

An Empirical Case of Gaussian Processes Learning in High Dimension: the Likelihood versus Leave-One-Out Rivalry

David Gaudrie¹, Rodolphe Le Riche², Tanguy Appriou^{1,2}

¹Stellantis

²CNRS LIMOS, Mines Saint-Étienne, Fr

SIAM UQ, Trieste, Feb 29th 2024



STELLANTIS



CIROQUO consortium



This work is licensed under a Creative Commons Attribution 4.0 International License.

Abstract I

Gaussian Processes (GPs) are semi-parametric models commonly employed in various applications such as statistical modeling, sensitivity analysis and Bayesian optimization. GPs are particularly useful in the context of small data. However, GPs suffer particularly from the curse of dimensionality: at a fixed number of data points, their predictive capability may decrease dramatically after 40 dimensions.

In this talk, we investigate such a phenomenon in details. We illustrate the loss of performance with increasing dimension on simple quadratic functions and analyze its underlying symptoms, in particular a tendency to become constant away from the data points.

We show that the fundamental problem is one of learning and not one of representation capacity: maximum likelihood, the dominant loss function for such models, can miss regions of optimality of the GP hyperparameters. Failure of maximum likelihood is related to statistical model inadequacy: a model with constant trend is sensitive to dimensionality when fitting quadratic functions while it much better handles dimension growth for linear functions or Gaussian trajectories generated with the right covariance. Our experiments also show that the leave-one-out loss function is less prone to the curse of dimensionality even for inadequate statistical models.

Abstract II

A first step towards analyzing the curse of dimensionality in this context is taken. It considers a uniform sampling of the data points. As dimension increases, the cross-covariance terms concentrate around a mean value. This mean value is calculated and defines a limit iso-covariance. The iso-covariance GP model has closed-form expressions for its prediction, likelihood and leave-one-out error. It allows to explain why the a priori mean must increase with dimension.

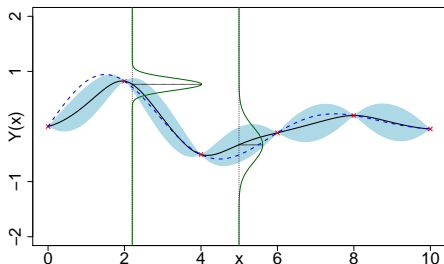
The kernel model

A classical model for *small data* (x^i, y_i) , $i = 1, \dots, n$, $n \leq \mathcal{O}(1000)$:

$$m(x) = \mu(x) + [k(x, x^1) \dots k(x, x^n)] K^{-1} \begin{pmatrix} y_1 - \mu(x^1) \\ \dots \\ y_n - \mu(x^n) \end{pmatrix}$$

where $K_{i,j} = k(x^i, x^j)$

Example: $m(x)$,
mean of a kriging (con-
ditionned Gaussian pro-
cess) model



Why we like kernel-based predictions

- Simple and efficient : Best Linear Unbiased Estimator knowing (x^i, y_i) , $i = 1, \dots, n$
- Classical. Similar forms appears in: weighted least squares, Bayesian linear model, radial basis function, Support Vector Machine, kernel based regression.
- It is a pre-trained model: has interpolation as a structural property

(contracted notation)

$$m(x) = \mu(x) + k(x, \mathbb{X})k(\mathbb{X}, \mathbb{X})^{-1}(y - \mu(\mathbb{X}))$$

where $\mathbb{X} \equiv \{x^1, \dots, x^n\}$

how does GP mean perform in high-dimension?

(GP \equiv Gaussian Process)

- dimension = dimension of x , written d
- Known to work well in low dimension ($d \leq 10$), still an open question beyond:
[Binois and Picheny, 2024, Durrande et al., 2012] for GPs,
[Le Riche and Picheny, 2021] for GPs in optimization.

