Post-Bayes inference for robust and conjugate models

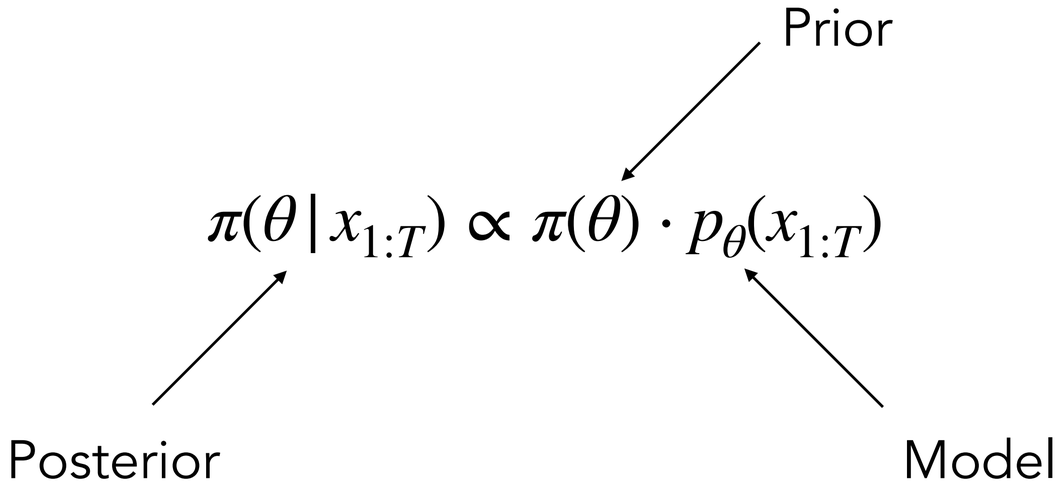
Matias Altamirano
University College London

Outline

1. Bayesian conjugate models: Benefits and Problems.
2. Weighted Score Matching (WSM) - Bayes.
3. WSM - Bayes for Regression

Bayesian Inference

Data $x_{1:T}$ generated from $\mathbb{P}$. We update our belief of a parameter of interest $\theta \in \Theta$ that index a statistical model $\{p_\theta, \theta \in \Theta\}$



Benefits: Optimal (Bayesian) update, Uncertainty Quantification, Inclusion of Expert Assessments

Bayesian Inference

Data $x_{1:T}$ generated from $\mathbb{P}$. We update our belief of a parameter of interest $\theta \in \Theta$ that index a statistical model $\{p_\theta, \theta \in \Theta\}$

Well specified prior

$$\pi(\theta \,|\, x_{1:T}) \propto \pi(\theta) \cdot p_\theta(x_{1:T})$$

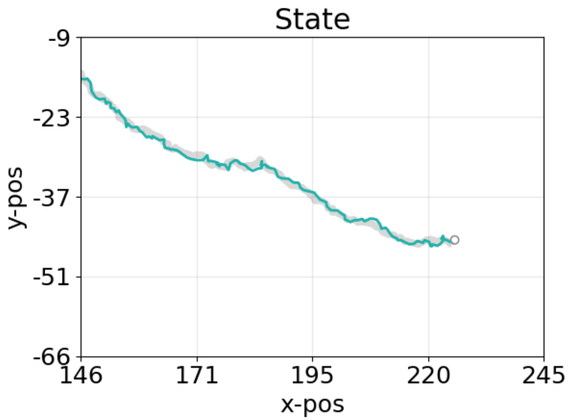
Enough computational power

Correct model

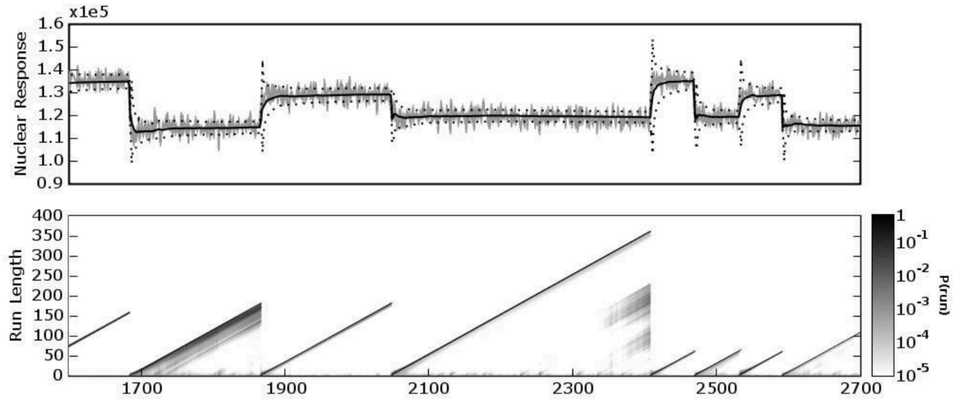
Benefits: Optimal (Bayesian) update, Uncertainty Quantification, Inclusion of Expert Assessments

Conjugate Bayesian methods

Kalman Filter



Bayesian Online Changepoint Detection



Gaussian Processes

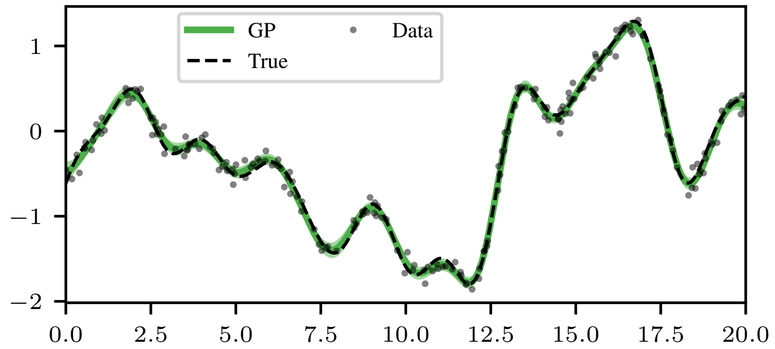


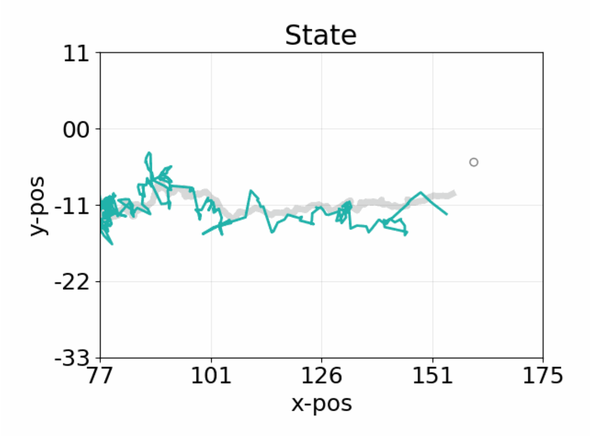
Figure taken from Adams, R. P., & MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

What do these Bayesian methods have in common?

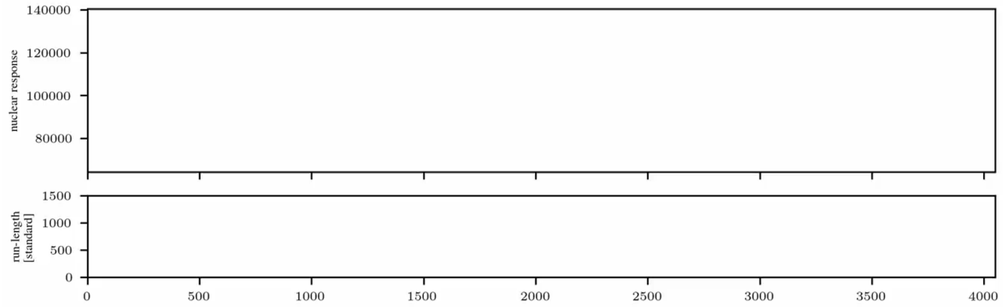
- Sequential Bayesian update requires closed-form posterior
- To achieve this, model choice is restricted
- Lack of flexibility to deal with extreme observations

Bayesian methods in the presence of Outliers

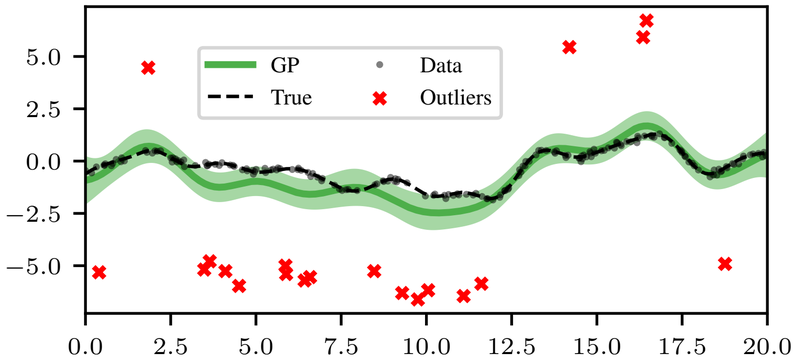
Kalman Filter



Bayesian Online Changepoint Detection



Gaussian Processes



Bayesian Inference

Data $x_{1:T}$ generated from $\mathbb{P}$. We update our belief of a parameter of interest $\theta \in \Theta$ that index a statistical model $\{p_\theta, \theta \in \Theta\}$

