Noisy MCMC Algorithms for Gibbs Random Fields

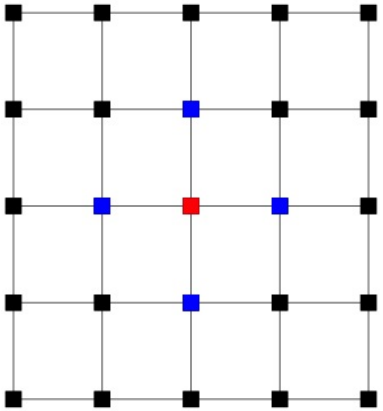
Aidan Boland
Pierre Alquire, Nial Friel, Richard Everitt (University of Reading)

12/2/14

Outline

Introduction

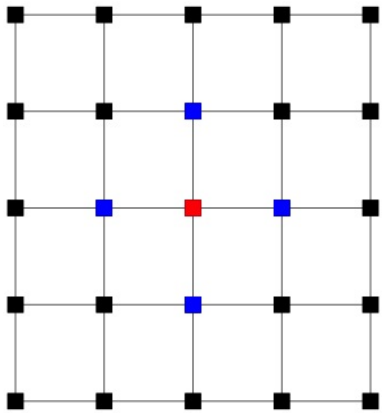
- What is a graph?
- What is a Gibbs random field?
  - Set of nodes with corresponding random variables (y).
  - $f(y|\theta) = \frac{1}{z(\theta)} \exp(\theta^T s(y))$
- They are tricky to work with due to intractable likelihood.
  - $z(\theta) = \sum_y \exp\{\theta^T s(y)\}$
  - Summation over all possible graphs $\{2^{n \times m}\}$



Ising example

Introduction

- What is a graph?
- What is a Gibbs random field?
  - Set of nodes with corresponding random variables (y).
  - $f(y|\theta) = \frac{1}{z(\theta)} \exp(\theta^T s(y))$
- They are tricky to work with due to intractable likelihood.
  - $z(\theta) = \sum_y \exp\{\theta^T s(y)\}$
    - Summation over all possible graphs $\left(2^{\frac{n(n-1)}{2}}\right)$



Ising example

Objectives

- Interested in the posterior distribution $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$.
- Markov chain Monte Carlo is a general approach to simulate from the posterior.
- Create a Markov chain whose stationary distribution matches the distribtuion of the posterior.

MCMC

- Simulate a Markov Chain $(\theta_n)_{n\in\mathbb{N}}$ using a transition kernel P where π is invariant under P, $(\pi P = \pi)$.
- Can then use approximation,

$$\frac{1}{N}\sum_{n=1}^{N} f(\theta_n) \approx \int_{\Theta} f(\theta)\pi(d\theta)$$

- To ensure the stationary distribution of the Markov chain matches the posterior distribution, we need conditions on P, such as uniform ergodicity.

$$\sup_{\theta_0} \left\| \delta_{\theta_0} P^n - \pi \right\| \leq C\rho^n$$

for some $C < \infty$ and $\rho < 1$, where $\|\cdot\|$ is the total variation distance.

MCMC and GRF's

- A natural kernel P exists for Gibbs random fields, however due to the intractability of the likelihood it is not feasible to draw $\theta_{n+1} \sim P(\cdot|\theta_n)$.
- We propose to replace P by an approximation $\hat{P}$.
- Obviously $\hat{P}$ should be 'close' to P.
- Using the study of stability of Markov chains, it is possible to put an upper bound on the difference between the Markov chains resulting from $\hat{P}$ and P.

Theorem (Mitrophanov (2005), Corollary 3.1)
Let us assume that

- **(H1)** the Markov chain with transition kernel P is uniformly ergodic:

$$\sup_{\theta_0} \|\delta_{\theta_0} P^n - \pi\| \leq C\rho^n$$

for some $C < \infty$ and $\rho < 1$.

Then we have, for any $n \in \mathbb{N}$, for any starting point θ_0,

$$\|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \left(\lambda + \frac{C\rho^\lambda}{1-\rho}\right) \|P - \hat{P}\|$$

where $\lambda = \left\lceil \frac{\log(1/C)}{\log(\rho)} \right\rceil$.

Application of theorem

- Will now show how the theorem can be applied in the case of Gibbs random fields.

