A Joint Kriging Model with Application to Constrained Classification

Didier Rullière, Marc Grossouvre

A Joint Kriging Model with Application to Constrained Classification

Didier Rullière*[1] and Marc Grossouvre†[1,2]

[1]Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, F - 42023 Saint-Etienne France.
[2]URBS.

September 27, 2023

Abstract

Interpolating or predicting data is of utmost importance in machine learning, and Gaussian Process Regression is one of the numerous techniques that is often used in practice. This paper considers the case of multi-input and multi-output data. It proposes a simple *Joint* Kriging model where common combination weights are applied to all output variables at the same time. This dramatically reduces the number of hyperparameters to be optimized, while keeping nice interpolating properties. An original constraint on predicted values is also introduced, useful for considering external information or adverse scenarios. Finally, it is shown that applied to membership degrees, the model is especially helpful for fuzzy classification problems. In particular, the model allows for prescribed average percentages of each class in predictions. Numerical illustrations are provided for both simulated and real data, and show the importance of the constraint on predicted values. The method also competes with state-of-the-art techniques on an open real world data set.

Keywords—Multi-output Kriging, Cokriging, Constrained classification, Spatial Prediction, multi-task Gaussian Process regression.

1 Introduction

Interpolating data is widely used in many fields of computer experiments, it is especially useful to predict values of one or several variables of interest, in the context of time-consuming or costly computer experiments. One

*didier.rulliere@emse.fr
†marcgrossouvre@urbs.fr

considers here a Kriging interpolation problem on several output variables, with specific constraints on predicted values, and with possible applications to classification. Let us detail the need to deal with such problem.

Kriging on several outputs. Kriging or Gaussian Process Regression is a method of interpolation, especially suited when there are only few observations that have to be interpolated. It is widely used in many fields of Machine Learning, originally for geostatistical studies and spatial interpolation, but also for computer experiments in many domains (finance, industry, environment, etc.). The most basic Kriging theory aims a predicting a single real-valued quantity of interest, the *output* (for instance gold concentration in the ground), depending on some explanatory variables that are referred to as *input values* or *locations* (for instance latitude, longitude and depth). From a statistical point of view, the Kriging method is based on a best linear unbiased combination of observed outputs, with an assumption that observations are random variables whose correlation depends on locations. From a Gaussian random field point of view, in a Gaussian setting, the interpolation is the mean of a conditional Gaussian random field, with confidence bands derived from the variance of the conditional random field (see e.g. Williams and Rasmussen, 2006, for an in-depth review).

The method has several advantages. First, it is interpretable: the prediction is a weighted average of observations, with quite logical behavior of weights. Second, the method fully interpolates the data, that is, predicts exactly an observed output if one uses the same input values. And third, it not only gives a prediction, but also confidence intervals for this prediction.

Among limitations of the method, and proposed extensions in the literature, one can cite the difficulty to handle a large number of observations (see e.g. Cressie and Johannesson, 2008; Banerjee et al., 2013; Rullière et al., 2018, and references therein), the difficulty to specify the covariance model and to estimate its hyperparameters (see e.g. Bachoc, 2013), the difficulty to treat multi-valued output (see Furrer and Genton (2011)).

In the present paper, we mainly consider this latter multi-valued output problem, which is clearly of practical interest. Co-Kriging techniques are built to treat several outputs, but there is usually one main output and others are used to improve the prediction of the main considered output. Furthermore, all cross-covariances between outputs at different locations have to be modelled, which creates $O(p^2)$ covariance models, where p is the number of outputs (see e.g. Goovaerts, 1998; Ver Hoef and Cressie, 1993; Furrer and Genton, 2011).

While highly parametrized models are useful in many situations, the prediction quality rely on the proper specification of the model and on the estimation of its parameters. A fine model with wrong parameters can some-

times be less efficient than a simpler model with more control on few parameters (see Rasmussen and Ghahramani, 2000). In this paper, one considers a model where Kriging is applied to multi-valued outputs in $\mathbb{R}^p$, but with a specific simplification leading to a single covariance function to tune instead of $O(p^2)$. This way, the prediction is feasible with a limited number of hyperparameters, interpretable and easy to explain.

One originality of the paper is that this kind of multi-valued interpolation is also applied to membership degrees in a classification setting. This allows to benefit from Kriging advantages (interpretation/explainability, interpolation, confidence intervals). Moreover, the proposed simplification of the model is especially useful since it keeps the property of membership degrees summing to one. At last, another novelty is to consider a specific constraint on predicted values. As detailed below, it will allow for proportion constraints in a classification setting.

Constraints on predicted values. There is a well known joke on actuaries: *How much is two plus two? An actuary will ask "What do you want it to equal?"*. At first glance, it seems dishonest to require constraints on predicted values, especially if these constraints are very precise. But such constraints can be useful in some situations, when having external information, for adverse modelling or for homogenization needs, as illustrated below. The constraint we consider in this paper focuses on the average of predicted values.

It can be very helpful to prescribe a specific value for the average of predicted values. Let us instantiate some examples: due to an industrial accident, one wishes to measure the pollution in the soil for different chemical products. Measures are done at some spatial places, but the number of measures is limited. One would like to infer the quantity of all chemical product everywhere in the soil. Knowing stockpiles of products before the accident, the total quantity of lost chemicals may be known for every chemical products. While Gaussian Process Regression is especially suited to predict one product dosage in the soil, it has difficulty to handle jointly a lot of products, as it needs to model many cross-covariances. Furthermore it cannot handle at all constraints like prescribing the average of predicted values. Another example is the case where one needs to build a prediction under an adverse scenario: even if the total quantity of lost chemical is unknown, it can be useful to get an idea of the distribution of pollutants in an adverse case of massive loss. Finally, in other investigations, there may be external knowledge to consider: for example a regional study might want to be in line with some given national statistics, if there is no reason that the regional statistics differs in average. One can observe data due to an exceptional situation (e.g. COVID), and one may want to use it knowing that the situation has returned to normal. Or one might want predictions

over different years or over different regions to coincide at least on average: for instance one may want that some disease incidence prediction does not differ, on average, over different medical centers. Another situation is the following one: imagine that one knows, under arbitrage free setting, that some predicted stock returns must be zero on average, or imagine that the regulator wants to force a prediction under specific shock scenarios. Fairness constraints can also be introduced to limit unfairness in algorithmic decision making (see e.g. Zafar et al., 2019).

In a nutshell, prescribing the average value of predictions is useful in many contexts, be it external information (known quantity of chemical, national statistic...), adverse modelling (regulation, simulation under specific scenarios...), or need to homogenize results (over different regions, observed years, fairness constraints...).

Constrained Kriging and classification. Once Kriging is adapted to multi-output prediction, applying jointly Kriging on membership degrees has several advantages for classification.

One advantage is that the interpolation property can be preserved, which is not necessarily the case for other classification or clustering techniques like kNN: even very near from one observation, if the chosen number of neighbors is greater than one, kNN can propose another class than the observed one. When applying Kriging, we will see that some constraints on predicted values, or alternatively the use of a nugget effect, may break this interpolation property. But Kriging still offers the choice between prescribing an interpolation constraint, or prescribing more specific constraint on predicted values. Another advantage of using a Joint Kriging model on membership degrees is to get an estimation of the uncertainty relying on the prediction. For instance, at a specific location one may predict 10 percent of chance that the class is one, but one can also give a confidence interval for this quantity.

Applied to classification, a constraint on average predicted membership degrees is also useful: imagine that one interpolates membership degrees of different classes, depending on some explanatory variables. One may want to predict the membership degrees for many possible values of explanatory variables. But due to some external knowledge, to adverse scenarios modelling or to homogenization needs, one may want to prescribe the proportion in each class over all predicted values. Predicted class (or more precisely predicted membership degree) may differ depending on explanatory variables, but one may want for sure that in average, the proportion in the different considered classes is given.

Some works can be found in the literature about clustering or classification under constraints: a survey on constrained classification can be found in Gordon (1996), see references therein. Some researches are treating size constraints for clustering Bradley et al. (2000); Höppner and Klawonn (2008);

Ganganath et al. (2014), some are treating the problem of fuzzy clustering with weights (membership degrees), as in the present paper, see for example Benatti et al. (2022). Fairness constraints are also considered in Zafar et al. (2019).

Regarding Kriging and classification, some works on classification using Gaussian settings can be found in a dedicated Chapter 3 in the book Rasmussen et al. (2006). In particular, for binary classification, membership probabilities can be approximated by a sigmoid-like transformations of some latent Gaussian Process. The approach can be generalized to multiclass problems, and bayesian inference can be conducted using analytic approximations of integrals, or solutions based on Monte Carlo sampling (see Williams and Barber, 1998; Rasmussen et al., 2006, and references therein). Other recent approaches involving Multi-task Gaussian process, using several latent Gaussian processes and bayesian inference with approximations or sampling, can be found in Dahl and Bonilla (2019); Panos et al. (2021).

Among works closer to what is proposed in the present paper, Indicator Kriging (IK) aims at determining the cdf of an underlying random field at an unknown location, as a weighted average of indicators. It uses as well linear combinations of transformed observations, but relies on a direct link between indicators and underlying random field, using thresholds. Hence, it seems not directly suited to classification for non ordinal data, without any hierarchy between classes. It also requires the observation of the latent process that generates the indicators (see e.g. Journel, 1983; Meer, 1996; Goovaerts, 2009; Chiang et al., 2013, and references therein). Extensions like indicator co-Kriging require a large number of cross-covariances (see Agarwal et al., 2021). In the present paper, the proposed method can be applied to non ordinal data, and does not require a specific model and thresholds between indicators of membership and an underlying real random field; furthermore in a simplified setting, the whole method can also rely on a single covariance function.

Proposal. To the best of our knowledge, the use of Kriging on several outputs with application to classification under constraints on predicted values has not been developed yet.