\Rightarrow this talk is an investigation of this question

Default GP model

A pragmatic choice, corresponding to a typical default:

- The kernel k is the isotropic Matérn 5/2

$$k_{\theta}(x, x') = \left(1 + \frac{\sqrt{5}\|x-x'\|}{\theta} + \frac{5\|x-x'\|^2}{3\theta^2} \right) \exp\left(-\frac{\sqrt{5}\|x-x'\|}{\theta}\right)$$

where θ is the “length-scale”

- $\mu(x) = \mu = \text{constant}$
- 2 hyperparameters: θ, μ . We can plot maps 😊
- The 2 hyperparameters, θ & μ , are learned by minimizing minus the log-likelihood

Default tests

A pragmatic choice, corresponding to a typical default:

- Number of data points : $n = 100$ (costly setting, e.g., CFD)
- Design of Experiments : uniform sampling within $[0, 1]^d$
- Sphere functions to start with:

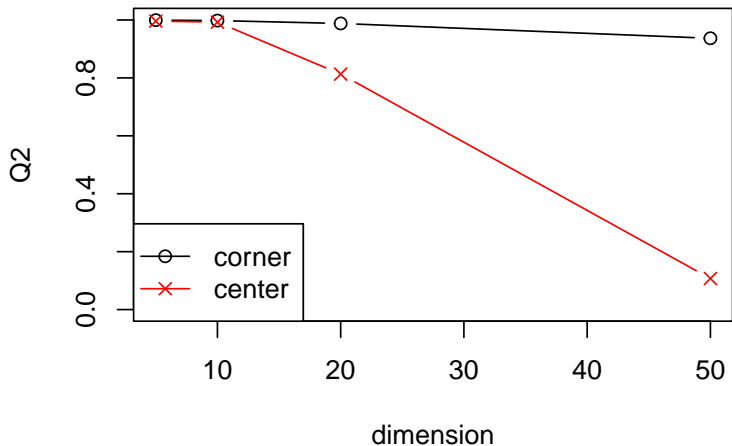
$$y_{05}(x) = \sqrt{\|x - 0.5 \times \mathbf{1}_d\|^2} \quad , \quad y_0(x) = \sqrt{\|x\|^2}$$

(isotropic like the GP)

- Figure of merit : $Q2 = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} (y_i^{\text{test}} - m(x_i^{\text{test}}))^2}{\sum_{i=1}^{n_{\text{test}}} (y_i^{\text{test}} - \overline{y^{\text{test}}})^2}$

A curse of dimensionality (1/3)

Prediction quality of the default GP for the (square root of) sphere with a minimum at the corner (y_0) or at the center (y_{05}):

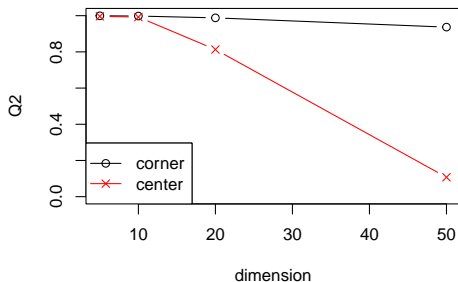


A curse of dimensionality (2/3)

- Why such a collapse in prediction only because of a translation in the quadratic function?

The intrinsic dimension of the problem does not change.

- Can we do something simple and general to avoid it ?



A curse of dimensionality (3/3)

The collapse in Q2 when d grows

- **also happens** for the Ackley and Rastrigin functions in 5D extended to any dimension by inactive variables
 $f(x) = f(x_1, \dots, x_5) \Rightarrow$ a geometrical effect ?
- **also happens** for other design of experiments : Latin Hypercube Sampling, balanced uniform sampling within nested slices of same volume.
- **does not happen** for linear functions, well-specified GPs (here with a constant a priori mean and Matérn 5/2 covariance), for misspecified GPs (exponential covariance) but the Q2 is very quickly low.

The phenomenon is general but our examples will only use the simple sphere case with uniform sampling.