Metropolis-Hastings algorithm

- Propose a new value $\theta' \sim h(\cdot|\theta)$
- Accept the θ' with probability:
  - $\alpha(\theta'|\theta) = \min\left(1, \frac{q_{\theta'}(y)\pi(\theta')h(\theta|\theta')}{q_{\theta}(y)\pi(\theta)h(\theta'|\theta)} \times \frac{Z(\theta)}{Z(\theta')}\right)$
- Depends on intractable ratio $\frac{Z(\theta)}{Z(\theta')}$

Exchange algorithm

- Introduce an auxiliary variable ($y' \sim f(\cdot|\theta')$).
- This algorithm samples from the augmented distribution
  - $\pi(\theta', y', \theta|y) \propto f(y|\theta)\pi(\theta)h(\theta'|\theta)f(y'|\theta')$
- The acceptance ratio then simplifies into:
  - $\hat{\alpha}(\theta'|\theta, y') = \min\left(1, \frac{q_{\theta'}(y)\pi(\theta')h(\theta|\theta')q_{\theta}(y')}{q_{\theta}(y)\pi(\theta)h(\theta'|\theta)q_{\theta'}(y')} \times \frac{Z(\theta)Z(\theta')}{Z(\theta')Z(\theta)}\right)$

Noisy exchange algorithm

- M-H
  - $$\frac{q_{\theta'}(y)\pi(\theta')h(\theta|\theta')}{q_{\theta}(y)\pi(\theta)h(\theta'|\theta)} \times \frac{Z(\theta)}{Z(\theta')}$$
- Exchange
  - $$\frac{q_{\theta'}(y)\pi(\theta')h(\theta|\theta')}{q_{\theta}(y)\pi(\theta)h(\theta'|\theta)} \times \frac{q_{\theta}(y')}{q_{\theta'}(y')}$$

Noisy exchange algorithm

- Exchange algorithm replaces the ratio of normalising contants.
  - $\mathbb{E}_{y' \sim f(\cdot|\theta')} \left(\frac{q_\theta(y')}{q_{\theta'}(y')} \right) = \frac{Z(\theta)}{Z(\theta')}$
- Idea of noisy exchange is to use a better estimate of ratio of normalizing constants.
  - $\frac{1}{N} \sum_{i=1}^{N} \frac{q_\theta(y')}{q_{\theta'}(y')} = \frac{Z(\theta)}{Z(\theta')}$
- As $N \to \infty$, the algorithm will create a Markov chain which converges to the true posterior distribution

Noisy exchange theoretical guarantees

- We have replaced α from the original Metropolis Hastings algorithm with an approximation $\hat{\alpha}$.
- Can apply Theorem to show convergence.

Corollary

Let us assume that

- **(H1)** the Markov chain with transition kernel P is uniformly ergodic holds,
- **(H2)** $\hat{\alpha}(\theta|\theta', y')$ satisfies:

$$\mathbb{E}_{y' \sim F_{\theta'}} \left| \hat{\alpha}(\theta|\theta', y') - \alpha(\theta|\theta') \right| \leq \delta(\theta, \theta'). \tag{1}$$

Then we have, for any $n \in \mathbb{N}$, for any starting point θ_0,

$$\|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \left(\lambda + \frac{C\rho^\lambda}{1-\rho} \right) \sup_\theta \int \mathrm{d}\theta' \, h(\theta'|\theta) \delta(\theta, \theta'),$$

where $\lambda = \left\lceil \frac{\log(1/C)}{\log(\rho)} \right\rceil$.

Noisy exchange theoretical guarantees

- For Gibbs random fields we have,

Lemma

$\hat{a}(\theta'|\theta, y')$ satisfies **(H2)** in the Corollary with

$$\mathbb{E}_{y' \sim f(\cdot|\theta')} \left| \hat{a}(\theta, \theta', y') - a(\theta, \theta') \right| \leq \delta(\theta, \theta')$$

$$= \frac{1}{\sqrt{N}} \frac{h(\theta|\theta')\pi(\theta')q_{\theta'}(y)}{h(\theta'|\theta)\pi(\theta)q_{\theta}(y)} \sqrt{\mathrm{Var}_{y' \sim f(y'|\theta')} \left(\frac{q_{\theta_n}(y')}{q_{\theta'}(y')} \right)}.$$

Noisy exchange theoretical guarantees

- Even further we can show

Theorem
Assuming the space Θ is bounded, then $\hat{\alpha}$ with,

$$\delta(\theta, \theta') \leq \frac{c_h^2 c_\pi^2 \mathcal{K}^4}{\sqrt{N}},$$

and

$$\sup_{\theta_0 \in \Theta} \|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \frac{\mathcal{C}}{\sqrt{N}}$$

where $\mathcal{C} = \mathcal{C}(c_\pi, c_h, \mathcal{K})$ is explicitly known.