By itself, as previously detailed, Joint Kriging on several outputs has advantages for interpretability, for its capacity to interpolate data, for the uncertainty measurement associated to each prediction, for the limited number of hyperparameters and the simplicity of their estimation, while allowing specific model characteristics as periodicity for example. The present approach yields directly closed form formulas, without the need of conditional density approximations or sampling. With constraints on predicted values, Kriging is useful for using external information, for adverse modelling, for the need to homogenize results or to introduce fairness constraints. Such

original constraints are not typically addressed by Kriging or Gaussian Process regression, but we show that with suitable adaptations, Kriging is an excellent choice for this task.

Applied to classification, the very same advantages hold: it allows for capacity constraints, as well as model uncertainty, interpretability, simple covariance modelling, and sometimes interpolation. Contrarily to Bayesian inference methods relying on integral approximations or Monte-Carlo sampling, this provides directly closed-form formulas for the predictor, using only basic linear algebra. It does not require creating a model where classes are derived from underlying Gaussian Processes and thresholds, so that it is also suited for non ordinal classes. It seems to us that using multi-outputs Kriging on classification offers many modelling perspectives, as well as practical results and performance: we will see that the proposed model compete with best available methods on an open data set.

The paper is structured as follows. In Section 2, one defines a simplified Kriging model that is suited for multi-valued outputs. The model is detailed in different cases: without specific constraint, similarly to Simple Kriging; with weights summing to 1, similarly to Ordinary Kriging; with constraints on weights summing to 1 and on average predicted values. In each case we derive optimal weights together with prediction mean and variance. An extension using an affine prediction is also developed. In Section 3, the proposed interpolation technique is applied to membership degrees, and it is shown that it preserves useful basic properties for the prediction. In Section 4, numerical applications of the proposed interpolation technique are given. One considers in particular a minimal application on a toy example, an illustration on a multi-valued time series on a real data set, and a more detailed real-world application on a classification problem. A conclusion closes the paper.

Appendix and supplementary material. For more readability, all proofs are gathered in a Section A provided in Appendix. A list of notations is given in a dedicated section B. All illustrations in the paper are generated with notebooks that are available as supplementary material at `https://gitlab.emse.fr/marc.grossouvre/jointkrigingsupplementary/`, in modifiable and executable format `.Rmd` and in already executed directly readable `.html` format (see Grossouvre and Rullière, 2023). Hence the results are fully reproducible, and all specifications for drawing figures are easy to retrieve.

2 Joint Kriging model

Let us consider a multi-valued random field $\mathbf{Y}(x) := (Y_1(x), \ldots, Y_p(x))^\top \in \mathbb{R}^p$, $x \in \mathcal{X}$ where $\mathcal{X}$ is a metric set of input points, typically $\mathcal{X} = \mathbb{R}^d$.

For the sake of clarity and using analogy with geostatistics, we will refer to x as *locations*, but of course $\mathcal{X}$ may contain any explanatory variable. The components $Y_1(.), \ldots, Y_p(.)$ will be referred to as the p considered *targets*.

Components of the random field $\mathbf{Y}(x)$ can be dependent. Furthermore $\mathbf{Y}$ or its components are not necessarily Gaussian. However one assumes that first and second order moments exists. One considers here that $\mathbf{Y}(x) \in \mathbb{R}^p$ and $\mathcal{X} = \mathbb{R}^d$, but other metric spaces would be possible, as soon as expectation and covariances between $\mathbf{Y}(x)$ and $\mathbf{Y}(x')$ can be derived.

Given n observations of $\mathbf{Y}(x_1), \ldots, \mathbf{Y}(x_n)$, we aim at predicting the values of the random field at some unobserved locations $x_1^\star, \ldots, x_q^\star$, i.e. we aim at giving a predictor of $\mathbf{Y}(x_1^\star), \ldots, \mathbf{Y}(x_q^\star)$. At an unobserved location $x^\star$, we define the *Joint Kriging predictor* as a predictor $\mathbf{M}(x) = (M_1(x), \ldots, M_p(x))^\top$ depending linearly on observations, where real coefficients apply *jointly* to all components of the observations:

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

where each $\alpha_i(x^\star) \in \mathbb{R}$. These weights $\boldsymbol{\alpha}(x^\star) := (\alpha_1(x^\star), \ldots, \alpha_n(x^\star))^\top$ are optimized in order to minimize some error that we will detail later on, under various possible constraints. Now, defining the $p \times n$ matrix $\mathbb{Y} := [\mathbf{Y}(x_1), \ldots, \mathbf{Y}(x_p)]$, Equation (1) also writes in a compact way

$$\mathbf{M}(x^\star) = \mathbb{Y}\boldsymbol{\alpha}(x^\star)\,. \tag{2}$$

The main assumption behind this Joint Kriging predictor is that the weights are impacting all components the same way: the first component $M_1(x^\star)$ is a linear combination of the observed first components, namely $Y_1(x_1), \ldots, Y_1(x_n)$, the second component $M_2(x^\star)$ is *the same* linear combination of the observed second components $Y_2(x_1), \ldots, Y_2(x_n)$, etc. In other words, the weights affect jointly (or simultaneously) all the components of observed $\mathbf{Y}(x_i)$, $i = 1, \ldots, n$, hence the chosen name of *Joint Kriging* model. This assumption can be quite logical, and easier to explain, to a decision maker: all targets are interpolated given observed targets, and the interpolation weights do not depend on the considered target. We will see in Section 3 that this key assumption is very useful, especially for classification under constraints. It would be technically possible to release this assumption, e.g. replacing weights $\alpha_i(x)$ by some $p \times p$ matrix for $i = 1, \ldots, n$: one would get closer to some general co-Kriging model with $O(p^2)$ covariance models, but it is not the purpose of the present paper.
Let us define the prediction error associated to a vector of weights $\boldsymbol{\alpha}(x^\star)$, at

a prediction location $x^\star$. This loss is defined as the scalar value

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

where $\|\mathbf{v}\|^2_{\mathbb{W}} := \mathbf{v}^\top \mathbb{W}\mathbf{v}$ is a squared norm with $\mathbb{W}$ a given symmetrical positive-definite matrix of real weights. For instance, if one changes the unit of the first target, say multiply it by 100, then it sounds logical that the resulting norm is unchanged. Thus, some weights matrix may seem reasonable: an inverse covariance matrix as in the Mahalanobis distance, or a diagonal matrix of inverse variances, etc. For simplicity, the reader may imagine that all p targets are already scaled and that $\mathbb{W}$ is the $p \times p$ identity matrix.

The main difficulty is to derive the optimal weights $\boldsymbol{\alpha}(x^\star)$ under the various constraints one would like to consider. At all prediction locations $x^\star_1, \ldots, x^\star_q$ one thus aims at determining the optimal weights, gathered in a $n \times q$ matrix

$$\mathbb{A} := \left[\boldsymbol{\alpha}(x^\star_1), \ldots, \boldsymbol{\alpha}(x^\star_q)\right] .$$

This is performed in the three following subsections, under different constraints.

2.1 Optimal weights without constraints

In this subsection, one defines optimal weights minimizing the prediction error, without supplementary constraints.

The following Proposition expresses the weights such that $\mathbf{M}(x^\star)$ is a best linear unbiased predictor (BLUP) of $\mathbf{Y}(x^\star)$, in the sense of minimizing the loss (3). The result looks exactly the same as in the simple Kriging model, but the components in the symmetric positive semidefinite matrix $\mathbb{K}$ and in the vector $\mathbf{h}(x^\star)$ here aggregate the values of all p observed targets. One retrieves usual Simple Kriging equations in the case where $p = 1$ and $\mathbb{W}$ is the identity matrix.

Proposition 1 (Simple Joint Kriging weights)**.** *The optimal weights* $\boldsymbol{\alpha}(x^\star)$ *minizing the loss of Equation* (3) *are given by the* $n \times 1$ *vector:*

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

or equivalently, using a matrix expression,

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

where the $n \times n$ *matrix* $\mathbb{K} := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W}\mathbb{Y}\right]$ *is assumed to be invertible, the* $n \times 1$ *vector* $\mathbf{h}(x^\star) := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W}\mathbf{Y}(x^\star)\right]$*, and the* $n \times q$ *matrix* $\mathbb{H} := \left[\mathbf{h}(x^\star_1), \ldots, \mathbf{h}(x^\star_q)\right]$*. If furthermore for any target* $j = 1, \ldots, p$*, for any location* $x \in \mathcal{X}$*,* $\mathrm{E}\left[Y_j(x)\right] = 0$*, then* $\mathbf{M}(x^\star)$ *is unbiased.*

Proof. The proof is postponed to Appendix. □

Note that the matrix $\mathbb{K}$ is necessarily a covariance matrix, since it is symmetric positive semidefinite. In particular, this appears clearly when $\mathrm{E}\left[Y_j(x)\right] = 0$ for any target $j = 1, \ldots, p$ and any location $x \in \mathcal{X}$.

2.2 Optimal weights summing to one

In this section, one considers an additional constraint. This constraint raises naturally when $Y_i(x)$ are not centered, and leads to weights summing to one, as in Ordinary Kriging, see Cressie (1988), namely for all $x^\star$,

$$\boldsymbol{\alpha}^\top(x^\star)\mathbf{1}_n = 1\,. \tag{6}$$

Where $\mathbf{1}_n$ is a $n \times 1$ vector of ones. It is clear that if for any $x \in \mathcal{X}$, $\mathrm{E}\left[\mathbf{Y}(x)\right] = \boldsymbol{\mu}$ then constraint (6) will imply that $\mathrm{E}\left[\mathbf{M}(x^\star)\right] = \boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$, so that $\mathrm{E}\left[\mathbf{M}(x^\star)\right] = \mathrm{E}\left[\mathbf{Y}(x^\star)\right]$. The reverse implication is also true if all μ_i are distinct from zero. Hence constraint (6) is a very natural constraint. It does not imply however that $\mathbf{M}(x^\star)$ is a convex combination of all $\mathbf{Y}(x_i)$, because some weights can still be negative.

