ROBUST ESTIMATION FOR GAUSSIAN PROCESSES AND BEYOND

Masha Naslidnyk
University College London

Probabilistic Numerics Spring School, 29th of March 2023

ROBUST ESTIMATION FOR GAUSSIAN PROCESSES AND BEYOND

Masha Naslidnyk
University College London

work with Ayush Bharti, Jeremias Knoblauch, and François-Xavier Briol. Under review @JRSSB.

CHALLENGE

— Paradigm shift in Bayesian inference exacerbates **model misspecification** —

Domain expertise
Bespoke models
Small, well-curated datasets

Black-box problems
Flexible models
Large, complex datasets

CHALLENGE

— Paradigm shift in Bayesian inference exacerbates **model misspecification** —

- Under misspecification, uncertainty quantification becomes brittle, especially in nonparametric models like **Gaussian Processes** (GPs).

CHALLENGE

— Paradigm shift in Bayesian inference exacerbates **model misspecification** —

CHALLENGE

— Paradigm shift in Bayesian inference exacerbates **model misspecification** —

CHALLENGE

— Paradigm shift in Bayesian inference exacerbates **model misspecification** —

- Under misspecification, uncertainty quantification becomes brittle, especially in nonparametric models like **Gaussian Processes** (GPs).

- Robust **distance-based methods** are a practical remedy, but previously proposed distances between **conditional distributions** are intractable!

CHALLENGE

— Paradigm shift in Bayesian inference exacerbates **model misspecification** —

METHODOLOGY

EMBEDDING CONDITIONAL DISTRIBUTIONS

EMBEDDING CONDITIONAL DISTRIBUTIONS

- For any $x \in \mathcal{X}$, $p(x, \cdot)$ is a **probability measure** on $\mathcal{Y}$: $\mathbb{P}(Y \in A | X = x) = \int_A p(x, \mathrm{d}y)$
- p can be mapped to a function $\mathcal{X} \to \mathcal{H}_\mathcal{Y}$, its **conditional kernel mean embedding**,

$$\mu_p(x)(y) = \mathbb{E}_{Y \sim p(x,\cdot)}[k_\mathcal{Y}(Y, y)] = \int_\mathcal{Y} k_\mathcal{Y}(y', y) p(x, \mathrm{d}y')$$

EXPECTED MEAN CONDITIONAL MEAN DISCREPANCY (EMCMD)

$$\mathbb{D}_\nu(\mu_p, \mu_q) = \mathbb{E}_{X\sim\nu}\|\mu_p(X) - \mu_q(X)\|_{\mathcal{H}_\mathcal{Y}}$$

EMCMD-BASED PARAMETER ESTIMATION

For a model p_θ and the true data-generating distribution q, we denote $\mu_\theta = \mu_{p_\theta}$ and target

$$\theta^* \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{D}_\nu(\mu_\theta, \mu_q)$$

EMCMD-BASED PARAMETER ESTIMATION

For a model p_θ and the true data-generating distribution q, we denote $\mu_\theta = \mu_{p_\theta}$ and target

$$\theta^* \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{D}_\nu(\mu_\theta, \mu_q)$$

EMCMD-BASED PARAMETER ESTIMATION

$$\hat{\mu}_q^n(x)(\cdot) = k_\mathcal{X}(x, x_{1:n})^\top \left(k_\mathcal{X}(x_{1:n}, x_{1:n}) + n\lambda \mathrm{Id}_n\right)^{-1} k_\mathcal{Y}(y_{1:n}, \cdot), \qquad \nu_l = \frac{1}{l}\sum_{i=1}^{l} \delta_{\tilde{x}_i}$$

EMCMD-BASED PARAMETER ESTIMATION

$$\hat{\mu}_q^n(x)(\cdot) = k_\mathcal{X}(x, x_{1:n})^\top \left(k_\mathcal{X}(x_{1:n}, x_{1:n}) + n\lambda \mathrm{Id}_n\right)^{-1} k_\mathcal{Y}(y_{1:n}, \cdot), \qquad \nu_l = \frac{1}{l}\sum_{i=1}^{l} \delta_{\tilde{x}_i}$$

$$\hat{\theta}_{n,l} \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{D}_{\nu_l}(\mu_\theta, \hat{\mu}_q^n)$$

