A robust and scalable approach to Bayesian doubly-intractable problems

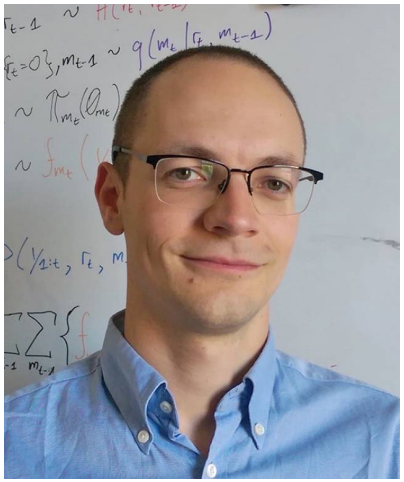
François-Xavier Briol
University College London & The Alan Turing Institute

Programme on Data-Driven Engineering,
Isaac Newton Institute.

Collaborators

Takuo Matsubara
(Newcastle)

Matias Altamirano
(UCL)



Jeremias Knoblauch
(UCL)

Chris Oates
(Newcastle)

The Setting

- Suppose we have access to some data assumed iid:

$$\{x_i\}_{i=1}^n \sim \mathbb{Q}$$

taking values on $\mathcal{X} \subseteq \mathbb{R}^d$ and consider a parametric model:

$$\{\mathbb{P}_\theta\}_{\theta \in \Theta}$$

where $\Theta \subseteq \mathbb{R}^p$ and p_θ is the density function or mass function of $\mathbb{P}_\theta$.

- Parameter estimation/Inference task:

Find θ^* such that $\mathbb{P}_{\theta^*} = \mathbb{Q}$ using the data $\{x_i\}_{i=1}^n$.

The Setting

- Suppose we have access to some data assumed iid:

$$\{x_i\}_{i=1}^n \sim \mathbb{Q}$$

taking values on $\mathcal{X} \subseteq \mathbb{R}^d$ and consider a parametric model:

$$\{\mathbb{P}_\theta\}_{\theta \in \Theta}$$

where $\Theta \subseteq \mathbb{R}^p$ and p_θ is the density function or mass function of $\mathbb{P}_\theta$.

- Parameter estimation/Inference task:

Find θ^* such that $\mathbb{P}_{\theta^*} = \mathbb{Q}$ using the data $\{x_i\}_{i=1}^n$.

Intractability in Bayesian Inference

Problem: Most Inference Methods are Likelihood-based...

- Most inference methods are likelihood-based. But what if our model is very complex and the likelihood (i.e. p_θ) is intractable?
- Example 1: Maximum Likelihood estimation:

$$\hat{\theta}^{\text{MLE}} := \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i).$$

- Example 2: Bayesian inference:

$$\pi(\theta | x_{1:n}) \quad \propto \quad \prod_{i=1}^{n} p_\theta(x_i) \pi(\theta).$$

[Quantities we can't evaluate will usually be in red!]

Problem: Most Inference Methods are Likelihood-based...

- Most inference methods are likelihood-based. But what if our model is very complex and the likelihood (i.e. p_θ) is intractable?
- Example 1: Maximum Likelihood estimation:

$$\hat{\theta}^{\text{MLE}} := \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i).$$

- Example 2: Bayesian inference:

$$\pi(\theta | x_{1:n}) \quad \propto \quad \prod_{i=1}^{n} p_\theta(x_i) \pi(\theta).$$

[Quantities we can't evaluate will usually be in red!]

Unnormalised Likelihood

- This often occurs because the likelihood is unnormalised, i.e.:

$$p_\theta(x) = \frac{\tilde{p}_\theta(x)}{C(\theta)} \propto \tilde{p}_\theta(x).$$

where we can easily evaluate $\tilde{p}_\theta$, but the normalisation constant:

$$C(\theta) = \int_{\mathcal{X}} \tilde{p}_\theta(x) dx \qquad \text{or} \qquad C(\theta) = \sum_{x \in \mathcal{X}} \tilde{p}_\theta(x).$$

is computationally intractable.

- Applications: deep energy models in machine learning, graphical models, nonparametric density models, lattice models in statistical physics or spatial statistics, ...

Unnormalised Likelihood

- This often occurs because the likelihood is unnormalised, i.e.:

$$p_\theta(x) = \frac{\tilde{p}_\theta(x)}{C(\theta)} \propto \tilde{p}_\theta(x).$$

where we can easily evaluate $\tilde{p}_\theta$, but the normalisation constant:

$$C(\theta) = \int_{\mathcal{X}} \tilde{p}_\theta(x) dx \qquad \text{or} \qquad C(\theta) = \sum_{x \in \mathcal{X}} \tilde{p}_\theta(x).$$

is computationally intractable.

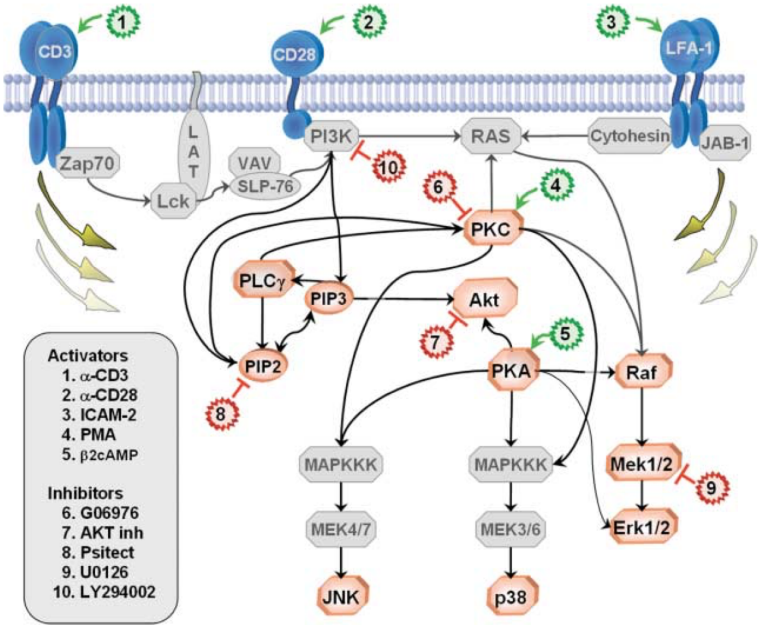
- Applications: deep energy models in machine learning, graphical models, nonparametric density models, lattice models in statistical physics or spatial statistics, ...

Biochemistry: Protein Signalling Networks

Graphical models representing interactions between proteins:

$$\tilde{p}_\theta(x) = \exp\Big(-\sum_{j=1}^{d} \theta_{(j)} x_{(j)} + \sum_{j<k} \theta_{(j,k)} x_{(j)} x_{(k)} \Big)$$

Both Θ and $\mathcal{X} \subseteq \mathbb{R}^d$ can be high-dimensional.



(Picture from [Sachs et al., 2005])

[1] Sachs et al. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. Science, 308(5721), 523–529.

[2] Yu et al. (2016). Statistical inference for pairwise graphical models using score matching. In Neural Information Processing Systems.