Under this constraint of weights summing to one, the following proposition gives the optimal weights. One retrieves exactly the same formulae as in ordinary Kriging, but the involved elements in matrices $\mathbb{K}$ and $\mathbb{H}$ are different: they are computed taking into account all p targets over all observations.

Proposition 2 (Ordinary Joint Kriging weights)**.** *Under the constraint of Equation* (6)*, the optimal weights* $\boldsymbol{\alpha}(x^\star)$ *minimizing the loss of Equation* (2) *are given by the* $n \times 1$ *vector* $\boldsymbol{\alpha}(x^\star)$*, together with the scalar* $\lambda(x^\star)$*:*

$$\left\{\begin{array}{rcl} \boldsymbol{\alpha}(x^\star) & = & \mathbb{K}^{-1}\left(\mathbf{h}(x^\star) + \lambda(x^\star)\mathbf{1}_n\right) \\ \lambda(x^\star) & = & \frac{1}{\delta}\left(1 - \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbf{h}(x^\star)\right) \end{array}\right. \tag{7}$$

Equivalently, using matrix expressions, one gets

$$\left\{\begin{array}{rcl} \mathbb{A} & = & \mathbb{K}^{-1}\left(\mathbb{H} + \mathbf{1}_n\boldsymbol{\lambda}^\top\right) \\ \boldsymbol{\lambda}^\top & = & \frac{1}{\delta}\left(\mathbf{1}_q{}^\top - \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbb{H}\right) \end{array}\right. \tag{8}$$

where $\mathbb{K} := \mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbb{Y}\right]$*,* $\mathbf{h}(x^\star) := \mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbf{Y}(x^\star)\right]$*, and with scalar* $\delta := \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbf{1}_n$*. For matrix expressions,* $\boldsymbol{\lambda} := (\lambda(x_1^\star), \ldots, \lambda(x_q^\star))^\top$*, and the* $n \times q$ *covariance matrix* $\mathbb{H} := \left[\mathbf{h}(x_1^\star), \ldots, \mathbf{h}(x_q^\star)\right]$*.*
If furthermore, for all targets $i = 1, \ldots, p$*, for all locations* $x \in \mathcal{X}$*,* $\mathrm{E}\left[Y_i(x)\right] = \mu_i$*, then* $\mathbf{M}(x^\star)$ *is unbiased.*

Proof. The proof is postponed to Appendix. □

An originality of the presentation of the result is that matrices can be expressed indifferently with compact expressions (using $\mathbb{K}$ and $\mathbf{h}(x^\star)$) or with more classical covariances (using $\widetilde{\mathbb{K}}$ and $\widetilde{\mathbf{h}}(x^\star)$), as stated in the next remark.

Remark 1 (Covariance matrices)**.** *Let us define the true unknown values of* $\mathbf{Y}$ *at all prediction points by* $\mathbb{Y}^\star := \left[\mathbf{Y}(x_1^\star), \ldots, \mathbf{Y}(x_q^\star)\right]$*. Assume* $\mathrm{E}\left[\mathbf{Y}(x)\right] = \boldsymbol{\mu}$ *for all* $x \in \mathcal{X}$*. Furthermore, assume that either weights sum to one that is to say* $\boldsymbol{\alpha}(x^\star)^\top \mathbf{1}_n = 1$*, or* $\boldsymbol{\mu} = \mathbf{0}_p$*.*

Then the matrices $\mathbb{K}$*,* $\mathbb{H}$ *and the vector* $\mathbf{h}(x^\star)$ *can be replaced by*

$$\left\{\begin{array}{lcl} \widetilde{\mathbb{K}} & = & \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbb{Y}\right] - \mathrm{E}\left[\mathbb{Y}^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbb{Y}\right] \\ \widetilde{\mathbb{H}} & = & \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbb{Y}^\star\right] - \mathrm{E}\left[\mathbb{Y}^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbb{Y}^\star\right] \\ \widetilde{\mathbf{h}}(x^\star) & = & \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbf{Y}(x^\star)\right] - \mathrm{E}\left[\mathbb{Y}^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbf{Y}(x^\star)\right] \end{array}\right. \tag{9}$$

everywhere in Proposition 2, without changing the optimal weights $\boldsymbol{\alpha}(x^\star)$*.*

The next remark shows that the model can be built from a single covariance function $k(x, x')$ between an implicitly weighted sum of all targets.

Remark 2 (Using correlation functions)**.** *Assume that there exists a positive definite matrix* $\mathbb{W}$*, such that the covariances*

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

depend only on some distance between x *and* x'*.*

Then the components of the covariances matrices $\widetilde{K}$ *and* $\widetilde{H}$ *can be derived from the covariance function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *by setting:*

$$\left\{\begin{array}{lcl} \widetilde{\mathbb{K}}_{ij} & = & k(x_i, x_j) \\ \widetilde{\mathbb{H}}_{ik} & = & k(x_i, x_k^\star) \end{array}\right. \tag{10}$$

In practice, with one hyperparameter for the variance $\sigma^2 > 0$*, and* d *positive hyperparameters* $\theta_1, \ldots, \theta_d$*, usually referred to as characteristic length-scales (see Rasmussen et al., 2006, bottom of p.14), one can set for example*

$$k(x, x') = \sigma^2 r_0\left(\|x - x'\|_{\boldsymbol{\theta}}\right), \tag{11}$$

where r_0 *a unit correlation function and* $\|x - x'\|_{\boldsymbol{\theta}}^2 = \sum_{i=1}^d \left(\frac{x_i - x_i'}{\theta_i}\right)^2$ *corresponds to a rescaled Euclidean norm. Notice that this expression does not depend on* $\mathbb{W}$ *any more, so that when using the above assumption we do not have to estimate* $\mathbb{W}$*.*

In Remark 2, $k(x, x) = \sigma^2$ for all $x \in \mathcal{X}$. It would be possible to consider a non homogeneous variance, this is not treated here for the sake of simplicity.

As a consequence of previous Remark 2, for a given matrix $\mathbb{W}$ (e.g. once the components variance is set to one, say), a noticeable advantage of the

Joint Kriging method is the limited number of hyperparameters to optimize. Despite the multivariate output of the $p \times 1$ response vector $\mathbf{Y}(x)$, $x \in \mathcal{X}$, there is only a few hyperparameters to choose for defining the covariances: for instance σ^2, $\boldsymbol{\theta}$, and the covariance family. This is quite different from co-Kriging techniques where all cross-covariances between components $Y_i(x)$ and $Y_j(x')$ should be defined, $i, j \in \{1, \ldots, p\}$, $x, x' \in \mathcal{X}$, which ends up in an order of $O(p^2)$ covariance functions, and many associated hyperparameters.

2.3 Optimal weights with constraint on predictions

The constraint we consider here is more original than the previous one: we would like that, given observations $\mathbb{Y}$, the average of predicted values has some prescribed value. Formally

$$\mathrm{E}\left[\mathbf{M}(X^\star) \,|\, \mathbb{Y}\right] = \mathbf{m}\,. \tag{12}$$

where $X^\star$ is a random variable taking values in prediction locations $\{x_1^\star, \ldots, x_q^\star\}$, and where $\mathbf{m}$ is a given $p \times 1$ vector. This constraint is quite original at it relies on predicted values for a given set of observations. The idea is to force the optimal weights to take into account this *a posteriori* constraint.

Notice the importance of the conditioning by $\mathbb{Y}$, otherwise if all $Y_i(x)$ are centered, say, then the constraint would not be possible to satisfy in general, since all $\mathbf{M}(x^\star)$ would be centered. We will see that this kind of constraint is particularly useful for classification, when one wishes to force classes proportions whatever the observed values.

Gathering all predictors in a $p \times q$ matrix $\mathbb{M} = \left[\mathbf{M}(x_1^\star), \ldots, \mathbf{M}(x_q^\star)\right]$, one can write $\mathbb{M} = \mathbb{Y}\mathbb{A}$. By conditioning on the value of $X^\star$, denoting $\pi_{x^\star} = \mathrm{P}\left[X^\star = x^\star\right]$, $x^\star \in \{x_1^\star, \ldots, x_q^\star\}$, Equation(12) simply writes

$$\mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m}\,. \tag{13}$$

where $\boldsymbol{\pi}$ is the $q \times 1$ column vector $\boldsymbol{\pi} = (\pi_{x_1^\star}, \ldots, \pi_{x_q^\star})$, and where the $n \times q$ matrix $\mathbb{A}$ gather all vectors of weights: $\mathbb{A} := \left[\boldsymbol{\alpha}(x_1^\star), \ldots, \boldsymbol{\alpha}(x_q^\star)\right]$. One specificity is that the resulting weights in matrix $\mathbb{A}$ will have to be solved all at once, for all q prediction locations. This is quite different from usual Kriging settings where prediction locations can be treated separately, if desired.

In the following proposition, we give the optimal weights matrix $\mathbb{A}$ when both constraints are considered at the same time: the constraint (12) on predicted values, and the constraint (6) on weights summing to one.

Proposition 3 (Joint Kriging weights under predicted values constraint)**.** *The Joint Kriging weights minizing the loss of Equation* (3) *under the constraint of weights summing to one of Equation* (6)*, and prescribed average predicted values of Equation* (12) *write:*

$$\mathbb{A} = \mathbb{K}^{-1}\left(\mathbb{H} + \mathbf{1}_n\boldsymbol{\lambda}^\top + \mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top\right) \tag{14}$$

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* $\delta := \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbf{1}_n \in \mathbb{R}$, *and* $\mathbb{H} := \left[\mathbf{h}(x_1^\star), ..., \mathbf{h}(x_q^\star)\right]$.

