Parameter Estimation in Gaussian Process Regression for Deterministic Functions

Toni Karvonen

Department of Mathematics and Statistics
University of Helsinki, Finland

UQSay #37
Paris-Saclay University

18 November 2021

UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE

Table of contents

Modelling Deterministic Functions

- Let $\Omega \subset \mathbb{R}^d$.
- Let $f : \Omega \to \mathbb{R}$ be a deterministic function.
- Suppose that f generates the noiseless data

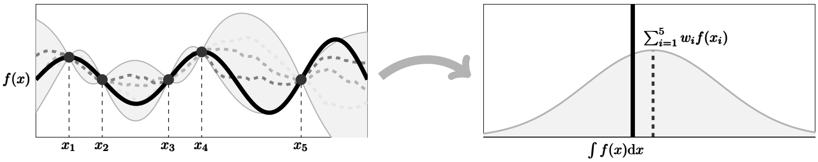
$$\mathcal{D}_N = \big\{(x_1, f(x_1)), \ldots, (x_N, f(x_N))\big\} \text{ at some } x_i \in \Omega.$$

In this talk we model f using a Gaussian process f_{GP}.

- f need not be a sample from (or in any way related to) f_{GP}.
- The assumption that there is no noise is crucial.

Motivation

- *Probabilistic numerics.* Provide quantification of epistemic uncertainty arising from discretisation in numerical approximation.
- *Modelling of computer experiments.* Predict the output of a complex and computationally expensive piece of code.
- *Bayesian optimisation.* Construct surrogates to an objective function.



Papers

(K2020) **Karvonen, Wynne, Tronarp, Oates & Särkkä (2020)**. Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):926–958.

(K2021a) **Karvonen (2021)**. Small sample spaces for Gaussian processes. *arXiv*:2103.03169.

(K2021b) **Karvonen (2021)**. Estimation of the scale parameter for a misspecified Gaussian process model. *arXiv*:2110.02810.

\+ Two papers in preparation.

Gaussian process interpolation I

- Let $K\colon \Omega \times \Omega \to \mathbb{R}$ be a positive-definite covariance kernel.
- Model f as a Gaussian process $f_{\mathrm{GP}} \sim \mathrm{GP}(0, K)$.
- Condition f_{GP} on the data $\mathcal{D}_N = \{(x_i, f(x_i))\}_{i=1}^N$.

The resulting conditional GP, $f_{\mathrm{GP}} \mid \mathcal{D}_N$, has the mean

$$s_{f,N}(x) := \mathbb{E}\big[f_{\mathrm{GP}}(x) \mid \mathcal{D}_N\big] = f_N^\mathsf{T} K_N^{-1} k_N(x) \tag{1}$$

and variance

$$P_N(x)^2 := \mathrm{Var}\big[f_{\mathrm{GP}}(x) \mid \mathcal{D}_N\big] = K(x,x) - k_N(x)^\mathsf{T} K_N^{-1} k_N(x), \tag{2}$$

where

$$(f_N)_i = f(x_i), \quad (k_N(x))_i = K(x, x_i) \quad \text{and} \quad (K_N)_{ij} = K(x_i, x_j).$$

Gaussian Process Interpolation II

- The conditional mean interpolates the data:

$$s_{f,N}(x_i) = f(x_i) \quad \text{for every} \quad i = 1, \ldots, N.$$

- The conditional variance vanishes at the data locations:

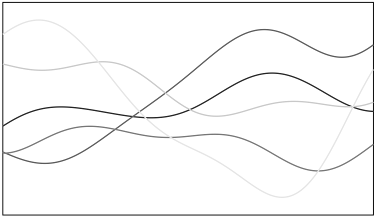
$$P_N(x_i)^2 = 0 \quad \text{for every} \quad i = 1, \ldots, N.$$

- The conditional variance does not depend on the data values $f(x_i)$.

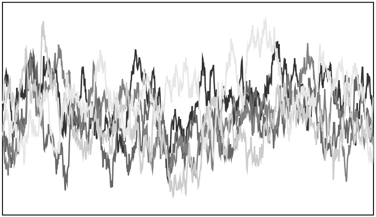
- Properties of the kernel K define the properties of f_{GP}.