$$\pi(\theta \,|\, x_{1:T}) \propto \pi(\theta) \cdot p_\theta(x_{1:T})$$

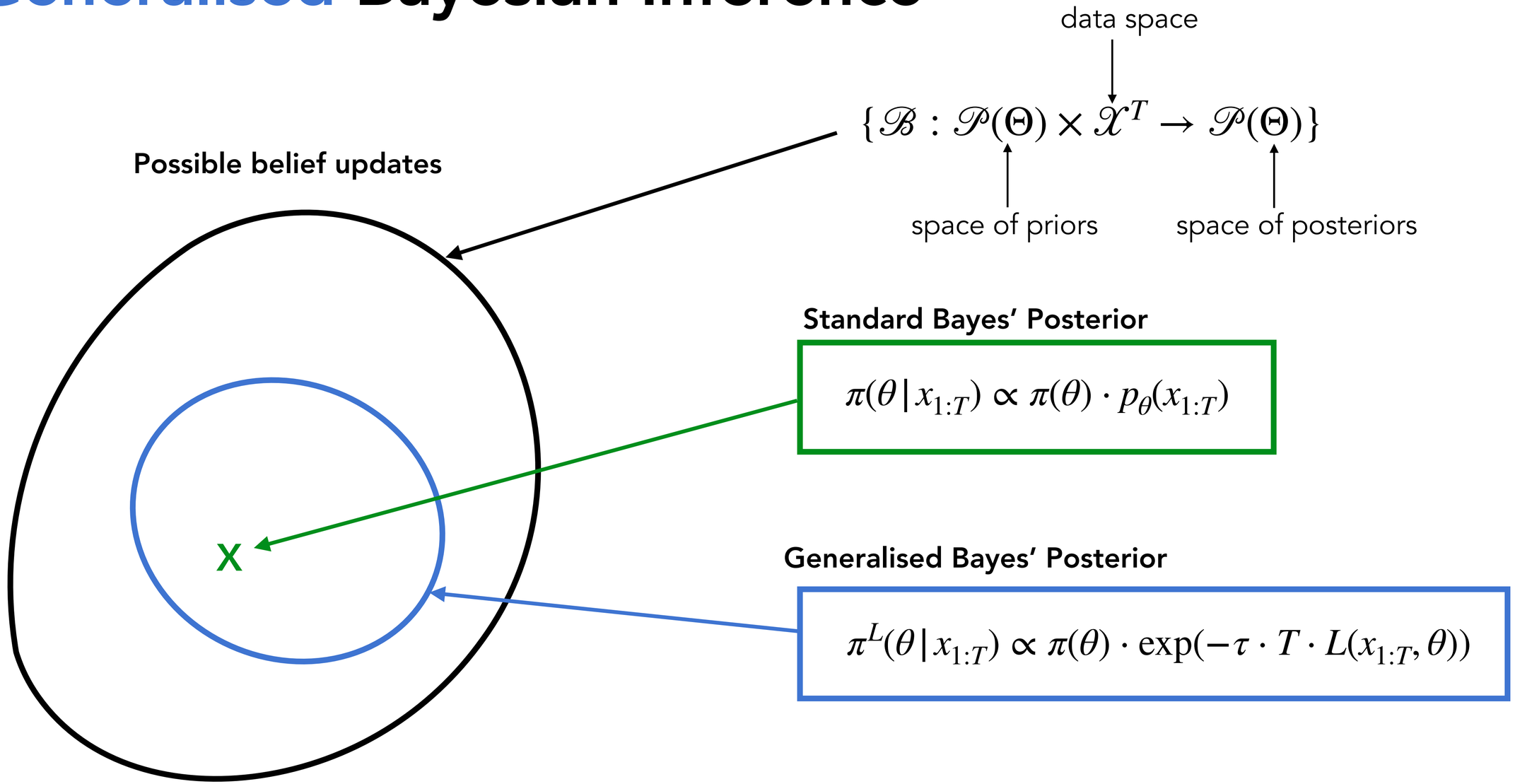
Hard computation

Misspecified model

Problems: Not optimal update, Incorrect Uncertainty Quantification

Solution

Generalised Bayesian Inference



See Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Generalised Bayesian Inference

$$\pi^L(\theta \,|\, x_{1:T}) \propto \pi(\theta) \cdot \exp(-T \cdot L(x_{1:T}, \theta))$$

Q: Can we select L to obtain robustness and tractability?

✓ Divergence based losses are good options! $D(\mathbb{P}, \mathbb{P}_\theta) \geq 0$ $D(\mathbb{P}, \mathbb{P}_\theta) = 0 \iff \mathbb{P} = \mathbb{P}_\theta$

Under mild conditions $\pi^L(\theta \,|\, x_{1:T})$ concentrates around $\theta^\star = \operatorname{argmin}_{\theta \in \Theta} L(\theta, x_{1:T})$

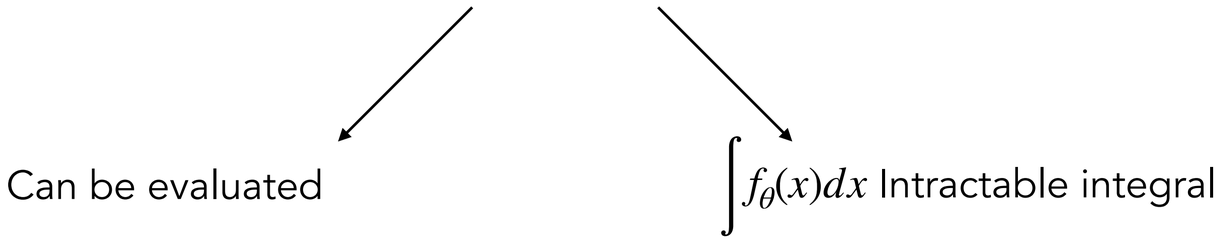
See Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Weighted Score Matching (WSM)

- Generalisation of Score Matching.

- Score Matching has been used for:
  - Paratemer estimation
  - Hypothesis testing
  - Diffusion Models

- Score Matching works with unnormalised models, i.e. $p_\theta(\,.\,) = f_\theta(\,.\,)/Z(\theta)$



Hyvarinen, A. (2006). Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6, 695-708.
Barp et al. (2019). Minimum Stein discrepancy estimators. Neural Information Processing Systems, 12964-12976.

Weighted Score Matching (WSM)

$$\mathscr{D}_w(\mathbb{P}, \mathbb{P}_\theta) = \mathbb{E}_{X\sim\mathbb{P}}\left[\|w^\top(X)(s_{p_\theta}(X) - s_p(X))\|_2^2\right]$$

We don't have access to the density of the true generative process!

- $s_p(x) = \nabla \log p(x)$ and $w : \mathbb{R}^d \to \mathbb{R}^{d\times d}$
- The case $w(X) = I_d$ is the original case studied by Hyvarinen, and Barp generalised this later (crucial for robustness).

Hyvarinen, A. (2006). Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6, 695-708.
Barp et al. (2019). Minimum Stein discrepancy estimators. Neural Information Processing Systems, 12964-12976.

Weighted Score Matching (WSM)

$$\mathscr{D}_w(\mathbb{P}, \mathbb{P}_\theta) = \mathbb{E}_{X\sim\mathbb{P}}[\|(w^\top s_{p_\theta})(X)\|_2^2 + (2\nabla\cdot(ww^\top s_{p_\theta}))(X)] + C$$

$$\widehat{\mathscr{D}_w}(\mathbb{P}_T, \mathbb{P}_\theta) = \sum_{t=1}^{T} \|(w^\top s_{p_\theta})(x_t)\|_2^2 + (2\nabla\cdot(ww^\top s_{p_\theta}))(x_t)$$

- $s_p(x) = \nabla \log p(x)$ and $w : \mathbb{R}^d \to \mathbb{R}^{d\times d}$
- The case $w(X) = I_d$ is the original case studied by Hyvarinen, and Barp generalised this later (crucial for robustness).

Hyvarinen, A. (2006). Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6, 695-708.
Barp et al. (2019). Minimum Stein discrepancy estimators. Neural Information Processing Systems, 12964-12976.

WSM-Bayes

$$\pi^{\mathscr{D}_w}(\theta \mid x_{1:T}) \propto \pi(\theta) \cdot \exp\{-T \cdot \mathscr{D}_w(\mathbb{P}_T, \mathbb{P}_\theta)\}$$

Q: Is WSM-Bayes robust and tractable?

Tractability for WSM-Bayes

Comparison	$\pi^{\mathscr{D}_w}(\theta \mid x_{1:T})$	$\Longleftrightarrow$	$\pi(\theta \mid x_{1:T})$
$p_\theta(\,.\,)$ tractable	$\pi^{\mathscr{D}_w}(\theta \mid x_{1:T})$	**Roughly as fast as**	$\pi(\theta \mid x_{1:T})$
$p_\theta(\,.\,) = f_\theta(\,.\,)/Z(\theta)$ + $Z(\theta)$ tractable	$\pi^{\mathscr{D}_w}(\theta \mid x_{1:T})$	**Faster to compute than**	$\pi(\theta \mid x_{1:T})$
$p_\theta(\,.\,)$ exponential family + $\pi(\theta)$ conjugate			

WSM-Bayes

$$\exp\{-T \cdot \mathscr{D}_w(\mathbb{P}_T, \mathbb{P}_\theta)\} = \exp\left\{ -\sum_{t=1}^{T} (\theta - \mu(x_t))^\top \Lambda(x_t)(\theta - \mu(x_t)) \right\}$$

