbol{\theta}, \nu|\mathbf{y}) &= \mathcal{Z}(\boldsymbol{\theta}) \exp(-\nu \mathcal{Z}(\boldsymbol{\theta})) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\boldsymbol{\theta})} \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})} \\ &= \exp\left(-\nu \mathcal{Z}(\boldsymbol{\theta})\right) \times f(\mathbf{y}; \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \times \frac{1}{\mathcal{Z}(\mathbf{y})}\end{aligned}$$

- Exact-approximate scheme constructed by estimating $\exp\left(-\nu \mathcal{Z}(\boldsymbol{\theta})\right)$
- The MacLaurin series expansion is

$$\exp(-\nu \mathcal{Z}(\boldsymbol{\theta})) = 1 + \sum_{n=1}^{\infty} \frac{(-\nu)^n}{n!} \mathcal{Z}(\boldsymbol{\theta})^n$$

- Suggesting an unbiased estimator of the form

$$\exp(\widehat{-\nu \mathcal{Z}}(\boldsymbol{\theta})) = 1 + \sum_{n=1}^{\infty} \frac{(-\nu)^n}{n!} \prod_{i=1}^{n} \hat{\mathcal{Z}}_i(\boldsymbol{\theta}),$$

- $n!$ grows faster than exponential, series finite a.s. with finite expectation

Unbiased Estimators via Exponential Tilting

- An approximate $\tilde{\mathcal{Z}}(\boldsymbol{\theta})$ can be exploited such that

Unbiased Estimators via Exponential Tilting

- An approximate $\tilde{\mathcal{Z}}(\boldsymbol{\theta})$ can be exploited such that

$$\begin{aligned}
\exp(-\nu\mathcal{Z}(\boldsymbol{\theta})) &= \exp(-\nu\tilde{\mathcal{Z}}(\boldsymbol{\theta})) \times \exp\left(\nu(\tilde{\mathcal{Z}}(\boldsymbol{\theta}) - \mathcal{Z}(\boldsymbol{\theta}))\right) \\
&= \exp(-\nu\tilde{\mathcal{Z}}(\boldsymbol{\theta})) \times \left(1 + \sum_{n=1}^{\infty} \frac{\nu^n}{n!} \left(\tilde{\mathcal{Z}}(\boldsymbol{\theta}) - \mathcal{Z}(\boldsymbol{\theta})\right)^n\right)
\end{aligned}$$

Unbiased Estimators via Exponential Tilting

- An approximate $\tilde{\mathcal{Z}}(\boldsymbol{\theta})$ can be exploited such that

$$\begin{aligned}\exp(-\nu \mathcal{Z}(\boldsymbol{\theta})) &= \exp(-\nu \tilde{\mathcal{Z}}(\boldsymbol{\theta})) \times \exp\left(\nu(\tilde{\mathcal{Z}}(\boldsymbol{\theta}) - \mathcal{Z}(\boldsymbol{\theta}))\right) \\ &= \exp(-\nu \tilde{\mathcal{Z}}(\boldsymbol{\theta})) \times \left(1 + \sum_{n=1}^{\infty} \frac{\nu^n}{n!} \left(\tilde{\mathcal{Z}}(\boldsymbol{\theta}) - \mathcal{Z}(\boldsymbol{\theta})\right)^n\right)\end{aligned}$$

- Yields a tilted estimator of the form

$$\exp\widehat{(-\nu \mathcal{Z}(\boldsymbol{\theta}))} = \exp(-\nu \tilde{\mathcal{Z}}(\boldsymbol{\theta})) \times \left(1 + \sum_{n=1}^{\infty} \frac{\nu^n}{n!} \prod_{i=1}^{n} \left(\tilde{\mathcal{Z}}(\boldsymbol{\theta}) - \hat{\mathcal{Z}}_i(\boldsymbol{\theta})\right)\right)$$

Unbiased Estimators via Exponential Tilting

- An approximate $\tilde{\mathcal{Z}}(\boldsymbol{\theta})$ can be exploited such that

$$\begin{aligned}\exp(-\nu\mathcal{Z}(\boldsymbol{\theta})) &= \exp(-\nu\tilde{\mathcal{Z}}(\boldsymbol{\theta})) \times \exp\left(\nu(\tilde{\mathcal{Z}}(\boldsymbol{\theta}) - \mathcal{Z}(\boldsymbol{\theta}))\right) \\ &= \exp(-\nu\tilde{\mathcal{Z}}(\boldsymbol{\theta})) \times \left(1 + \sum_{n=1}^{\infty} \frac{\nu^n}{n!} \left(\tilde{\mathcal{Z}}(\boldsymbol{\theta}) - \mathcal{Z}(\boldsymbol{\theta})\right)^n\right)\end{aligned}$$

