Some considerations on Kriging, Constraints and Classification

Didier Rullière, Mines Saint-Etienne, LIMOS, drulliere@emse.fr
joint work with Marc Grossouvre, URBS

JS2O Days, 2025, April 2-4
Journées de Statistique et Optimisation en Occitanie

picture: mining headframe (chevalement) at Saint-Etienne

Outline

1. Kriging: a brief overview

2. Constrained Multi-Output Kriging

3. Constrained Classification

4. Numerical Illustrations

Kriging: a brief overview

The Origins of Kriging

Geostatistical problem

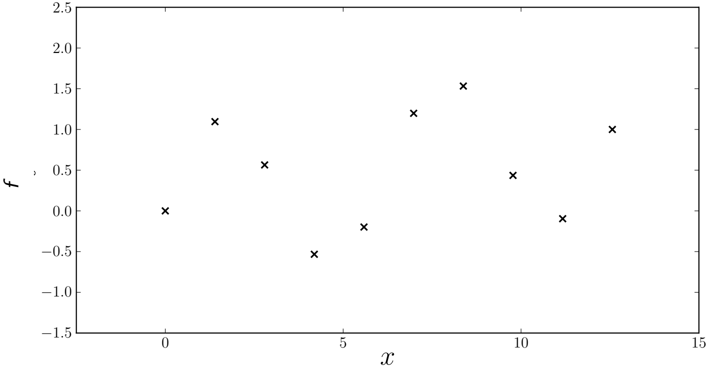
- How to predict gold concentration everywhere in the ground,
- ... if it is only measured on few specific sites?

- This is an interpolation problem.

Gaussian Process Regression (1/5)

Gaussian approach: Gaussian Process Regression (±ML community)

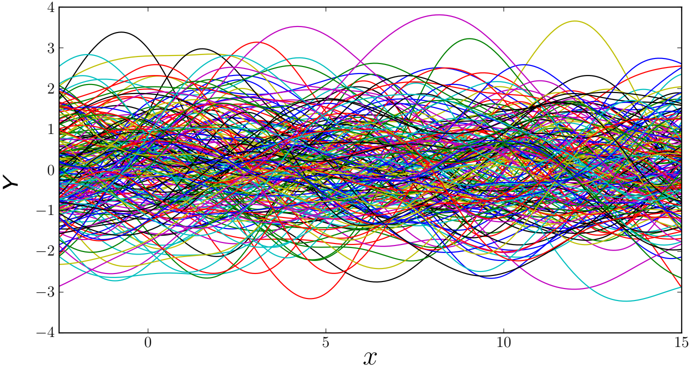
Assume we have observed a function $f(.)$ over a set of points $\mathbb{X} := (x_1, \ldots, x_n)^\top$:



Here x in 1D, f in 1D. The vector of observations is $\mathbf{y} := f(\mathbb{X})$, i.e. $y_i := f(x_i)$.

Gaussian Process Regression (2/5)

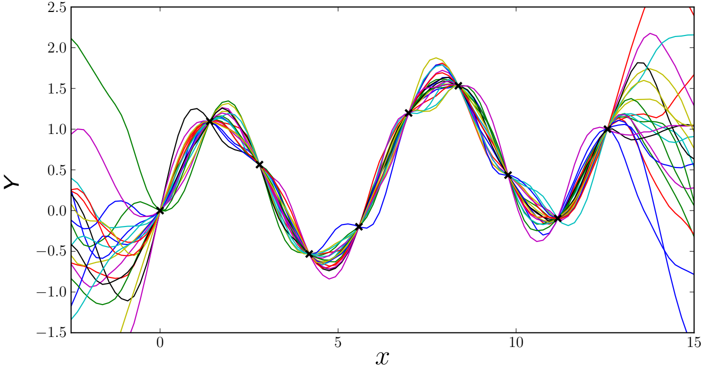
Since $f(.)$ in unknown, we make the assumption that it is close to the sample path of a Gaussian process $Y \sim \mathcal{N}(\mu(.), k(.,.))$, with trend μ and covariance function k:



here $\mu(x) = 0$.

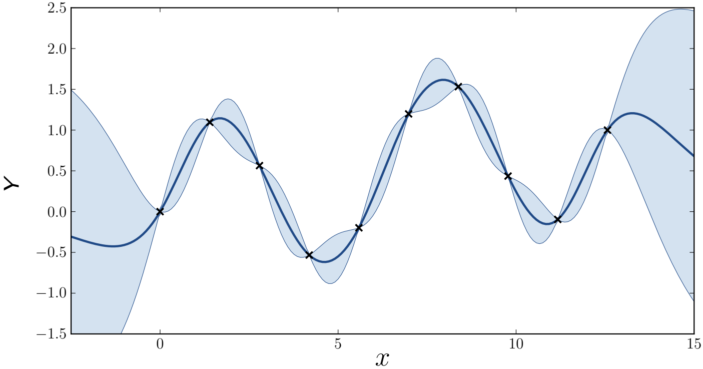
Gaussian Process Regression (3/5)

If we remove all the samples that do not interpolate the observations we obtain:



Gaussian Process Regression (4/5)

It can summarized by a mean function and 95% confidence intervals.



- Kriging Mean: blue thick line
- Kriging Standard Deviation: proportional to confidence band width.

You can play here: `https://durrande.shinyapps.io/gp_playground/` thanks Nicolas!

... here $x \in \mathbb{R}$, but it is easy to extend to $\mathbf{x} \in \mathbb{R}^d$...

Gaussian Process Regression (5/5): equations

Conditional (posterior) distribution of $Y(\mathbf{x}^\star)$ given $\mathbf{Y} = (Y(\mathbf{x}_1), ..., Y(\mathbf{x}_n))^\top$

By definition, $(Y(\mathbf{x}^\star), \mathbf{Y})$ is multivariate normal.
Hence the distribution of $Y(\mathbf{x}^\star)|\mathbf{Y} = \mathbf{y}$ is $\mathcal{N}(m(.), c(., .))$ with :

$$\left\{\begin{array}{lcl} m(\mathbf{x}^\star) & = & \mu(\mathbf{x}^\star) + \mathbf{h}(\mathbf{x}^\star)^\top \mathbb{K}^{-1} (\mathbf{y} - \mu(\mathbb{X})) \\ c(\mathbf{x}^\star, \mathbf{x}^{\star\prime}) & = & k(\mathbf{x}^\star, \mathbf{x}^{\star\prime}) - \mathbf{h}(\mathbf{x}^\star)^\top \mathbb{K}^{-1} \mathbf{h}(\mathbf{x}^{\star\prime}) \end{array}\right.$$

where key ingredients are:

- $\mathbb{K}$ the $n \times n$ covariance matrix between $\mathbf{Y}$ and $\mathbf{Y}$
- $\mathbf{h}(\mathbf{x}^\star)$ the $n \times 1$ covariance vector between $\mathbf{Y}$ and $Y(\mathbf{x}^\star)$
- both deduced from $k(., .)$ the covariance function of the (prior) Gaussian Process

Simple Kriging, Gaussian case

The Simple Kriging predictor mean and variance are

$$\left\{\begin{array}{lclcl} \mathrm{E}\left[\, Y(\mathbf{x}^\star)|\mathbf{Y}{=}\mathbf{y} \,\right] & = & m(\mathbf{x}^\star) & = & \mathbf{h}(\mathbf{x}^\star)^\top \mathbb{K}^{-1} \mathbf{y} \\ \mathrm{Var}\left[\, Y(\mathbf{x}^\star)|\mathbf{Y}{=}\mathbf{y} \,\right] & = & c(\mathbf{x}^\star, \mathbf{x}^\star) & = & \sigma(\mathbf{x}^\star)^2 - \mathbf{h}(\mathbf{x}^\star)^\top \mathbb{K}^{-1} \mathbf{h}(\mathbf{x}^\star) \end{array}\right.$$

with $\sigma(\mathbf{x}^\star)^2 = k(\mathbf{x}^\star, \mathbf{x}^\star)$, for a centered process, when $\mu(\mathbf{x}) = 0$ for all $\mathbf{x}$.