Proof. The proof is postponed to Appendix. □

Note that it is also possible to compute a model with the only constraint $\mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m}$, without requiring weights summing to one. In view of further classification applications, we do not develop it here and keep both constraints.

Again, an originality is that the previous result can be expressed using compact expressions for $\mathbb{K}$ and $\mathbb{H}$ or more classical covariances, as stated in the following remark. As a result, under suitable assumptions, all these covariance matrices can be filled using a covariance function $k(.,.)$, as detailed in Remark 2.

Remark 3 (Covariance matrices with two constraints)**.** *Under the assumptions of Remark 1 and using the same notations, the matrices* $\mathbb{K}$ *and* $\mathbb{H}$ *can be replaced by* $\widetilde{\mathbb{K}}$ *and* $\widetilde{\mathbb{H}}$ *everywhere in Proposition 3, without changing the optimal weights* $\mathbb{A}$.

Proof. The proof is postponed to Appendix. □

2.4 Optimal weights with affine extension

Up to this point, one has only considered linear predictors, where a predictor is a linear combination of observed responses $\mathbf{Y}(x_1), \ldots, \mathbf{Y}(x_n)$, under various constraints. In this subsection, one considers the case where the prediction involves one supplementary term.

Remind the requirement on predicted values: they should be $\mathbf{m}$ in average. The hidden idea behind the knowledge of vector $\mathbf{m}$ is that there is some external information giving an hint on the prediction. This information may come for instance from some known overall statistics on the territory, some expert knowledge or from an expectancy estimator. Let us denote by $\mathbf{Z}$ the $p \times 1$ random vector containing this external source of information.

With this in mind, it is quite natural to define an affine prediction:

$$\mathbf{M}^+(x^\star) := \alpha_0(x^\star)\mathbf{Z} + \sum_{i=1}^{n}\alpha_i(x^\star)\mathbf{Y}(x_i)\,, \tag{15}$$

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

The sum of weights constraint on the new vector $\boldsymbol{\alpha}^+ = (\alpha_0(x^\star), \ldots, \alpha_n(x^\star))$ can be written

$$\mathbf{1}_{n+1}{}^\top \boldsymbol{\alpha}^+(x^\star) = 1\,. \tag{16}$$

This way, if the p components of $\mathbf{m}$ and $\mathbf{Y}(x_i)$, $i = 1, \ldots, n$ are percentages summing to one, then the p components of the predictor $\mathbf{M}(x^\star)$ will also sum to one.

For the second constraint on average predicted values, previously detailed in Equation (12), there is an implicit conditioning by $\mathbf{Z} = \mathbf{m}$. this constraint may write, with $X^\star$ a r.v. defined on $\{x_1^\star, \ldots, x_q^\star\}$:

$$\mathrm{E}\left[\mathbf{M}^+(X^\star) \,\middle|\, \mathbf{Z} = \mathbf{m},\, \mathbb{Y}\right] = \mathbf{m}\,. \tag{17}$$

Finally, provided covariances between $\mathbf{Z}$ and $\mathbf{Y}(x)$ are given, $x \in \mathcal{X}$, then the setting is absolutely the same as in previous Propositions 1, 2 and 3, excepts one observation $\mathbf{Z} = \mathbf{m}$ is added in the vectors of observations $\mathbf{Y}(x_1), \ldots, \mathbf{Y}(x_n)$. The covariance matrices are also updated. This is detailed in the following Proposition.

Proposition 4 (Affine version of predictors)**.** *Assume that the following covariance vectors are given*

$$\left\{\begin{array}{lcl} \mathbf{P}^\top & := & \mathrm{E}\left[\mathbf{Z}^\top \mathbb{W}\mathbb{Y}\right] - \mathrm{E}\left[\mathbf{Z}^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbb{Y}\right] \\ \mathbf{Q}^\top & := & \mathrm{E}\left[\mathbf{Z}^\top \mathbb{W}\mathbb{Y}^\star\right] - \mathrm{E}\left[\mathbf{Z}^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbb{Y}^\star\right] \\ \sigma_Z^2 & := & \mathrm{E}\left[\mathbf{Z}^\top \mathbb{W}\mathbf{Z}\right] - \mathrm{E}\left[\mathbf{Z}^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbf{Z}\right] \end{array}\right. \tag{18}$$

Then affine predictors corresponding to the simple unconstrained case, to the ordinary case with one constraint, to the case with two constraints can be obtained by replacing $\mathbb{Y}$, $\mathbb{K}$, $\mathbb{H}$ by

$$\mathbb{Y}^+ = \begin{bmatrix} \mathbf{m} & \mathbb{Y} \end{bmatrix}, \quad \mathbb{K}^+ = \begin{bmatrix} \sigma_Z^2 & \mathbf{P}^\top \\ \mathbf{P} & \mathbb{K} \end{bmatrix}, \quad \mathbb{H}^+ = \begin{bmatrix} \mathbf{Q}^\top \\ \mathbb{H} \end{bmatrix}, \tag{19}$$

in Propositions 1, 2 and 3 respectively.

Proof. The proof is straightforward, hence not appearing in the Appendix. □

Notice that the previous Proposition 4 can be easily extended to several sources of information $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$. For the sake of simplicity, this is not developed here.

Recall that, from Remark 2, that matrices $\mathbb{K}, \mathbb{H}$ can be derived from simple correlation functions. Now it remains to derive one expression for $\mathbf{P}$ and $\mathbf{Q}$.

Remark 4 (Extra covariances for affine prediction)**.** *Let* $\mathbf{P} = (P_1, \dots, P_n)$, $\mathbf{Q} = (Q_1, \dots, Q_q)$ *and* $\mathbf{Z} = (Z_1, \dots, Z_p)$. *To fix ideas, in the case where* $\mathbb{W}$ *is the identity matrix, then for* $i = 1, \dots, n$, $j = 1, \dots, q$

$$\left\{ \begin{array}{rcl} P_i & = & \sum_{k=1}^{p} \mathrm{Cov}\left[Z_k, Y_k(x_i)\right] \\ Q_j & = & \sum_{k=1}^{p} \mathrm{Cov}\left[Z_k, Y_k(x_j^\star)\right] \end{array} \right.$$

This is easy to adapt when $\mathbb{W}$ *is not the identity matrix, by considering the components of* $\mathbb{W}^{1/2}\mathbf{Z}$ *and* $\mathbb{W}^{1/2}\mathbf{Y}(x)$, $x \in \mathcal{X}$.

Assuming that the above covariance does not depend on x_i *nor on* $x_j^\star$ *(i.e. the general source of information gives hints on the whole process, not on a particular location), then one can propose*

$$\left\{ \begin{array}{l} P_i = \rho\sigma\sigma_Z\,, \quad i = 1, \dots n \\ Q_j = \rho\sigma\sigma_Z\,, \quad j = 1, \dots q \end{array} \right. \tag{20}$$

This corresponds for example to a model where $\frac{Y_k(x)}{\sigma} = \rho\frac{Z_k}{\sigma_Z} + G_k(x)$, $k = 1, \dots, p$, *where all* $G_k(x)$ *are independent from all* Z_k.

Other assumptions can be chosen, leading to different vectors $\mathbf{P}$ *and* $\mathbf{Q}$.

In the previous remark 4, the parameter $\rho \in [-1, 1]$ measures how redundant is the information provided by $\mathbf{Z}$, and can even be set to 0 if one considers that the external information source is completely independent from observations. The parameter σ_Z measures how certain is the external information: when σ_Z is high, the added information cannot be trusted and one retrieves the linear predictor, when σ_Z is low, the added information is trustable, so that far from observed locations, $\mathbf{M}(x)$ gets nearer to $\mathbf{m}$. In practice one can set $0 < \sigma_Z \ll \sigma$ to see the maximal difference with the linear predictor. One can even optimize this parameter σ_Z to smoothly switch from a linear to an affine model.

2.5 Joint Kriging Mean and Variance

In this subsection we derive the mean predictor and the prediction error, assuming the optimal weights have been calculated with chosen constraints, as detailed in previous subsections.

Consider $\mathbf{M}(x^\star)$ and $\boldsymbol{\alpha}(x^\star)$ a Joint Kriging predictor and the associated weights with or without constraints. In the following, we call *Joint Kriging mean* the value of the the predictor $\mathbf{M}(x^\star)$ and *Joint Kriging variance* the value of the quadratic error $\Delta(x^\star)$. Let us recall that:

$$\left\{ \begin{array}{rcl} \mathbf{M}(x^\star) & := & \mathbb{Y}\boldsymbol{\alpha}(x^\star)\,, \\ \Delta(x^\star) & := & \mathrm{E}\left[\|\mathbf{M}(x^\star) - \mathbf{Y}(x^\star)\|_{\mathbb{W}}^2\right]\,, \end{array} \right.$$

where $\mathbb{Y} = [\mathbf{Y}(x_1), \ldots, \mathbf{Y}(x_p)]$ is the $p \times n$ matrix of observations. If $p = 1$ and if $\mathbb{W}$ is the identity matrix, Joint Kriging mean and Joint Kriging variance are exactly the Kriging mean and the Kriging variance usually known in Kriging.

The following Proposition gives a closed formula to compute the Joint Kriging variance.