- Yields a tilted estimator of the form

$$\exp(\widehat{-\nu\mathcal{Z}}(\boldsymbol{\theta})) = \exp(-\nu\tilde{\mathcal{Z}}(\boldsymbol{\theta})) \times \left(1 + \sum_{n=1}^{\infty} \frac{\nu^n}{n!} \prod_{i=1}^{n} \left(\tilde{\mathcal{Z}}(\boldsymbol{\theta}) - \hat{\mathcal{Z}}_i(\boldsymbol{\theta})\right)\right)$$

- Approximation $\exp(-\nu\tilde{\mathcal{Z}}(\boldsymbol{\theta}))$ corrected by exponential tilt

Russian Roulette

Russian Roulette

Russian Roulette

- Require unbiased truncation of infinite sum $\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=0}^{\infty} \phi_i(\boldsymbol{\theta})$

Russian Roulette

- Require unbiased truncation of infinite sum $\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=0}^{\infty} \phi_i(\boldsymbol{\theta})$
- Poisson truncation and Generalised Poisson truncation (infinite variance for Geometric series)

Russian Roulette

- Require unbiased truncation of infinite sum $\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=0}^{\infty} \phi_i(\boldsymbol{\theta})$
- Poisson truncation and Generalised Poisson truncation (infinite variance for Geometric series)
- Russian Roulette employed extensively in simulation of neutron scattering and computer graphics

Russian Roulette

- Require unbiased truncation of infinite sum $\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=0}^{\infty} \phi_i(\boldsymbol{\theta})$
- Poisson truncation and Generalised Poisson truncation (infinite variance for Geometric series)
- Russian Roulette employed extensively in simulation of neutron scattering and computer graphics
- Assign probabilities $\{q_j,\ j \geq 1\}$ $q_j \in (0,1]$ and $\mathcal{U}(0,1)$ i.i.d. r.v's $\{U_j,\ j \geq 1\}$

Russian Roulette

- Require unbiased truncation of infinite sum $\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=0}^{\infty} \phi_i(\boldsymbol{\theta})$
- Poisson truncation and Generalised Poisson truncation (infinite variance for Geometric series)
- Russian Roulette employed extensively in simulation of neutron scattering and computer graphics
- Assign probabilities $\{q_j,\ j \geq 1\}$ $q_j \in (0, 1]$ and $\mathcal{U}(0, 1)$ i.i.d. r.v's $\{U_j,\ j \geq 1\}$
- Find the first time $k \geq 1$ such that $U_k \geq q_k$

Russian Roulette

- Require unbiased truncation of infinite sum $\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=0}^{\infty} \phi_i(\boldsymbol{\theta})$
- Poisson truncation and Generalised Poisson truncation (infinite variance for Geometric series)
- Russian Roulette employed extensively in simulation of neutron scattering and computer graphics
- Assign probabilities $\{q_j,\ j \geq 1\}$ $q_j \in (0, 1]$ and $\mathcal{U}(0, 1)$ i.i.d. r.v's $\{U_j,\ j \geq 1\}$
- Find the first time $k \geq 1$ such that $U_k \geq q_k$
- Russian Roulette estimate of $\mathcal{S}(\boldsymbol{\theta})$ is

$$\hat{\mathcal{S}}(\boldsymbol{\theta}) = \sum_{j=0}^{k} \frac{\phi_j(\boldsymbol{\theta})}{\prod_{i=1}^{j-1} q_i},$$

Russian Roulette

- Require unbiased truncation of infinite sum $\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=0}^{\infty} \phi_i(\boldsymbol{\theta})$
- Poisson truncation and Generalised Poisson truncation (infinite variance for Geometric series)
- Russian Roulette employed extensively in simulation of neutron scattering and computer graphics
- Assign probabilities $\{q_j, \ j \geq 1\}$ $q_j \in (0, 1]$ and $\mathcal{U}(0, 1)$ i.i.d. r.v's $\{U_j, \ j \geq 1\}$
- Find the first time $k \geq 1$ such that $U_k \geq q_k$
- Russian Roulette estimate of $\mathcal{S}(\boldsymbol{\theta})$ is

$$\hat{\mathcal{S}}(\boldsymbol{\theta}) = \sum_{j=0}^{k} \frac{\phi_j(\boldsymbol{\theta})}{\prod_{i=1}^{j-1} q_i},$$

- If $\lim_{n\to\infty} \prod_{j=1}^{n} q_j = 0$, Russian Roulette terminates with probability one

Russian Roulette

- Require unbiased truncation of infinite sum $\mathcal{S}(\boldsymbol{\theta}) = \sum_{i=0}^{\infty} \phi_i(\boldsymbol{\theta})$
- Poisson truncation and Generalised Poisson truncation (infinite variance for Geometric series)
- Russian Roulette employed extensively in simulation of neutron scattering and computer graphics
- Assign probabilities $\{q_j,\ j \geq 1\}$ $q_j \in (0,1]$ and $\mathcal{U}(0,1)$ i.i.d. r.v's $\{U_j,\ j \geq 1\}$
- Find the first time $k \geq 1$ such that $U_k \geq q_k$
- Russian Roulette estimate of $\mathcal{S}(\boldsymbol{\theta})$ is

$$\hat{\mathcal{S}}(\boldsymbol{\theta}) = \sum_{j=0}^{k} \frac{\phi_j(\boldsymbol{\theta})}{\prod_{i=1}^{j-1} q_i},$$

