Beyond Generalised Bayes: Prediction-Centric Alternatives

Chris. J. Oates

`https://postbayes.github.io/seminar/`

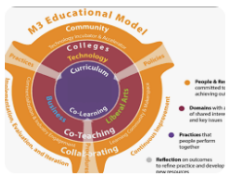
March 2025

“model” $\neq$ “statistical model”

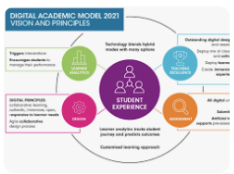
Different communities use different conventions and standards in defining a “model”:

- discrete/continuous;
- deterministic/stochastic;
- based on mathematical equations/computer simulation;
- ...

Growth Engineering
Essential Learning Models ...



ResearchGate
M3 Educational Model Overview ...



The University of Law
Learning techniques | Academic ...


St John the Baptist Cat...
St John the Baptist C...

Abu Hanifah Foundation
Holistic Educational Model – Abu ...

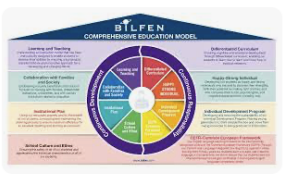
ResearchGate
The Proposed Teachin...

David White
Pedagogy, Presence and Pla...

UWC Great Britain
UWC Great Britain - UWC ...

Growth Engineering
Essential Learning Models ...

Disciplinary and Interdisciplinary Scie...
Modeling competence in science ...



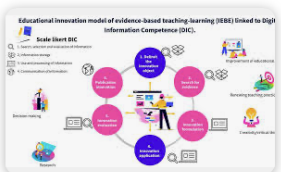
Bilfen
Bilfen Compherensive Education Model

ResearchGate
Digital education mode...

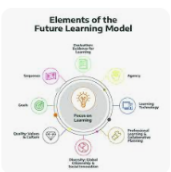

Child and Adolescent Psychiatry and Mental ...
child mental health among K-12 students...

MDPI
Sustainable Teaching Model ...

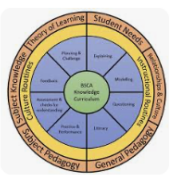
Growth Engineering
Essential Learning Mod...



MDPI
Evidence-Based Educational Innovation ...



SEK Education Group
Modelo educativo | SEK

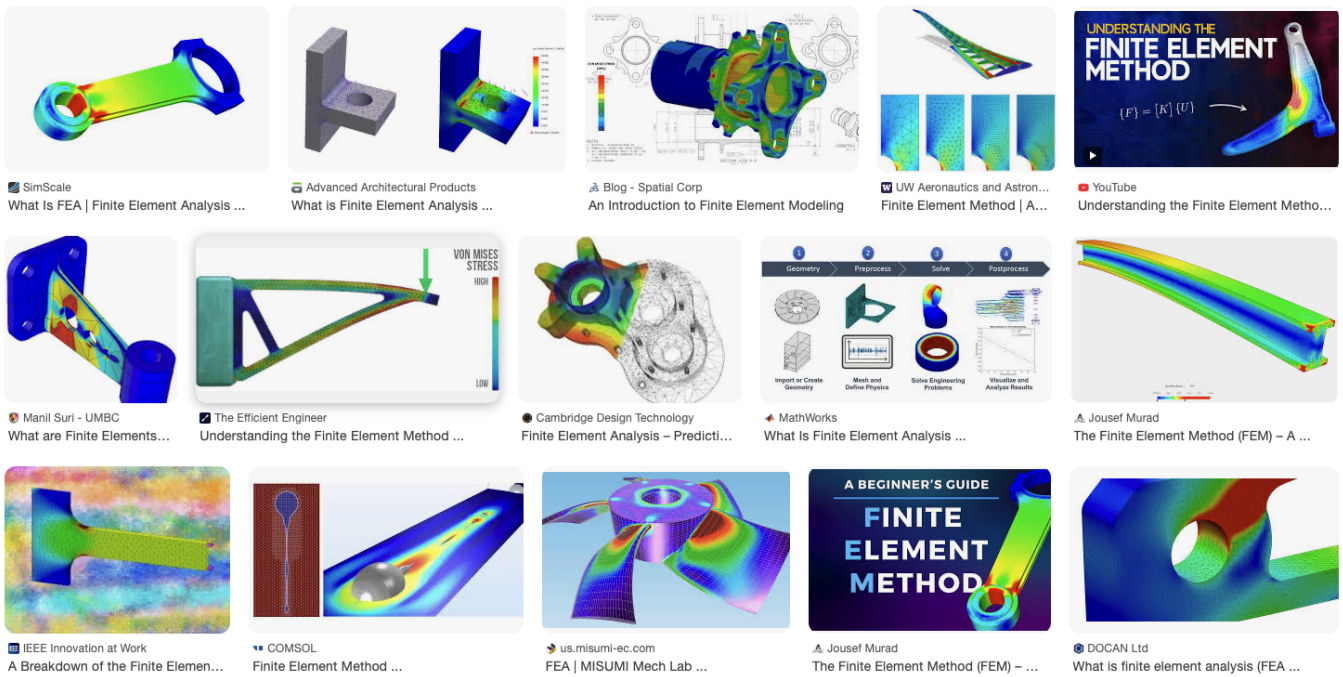


Bolton St Catherine's Aca...
Bolton St Catherine's Ac...

“model” $\neq$ “statistical model”

Different communities use different conventions and standards in defining a “model”:

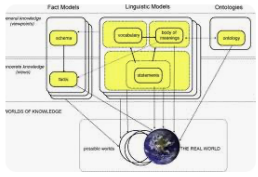
- discrete/continuous;
- deterministic/stochastic;
- based on mathematical equations/computer simulation;
- ...



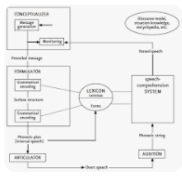
“model” $\neq$ “statistical model”

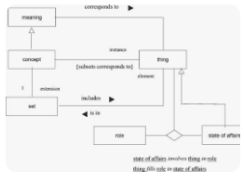
Different communities use different conventions and standards in defining a “model”:

- discrete/continuous;
- deterministic/stochastic;
- based on mathematical equations/computer simulation;
- ...



E ScienceDirect.com
Linguistic Model - an overview ...

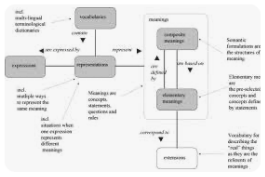

OpenEdition Journals
Model/Models in linguistics

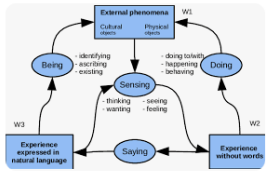


E ScienceDirect.com
Linguistic Model - an overview ...

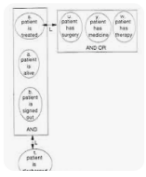
ResearchGate
Linguistic modeling appli...

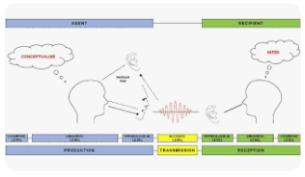
W Wikipedia
Logico-linguistic mod...

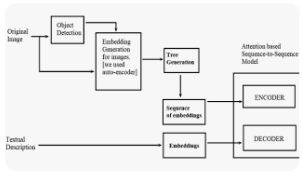

E ScienceDirect.com
Linguistic Model - an overview ...



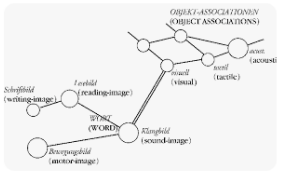
ResearchGate
Data flow diagram of a linguistic model ...


Wikipedia
Logico-linguistic m...

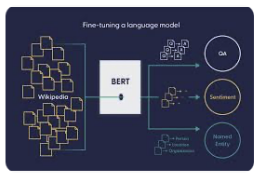

SLT info
Communication Model



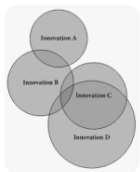
link.springer.com
Visual Linguistic Model and Its ...



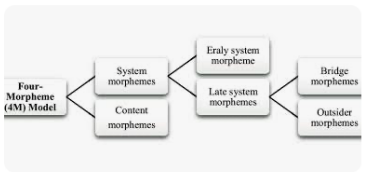
ResearchGate
1891) linguistic model: "The word ...



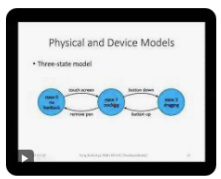
Deepset
deepset Blog | What Is a Language M...



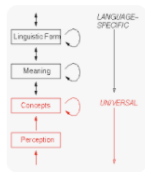
Wikipedia
Wave model - Wiki...



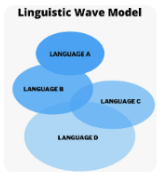
ResearchGate
4-M model for linguistic analysis ...



YouTube
Linguistic Model, Backus–Naur ...



legacy.cs.indiana.edu
Models of Languag...

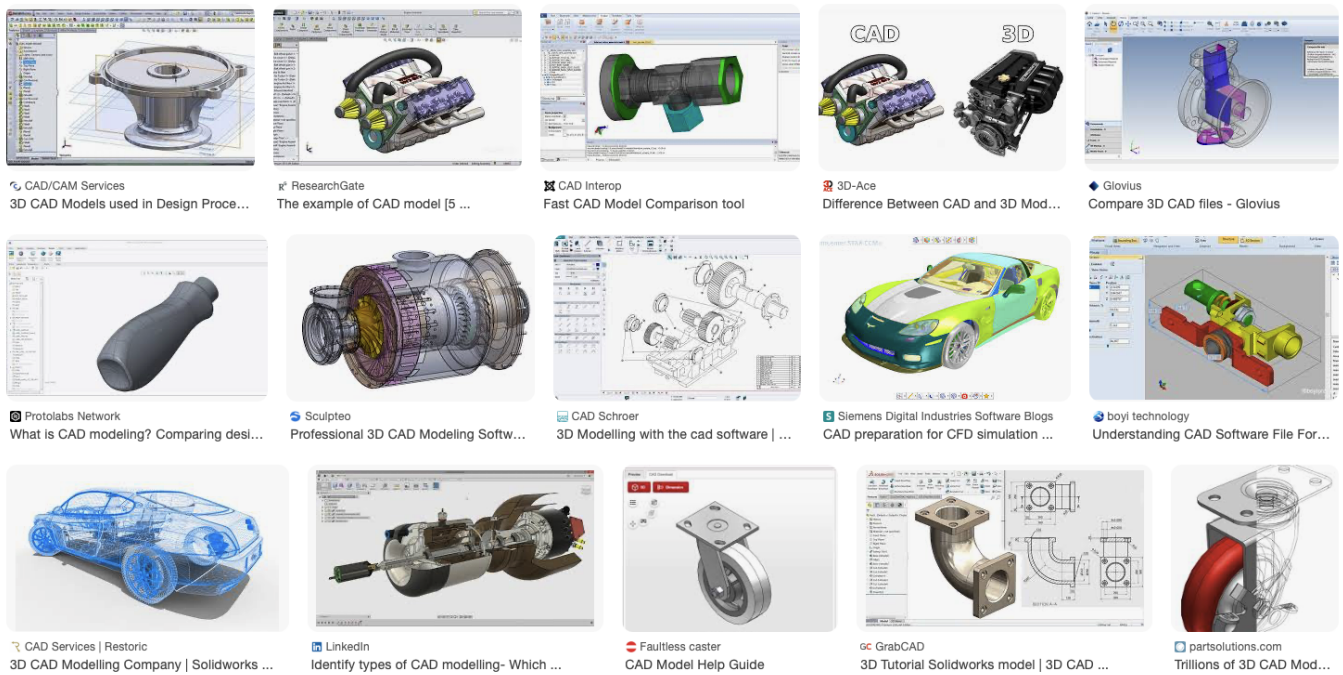


Globe Language
Wave Model - Globe ...

"model" $\neq$ "statistical model"

Different communities use different conventions and standards in defining a "model":

- discrete/continuous;
- deterministic/stochastic;
- based on mathematical equations/computer simulation;
- ...



"model" $\neq$ "statistical model"

Different communities use different conventions and standards in defining a "model":

- discrete/continuous;
- deterministic/stochastic;
- based on mathematical equations/computer simulation;
- ...

Focus on deterministic models

$$M_\theta : \mathcal{X} \to \mathcal{Y}$$

with parameters denoted $\theta \in \Theta$.

Challenge: How to use such a "model" for statistical inference and (causal) prediction?

Usual Solution: Turn the "model" M_θ into a "statistical model"

$$P_\theta : \qquad y_i = M_\theta(x_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

using knowledge of the equipment used to make the measurement.

Unfortunately a good "model" can lead to a misspecified "statistical model"...

"model" $\neq$ "statistical model"

Different communities use different conventions and standards in defining a "model":

- discrete/continuous;
- deterministic/stochastic;
- based on mathematical equations/computer simulation;
- ...

Focus on deterministic models

$$M_\theta : \mathcal{X} \to \mathcal{Y}$$

with parameters denoted $\theta \in \Theta$.

Challenge: How to use such a "model" for statistical inference and (causal) prediction?

Usual Solution: Turn the "model" M_θ into a "statistical model"

$$P_\theta : \qquad y_i = M_\theta(x_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

using knowledge of the equipment used to make the measurement.

Unfortunately a good "model" can lead to a misspecified "statistical model"...

"model" $\neq$ "statistical model"

Different communities use different conventions and standards in defining a "model":

- discrete/continuous;
- deterministic/stochastic;
- based on mathematical equations/computer simulation;
- ...

Focus on deterministic models

$$M_\theta : \mathcal{X} \to \mathcal{Y}$$

with parameters denoted $\theta \in \Theta$.

Challenge: How to use such a "model" for statistical inference and (causal) prediction?