Bayesian Doubly-Intractable Problems

- Recall that $p_\theta(x) = \frac{\tilde{p}_\theta(x)}{C(\theta)}$. Then, the Bayesian posterior becomes:

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \left(\prod_{i=1}^{n} \frac{\tilde{p}_\theta(x_i)}{C(\theta)} \right) \pi(\theta).$$

where

$$C_{\text{post}} = \int_\Theta \left(\prod_{i=1}^{n} \frac{\tilde{p}_\theta(x_i)}{C(\theta)} \right) \pi(\theta) d\theta$$

- The problem is known as doubly-intractable because it contains two intractable normalisation constants: $C(\theta)$ and C_{post}.

Bayesian Doubly-Intractable Problems

- Recall that $p_\theta(x) = \frac{\tilde{p}_\theta(x)}{C(\theta)}$. Then, the Bayesian posterior becomes:

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \left(\prod_{i=1}^{n} \frac{\tilde{p}_\theta(x_i)}{C(\theta)} \right) \pi(\theta).$$

where

$$C_{\text{post}} = \int_\Theta \left(\prod_{i=1}^{n} \frac{\tilde{p}_\theta(x_i)}{C(\theta)} \right) \pi(\theta) d\theta$$

- The problem is known as doubly-intractable because it contains two intractable normalisation constants: $C(\theta)$ and C_{post}.

Intractable Constants in Bayesian Statistics

- Intractable constants? So what? In the Bayesian world, we are used to this, and can use MCMC steps with ratios to remove the issue.
- Suppose we have a standard Bayesian posterior

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta).$$

Suppose also that our MCMC method proposes θ^* and we are at θ, then the Metropolis-Hastings acceptance probability is:

$$\begin{aligned}
&\min\left(1, \frac{\pi(\theta^*|x_{1:n})P(\theta^*,\theta)}{\pi(\theta|x_{1:n})P(\theta,\theta^*)}\right) \\
&= \min\left(1, \frac{\cancel{C_{\text{post}}} \prod_{i=1}^{n} p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\cancel{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right) \\
&= \min\left(1, \frac{\prod_{i=1}^{n} p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right)
\end{aligned}$$

Intractable Constants in Bayesian Statistics

- Intractable constants? So what? In the Bayesian world, we are used to this, and can use MCMC steps with ratios to remove the issue.
- Suppose we have a standard Bayesian posterior

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta).$$

Suppose also that our MCMC method proposes θ^* and we are at θ, then the Metropolis-Hastings acceptance probability is:

$$\min\left(1, \frac{\pi(\theta^*|x_{1:n})P(\theta^*,\theta)}{\pi(\theta|x_{1:n})P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{C_{\text{post}} \prod_{i=1}^{n} p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{C_{\text{post}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{\prod_{i=1}^{n} p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right)$$

Intractable Constants in Bayesian Statistics

- Intractable constants? So what? In the Bayesian world, we are used to this, and can use MCMC steps with ratios to remove the issue.
- Suppose we have a standard Bayesian posterior

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta).$$

Suppose also that our MCMC method proposes θ^* and we are at θ, then the Metropolis-Hastings acceptance probability is:

$$\min\left(1, \frac{\pi(\theta^*|x_{1:n})P(\theta^*,\theta)}{\pi(\theta|x_{1:n})P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{\cancel{C_{\text{post}}} \prod_{i=1}^{n} p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\cancel{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{\prod_{i=1}^{n} p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right)$$

Intractable Constants in Bayesian Statistics

- Intractable constants? So what? In the Bayesian world, we are used to this, and can use MCMC steps with ratios to remove the issue.
- Suppose we have a standard Bayesian posterior

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta).$$

Suppose also that our MCMC method proposes θ^* and we are at θ, then the Metropolis-Hastings acceptance probability is:

$$\begin{aligned} &\min\left(1, \frac{\pi(\theta^*|x_{1:n})P(\theta^*,\theta)}{\pi(\theta|x_{1:n})P(\theta,\theta^*)}\right) \\ = &\min\left(1, \frac{\cancel{C_{\text{post}}} \prod_{i=1}^{n} p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\cancel{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right) \\ = &\min\left(1, \frac{\prod_{i=1}^{n} p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\prod_{i=1}^{n} p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right) \end{aligned}$$

The Issue with Double Intractability

- But what if we are in the doubly-intractable scenario?

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} \frac{1}{C(\theta)} \tilde{p}_\theta(x_i)\pi(\theta).$$

- The main issue is that our second intractable constant depends on θ:

$$\min\left(1, \frac{\pi(\theta^*|x_{1:n})P(\theta^*,\theta)}{\pi(\theta|x_{1:n})P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{\cancel{C_{\text{post}}} \prod_{i=1}^{n} C(\theta)\tilde{p}_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\cancel{C_{\text{post}}} \prod_{i=1}^{n} C(\theta^*)\tilde{p}_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{\prod_{i=1}^{n} C(\theta)p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\prod_{i=1}^{n} C(\theta^*)p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right).$$

- This is still intractable!

The Issue with Double Intractability

- But what if we are in the doubly-intractable scenario?

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} \frac{1}{C(\theta)} \tilde{p}_\theta(x_i)\pi(\theta).$$

- The main issue is that our second intractable constant depends on θ:

$$\min\left(1, \frac{\pi(\theta^*|x_{1:n})P(\theta^*,\theta)}{\pi(\theta|x_{1:n})P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{C_{\text{post}} \prod_{i=1}^{n} C(\theta)\tilde{p}_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{C_{\text{post}} \prod_{i=1}^{n} C(\theta^*)\tilde{p}_{\theta}(x_i)\pi(\theta)P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{\prod_{i=1}^{n} C(\theta)p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\prod_{i=1}^{n} C(\theta^*)p_{\theta}(x_i)\pi(\theta)P(\theta,\theta^*)}\right).$$

- This is still intractable!

The Issue with Double Intractability

- But what if we are in the doubly-intractable scenario?

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} \frac{1}{C(\theta)} \tilde{p}_\theta(x_i)\pi(\theta).$$

- The main issue is that our second intractable constant depends on θ:

$$\min\left(1, \frac{\pi(\theta^*|x_{1:n})P(\theta^*,\theta)}{\pi(\theta|x_{1:n})P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{\cancel{C_{\text{post}}} \prod_{i=1}^{n} C(\theta)\tilde{p}_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\cancel{C_{\text{post}}} \prod_{i=1}^{n} C(\theta^*)\tilde{p}_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right)$$

$$= \min\left(1, \frac{\prod_{i=1}^{n} C(\theta)p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\prod_{i=1}^{n} C(\theta^*)p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right).$$

- This is still intractable!

The Issue with Double Intractability

- But what if we are in the doubly-intractable scenario?

$$\pi(\theta|x_{1:n}) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta) = \frac{1}{C_{\text{post}}} \prod_{i=1}^{n} \frac{1}{C(\theta)} \tilde{p}_\theta(x_i)\pi(\theta).$$

- The main issue is that our second intractable constant depends on θ:

$$\begin{aligned}
&\min\left(1, \frac{\pi(\theta^*|x_{1:n})P(\theta^*,\theta)}{\pi(\theta|x_{1:n})P(\theta,\theta^*)}\right) \\
&= \min\left(1, \frac{\cancel{C_{\text{post}}} \prod_{i=1}^{n} C(\theta)\tilde{p}_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\cancel{C_{\text{post}}} \prod_{i=1}^{n} C(\theta^*)\tilde{p}_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right) \\
&= \min\left(1, \frac{\prod_{i=1}^{n} C(\theta)p_{\theta^*}(x_i)\pi(\theta^*)P(\theta^*,\theta)}{\prod_{i=1}^{n} C(\theta^*)p_\theta(x_i)\pi(\theta)P(\theta,\theta^*)}\right).
\end{aligned}$$

- This is still intractable!