- If $\lim_{n\to\infty} \prod_{j=1}^{n} q_j = 0$, Russian Roulette terminates with probability one
- Note $E\{\hat{\mathcal{S}}(\boldsymbol{\theta})\} = \mathcal{S}(\boldsymbol{\theta})$, variance finite under certain known conditions

Summary..... so far

- Use Geometric or Exponential tilted correction of approximate likelihood

Summary..... so far

- Use Geometric or Exponential tilted correction of approximate likelihood
- Randomly and unbiasedly truncate tilt using Russian Roulette

Summary..... so far

- Use Geometric or Exponential tilted correction of approximate likelihood
- Randomly and unbiasedly truncate tilt using Russian Roulette
- Plug estimate into MCMC transition kernel targeting absolute measure

Summary..... so far

- Use Geometric or Exponential tilted correction of approximate likelihood
- Randomly and unbiasedly truncate tilt using Russian Roulette
- Plug estimate into MCMC transition kernel targeting absolute measure
- Obtain Monte Carlo estimates using state dependent sign correction

Ising Spin Model

- Consider Ising model of spins $x_i \in \{-1, +1\}$ of the form

$$p(\mathbf{x}|\theta_1, \theta_2) = \frac{1}{\mathcal{Z}(\theta_1, \theta_2)} \exp\left(\theta_1 \sum_i x_i + \theta_2 \sum_{i \sim j} x_i x_j\right)$$

Ising Spin Model

- Consider Ising model of spins $x_i \in \{-1, +1\}$ of the form

$$p(\mathbf{x}|\theta_1, \theta_2) = \frac{1}{\mathcal{Z}(\theta_1, \theta_2)} \exp\left(\theta_1 \sum_i x_i + \theta_2 \sum_{i \sim j} x_i x_j\right)$$

- Partition function intractable 10×10 torus $\sim 1.26 \times 10^{30}$ states

Ising Spin Model

- Consider Ising model of spins $x_i \in \{-1, +1\}$ of the form

$$p(\mathbf{x}|\theta_1, \theta_2) = \frac{1}{\mathcal{Z}(\theta_1, \theta_2)} \exp\left(\theta_1 \sum_i x_i + \theta_2 \sum_{i \sim j} x_i x_j\right)$$

- Partition function intractable 10×10 torus $\sim 1.26 \times 10^{30}$ states
- 10×10 torus spin state simulated with $\theta_1 = 0$ and $\theta_2 = 0.2$

Ising Spin Model

- Consider Ising model of spins $x_i \in \{-1, +1\}$ of the form

$$p(\mathbf{x}|\theta_1, \theta_2) = \frac{1}{\mathcal{Z}(\theta_1, \theta_2)} \exp\left(\theta_1 \sum_i x_i + \theta_2 \sum_{i \sim j} x_i x_j\right)$$

- Partition function intractable 10×10 torus $\sim 1.26 \times 10^{30}$ states
- 10×10 torus spin state simulated with $\theta_1 = 0$ and $\theta_2 = 0.2$
- SMC Sampling (AIS) employed for $\hat{\mathcal{Z}}(\theta_1, \theta_2)$, 1.5K samples in IS

Ising Spin Model

- Consider Ising model of spins $x_i \in \{-1, +1\}$ of the form

$$p(\mathbf{x}|\theta_1, \theta_2) = \frac{1}{\mathcal{Z}(\theta_1, \theta_2)} \exp\left(\theta_1 \sum_i x_i + \theta_2 \sum_{i \sim j} x_i x_j\right)$$

- Partition function intractable 10×10 torus $\sim 1.26 \times 10^{30}$ states
- 10×10 torus spin state simulated with $\theta_1 = 0$ and $\theta_2 = 0.2$
- SMC Sampling (AIS) employed for $\hat{\mathcal{Z}}(\theta_1, \theta_2)$, 1.5K samples in IS
- Russian Roulette parameters $c = 0.2$, $r = 0.8$, Uniform prior on θ_2

Ising Spin Model

- Consider Ising model of spins $x_i \in \{-1, +1\}$ of the form

$$p(\mathbf{x}|\theta_1, \theta_2) = \frac{1}{\mathcal{Z}(\theta_1, \theta_2)} \exp\left(\theta_1 \sum_i x_i + \theta_2 \sum_{i \sim j} x_i x_j\right)$$

