DEPARTMENT OF
STATISTICS

Large Sample Asymptotics of the Pseudo-Marginal Algorithm

Sebastian Schmon

Department of Statistics
University of Oxford

25 April 2019

Introduction
Bayesian Inference and MCMC

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

- Likelihood function $p_\theta(y)$ and prior distribution of density $p(\theta)$.
- Bayesian inference relies on the posterior

$$\pi(\theta) = p(\theta|y) = \frac{p_\theta(y)\, p(\theta)}{\int_\Theta p_{\theta'}(y)\, p(\theta')\, d\theta'}.$$

- For non-trivial models, inference often relies on MCMC.
- Basic Idea: Simulate an ergodic Markov chain $\{\vartheta_i\}_{i\geq 1}$ of limiting distribution $\pi(\theta)$.
- Problem: Metropolis-Hastings (MH) cannot be implemented if $p_\theta(y)$ cannot be evaluated.

Introduction
Pseudo-Marginal Approach (I)

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

"Idea": Replace $p_\vartheta(y)$ by an estimate $\widehat{p}_\vartheta(y)$ in MH.

At iteration i

- Sample $\vartheta \sim q(\cdot \mid \vartheta_{i-1})$.
- Compute an estimate $\widehat{p}_\vartheta(y)$ of $p_\vartheta(y)$.

Introduction
Pseudo-Marginal Approach (II)

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

With probability

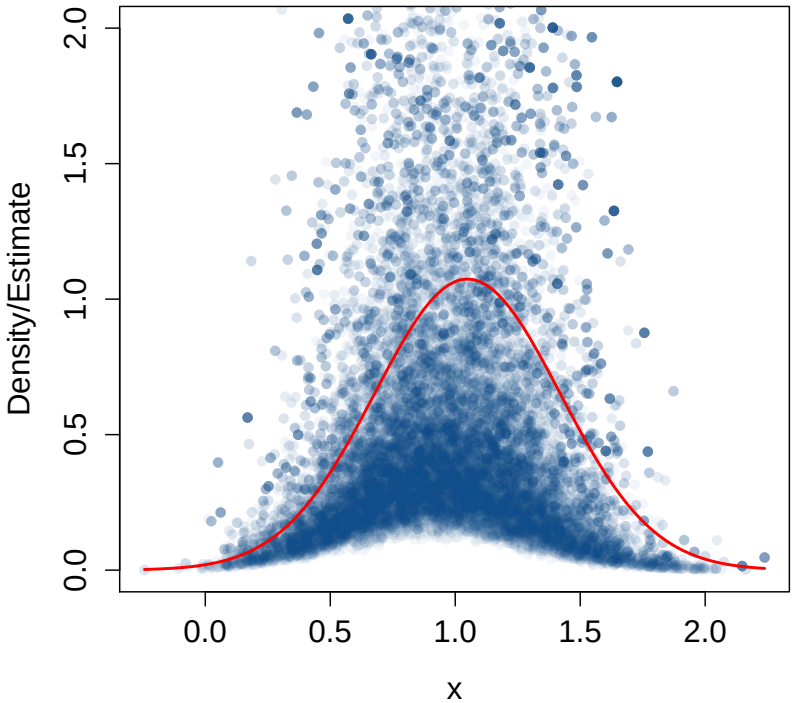
$$\min\left\{1, \frac{\widehat{p}_{\vartheta}(y)\, p(\vartheta)}{\widehat{p}_{\vartheta_{i-1}}(y)\, p(\vartheta_{i-1})} \frac{q(\vartheta_{i-1}|\vartheta)}{q(\vartheta|\vartheta_{i-1})}\right\}$$