Proposition 5 (Joint Kriging variance with arbitrary weights)**.** *Let $\boldsymbol{\alpha}(x^\star)$ be any vector of weights, possibly satisfying supplementary constraints.*

The associated Joint Kriging variance writes:

$$\Delta(x^\star) = \boldsymbol{\alpha}(x^\star)^\top \mathbb{K} \boldsymbol{\alpha}(x^\star) - 2\boldsymbol{\alpha}(x^\star)^\top \mathbf{h}(x^\star) + v(x^\star)\,, \tag{21}$$

or using a matrix expression, denoting $\boldsymbol{\Delta} := (\Delta(x_1^\star), \ldots, \Delta(x_1^\star))^\top$, we get

$$\boldsymbol{\Delta} = \operatorname{diag}\left[\mathbb{A}^\top \mathbb{K} \mathbb{A}\right] - 2\operatorname{diag}\left[\mathbb{A}^\top \mathbb{H}\right] + \operatorname{diag}[\mathbb{K}^\star]\ ,$$

where the $n \times n$ matrix $\mathbb{K} := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbb{Y}\right]$, the $q \times q$ matrix $\mathbb{K}^\star := \mathrm{E}\left[\mathbb{Y}^{\star\top} \mathbb{W} \mathbb{Y}^\star\right]$, the $n \times 1$ vector $\mathbf{h}(x^\star) := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbf{Y}(x^\star)\right]$, and the scalar $v(x^\star) := \mathrm{E}\left[\mathbf{Y}(x^\star)^\top \mathbb{W} \mathbf{Y}(x^\star)\right]$ are assumed to be known. diag[.] *is the vector whose entries are the diagonal of the considered matrix.*

Proof. The proof is postponed to Appendix. □

Note that the above Proposition 5 can be directly adapted to the affine case of Proposition 4, by replacing $\mathbb{Y}, \mathbb{K}, \mathbb{H}$ by $\mathbb{Y}^+, \mathbb{K}^+, \mathbb{H}^+$, v being unchanged: one can interpret the predictor to be a linear predictor with one more observation, with correct covariances.

As previously stated in Remarks 1 and 3, and using the same notations, one can replace $\mathbb{K}, \mathbb{H}, \mathbf{h}$ with $\widetilde{\mathbb{K}}, \widetilde{\mathbb{H}}, \widetilde{\mathbf{h}}$, provided that the following new quantities are defined:

$$\begin{cases} \widetilde{\mathbb{K}}^\star & = & \mathrm{E}\left[\mathbb{Y}^{\star\top} \mathbb{W} \mathbb{Y}^\star\right] - \mathrm{E}\left[\mathbb{Y}^{\star\top}\right] \mathbb{W}\, \mathrm{E}\left[\mathbb{Y}^\star\right] \\ \widetilde{v}(x^\star) & = & \mathrm{E}\left[\mathbf{Y}(x^\star)^\top \mathbb{W} \mathbf{Y}(x^\star)\right] - \mathrm{E}\left[\mathbf{Y}(x^\star)^\top\right] \mathbb{W}\, \mathrm{E}\left[\mathbf{Y}(x^\star)\right]\ . \end{cases}$$

The result is stated in Remark 5 below. Hence, in practice all these covariances can be filled using a given covariance function $k(x, x')$, under suitable assumptions, as detailed in Remark 2.

Remark 5 (Covariance matrices in Joint Kriging mean and variance)**.** *Under the assumptions of Remark 1 and using the same notations, the matrices $\mathbb{K}$, $\mathbb{H}$, $\mathbb{K}^\star$, the vector $\mathbf{h}(x^\star)$ and the scalar $v(x^\star)$ can be replaced by $\widetilde{\mathbb{K}}$, $\widetilde{\mathbb{H}}$, $\widetilde{\mathbb{K}}$, $\widetilde{\mathbf{h}}(x^\star)$ and $\widetilde{v}(x^\star)$ everywhere in Proposition 5, without changing the Joint Kriging mean and variance.*

Proof. The proof is postponed to Appendix. □

Now, remark that Proposition 5 gives only an overall error

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

which is a weighted sum of errors over all components of $\mathbf{M}(x^\star)$.

This is a strength of the method, since the quantity to optimize is real-valued, which allows using standard covariance functions as detailed in Remark 2. This is also an important limitation, because in practice, one surely needs prediction errors for each component of $\mathbf{M}(x)$:

$$\delta_i(x^\star) := \mathrm{E}\left[\|M_i(x^\star) - Y_i(x^\star)\|^2\right], \quad i = 1, \dots, p\,.$$

The following Proposition 6 shows that one can get this error $\delta_i(x^\star)$ for each component $i = 1, \dots, p$. It relies on a supplementary assumption on the matrix $\mathbb{W}$, but this assumption is only useful for determining the confidence bands for each component of the predictor $\mathbf{M}(x)$, not for computing $\mathbf{M}(x)$ itself.

Proposition 6 (Variance sharing)**.** *Assume that transformed observations $\widetilde{\mathbf{Y}}(x) := \mathbb{W}^{1/2}\mathbf{Y}(x)$ are such that components of $\widetilde{\mathbf{Y}}$ are uncorrelated and bear the same share of the covariance function k, that is to say:*

$$\mathrm{Cov}\left[\widetilde{Y}_i(x), \widetilde{Y}_j(x')\right] = \frac{1}{p}k(x, x')\mathbb{1}_{\{i=j\}}, \quad i, j \in \{1, \dots, p\},\ x, x' \in \mathcal{X}, \quad (22)$$

Assume also that $\mathrm{E}[\mathbf{Y}(x)] = \boldsymbol{\mu}$ for all $x \in \mathcal{X}$. Furthermore, assume that either the weights sum to one, or $\boldsymbol{\mu} = \mathbf{0}_p$. Then the local errors write

$$\delta_i(x^\star) = \frac{\sigma_i^2}{\sigma^2}\Delta(x^\star), \quad i = 1, \dots, p\,. \quad (23)$$

where $\sigma_i^2 := \mathrm{Var}[Y_i(x)]$ is the variance of the component $Y_i(x)$, assumed to be constant over x.

Proof. The proof is postponed to Appendix. □

The result of Proposition 6 is quite logical, it simply states that for a well chosen matrix $\mathbb{W}$, the error $\delta_i(x^\star)$ is proportional to the unit global error $\sigma^{-2}\Delta(x^\star)$: one simply has to apply the variance σ_i^2 of the component instead of the variance σ^2 of the aggregated weighted components.

3 Application to constrained classification

In this section, we now apply multi-output prediction to membership degrees for fuzzy classification. We show that the Joint Kriging predictor, together with constraints on weights and predicted values, is especially suited to this task since it remains very simple, and constraints are very logical in a classification setting.

Consider a classification problem with p possible labels. Labels are depending on some explanatory variables $x \in \mathcal{X}$, so that one may observe labels $\ell(x_1), \ldots, \ell(x_n)$ taking values in $\{1, \ldots, p\}$.

At an unobserved location $x^\star$, for a predictor $L(x^\star) \in \{1, \ldots, p\}$ of $\ell(x^\star)$, the reader may convince himself that, for given probabilities p_j, $j \in \{1, \ldots, p\}$, constraints on predicted classification such as

$$\mathrm{P}\left[L(X^\star) = j \,|\, \text{observed labels}\right] = p_j \,, \tag{24}$$

are not so easy to handle, even if $X^\star$ is a uniformly distributed random variable over prediction points $x_1^\star, \ldots, x_q^\star$. This is because such constraints usually act in a non-linear way on the predictor $L(X^\star)$, and the predictor $L(.)$ itself can be some complicated function of observed labels. Existing predictors, such as Indicator Kriging, may be unable to deal with such constraints. Furthermore, they may not be appropriate in cases where the considered labels do not correspond to ordinal classes.

The originality here is to use Joint Kriging to propose a fuzzy classification under both constraints: membership degrees summing to one, and prescribed average of predicted degrees. Indeed, in a practical context, it is natural to require that over all predictions, predicted values are distributed like the observed ones, for instance. Or the sampling bias of observations is known so that the expected label percentages are known. Or sometimes an external source of information gives the expected label percentages. It can be the case for a regional study, knowing some statistics at a national level. At last, it can be used for modelling adverse scenarios.

Technically, in a classification problem, each label $\ell \in \{1, \ldots, p\}$ can be converted into a $p \times 1$ vector of indicator functions, namely

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

This transformation is well known in the machine learning community as *label binarization* (see also *dummy variables* or *one-hot encoding*), and is implemented in many languages. Obviously using this simple representation, ${\mathbf{1}_p}^\top \mathbf{Y} = 1$.

In practice, it is common to observe true label values depending on some explanatory variables $x \in \mathcal{X}$. But it may also happen that one observes uncertain labels: multiple and distinct observed labels for a same $x \in \mathcal{X}$, uncertainty relying on the value of x, etc. To handle this problem, one generalizes slightly the previous label binarization: one assumes here that observations consist in a distribution of possible labels, so that one observes n vectors $\mathbf{Y}(x_1), \ldots, \mathbf{Y}(x_n)$, such that the components of each vector are summing to one: ${\mathbf{1}_p}^\top \mathbf{Y}(x_i) = 1$, $i = 1, \ldots, n$. In other words, the p components of $\mathbf{Y}(x_i)$ represent the membership degrees of the p possible classes, at an observed location x_i, $i = 1, \ldots, n$. Using previous notations, recall that $\mathbb{Y} = [\mathbf{Y}(x_1), \ldots, \mathbf{Y}(x_n)]$, so that observed membership degrees satisfy

$$ {\mathbf{1}_p}^\top \mathbb{Y} = {\mathbf{1}_n}^\top . \tag{25} $$

Now, using a Joint Kriging model, one can infer the membership degree of an unobserved location $x^\star$, using the predictor of Equation (1):

$$ \mathbf{M}(x^\star) := \sum_{i=1}^{n} \alpha_i(x^\star) \mathbf{Y}(x_i) \tag{26} $$

The next remark details the impact of both constraints, weights summing to one and prescribed average predicted values, in the particular setting of membership degrees that are summing to one. Despite being very straightforward, it aims in particular at recalling the previous results in the classification context.