- Partition function intractable 10×10 torus $\sim 1.26 \times 10^{30}$ states
- 10×10 torus spin state simulated with $\theta_1 = 0$ and $\theta_2 = 0.2$
- SMC Sampling (AIS) employed for $\hat{\mathcal{Z}}(\theta_1, \theta_2)$, 1.5K samples in IS
- Russian Roulette parameters $c = 0.2$, $r = 0.8$, Uniform prior on θ_2
- Acceptance rate of chain tuned to 45%

Ising Spin Model

- Consider Ising model of spins $x_i \in \{-1, +1\}$ of the form

$$p(\mathbf{x}|\theta_1, \theta_2) = \frac{1}{\mathcal{Z}(\theta_1, \theta_2)} \exp\left(\theta_1 \sum_i x_i + \theta_2 \sum_{i \sim j} x_i x_j\right)$$

- Partition function intractable 10×10 torus $\sim 1.26 \times 10^{30}$ states
- 10×10 torus spin state simulated with $\theta_1 = 0$ and $\theta_2 = 0.2$
- SMC Sampling (AIS) employed for $\hat{\mathcal{Z}}(\theta_1, \theta_2)$, 1.5K samples in IS
- Russian Roulette parameters $c = 0.2$, $r = 0.8$, Uniform prior on θ_2
- Acceptance rate of chain tuned to 45%
- 20K samples, ESS 1.6K, 0.2% sign violation rate (39)

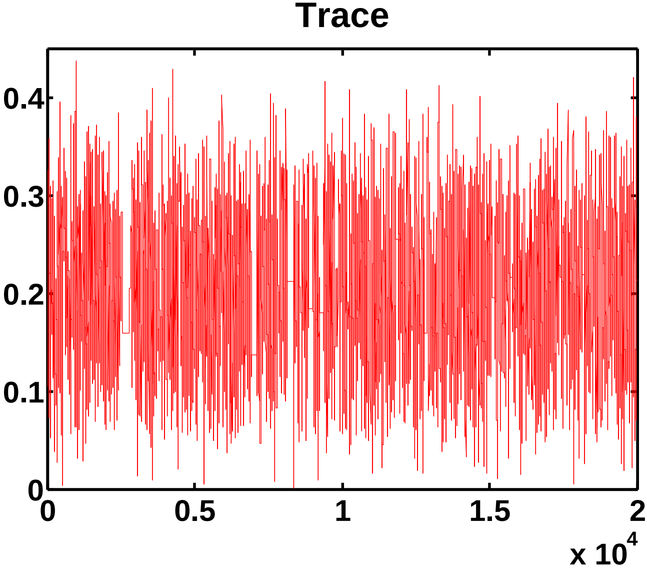
Ising Spin Model

- Consider Ising model of spins $x_i \in \{-1, +1\}$ of the form

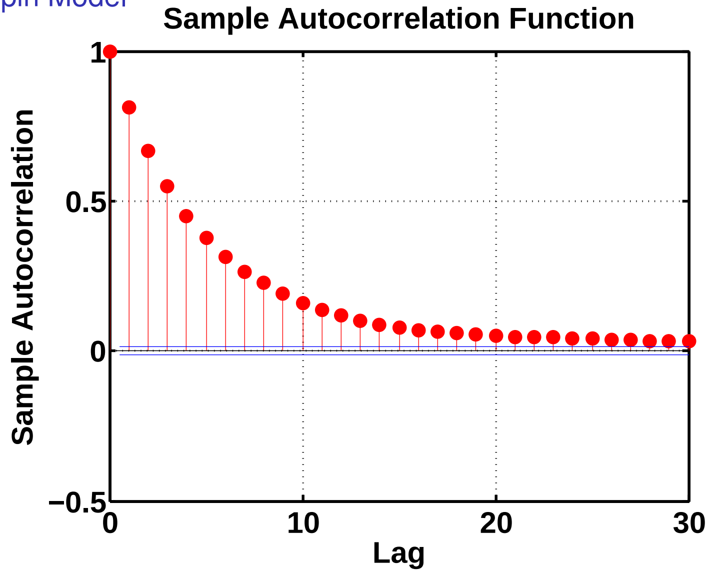
$$p(\mathbf{x}|\theta_1, \theta_2) = \frac{1}{\mathcal{Z}(\theta_1, \theta_2)} \exp\left(\theta_1 \sum_i x_i + \theta_2 \sum_{i \sim j} x_i x_j\right)$$

- Partition function intractable 10×10 torus $\sim 1.26 \times 10^{30}$ states
- 10×10 torus spin state simulated with $\theta_1 = 0$ and $\theta_2 = 0.2$
- SMC Sampling (AIS) employed for $\hat{\mathcal{Z}}(\theta_1, \theta_2)$, 1.5K samples in IS
- Russian Roulette parameters $c = 0.2$, $r = 0.8$, Uniform prior on θ_2
- Acceptance rate of chain tuned to 45%
- 20K samples, ESS 1.6K, 0.2% sign violation rate (39)
- Posterior mean 0.2028 ± 0.0646, uncorrected 0.2032 ± 0.0649