THEORETICAL GUARANTEES

$$\theta^* \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{D}_{\nu}(\mu_\theta, \mu_q) \longrightarrow \hat{\theta}_{n,l} \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{D}_{\nu_l}(\mu_\theta, \hat{\mu}_q^n)$$

CONSISTENCY AND GENERALISATION

CONSISTENCY AND GENERALISATION

- Consistency: $\mathbb{D}_{\nu_l}(\mu_\theta, \hat{\mu}_q^n) \xrightarrow{P} \mathbb{D}_\nu(\mu_\theta, \mu_q)$ e.f., uniformly in θ; $\hat{\theta}_{n,l} \xrightarrow{P} \theta^*$ e.f.
- Generalisation bound: $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_q) \xrightarrow{P} \mathbb{D}_\nu(\mu_{\theta^*}, \mu_q)$ e.f.

CONSISTENCY AND GENERALISATION

If $q = p_{\theta_0}$ for some θ_0, $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_{\theta_0}) \xrightarrow{P} 0$ e.f.

- Generalisation bound: $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_q) \xrightarrow{P} \mathbb{D}_\nu(\mu_{\theta^*}, \mu_q)$ e.f.

ROBUSTNESS

almost almost

If $q = p_{\theta_0}$ for some θ_0, $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_{\theta_0}) \xrightarrow{P} 0$ e.f.

ROBUSTNESS

If $q \overset{\text{almost}}{=} p_{\theta_0}$ for some θ_0, $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_{\theta_0}) \xrightarrow{P} 0$ e.f.

almost

ε-Huber contamination, $\varepsilon \in [0, 1]$:

ROBUSTNESS

If $q = (1-\varepsilon)p_{\theta_0} + \varepsilon r$ for some θ_0, $\mathbb{D}_\nu(\mu_{\hat{\theta}_{n,l}}, \mu_{\theta_0}) \xrightarrow{P} C_0\varepsilon$ e.f.

ε-Huber contamination, $\varepsilon \in [0,1]$:

NUMERICAL EXPERIMENTS: ROBUST EMPIRICAL BAYES FOR GAUSSIAN PROCESSES

SETTING

- p_θ is a marginal of a Gaussian process $\mathcal{GP}(0, c_\theta)$
- $c_\theta(x, x') = \theta c(x, x')$

GP REGRESSION WITH STUDENT-T MEASUREMENT ERROR

- Controlled environment: true function is a sample from a GP.
- 95% of data have Gaussian noise
- 5% of data have student-t noise
- Modeller assumes the noise is Gaussian.

GP REGRESSION WITH STUDENT-T MEASUREMENT ERROR

- 95% of data have noise $\sim \sigma_\varepsilon^2 \mathcal{N}(0,1)$
- 5% of data have noise $\sim \sigma_\varepsilon^2 \text{Student-t}(3)$
- **ML and LOO-CV are not robust to outliers; eMCMD is!**

CALIBRATION DIAGRAMS FOR WIND-FARM MODELLING

- Input: location of wind farms and wind angle. Output: energy produced. Noise is assumed Gaussian, but isn't likely to be—so the model is misspecified.
- **ML is underconfident; eMCMD is better calibrated.**

MISSPECIFIED BAYESIAN OPTIMISATION

Task: locate the maximum of a function.

MISSPECIFIED BAYESIAN OPTIMISATION

even an outlier that does **not** affect the maximum can hinder optimisation…

MISSPECIFIED BAYESIAN OPTIMISATION

even an outlier that does **not** affect the maximum can hinder optimisation…

eMCMD, unlike ML, is robust to outliers, so this doesn't happen!