Usual Solution: Turn the "model" M_θ into a "statistical model"

$$P_\theta : \qquad y_i = M_\theta(x_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

using knowledge of the equipment used to make the measurement.

Unfortunately a good "model" can lead to a misspecified "statistical model"...

"model" $\neq$ "statistical model"

Different communities use different conventions and standards in defining a "model":

- discrete/continuous;
- deterministic/stochastic;
- based on mathematical equations/computer simulation;
- ...

Focus on deterministic models

$$M_\theta : \mathcal{X} \to \mathcal{Y}$$

with parameters denoted $\theta \in \Theta$.

Challenge: How to use such a "model" for statistical inference and (causal) prediction?

Usual Solution: Turn the "model" M_θ into a "statistical model"

$$P_\theta : \qquad y_i = M_\theta(x_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

using knowledge of the equipment used to make the measurement.

Unfortunately a good "model" can lead to a misspecified "statistical model"...

"model" $\neq$ "statistical model"

Different communities use different conventions and standards in defining a "model":

- discrete/continuous;
- deterministic/stochastic;
- based on mathematical equations/computer simulation;
- ...

Focus on deterministic models

$$M_\theta : \mathcal{X} \to \mathcal{Y}$$

with parameters denoted $\theta \in \Theta$.

Challenge: How to use such a "model" for statistical inference and (causal) prediction?

Usual Solution: Turn the "model" M_θ into a "statistical model"

$$P_\theta : \qquad y_i = M_\theta(x_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

using knowledge of the equipment used to make the measurement.

Unfortunately a good "model" can lead to a misspecified "statistical model"...

Model Misspecification in Cell Signalling

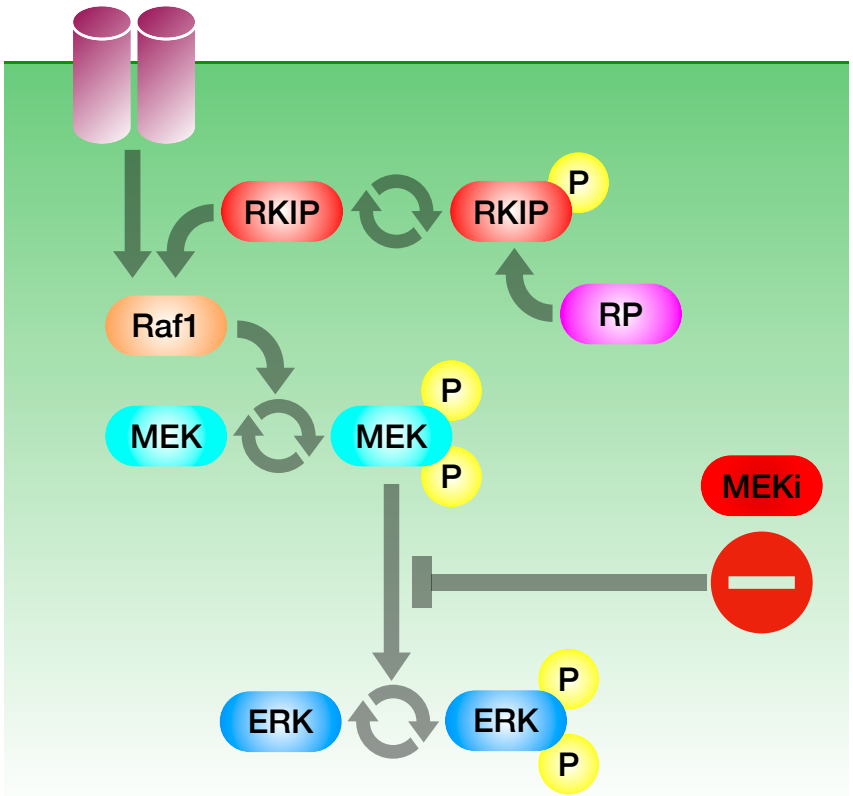


Figure: ERK signalling model.

Systems biology has invested decades of effort into the design of detailed ODE descriptions of cellular signalling pathways, with thousands of models hosted on repositories such as `BioModels` [Malik-Sheriff et al., 2020].

e.g. ERK signalling is modelled as

$$\frac{\mathrm{d}u}{\mathrm{d}x} = f_\theta(x, u), \qquad \theta \in \mathbb{R}^{11}$$

Data are (reasonably, as far as this talk is concerned) treated as noisy observations of molecular concentrations $u(x)$ at discrete times $x_{1:n}$.

Model Misspecification in Cell Signalling

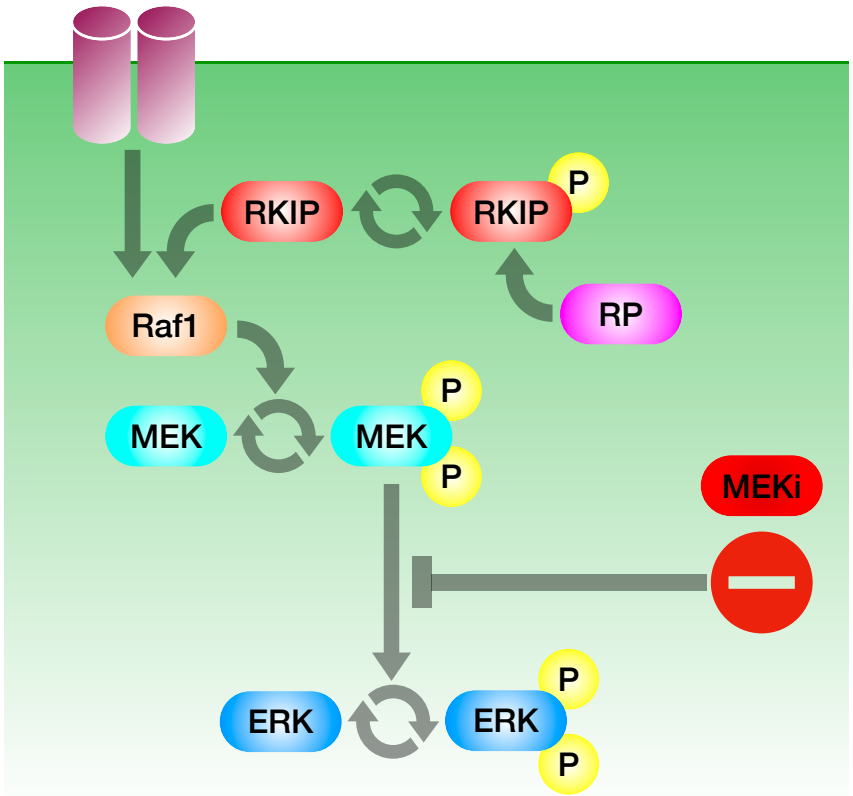


Figure: ERK signalling model.

Systems biology has invested decades of effort into the design of detailed ODE descriptions of cellular signalling pathways, with thousands of models hosted on repositories such as `BioModels` [Malik-Sheriff et al., 2020].

e.g. ERK signalling is modelled as

$$\frac{\mathrm{d}u}{\mathrm{d}x} = f_\theta(x, u), \qquad \theta \in \mathbb{R}^{11}$$

Data are (reasonably, as far as this talk is concerned) treated as noisy observations of molecular concentrations $u(x)$ at discrete times $x_{1:n}$.

Model Misspecification in Cell Signalling

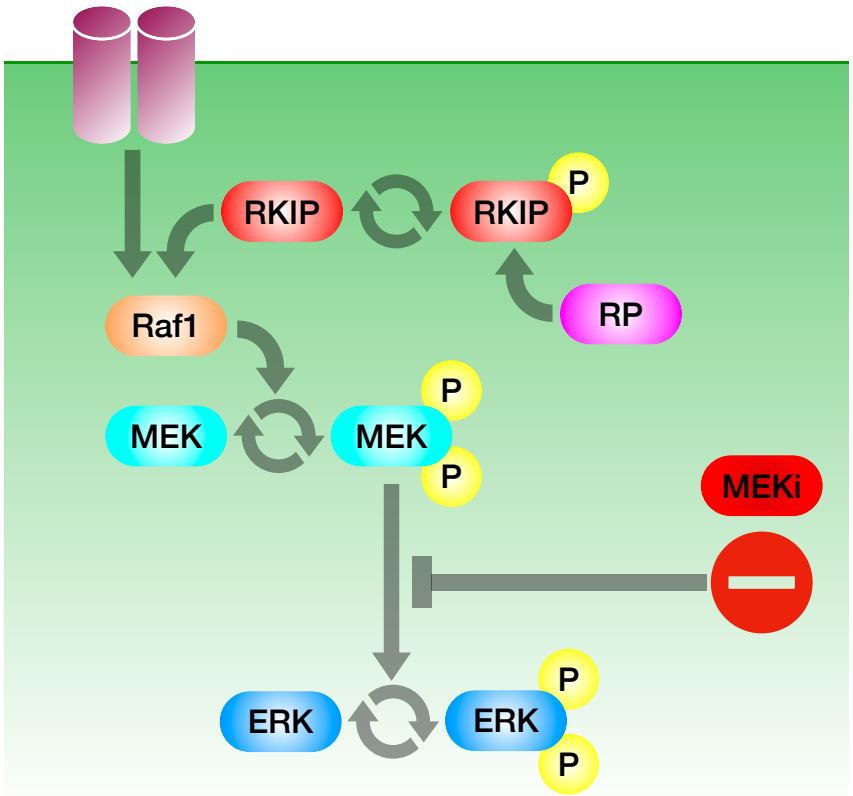


Figure: ERK signalling model.

Systems biology has invested decades of effort into the design of detailed ODE descriptions of cellular signalling pathways, with thousands of models hosted on repositories such as `BioModels` [Malik-Sheriff et al., 2020].

e.g. ERK signalling is modelled as

$$\frac{\mathrm{d}u}{\mathrm{d}x} = f_\theta(x, u), \qquad \theta \in \mathbb{R}^{11}$$

Data are (reasonably, as far as this talk is concerned) treated as noisy observations of molecular concentrations $u(x)$ at discrete times $x_{1:n}$.

Model Misspecification in Cell Signalling

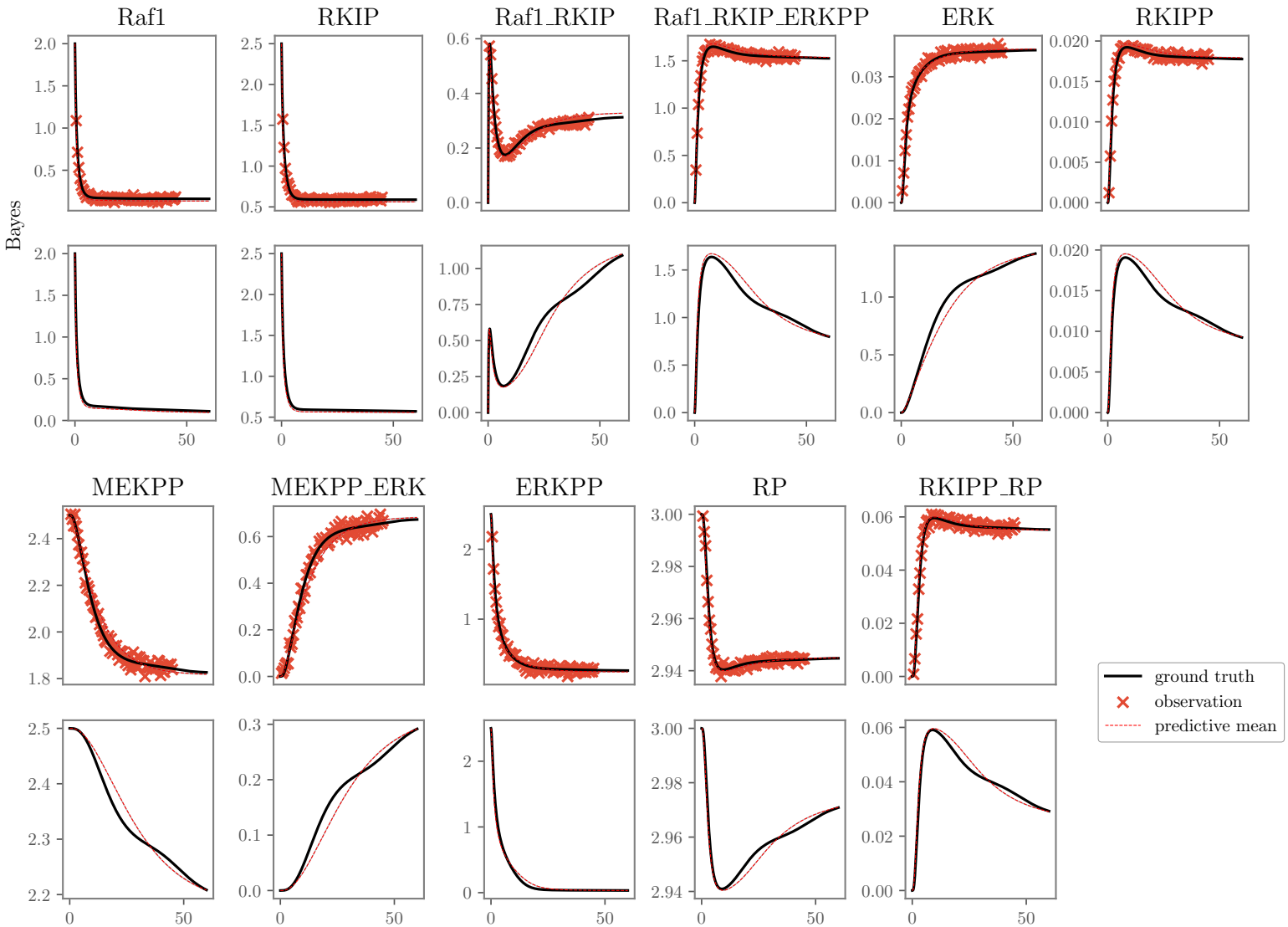


Figure: Posterior predictive for the ERK signalling model.