Exponential family

$$\pi^{\mathscr{D}_w}(\theta \,|\, x_{1:T}) \propto \exp\left\{ -\sum_{t=1}^{T} (\theta - \mu(x_t))^\top \Lambda(x_t)(\theta - \mu(x_t)) \right\} \exp\left\{-(\theta - \mu_0)^\top \Lambda_0 (\theta - \mu_0)\right\}$$

$$= \mathcal{N}(\theta; \mu_w(x_{1:T}), \Sigma_w(x_{1:T})))$$

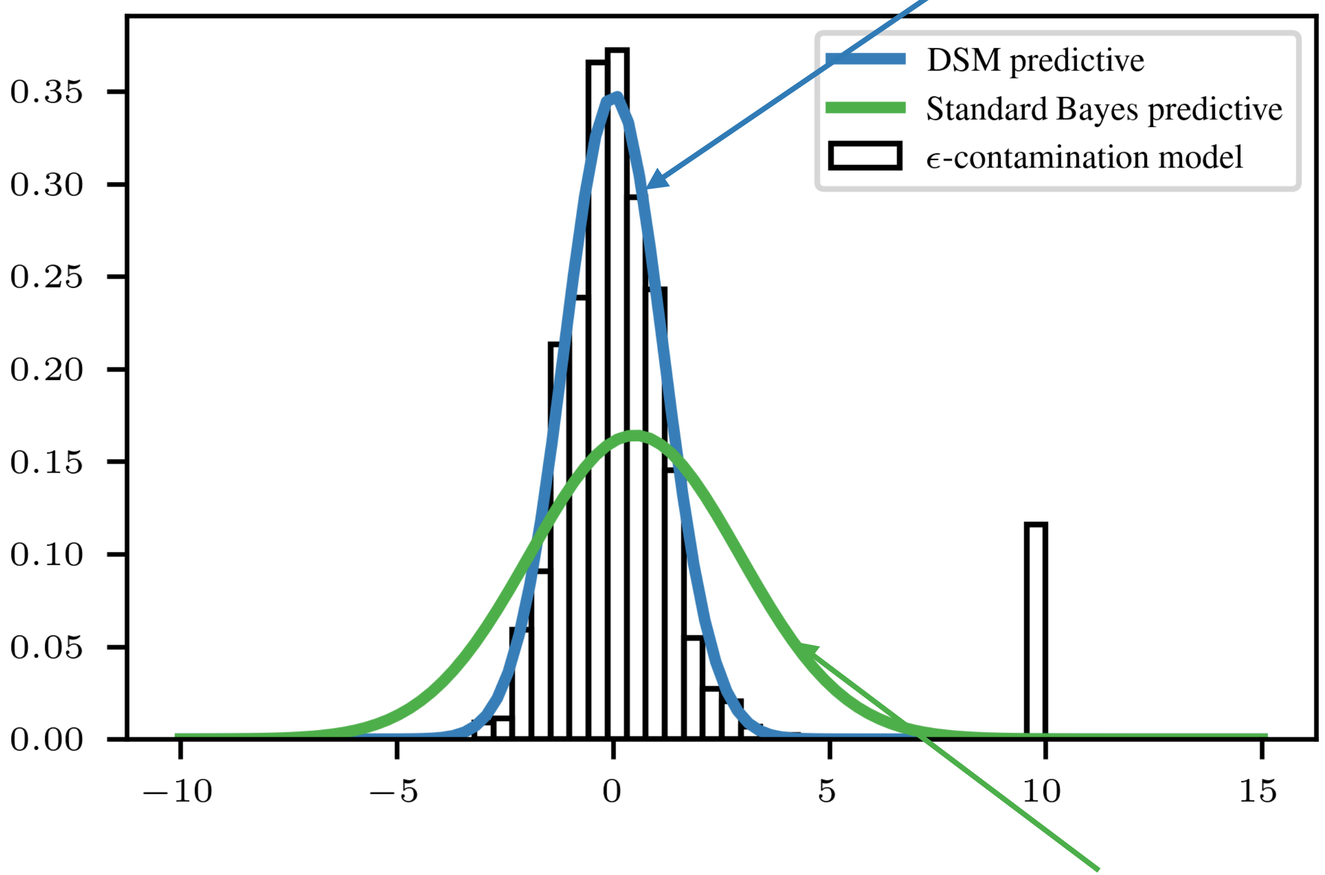
Conjugate Prior!

Tractability for WSM-Bayes

Comparison	$\pi^{\mathscr{D}_w}(\theta \mid x_{1:T})$	$\Longleftrightarrow$	$\pi(\theta \mid x_{1:T})$
$p_\theta(\,.\,)$ tractable	$\pi^{\mathscr{D}_w}(\theta \mid x_{1:T})$	**Roughly as fast as**	$\pi(\theta \mid x_{1:T})$
$p_\theta(\,.\,) = f_\theta(\,.\,)/Z(\theta)$ + $Z(\theta)$ tractable	$\pi^{\mathscr{D}_w}(\theta \mid x_{1:T})$	**Faster to compute than**	$\pi(\theta \mid x_{1:T})$
$p_\theta(\,.\,)$ exponential family + $\pi(\theta)$ conjugate	$\pi^{\mathscr{D}_w}(\theta \mid x_{1:T})$	**as fast as / faster to compute than**	$\pi(\theta \mid x_{1:T})$

Robustness of WSM-Bayes

Concentrated in the ‘correct’ data



Influenced by the contamination

Robustness of WSM-Bayes

Setting: $p^{\varepsilon} = (1-\varepsilon)\cdot p + \varepsilon \cdot c$

$$x_{1:T} \sim p^{\varepsilon} \longrightarrow \pi^{\mathscr{D}_w}(\theta \,|\, x_{1:T})$$

$$z_{1:T} \sim p \longrightarrow \pi^{\mathscr{D}_w}(\theta \,|\, z_{1:T})$$

Robustness:

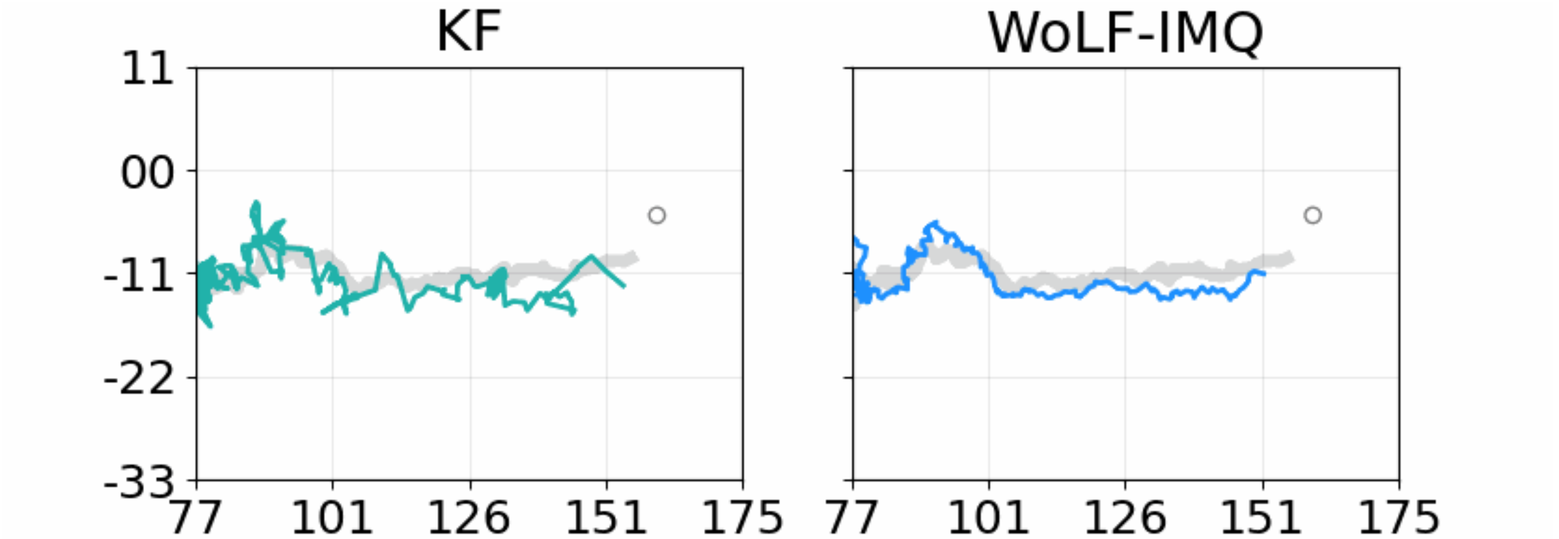
$$\sup_{c \in \delta} \left\{ \text{distance}(\pi^{\mathscr{D}_w}(\theta \,|\, x_{1:T}), \pi^{\mathscr{D}_w}(\theta \,|\, z_{1:T})) \right\} \leq c(\delta) \cdot \varepsilon$$

$$\sup_{\theta \in \Theta} \left| (\pi^{\mathscr{D}_w}(\theta \,|\, x_{1:T}) \,| - \pi^{\mathscr{D}_w}(\theta \,|\, z_{1:T})) \right|$$

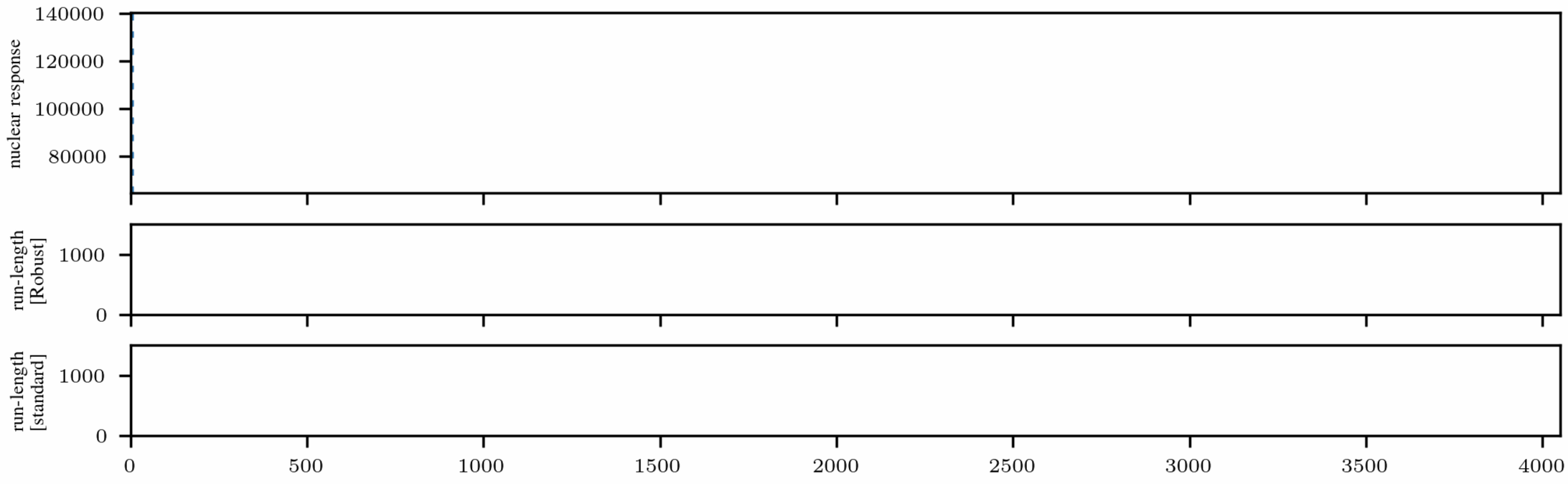
Under some conditions on w, WSM-Bayes is robust!