Gaussian process priors

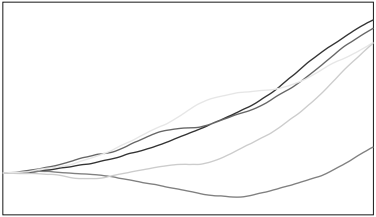
Gaussian: $K(x, y) = \mathrm{e}^{-(x-y)^2/2}$



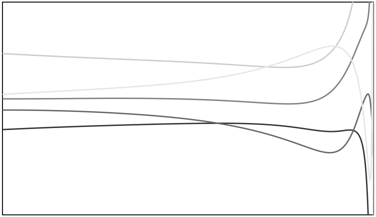
Matérn: $K(x, y) = \mathrm{e}^{-|x-y|}$

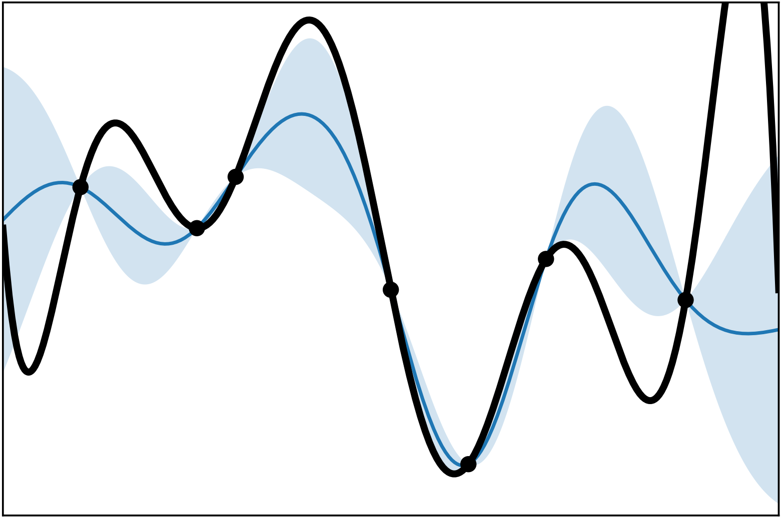


BM: $K(x, y) = \frac{\min\{x,y\}^3}{3} + \frac{|x-y| \min\{x,y\}^2}{2}$



Hardy: $K(x, y) = \frac{1}{1-xy}$





Objective

- Suppose that the prior covariance is parametric: $f_{\mathrm{GP}} \sim \mathrm{GP}(0, K_\theta)$.
- The conditional process is

$$f_{\mathrm{GP}} \mid \mathcal{D}_N \sim \mathrm{GP}(s_{\theta,f,N}, P^2_{\theta,N}).$$

- For any $a \in (0,1)$,

$$\mathbb{P}\left[|f_{\mathrm{GP}}(x) - s_{\theta,f,N}(x)| \leq c(a) P_{\theta,N}(x,x) \;\middle|\; \mathcal{D}_N \right] = 1 - a.$$

- Compute hyperparameter estimates $\theta(f,N)$ of θ.

Objective: Understand the behaviour of (i) $\theta(f,N)$ and (ii) the standard score

$$\frac{|f(x) - s_{\theta(f,N),f,N}(x)|}{P_{\theta(f,N),N}(x,x)} \quad \text{as} \quad N \to \infty \tag{3}$$

for different f and hyperparameter estimation methods.

Table of contents

Reproducing kernel Hilbert spaces

Reproducing kernel Hilbert space

Every covariance kernel $K\colon \Omega \times \Omega \to \mathbb{R}$ induces a unique *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}(K)$ with inner product $\langle \cdot, \cdot \rangle_K$. The RKHS consists of functions $g\colon \Omega \to \mathbb{R}$ and the kernel has the *reproducing property*

$$\big\langle g, K(\cdot, x) \big\rangle_K = g(x) \quad \text{for any} \quad g \in \mathcal{H}(K) \ \text{ and } \ x \in \Omega.$$

The GP conditional moments are related to optimal interpolation in $\mathcal{H}(K)$.