Model Misspecification in Cell Signalling

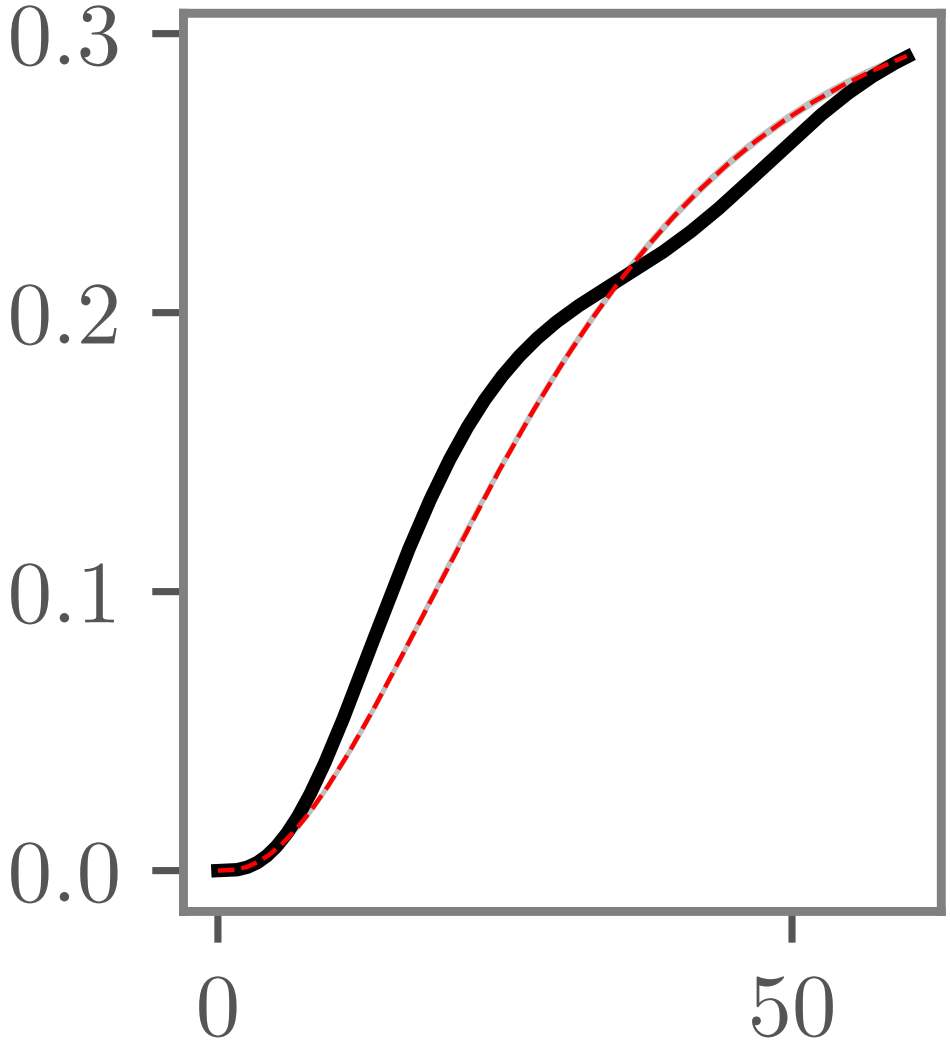


Figure: Posterior predictive for the ERK signalling model.

Model Misspecification in Cell Signalling

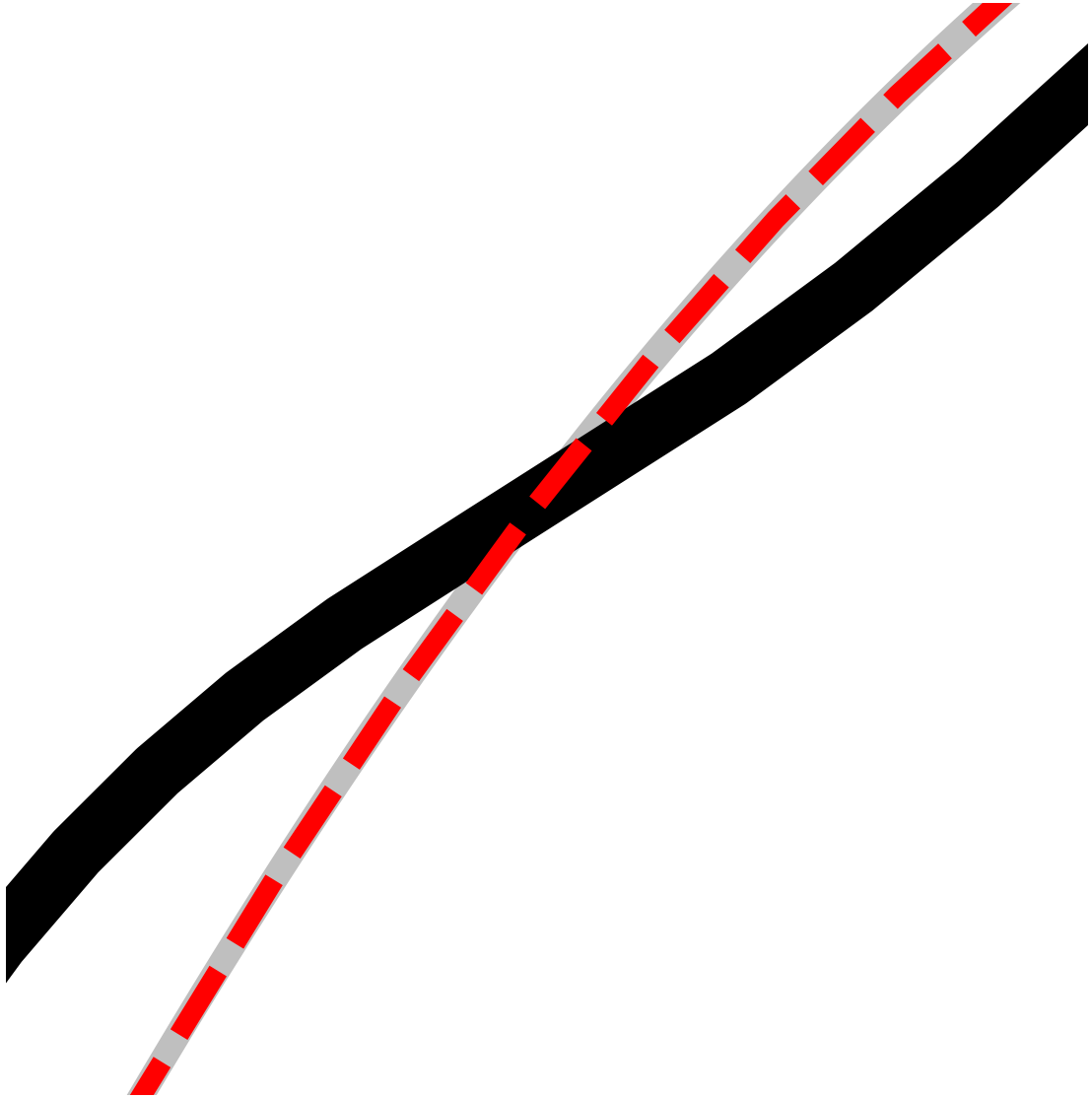


Figure: Posterior predictive for the ERK signalling model.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- **causal prediction impossible in this framework**.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- **causal prediction impossible in this framework**.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- **causal prediction impossible in this framework**.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- **causal prediction impossible in this framework**.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- **causal prediction impossible in this framework**.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- **causal prediction impossible in this framework**.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- **causal prediction impossible in this framework**.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- causal prediction impossible in this framework.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- causal prediction impossible in this framework.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- causal prediction impossible in this framework.

TL/DR: Better Bayesian Inference Does Not Help

Bayesian inference for misspecified models has been widely studied.

e.g. Kennedy and O'Hagan [2001]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- misspecified model $M_\theta : \mathcal{X} \to \mathcal{Y}$
- the residual $R : \mathcal{X} \to \mathcal{Y}$ (difference between real world and model)
- prior for the residual, e.g. $R \sim \mathcal{GP}$
- augmented statistical model, e.g.

$$y_i = \underbrace{M_\theta(x_i)}_{\text{"model"}} + \underbrace{R(x_i)}_{\text{residual}} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Limitations

- high data requirement to learn the residual R;
- **causal prediction impossible in this framework**.

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathsf{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Recovering Standard Bayes
Standard Bayes has $\mathsf{L}_n(\theta, y_{1:n}) = -\log p_\theta(y_{1:n})$ and $\lambda_n = 1$.

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathsf{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Recovering Standard Bayes
Standard Bayes has $\mathsf{L}_n(\theta, y_{1:n}) = -\log p_\theta(y_{1:n})$ and $\lambda_n = 1$.

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathsf{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Recovering Standard Bayes
Standard Bayes has $\mathsf{L}_n(\theta, y_{1:n}) = -\log p_\theta(y_{1:n})$ and $\lambda_n = 1$.

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Recovering Standard Bayes
Standard Bayes has $\mathsf{L}_n(\theta, y_{1:n}) = -\log p_\theta(y_{1:n})$ and $\lambda_n = 1$.

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Recovering Standard Bayes
Standard Bayes has $\mathsf{L}_n(\theta, y_{1:n}) = -\log p_\theta(y_{1:n})$ and $\lambda_n = 1$.

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Recovering Standard Bayes
Standard Bayes has $\mathsf{L}_n(\theta, y_{1:n}) = -\log p_\theta(y_{1:n})$ and $\lambda_n = 1$.

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathsf{KL}(Q, Q_0)}_{\text{regularisation}} .$$

Recovering Standard Bayes
Standard Bayes has $\mathsf{L}_n(\theta, y_{1:n}) = -\log p_\theta(y_{1:n})$ and $\lambda_n = 1$.

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathsf{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Recovering Standard Bayes
Standard Bayes has $\mathsf{L}_n(\theta, y_{1:n}) = -\log p_\theta(y_{1:n})$ and $\lambda_n = 1$.

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \underset{Q \in \mathcal{P}(\Theta)}{\arg\min} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathsf{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Other Choices of L_n and λ_n
For various other choices of L_n and λ_n, generalised Bayesian methods can produce robust posteriors suitable for dealing with certain forms of statistical model misspecification [see Hooker and Vidyashankar, 2014, Ghosh and Basu, 2016, Knoblauch et al., 2018, Schmon et al., 2020, Chérief-Abdellatif and Alquier, 2020, Dellaporta et al., 2022, Husain and Knoblauch, 2022, Altamirano et al., 2023, 2024, Duran-Martin et al., 2024].

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Concentration of Generalised Posterior
For convenient choices of L_n and λ_n, the magnitude of the data-fit term generally increases with n. As a result, $Q_n^\dagger \to \delta_{\theta^\dagger}$ [Miller, 2021].

So need to tune the learning rate...

TL/DR: Generalised Bayesian Inference is a Bit Complicated

Generalisations of Bayesian inference have been proposed for when the model is misspecified.

e.g. generalised Bayesian inference [Bissiri et al., 2016, Knoblauch et al., 2022]:

- parameter $\theta \in \Theta$
- IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- loss function $\mathsf{L}_n : \Theta \times \mathcal{Y}^n \to \mathbb{R}$
- prior $Q_0 \in \mathcal{P}(\Theta)$
- generalised posterior

$$Q_n^\dagger = \underset{Q \in \mathcal{P}(\Theta)}{\arg\min} \underbrace{\int \mathsf{L}_n(\theta, y_{1:n}) \, \mathrm{d}Q(\theta)}_{\text{average data fit}} + \underbrace{\lambda_n \, \mathsf{KL}(Q, Q_0)}_{\text{regularisation}}.$$

Tuning the Learning Rate
Several ideas have been proposed to select the learning rate λ_n - c.f. Ryan Martin's talk. But these involve approximations and/or can be computationally demanding.

Tuning the learning rate is complicated - seek alternative to generalised Bayes...?

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Nonparametric Maximum Likelihood
This approach solves

$$\arg\min_{Q \in \mathcal{P}(\Theta)} -\frac{1}{n} \sum_{i=1}^{n} \log p_Q(y_i)$$

where p_Q is a density for the mixture model P_Q [see Chapter 5 of Lindsay, 1995].

- approximates $\mathrm{KL}(P_\star, P_Q)$ when $y_{1:n}$ is a collection of n independent samples from $P_\star \in \mathcal{P}(\mathcal{Y})$
- lack of regularisation causes computational difficulties and non-identifiability [see e.g. Laird, 1978], as the minimising measure will generally be fully atomic, see Lindsay [1995, e.g. Theorem 21 in Chapter 5] and Jordan-Squire [2015].

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Nonparametric Maximum Likelihood
This approach solves

$$\arg\min_{Q \in \mathcal{P}(\Theta)} -\frac{1}{n} \sum_{i=1}^{n} \log p_Q(y_i)$$

where p_Q is a density for the mixture model P_Q [see Chapter 5 of Lindsay, 1995].

- approximates $\mathrm{KL}(P_\star, P_Q)$ when $y_{1:n}$ is a collection of n independent samples from $P_\star \in \mathcal{P}(\mathcal{Y})$
- lack of regularisation causes computational difficulties and non-identifiability [see e.g. Laird, 1978], as the minimising measure will generally be fully atomic, see Lindsay [1995, e.g. Theorem 21 in Chapter 5] and Jordan-Squire [2015].

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Nonparametric Maximum Likelihood
This approach solves

$$\arg\min_{Q \in \mathcal{P}(\Theta)} -\frac{1}{n} \sum_{i=1}^{n} \log p_Q(y_i)$$

where p_Q is a density for the mixture model P_Q [see Chapter 5 of Lindsay, 1995].

