Waterloo Statistics and Biostatistics Seminar

Resolving the Incongruity between Variational Inference and Markov chain Monte Carlo with the Occlusion Process

Max Hird, University of Waterloo

Joint work with Florian Maire, Université de Montréal

Outline

1. Introduction to Variational Inference (VI) and Markov chain Monte Carlo (MCMC)

2. The incongruity between VI and MCMC

3. Resolving the incongruity with the occlusion process

Interrupt me at any point if you have questions

Variational Inference

VI is optimisation over the space of distributions

Find $q^* = \arg \min_{q \in \mathscr{P}(\mathbb{R}^d)} d(q, \pi)$

Purpose is to approximate $\mathbb{E}_\pi[f(X)]$ with $\mathbb{E}_{q^*}[f(X)]$

Members of $\mathscr{P}(\mathbb{R}^d)$ cannot be stored on a computer

Find $\theta^* = \arg \min_{\theta \in \Theta} d(q_\theta, \pi)$ and compute $\mathbb{E}_{q_{\theta^*}}[f(X)]$

Variational Inference cont.

VI is **biased** i.e. $q_{\theta*} \neq \pi$ in general

Optimisation often engenders a ‘nice’ $q_{\theta*}$:

- Its properties can be ‘read off’
- Sampleable independently
  - $\Rightarrow$ we can form Monte Carlo estimates of $\mathbb{E}_{q_{\theta^*}}[f(X)]$

VI is **fast** (once we’ve found $q_{\theta*}$)

Markov chain Monte Carlo

Assume that our knowledge of π is such that we can't sample independently

MCMC consists of constructing Markov chains $\{X_t\}_{t=1}^n$ such that

$$\hat{f}_n := \frac{1}{n}\sum_{t=1}^{n} f(X_t) \to \mathbb{E}_\pi[f(X)]$$

in some appropriate sense

This is usually done by constructing a π-invariant Markov kernel (usually time homogeneous)

By 'Markov Kernel' we simply mean the conditional density defining the chain:

$$K(x \to A) := \mathbb{P}(X_{k+1} \in A \,|\, X_k = x)$$

for all $x \in \mathbb{R}^d$ and $A \in \mathscr{B}(\mathbb{R}^d)$ (and $k \in \mathbb{N}$)

Markov chain Monte Carlo Evaluation

If $X_0 \sim \pi$ then

$$\mathrm{Var}(\hat{f}_n) = \frac{\mathrm{Var}_\pi(f)}{n}\left(1 + 2\sum_{t=1}^{n-1}\frac{n-t}{n}\mathrm{Corr}\left(f(X_1), f(X_{t+1})\right)\right)$$

We can judge the efficiency of an MCMC output $\{X_t\}_{t=1}^n$, $X_0 \sim \pi$ using the **Effective Sample Size** (ESS)

The ESS is defined as

$$\mathrm{ESS} = \frac{n\mathrm{Var}_\pi(f)}{\lim_{n\to\infty} n\mathrm{Var}(\hat{f}_n)}$$

i.e. it's the amount of independent samples to achieve a variance equal to that of the Markov chain estimator

It is estimated using a chain initialised at $X_0 \sim \pi$

Markov chain Monte Carlo cont.

MCMC is **slow** because it is inherently serial

i.e. to get X_n we need X_{n-1} for which we need X_{n-2} etc.

But it is **asymptotically exact**

VI is potentially **fast** but **biased**

MCMC is **asymptotically exact** but **slow**

Strengths of one method cover the weaknesses of the other

Given q_θ and a Markov kernel can we achieve a better method than either in isolation?

We assume that we want to use VI within MCMC for the exactness

Incongruity between VI and MCMC

MCMC is fundamentally **local**

This is both its blessing and its curse

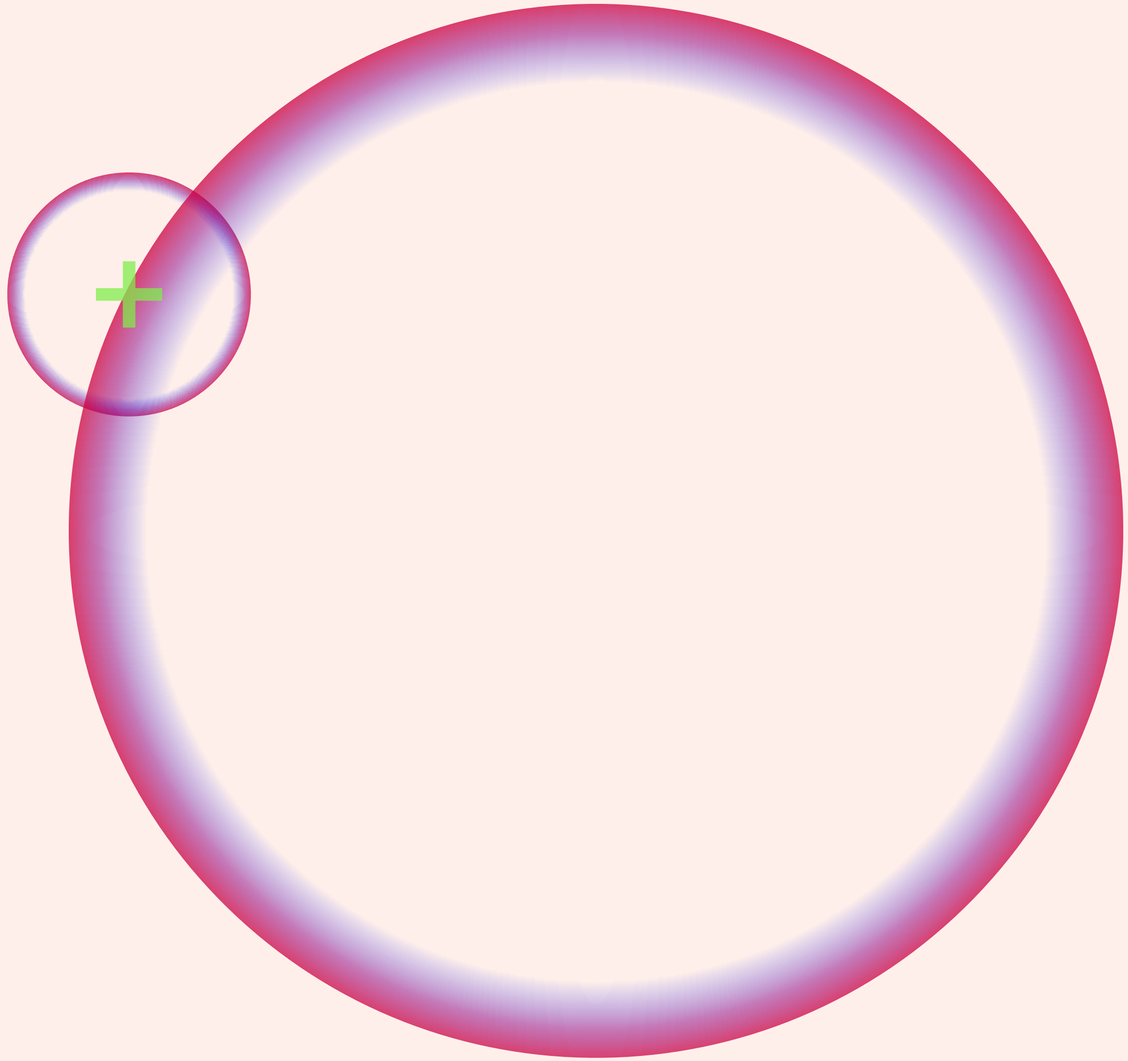
VI produces an object (q_θ) that is **global**

i.e. q_θ attempts to imitate the entirety of π since $d(q_\theta, \pi)$ usually compares q_θ and π at all points in the state space