Simple Kriging: the statistical approach (1/2)

Another Approach: Best Linear Unbiased Prediction (±Geostat community)

Define the linear predictor

$$M(\mathbf{x}^\star) := \sum_{i=1}^{n} \alpha_i(\mathbf{x}^\star) Y(\mathbf{x}_i) = \boldsymbol{\alpha}(\mathbf{x}^\star)^\top \mathbf{Y}\,.$$

Now let us minimize on $\boldsymbol{\alpha}(\mathbf{x}^\star) = (\alpha_1(\mathbf{x}^\star), ..., \alpha_n(\mathbf{x}^\star))$ the loss

$$\begin{aligned}\Delta(\mathbf{x}^\star) \quad &:= \quad \mathrm{E}\left[(M(\mathbf{x}^\star) - Y(\mathbf{x}^\star))^2\right] \\ &= \quad \boldsymbol{\alpha}(\mathbf{x}^\star)^\top \mathbb{K} \boldsymbol{\alpha}(\mathbf{x}^\star) - 2\boldsymbol{\alpha}(\mathbf{x}^\star)^\top \mathbf{h}(\mathbf{x}^\star) + \mathrm{constant}\,.\end{aligned}$$

This leads to the vector of weights

$$\boldsymbol{\alpha}(\mathbf{x}^\star) = \mathbb{K}^{-1}\mathbf{h}(\mathbf{x}^\star)\,,$$

where $\mathbf{h}(\mathbf{x}^\star)$ is the covariance vector between $Y(\mathbf{x}^\star)$ and the vector $\mathbf{Y}$, and $\mathbb{K}$ is the covariance matrix of $\mathbf{Y}$. $Y(.)$ centered here.

Simple Kriging: the statistical approach (2/2)

Predictor and variance

From that follows the expression of $M(\mathbf{x}^\star)$ and $\Delta(\mathbf{x}^\star)$:

$$\begin{cases} M(\mathbf{x}^\star) & = \quad \mathbf{h}(\mathbf{x}^\star)^\top \mathbb{K}^{-1}\mathbf{Y} \\ \Delta(\mathbf{x}^\star) & = \quad \sigma(\mathbf{x}^\star)^2 - \mathbf{h}(\mathbf{x}^\star)^\top \mathbb{K}^{-1}\mathbf{h}(\mathbf{x}^\star) \end{cases}$$

One retrieves exactly the Simple Kriging mean and variance. ☺
Notice that $\Delta(\mathbf{x}^\star)$ does not depend on observed responses $\mathbf{Y}$.

Pros and Cons of both approaches

- GPR more intuitive for varying $\mathbf{x}^\star$
- GPR more suited to Bayesian analysis and interpretation, more visual
- Stat Approach not limited to Gaussian case
- Stat Approach easier to extend (other combination, criterions, penalization, etc.)

Because GPR/Kriging predicts a full, spatially varying, distribution,
it is of great use in decision making.

GPR/Kriging Problems

Here are few selected problems:

- Model selection:
  how to choose prior process, prior covariance function family, prior covariance function parameters ?
- Computation:
  how to compute the predictor when the matrices are huge?
- Adaptation:
  how to adapt to specific settings (monotony, uncertainty, extremes, high dimension, **multiple outputs**, **constraints**...)

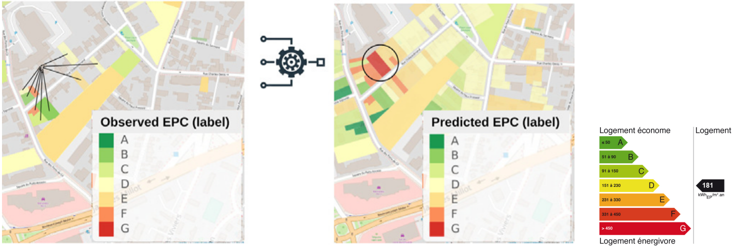
We focus here on the last problems: **multiple outputs**, **constraints**.

Main differences with usual Kriging models are highlighted with a symbol NEW .

Constrained Multi-Output Kriging

A problem at the origin of this study...

Predict energy performance without physical inspection

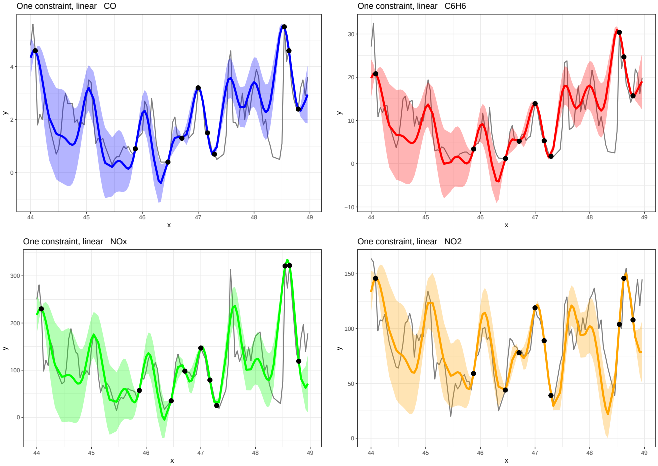


= Classification problem with knowledge on average percentages of each class.

Idea of predicting several membership degrees and with constraints.

Multi-Output Kriging: example

Example of multi-output, e.g. at $x \simeq 44$, an observation $\mathbf{Y}(x) = (Y_1(x), ..., Y_4(x))^\top$



Here time series, $x \in \mathbb{R}$ for visual illustration, but in general $\mathbf{x} \in \mathbb{R}^d$.

Multi-Output Kriging: a remark

Consider a vector-valued process

$$\mathbf{Y}(\mathbf{x}) = \left(Y_1(\mathbf{x}), \dots, Yp(\mathbf{x})\right)^\top \in \mathbb{R}^p.$$

The good

- Observations at $\mathbf{x}_1, \dots, \mathbf{x}_n$ can be arranged in a single *np* vector

$$\widetilde{\mathbf{Y}} := \left(Y_1(\mathbf{x}_1), \dots, Yp(\mathbf{x}_1), \dots, Y_1(\mathbf{x}_n), \dots, Yp(\mathbf{x}_n)\right)^\top \in \mathbb{R}^{np}$$

- We aim at predicting some $Y_{i^\star}(\mathbf{x}^\star) \in \mathbb{R}$
- Same as predicting one value from np observations ☺

The bad

- $\mathcal{O}\left(p^2\right)$ cross-covariances for each pair of points $(\mathbf{x}_i, \mathbf{x}_j)$, need to model this
- Many hyper-parameters for these cross-covariances, hard to tune
- Large complexity $\mathcal{O}\left(n^3 p^3\right)$ for the inversion of $\mathrm{Var}\left[\widetilde{\mathbf{Y}}\right]$
- Not suited to classification, as we will see...

Constrained Multi-Output Kriging: motivation

Why another study, why multiple outputs, why specific constraints?

- Multi-output:
  Studying multiple outputs is useful:
  - Observations of $p > 1$ variables, possibly dependent
  - Need for a model with with **not too many hyperparameters**, not $\mathcal{O}\left(p^2\right)$

- Constraints:
  Prescribing e.g. the average value of predictions is useful:
  - **external information** (known quantity of chemical loss, national statistic...)
  - **adverse modelling** (regulation, simulation under specific scenarios...)
  - need to homogenize results (over different regions, observed years, fairness constraints...)