What Can We Do?

- **Approximate likelihoods:** approximate likelihood with a tractable function. Examples incl. pseudo-likelihood and composite likelihoods.

  Issue: our model is now misspecified...

- **Advanced MCMC schemes:** based on unbiased estimate of $C(\theta)$. Examples include the "Russian roulette" and the exchange algorithm.

  Issue: these are fiddly and computationally expensive...

- **Simulation-based inference:** including Bayesian synthetic likelihoods or approximate Bayesian computation.

  Issue: Approximate method and computationally expensive...

What Can We Do?

- **Approximate likelihoods:** approximate likelihood with a tractable function. Examples incl. pseudo-likelihood and composite likelihoods.

  Issue: our model is now misspecified...

- **Advanced MCMC schemes:** based on unbiased estimate of $C(\theta)$. Examples include the "Russian roulette" and the exchange algorithm.

  Issue: these are fiddly and computationally expensive...

- **Simulation-based inference:** including Bayesian synthetic likelihoods or approximate Bayesian computation.

  Issue: Approximate method and computationally expensive...

What Can We Do?

- **Approximate likelihoods:** approximate likelihood with a tractable function. Examples incl. pseudo-likelihood and composite likelihoods.

  Issue: our model is now misspecified...

- **Advanced MCMC schemes:** based on unbiased estimate of $C(\theta)$. Examples include the "Russian roulette" and the exchange algorithm.

  Issue: these are fiddly and computationally expensive...

- **Simulation-based inference:** including Bayesian synthetic likelihoods or approximate Bayesian computation.

  Issue: Approximate method and computationally expensive...

Misspecification in Bayesian Inference

Most Intractable Models are Misspecified?

- Most intractable models are large-scale models of complex phenomenon. Scientists/engineers can do their best to make these as accurate as possible, they will most likely often be misspecified:

$$\nexists\, \theta^* \in \Theta \quad \text{such that} \quad \mathbb{P}_{\theta^*} = \mathbb{Q}$$

- **Q:** What is the impact of misspecification for these models? Can we do anything to prevent significant issues?

Most Intractable Models are Misspecified?

- Most intractable models are large-scale models of complex phenomenon. Scientists/engineers can do their best to make these as accurate as possible, they will most likely often be misspecified:

$$\nexists\, \theta^* \in \Theta \quad \text{such that} \quad \mathbb{P}_{\theta^*} = \mathbb{Q}$$

- **Q:** What is the impact of misspecification for these models? Can we do anything to prevent significant issues?

Bayesian Methods Struggle Under Misspecification

- Perhaps we still have a model which is decent but not perfect?

$$\exists\, \theta^* \in \Theta \quad \text{such that} \quad \mathbb{P}_{\theta^*} \approx \mathbb{Q}$$

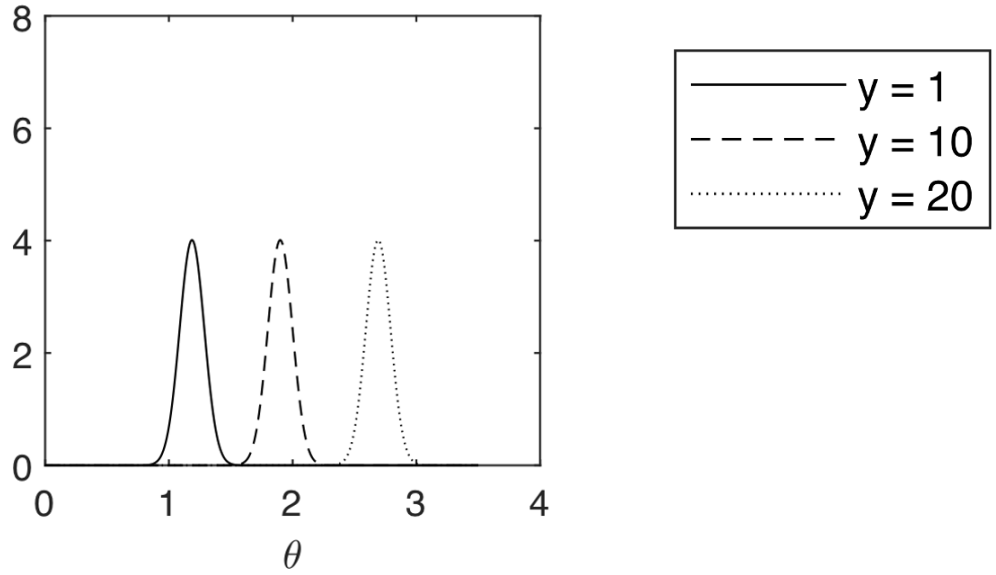
- This can create a lot of challenges for Bayes since it uses the likelihood to update beliefs!

Bayesian Methods Struggle Under Misspecification

- Perhaps we still have a model which is decent but not perfect?

$$\exists\, \theta^* \in \Theta \quad \text{such that} \quad \mathbb{P}_{\theta^*} \approx \mathbb{Q}$$

- This can create a lot of challenges for Bayes since it uses the likelihood to update beliefs!
- Toy example: let $n = 100$, the data is $N(1, 1)$ but 10% of the data points are fixed at y. We want to estimate θ for $\mathbb{P}_\theta = N(\theta, 1)$.



Objectives

The objective for this talk will be to develop Bayesian methods for doubly-intractable problem which are scalable and robust.

(1) Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. Journal of the Royal Statistical Society B: Statistical Methodology, 84(3), 997–1022.

(2) Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2022). Generalised Bayesian inference for discrete intractable likelihood. arXiv:2206.08420. (under revisions at JASA)

(3) Altamirano, M., Briol, F.-X., & Knoblauch, J. (2023). Robust and scalable Bayesian online changepoint detection. arXiv:2302.04759. (under review at ICML)

Generalised Bayesian Inference for Doubly-Intractable Problems

Inference with Discrepancies

- Let $\mathbb{Q}^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and assume you have a discrepancy D which gives the "distance" between probability measures $\mathbb{P}, \mathbb{Q}$.
- You may have seen discrepancies used for:
  1. Minimum distance estimators (MDE):

$$\hat{\theta}_n^D = \arg\inf_{\theta \in \Theta} D(\mathbb{P}_\theta, \mathbb{Q}^n).$$

  2. Approximate Bayesian computation: repeatedly sample $\tilde{\theta}$ from a prior $p(\theta)$, then accept if $D(\mathbb{P}_{\tilde{\theta}}, \mathbb{Q}^n) < \varepsilon$ for some threshold $\varepsilon > 0$.

Inference with Discrepancies

- Let $\mathbb{Q}^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and assume you have a discrepancy D which gives the "distance" between probability measures $\mathbb{P}, \mathbb{Q}$.
- You may have seen discrepancies used for:
  1. Minimum distance estimators (MDE):

$$\hat{\theta}_n^D = \arg\inf_{\theta \in \Theta} D(\mathbb{P}_\theta, \mathbb{Q}^n).$$

  2. Approximate Bayesian computation: repeatedly sample $\tilde{\theta}$ from a prior $p(\theta)$, then accept if $D(\mathbb{P}_{\tilde{\theta}}, \mathbb{Q}^n) < \varepsilon$ for some threshold $\varepsilon > 0$.