- approximates $\mathrm{KL}(P_\star, P_Q)$ when $y_{1:n}$ is a collection of n independent samples from $P_\star \in \mathcal{P}(\mathcal{Y})$
- lack of regularisation causes computational difficulties and non-identifiability [see e.g. Laird, 1978], as the minimising measure will generally be fully atomic, see Lindsay [1995, e.g. Theorem 21 in Chapter 5] and Jordan-Squire [2015].

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Nonparametric Maximum Likelihood
This approach solves

$$\arg\min_{Q \in \mathcal{P}(\Theta)} -\frac{1}{n} \sum_{i=1}^{n} \log p_Q(y_i)$$

where p_Q is a density for the mixture model P_Q [see Chapter 5 of Lindsay, 1995].

- approximates $\mathrm{KL}(P_\star, P_Q)$ when $y_{1:n}$ is a collection of n independent samples from $P_\star \in \mathcal{P}(\mathcal{Y})$
- lack of regularisation causes computational difficulties and non-identifiability [see e.g. Laird, 1978], as the minimising measure will generally be fully atomic, see Lindsay [1995, e.g. Theorem 21 in Chapter 5] and Jordan-Squire [2015].

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Nonparametric Maximum Likelihood
This approach solves

$$\arg\min_{Q \in \mathcal{P}(\Theta)} -\frac{1}{n} \sum_{i=1}^{n} \log p_Q(y_i)$$

where p_Q is a density for the mixture model P_Q [see Chapter 5 of Lindsay, 1995].

- approximates $\mathrm{KL}(P_\star, P_Q)$ when $y_{1:n}$ is a collection of n independent samples from $P_\star \in \mathcal{P}(\mathcal{Y})$
- lack of regularisation causes computational difficulties and non-identifiability [see e.g. Laird, 1978], as the minimising measure will generally be fully atomic, see Lindsay [1995, e.g. Theorem 21 in Chapter 5] and Jordan-Squire [2015].

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Nonparametric Maximum Likelihood
This approach solves

$$\arg\min_{Q \in \mathcal{P}(\Theta)} -\frac{1}{n} \sum_{i=1}^{n} \log p_Q(y_i)$$

where p_Q is a density for the mixture model P_Q [see Chapter 5 of Lindsay, 1995].

- approximates $\mathrm{KL}(P_\star, P_Q)$ when $y_{1:n}$ is a collection of n independent samples from $P_\star \in \mathcal{P}(\mathcal{Y})$
- lack of regularisation causes computational difficulties and non-identifiability [see e.g. Laird, 1978], as the minimising measure will generally be fully atomic, see Lindsay [1995, e.g. Theorem 21 in Chapter 5] and Jordan-Squire [2015].

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Regularised Nonparametric Maximum Likelihood
Jankowiak et al. [2020b,a], Sheth and Khardon [2020] studied

$$\arg\min_{Q \in \mathcal{Q}} -\frac{1}{n} \sum_{i=1}^{n} \log\left(p_Q(y_i)^\alpha\right) + \lambda_n \, \mathrm{KL}(Q, Q_0).$$

- the choice $\lambda_n = \frac{\alpha}{n}$ can be linked to approximation of the standard Bayesian posterior via α-divergences Li and Gal [2017], Villacampa-Calvo and Hernandez-Lobato [2020]
- considered in the context of Gaussian processes and deep Gaussian processes with $\alpha = 1$ and various choices for λ_n Jankowiak et al. [2020a,b], Sheth and Khardon [2020].

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Regularised Nonparametric Maximum Likelihood
Jankowiak et al. [2020b,a], Sheth and Khardon [2020] studied

$$\arg\min_{Q \in \mathcal{Q}} -\frac{1}{n} \sum_{i=1}^{n} \log\left(p_Q(y_i)^\alpha\right) + \lambda_n \, \mathrm{KL}(Q, Q_0).$$

- the choice $\lambda_n = \frac{\alpha}{n}$ can be linked to approximation of the standard Bayesian posterior via α-divergences Li and Gal [2017], Villacampa-Calvo and Hernandez-Lobato [2020]
- considered in the context of Gaussian processes and deep Gaussian processes with $\alpha = 1$ and various choices for λ_n Jankowiak et al. [2020a,b], Sheth and Khardon [2020].

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Regularised Nonparametric Maximum Likelihood
Jankowiak et al. [2020b,a], Sheth and Khardon [2020] studied

$$\arg\min_{Q \in \mathcal{Q}} -\frac{1}{n} \sum_{i=1}^{n} \log\left(p_Q(y_i)^\alpha\right) + \lambda_n \, \mathrm{KL}(Q, Q_0).$$

- the choice $\lambda_n = \frac{\alpha}{n}$ can be linked to approximation of the standard Bayesian posterior via α-divergences Li and Gal [2017], Villacampa-Calvo and Hernandez-Lobato [2020]
- considered in the context of Gaussian processes and deep Gaussian processes with $\alpha = 1$ and various choices for λ_n Jankowiak et al. [2020a,b], Sheth and Khardon [2020].

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Predictive Variational Inference
Lai and Yao [2024] considered

$$\arg\min_{Q \in \mathcal{Q}} \underbrace{-\frac{1}{n}\sum_{i=1}^{n} \log p_Q(y_i)}_{\text{predictive fit}} + \underbrace{\lambda_n \mathrm{KL}(Q, Q_n^\dagger)}_{\text{regularisation}}$$

where

- predictive fit assessed using log-predictive density (or any proper scoring rule)
- regularisation is toward the standard Bayesian posterior $Q_n^\dagger$
- for computation, parametric VI is used.

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Predictive Variational Inference
Lai and Yao [2024] considered

$$\arg\min_{Q \in \mathcal{Q}} \underbrace{-\frac{1}{n}\sum_{i=1}^{n} \log p_Q(y_i)}_{\text{predictive fit}} + \underbrace{\lambda_n \mathrm{KL}(Q, Q_n^\dagger)}_{\text{regularisation}}$$

where

- predictive fit assessed using log-predictive density (or any proper scoring rule)
- regularisation is toward the standard Bayesian posterior $Q_n^\dagger$
- for computation, parametric VI is used.

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Predictive Variational Inference
Lai and Yao [2024] considered

$$\arg\min_{Q \in \mathcal{Q}} \underbrace{-\frac{1}{n}\sum_{i=1}^{n} \log p_Q(y_i)}_{\text{predictive fit}} + \underbrace{\lambda_n \mathrm{KL}(Q, Q_n^\dagger)}_{\text{regularisation}}$$

where

- predictive fit assessed using log-predictive density (or any proper scoring rule)
- regularisation is toward the standard Bayesian posterior $Q_n^\dagger$
- for computation, parametric VI is used.

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Predictive Variational Inference
Lai and Yao [2024] considered

$$\arg\min_{Q \in \mathcal{Q}} \underbrace{-\frac{1}{n}\sum_{i=1}^{n} \log p_Q(y_i)}_{\text{predictive fit}} + \underbrace{\lambda_n \mathrm{KL}(Q, Q_n^\dagger)}_{\text{regularisation}}$$

where

- predictive fit assessed using log-predictive density (or any proper scoring rule)
- regularisation is toward the standard Bayesian posterior $Q_n^\dagger$
- for computation, parametric VI is used.

Prediction-Centric Alternatives

Setting Given a model P_θ that is useful (e.g. for causal prediction) but misspecified.

Step 1: Mitigate Misspecification Form a mixture model

$$P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y}).$$

Step 2: Learn Q For example, by matching the predictive distribution of P_Q to the dataset.

Example: Predictive Variational Inference
Lai and Yao [2024] considered

$$\arg\min_{Q \in \mathcal{Q}} \underbrace{-\frac{1}{n}\sum_{i=1}^{n} \log p_Q(y_i)}_{\text{predictive fit}} + \underbrace{\lambda_n \mathrm{KL}(Q, Q_n^\dagger)}_{\text{regularisation}}$$

where

- predictive fit assessed using log-predictive density (or any proper scoring rule)
- regularisation is toward the standard Bayesian posterior $Q_n^\dagger$
- for computation, parametric VI is used.

Prediction-Centric Uncertainty Quantification

Our Take: *Prediction-Centric Uncertainty Quantification* (PCUQ).

Joint work with Zheyang Shen (Newcastle), Jeremias Knoblauch (UCL), and Sam Power (Bristol)

- (for now) IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- empirical measure of the dataset $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$
- mixture model $P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y})$
- prediction-centric posterior

$$Q_n = \underset{Q \in \mathcal{P}(\Theta)}{\arg\min} \; \frac{1}{2} \underbrace{\mathrm{MMD}^2(P_n, P_Q)}_{\text{predictive fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}$$

Measuring Predictive Fit

The use of *maximum mean discrepancy* (MMD) [see e.g. Gretton et al., 2012], as opposed to other statistical divergences, confers outlier-robustness to PCUQ, which may be valuable in the misspecified context, and carries computational advantages, enabling the use of powerful emerging sampling methods based on gradient flows [Wild et al., 2023].

Prediction-Centric Uncertainty Quantification

Our Take: *Prediction-Centric Uncertainty Quantification* (PCUQ).

Joint work with Zheyang Shen (Newcastle), Jeremias Knoblauch (UCL), and Sam Power (Bristol)

- (for now) IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- empirical measure of the dataset $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$
- mixture model $P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y})$
- prediction-centric posterior

$$Q_n = \underset{Q \in \mathcal{P}(\Theta)}{\arg\min} \; \frac{1}{2} \underbrace{\mathrm{MMD}^2(P_n, P_Q)}_{\text{predictive fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}$$

Measuring Predictive Fit

The use of *maximum mean discrepancy* (MMD) [see e.g. Gretton et al., 2012], as opposed to other statistical divergences, confers outlier-robustness to PCUQ, which may be valuable in the misspecified context, and carries computational advantages, enabling the use of powerful emerging sampling methods based on gradient flows [Wild et al., 2023].

Prediction-Centric Uncertainty Quantification

Our Take: *Prediction-Centric Uncertainty Quantification* (PCUQ).

Joint work with Zheyang Shen (Newcastle), Jeremias Knoblauch (UCL), and Sam Power (Bristol)

- (for now) IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- empirical measure of the dataset $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$
- mixture model $P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y})$
- prediction-centric posterior

$$Q_n = \underset{Q \in \mathcal{P}(\Theta)}{\arg\min} \; \frac{1}{2} \underbrace{\mathrm{MMD}^2(P_n, P_Q)}_{\text{predictive fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}$$

Measuring Predictive Fit

The use of *maximum mean discrepancy* (MMD) [see e.g. Gretton et al., 2012], as opposed to other statistical divergences, confers outlier-robustness to PCUQ, which may be valuable in the misspecified context, and carries computational advantages, enabling the use of powerful emerging sampling methods based on gradient flows [Wild et al., 2023].

Prediction-Centric Uncertainty Quantification

Our Take: *Prediction-Centric Uncertainty Quantification* (PCUQ).

Joint work with Zheyang Shen (Newcastle), Jeremias Knoblauch (UCL), and Sam Power (Bristol)

- (for now) IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- empirical measure of the dataset $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$
- mixture model $P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y})$
- prediction-centric posterior

$$Q_n = \underset{Q \in \mathcal{P}(\Theta)}{\arg\min} \; \frac{1}{2} \underbrace{\mathrm{MMD}^2(P_n, P_Q)}_{\text{predictive fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}$$

Measuring Predictive Fit

The use of *maximum mean discrepancy* (MMD) [see e.g. Gretton et al., 2012], as opposed to other statistical divergences, confers outlier-robustness to PCUQ, which may be valuable in the misspecified context, and carries computational advantages, enabling the use of powerful emerging sampling methods based on gradient flows [Wild et al., 2023].

Prediction-Centric Uncertainty Quantification

Our Take: *Prediction-Centric Uncertainty Quantification* (PCUQ).

Joint work with Zheyang Shen (Newcastle), Jeremias Knoblauch (UCL), and Sam Power (Bristol)

- (for now) IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- empirical measure of the dataset $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$
- mixture model $P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y})$
- prediction-centric posterior

$$Q_n = \underset{Q \in \mathcal{P}(\Theta)}{\arg\min} \; \frac{1}{2} \underbrace{\mathrm{MMD}^2(P_n, P_Q)}_{\text{predictive fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}$$

Measuring Predictive Fit

The use of *maximum mean discrepancy* (MMD) [see e.g. Gretton et al., 2012], as opposed to other statistical divergences, confers outlier-robustness to PCUQ, which may be valuable in the misspecified context, and carries computational advantages, enabling the use of powerful emerging sampling methods based on gradient flows [Wild et al., 2023].

Prediction-Centric Uncertainty Quantification

Our Take: *Prediction-Centric Uncertainty Quantification* (PCUQ).

Joint work with Zheyang Shen (Newcastle), Jeremias Knoblauch (UCL), and Sam Power (Bristol)

- (for now) IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- empirical measure of the dataset $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$
- mixture model $P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y})$
- prediction-centric posterior

$$Q_n = \underset{Q \in \mathcal{P}(\Theta)}{\arg\min} \; \frac{1}{2} \underbrace{\mathrm{MMD}^2(P_n, P_Q)}_{\text{predictive fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}$$

Measuring Predictive Fit

The use of *maximum mean discrepancy* (MMD) [see e.g. Gretton et al., 2012], as opposed to other statistical divergences, confers outlier-robustness to PCUQ, which may be valuable in the misspecified context, and carries computational advantages, enabling the use of powerful emerging sampling methods based on gradient flows [Wild et al., 2023].

Prediction-Centric Uncertainty Quantification

Our Take: *Prediction-Centric Uncertainty Quantification* (PCUQ).