Results

Kalman Filter



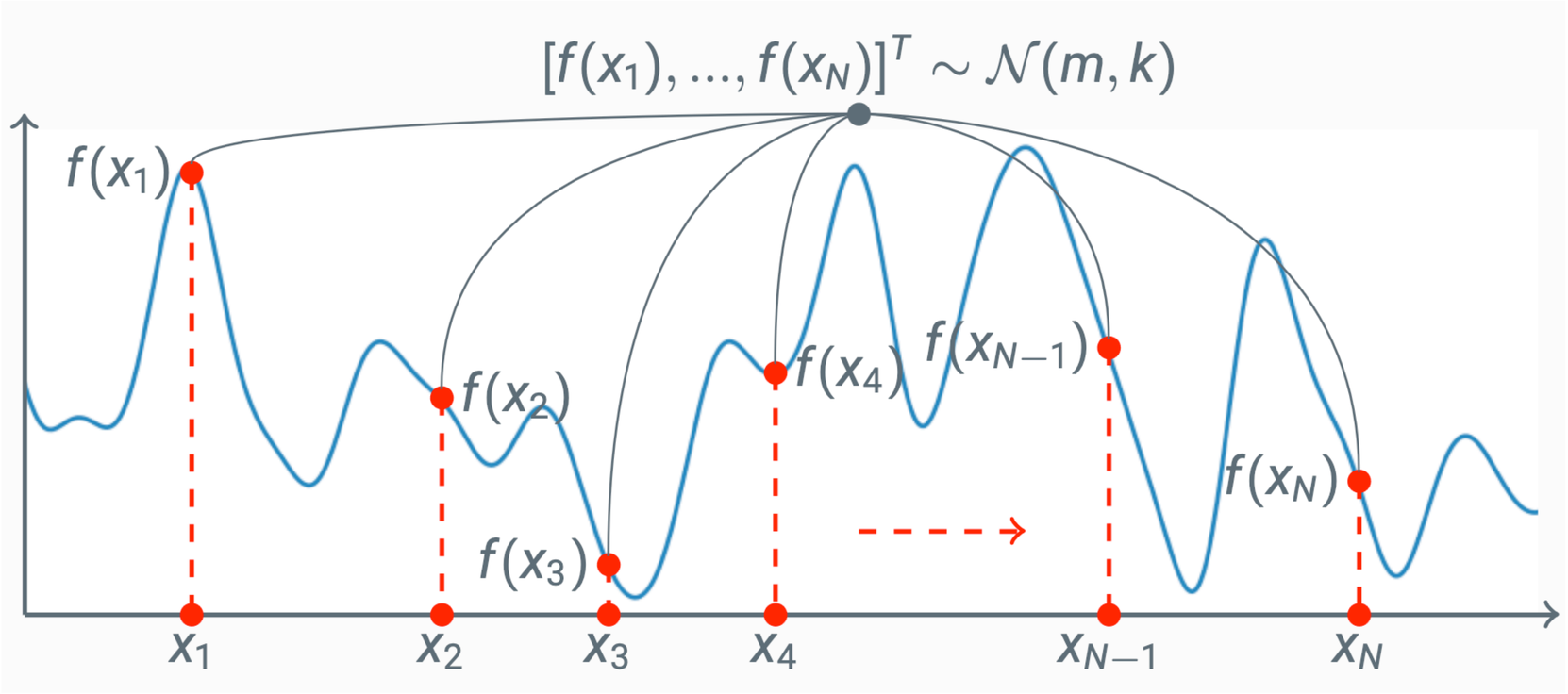
Bayesian Online changepoint Detection



Gaussian Processes

Gaussian Processes

$f \sim GP(m, k) \Leftrightarrow$ For every subset $\mathbf{x} = \{x_i\}_{i=1}^{n}$, then $\mathbf{f} = [f(x_1), \ldots, f(x_n)] \sim \mathcal{N}(m, k)$

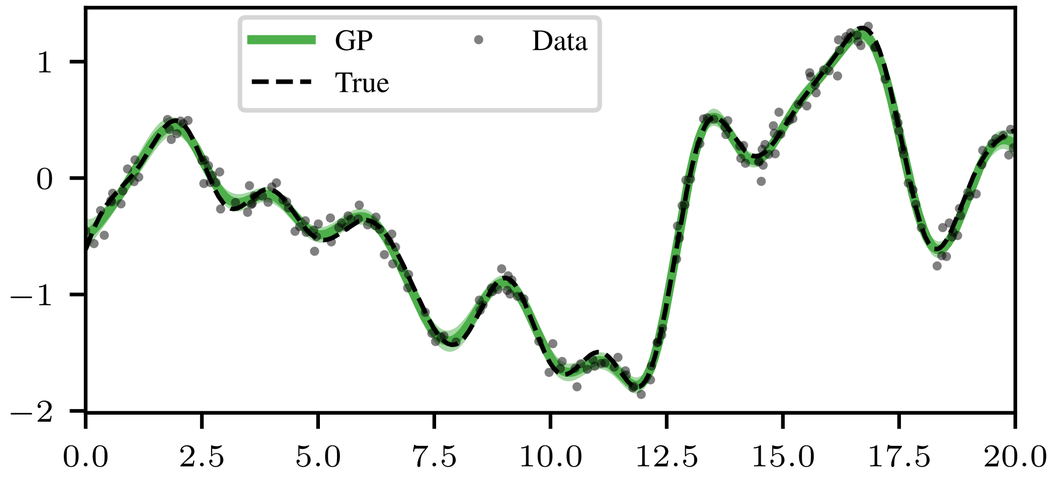


Gaussian Process Regression

Suppose $f \sim GP(m, k)$ and $\epsilon \sim N(0, \sigma^2 I_n)$: $\Rightarrow$ $p(\mathbf{f} \mid \mathbf{x}) = N(\mathbf{f}; \mathbf{m}, K)$ $\quad p(\mathbf{y} \mid \mathbf{f}, \mathbf{x}) = N(\mathbf{y}; \mathbf{f}, \sigma^2 I_n)$

Posterior: $p(\mathbf{f} \mid \mathbf{y}, \mathbf{x}) \propto p(\mathbf{y} \mid \mathbf{f}, \mathbf{x}) \cdot p(\mathbf{f} \mid \mathbf{x}) = N(\mathbf{f}; \mu, \Sigma)$

$$\mu = \mathbf{m} + K(K + \sigma^2 I_n)^{-1}(\mathbf{y} - \mathbf{m}) \qquad \Sigma = K(K + \sigma^2 I_n)^{-1}\sigma^2 I_n$$



WSM for Regression

For regression setting, we need to extend wsm (now $w : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$):

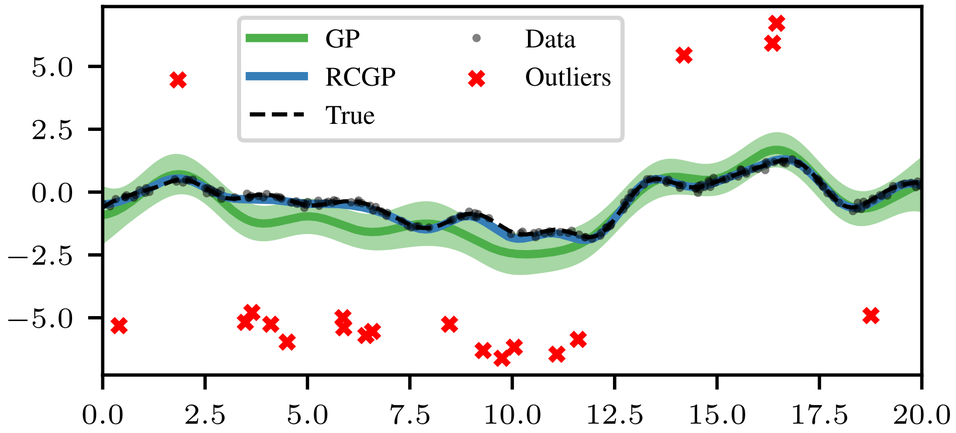
$$\mathscr{D}_w(p, q) := \mathbb{E}_{X \sim q_x}\left[\mathbb{E}_{Y \sim q(\cdot|X)}\left[\|(w(\nabla \log p - \nabla \log q))(X, Y)\|_2^2\right]\right]$$

$$L_n^w(\mathbf{f}, \mathbf{y}, \mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left((w \nabla \log p_f)^2 + 2\nabla_y(w^2 \nabla \log p_f)\right)(x_i, y_i)$$

Likelihood

Solution: Robust and Conjugate GP (RCGP)!