Remark 6 (Constraints impact)**.** *Consider the two previous constraints on weights and predicted values, namely constraint of Equation* (6) *and Equation* (12) *. Consider also the membership degree assumption given in Equation* (25)*,* ${\mathbf{1}_p}^\top \mathbb{Y} = {\mathbf{1}_q}^\top$*. Then the Joint Kriging model imply that*

- *Predicted membership degrees are summing to one:*

  $$ {\mathbf{1}_p}^\top \mathbf{M}(x^\star) = 1 , $$

  for any prediction point $x^\star \in \mathcal{X}$*. In particular* ${\mathbf{1}_p}^\top \mathbb{M} = {\mathbf{1}_q}^\top$*.*

- *Average membership degree over prediction points can be chosen:*

  $$ \mathrm{E}\left[\mathbf{M}(X^\star) \,|\, \mathbb{Y}\right] = \mathbf{m} , $$

  where $\mathbf{m}$ *is a prescribed average of predicted membership degrees of each class, with* ${\mathbf{1}_p}^\top \mathbf{m} = 1$*, and* $X^\star$ *a random variable over all prediction points.*

Proof. The proof is postponed to Appendix. □

As noticed before, although predicted membership degrees are summing to one, no guarantee of positivity is forced for the predicted values. In practice, as it is the case with numerous machine learning methods, a post-treatment of results may be required in specific cases involving negative membership degrees (see e.g. the use of softmax function with neural networks).

Recall that, computing the Joint Kriging predictor requires the knowledge of cross moments of $\mathbf{Y}(x)$ and $\mathbf{Y}(x')$, namely $\mathrm{E}\left[\mathbf{Y}(x)^\top \mathbb{W} \mathbf{Y}(x')\right]$ for all observed and predicted locations, i.e. for $x, x' \in \{x_1, \ldots, x_n\} \cup \{x_1^\star, \ldots, x_q^\star\}$, where $\mathbb{W}$ is a given definite-positive matrix. In practice, the Remark 2 can be used to build all covariance matrices from a single covariance function: it is enough to choose a covariance function $k(x, x')$ with few associated hyperparameters to get all needed covariance matrices.

Other techniques could be used for deriving covariances. Ordinal labels may derive from, say, an underlying 1D Gaussian Process $U(x)$, $x \in \mathcal{X}$, by setting $\mathbf{Y}(x) = \left(\mathbb{1}_{\{g(U(x))=j\}}\right)_{j=1,\ldots,p}$, where $g : \mathbb{R} \to \{1, \ldots, p\}$ is a given function. The derivation of all covariances from the ones of $U(.)$ is feasible, as with Indicator Kriging, but out of the scope of the present paper. And finally, it also ends up in the choice of a covariance function $k(x, x')$.

4 Numerical illustrations

In this section one considers different numerical illustrations, for both prediction and classification. The first illustration focuses on the impact of constraints with one output, the second one on the behavior of the predictor with multiple outputs. The third illustration gives an application to classification and a benchmark with numerous competitors. All the illustrations are created in `R markdown` notebooks, one per subsection, available as a supplementary material at `https://gitlab.emse.fr/marc.grossouvre/jointkrigingsupplementary/` (Grossouvre and Rullière, 2023). Notebooks are given in both an executable format and an executed `html` format. The presented figures are directly created from the notebooks and results are fully reproducible.

4.1 A simplified toy example

One considers here the very simple case where there is a single target: the output $\mathbf{Y}(x)$ is belonging to $\mathbb{R}^p$, with $p = 1$. The interest for testing the Joint Kriging with one single target is to discuss the impact of the constraint on predicted values, and the impact of the affine prediction.

For $p = 1$, Simple Joint Kriging and Ordinary Joint Kriging are identical to common Simple Kriging and Ordinary Kriging, but the constraint on predicted values leads to a new original predictor. We keep here the vector

bold font for vectors $\mathbf{Y}(x) \in \mathbb{R}^p$ and $\mathbf{m} \in \mathbb{R}^p$, even though $p = 1$, in order to keep the very same notations as in the rest of the paper.

Let us consider that the process $\mathbf{Y}(x)$ aims at approximating an hidden function, with say $a = 1$ and $b = 4$,

$$f(x) := a + \sin(x/b)\,.$$

Observed locations $x_1, \ldots, x_n$ are randomly chosen with a uniform distribution over the interval $[-10, 5]$, and q prediction locations $x_1^\star, \ldots, x_q^\star$ are chosen regularly spaced over the interval $[-3, 10]$. Both intervals are purposely shifted so that some prediction points are far from observations, and vice-versa.

Observed responses in $\mathbb{R}^p$, with $p = 1$, are $\mathbf{Y}(x_i) = f(x_i)$, $i = 1, \ldots, n$ with $n = 10$. Prediction is made over a set of $q = 100$ points. One defines $X^\star$ a discrete uniform random variable over all prediction points.

The purpose here is not to interpolate as precisely as possible the hidden function f given few observations, but only to illustrate the differences between various possible interpolators, and the impact of requiring a prescribed average values for predicted values.

The prescribed value for $\mathbf{m} \in \mathbb{R}^p$, with $p = 1$, is the scalar $\mathbf{m} = 1.5$. The covariances between $\mathbf{Y}(.)$ are modelled as prescribed in Remark 2, from a single covariance function, using a squared exponential kernel. One could also pick a kernel that reflects f periodicity. However, the purpose is not to do the best possible prediction but, rather, to understand the impact of various constraints.

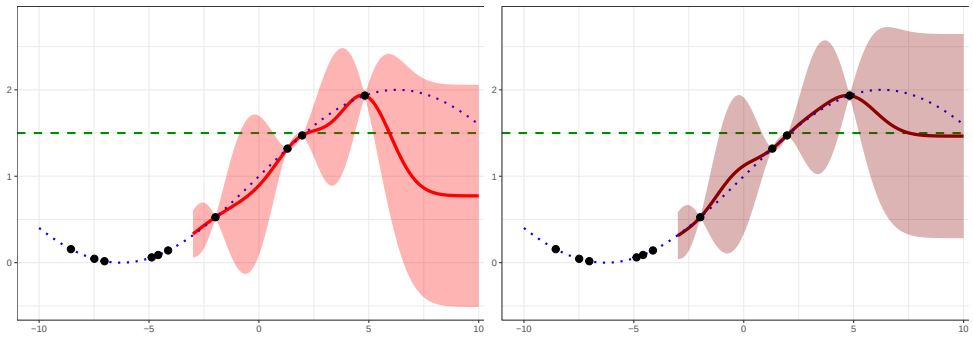
$$\mathrm{Cov}\left[\mathbf{Y}(x), \mathbf{Y}(x')\right] = k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\theta^2}\right)\,.$$

We set $\sigma^2 = 0.6$ mainly for the visibility of the confidence band in presented figures, and $\theta = 1.2$.

In the Figure 1, one exclusively considers the constraint of sum of weights, which is assumed to be one: ${\mathbf{1}_n}^\top \boldsymbol{\alpha} = 1$. The predictor $\mathbf{M}(x)$ appears in red thick line, together with confidence intervals built from the variance $\Delta(x)$.

The Panel 1a presents the result of ordinary Kriging exposed in Proposition 2. As is well known, when the location x is large (and far from observed locations), the ordinary Kriging mean tends to return to the estimated mean of the observations. The average value of the Kriging mean $\mathrm{E}\left[\mathbf{M}(X^\star) \,|\, \mathbb{Y}\right] \simeq 1.12$ is quite different from the value $\mathbf{m} = 1.5$ (horizontal dashed line), which is natural as this constraint has not been taken into account yet.

The Panel 1b uses the Proposition 4 to add a supplementary affine term to the linear combination, while preserving the sum of weights being



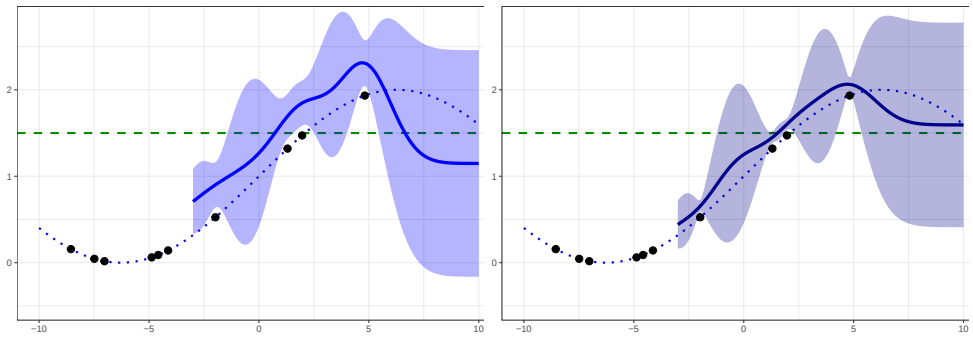
(a) One constraint, linear (b) One constraint, affine

Figure 1: Prediction with one constraint: weights summing to one. Left: linear predictor, and right: affine predictor. In both cases, the average of predicted value is distinct from the prescribed value $\mathbf{m} = 1.5$ (horizontal dashed line). Observations are black dots, the thin dotted blue line is the underlying function. In the right panel, one applies the assumption in Remark 4 with $\rho = 0$ and $\sigma_Z = \sigma/10$.

one. The affine term is derived from a random variable $\mathbf{Z}$, and we choose $\sigma_Z = \sigma/10$, so that this external information is assumed trustable (small variance). Given $\mathbf{Z} = \mathbf{m}$, the consequence is that, far from observed locations, the prediction tends to put all weight on this external source of information, so that the prediction gets closer to $\mathbf{m}$, as one can see at the extreme right of this Figure 1b. This also makes the average $\mathrm{E}[\mathbf{M}(X^\star) \,|\, \mathbb{Y}] \simeq 1.37$ closer to $\mathbf{m} = 1.5$, but the values of those two quantities remain distinct. Another consequence of the affine term is the reduction of the confidence band width, as a new source of information has been added.