Langevin

- Langevin diffusion is defined by the stochastic differential equation

$$d\theta(t) = \nabla \log \pi(\theta(t))dt/2 + db(t),$$

- Not possilbe to solve so a discretized version is used

$$\theta_{i+1} = \theta_i + \frac{\Sigma}{2}\nabla \log \pi(\theta_i) + \epsilon \qquad \epsilon \sim N(0, \Sigma)$$

- Unavailable for GRF's since the gradient, $\nabla \log \pi(\theta_i)$, is in tractable.
- We can use a monte carlo estimate of the gradient to create a noisy Langevin algorithm.

Noisy Langevin

- $\nabla \log \pi(\theta|y) = s(y) - \mathbb{E}_{y|\theta}(s(y)) + \nabla \log \pi(\theta)$
- Can estimate the gradient using monte carlo.
  - Draw $(y'_1, ..., y'_N) \sim f(\cdot|\theta)$
  - $\widehat{\nabla} \log \pi(\theta|y) = s(y) - \frac{1}{N} \sum_i^N (s(y'_i)) + \nabla \log \pi(\theta)$
- The noisy langevin algorithm is then:

$$\theta_{i+1} = \theta_i + \frac{\Sigma}{2} \widehat{\nabla} \log \pi(\theta_i|y) + \epsilon \qquad \epsilon \sim N(0, \Sigma)$$

Noisy Langevin theoretical guarantees

Corollary

- **(H1)** the Markov chain with transition kernel P is uniformly ergodic holds,
- **(H3)** the gradient estimator satisfies, for any θ,

$$\mathbb{E}_{y' \sim F_{\theta_n}} \left\{ \exp \left[\frac{1}{2} \left\| \Sigma^{\frac{1}{2}} (\nabla \log \pi(\theta) - \hat{\nabla}^{y'} \log \pi(\theta)) \right\|^2 \right] - 1 \right\} \leq \delta \qquad (2)$$

for some $\delta > 0$.

Then we have, for any $n \in \mathbb{N}$, for any starting point θ_0,

$$\| \delta_{\theta_0} P_\Sigma^n - \delta_{\theta_0} \hat{P}_\Sigma^n \| \leq \left(\lambda + \frac{C \rho^\lambda}{1 - \rho} \right) \sqrt{\frac{\delta}{2}}.$$

where $\lambda = \left\lceil \frac{\log(1/C)}{\log(\rho)} \right\rceil$.

Noisy Langevin theoretical guarantees

Lemma

As soon as $N > 4k\mathcal{S}^2\|\Sigma\|^2$, assumption **(H3)** is satisfied with

$$\delta = \exp\left(\frac{k\log(N)}{4\mathcal{S}^2\|\Sigma\|^2 N}\right) - 1 + \frac{4k\sqrt{\pi}\mathcal{S}\|\Sigma\|}{N} \sim_{N\to\infty} \frac{k\log\left(\frac{N}{k}\right)}{4\mathcal{S}^2\|\Sigma\|^2 N}$$

(where $\|\Sigma\| = \sup\{\|\Sigma x\|, \|x\| = 1\}$).

MALA-Exchange

- Extend the noisy langevin by introducing an accept/reject step.
- Combine with the exchange algorithm.
- The accept/reject step ensures the Markov chain targets the true density.
  - Draw $y' = (y'_1, ..., y'_N) \sim f(\cdot|\theta)$, and calculate $\widehat{\nabla} \log \pi(\theta|y)$
  - Draw $\theta' = \theta_i + \frac{c}{2}\widehat{\nabla} \log \pi(\theta_i) + \epsilon$
    where $\epsilon \sim N(0, C)$
  - Accept θ' with probability:
    $$\min\left(1, \frac{q_{\theta'}(y)\pi(\theta')h(\theta|\theta')q_\theta(y')}{q_\theta(y)\pi(\theta)h(\theta'|\theta)q_{\theta'}(y')}\right)$$