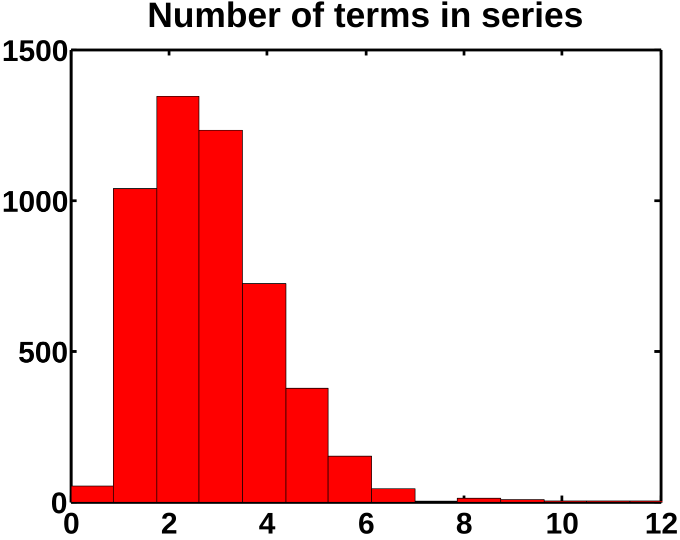
Ising Spin Model



Ising Spin Model



Ising Spin Model



Fisher-Bingham Distribution

- Embedded normal on manifold S_d with $p(\mathbf{y}|\mathbf{A}) \propto \exp\{\mathbf{y}'\mathbf{A}\mathbf{y}\}$

Fisher-Bingham Distribution

- Embedded normal on manifold S_d with $p(\mathbf{y}|\mathbf{A}) \propto \exp\{\mathbf{y}'\mathbf{A}\mathbf{y}\}$

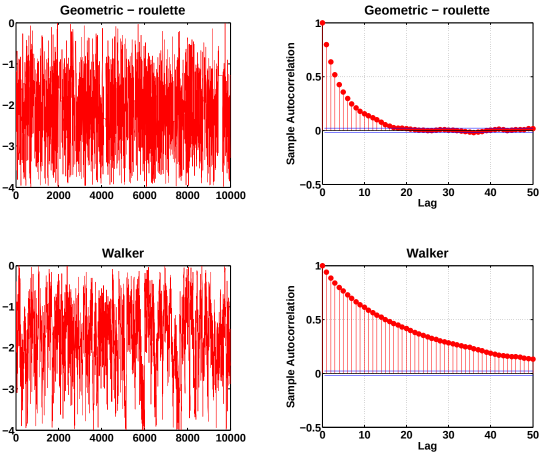
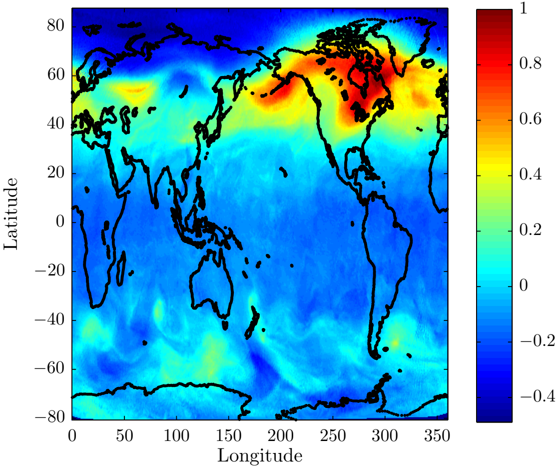


Figure: Sample traces and autocorrelation plots for the geometric tilting with roulette truncation ((a) and (b)) and Walker's method ((c) and (d))).

Large Scale GMRF Ozone Column Model



Large Scale GMRF Ozone Column Model

- Previously analysed in Cressie, (2008) comprised of 173,405 ozone measurements

Large Scale GMRF Ozone Column Model

- Previously analysed in Cressie, (2008) comprised of 173,405 ozone measurements
- Data and spatial extent has precluded full Bayesian analysis to date

Large Scale GMRF Ozone Column Model

- Previously analysed in Cressie, (2008) comprised of 173,405 ozone measurements
- Data and spatial extent has precluded full Bayesian analysis to date
- Matern covariance function triangulated over 196,002 vertices on sphere

Large Scale GMRF Ozone Column Model

- Previously analysed in Cressie, (2008) comprised of 173,405 ozone measurements
- Data and spatial extent has precluded full Bayesian analysis to date
- Matern covariance function triangulated over 196,002 vertices on sphere

$$p(\mathbf{x}|\theta) = \frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2}\left[\log|\mathbf{C}_\theta| + \mathbf{x}^{\mathsf{T}}\mathbf{C}_\theta^{-1}\mathbf{x}\right]\right\}$$