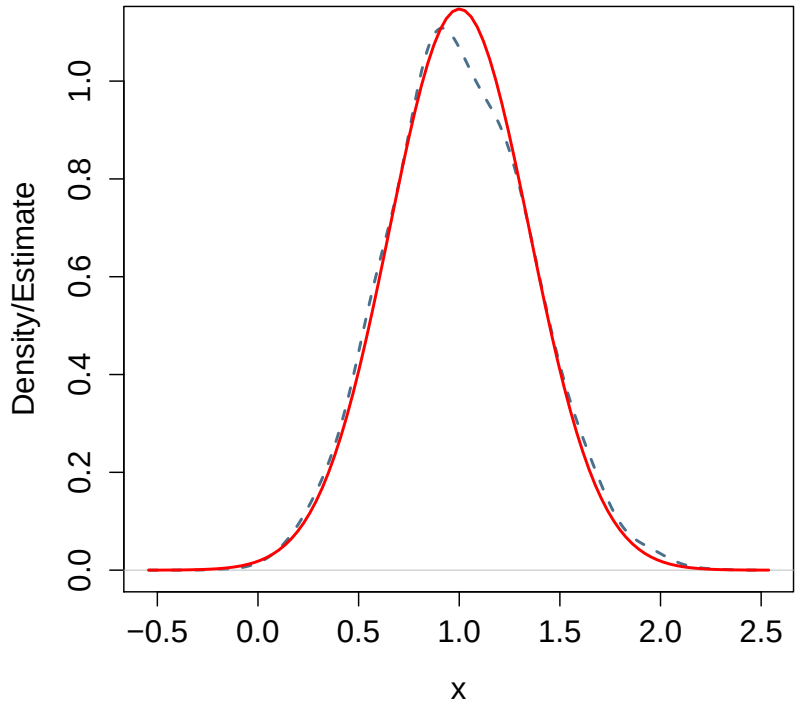
$$= \min\left\{1, \underbrace{\frac{p_{\vartheta}(y)}{p_{\vartheta_{i-1}}(y)} \frac{p(\vartheta)}{p(\vartheta_{i-1})} \frac{q(\vartheta_{i-1}|\vartheta)}{q(\vartheta|\vartheta_{i-1})}}_{\text{exact MH ratio}} \times \underbrace{\frac{\widehat{p}_{\vartheta}(y)/p_{\vartheta}(y)}{\widehat{p}_{\vartheta_{i-1}}(y)/p_{\vartheta_{i-1}}(y)}}_{\text{noise}}\right\},$$

set $\vartheta_i = \vartheta$, $\widehat{p}_{\vartheta_i}(y) = \widehat{p}_{\vartheta}(y)$ otherwise set $\vartheta_i = \vartheta_{i-1}$, $\widehat{p}_{\vartheta_i}(y) = \widehat{p}_{\vartheta_{i-1}}(y)$.

Introduction

Illustration: True density vs. estimate





Introduction
Pseudo-Marginal Approach (III)

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

- Proposition: If $\widehat{p}_{\vartheta}(y)$ is a non-negative unbiased estimator of $p_{\theta}(y)$ then the pseudo-marginal MH admits $\pi(\theta)$ as invariant density.
- Let $Z = \log\{\widehat{p}_{\theta}(y)/p_{\theta}(y)\}$ be the error in log-likelihood estimator and introduce an auxiliary target density on $\Theta \times \mathbb{R}$

$$\overline{\pi}(\theta, z) = \pi(\theta) \underbrace{\exp(z) g_{\theta}(z)}_{\text{unbiasedness} \Leftrightarrow \int(\cdot)dz=1}$$

  where $Z \sim g_{\theta}$; e.g. importance sampling or particle filter.
- In this notation $\hat{p}_{\theta}(y) = p_{\theta}(y)e^{Z}$.

Introduction
Pseudo-Marginal Approach (IV)

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

Pseudo-marginal MH is a standard MH of target $\overline{\pi}(\theta, z)$ and proposal $q(\vartheta|\theta)\, g_\vartheta(z)$ as

$$\frac{\overline{\pi}(\vartheta, w)}{\overline{\pi}(\theta, z)} \frac{q(\theta|\vartheta)\, g_\theta(z)}{q(\vartheta|\theta)\, g_\vartheta(w)} = \frac{\widehat{p}_\vartheta(y)}{\widehat{p}_\theta(y)} \frac{p(\vartheta)}{p(\theta)} \frac{q(\theta|\vartheta)}{q(\vartheta|\theta)}$$

Validity of pseudo-marginal MH does not rely on conditions on the variance of $\widehat{p}_\vartheta(y)$ but its quantitative properties do!