Noisy MALA exchange

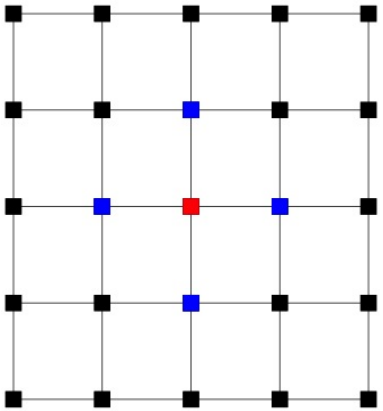
- Similar to noisy exchange, we can replace the ratio of normalising constants with a Monte carlo approximation.
- No extra computaional effort as $y'_1, ..., y'_N$ is already required for the MALA exchange.
- $$\frac{1}{N}\sum_{i=1}^{N}\frac{q_\theta(y')}{q_{\theta'}(y')} = \frac{Z(\theta)}{Z(\theta')}$$

Results

- Ising study,
  - single parameter model.
- Ergm study,
  - two parameter model.

Ising

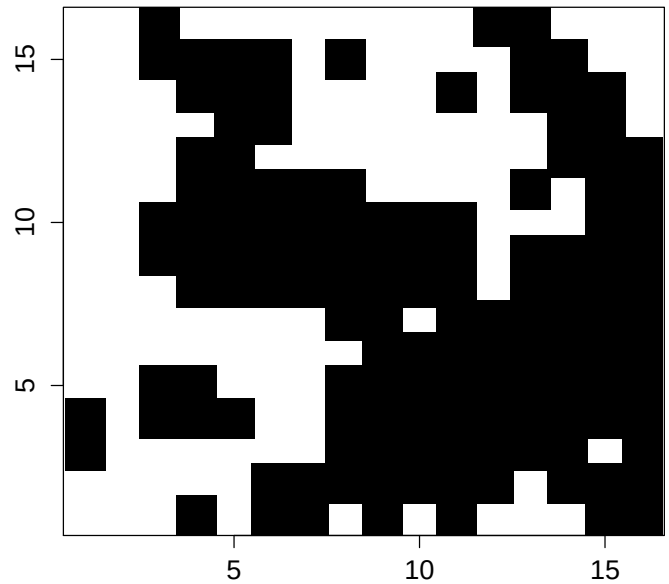
- Single parameter model.
- Defined on a rectangular lattice.
- Models spatial distribution of binary variables.
- $f(y|\theta) = \frac{1}{Z(\theta)} exp\left\{\theta \sum_{j=1}^{N} \sum_{i \sim j} y_i y_j\right\}$
  - $i \sim j$ denotes that i and j are neighbours.