Therefore either

- Integrate VI into an MCMC method that has a **global** approximation to π as a tuning parameter
  - e.g. Rejection sampling, importance sampling, importance Metropolis-Hastings
  - These are known to fail catastrophically
- Or come up with some clever solution

A Naïve combination of VI and MCMC



The Occlusion Process

Recall

$$\mathrm{Var}(\hat{f}_n) = \frac{\mathrm{Var}_\pi(f)}{n}\left(1 + 2\sum_{t=1}^{n-1}\frac{n-t}{n}\mathrm{Corr}\left(f(X_1), f(X_{t+1})\right)\right)$$

The Occlusion Process takes as an input an MCMC kernel and a variational approximation $Q \in \mathscr{P}(\mathbb{R}^d)$

It reduces variance by selectively replacing terms in $\{X_t\}_{t=1}^n$ with samples from Q

Samples from Q are easy to generate and can be done in parallel to the Markov chain, increasing the probability that a larger number of states will be replaced

The Process cont.

Partition the state space X into $\left\{\mathsf{X}_i : i \in [R]\right\}$

Define π_i as π restricted to X_i i.e $\pi_i(A) := \pi\left(\mathsf{X}_i\right)^{-1} \pi\left(A \cap \mathsf{X}_i\right)$

Define $\rho : \mathsf{X} \to [R]$ with $\rho(x) \mapsto$ index of the part that x is in

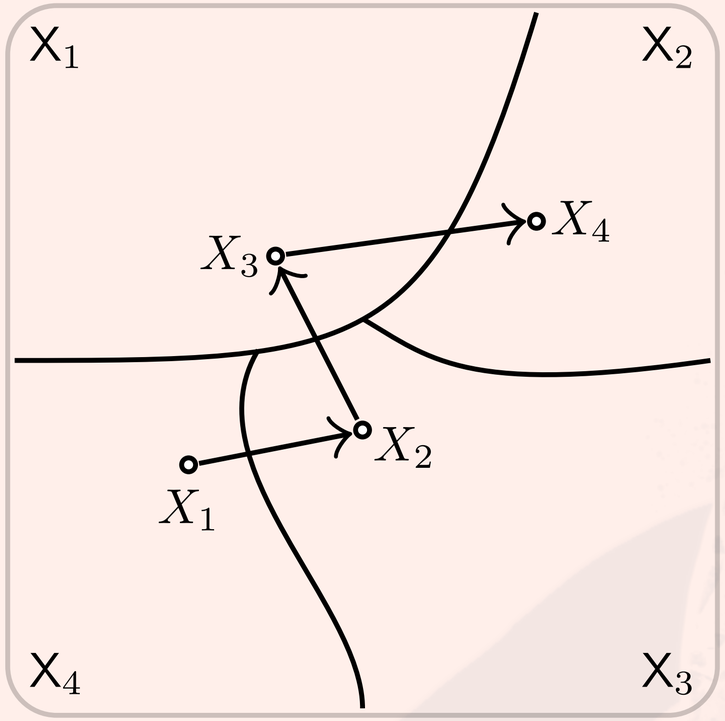
Idea: when $X_t \in \mathsf{X}_i$ attempt to sample $Y_t \sim \pi_i$ using Q

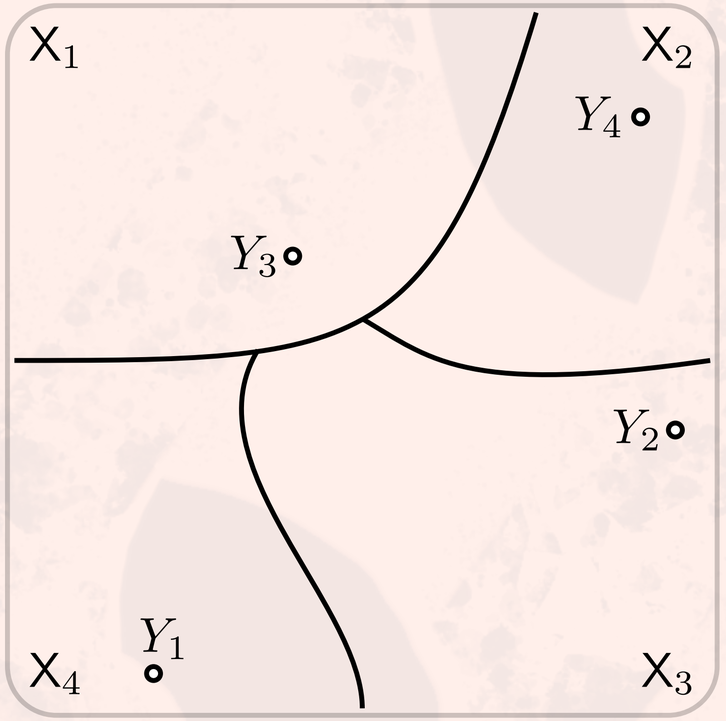
When successful replace X_t with Y_t **in the estimator** $\hat{f}_n$