Inference with Discrepancies

- Let $\mathbb{Q}^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and assume you have a discrepancy D which gives the "distance" between probability measures $\mathbb{P}, \mathbb{Q}$.
- You may have seen discrepancies used for:
  1. Minimum distance estimators (MDE):

$$\hat{\theta}_n^D = \arg\inf_{\theta \in \Theta} D(\mathbb{P}_\theta, \mathbb{Q}^n).$$

  2. Approximate Bayesian computation: repeatedly sample $\tilde{\theta}$ from a prior $p(\theta)$, then accept if $D(\mathbb{P}_{\tilde{\theta}}, \mathbb{Q}^n) < \varepsilon$ for some threshold $\varepsilon > 0$.

Generalised Bayesian Inference

- Generalised Bayesian inference: Given a prior $\pi(\theta)$ and some $\beta > 0$, we can construct a generalised posterior as:

$$\pi(\theta|x_{1:n}) \propto \prod_{i=1}^{n} p_\theta(x_i)\, \pi(\theta).$$

$$\pi^D(\theta|x_{1:n}) \propto \exp(-\beta D(\mathbb{P}_\theta, \mathbb{Q}^n))\, \pi(\theta).$$

Bissiri, P., Holmes, C., & Walker, S. (2016). A general framework for updating belief distributions. J. Royal Stat. Soc. Series B, 78, 1103–1130.

Knoblauch, J., Jewson, J., & Damoulas, T. (2022). An optimization-centric view on Bayes' rule: reviewing and generalizing variational inference. JMLR, 23(132), 1–109.

Generalised Bayesian Inference

$$\pi^D(\theta|x_{1:n}) \propto \exp(-\beta D(\mathbb{P}_\theta, \mathbb{Q}^n))\, \pi(\theta).$$

- Clearly $D(\mathbb{P}_\theta, \mathbb{Q}^n)$ encourages our posterior to concentrate on regions where the data agrees with the model.
- Lots of questions regarding interpretation of π^D that I will ignore...
- **Q:** How do you select a discrepancy to get scalability and robustness?

Case 1: Continuous Data

Bayesian Inference with Score Functions

- We ideally want to create discrepancy such that the normalisation constant of $p_\theta(x) = \tilde{p}_\theta(x)/C(\theta)$ does not need to be computed.
- Suppose that the data is continuous; then notice that:

$$\begin{aligned}\nabla_x \log p_\theta(x) &= \nabla_x \log \frac{\tilde{p}_\theta(x)}{C(\theta)} \\ &= \nabla_x \log \tilde{p}_\theta(x) - \cancel{\nabla_x \log C(\theta)} \\ &= \nabla_x \log \tilde{p}_\theta(x).\end{aligned}$$

This quantity is known as the score function, and we can simply construct a discrepancy which compares scores.

Bayesian Inference with Score Functions

- We ideally want to create discrepancy such that the normalisation constant of $p_\theta(x) = \tilde{p}_\theta(x)/C(\theta)$ does not need to be computed.
- Suppose that the data is continuous; then notice that:

$$\begin{aligned}\nabla_x \log p_\theta(x) &= \nabla_x \log \frac{\tilde{p}_\theta(x)}{C(\theta)} \\ &= \nabla_x \log \tilde{p}_\theta(x) - \cancel{\nabla_x \log C(\theta)} \\ &= \nabla_x \log \tilde{p}_\theta(x).\end{aligned}$$

This quantity is known as the score function, and we can simply construct a discrepancy which compares scores.

Bayesian Inference with Score Functions

- Two natural options: the kernel Stein discrepancy and the score-matching divergence. They are both of the from:

$$D(\mathbb{P}, \mathbb{Q}) = d(\nabla_x \log p, \nabla_x \log q)$$

for some appropriate distance d.

[1] Hyvärinen, A. (2006). Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6, 695–708.

[2] Chwialkowski et. al. (2016). A kernel test of goodness of fit. International Conference on Machine Learning, 2606–2615.

The Hyvarinen/Fisher/Score-matching Discrepancy

- The (diffusion) score-matching (SM) divergence is of the form:

$$\begin{aligned}\mathrm{SM}(\mathbb{P}_\theta, \mathbb{Q}) &= \int_{\mathcal{X}} \|m^\top(x)(\nabla_x \log p_\theta(x) - \nabla_x \log q(x))\|_2^2 \mathbb{Q}(dx) \\ &= \int_{\mathcal{X}} \|m^\top(x)(\nabla_x \log p_\theta(x)\|_2^2 \\ &\qquad + 2\nabla_x \cdot (m(x)m^\top(x)\nabla_x \log p_\theta(x))\mathbb{Q}(dx) + C_q\end{aligned}$$

Except: (i) $\nabla_x \log p_\theta(x) = \nabla_x \log \tilde{p}_\theta(x)$, and

(ii) C_q is independent of θ!

- The case $m(x) = I_d$ is the original case studied by Hyvarinen. The generalisation is crucial for robustness.

Barp et al. (incl. FXB) (2019). Minimum Stein discrepancy estimators. Neural Information Processing Systems, 12964–12976.

The Hyvarinen/Fisher/Score-matching Discrepancy

- The (diffusion) score-matching (SM) divergence is of the form:

$$\begin{aligned}\mathrm{SM}(\mathbb{P}_\theta, \mathbb{Q}) &= \int_{\mathcal{X}} \|m^\top(x)(\nabla_x \log p_\theta(x) - \nabla_x \log q(x))\|_2^2 \mathbb{Q}(dx) \\ &= \int_{\mathcal{X}} \|m^\top(x)(\nabla_x \log p_\theta(x)\|_2^2 \\ &\qquad + 2\nabla_x \cdot (m(x)m^\top(x)\nabla_x \log p_\theta(x))\mathbb{Q}(dx) + C_q\end{aligned}$$

Except: (i) $\nabla_x \log p_\theta(x) = \nabla_x \log \tilde{p}_\theta(x)$, and

(ii) C_q is independent of θ!

- The case $m(x) = I_d$ is the original case studied by Hyvarinen. The generalisation is crucial for robustness.

Barp et al. (incl. FXB) (2019). Minimum Stein discrepancy estimators. Neural Information Processing Systems, 12964–12976.

The Kernel Stein Discrepancy

- Consider a kernel k (e.g. $k(x,y) = \exp(-(x-y)^2/l^2)$). The kernel Stein discrepancy (KSD) is a divergence of the form:

$$\begin{aligned}
\mathrm{KSD}^2(\mathbb{P}_\theta, \mathbb{Q}) &= \int_{\mathcal{X}} \int_{\mathcal{X}} k_0(x,x') \mathbb{Q}(dx) \mathbb{Q}(dx') \\
k_0(x,y) &= \langle \nabla_x, \nabla_y k(x,y) \rangle + \langle \nabla_x k(x,y), \nabla_y \log p_\theta(y) \rangle \\
&\quad + \langle \nabla_y k(x,y), \nabla_x \log p_\theta(x) \rangle \\
&\quad + k(x,y) \langle \nabla_x \log p_\theta(x), \nabla_y \log p_\theta(y) \rangle .
\end{aligned}$$

Anastasiou et al. (incl. FXB) (2022). Stein's method meets computational statistics: A review of some recent developments. Statistical Science, 38 (1), 120-139.