Joint work with Zheyang Shen (Newcastle), Jeremias Knoblauch (UCL), and Sam Power (Bristol)

- (for now) IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- empirical measure of the dataset $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$
- mixture model $P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y})$
- prediction-centric posterior

$$Q_n = \underset{Q \in \mathcal{P}(\Theta)}{\arg\min} \; \frac{1}{2} \underbrace{\mathrm{MMD}^2(P_n, P_Q)}_{\text{predictive fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}$$

Regularisation Target

Q_0 acts on Q_n in essentially the same way that Q_0 acts on Gibbs measures like $Q_n^\dagger$, as a reference measure in a Radon–Nikodym derivative [Bissiri et al., 2016, Knoblauch et al., 2022]. That is, once can reason about 'updating belief distributions' using PCUQ.

Prediction-Centric Uncertainty Quantification

Our Take: *Prediction-Centric Uncertainty Quantification* (PCUQ).

Joint work with Zheyang Shen (Newcastle), Jeremias Knoblauch (UCL), and Sam Power (Bristol)

- (for now) IID data $y_{1:n} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$
- empirical measure of the dataset $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$
- mixture model $P_Q = \int P_\theta \, \mathrm{d}Q(\theta) \in \mathcal{P}(\mathcal{Y})$
- prediction-centric posterior

$$Q_n = \operatorname*{arg\,min}_{Q \in \mathcal{P}(\Theta)} \frac{1}{2} \underbrace{\mathrm{MMD}^2(P_n, P_Q)}_{\text{predictive fit}} + \underbrace{\lambda_n \, \mathrm{KL}(Q, Q_0)}_{\text{regularisation}}$$

Learning Rate

Compared to generalised Bayes, **PCUQ depends less critically on the learning rate** λ_n. i.e. support of Q_n is not a singleton set when $\lambda_n \to 0$.

Model Misspecification in Cell Signalling

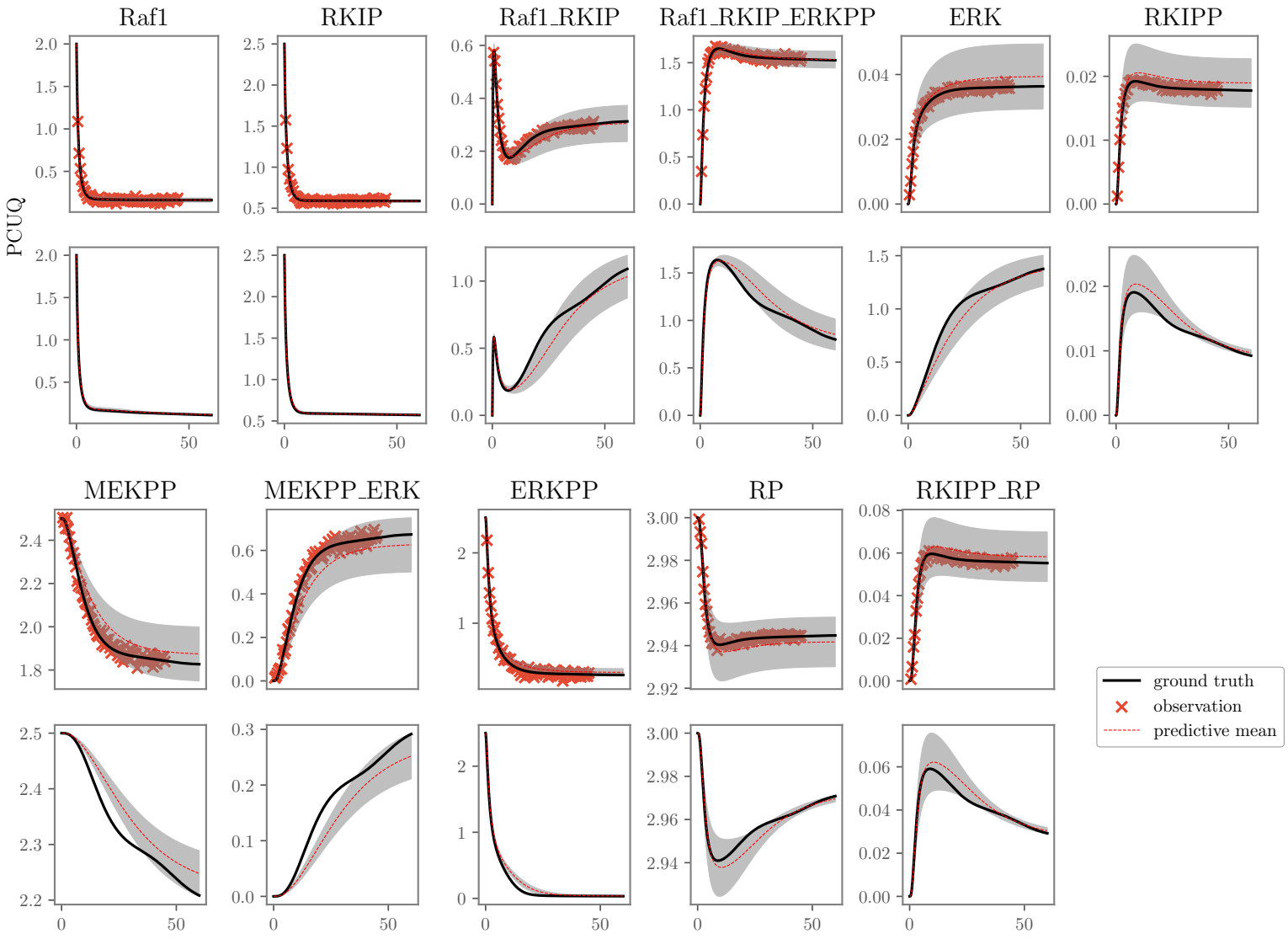


Figure: PCUQ predictive for the ERK signalling model.

Model Misspecification in Cell Signalling

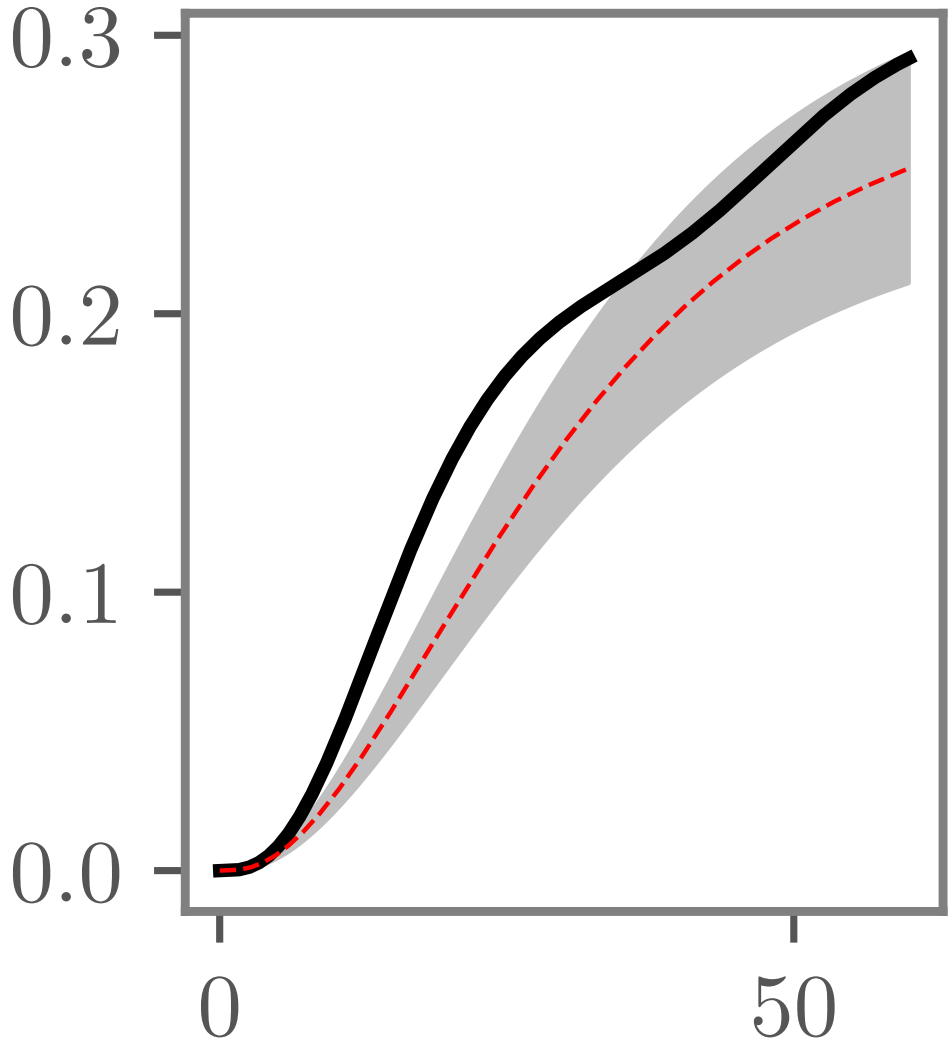


Figure: PCUQ predictive for the ERK signalling model.

A Bit More Detail

Predictive Fit via MMD

To define the *predictive fit* for PCUQ, we need:

- kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, e.g. $k(y, y') = (yy') + (yy')^2$
- *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}(k)$ [see Berlinet and Thomas-Agnan, 2011, for background]
- *kernel mean embedding*

$$\mu_k(P) := \int k(\cdot, y) \, \mathrm{d}P(y) \in \mathcal{H}(k).$$

The divergence of a candidate $P \in \mathcal{P}(\mathcal{Y})$ from the data-generating distribution $P_\star$ can be quantified using *maximum mean discrepancy* (MMD):

$$\begin{aligned} \mathrm{MMD}(P_\star, P) &= \|\mu_k(P_\star) - \mu_k(P)\|_{\mathcal{H}(k)} \\ &\overset{\text{e.g.}}{=} \left\| \begin{bmatrix} \mathbb{E}_{Y \sim P_\star}[Y] \\ \mathbb{E}_{Y \sim P_\star}[Y^2] \end{bmatrix} - \begin{bmatrix} \mathbb{E}_{Y \sim P}[Y] \\ \mathbb{E}_{Y \sim P}[Y^2] \end{bmatrix} \right\| \end{aligned}$$

The MMD is a proper metric if k is a *characteristic* kernel [Sriperumbudur et al., 2011]; our use of MMD is justified by its interpretation as a statistical divergence induced by a *proper scoring rule* [Dawid, 1986].

Predictive Fit via MMD

To define the *predictive fit* for PCUQ, we need:

- kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, e.g. $k(y, y') = (yy') + (yy')^2$
- *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}(k)$ [see Berlinet and Thomas-Agnan, 2011, for background]
- *kernel mean embedding*

$$\mu_k(P) := \int k(\cdot, y) \, \mathrm{d}P(y) \in \mathcal{H}(k).$$

The divergence of a candidate $P \in \mathcal{P}(\mathcal{Y})$ from the data-generating distribution $P_\star$ can be quantified using *maximum mean discrepancy* (MMD):

$$\begin{aligned} \mathrm{MMD}(P_\star, P) &= \|\mu_k(P_\star) - \mu_k(P)\|_{\mathcal{H}(k)} \\ &\overset{\text{e.g.}}{=} \left\| \begin{bmatrix} \mathbb{E}_{Y \sim P_\star}[Y] \\ \mathbb{E}_{Y \sim P_\star}[Y^2] \end{bmatrix} - \begin{bmatrix} \mathbb{E}_{Y \sim P}[Y] \\ \mathbb{E}_{Y \sim P}[Y^2] \end{bmatrix} \right\| \end{aligned}$$

The MMD is a proper metric if k is a *characteristic* kernel [Sriperumbudur et al., 2011]; our use of MMD is justified by its interpretation as a statistical divergence induced by a *proper scoring rule* [Dawid, 1986].

Predictive Fit via MMD

To define the *predictive fit* for PCUQ, we need:

- kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, e.g. $k(y, y') = (yy') + (yy')^2$
- *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}(k)$ [see Berlinet and Thomas-Agnan, 2011, for background]
- *kernel mean embedding*

$$\mu_k(P) := \int k(\cdot, y) \, \mathrm{d}P(y) \in \mathcal{H}(k).$$

The divergence of a candidate $P \in \mathcal{P}(\mathcal{Y})$ from the data-generating distribution $P_\star$ can be quantified using *maximum mean discrepancy* (MMD):

$$\begin{aligned} \mathrm{MMD}(P_\star, P) &= \|\mu_k(P_\star) - \mu_k(P)\|_{\mathcal{H}(k)} \\ &\overset{\text{e.g.}}{=} \left\| \begin{bmatrix} \mathbb{E}_{Y \sim P_\star}[Y] \\ \mathbb{E}_{Y \sim P_\star}[Y^2] \end{bmatrix} - \begin{bmatrix} \mathbb{E}_{Y \sim P}[Y] \\ \mathbb{E}_{Y \sim P}[Y^2] \end{bmatrix} \right\| \end{aligned}$$

The MMD is a proper metric if k is a *characteristic* kernel [Sriperumbudur et al., 2011]; our use of MMD is justified by its interpretation as a statistical divergence induced by a *proper scoring rule* [Dawid, 1986].