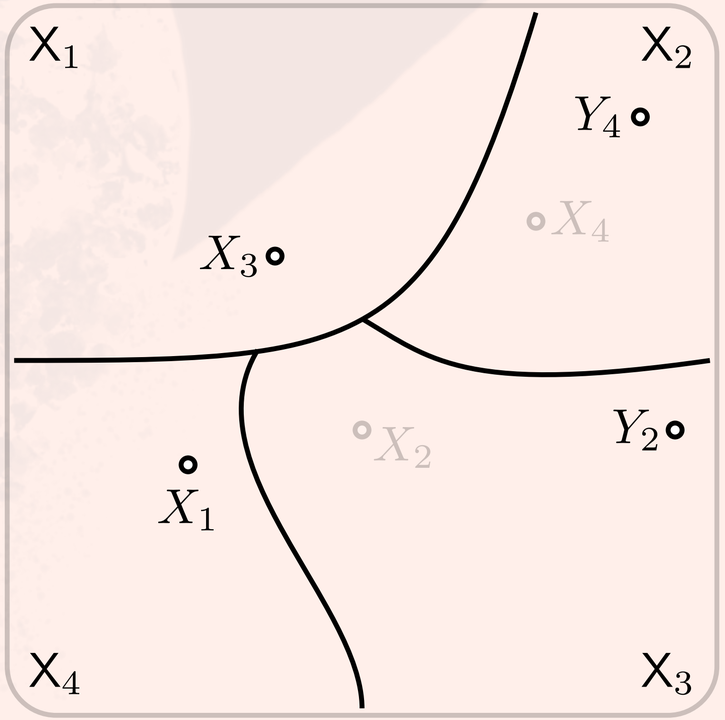
We say Y_t has ***occluded*** X_t

Y_t is ***less correlated*** with $\left\{X_t\right\}_{t=1}^{n}$ than X_t

The Process cont.







The Process

In general we will not be able to sample $Y_t \sim \pi_{\rho(X_t)}$ for all $t \in [n]$

$$K_{\text{occ}}\left(\left(x, s, y\right) \rightarrow \left(dx', s', dy'\right)\right) =$$

$$K\left(x \rightarrow dx'\right)\left(\alpha\left(\rho\left(x'\right)\right)\mathbf{1}\left\{s' = 1\right\} + \left(1 - \alpha\left(\rho\left(x'\right)\right)\right)\mathbf{1}\left\{s' = 0\right\}\right)\pi_{\rho(x')}\left(dy'\right)$$

Occluded estimator:

$$\frac{1}{n}\sum_{t=1}^{n} f_{\text{occ}}\left(X_t, S_t, Y_t\right)$$

where

$$f_{\text{occ}}\left(X_t, S_t, Y_t\right) := \mathbf{1}\left\{S_t = 0\right\} f\left(X_t\right) + \mathbf{1}\left\{S_t = 1\right\} f\left(Y_t\right)$$

K_{occ} is π_{occ}-invariant such that $\mathbb{E}_{\pi_{\text{occ}}}[f(X, S, Y)] = \mathbb{E}_{\pi}[f(X)]$

Sampling from π_i with Q

Let $C > 0$

1. Sample $Y \sim Q$ and $U \sim$ Uniform $[0,1]$ (independently)
2. If

$$U \leq \frac{1}{C} \frac{d\pi}{dQ}(Y) \leq 1$$

accept Y otherwise go to step 1.

Defining $\mathsf{X}_C := \left\{ y : \frac{1}{C} \frac{d\pi}{dQ}(y) \leq 1 \right\}$ then Y is sampled from π restricted to X_C [Tierney, 1994, Section 2.3.4]

Sampling from π_i with Q cont.

Let $0 =: C_0 < C_1 < \cdots < C_{R-1} < C_R := \infty$ partition the 'Radon-Nikodym' space $[0,\infty]$

Define

$$\mathsf{X}_i := \left\{ x : \frac{d\pi}{dQ}(x) \in [C_{i-1}, C_i) \right\}$$

1. Sample $Y \sim Q$ and $U \sim$ Uniform $[0,1]$ (independently)
2. Get the smallest C_i greater than $d\pi/dQ\,(Y)$ (s.t. $Y \in \mathsf{X}_i$)
3. If

$$U \leq \frac{1}{C_i} \frac{d\pi}{dQ}(Y)$$

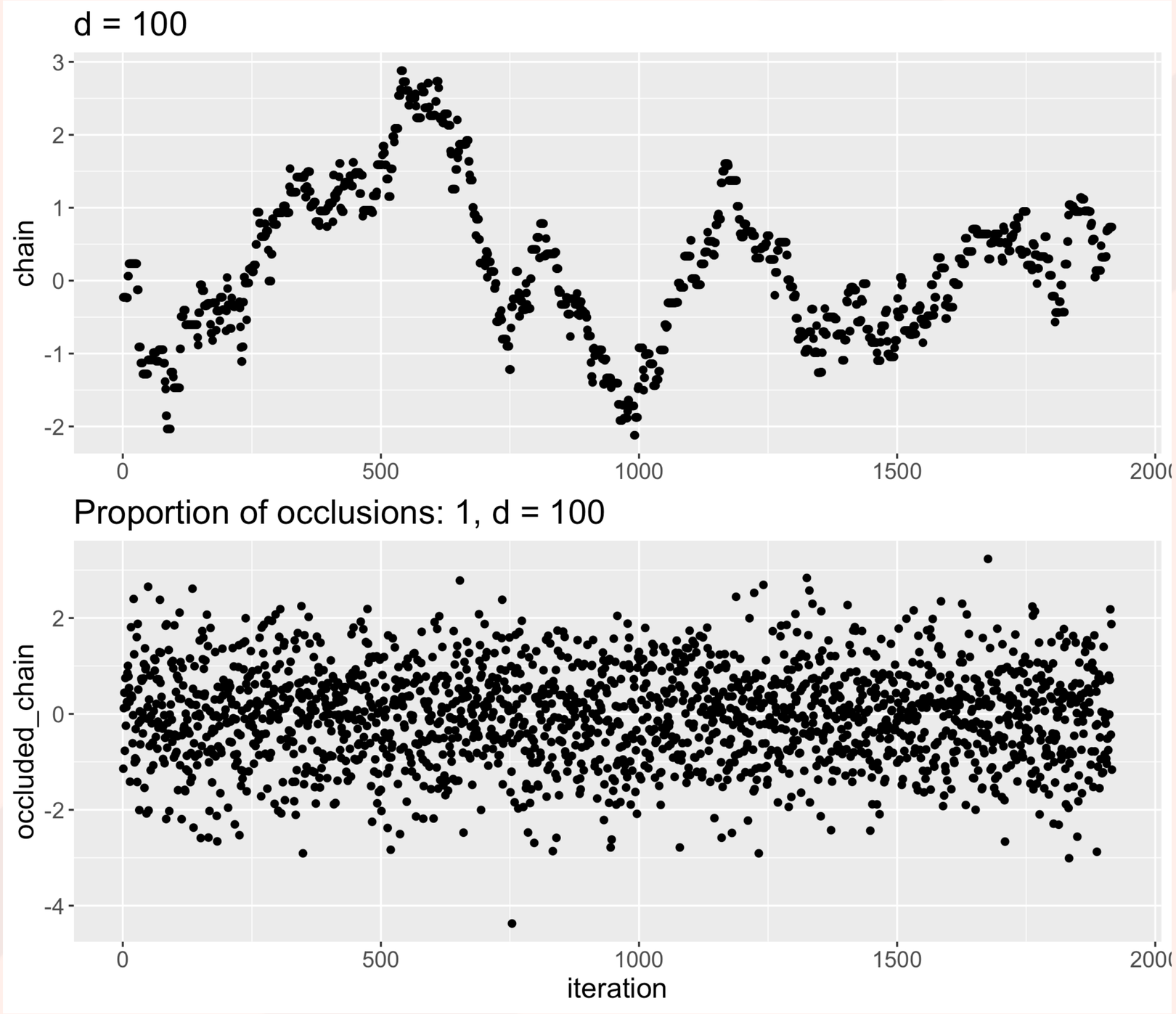
accept Y otherwise go to step 1.