Key idea: replace standard posterior with **robust posterior**



$$p^w(\mathbf{f}\,|\,\mathbf{y},\mathbf{x}) \propto \exp\left(-nL_n^w(\mathbf{f},\mathbf{y},\mathbf{x})\right)\cdot p(\mathbf{f}\,|\,\mathbf{x})$$

$$L_n^w(\mathbf{f},\mathbf{y},\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left((w\,\nabla\log p_f)^2 + 2\,\nabla_y(w^2\,\nabla\log p_f)\right)(x_i,y_i)$$

Suppose $f \sim GP(m,k)$ and $\epsilon \sim N(0,\sigma^2 I_n)$:

$$p^w(\mathbf{f}\,|\,\mathbf{y},\mathbf{x}) = N(\mathbf{f};\mu^R,\Sigma^R)$$

$$\mu^R = \mathbf{m} + K(K+\sigma^2 J_{\mathbf{w}})^{-1}(\mathbf{y}-\mathbf{m}_{\mathbf{w}})$$

$$\Sigma^R = K(K+\sigma^2 J_{\mathbf{w}})^{-1}\sigma^2 J_{\mathbf{w}}$$

$$J_{\mathbf{w}} = \mathrm{diag}(\mathbf{w}^{-2})$$

$$\mathbf{m}_{\mathbf{w}} = \mathbf{m} + \sigma^2\,\nabla_y \log(\mathbf{w}^2)$$

Robustness of RCGP

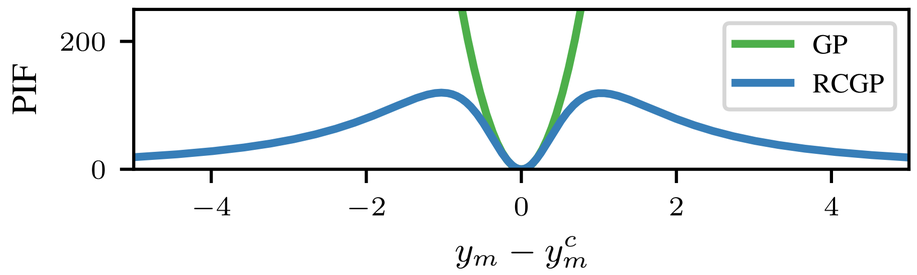
Posterior Influence Function: measures the impact of a single outlier on the posterior:

$$\mathrm{PIF}(y_m^c, D) = \mathrm{KL}\left(p(f \mid D), p(f \mid D_m^c)\right)$$

$D = \{(x_i, y_i\}_{i=1}^n$ $\qquad$ $D_m^c = (D \backslash \{(x_m, y_m)\}) \cup \{(x_m, y_m^c)\}$

Theorem (simplified): Suppose $w(x, y) = (1 + (y - m(x))^2/c^2)^{-\frac{1}{2}}$ for some $c > 0$, then **RCGPs** are robust since:

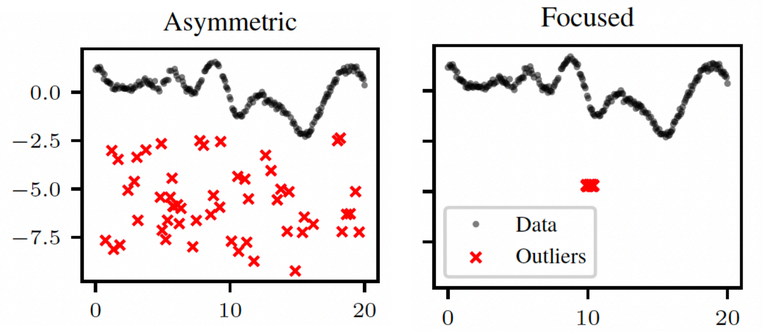
$$\sup_{y_m^c \in \mathbb{R}} \mathrm{PIF}_{\mathrm{RCGP}}(y_m^c, D) < \infty$$



RCGP is Robust and Scalable!

Mean Absolute Error

	GP	RCGP	t-GP	m-GP
Focused Outliers				
Synthetic	0.19 (0.00)	**0.16 (0.00)**	0.20 (0.00)	0.23 (0.0)
Boston	0.27 (0.12)	**0.22 (0.03)**	0.25 (0.01)	0.27 (0.0)
Energy	0.06 (0.06)	**0.02 (0.00)**	0.03 (0.00)	0.24 (0.0)
Yacht	0.28 (0.19)	**0.10 (0.06)**	0.24 (0.08)	0.24 (0.0)
Asymmetric Outliers				
Synthetic	1.14 (0.00)	0.82 (0.00)	1.06 (0.00)	**0.61 (0.0)**
Boston	0.64 (0.04)	**0.49 (0.01)**	0.52 (0.00)	0.52 (0.0)
Energy	0.55 (0.05)	0.50 (0.16)	0.44 (0.04)	**0.41 (0.0)**
Yacht	0.54 (0.06)	**0.36 (0.05)**	0.41 (0.00)	0.40 (0.0)



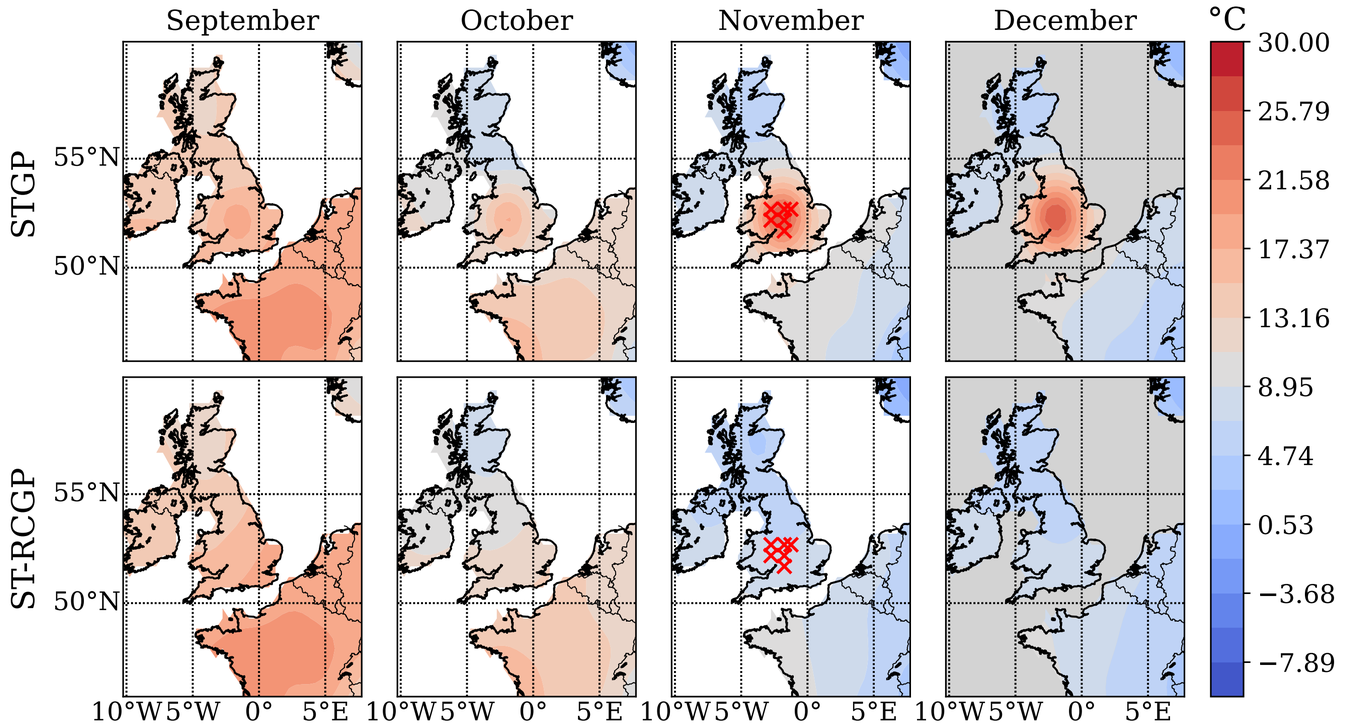
RCGP is robust!

Time [s]

	GP	RCGP	t-GP	m-GP
Synthetic	1.5 (0.1)	1.2 (0.0)	2.2 (0.0)	3.0 (0.0)
Boston	1.9 (0.5)	5.1 (0.9)	30.7 (6.1)	16.7 (1.7)
Energy	3.8 (0.9)	4.6 (2.0)	34.0 (11)	33.8 (0.3)
Yacht	1.6 (0.3)	2.1 (0.2)	5.6 (0.7)	4.5 (0.4)

RCGP is much faster!

RCGP in spatio-temporal setting



Summary

1. Standard Bayesian inference is either robust or scalable, not both.

2. We can rectify this by using post-Bayesian methods.

Any Questions?