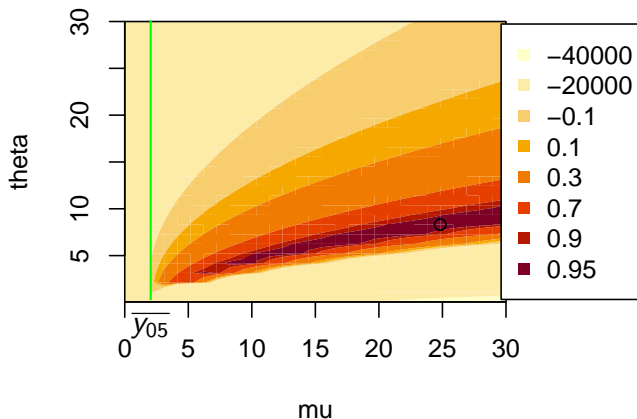
Representability question (1/2)

Is there a GP model, i.e., a choice of (μ, θ) , i.e., a mean and covariance choice within the parameterized set, that correctly represents y_{05} ($0.95 \leq Q2 \leq 1$) when $n = 100$ uniform and $d = 50$?

- If yes, how to find the right (μ, θ) ?
- If no,
 - Look at another covariance structure (like in [Binois and Picheny, 2024, Durrande et al., 2012, Bouhlel et al., 2016])
 - Look at other design of experiments

Representability question (2/2)

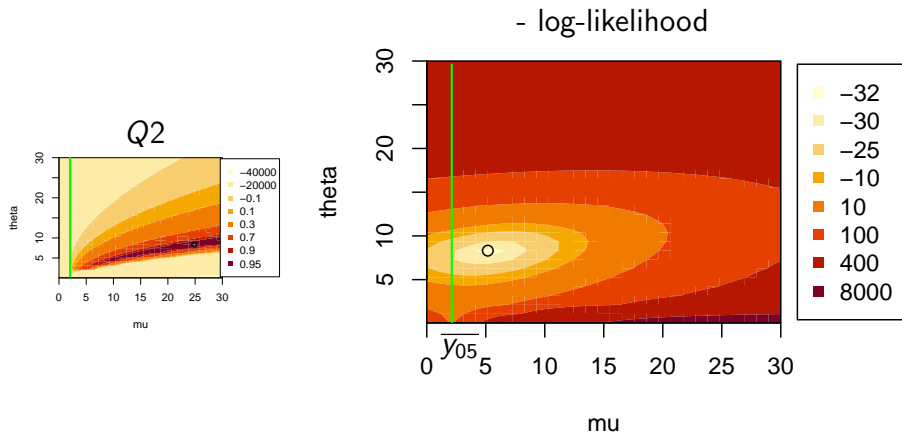
A Q2 map provides the answer in our simple context, $d = 50$:



There exists a narrow valley of good models, the best one at $(\mu^* = 24.8, \theta^* = 8.3)$, $Q2 = 0.99$. Note : μ^* much larger than $\overline{y_{05}}$.

MLE and the centered sphere y_{05} (1/2)

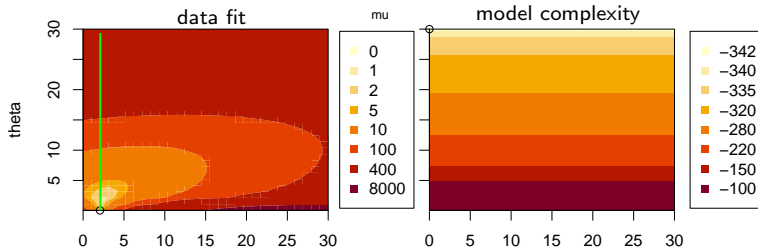
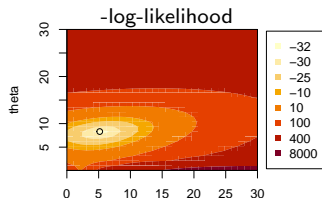
Why minimizing minus log-likelihood does not find the right model (empirical, $d = 50$) ?