So more regions means a high acceptance probability, but possible more autocorrelation

Concrete implementation with Parallel Computation

Modern MCMC is run with independent chains on every core, see e.g. [Margossian and German 2023]

- Our experiments use 7 cores:
  - Normal chain: run a chain on every core
  - Occluded process:
    - Run the MCMC on k cores
    - Attempt to sample from π_i's using Q on $7 - k$ cores
    - For a given region occlude uniformly at random


d = 100
chain
3
2
1
0
-1
-2
0
500
1000
1500
2000
Proportion of occlusions: 1, d = 100
occluded_chain
2
0
-2
-4
0
500
1000
1500
2000
iteration

Inherited properties of the occlusion process

What properties of $\left(K, f, \pi\right)$ does $\left(K_{\text{occ}}, f_{\text{occ}}, \pi_{\text{occ}}\right)$ inherit?

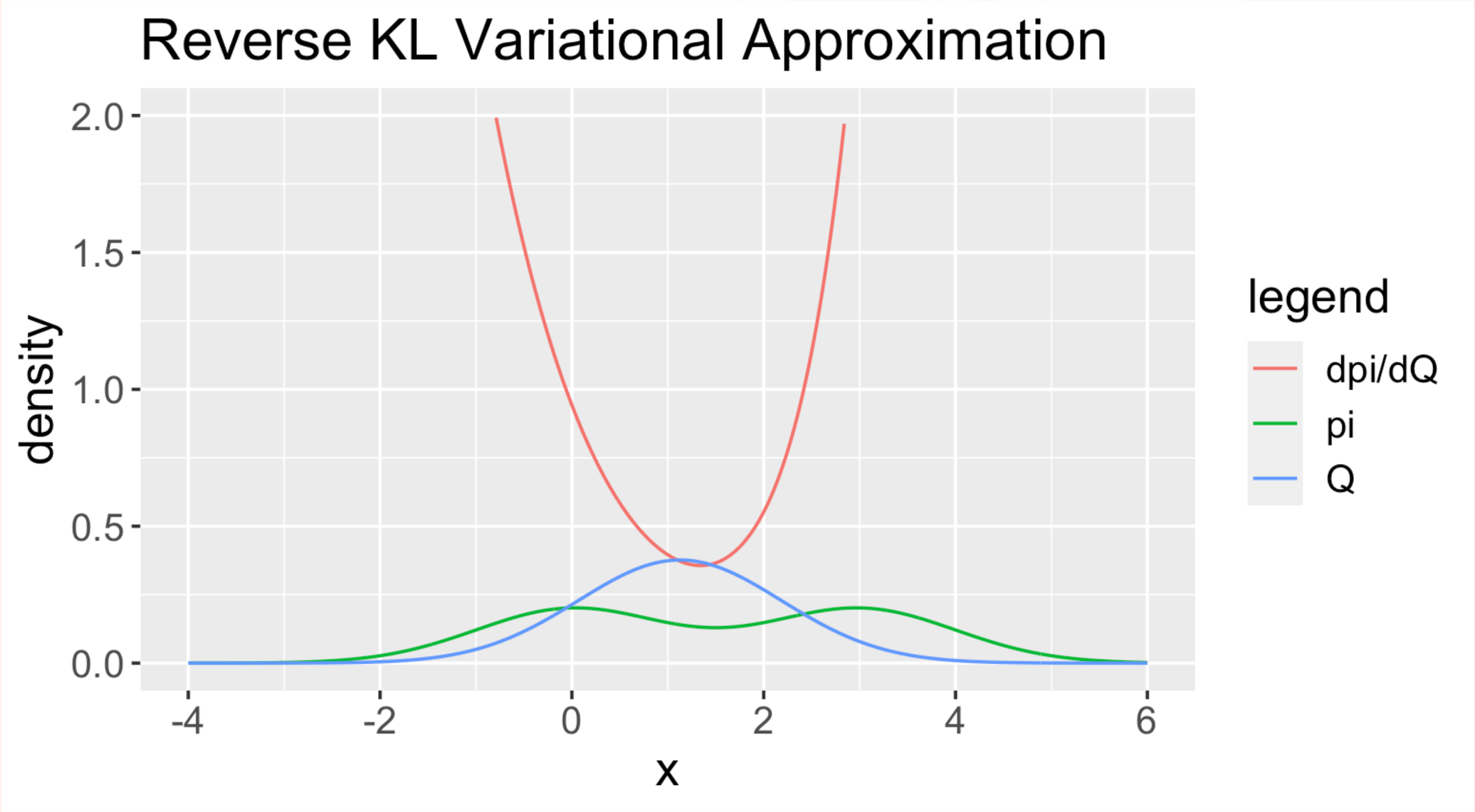
1. LLN for all $g \in L^1\left(\pi\right)$ $\iff$ LLN for all $g_{\text{occ}} \in L^1\left(\pi_{\text{occ}}\right)$
2. K converges in a normed function space with rate $r\left(t\right)$ $\implies$ K_{occ} converges in a normed function space with rate $r\left(t\right)$
3. K converges in a normed measure space with rate $r\left(t\right)$ $\implies$ K_{occ} converges in a normed measure space with rate $r\left(t\right)$
4. K is geometrically ergodic with rate $\exp\left(-\lambda t\right)$ $\implies$ K_{occ} is geometrically ergodic with rate $\exp\left(-\lambda t\right)$
5. K is geometrically ergodic and π-reversible and $f \in L^2\left(\pi\right)$ $\implies$ the occlusion process satisfies a CLT