Ising example

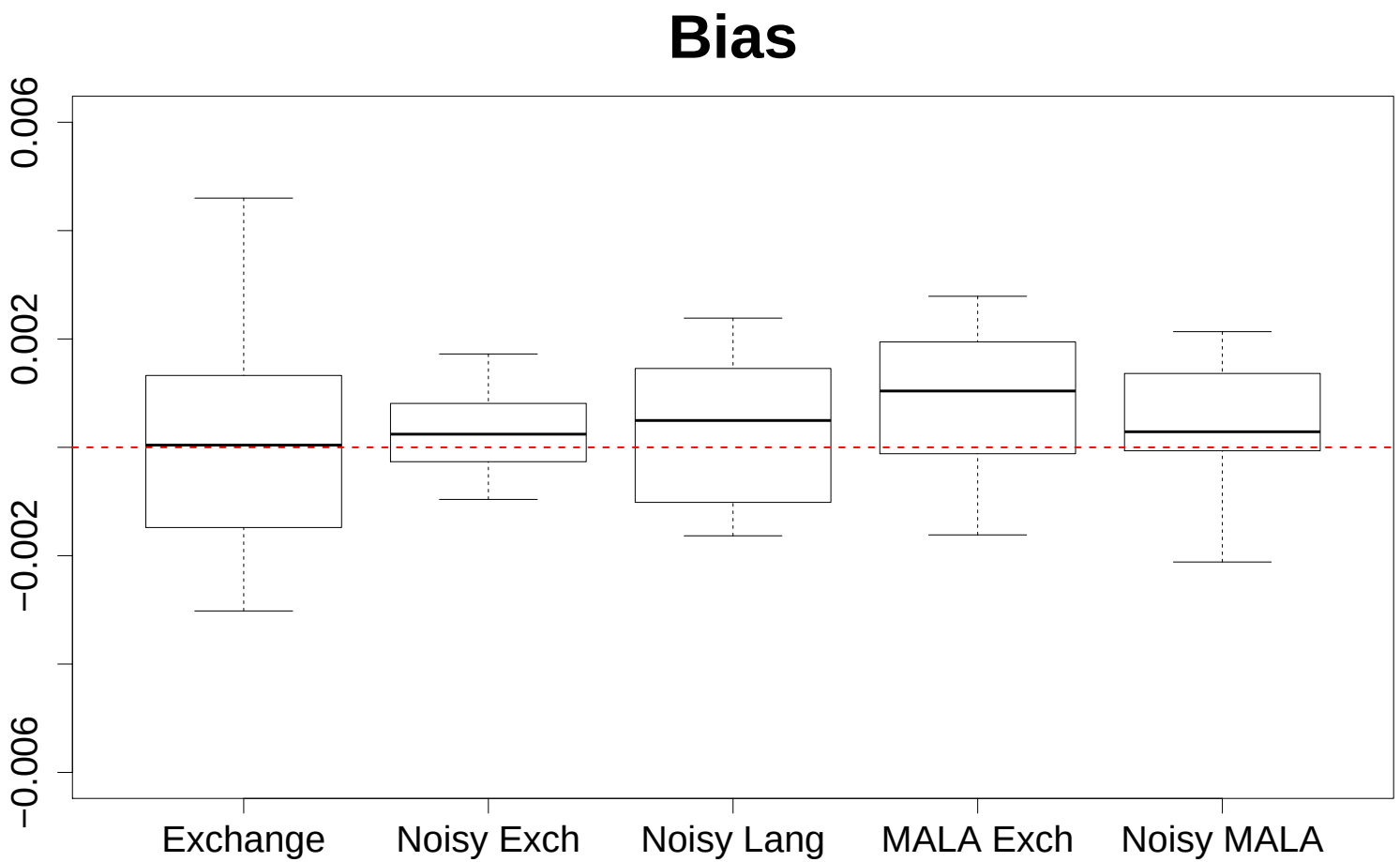
Ising

- 20 simulated graphs.
- Each a 16x16 lattice.
- True posterior can be calcuated.



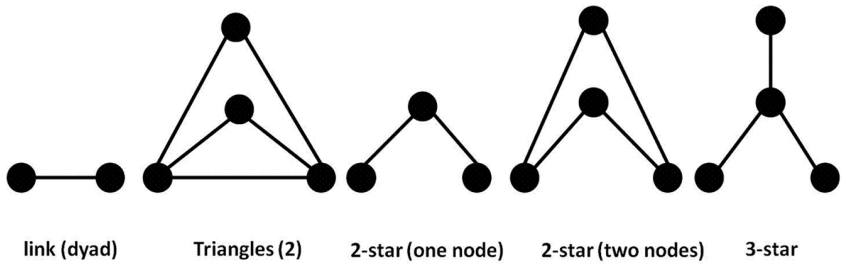
Ising data

Ising


Bias
0.006
0.002
−0.002
−0.006
Exchange
Noisy Exch
Noisy Lang
MALA Exch
Noisy MALA

ERGM

- Exponential Random Graph Model
- Used in analysis of social networks.
- $f(y|\theta) = \frac{q_\theta(y)}{Z(\theta)} = \frac{exp(\sum_{i=1}^{m} \theta_i s_i(y))}{Z(\theta)}$
  - y observed graph.
  - $s(y)$ vector of sufficient statistics.



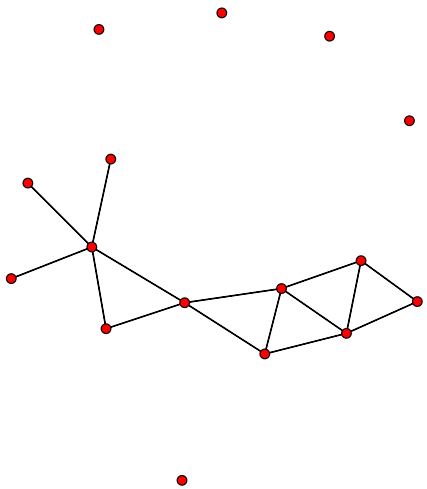
Florentine business

- Florentine Business dataset (around 1430).
- 16 families, each represented by a node.

- Edge between two nodes if the corresponding families have a business connection.
- Fit a 2-dimensional model

$$f(y|\theta) = \frac{1}{Z(\theta)} exp(\theta_1 s_1(y) + \theta_2 s_2(y))$$

  - $s_1(y)$ is number of edges in the graph and $s_2(y)$ is the number of two-stars.