$s_{f,N}$ = minimum-norm interpolant

$$s_{f,N} = \operatorname*{arg\,min}_{s \in \mathcal{H}(K)} \big\{ \|s\|_K \ : \ s(x_i) = f(x_i) \text{ for every } i = 1, \ldots, N \big\}$$

$P_N(x,x)$ = worst-case error

$$P_N(x,x) = \sup_{\|g\|_K \leq 1} |g(x) - s_{g,N}(x)|$$

Sobolev spaces

The Sobolev space $H^\alpha(\mathbb{R}^d)$ consists of functions $g \in L^2(\mathbb{R}^d)$ such that

$$\|g\|_\alpha^2 := \int_{\mathbb{R}^d} \left(1 + \|\xi\|^2\right)^\alpha |\widehat{g}(\xi)|^2 \,\mathrm{d}\xi < \infty.$$

- For $\Omega \subset \mathbb{R}^d$ the space $H^\alpha(\Omega)$ is defined via restrictions.
- If $\alpha > n + d/2$ for $n \in \mathbb{N}$, then $H^\alpha(\mathbb{R}^d) \subset C^n(\mathbb{R}^d)$.

Sobolev kernel

Let $\alpha > d/2$ and $\Omega \subset \mathbb{R}^d$. A kernel $K \colon \Omega \times \Omega \to \mathbb{R}$ is a *Sobolev kernel of order α* if its RKHS $\mathcal{H}(K)$ is norm-equivalent ($\simeq$) to $H^\alpha(\Omega)$.

A Matérn kernel

$$K(x, y) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,\|x - y\|}{\lambda} \right)^\nu \mathrm{K}_\nu \left(\frac{\sqrt{2\nu}\,\|x - y\|}{\lambda} \right)$$

of smoothness $\nu > 0$ is a Sobolev kernel of order $\nu + d/2$.

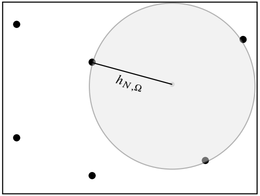
Sobolev rates

Suppose that

- $\Omega \subset \mathbb{R}^d$ is bounded and sufficiently regular (e.g., $\Omega = [0,1]^d$).
- The points $x_1, \ldots, x_N$ are quasi-uniform: the *fill-distance*

$$h_{N,\Omega} = \sup_{x \in \Omega} \min_{i=1,\ldots,N} \|x - x_i\|$$

is of order $N^{-1/d}$ as $N \to \infty$.



Theorem (from approximation theory)

Let $d/2 < \beta \leq \alpha$. Suppose that $\mathcal{H}(K) \simeq H^{\alpha}(\Omega)$ and $f \in H^{\beta}(\Omega)$. Then

$$\sup_{x \in \Omega} |f(x) - s_{f,N}(x)| \leq C_1 \|f\|_{\beta} N^{-\beta/d+1/2} \tag{4}$$

and

$$C_2 N^{-\alpha/d+1/2} \leq P_N(x) \leq C_3 N^{-\alpha/d+1/2} \quad \text{for} \quad x \notin \{x_i\}_{i=1}^{\infty}. \tag{5}$$

Why parameter estimation is necessary

If K is a Sobolev kernel of order α and $f \in H^{\beta}(\Omega)$ for $\beta \leq \alpha$, then

$$\frac{|f(x) - s_{f,N}(x)|}{P_N(x,x)} \leq \frac{C_1 \, \|f\|_{\beta} \, N^{-\beta/d+1/2}}{C_2 N^{-\alpha/d+1/2}} = C_4 N^{(\alpha-\beta)/d}.$$

$\Longrightarrow$ If $f \notin \mathcal{H}(K) \simeq H^{\alpha}(\Omega)$, it may be necessary estimate kernel parameters in order to prevent overconfidence:

$$\frac{\alpha - \beta}{d} > 0.$$

Maximum likelihood estimation

The log-likelihood function is

$$\ell(\theta) = -\frac{1}{2}\left[f_N^\mathsf{T} K_{\theta,N}^{-1} f + \log\det K_{\theta,N} + N\log(2\pi) \right],$$

where $(f_N)_i = f(x_i)$ and $(K_{\theta,N})_{ij} = K_\theta(x_i, x_j)$.