SM-Bayes and KSD-Bayes

$$\pi^{\mathsf{KSD}}(\theta|x_{1:n}) \propto \exp(-\beta \mathsf{KSD}^2(\mathbb{P}_\theta, \mathbb{Q}^n))\pi(\theta)$$
$$\pi^{\mathsf{SM}}(\theta|x_{1:n}) \propto \exp(-\beta \mathsf{SM}(\mathbb{P}_\theta, \mathbb{Q}^n))\pi(\theta).$$

- SM and KSD do not require $C(\theta)$! We are back to standard MCMC...
- Under relatively mild conditions, we can also show that they are consistent and satisfies a Bernstein-von-Mises theorem.
- **Q:** Why did we bother with two discrepancies?
  A: They are broadly applicable in different settings. SM has cost $\mathcal{O}(nd^2)$ whereas KSD has cost $\mathcal{O}(n^2 d)$.

SM-Bayes and KSD-Bayes

$$\pi^{\mathrm{KSD}}(\theta|x_{1:n}) \propto \exp(-\beta \mathrm{KSD}^2(\mathbb{P}_\theta, \mathbb{Q}^n))\pi(\theta)$$
$$\pi^{\mathrm{SM}}(\theta|x_{1:n}) \propto \exp(-\beta \mathrm{SM}(\mathbb{P}_\theta, \mathbb{Q}^n))\pi(\theta).$$

- SM and KSD do not require $C(\theta)$! We are back to standard MCMC...
- Under relatively mild conditions, we can also show that they are consistent and satisfies a Bernstein-von-Mises theorem.
- **Q:** Why did we bother with two discrepancies?
  A: They are broadly applicable in different settings. SM has cost $\mathcal{O}(nd^2)$ whereas KSD has cost $\mathcal{O}(n^2 d)$.

SM-Bayes and KSD-Bayes

$$\pi^{\mathsf{KSD}}(\theta|x_{1:n}) \propto \exp(-\beta \mathsf{KSD}^2(\mathbb{P}_\theta, \mathbb{Q}^n))\pi(\theta)$$
$$\pi^{\mathsf{SM}}(\theta|x_{1:n}) \propto \exp(-\beta \mathsf{SM}(\mathbb{P}_\theta, \mathbb{Q}^n))\pi(\theta).$$

- SM and KSD do not require $C(\theta)$! We are back to standard MCMC...
- Under relatively mild conditions, we can also show that they are consistent and satisfies a Bernstein-von-Mises theorem.
- **Q:** Why did we bother with two discrepancies?
  A: They are broadly applicable in different settings. SM has cost $\mathcal{O}(nd^2)$ whereas KSD has cost $\mathcal{O}(n^2 d)$.

SM-Bayes and KSD-Bayes

$$\pi^{\mathsf{KSD}}(\theta|x_{1:n}) \propto \exp(-\beta \mathsf{KSD}^2(\mathbb{P}_\theta, \mathbb{Q}^n))\pi(\theta)$$
$$\pi^{\mathsf{SM}}(\theta|x_{1:n}) \propto \exp(-\beta \mathsf{SM}(\mathbb{P}_\theta, \mathbb{Q}^n))\pi(\theta).$$

- SM and KSD do not require $C(\theta)$! We are back to standard MCMC...
- Under relatively mild conditions, we can also show that they are consistent and satisfies a Bernstein-von-Mises theorem.
- **Q:** Why did we bother with two discrepancies?
  A: They are broadly applicable in different settings. SM has cost $\mathcal{O}(nd^2)$ whereas KSD has cost $\mathcal{O}(n^2 d)$.

Conjugate Bayesian Inference for Intractable Models

- Much more importantly, in the case of a Gaussian prior: $\pi(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$, and of an exponential family model:

$$p_\theta(x) = \frac{\exp(\eta(\theta)^\top r(x) + b(x))}{C(\theta)}$$

  then the generalised Bayesian posterior is also Gaussian:

$$\pi^{\mathrm{KSD}}(\theta|x_{1:n}) = \mathcal{N}(\theta; \mu_{\mathrm{KSD}}, \Sigma_{\mathrm{KSD}}), \qquad \pi^{\mathrm{SM}}(\theta|x_{1:n}) = \mathcal{N}(\theta; \mu_{\mathrm{SM}}, \Sigma_{\mathrm{SM}})$$

- The expressions are ugly, but they are all tractable and depend on $x_{1:n}, \tilde{p}_\theta, \mu, \Sigma$ and either k or m. Crucially, they don't depend on $C(\theta)$.
- This is by far the most important result of the whole talk!!

Conjugate Bayesian Inference for Intractable Models

- Much more importantly, in the case of a Gaussian prior: $\pi(\theta) = \mathcal{N}(\theta; \mu, \Sigma)$, and of an exponential family model:

$$p_\theta(x) = \frac{\exp(\eta(\theta)^\top r(x) + b(x))}{C(\theta)}$$

  then the generalised Bayesian posterior is also Gaussian:

$$\pi^{\mathrm{KSD}}(\theta|x_{1:n}) = \mathcal{N}(\theta; \mu_{\mathrm{KSD}}, \Sigma_{\mathrm{KSD}}), \qquad \pi^{\mathrm{SM}}(\theta|x_{1:n}) = \mathcal{N}(\theta; \mu_{\mathrm{SM}}, \Sigma_{\mathrm{SM}})$$

- The expressions are ugly, but they are all tractable and depend on $x_{1:n}, \tilde{p}_\theta, \mu, \Sigma$ and either k or m. Crucially, they don't depend on $C(\theta)$.
- This is by far the most important result of the whole talk!!