Pseudo-Marginal Approach
Optimising the computing time

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

- Hence, the question arises which number of particles, N, balances the computational effort and gain in precision through lower variance.
- To be able to address this problem current research papers make the simplifying assumption that $g(\mathrm{d}z \mid \theta) = \varphi(\mathrm{d}z; -\sigma^2/2, \sigma^2)$, that $\sigma^2 \propto 1/N$

$$\mathrm{CT}(h, P_\sigma) := \frac{\tau(h, P_\sigma)}{\sigma^2}.$$

Pseudo-Marginal Approach
Optimising the computing time

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

- Recent approaches to find optimality criteria of Pseudo-Marginal schemes (e.g. Doucet et al. 2015 and Sherlock et al. 2015) rely on the assumption that the noise induced by the likelihood estimate is independent of the current state of the chain
- This assumption is not fulfilled in most practical settings. However, it can be justified theoretically under very mild assumptions!

Pseudo-Marginal Approach
Large Sample Asymptotics

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

- "Idea": Use the regularity structure of the Bernstein-von Mises theorem, i.e.

$$\int \left| \pi^T(\theta) - \varphi(\theta; \widehat{\theta}^T, \overline{\Sigma}/T) \right| \mathrm{d}\theta \xrightarrow{\mathbb{P}} 0, \quad \text{as} \quad T \to \infty$$

  and use classical large data asymptotics to explore the behaviour of the algorithm.

Pseudo-Marginal Approach
Large Sample Asymptotics

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

- In addition use the fact that under certain regularity condition the noise density converges to

$$g_\theta(z) = \varphi\left(z, -\frac{\sigma^2}{2}, \sigma^2\right)$$

  as the number of particles goes to infinity together with the data (see Berard et.al (2014) or Deligiannidis et. al. (2015)).
- Usually number of particles $N = \lfloor \gamma T \rfloor$ and $T \to \infty$.

Large Sample Asymptotics
Theorem

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

- Under mild regularity condition a scaled random walk Metropolis chain $(\boldsymbol{\theta}_n, Z_n)_{n\in\mathbb{N}}$ converges weakly (on $\mathbb{R}^\infty$) to a Markov chain with new equilibrium distribution

$$\begin{aligned}\widetilde{\pi}_\infty(\widetilde{\theta}, z) &= \varphi(\widetilde{\theta}, 0, \overline{\Sigma}) e^z \varphi\left(z, -\frac{\sigma^2(\overline{\theta})}{2}, \sigma^2(\overline{\theta})\right) \\ &= \varphi(\widetilde{\theta}, 0, \overline{\Sigma}) \varphi\left(z, \frac{\sigma^2(\overline{\theta})}{2}, \sigma^2(\overline{\theta})\right)\end{aligned}$$

- This is saying that, asymptotically, for any n $(\boldsymbol{\theta}_n, Z_n)$ are a pair of independent normal distributions, thus justifying the independence assumption theoretically.

Large Sample Asymptotics
Theorem

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

The transition kernel is given by

$$\tilde{P}_{\ell,\sigma}(\widetilde{\theta}, z; \mathrm{d}\widetilde{\theta}', \mathrm{d}z') = \widetilde{q}(\widetilde{\theta}, \mathrm{d}\widetilde{\theta}')\varphi\left(\mathrm{d}z', -\sigma^2/2, \sigma^2\right)\tilde{\alpha}(\widetilde{\theta}, z; \widetilde{\theta}', z') + \widetilde{\rho}(\widetilde{\theta}, z)\delta_{(\widetilde{\theta}, z)}(\mathrm{d}\widetilde{\theta}', \mathrm{d}z').$$

where $\sigma := \sigma(\bar{\theta})$ and

$$\tilde{\alpha}(\widetilde{\theta}, z; \widetilde{\theta}', z') = \min\left\{1, \frac{\varphi(\widetilde{\theta}'; 0, \Sigma)}{\varphi(\widetilde{\theta}; 0, \Sigma)} \frac{\widetilde{q}(\widetilde{\theta}', \widetilde{\theta})}{\widetilde{q}(\widetilde{\theta}, \widetilde{\theta}')} \exp\left(z' - z\right)\right\}.$$