Fix K and use the simple parametrisation $K_\sigma = \sigma^2 K$ for $\sigma \geq 0$. Then

$$s_{\sigma,f,N}(x) = s_{f,N}(x) \quad \text{and} \quad P_{\sigma,N}(x,y) = \sigma P_N(x,y).$$

$$\sigma_{\mathrm{ML}}(f,N) = \operatorname*{arg\,max}_{\sigma\geq 0} \ell(\sigma) = \sqrt{\frac{f_N^\mathsf{T} K_N^{-1} f_N}{N}} = \frac{\|s_{f,N}\|_K}{\sqrt{N}} \tag{6}$$

and

$$\text{standard score} = \frac{|f(x) - s_{f,N}(x)|}{\sigma_{\mathrm{ML}}(f,N) P_N(x,x)} \tag{7}$$

First results — $f \in \mathcal{H}(K)$

Proposition (Proposition 3.1 in K2020)

If $f \in \mathcal{H}(K)$ and $f(x_i) \neq 0$ for some x_i, then there is $c > 0$ such that

$$c\, N^{-1/2} \leq \sigma_{\mathrm{ML}}(f, N) \leq \|f\|_K\, N^{-1/2}. \tag{8}$$

Theorem (Theorem 3.2 in K2020)

If $f \in \mathcal{H}(K)$ and $f(x_i) \neq 0$ for some x_i, then there is $C > 0$ such that

$$\frac{|f(x) - s_{f,N}(x)|}{\sigma_{\mathrm{ML}}(f, N) P_N(x, x)} \leq C\sqrt{N}. \tag{9}$$

At most "slow" (i.e., $\sqrt{N}$) overconfidence if $f \in \mathcal{H}(K)$.

Second result — $f \in H^{\beta}(\Omega)$ and $\mathcal{H}(K) \simeq H^{\alpha}(\Omega)$

Let $\beta \in (d/2, \alpha]$. Suppose that Ω is regular and $\{x_i\}_{i=1}^{\infty}$ are quasi-uniform.

Proposition (Proposition 4.5 in K2020)

If $\mathcal{H}(K) \simeq H^{\alpha}(\Omega)$ and $f \in H^{\beta}(\Omega)$, then

$$\sigma_{\mathrm{ML}}(f, N) \leq C_1 N^{(\alpha-\beta)/d-1/2} \, \|f\|_{H^{\beta}(\Omega)} . \tag{10}$$

$\Longrightarrow$ the behaviour of $\sigma_{\mathrm{ML}}(f, N)$ tells how "far" from $\mathcal{H}(K)$ the function f is.

Theorem (Theorem 4.10 in K2020)

If $\mathcal{H}(K) \simeq H^{\alpha}(\Omega)$ and f has "exact smoothness" β for $\lfloor \beta \rfloor > d/2$, then

$$\frac{|f(x) - s_{f,N}(x)|}{\sigma_{\mathrm{ML}}(f, N) P_N(x, x)} \leq C_2(f) (\log N)^{\alpha/(2\beta)} \sqrt{N}. \tag{11}$$

At most "slow" (i.e., $\approx \sqrt{N}$) overconfidence if f is rougher than $\mathcal{H}(K)$.

Some implications

- Overconfidence cannot be ruled out, but at least it cannot be overly severe: the standard score is approximately $\mathcal{O}(N^{1/2})$.

- Simple maximum likelihood estimation of a scaling parameter provides strong protection against smoothness misspecification.

- MLE does not detect undersmoothing by the model:

$$\sigma_{\mathrm{ML}}(f, N) \asymp N^{-1/2} \quad \text{for every non-zero} \quad f \in \mathcal{H}(K).$$

- It can be shown **(Theorem 4.11 in K2020)** that underconfidence occurs if there is (roughly speaking) sufficient undersmoothing:

$$\mathcal{H}(K) \simeq H^{\alpha}(\Omega) \quad \text{and} \quad f \in H^{2\alpha}(\Omega).$$

Numerical results: confidence intervals

$$\mathcal{H}(K) \simeq H^{2}([0,1]) \text{ and } f \in H^{\beta}([0,1]) \text{ for } \beta = \frac{2}{2}, \frac{3}{2}, \ldots, \frac{6}{2}.$$

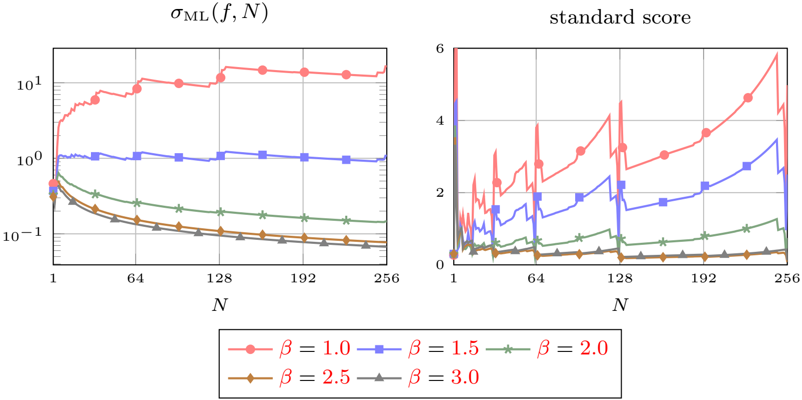


Table of contents

Sample path properties

Theorem (e.g., Steinwart 2019)