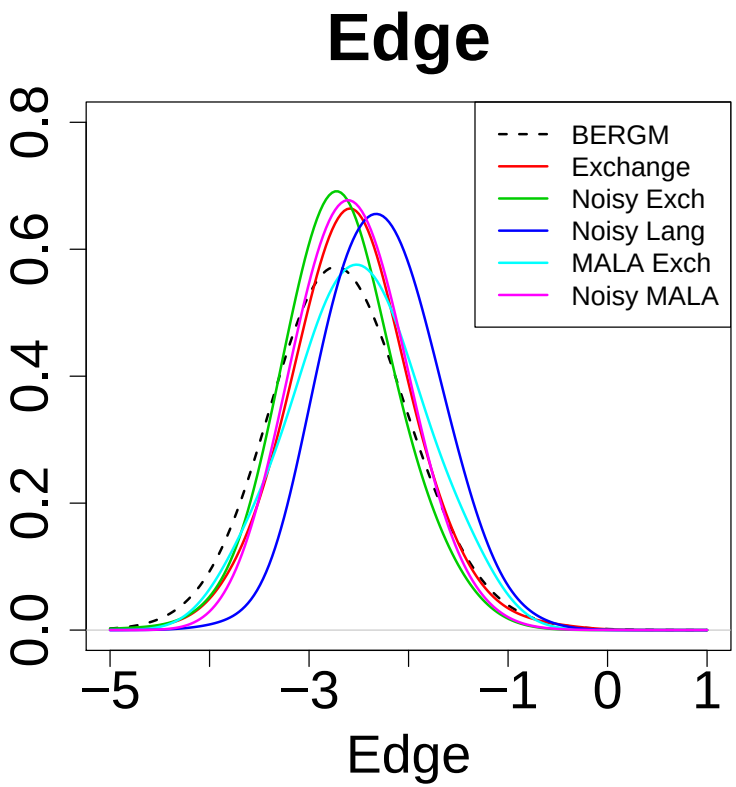
Florentine data

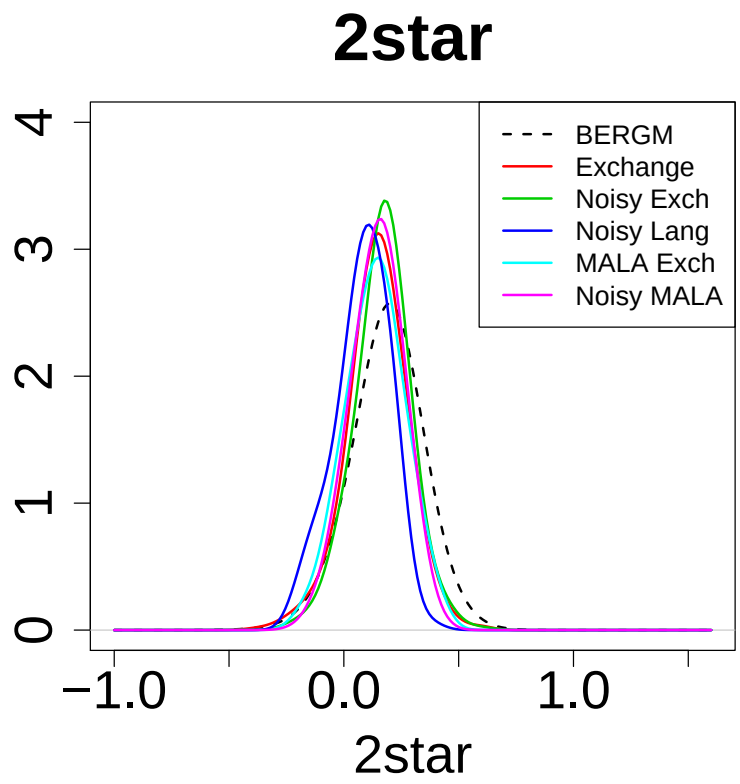
Florentine business

Method	Edge Mean	Edge SD	2-star Mean	2-star SD
BERGM	-2.675	0.647	0.188	0.155
Exchange	-2.573	0.568	0.146	0.133
Noisy Exchange	-2.686	0.526	0.167	0.122
Noisy Langevin	-2.281	0.513	0.081	0.119
MALA Exchange	-2.518	0.62	0.136	0.128
Noisy MALA	-2.584	0.498	0.144	0.113