Large Sample Asymptotics
Theorem

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

The transition kernel is given by

$$\tilde{P}_{\ell,\sigma}(\widetilde{\theta}, z; \mathrm{d}\widetilde{\theta}', \mathrm{d}z') = \widetilde{q}(\widetilde{\theta}, \mathrm{d}\widetilde{\theta}')\varphi\left(\mathrm{d}z', -\sigma^2/2, \sigma^2\right)\tilde{\alpha}(\widetilde{\theta}, z; \widetilde{\theta}', z') + \widetilde{\rho}(\widetilde{\theta}, z)\delta_{(\widetilde{\theta}, z)}(\mathrm{d}\widetilde{\theta}', \mathrm{d}z').$$

where $\sigma := \sigma(\bar{\theta})$ and

$$\tilde{\alpha}(\widetilde{\theta}, z; \widetilde{\theta}', z') = \min\left\{1, \frac{\varphi(\widetilde{\theta}'; 0, \Sigma)}{\varphi(\widetilde{\theta}; 0, \Sigma)} \frac{\widetilde{q}(\widetilde{\theta}', \widetilde{\theta})}{\widetilde{q}(\widetilde{\theta}, \widetilde{\theta}')} \underbrace{\exp\left(z' - z\right)}_{\text{difference to "marginal" chain}}\right\}.$$

Optimal Parameters
Setting

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

- Consider a normal random walk with proposal

$$q(\theta, \theta') = \varphi\left(\theta'; \theta, \frac{\ell^2}{d} I_d\right).$$