MISSPECIFIED BAYESIAN OPTIMISATION

- Hartmann 6D function.
- Outliers don't change the maximum.
- Still, **eMCMD reaches the maximum faster than ML...**
- **...even when no outliers are added, since there is still misspecification.**

CONCLUSION AND FINAL REMARKS

- Introduced eMCMD, $\mathbb{D}_\nu(\mu_p, \mu_q) = \mathbb{E}_{X\sim\nu}\|\mu_p(X) - \mu_q(X)\|_{\mathcal{H}_\mathcal{Y}}$
- Proved optimality of the approximation $\hat{\theta}_{n,l} \in \operatorname*{argmin}_{\theta\in\Theta} \mathbb{D}_{\nu_l}(\mu_\theta, \hat{\mu}_q^n)$
- Optimality results hold even if we need to estimate *both* μ_θ and μ_q
- What's next? Robust (generalised) Bayesian inference!

P.S. OPTIMAL WEIGHTS FOR MMD ESTIMATION

$\mathbb{U} \longrightarrow$ $\longrightarrow \mathbb{P}_\theta$

$$G_\theta : \mathcal{U} \to \mathcal{X}$$

- Setting: likelihood-free inference. All we can do is sample from the model $\mathbb{P}_\theta$,

$$u_i \sim \mathbb{U}, \quad x_i = G_\theta(u_i)$$

- Minimise the distance (MMD) between empirical distributions,

$$\hat{\mathbb{Q}}^m = \frac{1}{m}\sum_{i=1}^m \delta_{y_i} \quad \text{and} \quad \hat{\mathbb{P}}^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i} = \frac{1}{n}\sum_{i=1}^n \delta_{G_\theta(u_i)}$$

Bharti, A., Naslidnyk, M., Key, O., Kaski, S., & Briol, F. X. (2023). Optimally-Weighted Estimators of the Maximum Mean Discrepancy for Likelihood-Free Inference. arXiv:2301.11674.

P.S. OPTIMAL WEIGHTS FOR MMD ESTIMATION

$\mathbb{U} \longrightarrow$ $\longrightarrow \mathbb{P}_\theta$

$G_\theta : \mathcal{U} \to \mathcal{X}$

- Approximating MMD involves approximating

$$\int_{\mathcal{X}} k(x, x')\mathbb{P}(\mathrm{d}x) = \int_{\mathcal{U}} k(G_\theta(u), x')\mathbb{U}(\mathrm{d}u) \approx w_i k(G_\theta(u_i), x')$$

- If k is Matérn of order ν, we can choose ν_0 s.t. for a Matérn c of order ν_0 the weights

$$w_{1:n} = c(u_{1:n}, u_{1:n})^{-1} \int_{\mathcal{U}} c(u_{1:n}, u)\mathbb{U}(\mathrm{d}u)$$

are optimal for approximating MMD (significantly improved sample complexity!)

Bharti, A., Naslidnyk, M., Key, O., Kaski, S., & Briol, F. X. (2023). Optimally-Weighted Estimators of the Maximum Mean Discrepancy for Likelihood-Free Inference. arXiv:2301.11674.

P.S. OPTIMAL WEIGHTS FOR MMD ESTIMATION

$\mathbb{U} \longrightarrow$ $\longrightarrow \mathbb{P}_\theta$

$G_\theta : \mathcal{U} \to \mathcal{X}$

- Approximating MMD involves approximating

$$\int_{\mathcal{X}} k(x, x')\mathbb{P}(\mathrm{d}x) = \int_{\mathcal{U}} k(G_\theta(u), x')\mathbb{U}(\mathrm{d}u) \approx w_i k(G_\theta(u_i), x')$$

- If k is Matérn of order ν, we can choose ν_0 s.t. for a Matérn c of order ν_0 the weights

$$w_{1:n} = c(u_{1:n}, u_{1:n})^{-1} \int_{\mathcal{U}} c(u_{1:n}, u)\mathbb{U}(\mathrm{d}u)$$

are optimal for approximating MMD (significantly improved sample complexity!)

Bharti, A., Naslidnyk, M., Key, O., Kaski, S., & Briol, F. X. (2023). Optimally-Weighted Estimators of the Maximum Mean Discrepancy for Likelihood-Free Inference. arXiv:2301.11674.

THANK YOU!

QUESTIONS?