- Classification in mind:
  Adapting to constrained classification:
  - Multi-output applied to membership degrees
  - Useful constraints: membership degrees sum to 1, prescribed percentages of each class.

Literature

Among a vast literature,

- Co-Kriging and multi-output Kriging: co-kriging, usually one main "primary" output, $\mathcal{O}\left(p^2\right)$ covariance models, cross co-variograms *Goovaerts*, 1998, *Ver Hoef and Cressie*, 1993, *Furrer and Genton*, 2011, *Genton and Kleiber*, 2015, *Chiles and Delfiner*, 2012, *Alvarez et al.*, 2012, *Wackernagel*, 2003, *Leroy et al.*, 2022...

- Indicator Kriging: with a latent process, post-treatments, many covariances models, *Journel*, 1983, *Meer*, 1996, *Goovaerts*, 2009, *Chiang et al.*, 2013, *Agarwal et al.*, 2021...

- Gaussian Process and classification: with latent GP, Bayesian inference and approximations, ordinal classes, *Williams and Barber*, 1998, *Rasmussen et al.*, 2006, *Dahl and Bonilla*, 2019, *Panos et al.*, 2021...

- Constraints: without Kriging, classification *Gordon*, 1996, *Bradley et al.*, 2000, *Höppner and Klawonn*, 2008, *Ganganath et al.*, 2014, fuzzy classification *Benatti et al.*, 2022, fairness constraints *Zafar et al.*, 2019...

Multi-Output Kriging

Framework

- Inputs: set of locations $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$.
- Outputs: multi-valued random field $\mathbf{Y}(\mathbf{x}) := \left(Y_1(\mathbf{x}), \ldots, Y_p(\mathbf{x})\right)^\top \in \mathbb{R}^p$.
- Observations: $\mathbf{Y}(\mathbf{x}_1), \ldots, \mathbf{Y}(\mathbf{x}_n)$, that is n vectors of size p.

Question

How to predict $\mathbf{Y}(.)$ at some unobserved locations $\mathbf{x}_1^\star, \ldots, \mathbf{x}_q^\star$?

Joint Kriging Model

$$\mathbf{M}(\mathbf{x}^\star) := \sum_{i=1}^{n} \alpha_i(\mathbf{x}^\star)\mathbf{Y}(\mathbf{x}_i)\,, \quad \alpha_i(\mathbf{x}^\star) \in \mathbb{R},\ i = 1 \ldots n \tag{1}$$

Simplifying assumption:

The weights are impacting all components the same way. NEW

Constrained Multi-Output Kriging

Recall the joint Kriging model,

$$\mathbf{M}(\mathbf{x}^\star) \quad := \quad \sum_{i=1}^{n} \alpha_i(\mathbf{x}^\star)\mathbf{Y}(\mathbf{x}_i) \quad = \quad \mathbb{Y}\boldsymbol{\alpha}(\mathbf{x}^\star) \tag{2}$$

where $\mathbb{Y} := [\mathbf{Y}(\mathbf{x}_1), \dots, \mathbf{Y}(\mathbf{x}_n)] \in \mathbb{R}^{p\times n}$ and $\boldsymbol{\alpha}(\mathbf{x}^\star) := (\alpha_1(\mathbf{x}^\star), \dots, \alpha_n(\mathbf{x}^\star))^\top \in \mathbb{R}^n$.

How to get optimal weights?

To get optimal weights $\mathbb{A} := \left[\boldsymbol{\alpha}(\mathbf{x}_1^\star), \dots, \boldsymbol{\alpha}(\mathbf{x}_q^\star)\right]$, they are optimized in order to:

- minimize some error:

$$\Delta(\mathbf{x}^\star) := \mathrm{E}\left[\|\mathbf{M}(\mathbf{x}^\star) - \mathbf{Y}(\mathbf{x}^\star)\|_{\mathbb{W}}^2\right] \ \in \mathbb{R}\,. \tag{3}$$

NEW

  where $\|\mathbf{v}\|_{\mathbb{W}}^2 := \mathbf{v}^\top \mathbb{W}\mathbf{v}$ and $\mathbb{W}$ a given real symmetrical positive-definite matrix.

- under various constraints:
  - **Constraint 1**: Sum of weights equal to 1, $\alpha_1(\mathbf{x}^\star) + \ldots + \alpha_n(\mathbf{x}^\star) = 1$
  - **Constraint 2**: Prescribed average $\mathbf{m}$ of predicted values $\mathbf{M}(\mathbf{x}_1^\star)$, ..., $\mathbf{M}(\mathbf{x}_q^\star)$ NEW

Optimal weights, no constraint

Without constraint one retrieves Simple Kriging equations, but $\mathbb{K}$, $\mathbf{h}(.)$ involve cross covariances of all components of $\mathbf{Y}(.)$.

Proposition (Simple Joint Kriging weights)

The optimal weights $\boldsymbol{\alpha}(\mathbf{x}^\star)$ minizing the loss $\Delta(\mathbf{x}^\star)$ are given by:

$$\boldsymbol{\alpha}(\mathbf{x}^\star) = \mathbb{K}^{-1}\mathbf{h}(\mathbf{x}^\star)\,, \tag{4}$$

or equivalently, using a matrix expression, under invertibility assumption,

$$\mathbb{A} = \mathbb{K}^{-1}\mathbb{H}\,, \tag{5}$$

where $\mathbb{K} := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W}\mathbb{Y}\right]$, $\mathbf{h}(\mathbf{x}^\star) := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W}\mathbf{Y}(\mathbf{x}^\star)\right]$, *and* $\mathbb{H} := \left[\mathbf{h}(\mathbf{x}_1^\star), \dots, \mathbf{h}(\mathbf{x}_q^\star)\right]$. NEW

If furthermore $\mathrm{E}\left[Y_j(x)\right] = 0$ *for all* $j = 1, \dots, p$, $x \in \mathcal{X}$, *then* $\mathbf{M}(\mathbf{x}^\star)$ *is unbiased.*

matrix sizes: $\boldsymbol{\alpha}(\mathbf{x}^\star) \in \mathbb{R}$, $\mathbb{K} \in \mathbb{R}^{n \times n}$, $\mathbf{h}(\mathbf{x}^\star) \in \mathbb{R}^n$, $\mathbb{H} \in \mathbb{R}^{n \times q}$.

Optimal weights, constraint 1

Considered constraint, similarly to ordinary Kriging

Constraint 1: Weights sum to one, $\boldsymbol{\alpha}^\top(\mathbf{x}^\star)\mathbf{1}_n = 1$, $\mathbf{x}^\star \in \mathcal{X}$.