In Figure 2, one considers both the constraint of sum of weights, which is assumed to be one: ${\mathbf{1}_n}^\top \boldsymbol{\alpha} = 1$, together with the prescribed average of predicted values $\mathrm{E}[\mathbf{M}(X^\star) \,|\, \mathbb{Y}] = \mathbf{m}$. The predictor $\mathbf{M}(x)$ appears in thick blue line, together with confidence intervals built from the variance $\boldsymbol{\Delta}(x)$.

The Panel 2a presents the result of ordinary Kriging exposed in Proposition 3. The average value of the Kriging mean $\mathrm{E}[\mathbf{M}(X^\star) \,|\, \mathbb{Y}] = 1.5$ is exactly the prescribed one $\mathbf{m} = 1.5$ (horizontal dashed line), which is natural as this constraint has been taken into account during the joint optimization of all $\boldsymbol{\alpha}(x_j^\star)$, $j = 1, \ldots, q$. However, the predictor is no more interpolating. This is logical: if $q = 1$, one has one only prediction point $x_1^\star$, the constraint $\mathrm{E}[\mathbf{M}(X^\star) \,|\, \mathbb{Y}] = \mathbf{m}$ becomes $\mathbf{M}(x_1^\star) = \mathbf{m}$, which is distinct to an observation $\mathbf{Y}(x_n)$, even if $x_1^\star$ gets closer to x_n. Another example: if on the one hand observation points and prediction points are the same, if on the other hand $\mathbf{m}$ is not the average value of observations, then at least one prediction must



(a) Two constraints, linear (b) Two constraints, affine

Figure 2: Prediction with two constraints: weights summing to one, and average of predicted values set to $m = 1.5$ (horizontal dashed line). The average of predictions is equal to this value $m = 1.5$ in both cases. Observations are black dots, the thin dotted blue line is the underlying function. In the right panel, one applies the assumption in Remark 4 with $\rho = 0$ and $\sigma_Z = \sigma/10$.

be different from the associated observation to satisfy the constraint.

The Panel 2b uses Proposition 4 to add a supplementary affine term to the previous linear predictor of Panel 2a, while preserving the sum of weights being equal to one. The affine term is derived from a random variable $\mathbf{Z}$, and we choose as previously $\sigma_Z = \sigma/10$, so that this external information is assumed trustable. As above, the average of predicted values is exactly the prescribed one, by construction. Again, given $\mathbf{Z} = \mathbf{m}$, the consequence is that, far from observed locations, the prediction tends to put all weights on this external source of information, so that the prediction gets closer to $\mathbf{m}$, as one can see at the extreme right of this Figure 2b. Another consequence of the affine term is the reduction of the confidence band width, as a new source of information has been added. With the prescribed average of predicted value, the predictor is not interpolating, but adding the affine term helps the prediction to get closer to observations.

In this simple toy example, one can check numerically that each prediction satisfies the constraints that it should. One can also clearly visualize the impact of the specific constraint on average predicted values, and the behavior of the predictor when adding an affine term.

The illustrations that have been presented in this subsection are available in the supplementary material notebook `Application1D`.

4.2 A multi-output time series example

In the previous example, we have illustrated the impact of constraints on the prediction of a one-dimensional output. Hence, the *joint* aspect of the estimation was not discussed. In the present example, one considers a multi-output data, so as to illustrate the specificity of the single hyperparameter estimation with multiple outputs. We choose one dimensional inputs in $\mathbb{R}$ to facilitate the interpolation representation, but considering more general inputs in $\mathbb{R}^d$, $d > 1$, would be easy. It would only change the number of hyperparameters to estimate, d instead of 1.

Imagine the following situation: a city wants to infer the history of some pollutants concentration at a particular crossroad based on a small series of measurements. This simple problem requires a model that takes time as input and multiple concentrations as output. Obviously, the end purpose would be to have a model with space and time as input but this is out of this illustration's framework.

Using the data *air quality* (see Vito, 2016), one tries to infer the concentration of several pollutants, from only few values. Studied pollutants in the data where chosen arbitrarily: CO, C6H6, NOx and NO2. The time range of learning data has been selected so that visually there is not too much missing data in the period (sensor stuck to an inferior bound or missing), it corresponds to hourly measurements from 23/04/2004 18.00.00 to 28/04/2004 17.00.00. Missing values are tagged with -200 values in this data, they have been all filtered before the study, as if they were not informative at all. The challenge is to predict all hourly measurements in the selected period from only $n = 10$ values.

The purpose here is not to give specific conclusions about the measured pollution, but only to illustrate the capacity of the Joint Kriging model to handle complex multi-valued data, with very few hyperparameters to optimize. The idea is to create a *joint* model that would be as simple as possible. Many refinements of the model could be suggested, but this is not the purpose of this example.

Let us model the covariances between components of $\mathbf{Y}(.)$ using Remark 2. The proposed method does not require the definition of each cross-correlation between a pollutant concentration at one location and a different pollutant concentration at a different location. It just takes one covariance function $k(x, x')$ between an implicitly weighted sum of all targets. We use the multiplication of two covariance kernels (hence it is positive semi-definite): a periodic kernel with period of one day, and a kernel of the Matérn 3/2 family (see Chapter 4 and Equation (4.31) in Rasmussen et al., 2006).

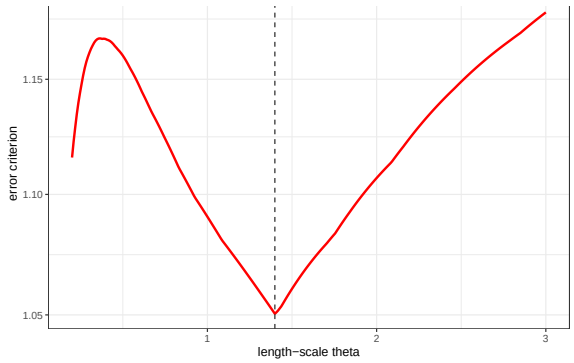


Figure 3: Optimization of the *single* correlation hyperparameter θ for the four selected pollutants, data extracted from Air quality data set.

$$k(x,x') = \sigma^2 \exp\left(-\sin^2(\pi|x-x'|)\right)\left(1+\frac{|x-x'|}{\theta}\right)\exp\left(-\frac{|x-x'|}{\theta}\right). \tag{27}$$

The parameterization has been simplified, e.g. factors $\sqrt{3}$ in Matérn covariance expressions are not used here: they have the same effect as a rescaling of the characteristic length-scale θ. Notice that despite the p dimensional output where $p = 4$ is the number of studied pollutants, the kernel $k(x,x')$ in Equation (27) depends only on two hyperparameters θ and σ^2. Since σ^2 impacts the uncertainty measurement but not the prediction itself, it is set to $\sigma^2 = 1$.

Let us consider first one single constraint: the sum of weights should be one. It corresponds to the Joint Ordinary Kriging predictor.

Figure 3 shows the optimization of the *single* length-scale hyperparameter θ. As this study does not aim at comparing the prediction accuracy with other methods, we did not use a separate test sample, but only a validation sample, keeping in mind that it may lead to an overfit. The validation data used for this single hyperparameter estimation is set to all hourly measurements in the selected period. For the hyperparameter optimization, a specific error has been chosen, where one optimizes the worst standardized mean absolute error over all $p = 4$ series: the errors have been standardized in order to make them unitless and scale invariant. The best estimation is $\hat{\theta} \simeq 1.4$, it is kept for all other illustrations of the subsection.

The optimization here depends quite heavily on the chosen observation locations, so that in practice, an averaged error on several training and validation data would probably be more stable: in many situations on real data, the error function is monotonic, either increasing and leading to extremely small optimized hyperparameter θ (the prediction then tends to return quickly to an average value), or either decreasing, leading to a

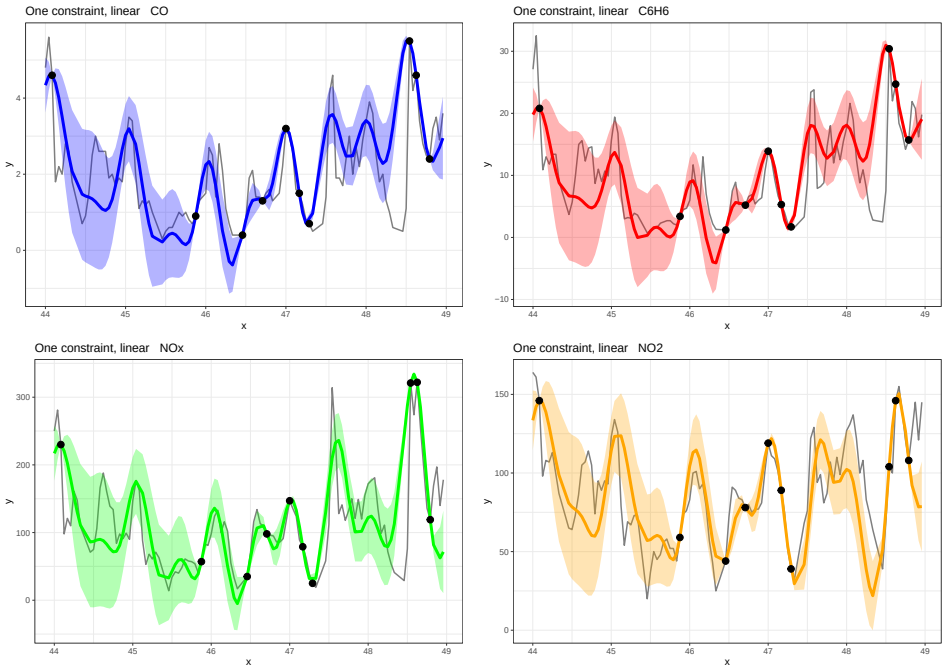


Figure 4: Joint Kriging interpolation: using the joint ordinary model with weights summing to one, with very few data points (black dots) and a single optimized length-scale hyperparameter obtained in Figure 3. Upper left: CO, upper right: C6H6, lower left: NOx, lower right: NO2. Predictions are in thick solid lines, true values are in thin black solid lines.

very large value of θ (the prediction then tends to smooth data a lot). Classical co-Kriging strategies that define a large number of cross-covariance hyperparameters would probably worsen the situation, highlighting the utility of a small number of hyperparameters.