Predictive Fit via MMD

To define the *predictive fit* for PCUQ, we need:

- kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, e.g. $k(y, y') = (yy') + (yy')^2$
- *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}(k)$ [see Berlinet and Thomas-Agnan, 2011, for background]
- *kernel mean embedding*

$$\mu_k(P) := \int k(\cdot, y) \, \mathrm{d}P(y) \in \mathcal{H}(k).$$

The divergence of a candidate $P \in \mathcal{P}(\mathcal{Y})$ from the data-generating distribution $P_\star$ can be quantified using *maximum mean discrepancy* (MMD):

$$\begin{aligned} \mathrm{MMD}(P_\star, P) &= \|\mu_k(P_\star) - \mu_k(P)\|_{\mathcal{H}(k)} \\ &\overset{\text{e.g.}}{=} \left\| \begin{bmatrix} \mathbb{E}_{Y \sim P_\star}[Y] \\ \mathbb{E}_{Y \sim P_\star}[Y^2] \end{bmatrix} - \begin{bmatrix} \mathbb{E}_{Y \sim P}[Y] \\ \mathbb{E}_{Y \sim P}[Y^2] \end{bmatrix} \right\| \end{aligned}$$

The MMD is a proper metric if k is a *characteristic* kernel [Sriperumbudur et al., 2011]; our use of MMD is justified by its interpretation as a statistical divergence induced by a *proper scoring rule* [Dawid, 1986].

Predictive Fit via MMD

To define the *predictive fit* for PCUQ, we need:

- kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, e.g. $k(y, y') = (yy') + (yy')^2$
- *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}(k)$ [see Berlinet and Thomas-Agnan, 2011, for background]
- *kernel mean embedding*

$$\mu_k(P) := \int k(\cdot, y) \, \mathrm{d}P(y) \in \mathcal{H}(k).$$

The divergence of a candidate $P \in \mathcal{P}(\mathcal{Y})$ from the data-generating distribution $P_\star$ can be quantified using *maximum mean discrepancy* (MMD):

$$\begin{aligned} \mathrm{MMD}(P_\star, P) &= \|\mu_k(P_\star) - \mu_k(P)\|_{\mathcal{H}(k)} \\ &\overset{\text{e.g.}}{=} \left\| \begin{bmatrix} \mathbb{E}_{Y \sim P_\star}[Y] \\ \mathbb{E}_{Y \sim P_\star}[Y^2] \end{bmatrix} - \begin{bmatrix} \mathbb{E}_{Y \sim P}[Y] \\ \mathbb{E}_{Y \sim P}[Y^2] \end{bmatrix} \right\| \end{aligned}$$

The MMD is a proper metric if k is a *characteristic* kernel [Sriperumbudur et al., 2011]; our use of MMD is justified by its interpretation as a statistical divergence induced by a *proper scoring rule* [Dawid, 1986].

Predictive Fit via MMD

To define the *predictive fit* for PCUQ, we need:

- kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, e.g. $k(y, y') = (yy') + (yy')^2$
- *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}(k)$ [see Berlinet and Thomas-Agnan, 2011, for background]
- *kernel mean embedding*

$$\mu_k(P) := \int k(\cdot, y) \, \mathrm{d}P(y) \in \mathcal{H}(k).$$

The divergence of a candidate $P \in \mathcal{P}(\mathcal{Y})$ from the data-generating distribution $P_\star$ can be quantified using *maximum mean discrepancy* (MMD):

$$\begin{aligned} \mathrm{MMD}(P_\star, P) &= \|\mu_k(P_\star) - \mu_k(P)\|_{\mathcal{H}(k)} \\ &\overset{\text{e.g.}}{=} \left\| \begin{bmatrix} \mathbb{E}_{Y \sim P_\star}[Y] \\ \mathbb{E}_{Y \sim P_\star}[Y^2] \end{bmatrix} - \begin{bmatrix} \mathbb{E}_{Y \sim P}[Y] \\ \mathbb{E}_{Y \sim P}[Y^2] \end{bmatrix} \right\| \end{aligned}$$

The MMD is a proper metric if k is a *characteristic* kernel [Sriperumbudur et al., 2011]; our use of MMD is justified by its interpretation as a statistical divergence induced by a *proper scoring rule* [Dawid, 1986].

Predictive Fit via MMD

The mixture model P_Q has kernel mean embedding

$$\mu_k(P_Q) = \iint k(\cdot, y)\, \mathrm{d}P_\theta(y)\mathrm{d}Q(\theta) = \int \mu_k(P_\theta)\, \mathrm{d}Q(\theta),$$

so the MMD between $P_\star$ and P_Q can be written as

$$\begin{aligned} \mathrm{MMD}^2(P_\star, P_Q) &= \left\| \int \{\mu_k(P_\star) - \mu_k(P_\theta)\}\, \mathrm{d}Q(\theta) \right\|^2_{\mathcal{H}(k)} \\ &= \iint \kappa_{P_\star}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta), \end{aligned} \tag{1}$$

where $\kappa_{P_\star} : \Theta \times \Theta \to \mathbb{R}$ is a kernel on Θ, and given by

$$\kappa_{P_\star}(\theta, \vartheta) = \langle \mu_k(P_\star) - \mu_k(P_\theta), \mu_k(P_\star) - \mu_k(P_\vartheta) \rangle_{\mathcal{H}(k)} .$$

Interpretation as Kernel Stein Discrepancy
This reveals one possible interpretation of (1) as a *kernel Stein discrepancy* [Chwialkowski et al., 2016, Liu et al., 2016, Gorham and Mackey, 2017] corresponding to the *Stein kernel* $\kappa_{P_\star}$ [Oates et al., 2017].

Predictive Fit via MMD

The mixture model P_Q has kernel mean embedding

$$\mu_k(P_Q) = \iint k(\cdot, y)\, \mathrm{d}P_\theta(y)\mathrm{d}Q(\theta) = \int \mu_k(P_\theta)\, \mathrm{d}Q(\theta),$$

so the MMD between $P_\star$ and P_Q can be written as

$$\begin{aligned}\mathrm{MMD}^2(P_\star, P_Q) &= \left\| \int \{\mu_k(P_\star) - \mu_k(P_\theta)\}\, \mathrm{d}Q(\theta) \right\|^2_{\mathcal{H}(k)} \\ &= \iint \kappa_{P_\star}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta), \end{aligned} \tag{1}$$

where $\kappa_{P_\star} : \Theta \times \Theta \to \mathbb{R}$ is a kernel on Θ, and given by

$$\kappa_{P_\star}(\theta, \vartheta) = \langle \mu_k(P_\star) - \mu_k(P_\theta), \mu_k(P_\star) - \mu_k(P_\vartheta) \rangle_{\mathcal{H}(k)} .$$

Interpretation as Kernel Stein Discrepancy
This reveals one possible interpretation of (1) as a *kernel Stein discrepancy* [Chwialkowski et al., 2016, Liu et al., 2016, Gorham and Mackey, 2017] corresponding to the *Stein kernel* $\kappa_{P_\star}$ [Oates et al., 2017].

Predictive Fit via MMD

The mixture model P_Q has kernel mean embedding

$$\mu_k(P_Q) = \iint k(\cdot, y)\, \mathrm{d}P_\theta(y)\mathrm{d}Q(\theta) = \int \mu_k(P_\theta)\, \mathrm{d}Q(\theta),$$

so the MMD between $P_\star$ and P_Q can be written as

$$\mathrm{MMD}^2(P_\star, P_Q) = \left\| \int \{\mu_k(P_\star) - \mu_k(P_\theta)\}\, \mathrm{d}Q(\theta) \right\|^2_{\mathcal{H}(k)}$$
$$= \iint \kappa_{P_\star}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta), \qquad (1)$$

where $\kappa_{P_\star} : \Theta \times \Theta \to \mathbb{R}$ is a kernel on Θ, and given by

$$\kappa_{P_\star}(\theta, \vartheta) = \langle \mu_k(P_\star) - \mu_k(P_\theta), \mu_k(P_\star) - \mu_k(P_\vartheta) \rangle_{\mathcal{H}(k)} .$$

Interpretation as Kernel Stein Discrepancy
This reveals one possible interpretation of (1) as a *kernel Stein discrepancy* [Chwialkowski et al., 2016, Liu et al., 2016, Gorham and Mackey, 2017] corresponding to the *Stein kernel* $\kappa_{P_\star}$ [Oates et al., 2017].

Predictive Fit via MMD

The mixture model P_Q has kernel mean embedding

$$\mu_k(P_Q) = \iint k(\cdot, y)\, \mathrm{d}P_\theta(y)\mathrm{d}Q(\theta) = \int \mu_k(P_\theta)\, \mathrm{d}Q(\theta),$$

so the MMD between $P_\star$ and P_Q can be written as

$$\begin{aligned} \mathrm{MMD}^2(P_\star, P_Q) &= \left\| \int \{\mu_k(P_\star) - \mu_k(P_\theta)\}\, \mathrm{d}Q(\theta) \right\|^2_{\mathcal{H}(k)} \\ &= \iint \kappa_{P_\star}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta), \end{aligned} \tag{1}$$

where $\kappa_{P_\star} : \Theta \times \Theta \to \mathbb{R}$ is a kernel on Θ, and given by

$$\kappa_{P_\star}(\theta, \vartheta) = \langle \mu_k(P_\star) - \mu_k(P_\theta), \mu_k(P_\star) - \mu_k(P_\vartheta) \rangle_{\mathcal{H}(k)} .$$

Interpretation as Kernel Stein Discrepancy

This reveals one possible interpretation of (1) as a *kernel Stein discrepancy* [Chwialkowski et al., 2016, Liu et al., 2016, Gorham and Mackey, 2017] corresponding to the *Stein kernel* $\kappa_{P_\star}$ [Oates et al., 2017].

Predictive Fit via MMD

The mixture model P_Q has kernel mean embedding

$$\mu_k(P_Q) = \iint k(\cdot, y)\, \mathrm{d}P_\theta(y)\mathrm{d}Q(\theta) = \int \mu_k(P_\theta)\, \mathrm{d}Q(\theta),$$

so the MMD between $P_\star$ and P_Q can be written as

$$\begin{aligned} \mathrm{MMD}^2(P_\star, P_Q) &= \left\| \int \{\mu_k(P_\star) - \mu_k(P_\theta)\}\, \mathrm{d}Q(\theta) \right\|^2_{\mathcal{H}(k)} \\ &= \iint \kappa_{P_\star}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta), \end{aligned} \tag{1}$$

where $\kappa_{P_\star} : \Theta \times \Theta \to \mathbb{R}$ is a kernel on Θ, and given by

$$\kappa_{P_\star}(\theta, \vartheta) = \langle \mu_k(P_\star) - \mu_k(P_\theta), \mu_k(P_\star) - \mu_k(P_\vartheta) \rangle_{\mathcal{H}(k)} .$$

Interpretation as Kernel Stein Discrepancy
This reveals one possible interpretation of (1) as a *kernel Stein discrepancy* [Chwialkowski et al., 2016, Liu et al., 2016, Gorham and Mackey, 2017] corresponding to the *Stein kernel* $\kappa_{P_\star}$ [Oates et al., 2017].

Predictive Fit via MMD

The mixture model P_Q has kernel mean embedding

$$\mu_k(P_Q) = \iint k(\cdot, y)\, \mathrm{d}P_\theta(y)\mathrm{d}Q(\theta) = \int \mu_k(P_\theta)\, \mathrm{d}Q(\theta),$$

so the MMD between $P_\star$ and P_Q can be written as

$$\begin{aligned} \mathrm{MMD}^2(P_\star, P_Q) &= \left\| \int \{\mu_k(P_\star) - \mu_k(P_\theta)\}\, \mathrm{d}Q(\theta) \right\|^2_{\mathcal{H}(k)} \\ &= \iint \kappa_{P_\star}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta), \end{aligned} \tag{1}$$

where $\kappa_{P_\star} : \Theta \times \Theta \to \mathbb{R}$ is a kernel on Θ, and given by

$$\kappa_{P_\star}(\theta, \vartheta) = \langle \mu_k(P_\star) - \mu_k(P_\theta), \mu_k(P_\star) - \mu_k(P_\vartheta) \rangle_{\mathcal{H}(k)}\,.$$

Sensible in the Well-Specified Context
If $P_\star = P_{\theta_\star}$ for some unique $\theta_\star \in \Theta$, then (1) is uniquely minimised by $Q = \delta_{\theta_\star}$ provided k is a characteristic kernel.

Predictive Fit via MMD

The mixture model P_Q has kernel mean embedding

$$\mu_k(P_Q) = \iint k(\cdot, y)\, \mathrm{d}P_\theta(y)\mathrm{d}Q(\theta) = \int \mu_k(P_\theta)\, \mathrm{d}Q(\theta),$$

so the MMD between $P_\star$ and P_Q can be written as

$$\begin{aligned} \mathrm{MMD}^2(P_\star, P_Q) &= \left\| \int \{\mu_k(P_\star) - \mu_k(P_\theta)\}\, \mathrm{d}Q(\theta) \right\|^2_{\mathcal{H}(k)} \\ &= \iint \kappa_{P_\star}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta), \end{aligned} \tag{1}$$

where $\kappa_{P_\star} : \Theta \times \Theta \to \mathbb{R}$ is a kernel on Θ, and given by

$$\kappa_{P_\star}(\theta, \vartheta) = \langle \mu_k(P_\star) - \mu_k(P_\theta), \mu_k(P_\star) - \mu_k(P_\vartheta) \rangle_{\mathcal{H}(k)}\,.$$

Estimation
Of course, the true data-generating distribution $P_\star$ in (1) is unknown and must be approximated. In PCUQ we use the empirical distribution P_n in lieu of $P_\star$.

Predictive Fit via MMD

The mixture model P_Q has kernel mean embedding

$$\mu_k(P_Q) = \iint k(\cdot, y)\, \mathrm{d}P_\theta(y)\mathrm{d}Q(\theta) = \int \mu_k(P_\theta)\, \mathrm{d}Q(\theta),$$

so the MMD between $P_\star$ and P_Q can be written as

$$\begin{aligned} \mathrm{MMD}^2(P_\star, P_Q) &= \left\| \int \{\mu_k(P_\star) - \mu_k(P_\theta)\}\, \mathrm{d}Q(\theta) \right\|^2_{\mathcal{H}(k)} \\ &= \iint \kappa_{P_\star}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta), \end{aligned} \tag{1}$$

where $\kappa_{P_\star} : \Theta \times \Theta \to \mathbb{R}$ is a kernel on Θ, and given by

$$\kappa_{P_\star}(\theta, \vartheta) = \langle \mu_k(P_\star) - \mu_k(P_\theta), \mu_k(P_\star) - \mu_k(P_\vartheta) \rangle_{\mathcal{H}(k)} .$$

Regularisation
A plug-in approximation necessitates additional regularisation, since otherwise minimisation of $Q \mapsto \mathrm{MMD}(P_n, P_Q)$ would result in a discrete distribution where each atom corresponds to a value of θ that explains one of the data points well.