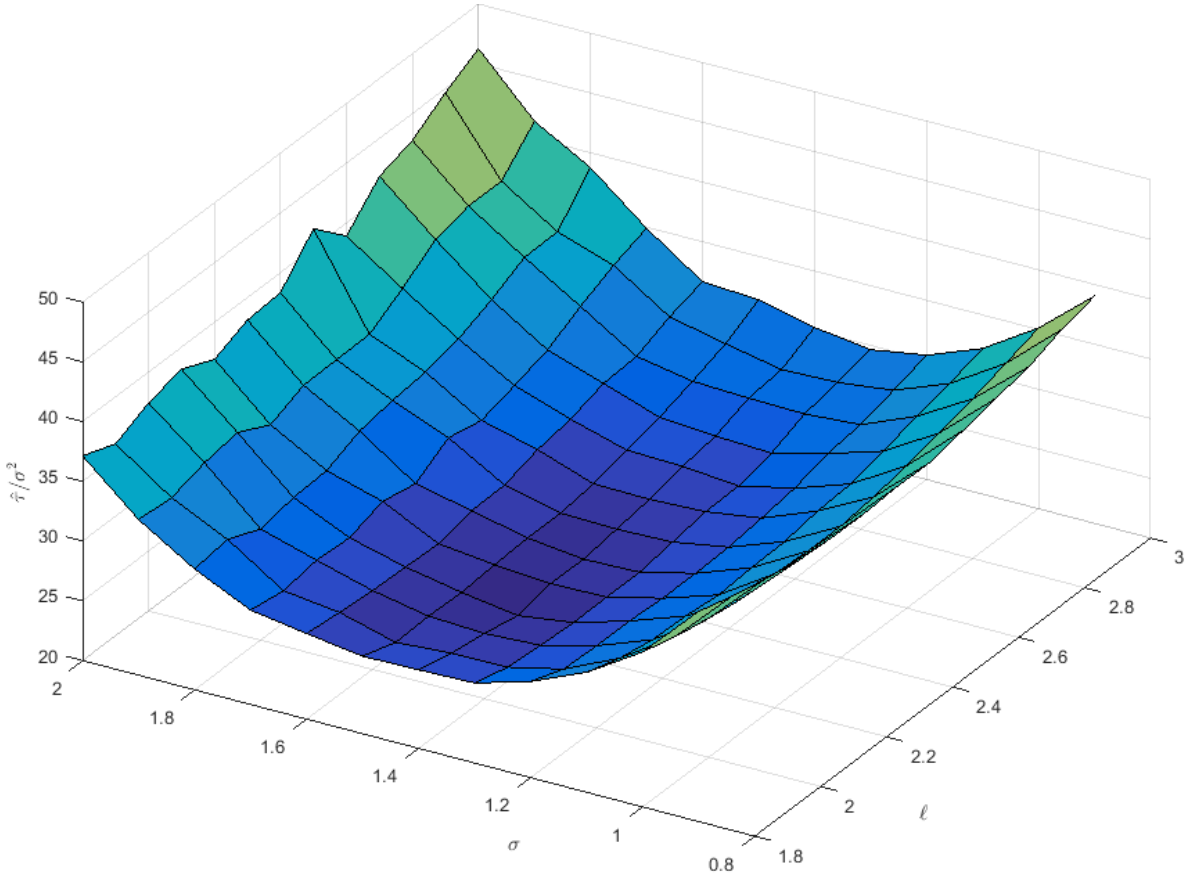
- We minimize the computing time

$$\mathrm{CT}(h, \widetilde{P}_{\ell,\sigma}) := \frac{\tau(h, \tilde{P}_{\ell,\sigma})}{\sigma^2}$$

where τ denotes the integrated autocorrelation time of $\tilde{P}_{\ell,\sigma}$.