Robust and Scalable Bayesian Online Changepoint Detection
Matias Altamirano, François-Xavier Briol, Jeremias Knoblauch
ICML 2023

Robust and Conjugate Gaussian Process Regression
Matias Altamirano, François-Xavier Briol, Jeremias Knoblauch
ICML 2024, Spotlight

Outlier-robust Kalman Filtering through Generalised Bayes
Gerardo Durán-Martín, ***Matias Altamirano***, Alexander Y Briol, and Leandro Sánchez-Betancourt, Jeremias Knoblauch, Matt Jones, François-Xavier, Kevin Murphy
ICML 2024

Robust and Conjugate Spatio-Temporal Gaussian Processes
William Laplante, **Matias Altamirano**, Andrew Duncan
François-Xavier Briol, Jeremias Knoblauch
ICML 2025

Applications of the Bayesian Nonparametric Learning framework and the Posterior Bootstrap

Harita Dellaporta
Post-Bayes seminar 06/08/2025

Outline

- Bayesian Nonparametric Learning
- Application 1: Simulation-based Inference
- Application 2: Measurement error models
- Discussion

Bayesian Nonparametric Learning and the Posterior Bootstrap

Notation

- Observations $x_{1:n} \overset{i.i.d.}{\sim} \mathbb{P}^\star$ ← Data-generating process over data space $\mathcal{X}$
- Model family $\mathscr{P}_\Theta := \{\mathbb{P}_\theta : \theta \in \Theta\}$

Bayesian Nonparametric Learning

Main idea

Standard Bayesian Inference

$$\underbrace{P(\theta \mid x_{1:n})}_{\text{Posterior}} \propto \underbrace{P(x_{1:n} \mid \theta)}_{\text{Likelihood}} \cdot \underbrace{P(\theta)}_{\text{Prior}}$$

- Likelihood/model is assumed to be correct
- Uncertainty is set directly on model parameters θ via the prior

Bayesian Nonparametric Learning (NPL) [Lyddon et al. (2018); Fong et al. (2019)]

$$\underbrace{\mathbb{P}^\star \sim P(\mathbb{P}^\star)}_{\text{Nonp. Prior on } \mathbb{P}^\star} \rightarrow \underbrace{P(\mathbb{P}^\star \mid x_{1:n})}_{\text{Nonp. Posterior on } \mathbb{P}^\star} \xrightarrow{\text{Via some loss}} \underbrace{\Pi_{\text{NPL}}(\theta)}_{\text{NPL Posterior on } \theta}$$

- Model not assumed to be correct
- Uncertainty set directly on $\mathbb{P}^\star$ and propagated to θ

Bayesian Nonparametric Learning

- Place a nonparametric prior *directly* on the data-generating process $\mathbb{P}^\star$:

$$\mathbb{P} \sim DP(\alpha, \mathbb{F}), \quad \mathbb{P} \mid x_{1:n} \sim DP(\alpha', \mathbb{F}')$$

where

$$\alpha' := \alpha + n, \quad \mathbb{F}' := \frac{\alpha}{\alpha + n}\mathbb{F} + \frac{n}{\alpha + n}\mathbb{P}_n, \quad \mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

- For a loss function $l : \mathcal{X} \times \Theta \to \mathbb{R}$, propagate uncertainty from $\mathbb{P}^\star$ to the parameter of interest θ through

$$\theta_l^\star(\mathbb{P}^\star) := \arg\min_{\theta \in \Theta} \mathbb{E}_{X \sim \mathbb{P}^\star}[l(X, \theta)]$$

- The push-forward measure $(\theta_l^\star)_\#(DP(\alpha', \mathbb{F}'))$ gives a posterior on Θ denoted by Π_{NPL}

Sampling?

Posterior bootstrap

For $j = 1,2,\ldots$

1. Sample $\mathbb{P}^{(j)} \sim DP(\alpha', \mathbb{F}')$

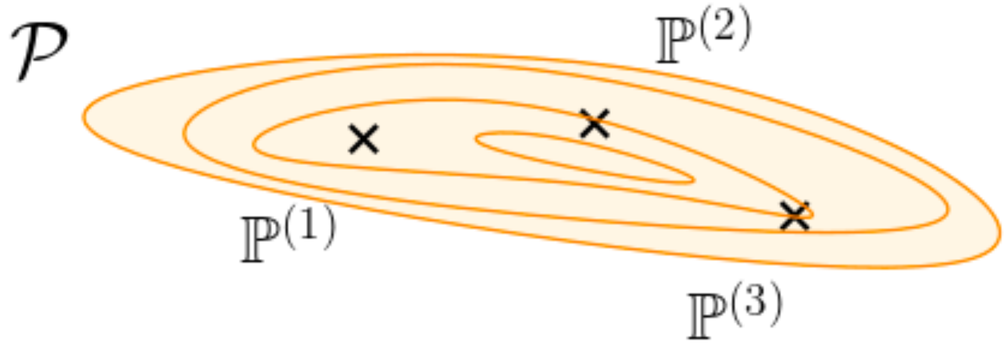
$$\alpha = 0: \quad w_{1:n}^{(j)} \sim \text{Dir}(1,\cdots,1), \quad \theta^{(j)} := \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} w_i^{(j)} l(x_i, \theta)$$

2. Find $\theta^{(j)} := \theta_l^\star(\mathbb{P}^{(j)}) = \arg\min_{\theta\in\Theta} \mathbb{E}_{X\sim\mathbb{P}^{(j)}}[l(X,\theta)]$

- $\alpha = 0 \rightarrow$ **Loss-likelihood bootstrap [Lyddon et al. 2019]**
- $\alpha = 0$ **and loss is negative log-likelihood** $\rightarrow$ **weighted likelihood bootstrap [Newton and Raftery 1994]**

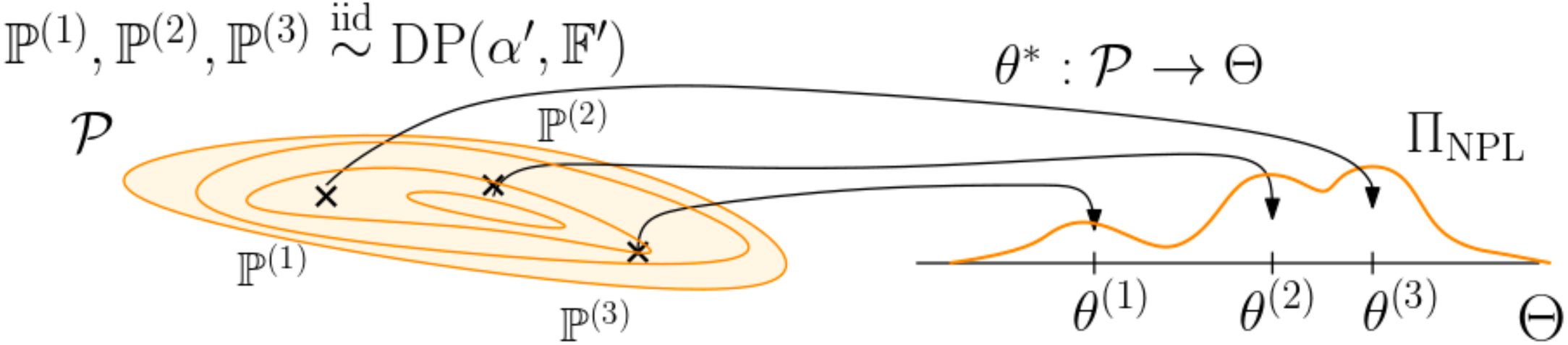
Posterior Bootstrap

$$\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \mathbb{P}^{(3)} \overset{\text{iid}}{\sim} \mathrm{DP}(\alpha', \mathbb{F}')$$



1. Draw $\mathbb{P}^{(j)} \sim DP(\alpha', \mathbb{F}')$

Posterior Bootstrap



1. Draw $\mathbb{P}^{(j)} \sim DP(\alpha', \mathbb{F}')$

2. Obtain $\theta^{(j)} := \theta_l^\star(\mathbb{P}^{(j)}) = \arg\min_{\theta \in \Theta} \mathbb{E}_{X \sim \mathbb{P}^{(j)}}[l(X, \theta)]$

Why NPL/Posterior Bootstrap?

- **Robustness** to model misspecification
- Direct inference for a **functional of interest** (e.g. median $l(x, \theta) = |\theta - x|$)
- Sampling from **multimodal** posteriors
- **Consistency:** under regularity conditions, the NPL posterior is consistent at $\theta_l^\star(\mathbb{P}^\star) := \arg\min_{\theta \in \Theta} \mathbb{E}_{X \sim \mathbb{P}^\star}[l(X, \theta)]$ regardless of the choice of $\mathbb{F}$ and its support

Applications of Posterior Bootstrap

- Medical Imaging [Goncharov et al. 2023]
- Ensemble tree modelling [Galvani et al. 2021]
- Transfer Learning [Lee et al. 2024]
- Normalising Flows [Ott and Williamson 2023]
- Generative Networks [Nie and Ročková 2023]

Application 1: Simulation-based Inference

[Flexibility of loss choice]



Dellaporta, C., Knoblauch, J., Damoulas, T. and Briol, F.X., 2022, May. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics* (pp. 943–970). PMLR.

Simulator-based models

- Independent sampling is possible, but the likelihood is unavailable
- Model is usually at best a rough approximation of a complex physical or biological phenomenon
- It will most likely not capture all of the key characteristics of the underlying data-generating process

Two main problems

1. Unavailability of the likelihood function
2. Model misspecification

Problem Setting

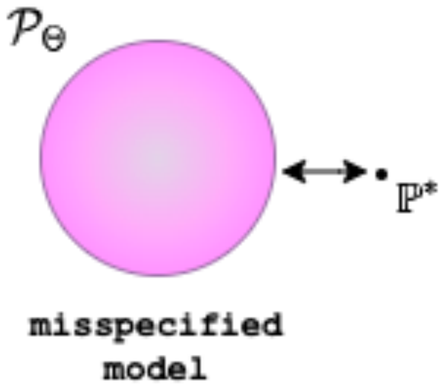
Simulation-based model family

Observed i.i.d data $x_{1:n} \sim \mathbb{P}^\star$

$$\mathscr{P}_\Theta := \{\mathbb{P}_\theta : \theta \in \Theta\} \subseteq \mathscr{P}$$