Experiments: Gaussian Mixture

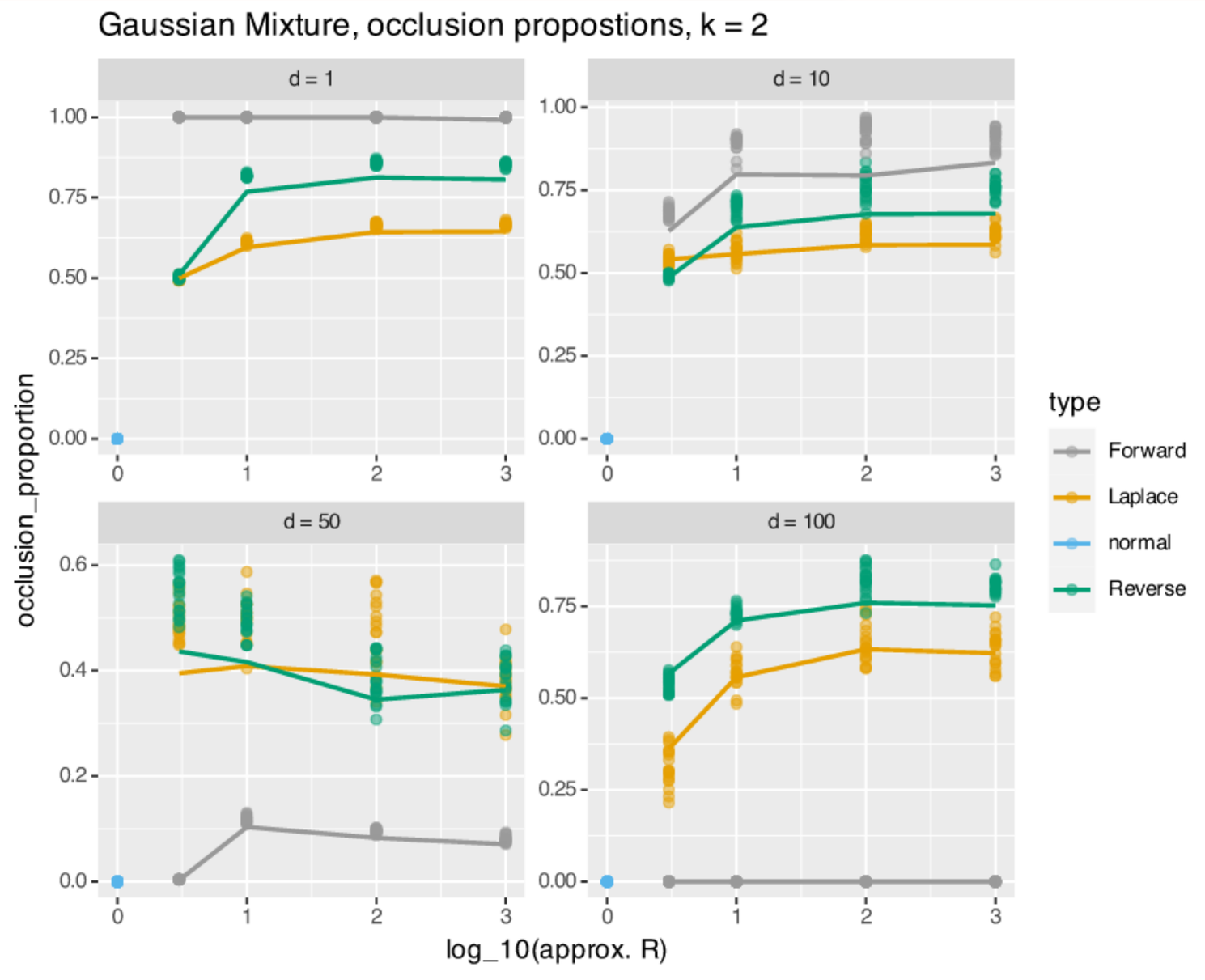
$$\pi(\mathrm{d}x) = 0.5 \times \mathcal{N}\left(\mathrm{d}x; 0, \mathbf{I}_d\right) + 0.5 \times \mathcal{N}\left(\mathrm{d}x; (3,0,\ldots,0)^\top, \mathbf{I}_d\right)$$

- $d \in \{1, 10, 50, 100\}$
- $R \in \{3, 10, 100, 1000\}$
- $Q =$ Normal
  - Laplace approximation
  - Gradient descent on the Reverse KL
  - Gradient descent on the Forward KL I.e. $\mu_Q = \mu_\pi$ and $\mathrm{Cov}_Q = \mathrm{Cov}_\pi$
- Regions:
  - Run a short MCMC chain and then split the histogram of the Radon-Nikodym's into equal bins
- $k = 2$

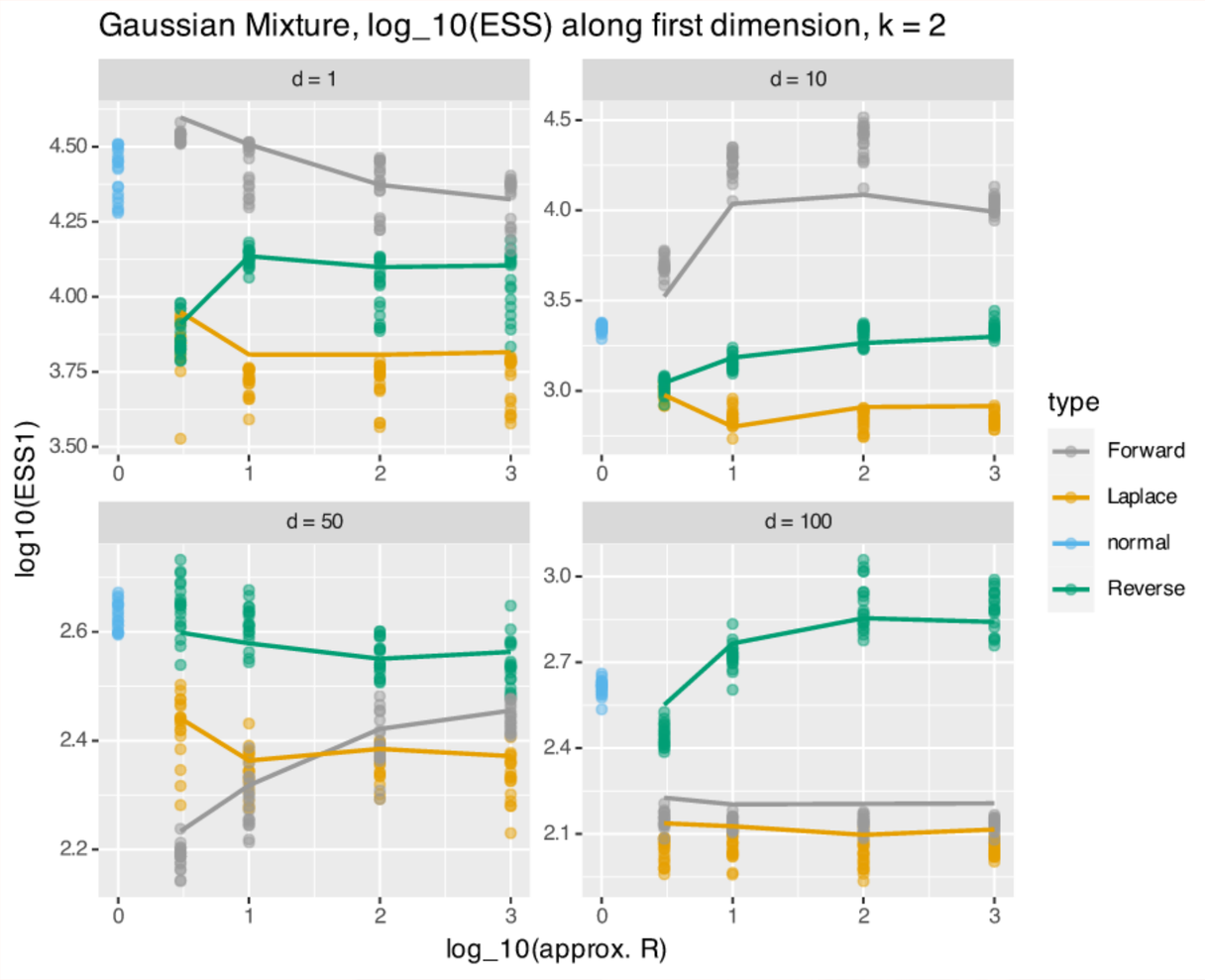
Experiments: Gaussian Mixture cont.