Approximating Q_n via Gradient Flow

The output Q_n of PCUQ is a minimiser of the entropy-regularised objective

$$\mathcal{F}_n(Q) = \mathcal{E}_n(Q) + \lambda_n \int \log q(\theta)\, \mathrm{d}Q(\theta), \tag{2}$$

where the *free energy* $\mathcal{E}_n(Q)$ is identical, after algebraic manipulation, to

$$\mathcal{E}_n(Q) \overset{+C}{=} \int v(\theta)\, \mathrm{d}Q(\theta) + \frac{1}{2} \iint \kappa_{P_n}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta),$$

and where q and q_0 are respectively densities for Q and Q_0.

Wasserstein Gradient Flow

For the entropy-regularised functional $\mathcal{F}_n$ (2), we can simulate a Wasserstein gradient flow via a McKean–Vlasov process [Ambrosio et al., 2008]

$$\mathrm{d}\theta_t = -\nabla_{\mathrm{W}} \mathcal{E}_n(Q^t)(\theta_t) + \sqrt{2\lambda_n}\mathrm{d}W_t, \tag{3}$$

$$\nabla_{\mathrm{W}} \mathcal{E}_n(Q^t)(\theta_t) = \nabla v(\theta_t) + \int \nabla_1 \kappa_{P_n}(\theta_t, \vartheta)\, \mathrm{d}Q^t(\vartheta)$$

where $Q^t = \mathrm{law}(\theta_t)$, ∇_{W} denotes the Wasserstein gradient, $(W_t)_{t \geq 0}$ is a Wiener process on $\mathbb{R}^p$ and, for the bivariate function κ_{P_n}, the notation $\nabla_1 \kappa_{P_n}$ denotes differentiation with respect to the first argument.

Approximating Q_n via Gradient Flow

The output Q_n of PCUQ is a minimiser of the entropy-regularised objective

$$\mathcal{F}_n(Q) = \mathcal{E}_n(Q) + \lambda_n \int \log q(\theta)\, \mathrm{d}Q(\theta), \tag{2}$$

where the *free energy* $\mathcal{E}_n(Q)$ is identical, after algebraic manipulation, to

$$\mathcal{E}_n(Q) \overset{+C}{=} \int v(\theta)\, \mathrm{d}Q(\theta) + \frac{1}{2} \iint \kappa_{P_n}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta),$$

and where q and q_0 are respectively densities for Q and Q_0.

Wasserstein Gradient Flow

For the entropy-regularised functional $\mathcal{F}_n$ (2), we can simulate a Wasserstein gradient flow via a McKean–Vlasov process [Ambrosio et al., 2008]

$$\mathrm{d}\theta_t = -\nabla_{\mathrm{W}} \mathcal{E}_n(Q^t)(\theta_t) + \sqrt{2\lambda_n}\mathrm{d}W_t, \tag{3}$$

$$\nabla_{\mathrm{W}} \mathcal{E}_n(Q^t)(\theta_t) = \nabla v(\theta_t) + \int \nabla_1 \kappa_{P_n}(\theta_t, \vartheta)\, \mathrm{d}Q^t(\vartheta)$$

where $Q^t = \mathrm{law}(\theta_t)$, ∇_{W} denotes the Wasserstein gradient, $(W_t)_{t \geq 0}$ is a Wiener process on $\mathbb{R}^p$ and, for the bivariate function κ_{P_n}, the notation $\nabla_1 \kappa_{P_n}$ denotes differentiation with respect to the first argument.

Approximating Q_n via Gradient Flow

The output Q_n of PCUQ is a minimiser of the entropy-regularised objective

$$\mathcal{F}_n(Q) = \mathcal{E}_n(Q) + \lambda_n \int \log q(\theta)\, \mathrm{d}Q(\theta), \tag{2}$$

where the *free energy* $\mathcal{E}_n(Q)$ is identical, after algebraic manipulation, to

$$\mathcal{E}_n(Q) \overset{+C}{=} \int v(\theta)\, \mathrm{d}Q(\theta) + \frac{1}{2} \iint \kappa_{P_n}(\theta, \vartheta)\, \mathrm{d}Q(\theta)\mathrm{d}Q(\vartheta),$$

and where q and q_0 are respectively densities for Q and Q_0.

Simulation as an Interacting Particle System
Discretise Q^t into a system of N evolving particles $\theta_t^1, \theta_t^2, \ldots, \theta_t^N$, whose evolution is governed by the following system of *stochastic differential equations* (SDEs):

$$\mathrm{d}\theta_t^i = -\left(\nabla v(\theta_t^i) + \frac{1}{N-1} \sum_{j \neq i} \nabla_1 \kappa_{P_n}(\theta_t^i, \theta_t^j)\right)\mathrm{d}t + \sqrt{2\lambda_n}\mathrm{d}W_t^i,$$

where $(W_t^i)_{t \geq 0}$ are N independent Wiener processes on $\mathbb{R}^p$. An Euler–Maruyama discretisation incurs per-iteration computational complexity $O(nN^2)$ and storage complexity (with caching) of $O(n + N)$.

Approximating Q_n via Gradient Flow

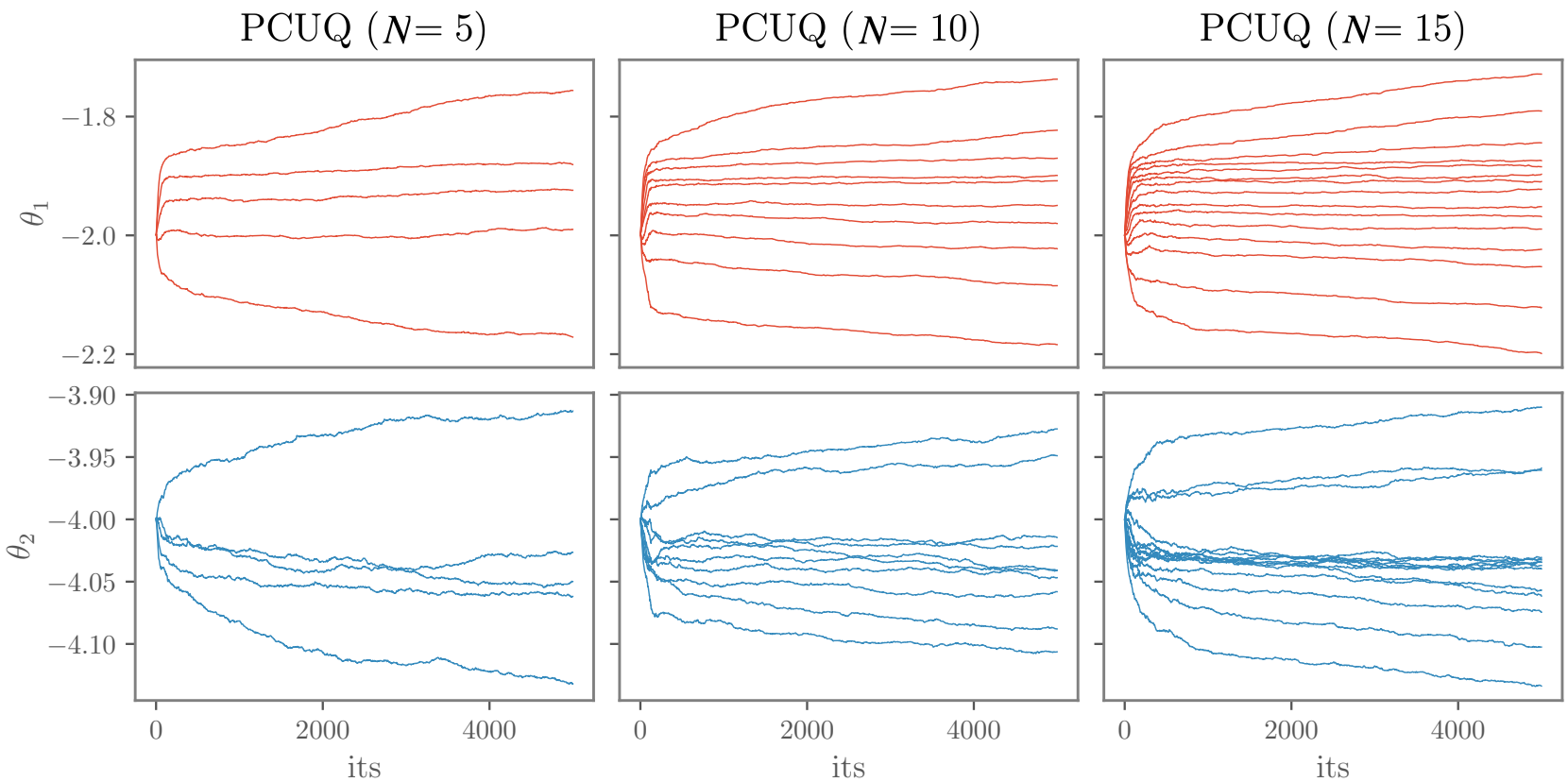


Figure: Interacting particle system for approximation of Q_n. (N = number of particles used)

Convergence of the Gradient Flow

Though theoretically convex, in practice gradients are small in low probability region; we mitigated this by initialising close to the Bayesian MAP $\theta^\dagger$.

Approximating Q_n via Gradient Flow

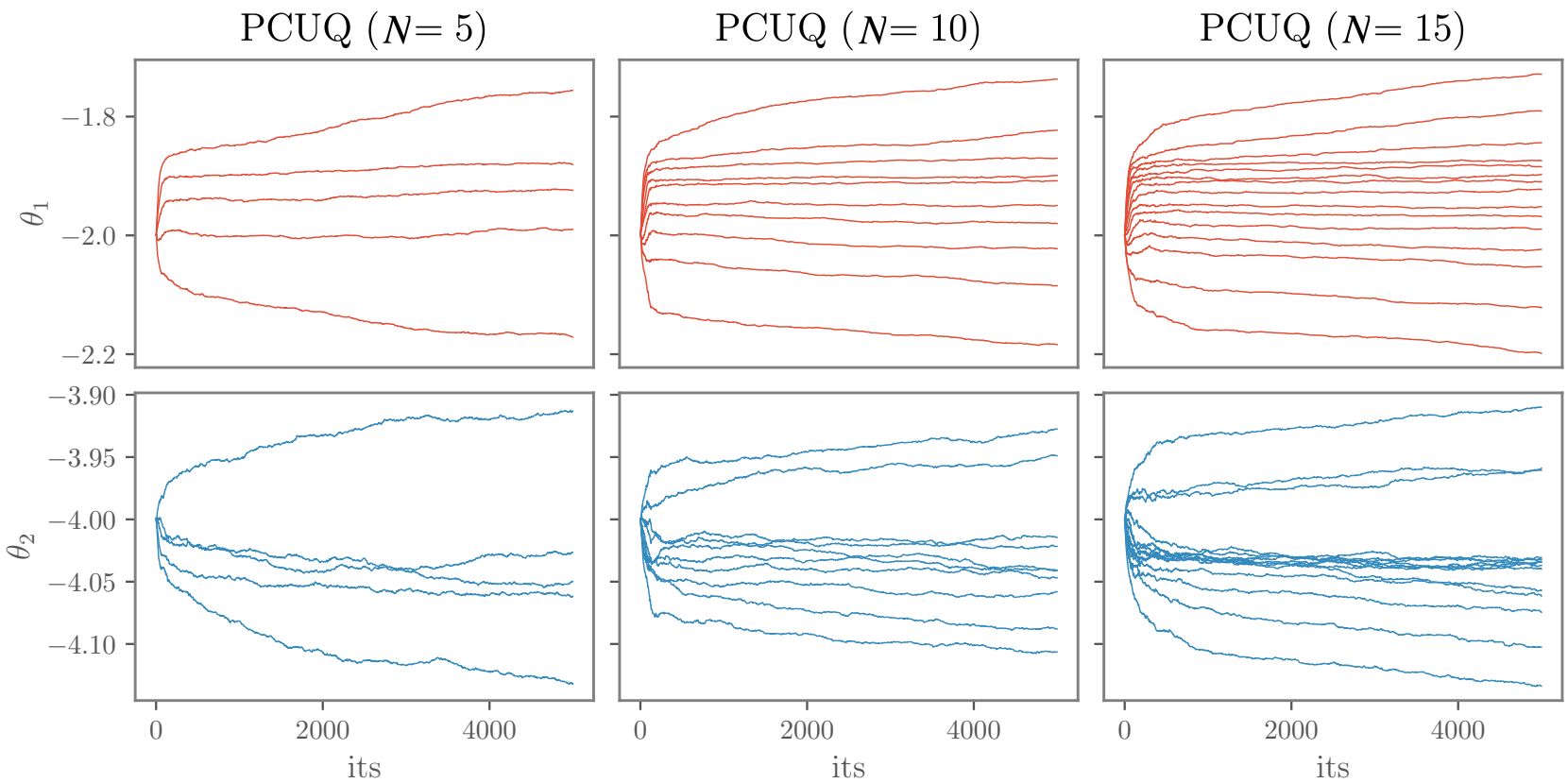


Figure: Interacting particle system for approximation of Q_n. (N = number of particles used)

Convergence of the Gradient Flow

Though theoretically convex, in practice gradients are small in low probability region; we mitigated this by initialising close to the Bayesian MAP $\theta^\dagger$.