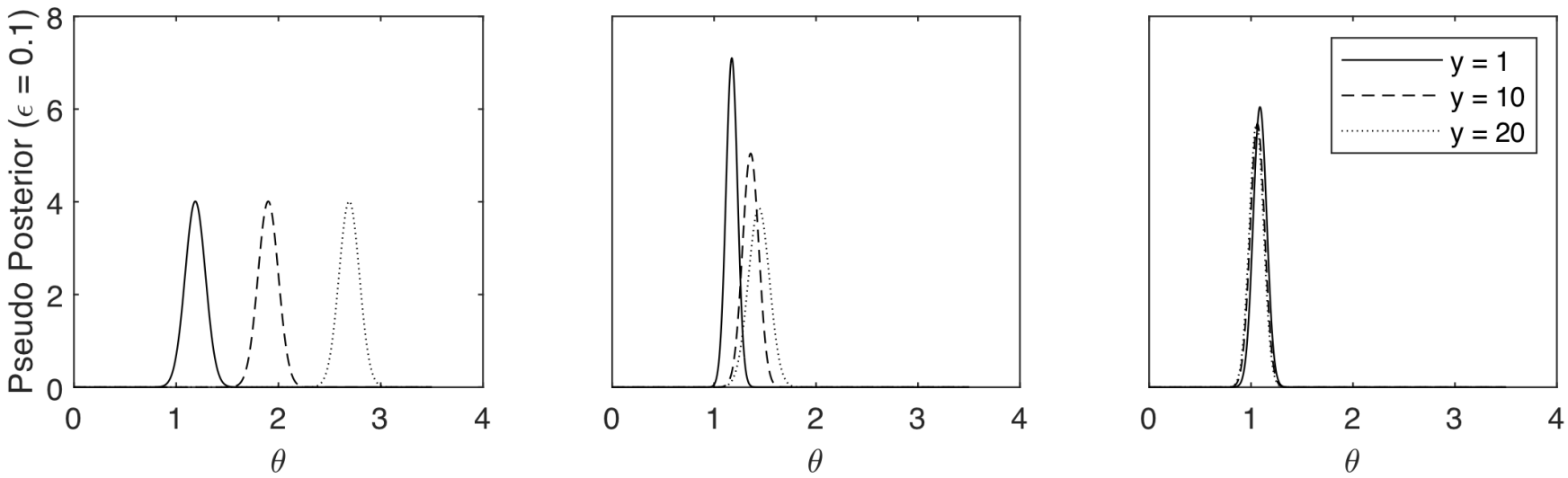
Robust Bayesian Inference

- We can formally show that k and m can induce robustness to mild model misspecification.
- Back to our Gaussian model, we consider KSD-Bayes:

($n = 100$ points, most are $N(1, 1)$ but 10% fixed at y. $\mathbb{P}_\theta = N(\theta, 1)$.)

Robust Bayesian Inference

- We can formally show that k and m can induce robustness to mild model misspecification.
- Back to our Gaussian model, we consider KSD-Bayes:



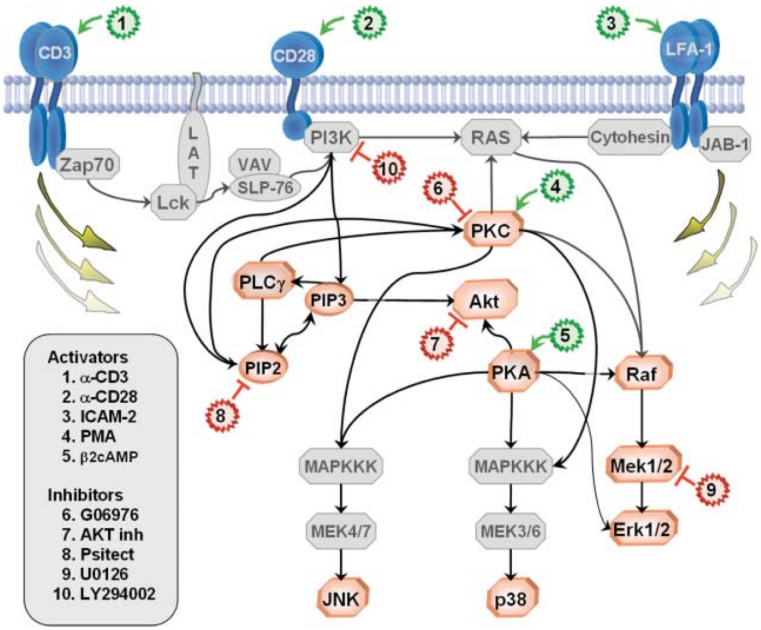
($n = 100$ points, most are $N(1, 1)$ but 10% fixed at y. $\mathbb{P}_\theta = N(\theta, 1)$.)

Biochemistry: Protein Signalling Networks

Graphical model representing interactions between proteins:

$$\tilde{p}_\theta(x) = \exp\Big(-\sum_{j=1}^{d} \theta_{(j)} x_{(j)} + \sum_{j<k} \theta_{(j,k)} x_{(j)} x_{(k)} \Big)$$

Both Θ and $\mathcal{X} \subseteq \mathbb{R}^d$ can be high-dimensional.

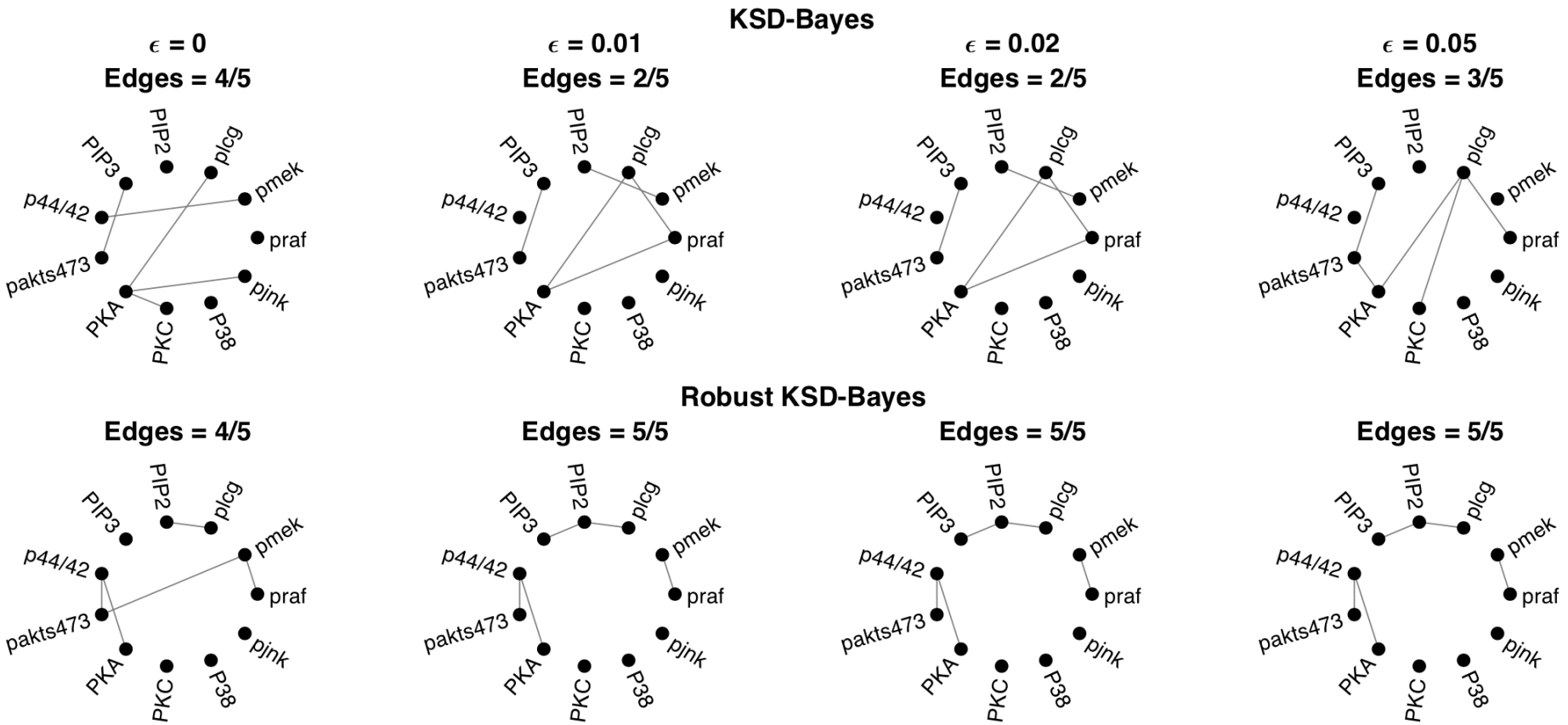


(Picture from [Sachs et al., 2005])

[1] Sachs et al. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. Science, 308(5721), 523–529.