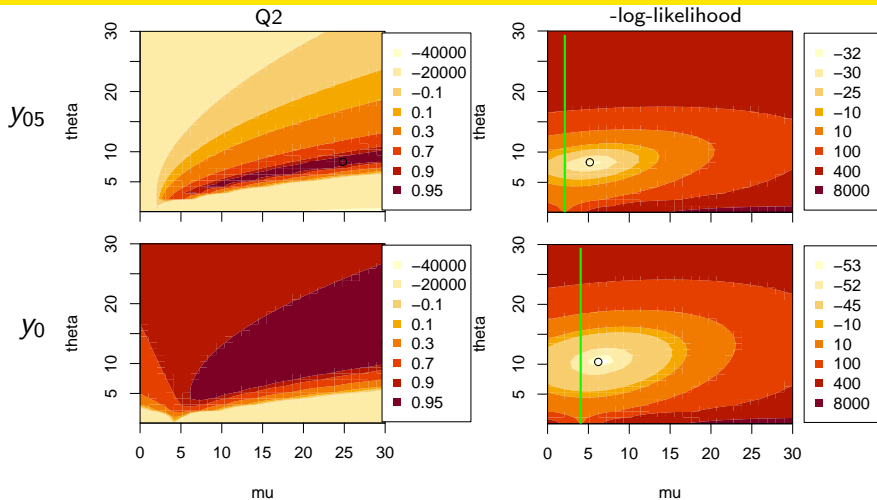
The MLE is not in the right valley

MLE and the centered sphere y_{05} (2/2)

$$-\mathcal{L}(\mu, \theta; \mathbb{X}, y) = \underbrace{\frac{1}{2}(y - \mu \mathbf{1})^\top K_\theta^{-1}(y - \mu \mathbf{1})}_{\text{data fit}} + \underbrace{\frac{1}{2} \log(\det(K_\theta)) + \frac{n}{2} \log(2\pi)}_{\text{model complexity}}$$



Sphere centered versus at corner, y_{05} vs. y_0 , ($d=50$)



Wider Q^2 optimal valley for y_0 .
Probabilistic model better fits y_0 than y_{05} .

Learning the right model

Try the leave-one-out loss because

- When the probabilistic model is not well specified, leave-one-out loss function should be preferred to the negative log-likelihood [Bachoc, 2013].
- Efficient (same complexity as likelihood) formula for calculating the leave-one-out error [Dubrule, 1983]:

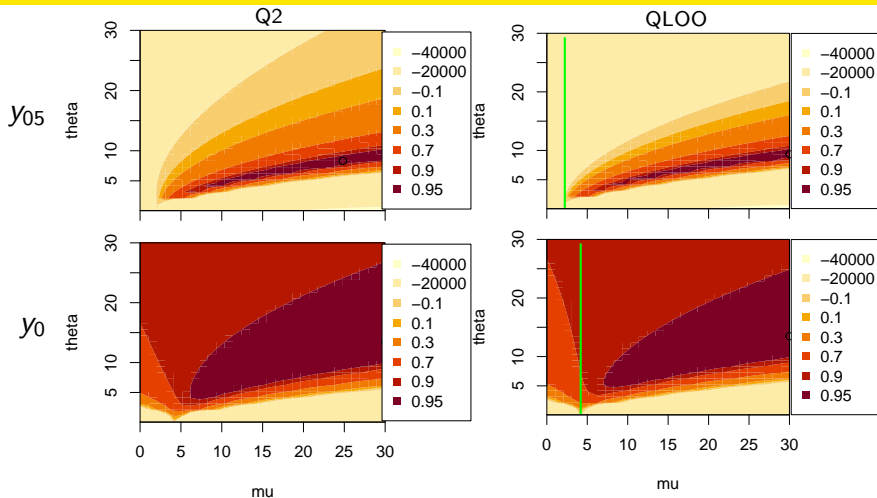
$$y_i - m_{(-i)}(x_i) = \frac{[k_{\theta}(\mathbb{X}, \mathbb{X})^{-1}(y - \mu \mathbf{1}_n)]_i}{[k_{\theta}(\mathbb{X}, \mathbb{X})^{-1}]_{i,i}}$$

- $\|y - m_{(-\cdot)}(\mathbb{X})\|^2$ is a quadratic form, like the Q2.