Extension to Dependent Data

- each y_i is associated with a covariate $x_i \in \mathcal{X}$ and generated according to an (unknown) conditional distribution $P_\star(\cdot|x_i)$
- have a conditional model $\{P_\theta(\cdot|x)\}_{\theta\in\Theta}$ for each $x \in \mathcal{X}$

Idea: Suppose that [Alquier and Gerber, 2024]

$$\{(x_i, y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} \bar{P}_\star(\mathrm{d}x, \mathrm{d}y) := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(\mathrm{d}x)\, P_0(\mathrm{d}y|x_i)$$

and consider the extended model

$$\bar{P}_\theta(\mathrm{d}x, \mathrm{d}y) := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(\mathrm{d}x)\, P_\theta(\mathrm{d}y|x_i).$$

Choice of Kernel
For example, if $\mathcal{X} \subset \mathbb{R}^{d_\mathcal{X}}$ and $\mathcal{Y} \subset \mathbb{R}^{d_\mathcal{Y}}$, we may consider the Gaussian kernel

$$k((x, y), (x', y')) = \exp\left(-\frac{\|x - x'\|^2}{\ell_\mathcal{X}^2} - \frac{\|y - y'\|^2}{\ell_\mathcal{Y}^2}\right)$$

with bandwidths $\ell_\mathcal{X}$ and $\ell_\mathcal{Y}$ to be specified. (We take $\ell_\mathcal{X} \to 0$, as recommended in Alquier and Gerber [2024].)

Extension to Dependent Data

- each y_i is associated with a covariate $x_i \in \mathcal{X}$ and generated according to an (unknown) conditional distribution $P_\star(\cdot|x_i)$
- have a conditional model $\{P_\theta(\cdot|x)\}_{\theta\in\Theta}$ for each $x \in \mathcal{X}$

Idea: Suppose that [Alquier and Gerber, 2024]

$$\{(x_i, y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} \bar{P}_\star(\mathrm{d}x, \mathrm{d}y) := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(\mathrm{d}x)\, P_0(\mathrm{d}y|x_i)$$

and consider the extended model

$$\bar{P}_\theta(\mathrm{d}x, \mathrm{d}y) := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(\mathrm{d}x)\, P_\theta(\mathrm{d}y|x_i).$$

Choice of Kernel

For example, if $\mathcal{X} \subset \mathbb{R}^{d_\mathcal{X}}$ and $\mathcal{Y} \subset \mathbb{R}^{d_\mathcal{Y}}$, we may consider the Gaussian kernel

$$k((x,y),(x',y')) = \exp\left(-\frac{\|x-x'\|^2}{\ell_\mathcal{X}^2} - \frac{\|y-y'\|^2}{\ell_\mathcal{Y}^2}\right)$$

with bandwidths $\ell_\mathcal{X}$ and $\ell_\mathcal{Y}$ to be specified. (We take $\ell_\mathcal{X} \to 0$, as recommended in Alquier and Gerber [2024].)

Extension to Dependent Data

- each y_i is associated with a covariate $x_i \in \mathcal{X}$ and generated according to an (unknown) conditional distribution $P_\star(\cdot|x_i)$
- have a conditional model $\{P_\theta(\cdot|x)\}_{\theta\in\Theta}$ for each $x \in \mathcal{X}$

Idea: Suppose that [Alquier and Gerber, 2024]

$$\{(x_i, y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} \bar{P}_\star(\mathrm{d}x, \mathrm{d}y) := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(\mathrm{d}x)\, P_0(\mathrm{d}y|x_i)$$

and consider the extended model

$$\bar{P}_\theta(\mathrm{d}x, \mathrm{d}y) := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(\mathrm{d}x)\, P_\theta(\mathrm{d}y|x_i).$$

Choice of Kernel
For example, if $\mathcal{X} \subset \mathbb{R}^{d_\mathcal{X}}$ and $\mathcal{Y} \subset \mathbb{R}^{d_\mathcal{Y}}$, we may consider the Gaussian kernel

$$k((x, y), (x', y')) = \exp\left(-\frac{\|x - x'\|^2}{\ell_\mathcal{X}^2} - \frac{\|y - y'\|^2}{\ell_\mathcal{Y}^2}\right)$$

with bandwidths $\ell_\mathcal{X}$ and $\ell_\mathcal{Y}$ to be specified. (We take $\ell_\mathcal{X} \to 0$, as recommended in Alquier and Gerber [2024].)

Summary

Summary

The main claims:

- There is a need for post-Bayesian methods to meet the needs of scientific communities for whom "model" $\neq$ "statistical model".
- Methodology should probably be tailored to specific communities.
- Prediction-centric approaches are not new, but they are an interesting alternative to generalised Bayes!

Notable omissions:

- Bayesian exponentially tilted empirical likelihood, conformal prediction, martingale posteriors (next chapter!), ...

If you would like to read more about our approach:

Prediction-Centric Uncertainty Quantification via MMD
Shen Z, Knoblauch J, Power S, Oates CJ
In: Artificial Intelligence and Statistics (AISTATS 2025)
`https://arxiv.org/abs/2410.11637`

Thank you for your attention!

Summary

The main claims:

- There is a need for post-Bayesian methods to meet the needs of scientific communities for whom "model" $\neq$ "statistical model".
- Methodology should probably be tailored to specific communities.
- Prediction-centric approaches are not new, but they are an interesting alternative to generalised Bayes!

Notable omissions:

- Bayesian exponentially tilted empirical likelihood, conformal prediction, martingale posteriors (next chapter!), ...

If you would like to read more about our approach:

Prediction-Centric Uncertainty Quantification via MMD
Shen Z, Knoblauch J, Power S, Oates CJ
In: Artificial Intelligence and Statistics (AISTATS 2025)
https://arxiv.org/abs/2410.11637

Thank you for your attention!

Summary

The main claims:

- There is a need for post-Bayesian methods to meet the needs of scientific communities for whom "model" $\neq$ "statistical model".
- Methodology should probably be tailored to specific communities.
- Prediction-centric approaches are not new, but they are an interesting alternative to generalised Bayes!

Notable omissions:

- Bayesian exponentially tilted empirical likelihood, conformal prediction, martingale posteriors (next chapter!), ...

If you would like to read more about our approach:

Prediction-Centric Uncertainty Quantification via MMD
Shen Z, Knoblauch J, Power S, Oates CJ
In: Artificial Intelligence and Statistics (AISTATS 2025)
https://arxiv.org/abs/2410.11637

Thank you for your attention!

Summary

The main claims:

- There is a need for post-Bayesian methods to meet the needs of scientific communities for whom "model" $\neq$ "statistical model".
- Methodology should probably be tailored to specific communities.
- Prediction-centric approaches are not new, but they are an interesting alternative to generalised Bayes!

Notable omissions:

- Bayesian exponentially tilted empirical likelihood, conformal prediction, martingale posteriors (next chapter!), ...

If you would like to read more about our approach:

Prediction-Centric Uncertainty Quantification via MMD
Shen Z, Knoblauch J, Power S, Oates CJ
In: Artificial Intelligence and Statistics (AISTATS 2025)
https://arxiv.org/abs/2410.11637

Thank you for your attention!

Summary

The main claims:

- There is a need for post-Bayesian methods to meet the needs of scientific communities for whom "model" $\neq$ "statistical model".
- Methodology should probably be tailored to specific communities.
- Prediction-centric approaches are not new, but they are an interesting alternative to generalised Bayes!

Notable omissions:

- Bayesian exponentially tilted empirical likelihood, conformal prediction, martingale posteriors (next chapter!), ...

If you would like to read more about our approach:

Prediction-Centric Uncertainty Quantification via MMD
Shen Z, Knoblauch J, Power S, Oates CJ
In: Artificial Intelligence and Statistics (AISTATS 2025)
https://arxiv.org/abs/2410.11637

Thank you for your attention!

Summary

The main claims:

- There is a need for post-Bayesian methods to meet the needs of scientific communities for whom "model" $\neq$ "statistical model".
- Methodology should probably be tailored to specific communities.
- Prediction-centric approaches are not new, but they are an interesting alternative to generalised Bayes!

Notable omissions:

- Bayesian exponentially tilted empirical likelihood, conformal prediction, martingale posteriors (next chapter!), ...

If you would like to read more about our approach:

Prediction-Centric Uncertainty Quantification via MMD
Shen Z, Knoblauch J, Power S, Oates CJ
In: Artificial Intelligence and Statistics (AISTATS 2025)
https://arxiv.org/abs/2410.11637

Thank you for your attention!

Summary

The main claims:

- There is a need for post-Bayesian methods to meet the needs of scientific communities for whom "model" $\neq$ "statistical model".
- Methodology should probably be tailored to specific communities.
- Prediction-centric approaches are not new, but they are an interesting alternative to generalised Bayes!

Notable omissions:

- Bayesian exponentially tilted empirical likelihood, conformal prediction, martingale posteriors (next chapter!), ...

If you would like to read more about our approach:

Prediction-Centric Uncertainty Quantification via MMD
Shen Z, Knoblauch J, Power S, Oates CJ
In: Artificial Intelligence and Statistics (AISTATS 2025)
`https://arxiv.org/abs/2410.11637`

Thank you for your attention!

Summary

The main claims:

- There is a need for post-Bayesian methods to meet the needs of scientific communities for whom "model" $\neq$ "statistical model".
- Methodology should probably be tailored to specific communities.
- Prediction-centric approaches are not new, but they are an interesting alternative to generalised Bayes!

Notable omissions:

- Bayesian exponentially tilted empirical likelihood, conformal prediction, martingale posteriors (next chapter!), ...

If you would like to read more about our approach:

Prediction-Centric Uncertainty Quantification via MMD
Shen Z, Knoblauch J, Power S, Oates CJ
In: Artificial Intelligence and Statistics (AISTATS 2025)
`https://arxiv.org/abs/2410.11637`

Thank you for your attention!

References I

P. Alquier and M. Gerber. Universal robust regression via maximum mean discrepancy. *Biometrika*, 111(1):71–92, 2024.

M. Altamirano, F.-X. Briol, and J. Knoblauch. Robust and scalable Bayesian online changepoint detection. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

M. Altamirano, F.-X. Briol, and J. Knoblauch. Robust and conjugate Gaussian process regression. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: In metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society, Series B*, 78(5):1103, 2016.

B.-E. Chérief-Abdellatif and P. Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Proceedings of the Symposium on Advances in Approximate Bayesian Inference*. PMLR, 2020.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

A. P. Dawid. Probability forecasting. In *Encyclopedia of Statistical Sciences*. Wiley Online Library, 1986.

C. Dellaporta, J. Knoblauch, T. Damoulas, and F.-X. Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022.

G. Duran-Martin, M. Altamirano, A. Shestopaloff, L. Sánchez-Betancourt, J. Knoblauch, M. Jones, F.-X. Briol, and K. P. Murphy. Outlier-robust Kalman filtering through generalised Bayes. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

References II

A. Ghosh and A. Basu. Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68:413–437, 2016.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

G. Hooker and A. N. Vidyashankar. Bayesian model robustness via disparities. *Test*, 23:556–584, 2014.

H. Husain and J. Knoblauch. Adversarial interpretation of Bayesian inference. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, 2022.

M. Jankowiak, G. Pleiss, and J. Gardner. Deep sigma point processes. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 2020a.

M. Jankowiak, G. Pleiss, and J. Gardner. Parametric Gaussian process regressors. In *Proceedings of the 37th International Conference on Machine Learning*, 2020b.

C. Jordan-Squire. *Convex Optimization over Probability Measures*. PhD thesis, University of Washington, 2015.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, 63(3):425–464, 2001.

J. Knoblauch, J. E. Jewson, and T. Damoulas. Doubly robust Bayesian inference for non-stationary streaming data with beta-divergences. 2018.

J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23 (132):1–109, 2022.

J. Lai and Y. Yao. Predictive variational inference: Learn the predictively optimal posterior distribution. *arXiv preprint arXiv:2410.14843*, 2024.

References III

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

Y. Li and Y. Gal. Dropout inference in Bayesian neural networks with alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

B. G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. Institute of Mathematical Statistics, 1995.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

R. S. Malik-Sheriff, M. Glont, T. V. N. Nguyen, K. Tiwari, M. G. Roberts, A. Xavier, M. T. Vu, J. Men, M. Maire, S. Kananathan, E. L. Fairbanks, J. P. Meyer, C. Arankalle, T. M. Varusai, V. Knight-Schrijver, L. Li, C. Dueñas-Roca, G. Dass, S. M. Keating, Y. M. Park, N. Buso, N. Rodriguez, M. Hucka, and H. Hermjakob. BioModels — 15 years of sharing computational models in life science. *Nucleic Acids Research*, 48(D1):D407–D415, 2020.

J. W. Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(1):7598–7650, 2021.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society, Series B*, 79(3):695–718, 2017.

S. M. Schmon, P. W. Cannon, and J. Knoblauch. Generalized posteriors in approximate Bayesian computation. In *Proceedings of the 3rd Symposium on Advances in Approximate Bayesian Inference*, 2020.

R. Sheth and R. Khardon. Pseudo-Bayesian learning via direct loss minimization with applications to sparse Gaussian process models. In *Proceedings of the 6th Symposium on Advances in Approximate Bayesian Inference*, 2020.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

References IV

C. Villacampa-Calvo and D. Hernandez-Lobato. Alpha divergence minimization in multi-class Gaussian process classification. *Neurocomputing*, 378:210–227, 2020.

V. D. Wild, S. Ghalebikesabi, D. Sejdinovic, and J. Knoblauch. A rigorous link between deep ensembles and (variational) Bayesian methods. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, 2023.