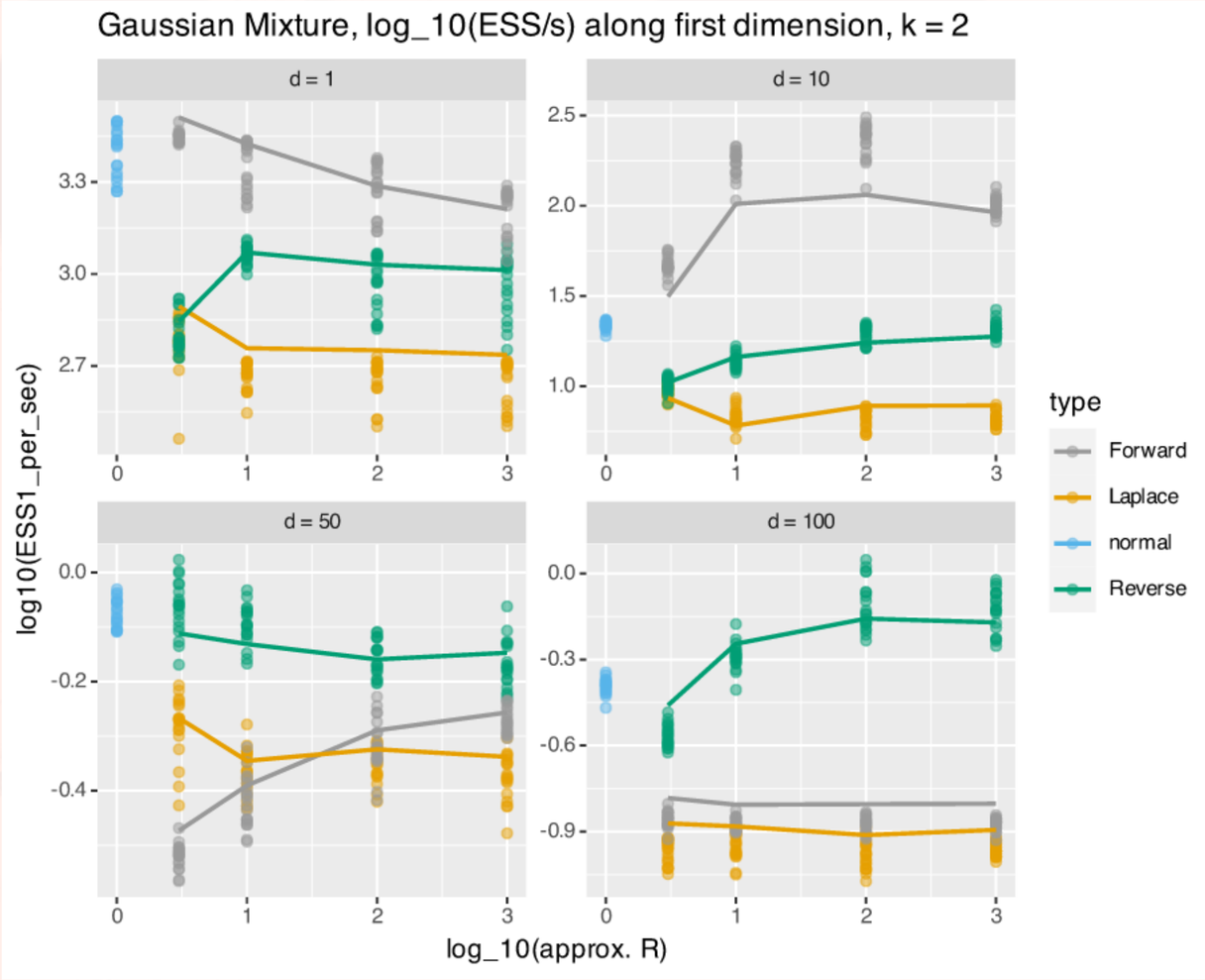


Experiments: Gaussian Mixture cont.