Proposition (Ordinary Joint Kriging weights)

Under the Constraint 1, the optimal weights $\boldsymbol{\alpha}(\mathbf{x}^\star)$ minimizing the loss $\Delta(\mathbf{x}^\star)$ are:

$$\left\{\begin{array}{lcl} \boldsymbol{\alpha}(\mathbf{x}^\star) & = & \mathbb{K}^{-1}\left(\mathbf{h}(\mathbf{x}^\star) + \lambda(\mathbf{x}^\star)\mathbf{1}_n\right) \\ \lambda(\mathbf{x}^\star) & = & \frac{1}{\delta}\left(1 - \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbf{h}(\mathbf{x}^\star)\right) \end{array}\right. \tag{6}$$

Equivalently, using matrix expressions, one gets

$$\left\{\begin{array}{lcl} \mathbb{A} & = & \mathbb{K}^{-1}\left(\mathbb{H} + \mathbf{1}_n\boldsymbol{\lambda}^\top\right) \\ \boldsymbol{\lambda}^\top & = & \frac{1}{\delta}\left(\mathbf{1}_q{}^\top - \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbb{H}\right) \end{array}\right. \tag{7}$$

where $\mathbb{K} := \mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbb{Y}\right]$, $\mathbf{h}(\mathbf{x}^\star) := \mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbf{Y}(\mathbf{x}^\star)\right]$, *and with scalar* $\delta := \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbf{1}_n$.
For matrix expressions, $\boldsymbol{\lambda} := (\lambda(\mathbf{x}_1^\star), \dots, \lambda(\mathbf{x}_q^\star))^\top$, *and* $\mathbb{H} := \left[\mathbf{h}(\mathbf{x}_1^\star), \dots, \mathbf{h}(\mathbf{x}_q^\star)\right]$. NEW

If furthermore, for all $i = 1, \dots, p$, $x \in \mathcal{X}$, $\mathrm{E}\left[Y_i(x)\right] = \mu_i$, *then* $\mathbf{M}(\mathbf{x}^\star)$ *is unbiased.*

Optimal weights, constraints 1+2

Considered constraints

- **Constraint 1**: Weights sum to one, $\boldsymbol{\alpha}^\top(\mathbf{x}^\star)\mathbf{1}_n = 1$, $\mathbf{x}^\star \in \mathcal{X}$.
- **Constraint 2**: Prescribed average **m** of predicted values:

$$\mathrm{E}\left[\mathbf{M}(X^\star)|\mathbb{Y}\right] = \mathbf{m}, \text{ with } X^\star \text{ r.v. on } \{\mathbf{x}_1^\star, \ldots, \mathbf{x}_q^\star\}, \text{ distribution } \boldsymbol{\pi}.$$

Note: unlike usual kriging methods, weights **must** be calculated simultaneously.

Proposition (Joint Kriging weights under predicted values constraint)

The Joint Kriging weights minizing the loss $\Delta(\mathbf{x}^\star)$ under the constraints 1+2 are:

$$\mathbb{A} = \mathbb{K}^{-1}\left(\mathbb{H} + \mathbf{1}_n\boldsymbol{\lambda}^\top + \mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top\right) \tag{8}$$

NEW *[weights must be solved all at once]*

with Lagrange multipliers, provided that $\left(\frac{1}{\delta}\mathbf{u}\mathbf{u}^\top - \mathbb{Y}\mathbb{K}^{-1}\mathbb{Y}^\top\right)$ *is invertible,*

$$\begin{aligned}
\boldsymbol{\lambda}' &= \gamma^{-1}\left(\frac{1}{\delta}\mathbf{u}\mathbf{u}^\top - \mathbb{Y}\mathbb{K}^{-1}\mathbb{Y}^\top\right)^{-1}\left(\mathbb{Y}\mathbb{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \frac{1}{\delta}\mathbf{u}\left(1 - \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbb{H}\boldsymbol{\pi}\right) - \mathbf{m}\right)\\
\boldsymbol{\lambda} &= \delta^{-1}\left(\mathbf{1}_q - \mathbb{H}^\top\mathbb{K}^{-1}\mathbf{1}_n - \boldsymbol{\pi}\boldsymbol{\lambda}'^\top\mathbf{u}\right)
\end{aligned}$$

where $\mathbf{u} := \mathbb{Y}\mathbb{K}^{-1}\mathbf{1}_n$, $\gamma := \boldsymbol{\pi}^\top\boldsymbol{\pi} \in \mathbb{R}$ *and* $\delta := \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbf{1}_n \in \mathbb{R}$. $\boldsymbol{\pi} = \left(\mathrm{P}\left[X^\star = \mathbf{x}_i^\star\right]\right)_i$.

Kriging mean and variance

There is no difficulty to compute Kriging Mean and Variance from the weights:

Proposition (Joint Kriging Mean)

Let $\alpha(\mathbf{x}^\star)$ be any vector of weights, possibly satisfying supplementary constraints. The associated Joint Kriging Mean writes:

$$\mathbf{M}(\mathbf{x}^\star) \quad := \quad \mathbb{Y}\alpha(\mathbf{x}^\star) \tag{9}$$

where $\mathbb{Y} = [\mathbf{Y}(\mathbf{x}_1), ..., \mathbf{Y}(\mathbf{x}_1)]$ is the $p \times n$ matrix of observations

Proposition (Joint Kriging Variance)

Let $\alpha(\mathbf{x}^\star)$ be any vector of weights, possibly satisfying supplementary constraints. The associated Joint Kriging variance writes:

$$\Delta(\mathbf{x}^\star) = \alpha(\mathbf{x}^\star)^\top \mathbb{K}\alpha(\mathbf{x}^\star) - 2\alpha(\mathbf{x}^\star)^\top \mathbf{h}(\mathbf{x}^\star) + v(\mathbf{x}^\star)\,, \tag{10}$$

with $\mathbb{K} := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W}\mathbb{Y}\right]$, $\mathbf{h}(\mathbf{x}^\star) := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W}\mathbf{Y}(\mathbf{x}^\star)\right]$, $v(\mathbf{x}^\star) := \mathrm{E}\left[\mathbf{Y}(\mathbf{x}^\star)^\top \mathbb{W}\mathbf{Y}(\mathbf{x}^\star)\right]$. NEW

Here aggregated error, also variance sharing results for the error of each component.

Affine extension (1/2)

External information source

Idea for prescribed $\mathbf{m}$: hidden external information (expert, other stat, etc.).
Let $\mathbf{Z}$ be the random vector containing this source of information.

Affine predictor

NEW The affine predictor is:

$$\mathbf{M}^{+}(\mathbf{x}^{\star}) := \alpha_0(\mathbf{x}^{\star})\mathbf{Z} + \sum_{i=1}^{n} \alpha_i(\mathbf{x}^{\star})\mathbf{Y}(x_i)\,, \tag{11}$$

Given $\mathbf{Z} = \mathbf{m}$, a constant term is included in the sum, hence the name *affine prediction*.

Updated constraints

- **Constraint 1**: Weights sum to one, $\mathbf{1}_{n+1}{}^{\top}\boldsymbol{\alpha}^{+}(\mathbf{x}^{\star}) = 1$, $\mathbf{x}^{\star} \in \mathcal{X}$.
  with $\boldsymbol{\alpha}^{+} = (\alpha_0(\mathbf{x}^{\star}), \ldots, \alpha_n(\mathbf{x}^{\star}))^{\top}$.
- **Constraint 2**: Prescribed average predicted values:

$$\mathrm{E}\left[\mathbf{M}^{+}(X^{\star}) | \mathbf{Z} = \mathbf{m}, \mathbb{Y}\right] = \mathbf{m}, \text{ with } X^{\star} \text{ r.v. on } \{\mathbf{x}_1^{\star}, \ldots, \mathbf{x}_q^{\star}\}$$ NEW

Affine extension (2/2)

Updated optimal weights of the affine predictor can be derived easily:

Proposition (Affine version of predictors)

Assume that the following covariance vectors are given

$$\left\{\begin{array}{lll} \mathbf{P}^\top & := & \mathrm{E}\left[\mathbf{Z}^\top \mathbb{W}\mathbb{Y}\right] - \mathrm{E}\left[\mathbf{Z}^\top\right] \mathbb{W}\,\mathrm{E}\left[\mathbb{Y}\right] \\ \mathbf{Q}^\top & := & \mathrm{E}\left[\mathbf{Z}^\top \mathbb{W}\mathbb{Y}^\star\right] - \mathrm{E}\left[\mathbf{Z}^\top\right] \mathbb{W}\,\mathrm{E}\left[\mathbb{Y}^\star\right] \\ \sigma_Z^2 & := & \mathrm{E}\left[\mathbf{Z}^\top \mathbb{W}\mathbf{Z}\right] - \mathrm{E}\left[\mathbf{Z}^\top\right] \mathbb{W}\,\mathrm{E}\left[\mathbf{Z}\right] \end{array}\right. \tag{12}$$

Then optimal weights of previous cases can be updated by replacing $\mathbb{Y}$, $\mathbb{K}$, $\mathbb{H}$ *by*

$$\mathbb{Y}^+ = \begin{bmatrix} \mathbf{m} & \mathbb{Y} \end{bmatrix}, \quad \mathbb{K}^+ = \begin{bmatrix} \sigma_Z^2 & \mathbf{P}^\top \\ \mathbf{P} & \mathbb{K} \end{bmatrix}, \quad \mathbb{H}^+ = \begin{bmatrix} \mathbf{Q}^\top \\ \mathbb{H} \end{bmatrix}, \tag{13}$$

NEW

In practice one can set, e.g., $\sigma_Z^2 \ll \sigma^2$, and **P**, **Q** filled with zeros.