Let $f_{\mathrm{GP}} \sim \mathrm{GP}(0, K)$ and suppose that K is a Sobolev kernel of order $\alpha > d/2$. Then

$$\mathbb{P}\big[f_{\mathrm{GP}} \in H^{\beta}(\Omega)\big] = 0 \quad \text{if} \quad \beta \geq \alpha - d/2$$

and

$$\mathbb{P}\big[f_{\mathrm{GP}} \in H^{\beta}(\Omega)\big] = 1 \quad \text{if} \quad \beta < \alpha - d/2.$$

The samples of the Gaussian process f_{GP} are therefore "$d/2$ less smooth" than the RKHS $\mathcal{H}(K)$.

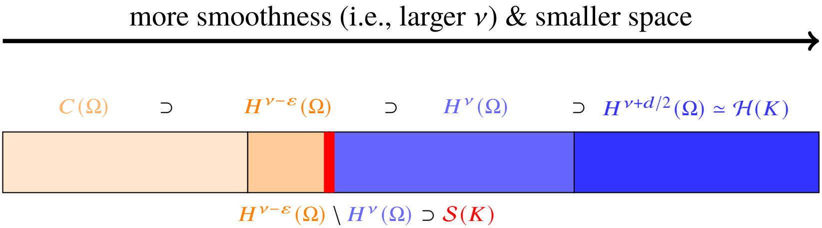
$\Longrightarrow$ Samples are *not* in the RKHS $\mathcal{H}(K) \simeq H^{\alpha}(\Omega)$!

$\Longrightarrow$ Samples have "exact smoothness" $\alpha - d/2$ if $\mathcal{H}(K) \simeq H^{\alpha}(\Omega)$.

Matérn kernels

Let K be a Matérn kernel of smoothness $\nu > 0$ and $\Omega \subset \mathbb{R}^d$ sufficiently regular. Then

- $\mathcal{H}(K)$ is norm-equivalent to the Sobolev space $H^{\nu+d/2}(\Omega)$.
- $\mathbb{P}[f_{\mathrm{GP}} \in H^{\nu-\varepsilon}(\Omega)] = 1 \quad$ if and only if $\quad \varepsilon > 0$.



The samples of f_{GP} can be thought of as elements of the "Sobolev slice"

$$\mathcal{S}(K) \subset H^{\nu-\varepsilon}(\Omega) \setminus H^{\nu}(\Omega) \quad \text{for every} \quad \varepsilon > 0.$$

Expected scale MLE for samples

Suppose that the data-generating function is a Gaussian process

$$f = f_{\mathrm{GP}}^* \sim \mathrm{GP}(0, R)$$

but the model is $f_{\mathrm{GP}} \sim \mathrm{GP}(0, K)$. Then

$$\mathbb{E}_{f_{\mathrm{GP}}^*}\big[\sigma_{\mathrm{ML}}(f_{\mathrm{GP}}^*, N)^2\big] = \mathbb{E}_{f_{\mathrm{GP}}^*}\left[\frac{(f_{\mathrm{GP},N}^*)^{\mathsf{T}} K_N^{-1} f_{\mathrm{GP},N}^*}{N}\right] = \frac{\mathrm{trace}(R_N K_N^{-1})}{N}.$$

Theorem (Theorem 4.2 in K2021b)

Suppose that $\mathcal{H}(K) \simeq H^{\alpha}([0,1]^d)$ and $\mathcal{H}(R) \simeq H^{\alpha_0}([0,1]^d)$ for $\alpha \geq \alpha_0 > d/2$. If the points $\{x_i\}_{i=1}^{\infty}$ are quasi-uniform, then

$$C_1 N^{2(\alpha-\alpha_0)/d} \leq \mathbb{E}_{f_{\mathrm{GP}}^*}\big[\sigma_{\mathrm{ML}}(f_{\mathrm{GP}}^*, N)^2\big] \leq C_2 N^{2(\alpha-\alpha_0)/d}.$$

Comparison to the deterministic case

Suppose that $\mathcal{H}(K) \simeq H^{\alpha}([0,1]^d)$ and that $\{x_i\}_{i=1}^{\infty}$ are quasi-uniform.