Normalize it like the Q2:

$$QLOO = 1 - \frac{\|y - m_{(-\cdot)}(\mathbb{X})\|^2}{\|y - \bar{y} \mathbf{1}_n\|^2}$$

Q2 versus LOO maps ($d = 50$)



The leave-one-out maps coincide closely with the Q2 maps for both y_0 and y_{05} . Both Q2 and LOO have a quadratic form.

Related analyses (1/2)

- Work on the consistency of the ML and Cross-Validation (CV) estimators when the underlying data comes from a GP and n is large: [Wahba, 1985, Stein, 1990, Naslidnyk et al., 2023].
- Cases when ML is superior to CV [Stein, 1990], Chap. 3 of [Santner et al., 2003]; and vice versa [Wahba, 1985]
- Modified versions of ML and CV tend towards each other: LOO accounting for terms correlations [Ginsbourger and Schärer, 2021] and Bayesian ML [Fong and Holmes, 2020].

Related analyses (2/2)

- In short: when the model is well-specified, ML is more accurate, but CV is more robust to model misspecification [Bachoc, 2013, Martin and Simpson, 2005, Naslidnyk et al., 2023].
- Learning a GP in a non-asymptotic case in terms of data, n , is still an open research question : [Karvonen and Oates, 2023] show that the correlation lengths may become infinite with a function that is a constant shift from $\mu(x)$.

Here, we provide a partial analysis for large d (and finite n) with a y not coming from a GP.

Concentration of covariances (1/3)

If the x^i are randomly sampled (uniformly, LHS, ...), as $d \nearrow$, the pairwise distances and the covariances concentrate at a given value.

Pairwise distances for a uniform distribution:

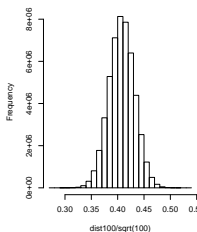
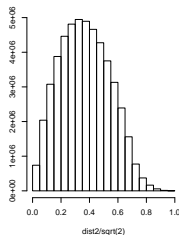
- Let X_i and $X'_i \sim \mathcal{U}([0, 1])$, $i = 1, \dots, d$ independently

- $R^2 := \sum_{i=1}^d (X_i - X'_i)^2$ squared euclidean distance between points

- $\mathbb{E}[(X_i - X'_i)^2] = 1/6$,
 $\mathbb{V}[(X_i - X'_i)^2] = 7/180$

- CLT:

$$R^2 = d \times R^2/d \underset{d \rightarrow \infty}{\sim} d \times \mathcal{N}(1/6, 7/(180d))$$



Concentration of covariances (2/3)

When $X \sim \mathcal{U}[0, 1]^d$, as the dimension d increases,

$$\|x^i - x^j\| \rightarrow \sqrt{d/6} \quad , \quad i \neq j,$$

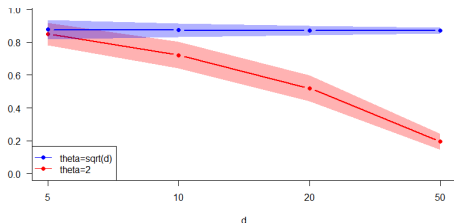
The radial Matérn 5/2 covariance becomes

$$k_\theta(x^i, x^j) \xrightarrow{d \nearrow} c(d, \theta) \equiv \left(1 + \sqrt{\frac{5d}{6}} \frac{1}{\theta} + \frac{5d}{18} \frac{1}{\theta^2} \right) \exp \left(-\sqrt{\frac{5d}{6}} \frac{1}{\theta} \right)$$

Concentration of covariances (3/3)

Tensorized Matérn 5/2 kernel : $k_{\theta}(x^i, x^j) = \bigotimes_{l=1}^d k_{j_{\theta}}(x_l^i, x_l^j)$