$\Longrightarrow$ Better tuning of predictor's behavior far from observations.

Covariances (1/2)

Recall that all optimal weights rely on cross-moments matrices like $\mathbb{K} := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbb{Y}\right]$. One can easily replace these objects by "centered" ones:

Remark (Covariance matrices)

Assume that the predictor is unbiased (e.g. process with constant mean under Constraint 1). Then covariance matrices can be replaced by the "centered" ones

$$\left\{ \begin{array}{lcl} \widetilde{\mathbb{K}} & = & \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbb{Y}\right] - \mathrm{E}\left[\mathbb{Y}^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbb{Y}\right] \\ \textit{etc}. & & \end{array} \right. \tag{14}$$

everywhere in previous Propositions, without changing the optimal weights $\alpha(\mathbf{x}^\star)$*.*

$\Longrightarrow$ Only need, a generic covariance function

$$k(\mathbf{x}, \mathbf{x}') := \mathrm{E}\left[\mathbf{Y}(\mathbf{x})^\top \mathbb{W} \mathbf{Y}(\mathbf{x}')\right] - \mathrm{E}\left[\mathbf{Y}(\mathbf{x})^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbf{Y}(\mathbf{x}')\right]$$

Covariances (2/2)

As $\Delta(\mathbf{x}^\star) \in \mathbb{R}$, the covariances rely on an implicit sum of components of $\mathbf{Y}(.)$ NEW :

$$k(\mathbf{x},\mathbf{x}') := \mathrm{E}\left[\mathbf{Y}(\mathbf{x})^\top \mathbb{W} \mathbf{Y}(\mathbf{x}')\right] - \mathrm{E}\left[\mathbf{Y}(\mathbf{x})^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbf{Y}(\mathbf{x}')\right]$$

Remark (Known cross-covariances)

If weights $\mathbb{W}$ and cross-covariances between $\mathbf{Y}(\mathbf{x})$ and $\mathbf{Y}(\mathbf{x}')$ are known. Then $k(\mathbf{x},\mathbf{x}')$ can be computed.

Remark (Unknown cross-covariances)

Otherwise, one can model directly $k(\mathbf{x},\mathbf{x}')$ with suitable hyperparameters: Assume that $\mathbb{W}$ is such that the covariances depend only on some distance between $\mathbf{x}$ and $\mathbf{x}'$:

$$k(\mathbf{x},\mathbf{x}') = \sigma^2 r\left(\|\mathbf{x}-\mathbf{x}'\|_{\boldsymbol{\theta}}\right)\,, \tag{15}$$

where $r(.)$ is a correlation function and $\|\mathbf{x}-\mathbf{x}'\|^2_{\boldsymbol{\theta}} = \sum_{i=1}^d \left(\frac{x_i - x_i'}{\theta_i}\right)^2$ is a rescaled Euclidean norm. Then all components of the covariances matrices $\widetilde{K}$, etc. can be derived from this covariance function k

- Does not depend on $\mathbb{W}$ any more: with this assumption, no need to estimate $\mathbb{W}$.
- This simplifies a lot the hyperparameters estimation.

Constrained Classification

Application to constraint classification (1/3)

Assumption on observations

- Label binarization. Each class label $\ell \in \{1, \dots, p\}$ can be converted into a vector of indicator functions (even for non ordinal classes):

$$\mathbf{Y} := \left(\mathbb{1}_{\{j=\ell\}}\right)_{j=1,\dots,p} .$$

- Observation of membership degrees. More generally, each observation $\mathbf{Y}(x_i)$ consists in a distribution of possible labels, where degrees sum to one

$$\text{“Constraint 3”:} \quad \mathbf{1}_p{}^\top \mathbf{Y}(x_i) = 1\,, \quad i = 1, \dots, n\,.$$

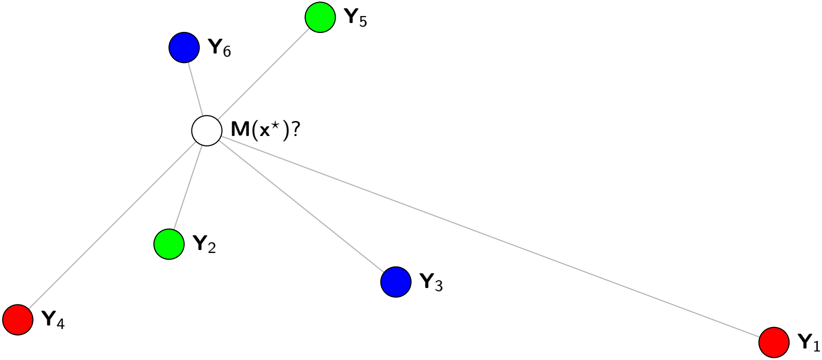
Non ordinal example of observations, $p = 3$ classes: $\{red, green, blue\}$

🔴 $\mathbf{Y} = \begin{bmatrix} \mathbf{1} \\ 0 \\ 0 \end{bmatrix}$, 🟢 $\mathbf{Y} = \begin{bmatrix} 0 \\ \mathbf{1} \\ 0 \end{bmatrix}$, 🔵 $\mathbf{Y} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{1} \end{bmatrix}$, and even 🔴🔵 $\mathbf{Y} = \begin{bmatrix} 1/2 \\ 0 \\ 1/2 \end{bmatrix}$

only requirement: components sum to one.

Application to constraint classification (2/3)

Illustration of the fuzzy classification



Predicted membership degrees at point $\mathbf{x}^\star$:

$$\mathbf{M}(\mathbf{x}^\star) = \underbrace{\alpha_1(\mathbf{x}^\star)}_{\in\mathbb{R}}\mathbf{Y}_1 + \ldots + \underbrace{\alpha_6(\mathbf{x}^\star)}_{\in\mathbb{R}}\mathbf{Y}_6 = 2\% \begin{bmatrix} \mathbf{1} \\ 0 \\ 0 \end{bmatrix} + \ldots + 40\% \begin{bmatrix} 0 \\ 0 \\ \mathbf{1} \end{bmatrix} = \begin{bmatrix} 10\% \\ 35\% \\ 55\% \end{bmatrix}$$

... weights $\alpha_i(\mathbf{x}^\star)$ are obtained by Joint Kriging formulas, under chosen constraints.

Application to constraint classification (3/3)

Constrained Classification: apply Joint Kriging model to membership degrees NEW :

Remark (Constraints impact)

Recall all constraints

- *Constraint 1: weights sum to one,*
- *Constraint 2: prescribed average of predictions,*
- *Constraint 3: observations are membership degrees,* $\mathbf{1}_p{}^\top \mathbb{Y} = \mathbf{1}_q{}^\top$.