Large Scale GMRF Ozone Column Model

- Previously analysed in Cressie, (2008) comprised of 173,405 ozone measurements
- Data and spatial extent has precluded full Bayesian analysis to date
- Matern covariance function triangulated over 196,002 vertices on sphere

$$\begin{aligned} p(\mathbf{x}|\theta) &= \frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2}\left[\log|\mathbf{C}_\theta| + \mathbf{x}^{\mathrm{T}}\mathbf{C}_\theta^{-1}\mathbf{x}\right]\right\} \\ &= \frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left\{\frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n}\mathrm{E}\{\mathbf{z}^{\mathrm{T}}(\mathbf{I}-\mathbf{C}_\theta)^n\mathbf{z}\} - \frac{1}{2}\sum_{m=0}^{\infty}\mathbf{x}^{\mathrm{T}}(\mathbf{I}-\mathbf{C}_\theta)^m\mathbf{x}\right\} \end{aligned}$$

Large Scale GMRF Ozone Column Model

- Previously analysed in Cressie, (2008) comprised of 173,405 ozone measurements
- Data and spatial extent has precluded full Bayesian analysis to date
- Matern covariance function triangulated over 196,002 vertices on sphere

$$\begin{aligned} p(\mathbf{x}|\theta) &= \frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left\{ -\frac{1}{2} \left[\log|\mathbf{C}_\theta| + \mathbf{x}^{\mathrm{T}} \mathbf{C}_\theta^{-1} \mathbf{x} \right] \right\} \\ &= \frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left\{ \frac{1}{2} \sum_{n=1}^{\infty} \frac{1}{n} \mathrm{E}\{\mathbf{z}^{\mathrm{T}} (\mathbf{I} - \mathbf{C}_\theta)^n \mathbf{z}\} - \frac{1}{2} \sum_{m=0}^{\infty} \mathbf{x}^{\mathrm{T}} (\mathbf{I} - \mathbf{C}_\theta)^m \mathbf{x} \right\} \end{aligned}$$

$$\begin{aligned} p(\mathbf{x} \mid \theta) &= \frac{1}{(2\pi)^{N/2}} \exp\left\{ \frac{1}{2} \log|\mathbf{Q}_\theta| - \frac{1}{2} \mathbf{x}^T \mathbf{Q}_\theta \mathbf{x} \right\} \\ &= \frac{1}{(2\pi)^{N/2}} \exp\left\{ \frac{1}{2} \mathbb{E}_{\mathbf{z}}\{\mathbf{z}^{\mathbf{T}} \log(\mathbf{Q}_\theta) \mathbf{z}\} - \frac{\mathbf{1}}{\mathbf{2}} \mathbf{x}^{\mathbf{T}} \mathbf{Q}_\theta \mathbf{x} \right\} \end{aligned}$$

Large Scale GMRF Ozone Column Model

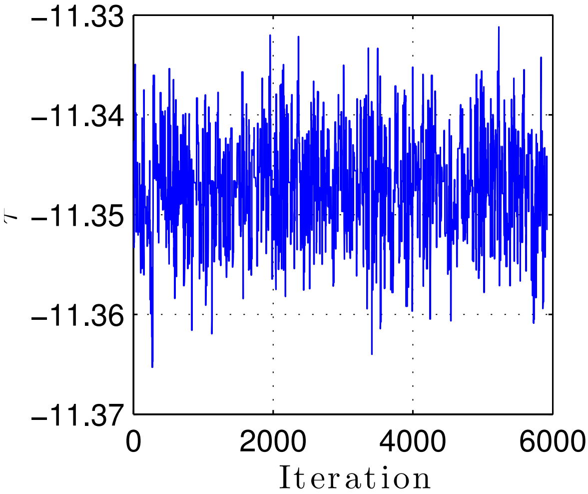
- Previously analysed in Cressie, (2008) comprised of 173,405 ozone measurements
- Data and spatial extent has precluded full Bayesian analysis to date
- Matern covariance function triangulated over 196,002 vertices on sphere

$$\begin{aligned} p(\mathbf{x}|\theta) &= \frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2}\left[\log|\mathbf{C}_\theta| + \mathbf{x}^{\mathrm{T}}\mathbf{C}_\theta^{-1}\mathbf{x}\right]\right\} \\ &= \frac{1}{(2\pi)^{\frac{N}{2}}} \exp\left\{\frac{1}{2}\sum_{n=1}^{\infty}\frac{1}{n}\mathrm{E}\{\mathbf{z}^{\mathrm{T}}(\mathbf{I}-\mathbf{C}_\theta)^n\mathbf{z}\} - \frac{1}{2}\sum_{m=0}^{\infty}\mathbf{x}^{\mathrm{T}}(\mathbf{I}-\mathbf{C}_\theta)^m\mathbf{x}\right\} \end{aligned}$$