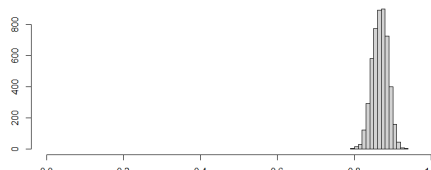
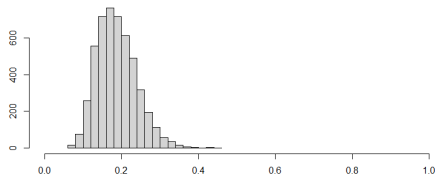
Covariances with the tensorized Matérn 5/2 kernel converge too:



Distribution of $k_{\theta}(x^i, x^j)$ when $d = 50$:

$$\theta = 2$$

$$\theta = \sqrt{d} \approx 7.1$$



Isocorrelation as a limit case

The limit covariance is $\hat{K} = \sigma^2 \hat{R}$ where $(c(d, \theta)$ simplified as c)

$$K \xrightarrow{d \nearrow} \hat{K} \equiv \sigma^2 \begin{bmatrix} 1 & c & \dots & c \\ c & 1 & c & \dots \\ \dots & \dots & \dots & \dots \\ c & \dots & c & 1 \end{bmatrix}, \quad \hat{R} = (1 - c)I + c\mathbf{1}_n\mathbf{1}_n^\top$$

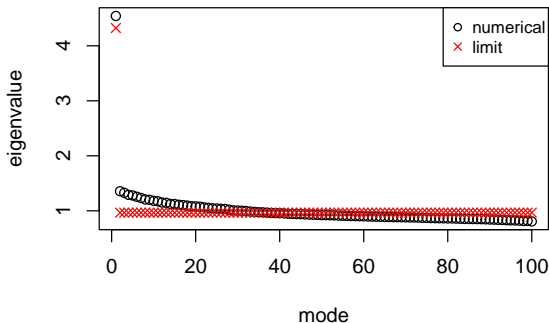
Calculations are analytical:

$$\hat{\lambda}_1 = \sigma^2[1 + (n - 1)c] \quad , \quad \hat{\lambda}_2 = \dots = \hat{\lambda}_n = \sigma^2(1 - c)$$
$$\hat{v}^1 = \mathbf{1}_n/\sqrt{n} \quad , \quad \hat{v}^i = [1 \ 0 \dots 0 \ \underset{i\text{-th}}{-1} \ 0 \dots 0]^\top / \sqrt{2}$$

$$\hat{R}^{-1} = \frac{1}{1 - c} \left(I - \frac{1}{1 + (n - 1)c} \mathbf{1}\mathbf{1}^\top \right)$$

Spectrum of K versus \hat{K}

The eigenvalues/vectors of K and \hat{K} are close.



- \times : spectrum of \hat{K} , the isocovariance matrix
- \circ : spectrum of K

Can the isocorrelation approximation help understand the geometrical problems of high-dimension?

Likelihood, isocorrelation case

In blue, the true function ; In red, the parameters $c(d, \theta)$ and μ

Project the true function onto the eigenvectors of \hat{R} : $y = \sum_{i=1}^n \beta_i \hat{v}^i$

Negative log-likelihood

$$-\hat{\mathcal{L}} = n \log(\sigma) + \frac{1}{2} \left[\frac{(\beta_1 - \mu \sqrt{n})^2}{1 + (n-1)c} + \frac{1}{1-c} \sum_{i=2}^n \beta_i^2 \right] + \frac{1}{2} (1 + (n-1)c) (1-c)^{n-1} + \frac{n}{2} \log(2\pi)$$