Then with constrained Joint Kriging model:

- *Predicted membership degrees are summing to one:*

$$\textit{Constraints } 1{+}3 \implies \mathbf{1}_p{}^\top \mathbf{M}(\mathbf{x}^\star) = 1\,, \quad \mathbf{x}^\star \in \mathcal{X}$$

- *Average class percentages over prediction points can be chosen:*

$$\textit{Constraints } 2{+}3 \implies \mathrm{E}\left[\mathbf{M}(X^\star)|\mathbb{Y}\right] = \mathbf{m}\,, \quad \textit{with } \mathbf{1}_p{}^\top \mathbf{m} = 1$$

$\mathbf{m}$ *is the prescribed average of each class, and* $X^\star$ *a rv over all prediction points.*

Numerical Illustrations

available notebooks

☺

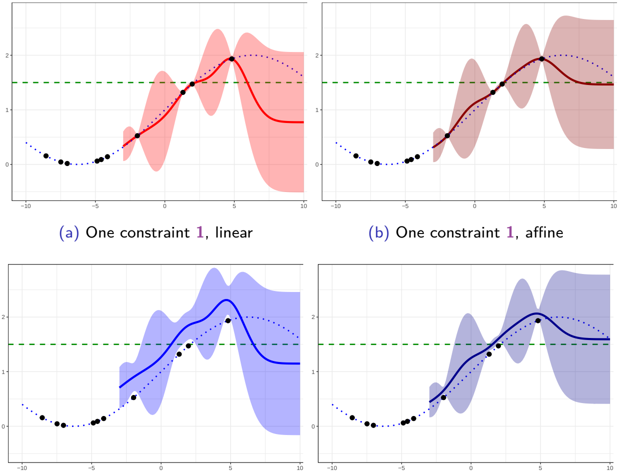
Available notebooks

All illustrations are generated with notebooks that are available at `https://gitlab.emse.fr/marc.grossouvre/jointkrigingsupplementary/`

In modifiable executable format `.Rmd` and in executed directly readable `.html` format.

We did our best to make results fully reproducible, and figures settings easy to retrieve.

A simple sinus function, 1D input, 1D output



(a) One constraint **1**, linear

(b) One constraint **1**, affine

(c) Two constraints **1**+**2**, linear

(d) Two constraints **1**+**2**, affine

Figure: Joint Kriging Prediction. prescribed value $\mathbf{m} = 1.5$ (horizontal dashed line). Observations are black dots, the thin dotted blue line is the underlying function.

Time-series Air Quality data, 1D input, 4D output, Constraint 1

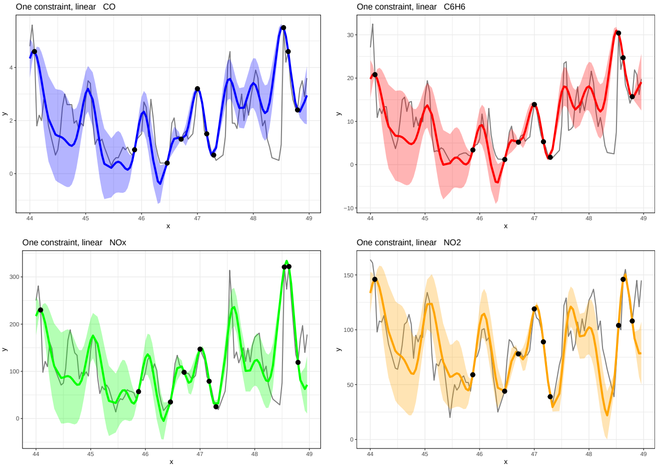


Figure: Joint Kriging interpolation with Constraint 1. data points (black dots). Top: CO, C6H6, bottom: NOx, NO2. Predictions thick solid lines, true values are in thin black solid lines.

Time-series Air Quality data, 1D input, 4D output, chosen covariance

Despite $p = 4$ outputs, one needs a single covariance structure (which involves linear combinations of components).
Multiply a periodic kernel with period of one day, and Matérn 3/2 kernel, parameter θ:

$$k(x,x') = \sigma^2 \exp\left(-\sin^2(\pi|x-x'|)\right)\left(1+\frac{|x-x'|}{\theta}\right)\exp\left(-\frac{|x-x'|}{\theta}\right). \qquad (16)$$

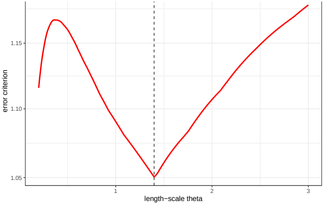


Figure: Optimization of the *single* correlation hyperparameter θ for the four selected pollutants, data extracted from Air quality data set.

Caution: depends quite heavily on the chosen observation locations, sometimes the error function is monotonic! easier to control with very few hyperparameters!

Time-series Air Quality data, 1D input, 4D output, Constraint 1+2 adverse

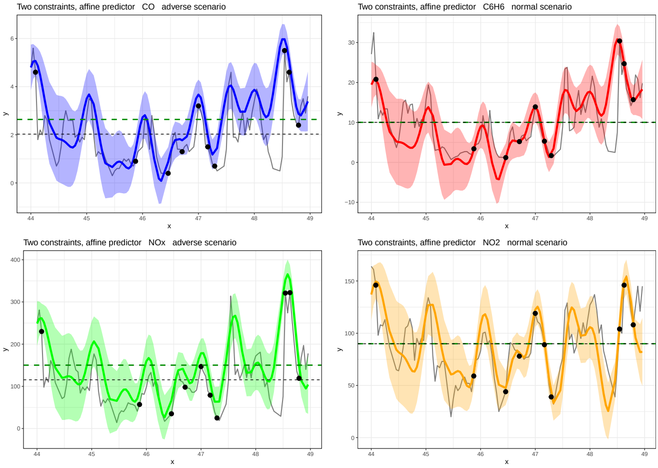


Figure: Adverse scenarios: constraints 1+2 and affine predictor. Left panels: adverse scenarios, average 130% of the true average, right panels: regular scenarios.

Quake classification problem, input 3D, output "2D"

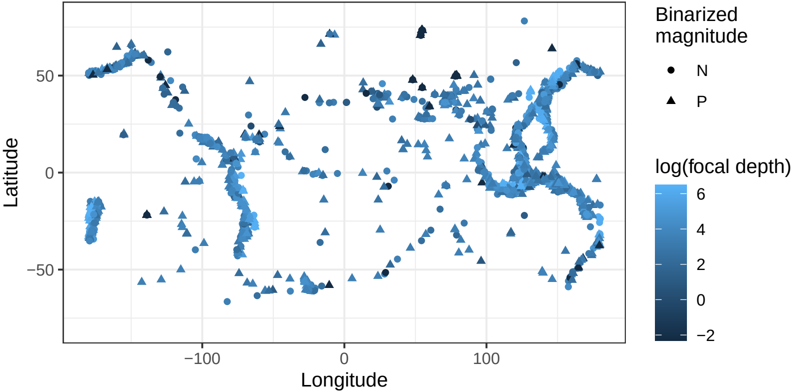


Figure: Earthquakes observations. An earthquake is a point with coordinates latitude, longitude and focal depth (given by the color). Triangles: earthquakes which magnitude is above average. Circles: below average.

data available at `www.openml.org/search?type=data&id=772`.

all notebooks: `https://gitlab.emse.fr/marc.grossouvre/jointkrigingsupplementary/` ‘

Quake classification problem - covariance

Chosen covariance function

A single covariance function (only one tested!):

$$k(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(-2\frac{\sin^2((x_1 - x_1')/2)}{\theta_1^2} - 2\frac{\sin^2((x_2 - x_2')/2)}{\theta_2^2}\right)\exp\left(-2\frac{(x_3 - x_3')^2}{\theta_3^2}\right)$$

periodicity of longitude, latitude, not focal depth. Small nugget (rounded magnitudes).