$\mathbb{P}_\theta$ is associated with simulator function $G_\theta : \mathscr{U} \to \mathscr{X}$ and probability measure $\mathbb{U}$ in $\mathscr{U}$ such that

$$u \sim \mathbb{U}, \quad x := G_\theta(u) \sim \mathbb{P}_\theta$$



Maximum Mean Discrepancy (MMD) based loss

$$\theta_l^\star(\mathbb{P}^\star) := \arg\min_{\theta\in\Theta} \mathbb{E}_{X\sim\mathbb{P}^\star}[l(X,\theta)]$$

$$\downarrow$$

$$\mathrm{MMD}(\mathbb{P}_\theta, \mathbb{P}^\star)$$

- Map $\theta^\star$ now corresponds to a minimum MMD estimator as in Briol et al. (2019); Chérief-Abdellatif and Alquier (2022)
- The MMD between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ in an RKHS $\mathscr{H}_k$ with associated reproducing kernel $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ can be expressed as

$$\mathrm{MMD}_k^2(\mathbb{P},\mathbb{Q}) := \int_{\mathscr{X}}\int_{\mathscr{X}} k(x,y)\mathbb{P}(dx)\mathbb{P}(dy) - 2\int_{\mathscr{X}}\int_{\mathscr{X}} k(x,y)\mathbb{P}(dx)\mathbb{Q}(dy)$$

$$+\int_{\mathscr{X}}\int_{\mathscr{X}} k(x,y)\mathbb{Q}(dx)\mathbb{Q}(dy)$$

Generalisation Error

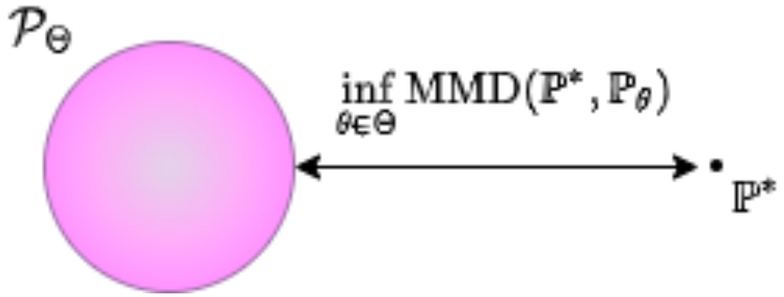
- Assumption: bounded kernel w.l.o.g. $|k(x, x')| \leq 1 \forall x, x' \in \mathcal{X}$
- Let $\nu := DP(\alpha', \mathbb{F}')$ (DP posterior)

$$\mathbb{E}_{x_{1:n} \sim \mathbb{P}^\star}[\mathbb{E}_{\mathbb{P} \sim \nu}[\mathrm{MMD}_k(\mathbb{P}^\star, \mathbb{P}_{\theta^\star(\mathbb{P})})]]$$

Generalisation Error

- Assumption: bounded kernel w.l.o.g. $|k(x, x')| \leq 1 \forall x, x' \in \mathcal{X}$
- Let $\nu := DP(\alpha', \mathbb{F}')$ (DP posterior)

$$\inf_{\theta \in \Theta} \text{MMD}_k(\mathbb{P}^\star, \mathbb{P}_\theta) \quad \leq \quad \mathbb{E}_{x_{1:n} \sim \mathbb{P}^\star}[\mathbb{E}_{\mathbb{P} \sim \nu}[\text{MMD}_k(\mathbb{P}^\star, \mathbb{P}_{\theta^\star(\mathbb{P})})]]$$

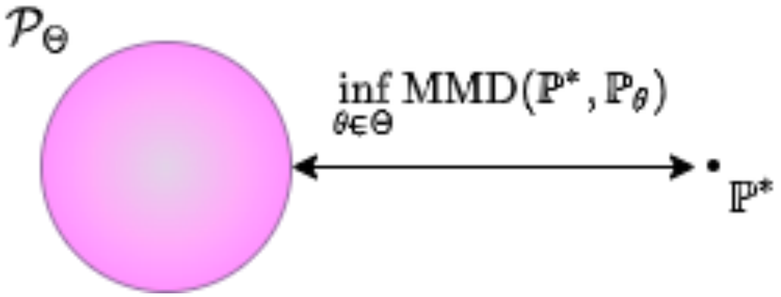


Generalisation Error

- Assumption: bounded kernel w.l.o.g. $|k(x, x')| \leq 1 \, \forall x, x' \in \mathcal{X}$
- Let $\nu := DP(\alpha', \mathbb{F}')$ (DP posterior)

$$0 \leq \overbrace{\mathbb{E}_{x_{1:n} \sim \mathbb{P}^\star}[\mathbb{E}_{\mathbb{P} \sim \nu}[\mathrm{MMD}_k(\mathbb{P}^\star, \mathbb{P}_{\theta^\star(\mathbb{P})})]] - \inf_{\theta \in \Theta} \mathrm{MMD}_k(\mathbb{P}^\star, \mathbb{P}_\theta)}^{\text{Generalisation Error}}$$

$$\leq \frac{2}{\sqrt{n}} + \frac{2}{\sqrt{\alpha + n + 2}} + \frac{4\alpha}{\alpha + n}$$



Extensions:

- Robustness to outliers for contaminated model: $\mathbb{P}^\star = (1 - \epsilon)\mathbb{P}_{\theta_0} + \epsilon\mathbb{Q}, \quad \epsilon \in (0.0.5)$
- Posterior consistency in the frequentist sense

Application 2: Measurement Error Models

[Flexibility of nonparametric prior choice]

Dellaporta, C. and Damoulas, T., 2023. Robust Bayesian Inference for Berkson and Classical Measurement Error Models. *arXiv preprint arXiv:2306.01468*.

Extension to Measurement Error Models

- Measurement Error (ME) in the covariates
- Covariate X is only observed via a noisy proxy W such that:

$$X = W + N, \quad \mathbb{E}[N] = 0$$

- Function $g_\theta : \mathcal{X} \to \mathbb{R}$ for any $\theta \in \Theta$ explains the relationship between X and Y such that:

$$Y = g_{\theta_0}(X) + E, \quad \mathbb{E}[E] = 0$$

Goal: Estimate θ_0 when only realisations of (W, Y) are available and the true ME distribution of N is potentially misspecified

Measurement Error (ME)

- We do not have observations from $\mathbb{P}^\star$ anymore!
- Target parameter: $\theta_l^\star(\mathbb{P}^\star_{X,Y}) = \arg\min_{\theta\in\Theta} \mathbb{E}_{(X,Y)\sim\mathbb{P}^\star_{X,Y}}[l(X, Y; \theta)]$
- ME uncertainty stems from the uncertainty on the true distribution of $X \mid W = w$ denoted by $\mathbb{P}^\star_{X|w}$
- DP prior for each $i = 1, \ldots, n$:

$$\mathbb{P}_i \sim DP(\alpha, \mathbb{F}_{w_i})$$

- DP posterior:

$$\mathbb{P}_i \mid w_i \sim DP(\alpha + 1, \mathbb{F}'_{w_i}), \quad \mathbb{F}'_{w_i} = \frac{1}{\alpha+1}\delta_{w_i} + \frac{\alpha}{\alpha+1}\mathbb{F}_{w_i}$$

Prior Specification

$$\mathbb{P}_i \mid w_i \sim DP(\alpha + 1, \mathbb{F}'_{w_i}), \quad \mathbb{F}'_{w_i} = \frac{1}{\alpha + 1}\delta_{w_i} + \frac{\alpha}{\alpha + 1}\mathbb{F}_{w_i}$$

- $\alpha = 0$: all the posterior centering measure mass is concentrated at the observation δ_{w_i} $\longrightarrow$ no measurement error
- $\alpha = 1$: weighting in $\mathbb{F}'_{w_i}$ is equally split between the prior centering measure $\mathbb{F}_{w_i}$ and δ_{w_i}
- $\alpha \to \infty$: no weighting on the observation w_i

Prior Specification

$$\mathbb{P}_i \mid w_i \sim DP(\alpha + 1, \mathbb{F}'_{w_i}), \quad \mathbb{F}'_{w_i} = \frac{1}{\alpha + 1}\delta_{w_i} + \frac{\alpha}{\alpha + 1}\mathbb{F}_{w_i}$$

- **Berkson** ME ($W \perp\!\!\!\perp N$)
  - $\mathbb{F}_N$ represents prior beliefs about the ME distribution
  - $x \sim \mathbb{F}_{w_i}$ is such that $x = w_i + \nu$ where $\nu \sim \mathbb{F}_N$
- **Classical** ME ($X \perp\!\!\!\perp N$)
  - $\mathbb{F}_{W|X}$ with density $f_{W|X}$ represents prior beliefs about the ME distribution, $\mathbb{F}_X$ with density f_X represents prior beliefs about the marginal distribution of the *true* covariate X
  - $x \sim \mathbb{F}_{w_i}$ is such that

$$f_{w_i}(x \mid w_i) = \frac{f_{W|X}(w_i \mid x) f_X(x)}{\int f_{W|X}(w_i \mid x) f_X(x) dx}$$

Discussion / Conclusion

- Flexibility of choosing the loss function and incorporating nonparametric prior beliefs
- Provable robustness via generalisation error guarantees
- Consistency and asymptotic normality in general and ME settings (to appear)
- Other applications: Distributionally Robust Optimisation (arXiv:2505.03585)
- Posterior Bootstrap is parallelisable but optimisation step might be costly
- Optimisation can affect posterior coverage

References

- **Bernton, E., Jacob, P.E., Gerber, M. and Robert, C.P.**, 2019. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2), pp.235–269.
- **Briol, F.X., Barp, A., Duncan, A.B. and Girolami, M.**, 2019. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*.
- **Chérief-Abdellatif, B.E. and Alquier, P.**, 2022. Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1), pp.181–213.
- **Dellaporta, C. and Damoulas, T.**, 2023. Robust Bayesian Inference for Berkson and Classical Measurement Error Models. *arXiv preprint arXiv:2306.01468*.
- **Dellaporta, C., Knoblauch, J., Damoulas, T. and Briol, F.X.**, 2022, May. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics* (pp. 943–970). PMLR.
- **Fong, E., Lyddon, S. and Holmes, C.**, 2019, May. Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *International Conference on Machine Learning* (pp. 1952–1962). PMLR.
- **Galvani, M., Bardelli, C., Figini, S. and Muliere, P.**, 2021. A Bayesian nonparametric learning approach to ensemble models using the proper Bayesian bootstrap. *Algorithms*, 14(1), p.11.
- **Goncharov, F., Barat, E. and Dautremer, T.**, 2023. Nonparametric posterior learning for emission tomography. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2), pp.452–479.
- **Lee, H., Nam, G., Fong, E. and Lee, J.**, 2024. Enhancing transfer learning with flexible nonparametric posterior sampling. *arXiv preprint arXiv:2403.07282*.
- **Lyddon, S., Walker, S. and Holmes, C.C.**, 2018. Nonparametric learning from Bayesian models with randomized objective functions. *Advances in Neural Information Processing Systems*, 31.
- **Lyddon, S.P., Holmes, C.C. and Walker, S.G.**, 2019. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2), pp.465–478.
- **Newton, M.A. and Raftery, A.E.**, 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(1), pp.3–26.
- **Nie, L. and Ročková, V.**, 2023. Deep bootstrap for Bayesian inference. *Philosophical Transactions of the Royal Society A*, 381(2247), p.20220154.
- **Ott, E. and Williamson, S.**, 2022. Nonparametric posterior normalizing flows. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*.