LOO, isocorrelation case

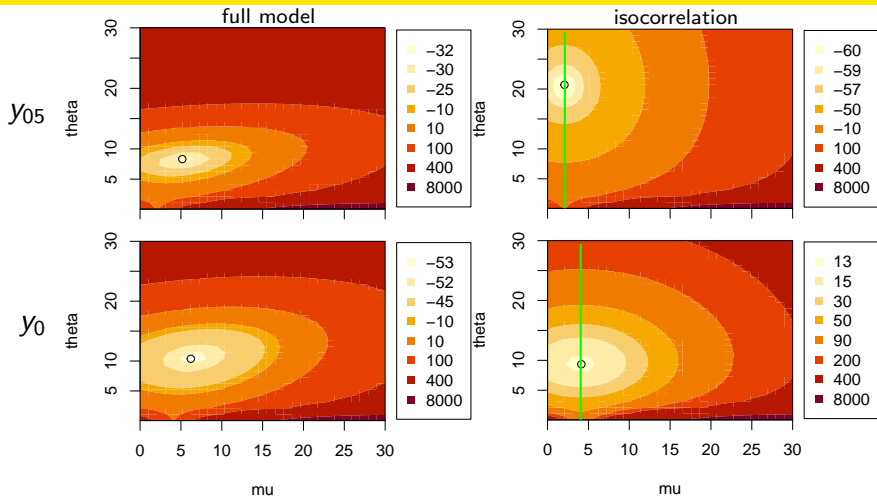
Project the true function onto the eigenvectors of \hat{R} : $y = \sum_{i=1}^n \beta_i \hat{v}^i$

LOO

$$y - m_{(-)}(\mathbb{X}) = \frac{1 - c}{1 + (n - 2)c} (\beta_1 - \mu \sqrt{n}) \hat{v}^1 + \frac{1 + (n - 1)c}{1 + (n - 2)c} \sum_{i=2}^n \beta_i \hat{v}^i$$

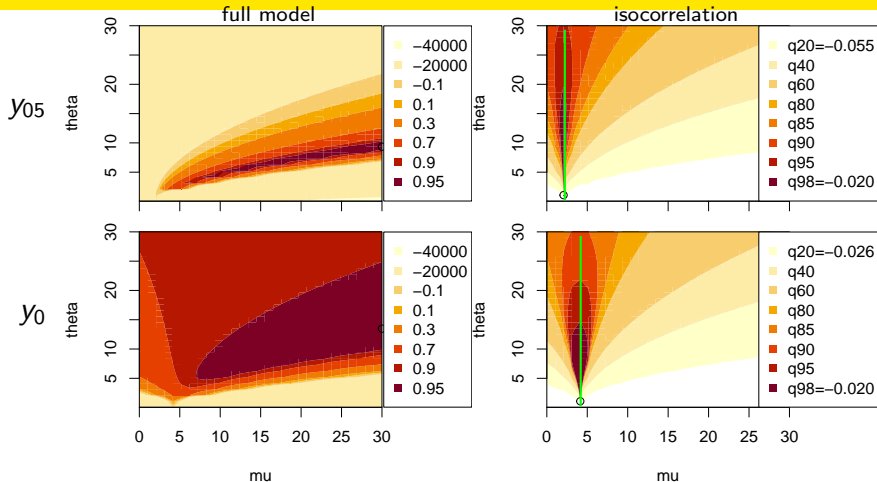
- The isocor. LOO error can be minimized in μ and c analytically.
- With the isocorrelation approximation, the likelihood and the LOO are very fast to calculate. We can do maps and complicated treatments with them.

Likelihood maps



Rough approximations. Best μ from isocorrelation is the empirical mean (can be seen from the formula as well).

QLOO maps

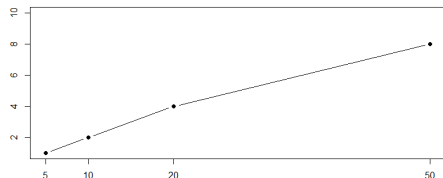


Very rough. Best isocorrelation model is the empirical mean.
Is the volume of the 98% level set of QLOO indicative of a problem that can be well modelled by the GP?