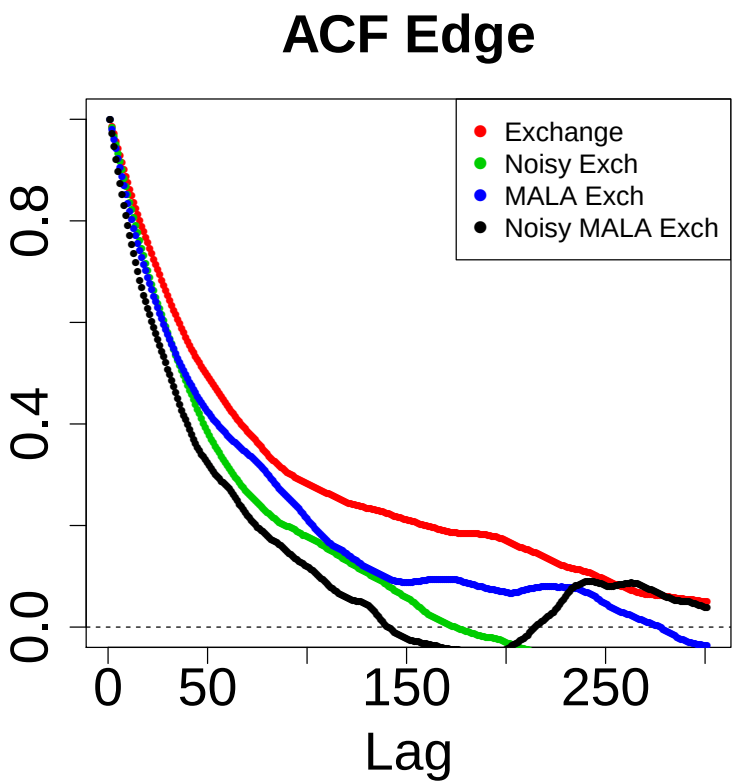
Table: Posterior means and standard deviation.

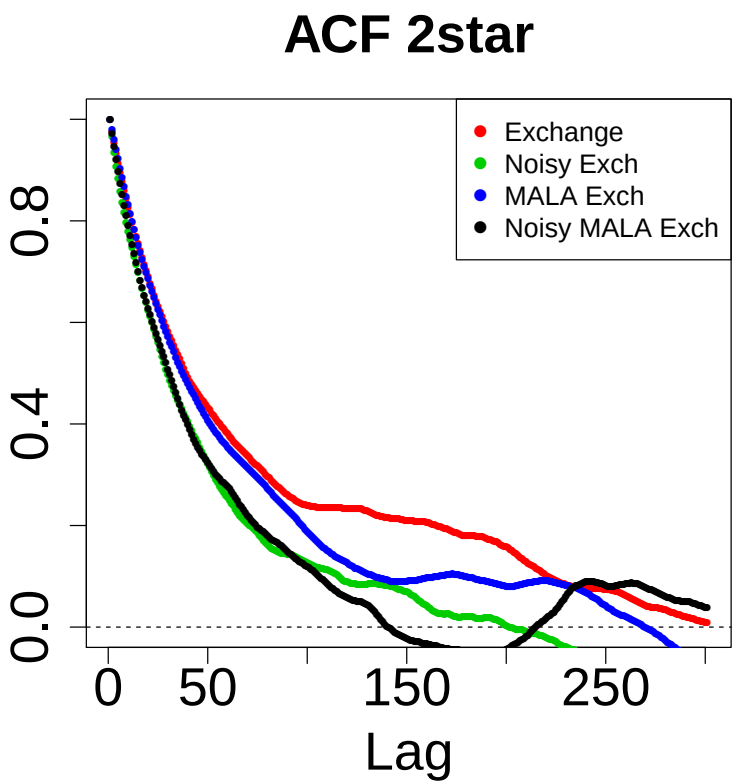
Florentine business





Florentine business





References

- Andrieu and Roberts (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics.*
- Mitrophanov (2005) *Sensitivity and convergence of uniformly ergodic Markov chains.* Journal of applied probability.
- Murray, Ghahramani and MacKay, (2006) *MCMC for doubly-intractable distribution.* In Proceedings of the 22nd annual conference on uncertainty in artificial intelligence.

END

- Any questions?

Funded under the programme for Research in Third-level Institutions and co-funded under the European Regional Development fund