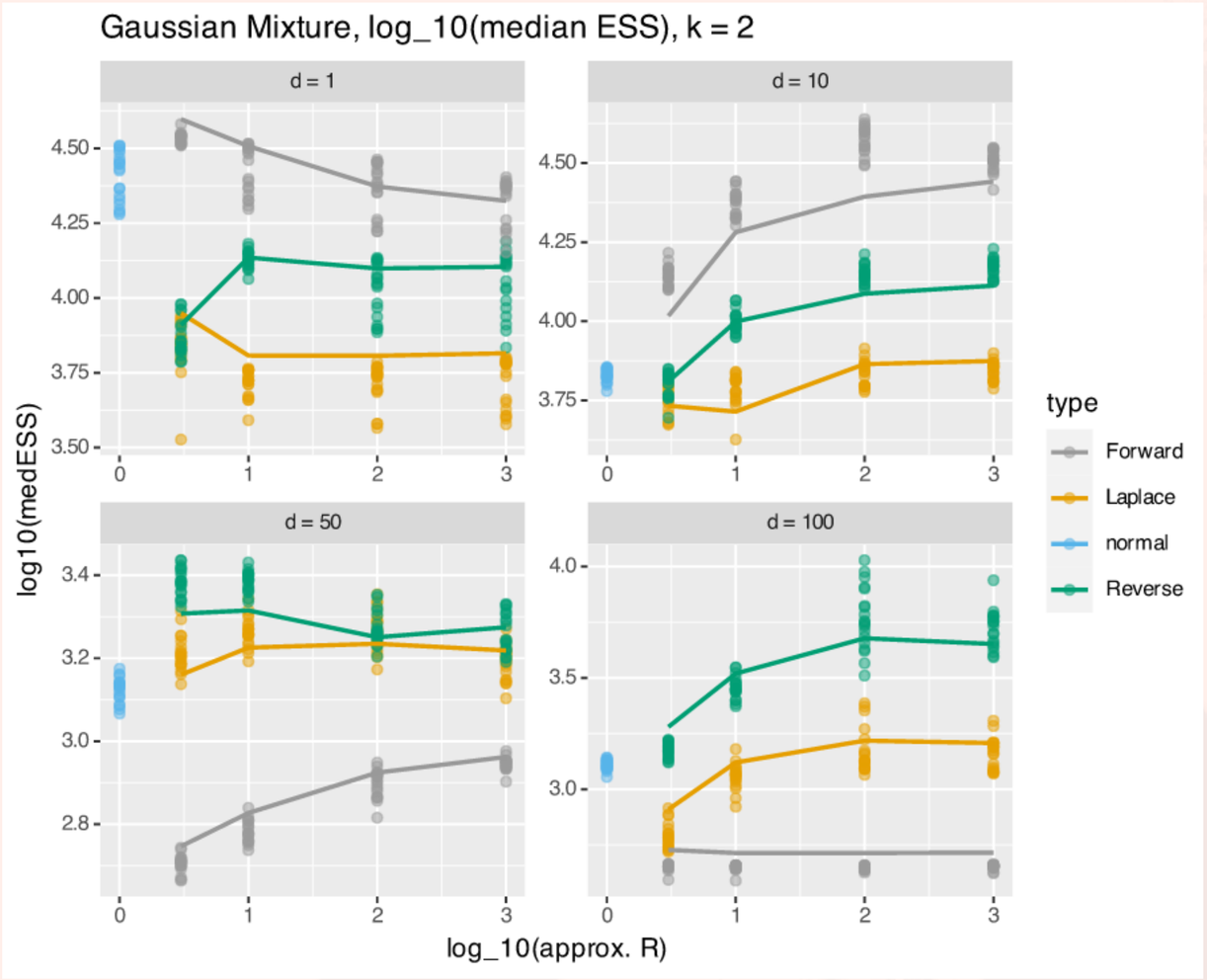


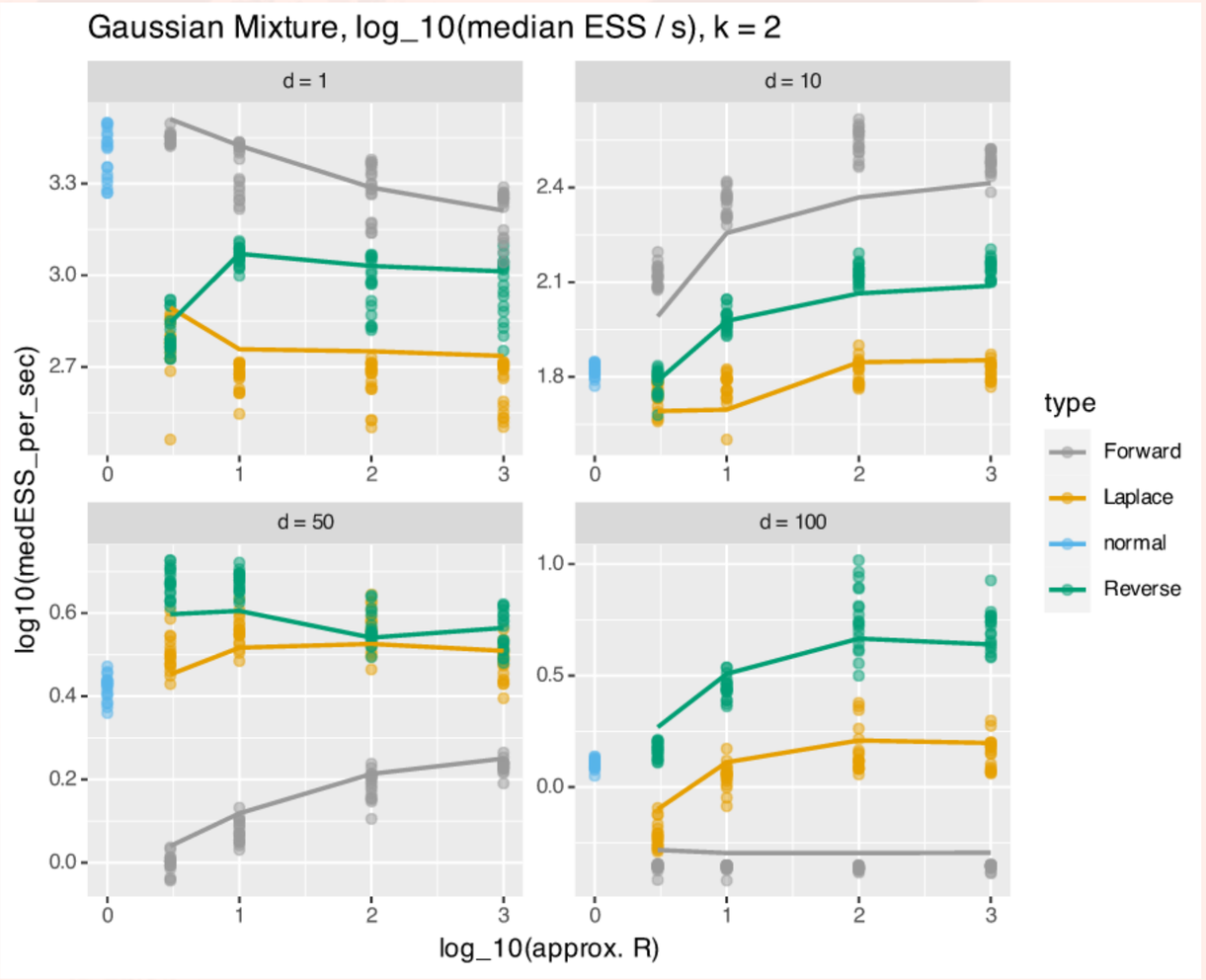
Experiments: Gaussian Mixture cont.





Experiments: Gaussian Mixture cont.