Parameter estimation

The hyperparameters estimation has been treated separately on other train/test splits to avoid overfitting the data (using 10 fold cross-validation).

Resulting values for $\boldsymbol{\theta}$ are 2.3 for latitude, 0.9 for longitude and 196.8 for focal depth.

Quake classification problem - prediction constraints 1+2

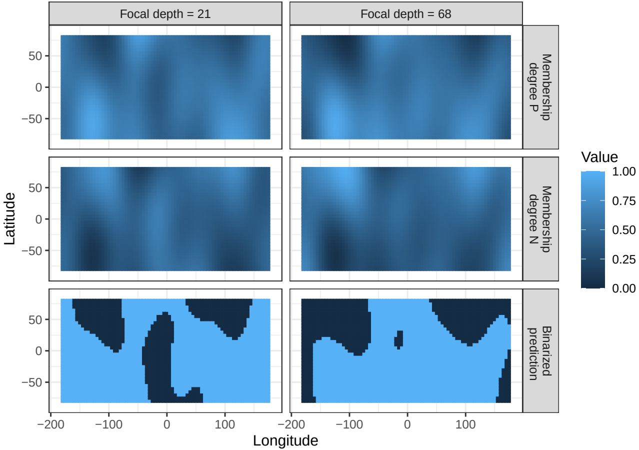


Figure: Joint Kriging with two constraints 1+2. Top: membership degree of “P: magnitude is above average”, bottom: membership degree of “N: magnitude is below average”, binarized prediction (1 if membership degree of P is greater than 0.5). Left: 21km focal depth (=Q1). Right: 68km (=Q3).

Quake classification problem - benchmark

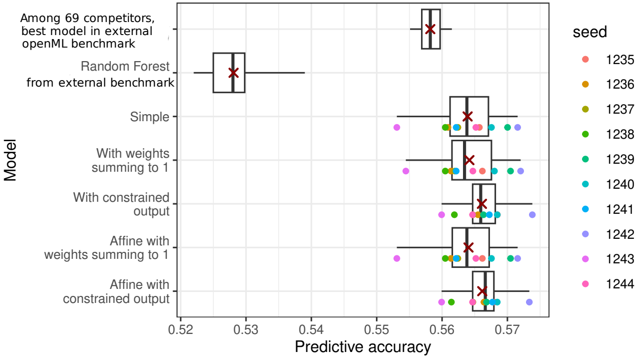


Figure: Performances for 10 runs. Top 2: best OpenML model (kernel logistic regression) and OpenML Random Forest. Bottom 5: Joint Kriging models. Dark red cross: average predictive accuracy, **the higher the better**.

OpenML benchmark: 69 models, `www.openml.org/search?type=task&id=4516`.

Quake classification problem - adverse scenario

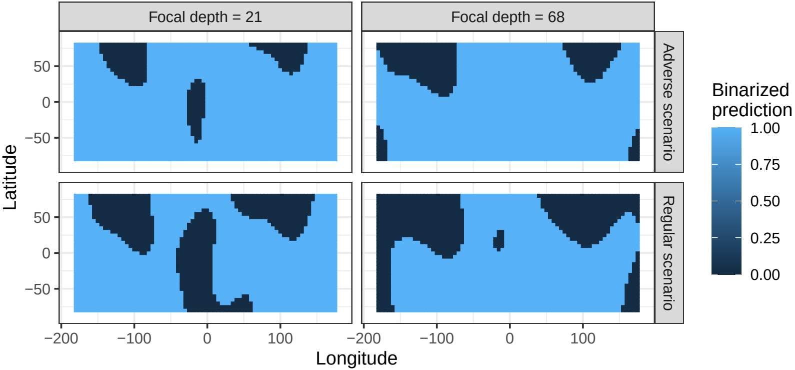


Figure: Adverse scenario, constraints 1+2. Top: adverse scenario, first class output average constrained to be 65%. Bottom: regular scenario, output average constrained to 55.5%. Left: 21km focal depth. Right: 68km focal depth.

Quake classification problem - four classes

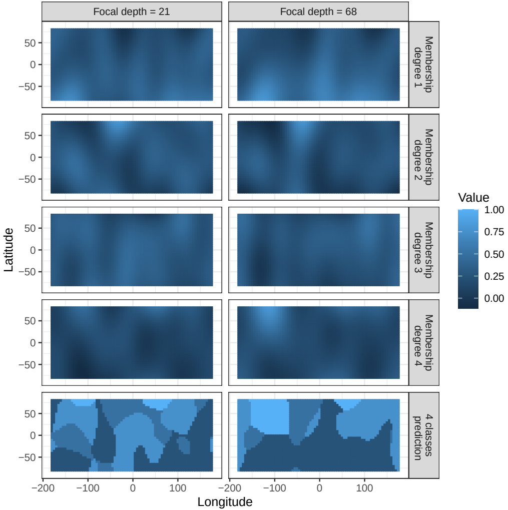


Figure: Affine Joint Kriging with constraints 1+2, magnitude thresholds 5.85, 5.95, 6.15. bottom: class of greatest membership degree. Remark longitude circular coherence.

Illustration on several datasets

Small benchmark on 9 datasets. Average ranks among 4 methods.

Indicator	Joint Kriging	kknn	parRF	multinom
Accuracy	**2.22**	2.72	2.28	2.78
Balanced Accuracy	2.22	2.78	**2.00**	3.00
Macro Average Precision	**2.22**	2.78	2.33	2.67
Micro Average Precision	**2.22**	2.72	2.39	2.67
Macro Average Recall	2.22	2.78	**2.00**	3.00

Table 4: Average rank of each model over the 9 datasets. The lower the better (in bold font). Green highlights ranks 1 or 2, orange ranks 3 or 4.

- Results seem competitive, even if more extensive benchmarks would be required.
- Not extensive enough for testing

Conclusion

what is done

- Multi-output Kriging model, not necessarily Gaussian.
- Specific simplification: weights apply jointly to all outputs.
- Specific constraints, especially on predicted values.
- Specific affine model.

Pros

- Simple: computable in closed-form, drastically reduces hyperparameters number.
- Useful: interpretable, can interpolate data, uncertainty measurements, specific covariances (e.g. periodicity). Constraints allows for external information, expert judgments, adverse modelling, or homogenization needs such as fairness constraints. For fuzzy classification, prescribed class percentages.
- Competitive: competes with state-of-the-art algorithms on an open benchmark.

Cons and perspectives

- Simplified: possible limitations for different regularities of outputs. Introduce more complex covariances, model with higher complexity...
- Needs hyperparameters optimization: specific estimation procedures...
- Non-convex: possible membership degrees outside $[0, 1]$. Convex constraints...
- Broken continuous interpolation property with Constraint 2. Modify predictor...

Thank you for your attention !

- Details and proofs in the preprint
  `https://hal.science/hal-04208454`.
- Available code & notebooks
  `https://gitlab.emse.fr/marc.grossouvre/jointkrigingsupplementary/`
- Do not hesitate to send comments or references!
  `drulliere@emse.fr`

Questions ?

References I

[1] Agarwal, G., Sun, Y., and Wang, H. J. (2021). Copula-based multiple indicator kriging for non-gaussian random fields. *Spatial Statistics*, 44:100524.

[2] Alvarez, M. A., Rosasco, L., Lawrence, N. D., and others (2012). Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266.

[3] Benatti, K. A., Pedroso, L. G., and Ribeiro, A. A. (2022). Theoretical analysis of classic and capacity constrained fuzzy clustering. *Information Sciences*, 616:127–140.