[2] Yu et al. (2016). Statistical inference for pairwise graphical models using score matching. In Neural Information Processing Systems.

Inference at Scale for Protein Signalling Networks



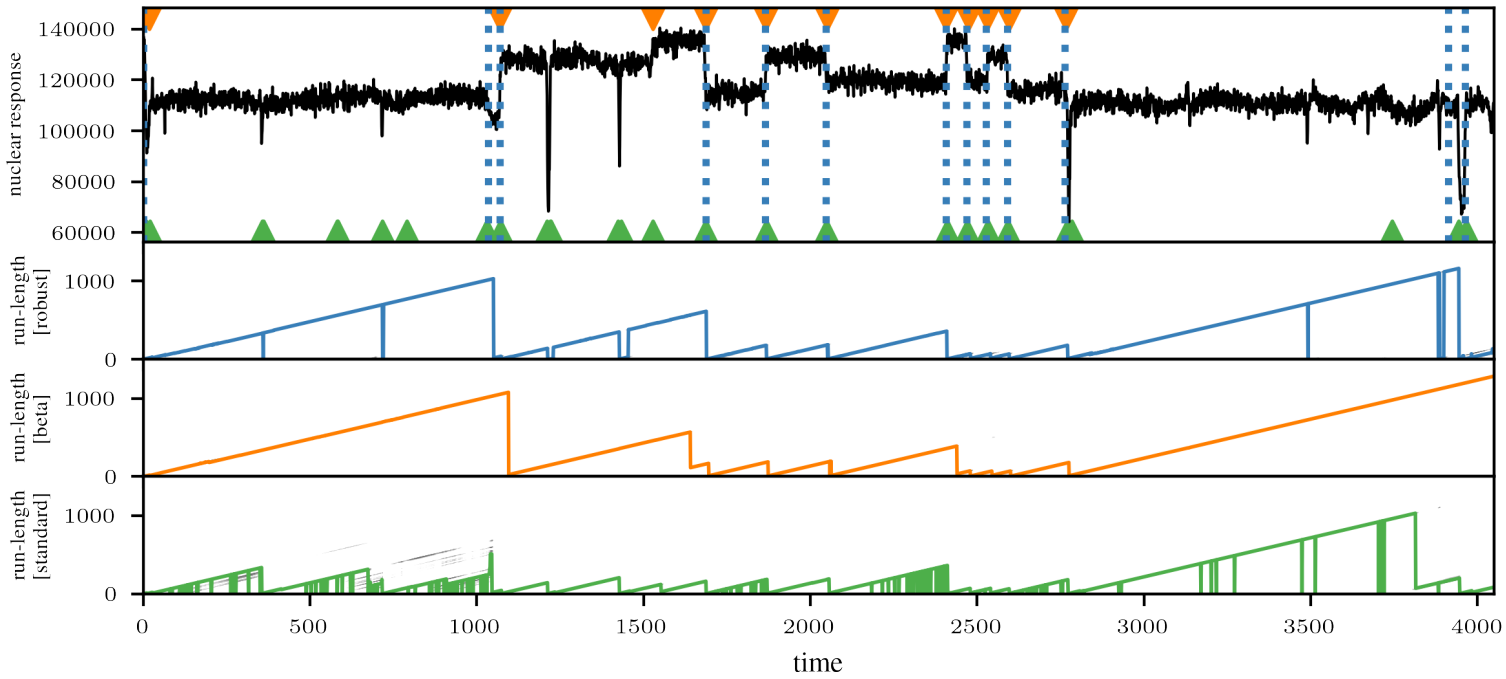
Dimension $d = 11$, $p = 68$ parameters and the dataset size $n = 7449$.

We are able to implement a conjugate inference scheme for KSD-Bayes!

Bayesian Online Changepoint Detection

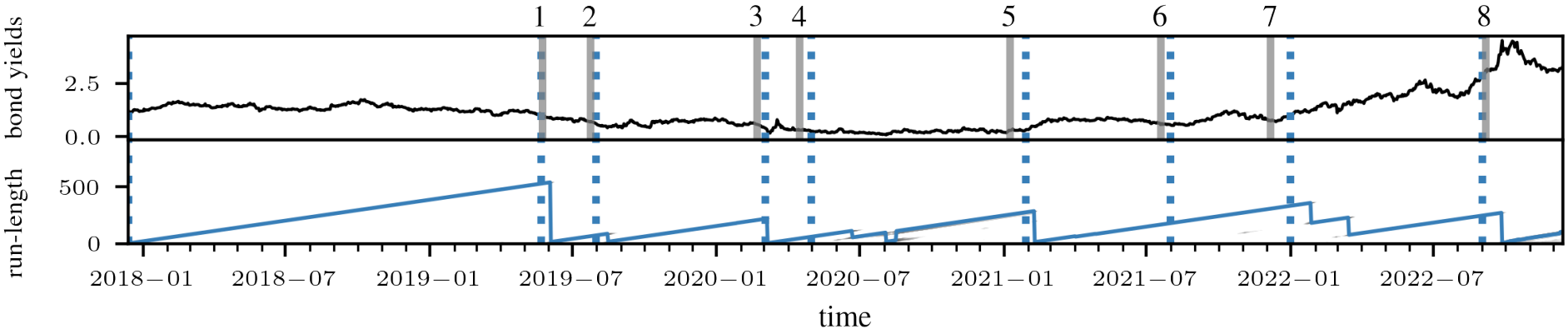
Model: iid data from $\mathbb{P}_\theta$, with different θ per segment, and a prior on distribution of change points.

Aim: Identify the changepoints in an online fashion. Scalability and robustness are both key!



Knoblauch et al. (2018). Doubly robust Bayesian inference for non-stationary streaming data with β-divergences. NeurIPS.

A Slightly More Depressing Example...



- UK's 10-year government bond yield, $\mathbb{P}_\theta$ is a Gamma per segment (standard Bayes posterior intractable, but conjugate for SM-Bayes).
- Not doubly intractable, but still benefit from robustness!
- Events (grey) includes times at which Boris Johnson and Liz Truss are sworn in as prime ministers, when Covid waves are officially announced, and the removal of Covid measures by the government.

Case 2: Discrete Data

The Challenge with Discrete Data

- Recall we are now in a scenario with a pmf $p_\theta(x) = \tilde{p}_\theta(x)/C(\theta)$.
- The continuous data case was tackled by focusing on the score:

$$\nabla_x \log p_\theta(x).$$

But we can't use the derivative ∇_x directly for discrete data!