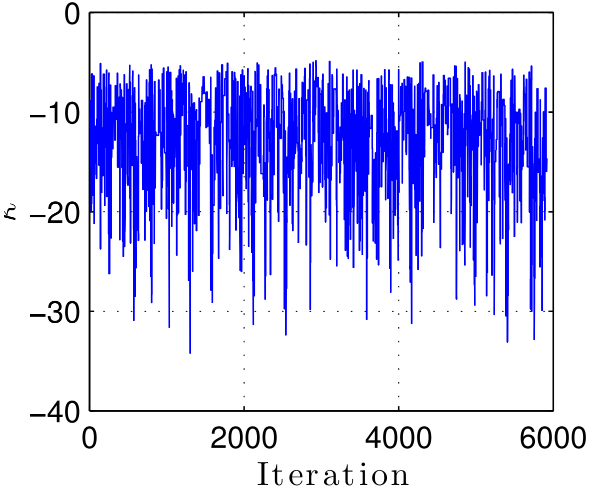
$$\begin{aligned} p(\mathbf{x} \mid \theta) &= \frac{1}{(2\pi)^{N/2}} \exp\left\{\frac{1}{2}\log|\mathbf{Q}_\theta| - \frac{1}{2}\mathbf{x}^T\mathbf{Q}_\theta\mathbf{x}\right\} \\ &= \frac{1}{(2\pi)^{N/2}} \exp\left\{\frac{1}{2}\mathbb{E}_{\mathbf{z}}\{\mathbf{z}^{\mathbf{T}}\log(\mathbf{Q}_\theta)\mathbf{z}\} - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{x}^{\mathbf{T}}\mathbf{Q}_\theta\mathbf{x}\right\} \end{aligned}$$

- Employ trace log construction described in Aune *et al* 2012, Statistics and Computing.

Large Scale GMRF Ozone Column Model



Large Scale GMRF Ozone Column Model



Large Scale GMRF Some Details

- Likelihood estimate is of form $\sum_{i=0}^{\infty} \alpha_i$, where α_i built from *independent* estimates of the *same* quantity

Large Scale GMRF Some Details

- Likelihood estimate is of form $\sum_{i=0}^{\infty} \alpha_i$, where α_i built from *independent* estimates of the *same* quantity
- Truncate at some *n*, so need *n* estimates $\hat{\alpha}$

Large Scale GMRF Some Details

- Likelihood estimate is of form $\sum_{i=0}^{\infty} \alpha_i$, where α_i built from *independent* estimates of the *same* quantity
- Truncate at some *n*, so need *n* estimates $\hat{\alpha}$
- Log-determinant estimate is of the form $\mathbb{E}_{\boldsymbol{z}}(\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z})$. Monte-Carlo estimate relies on *independent* estimates of $\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z}$

Large Scale GMRF Some Details

- Likelihood estimate is of form $\sum_{i=0}^{\infty} \alpha_i$, where α_i built from *independent* estimates of the *same* quantity
- Truncate at some n, so need n estimates $\hat{\alpha}$
- Log-determinant estimate is of the form $\mathbb{E}_{\boldsymbol{z}}(\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z})$. Monte-Carlo estimate relies on *independent* estimates of $\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z}$
- Matrix logarithm in form $\log(\boldsymbol{Q})\boldsymbol{z} \approx \sum_{i=1}^{N} \boldsymbol{A}_i^{-1}\boldsymbol{x}$. Solve N *independent* linear systems

Large Scale GMRF Some Details

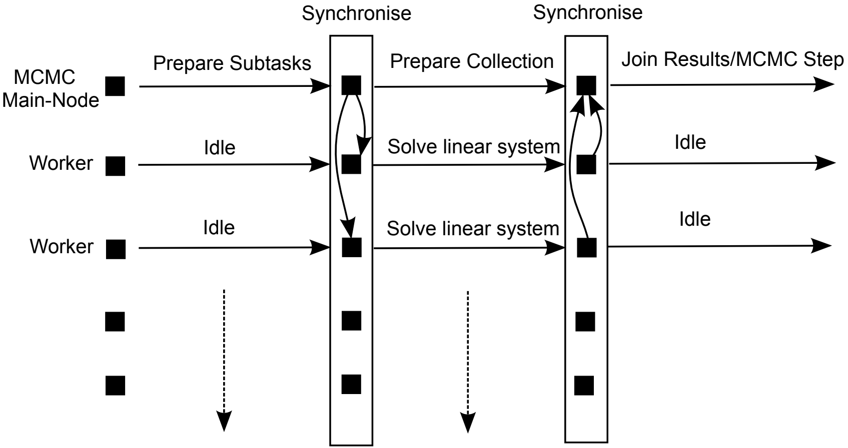
- Likelihood estimate is of form $\sum_{i=0}^{\infty} \alpha_i$, where α_i built from *independent* estimates of the *same* quantity
- Truncate at some n, so need n estimates $\hat{\alpha}$
- Log-determinant estimate is of the form $\mathbb{E}_{\boldsymbol{z}}(\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z})$. Monte-Carlo estimate relies on *independent* estimates of $\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z}$
- Matrix logarithm in form $\log(\boldsymbol{Q})\boldsymbol{z} \approx \sum_{i=1}^{N} \boldsymbol{A}_i^{-1}\boldsymbol{x}$. Solve N *independent* linear systems
- For extremely large matrices $\boldsymbol{A}$, matrix vector product $\boldsymbol{Ax}$ could be parallelised on multicore-machines.