Optimal Parameters
Simulation Results

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS



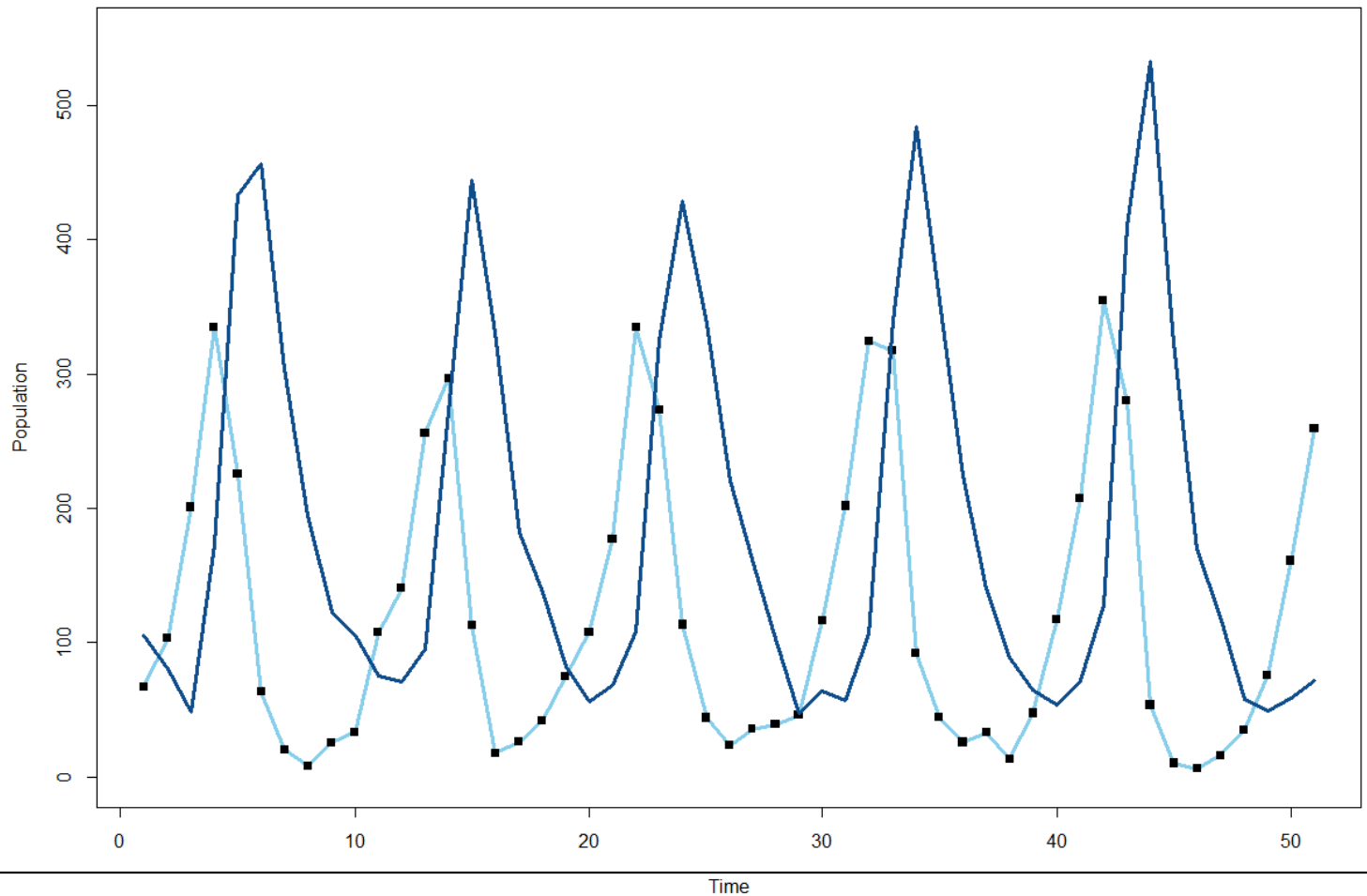
Optimal Parameters
Simulation Results

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

Dimension d	$\hat{\ell}_{\text{opt}}$	$\hat{\sigma}_{\text{opt}}$	$p_{\text{jump}}(\hat{\sigma}_{\text{opt}}, \hat{\ell}_{\text{opt}})$
1	2.2	1.1	25.61%
2	2	1.2	21.70%
3	2	1.3	18.55%
5	2.1	1.4	16.12%
10	2.2	1.4	13.81%
15	2.3	1.5	12.20%
20	2.3	1.6	11.54%
30	2.3	1.6	10.10%
50	2.4	1.8	8.00%