- **Idea:** Look at an analoguous quantity based on ratios:

$$\frac{\nabla^- p_\theta(x)}{p_\theta(x)} = \begin{bmatrix} \frac{p_\theta(x) - p_\theta(x^{1-})}{p_\theta(x)} \\ \cdots \\ \frac{p_\theta(x) - p_\theta(x^{d-})}{p_\theta(x)} \end{bmatrix} = \begin{bmatrix} \frac{\tilde{p}_\theta(x) - \tilde{p}_\theta(x^{1-})}{\tilde{p}_\theta(x)} \\ \cdots \\ \frac{\tilde{p}_\theta(x) - \tilde{p}_\theta(x^{d-})}{\tilde{p}_\theta(x)} \end{bmatrix} = \frac{\nabla^- \tilde{p}_\theta(x)}{\tilde{p}_\theta(x)}$$

The Challenge with Discrete Data

- Recall we are now in a scenario with a pmf $p_\theta(x) = \tilde{p}_\theta(x)/C(\theta)$.
- The continuous data case was tackled by focusing on the score:

$$\nabla_x \log p_\theta(x).$$

But we can't use the derivative ∇_x directly for discrete data!

- **Idea:** Look at an analoguous quantity based on ratios:

$$\frac{\nabla^- p_\theta(x)}{p_\theta(x)} = \begin{bmatrix} \frac{p_\theta(x) - p_\theta(x^{1-})}{p_\theta(x)} \\ \cdots \\ \frac{p_\theta(x) - p_\theta(x^{d-})}{p_\theta(x)} \end{bmatrix} = \begin{bmatrix} \frac{\tilde{p}_\theta(x) - \tilde{p}_\theta(x^{1-})}{\tilde{p}_\theta(x)} \\ \cdots \\ \frac{\tilde{p}_\theta(x) - \tilde{p}_\theta(x^{d-})}{\tilde{p}_\theta(x)} \end{bmatrix} = \frac{\nabla^- \tilde{p}_\theta(x)}{\tilde{p}_\theta(x)}$$

Score-matching for Discrete Data

- Discrete Fisher divergence (i.e. discrete score-matching):

$$\begin{aligned}\mathrm{DFD}(\mathbb{P}_\theta, \mathbb{Q}) &:= \int_{\mathcal{X}} \left\| \frac{\nabla^- p_\theta(x)}{p_\theta(x)} - \frac{\nabla^- q(x)}{q(x)} \right\|_2^2 \mathbb{Q}(dx) \\ &= \int_{\mathcal{X}} \left\| \frac{\nabla^- \tilde{p}_\theta(x)}{\tilde{p}_\theta(x)} \right\|_2^2 + 2\nabla^+ \cdot \left(\frac{\nabla^- \tilde{p}_\theta(x)}{\tilde{p}_\theta(x)} \right) \mathbb{Q}(dx) + C_q\end{aligned}$$

- Even more scalable than before: usually $\mathcal{O}(nd)$, and sometimes $\mathcal{O}(d)$!
- Removes one level of intractability: can use standard MCMC.
- Unfortunately, no known conjugate priors...

Genomics: Multivariate Count Data

- The use of Bayesian methods for multivariate count data is extremely limited due to the challenges brought by doubly-intractable problems.
- For example, the Conway-Maxwell Poisson graphical model is extremely popular but does not yet have a Bayesian treatment:

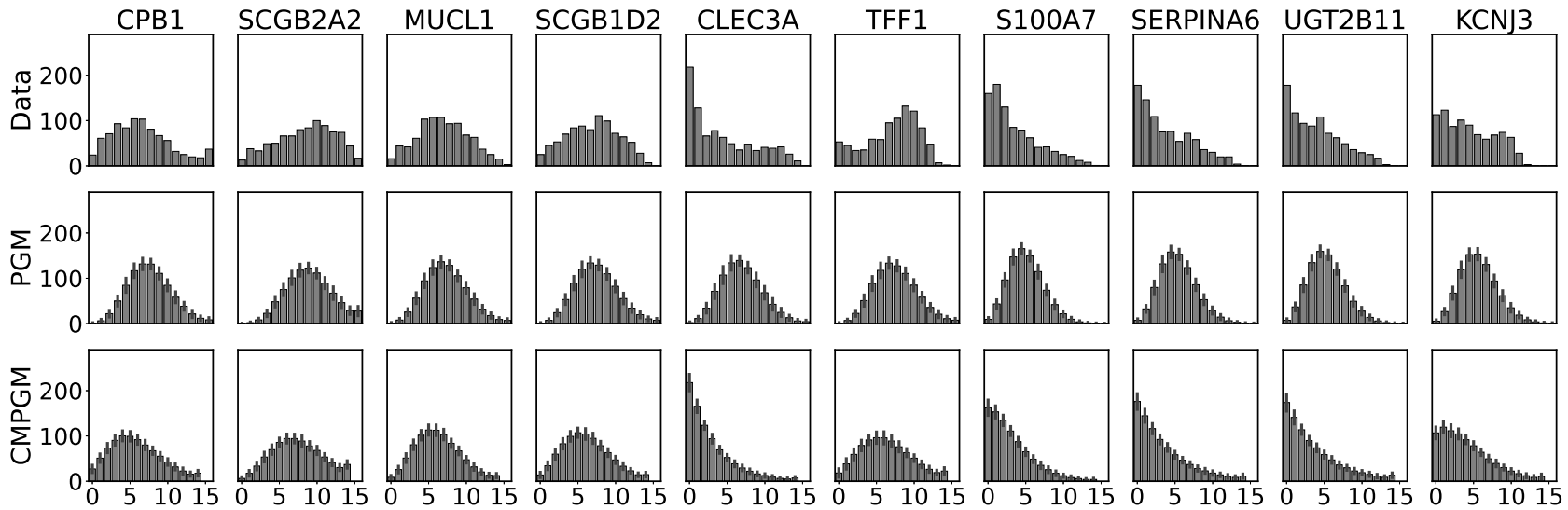
$$\tilde{p}_\theta(x) = \exp\left(\sum_{j=1}^{d} \theta_{(j)} x_{(j)} - \sum_{j=1}^{d} \log(x_{(j)}!) - \sum_{j=1}^{d} \sum_{k \in N_j} \theta_{(i,j)} x_{(j)} x_{(k)}\right)$$

where N_j are the neighbours in the set $\{j+1, \ldots, d\}$ and $\mathcal{X} \subset \mathbb{N}^d$.

[1] Shmueli et al. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. J. Royal Stat. Soc. Series C, 54(1), 127–142.

[2] Inouye et al. (2017). A review of multivariate distributions for count data derived from the Poisson distribution. Wiley Interdisciplinary Reviews: Computational Statistics, 9(3): e1398.

Genomics: Multivariate Count Data



- RNA sequencing data for breast cancer: $n = 878, d = 10, p = 74$. Aim is to discover genetic substructures of cancer. Data is total count of gene profiles found in biological samples.
- Both Poisson and Conway-Maxwell-Poisson graphical models can be fitted in reasonable time ($\approx$ 30mins) with Hamiltonian Monte Carlo!

Conclusion

Conclusion

- Going beyond the realm of standard Bayes and entering generalised Bayes gives a lot of felixibility to design methods which are:
  - Robust to model misspecification!
  - Much more computationally scalable!
  - Theoretically sound and provide good UQ (see the papers).

[1] Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. Journal of the Royal Statistical Society B: Statistical Methodology, 84(3), 997–1022.

[2] Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2022). Generalised Bayesian inference for discrete intractable likelihood. arXiv:2206.08420. (under revisions at JASA)

[3] Altamirano, M., Briol, F.-X., & Knoblauch, J. (2023). Robust and scalable Bayesian online changepoint detection. arXiv:2302.04759. (under review at ICML)