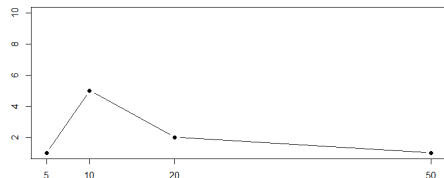
The increasing a priori mean effect

As d increases, the best μ according to Q2 grows beyond the MLE μ .

y_{05} , best μ from Q2 maximization



best μ from MLE



$$m(x) \approx \mu + r(x, \mathbb{X}) \hat{R}^{-1} (y - \mu \mathbf{1}_n) = \mu + \sum_{i=1}^n \hat{\gamma}_i r(x, x^i)$$

$$\hat{\gamma}_i = \frac{\bar{y} - \mu}{1 + (n-1)c} - \frac{\beta_i}{(1-c)\sqrt{2}}, \quad i = 2, \dots, n$$

The terms $\hat{\gamma}_i$ fade away if μ is not large enough i.e., $m(x)$ becomes insensitive to x .

Summary

- We have studied empirically the effects of the increase in dimension on prediction by kernel methods.
- In high-dimension, LOO leads to better models (i.t.o. Q2) than MLE.
- The a priori mean increases with dimension (beyond what MLE says).
- An analytical isocorrelation approximation, justified by the geometry of sampling in high dimension, has been described.

- Provide theoretical arguments to understand the accordance between Q2 and LOO in high-dimension: for which functions does it hold?
- Is the volume of a good level set of QLOO (with and without isocorrelation approximation) indicative of the difficulty to model the function in high-dimension ?
- Is the isocorrelation model useful to size θ as a function of d ?
- Study a probabilistic model of the covariance matrix when d is large (beyond the sheer mean c).

References I



Bachoc, F. (2013).

Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification.

[Computational Statistics & Data Analysis](#), 66:55–69.



Binois, M. and Picheny, V. (2024).

Combining additivity and active subspaces for high-dimensional gaussian process modeling.



Bouhlel, M. A., Bartoli, N., Otsmane, A., and Morlier, J. (2016).

Improving kriging surrogates of high-dimensional design models by partial least squares dimension reduction.

[Structural and Multidisciplinary Optimization](#), 53:935–952.



Dubrule, O. (1983).

Cross validation of Kriging in a unique neighborhood.

[Mathematical Geology](#), 15(6):687–699.



Durrande, N., Ginsbourger, D., and Roustant, O. (2012).

Additive covariance kernels for high-dimensional gaussian process modeling.

In [Annales de la Faculté des sciences de Toulouse: Mathématiques](#), volume 21, pages 481–499.



Fong, E. and Holmes, C. C. (2020).

On the marginal likelihood and cross-validation.

[Biometrika](#), 107(2):489–496.





Ginsbourger, D. and Schärer, C. (2021).


Fast calculation of gaussian process multiple-fold cross-validation residuals and their covariances.

[arXiv preprint arXiv:2101.03108](#).


References II


 Karvonen, T. and Oates, C. J. (2023).
Maximum likelihood estimation in gaussian process regression is ill-posed.
[Journal of Machine Learning Research](#), 24(120):1–47.


 Le Riche, R. and Picheny, V. (2021).
Revisiting Bayesian optimization in the light of the COCO benchmark.
[Structural and Multidisciplinary Optimization](#), 64(5):3063–3087.

 Martin, J. D. and Simpson, T. W. (2005).
Use of kriging models to approximate deterministic computer models.
[AIAA journal](#), 43(4):853–863.

 Naslidnyk, M., Kanagawa, M., Karvonen, T., and Mahsereci, M. (2023).
Comparing scale parameter estimators for gaussian process regression: Cross validation and maximum likelihood.
[arXiv preprint arXiv:2307.07466](#).

 Santner, T. J., Williams, B., and Notz, W. (2003).
[The design and analysis of computer experiments](#).
Springer, New York.

 Stein, M. L. (1990).
A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process.
[The Annals of Statistics](#), pages 1139–1157.

 Wahba, G. (1985).
A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem.
[The annals of statistics](#), pages 1378–1402.