Application to Systems Biology
Stochastic Lotka-Volterra Model

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS



Application to Systems Biology
Stochastic Lotka-Volterra Model

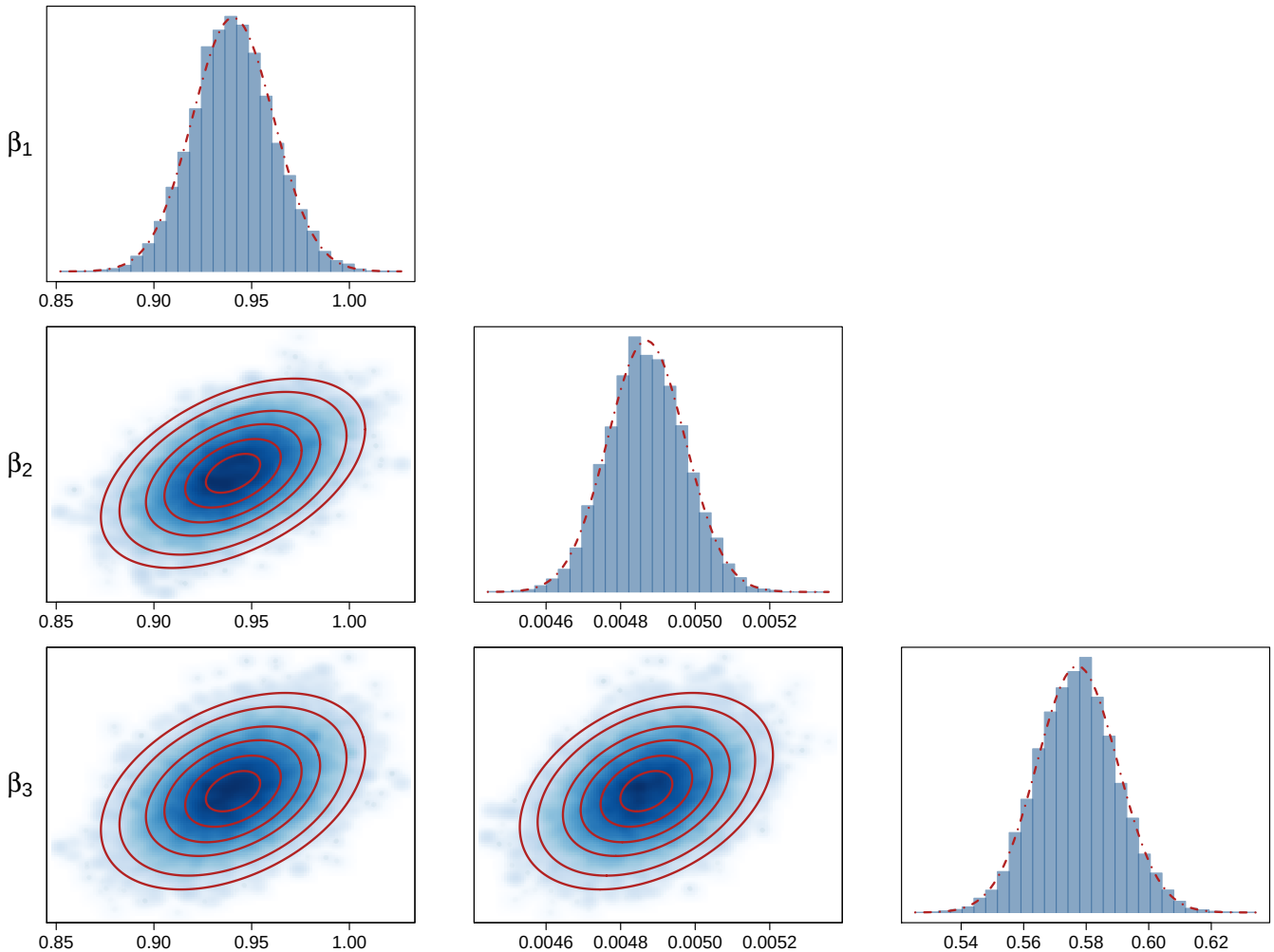
UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

Transition equations: As $h \to 0$

$$\mathbb{P}\left(X_{1,t+h} - X_{1,t} = 1, X_{2,t+h} - X_{2,t} = 0 \mid X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t}\right) = \beta_1 x_{1,t} + o(h)$$
$$\mathbb{P}\left(X_{1,t+h} - X_{1,t} = -1, X_{2,t+h} - X_{2,t} = 1 \mid X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t}\right) = \beta_2 x_{1,t} x_{2,t} + o(h)$$
$$\mathbb{P}\left(X_{1,t+h} - X_{1,t} = 0, X_{2,t+h} - X_{2,t} = -1 \mid X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t}\right) = \beta_3 x_{2,t} + o(h)$$

Optimal Parameters

Check: Bernstein-von Mises



Application to Systems Biology
Stochastic Lotka-Volterra Model

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

- Minimum computing time is recorded at 225 and 250 particles respectively.
- This corresponds to $\sigma = 1.47$ and $\sigma = 1.34$, both with approximately 23% acceptance rate.
- Confirms earlier research suggesting that σ is between 0.96 and 1.8.

References

Bérard, J.; Del Moral, P. & Doucet, A. *A lognormal central limit theorem for particle approximations of normalizing constants* Electron. J. Probab, 2014, 19, 1-28

Deligiannidis, G.; Doucet, A. & Pitt, M. K. *The Correlated Pseudo-Marginal Method* arXiv preprint arXiv:1511.04992, 2015

Doucet, A.; Pitt, M. K.; Deligiannidis, G. & Kohn, R. *Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator* Biometrika, 2015, 102 (2), 295-313

Pitt, M. K., dos Santos Silva, R., Giordani, P., & Kohn, R. *On some properties of Markov chain Monte Carlo simulation methods based on the particle filter* Journal of Econometrics, 2012, 171(2), 134-151.

Sherlock, C.; Thiery, A. H.; Roberts, G. O.; Rosenthal, J. S. & others *On the efficiency of pseudo-marginal random walk Metropolis algorithms* The Annals of Statistics, 2015, 43, 238-275