Recall that the samples "have smoothness $\alpha - d/2$".

Deterministic

Let $f \in H^{\alpha - d/2}([0,1]^d)$ so that f *does not have less smoothness* than the samples of $f_{\mathrm{GP}} \sim \mathrm{GP}(0, K)$. For $\beta = \alpha - d/2$ we get

$$\sigma_{\mathrm{ML}}(f, N)^2 \leq C_1 N^{2(\alpha - \beta)/d - 1} \|f\|_{H^{\beta}(\Omega)}^2 = C_1 \|f\|_{H^{\alpha - d/2}(\Omega)}^2. \quad (12)$$

Stochastic

Let $f = f_{\mathrm{GP}}^* \sim \mathrm{GP}(0, R)$ for R such that $\mathcal{H}(R) \simeq H^{\alpha}([0,1]^d)$. Then $f_{\mathrm{GP}} \sim \mathrm{GP}(0, K)$ and f_{GP}^* *have similar paths*. For $\alpha_0 = \alpha$ we get

$$\mathbb{E}_{f_{\mathrm{GP}}^*}\left[\sigma_{\mathrm{ML}}(f_{\mathrm{GP}}^*, N)^2\right] \asymp N^{(\alpha - \alpha_0)/d} = 1. \quad (13)$$

Table of contents

Is cross-validation better than MLE?

The leave-one-out cross-validated estimate of σ is

$$\sigma_{\mathrm{CV}}(f, N) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{f(x_i) - s_{f,N\setminus i}(x_i)}{P_{N\setminus i}(x_i, x_i)} \right)^2,$$

where $s_{f,N\setminus i}$ and $P_{N\setminus i}$ are the GP conditional mean and std based on the data set $\mathcal{D}_N \setminus \{(x_i, f(x_i))\}$.

- Recall that for any non-zero $f \in \mathcal{H}(K)$ we have

$$\sigma_{\mathrm{ML}}(f, N) \asymp N^{-1/2} \quad \text{for any non-zero} \quad f \in \mathcal{H}(K).$$

- At least in some cases it can be proved[1] that the rate of decay of

$$\sigma_{\mathrm{CV}}(f, N) \quad \text{depends on} \quad f \in \mathcal{H}(K).$$

$\Longrightarrow$ Cross-validation is more sensitive to the smoothness of f than MLE.

[1] Work in progress with M. Naslidnyk, M. Mahsereci and M. Kanagawa.

Estimating the lengthscale parameter

The kernel K is stationary if it can be written as

$$K(x, y) = \Phi\left(\frac{x - y}{\lambda}\right) \quad \text{for some} \quad \Phi\colon \mathbb{R}^d \to \mathbb{R},$$

where $\lambda > 0$ is the lengthscale parameter.

Theorem (to appear in a paper with C. Oates)

Let $N \geq 2$ and suppose that $\mathcal{H}(K) \simeq H^\alpha(\mathbb{R}^d)$ for some $\alpha > d/2$. If the data vector is constant,

$$f_N = (c, \ldots, c) \in \mathbb{R}^N \quad \text{for some} \quad c \in \mathbb{R},$$

then

$$\lambda_{\mathrm{ML}} = \infty.$$

Conclusion

- Simple MLE of the scaling parameter σ in $K_\sigma(x, y) = \sigma^2 K(x, y)$ provides protection against misspecification.
- Overconfidence is possible but at least it cannot be too severe.
- Samples from a GP are elements of a "small" set of functions. This is manifested in the sample results being "nicer".
- Cross-validation may be more sensitive to the smoothness of f than MLE.

Additional references

- **Lukić & Beder (2001)**. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969.
- **Narcowich, Ward & Wendland (2006)**. Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, 24(2):175–186.
- **Bachoc (2013)**. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69.
- **Xu & Stein (2017)**. Maximum likelihood estimation for a smooth Gaussian random field model. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):138-175.
- **Steinwart (2019)**. Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395.
- **Hadji & Szábo (2019)**. Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel? *arXiv:2002.01381*.
- **Teckentrup (2020)**. Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1310–1337.
- **Wang (2020)**. On the inference of applying Gaussian process modeling to a deterministic function. *arXiv:2002.01381*.

Thank you for your attention!