[4] Bradley, P. S., Bennett, K. P., and Demiriz, A. (2000). Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0.

[5] Chiang, J.-L., Liou, J.-J., Wei, C., and Cheng, K.-S. (2013). A feature-space indicator kriging approach for remote sensing image classification. *IEEE transactions on geoscience and remote sensing*, 52(7):4046–4055.

[6] Chiles, J.-P. and Delfiner, P. (2012). *Geostatistics: modeling spatial uncertainty*, volume 713. John Wiley & Sons, New York.

[7] Dahl, A. and Bonilla, E. V. (2019). Grouped gaussian processes for solar power prediction. *Machine Learning*, 108(8-9):1287–1306.

[8] Furrer, R. and Genton, M. G. (2011). Aggregation-cokriging for highly multivariate spatial data. *Biometrika*, 98(3):615–631.

References II

[9] Ganganath, N., Cheng, C.-T., and Tse, C. K. (2014). Data clustering with cluster size constraints using a modified k-means algorithm. In *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 158–161.

[10] Genton, M. G. and Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science*, 30(2).

[11] Goovaerts, P. (1998). Ordinary cokriging revisited. *Mathematical Geology*, 30:21–42.

[12] Goovaerts, P. (2009). Auto-ik: A 2d indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences. *Computers & Geosciences*, 35(6):1255–1270.

[13] Gordon, A. (1996). A survey of constrained classification. *Computational Statistics & Data Analysis*, 21(1):17–29.

[14] Höppner, F. and Klawonn, F. (2008). Clustering with size constraints. In Jain, L. C., Sato-Ilic, M., Virvou, M., Tsihrintzis, G. A., Balas, V. E., and Abeynayake, C., editors, *Computational Intelligence Paradigms: Innovative Applications*, pages 167–180. Springer Berlin Heidelberg, Berlin, Heidelberg.

[15] Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology*, 15:445–468.

References III

[16] Leroy, A., Latouche, P., Guedj, B., and Gey, S. (2022). MAGMA: inference and prediction using multi-task gaussian processes with common mean. *Machine Learning*, 111(5):1821–1849.

[17] Meer, F. V. D. (1996). Classification of remotely-sensed imagery using an indicator kriging approach: application to the problem of calcite-dolomite mineral mapping. *International Journal of Remote Sensing*, 17(6):1233–1249.

[18] Panos, A., Dellaportas, P., and Titsias, M. K. (2021). Large scale multi-label learning using gaussian processes. *Machine Learning*, 110:965–987.

[19] Rasmussen, C. E., Williams, C. K., et al. (2006). *Gaussian processes for machine learning*, volume 1. Springer.

[20] Ver Hoef, J. M. and Cressie, N. (1993). Multivariable spatial prediction. *Mathematical Geology*, 25:219–240.

[21] Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, Berlin.

[22] Williams, C. K. and Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on pattern analysis and machine intelligence*, 20(12):1342–1351.

[23] Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778.

Notations I

locations

$\mathcal{X}$ set of locations (inputs/design points).
n, q number of observed locations, of prediction locations.
$\mathbf{x}$ any location. $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are all observed locations.
$\mathbf{x}^\star$ any prediction location. $\mathbf{x}_1^\star, \ldots, \mathbf{x}_q^\star$ are all prediction locations.
$X^\star$ a random variable over prediction locations.
$\boldsymbol{\pi} = (\pi_{\mathbf{x}_1^\star}, \ldots, \pi_{\mathbf{x}_q^\star})$ the $q \times 1$ distribution of $X^\star$ over prediction locations.
$\gamma = \boldsymbol{\pi}^\top \boldsymbol{\pi}$ an intermediate real value used in calculations.

outputs

p number of outputs (i.e. number of outputs).
$\mathbf{Y}(\mathbf{x})$ the $p \times 1$ vector of outputs at location $\mathbf{x}$.
$\boldsymbol{\mu} = \mathrm{E}\left[\mathbf{Y}(\mathbf{x})\right]$ the $p \times 1$ mean of $\mathbf{Y}(\mathbf{x})$, when constant over $\mathbf{x}$.
$\mathbb{Y} = [\mathbf{Y}(\mathbf{x}_1), \ldots, \mathbf{Y}(\mathbf{x}_n)]$ all the $p \times n$ values of observed outputs.
$\mathbb{Y}^\star = \left[\mathbf{Y}(\mathbf{x}_1^\star), \ldots, \mathbf{Y}(\mathbf{x}_q^\star)\right]$ all $p \times q$ unknown outputs at prediction locations.

prediction

$\mathbf{M}(\mathbf{x}^\star)$ a $p \times 1$ predictor of $\mathbf{Y}(\mathbf{x}^\star)$
$\mathbb{M} = \left[\mathbf{M}(\mathbf{x}_1^\star), \ldots, \mathbf{M}(\mathbf{x}_q^\star)\right]$ the $p \times q$ matrix of all predictions.
$\boldsymbol{\alpha}(\mathbf{x}^\star)$ the $n \times 1$ linear weights for the prediction in $\mathbf{x}^\star$.
$\mathbb{A} = \left[(\boldsymbol{\alpha}(\mathbf{x}_1^\star), \ldots, \boldsymbol{\alpha}(\mathbf{x}_q^\star)\right]$ the $n \times q$ matrix of weights for all predictions.
$\mathbf{m}$ a given constant $p \times 1$ vector of prescribed mean predicted values.

Notations II

$\Delta(\mathbf{x}^\star)$ loss to be minimized for finding $\mathbf{M}(\mathbf{x}^\star)$.
$\boldsymbol{\lambda}$ a $q \times 1$ vector of Lagrange multipliers (relative to sum of weights)
$\boldsymbol{\lambda}'$ a $p \times 1$ vector of Lagrange multipliers (relative to predicted values)
$\mathbf{u} = \mathbb{Y}\mathbb{K}^{-1}\mathbf{1}_n$ an intermediate $p \times 1$ vector in calculations.
$\mathbf{Z}$ an additional $p \times 1$ factor for affine predictions.

covariances

$k(.,.)$ a covariance function.
$\mathbb{W}$ a given symmetric positive definite matrix for computing norms.
$\mathbf{h}(\mathbf{x}^\star) = \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W}\mathbf{Y}(\mathbf{x}^\star)\right]$ a $n \times 1$ covariance vector.
$\mathbb{H} = (\mathbf{h}(\mathbf{x}_1^\star, \ldots, \mathbf{h}(\mathbf{x}_q^\star)))$ a $n \times q$ covariance matrix.
$\mathbb{K} = \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W}\mathbb{Y}\right]$ a $n \times n$ covariance matrix.
$\widetilde{\mathbb{K}}, \widetilde{\mathbf{h}}(\mathbf{x}^\star), \widetilde{\mathbb{H}}$ other covariances using centred expressions.
$\delta = \mathbf{1}_n{}^\top \mathbb{K}^{-1}\mathbf{1}_n$ an intermediate real value in calculations.
$\mathbf{P}$ additional $n \times 1$ covariance vector between $\mathbf{Z}$ and $\mathbf{Y}(\mathbf{x}_i)$
$\mathbf{Q}$ additional $q \times 1$ covariance vector between $\mathbf{Z}$ and $\mathbf{Y}(\mathbf{x}_j^\star)$

miscellaneous

$\mathbf{v}$ a generic vector for defining norm or checking psd characteristic.
$\mathbf{1}_n, \mathbf{1}_p, \mathbf{1}_q$ a vector of ones of size n, p, q respectively.
$\mathbf{0}_n, \mathbf{0}_p, \mathbf{0}_q$ a vector of zeros of size n, p, q respectively.