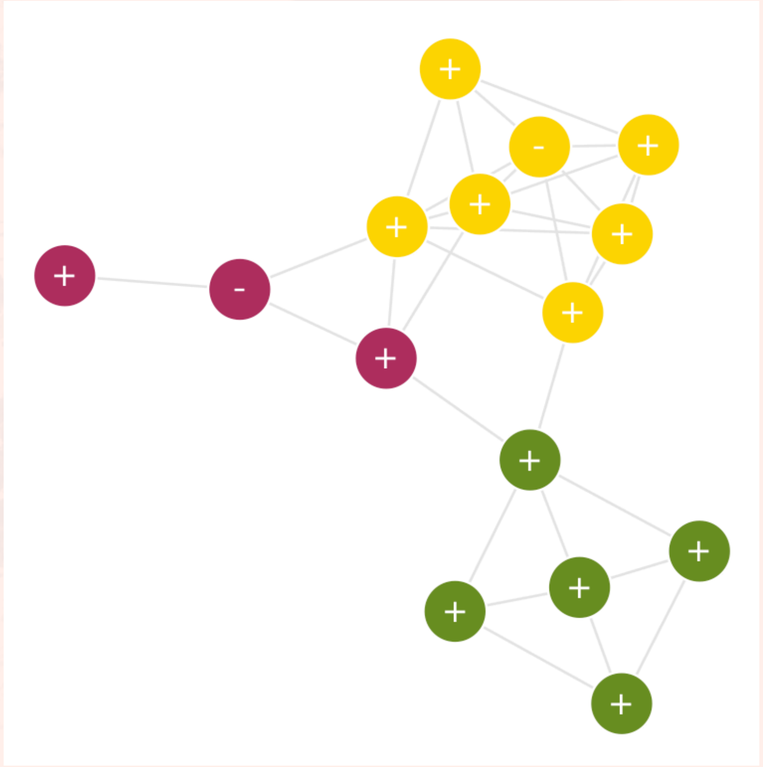
Experiments: Ising Model on Arbitrary Graph

Spins on a graph (V, E)

State: $\sigma \in \{-1, +1\}^N$ where $N = |V|$

Potential:

$$U(\sigma) := -\sum_{i=1}^{N} \sum_{j \in S_i} J_{ij} \sigma_i \sigma_j$$



Mass function:

$$\pi(A) := \sum_{\sigma \in A} \frac{\exp(-\beta U(\sigma))}{Z(\beta)}$$

where $\beta \in [0, \infty]$ is the inverse temperature

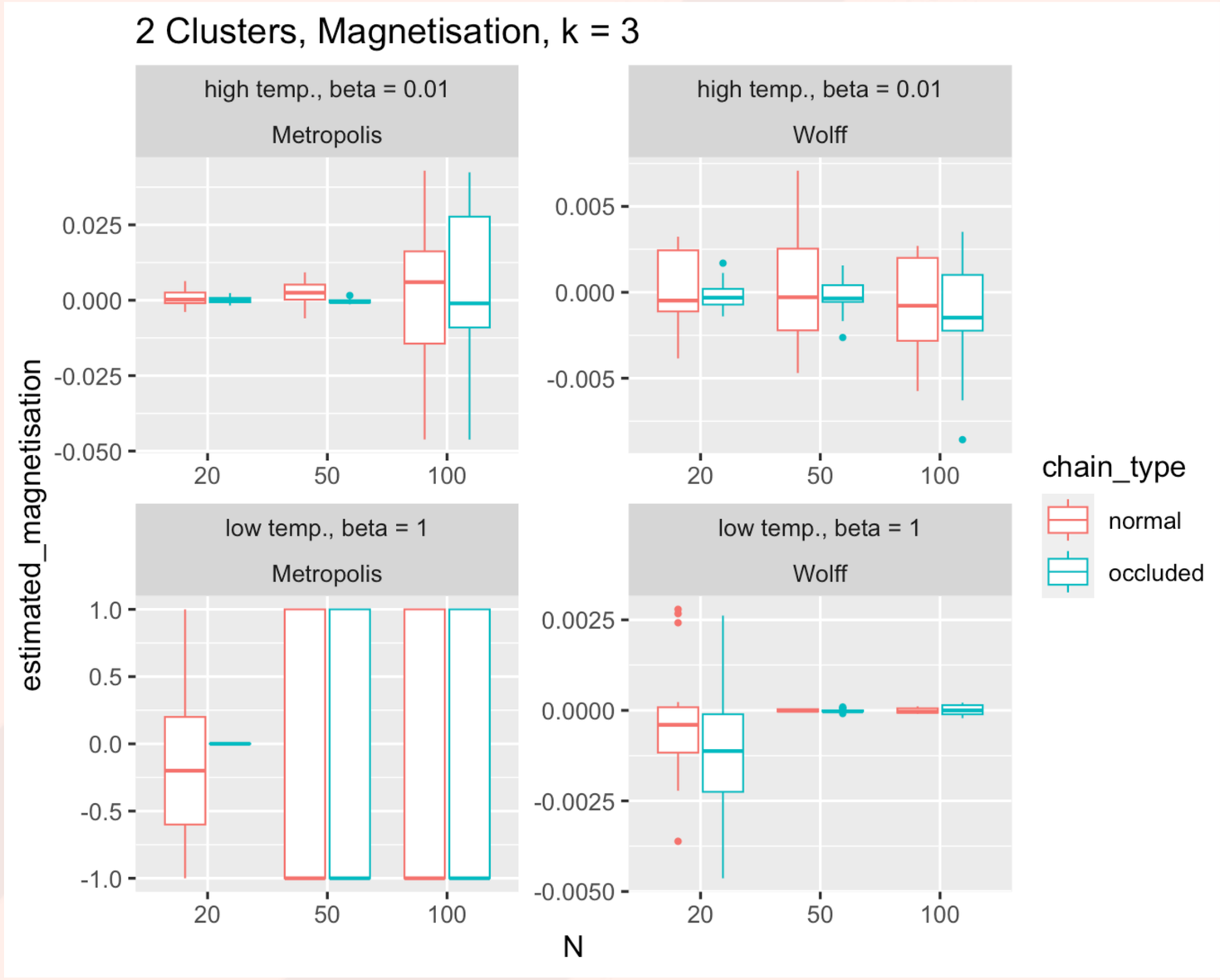
Physicists care about the average magnetisation: $f(\sigma) := \frac{1}{N} \sum_{i=1}^{N} \sigma_i$

$\pi(\sigma) = \pi(-\sigma)$ and $f(-\sigma) = -f(\sigma)$ so $\mathbb{E}_\pi[f(\sigma)] = 0$

Experiments: Ising Model on Arbitrary Graph cont.

- $N \in \{20, 50, 100\}$
- $\beta \in \{0.01, 100\}, J_{ij} = 1$ for all $i, j \in [N]$
- Regions:
  - Run a short MCMC chain and then split the histogram of the Radon-Nikodym's into equal bins
- MCMC algorithm ∈ {Metropolis, Wolff}
- $k = 3$

Experiments: Ising Model on Arbitrary Graph cont.



Summary

VI is fast but biased, MCMC is slow but asymptotically unbiased

However combining the two must take care to avoid the incongruity

We propose the occlusion process:

General purpose method that takes any MCMC and variational distribution

Leverages the exactness of the MCMC and the ability to sample easily from the variational distribution

Works by decorrelating states in the estimator