Large Scale GMRF Some Details

- Likelihood estimate is of form $\sum_{i=0}^{\infty} \alpha_i$, where α_i built from *independent* estimates of the *same* quantity
- Truncate at some n, so need n estimates $\hat{\alpha}$
- Log-determinant estimate is of the form $\mathbb{E}_{\boldsymbol{z}}(\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z})$. Monte-Carlo estimate relies on *independent* estimates of $\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z}$
- Matrix logarithm in form $\log(\boldsymbol{Q})\boldsymbol{z} \approx \sum_{i=1}^{N} \boldsymbol{A}_i^{-1}\boldsymbol{x}$. Solve N *independent* linear systems
- For extremely large matrices $\boldsymbol{A}$, matrix vector product $\boldsymbol{Ax}$ could be parallelised on multicore-machines.
- Independent tasks, light-weight, and based on same data. Only scalar parameters differ.

Large Scale GMRF Some Details

- Likelihood estimate is of form $\sum_{i=0}^{\infty} \alpha_i$, where α_i built from *independent* estimates of the *same* quantity
- Truncate at some n, so need n estimates $\hat{\alpha}$
- Log-determinant estimate is of the form $\mathbb{E}_{\boldsymbol{z}}(\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z})$. Monte-Carlo estimate relies on *independent* estimates of $\boldsymbol{z}^T \log(\boldsymbol{Q})\boldsymbol{z}$
- Matrix logarithm in form $\log(\boldsymbol{Q})\boldsymbol{z} \approx \sum_{i=1}^{N} \boldsymbol{A}_i^{-1}\boldsymbol{x}$. Solve N *independent* linear systems
- For extremely large matrices $\boldsymbol{A}$, matrix vector product $\boldsymbol{A}\boldsymbol{x}$ could be parallelised on multicore-machines.
- Independent tasks, light-weight, and based on same data. Only scalar parameters differ.
- Both MCMC speed/mixing and problem size scale with number of nodes in cluster


Synchronise
Synchronise
MCMC Main-Node
Prepare Subtasks
Prepare Collection
Join Results/MCMC Step
Worker
Idle
Solve linear system
Idle
Worker
Idle
Solve linear system
Idle

Conclusions and Discussion

- Presented general methodology for Exact-Approximate MCMC

Conclusions and Discussion

- Presented general methodology for Exact-Approximate MCMC
- Exploits results from QCD literature, Russian Roulette, Absolute Measure Target

Conclusions and Discussion

- Presented general methodology for Exact-Approximate MCMC
- Exploits results from QCD literature, Russian Roulette, Absolute Measure Target
- Exact MCMC on massive scale models feasible

Conclusions and Discussion

- Presented general methodology for Exact-Approximate MCMC
- Exploits results from QCD literature, Russian Roulette, Absolute Measure Target
- Exact MCMC on massive scale models feasible
- Would be good to have general solution to Sign Problem beyond restrictive bounds

Conclusions and Discussion

- Presented general methodology for Exact-Approximate MCMC
- Exploits results from QCD literature, Russian Roulette, Absolute Measure Target
- Exact MCMC on massive scale models feasible
- Would be good to have general solution to Sign Problem beyond restrictive bounds
- Quality of mixing dependent on estimates of partition function

Conclusions and Discussion

- Presented general methodology for Exact-Approximate MCMC
- Exploits results from QCD literature, Russian Roulette, Absolute Measure Target
- Exact MCMC on massive scale models feasible
- Would be good to have general solution to Sign Problem beyond restrictive bounds
- Quality of mixing dependent on estimates of partition function
- Signed measure relaxes absolute bound in Generalised Poisson Estimators

Conclusions and Discussion

- Presented general methodology for Exact-Approximate MCMC
- Exploits results from QCD literature, Russian Roulette, Absolute Measure Target
- Exact MCMC on massive scale models feasible
- Would be good to have general solution to Sign Problem beyond restrictive bounds
- Quality of mixing dependent on estimates of partition function
- Signed measure relaxes absolute bound in Generalised Poisson Estimators
- Massively parallelizable - a very good thing

Acknowledgements

- Girolami funded by EPSRC Established Fellowship and Royal Society Wolfson Research Merit Award