Contact: h.dellaporta@ucl.ac.uk

Example: Contaminated G-and-k distribution

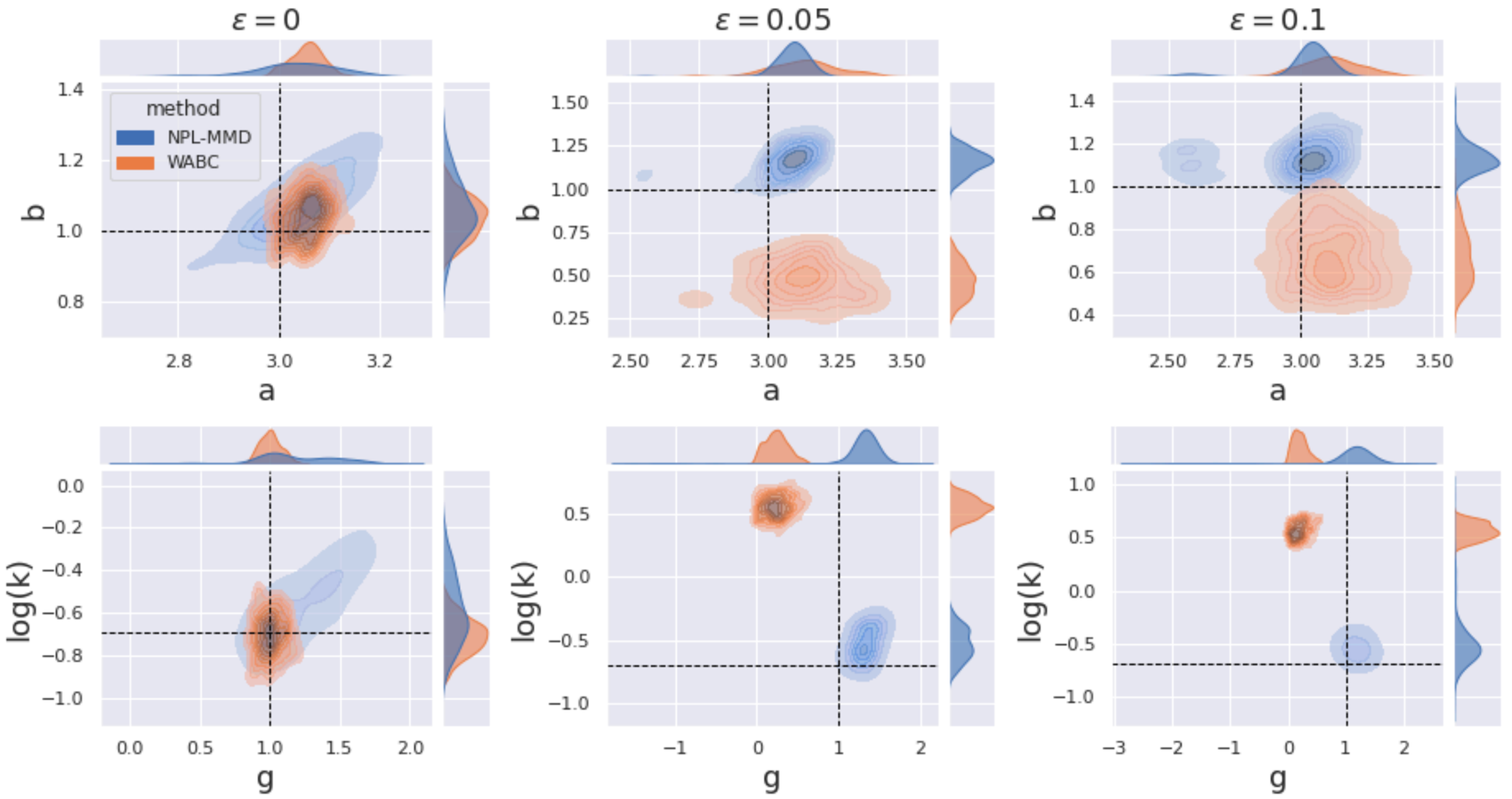
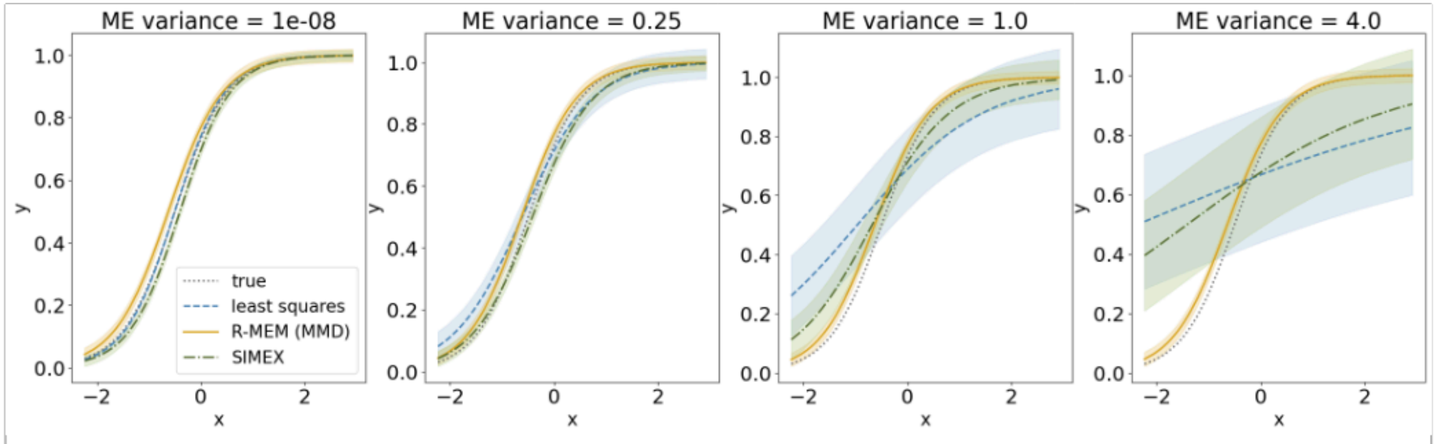
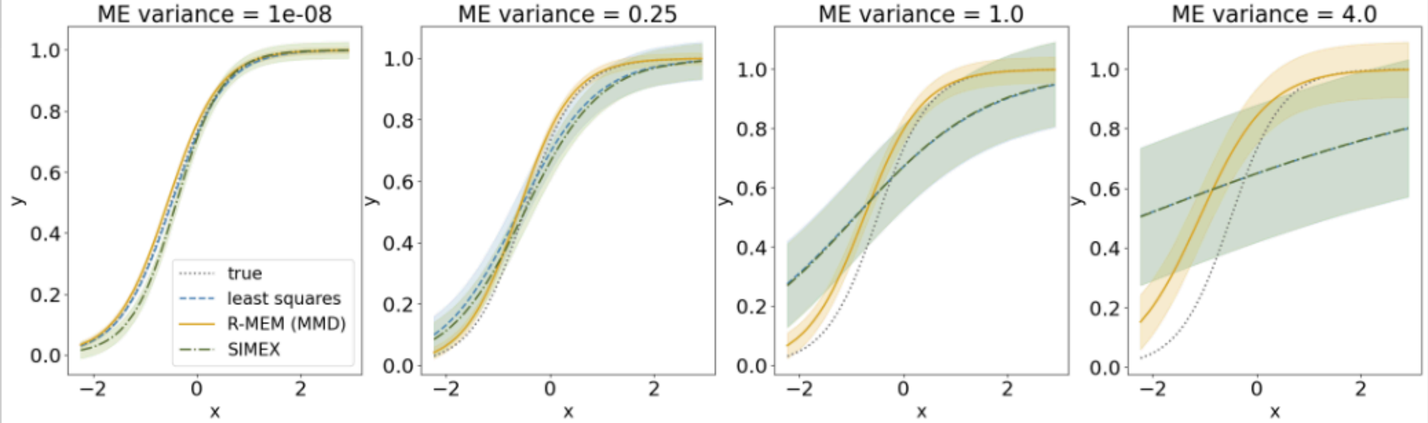


Figure: Comparison of posterior marginal distributions obtained using the MMD Posterior Bootstrap (NPL-MMD) and the Wasserstein-ABC (WABC) method in Bernton et al. (2019).

Example: Nonlinear Regression with Classical ME



(a) ME variance is correctly specified



(b) ME variance is misspecified

$$y = \frac{exp(a+bx)}{1+exp(a+bx)} + \epsilon, \quad \epsilon \sim N(0,\sigma_\epsilon^2)$$

$$w = x + \nu, \quad \nu \sim \mathcal{N}(0,\sigma_\nu^2)$$