Figure 4 presents the simultaneous predictions of the four pollutant concentrations with Joint Kriging, the only constraint being that weights sum to 1. Confidence band associated with a given pollutant is represented proportional to the standard deviation of this pollutant's concentration as detailed in Proposition 6. Pollutant concentrations have very different orders of magnitude but when applying Proposition 6, the obtained confidence bands looks quite comparable between series, as desired.

With very few hyperparameters and with a rough covariance model, the result has a lot of room for improvement. Nevertheless, despite the single model hyperparameter θ, and considering the limited number of observations $n = 10$, the predictions of the $p = 4$ concentrations seem quite reasonable. By construction each prediction is a combination of observed values of the considered pollutant, with weights summing to one. No other

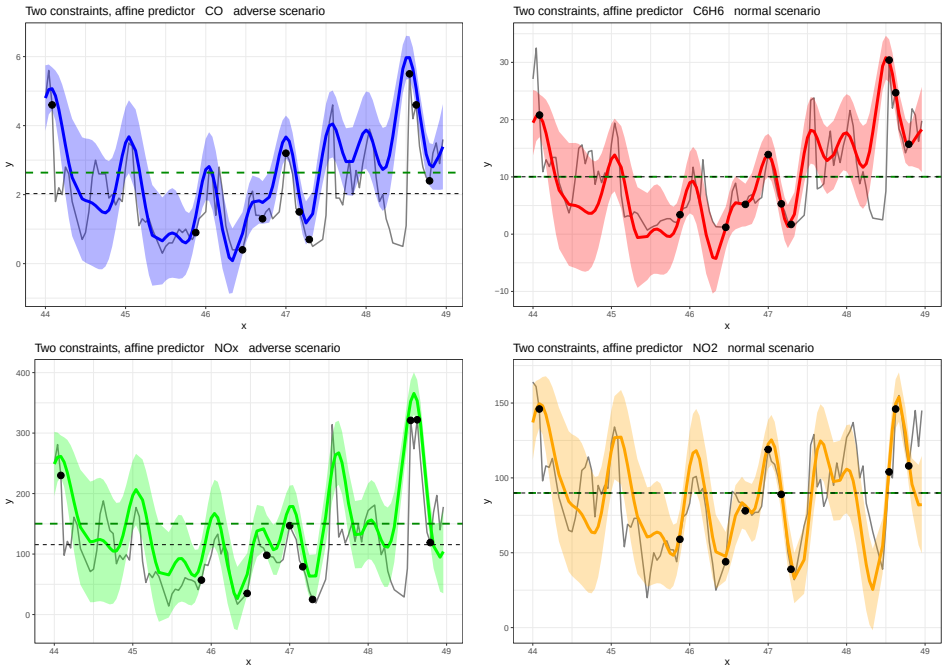


Figure 5: Adverse scenarios: interpolation using the joint affine model with two constraints (weights summing to one, prescribed average predictions), with very few data points (black dots) and a single optimized length-scale hyperparameter obtained in Figure 3. Upper left: CO, upper right: C6H6, lower left: NOx, lower right: NO2. Predictions are in thick solid lines, true values are in thin black solid lines. Left panels are adverse scenarios where the average of predictions (thick dark green horizontal dashed line) is set to 130% of the true average (thin black horizontal dashed line), right panels are scenarios where the average of predictions is set to 100% of the true average.

constraint is added on Figure 4, so that the average of predictions does not correspond at all to a specific prescribed value.

Figure 5 presents the simultaneous predictions of the four pollutant concentrations with Joint Kriging on which both constraints on the weights and on the predicted values are imposed using the affine model of Remark 4. Left panels show an adverse scenarios where the average of predictions is set to 130% of the true average, right panels are normal scenarios where the average of predictions is set to 100% of the true average. Using this setting, the interpolation property is lost, as seen in the previous example of Section 4.1, but the $n = 10$ observations still have a large influence, and the global shape of the prediction is preserved. By construction, the average of predicted values (thick solid line) is exactly the prescribed one (horizontal thick dashed dark green line).

Considering this single hyperparameter model, with a basic covariance model, the results also seem reasonable when using two constraints. In the left panels, the average of predicted values is exactly set to 30% more than the observed average of pollutant concentration, which is a lot. However the visual differences between true and predicted sequences looks surprisingly moderate, even in this adverse scenario: despite satisfying all constraints, the model still offers a good fit with observations.

The goal of this numerical experiment is to demonstrate the Joint Kriging model's ability to handle complex multi-valued data. It also illustrates the advantage of having a limited number of hyperparameters. One sees here that with a quite simple model, in a difficult problem (predicting four quite erratic time series from 10 observations), the model performs reasonably well. Furthermore, it allows for introducing some constraints, like setting an adverse scenario of 30% increase of the pollutant concentration.

The illustrations that have been presented in this subsection are available in the supplementary material notebook `ApplicationAirQuality`.

4.3 A constrained classification example

We present in this subsection the specific case of multi-dimensional outputs derived from a classification problem. As presented in section 3, Joint Kriging can be implemented for fuzzy classification. Different modalities of a classification variable are regarded as multiple output variables with values in $[0, 1]$.

Imagine the case of an event that may occur at a given location of a territory, with a measurable intensity. We are interested in classifying the intensity of this event, if it occurs, into multiple classes, depending on some thresholds. In the following, this event is an earthquake and its intensity is its Richter magnitude.

The Quake data set given in Simonoff (1996), visualized in Figure 6, describes 2 178 earthquakes with their latitude, longitude, focal depth and magnitude. A given location x has coordinates latitude, longitude and focal depth. For a single observation at location x, the target $\mathbf{Y}(x) = (Y_1(x), Y_2(x))^\top$ is equal to $(1, 0)^\top$ if an earthquake is occurring here with a magnitude above the data set average magnitude, or $(0, 1)^\top$ otherwise. If a location x is observed repeatedly, the membership degrees at x are averaged out over observations. It makes sense to impose that membership degrees are summing to one, so that ${\mathbf{1}_p}^\top \mathbf{Y}(x) = 1$. Extensions with more thresholds would be easy to conduct, see Figure 10, we keep here $p = 2$ for comparison to existing benchmarks. The binarized data is available at `www.openml.org/search?type=data&id=772`, on openML website (see Bischl et al., 2021).

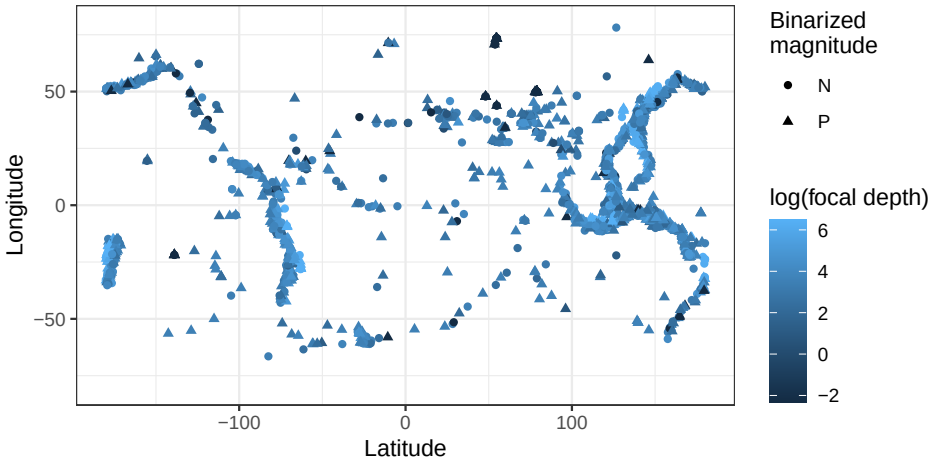


Figure 6: Earthquakes observations. An earthquake is a point with coordinates latitude, longitude and focal depth (given by the color). Triangles represent earthquakes which magnitude is above average. Circles represent earthquakes which magnitude is below average.

The purpose here is to compare the performance of Joint Kriging with a set of 69 other models' performances. The study available at `www.openml.org/search?type=task&id=4516` compares models, called flows in openML, performing 10 times a 10-fold cross-validation and computing the predictive accuracy as performance indicator (see tab `Analysis`, measure `predictive_accuracy`, the user must click on "fetch next 100 runs" several times in order to see all benchmarked models).

Remember from Remark 2 that although we are constructing a bivariate model, we need a single covariance kernel. The latter should be periodic with respect to the latitude and longitude, not with respect to focal depth. A simple way to define an admissible kernel is to multiply 3 kernels associated with the 3 dimensions (see Williams and Rasmussen, 2006):

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

The hyperparameters estimation has been treated separately on other train/test splits to avoid overfitting the data. The resulting values for $\boldsymbol{\theta}$ are 2.3 for latitude, 0.9 for longitude and 196.8 for focal depth.

In order to visualize the algorithm's behaviour, we predict on a grid of latitude, longitude and focal depth values. In addition to imposing the sum of membership degrees to be 1, we set the output mean expectation to be the same as in the data set. Predictions on a grid of latitude, longitude

and focal depth are presented on Figure 7. One can observe that maps representing membership degrees (first two rows) can be deduced from each other by $y = 1 - x$. The third row shows a segmentation of the plane into areas where membership degree for "P: magnitude is greater than average" is greater than 0.5, and areas where the converse is true. This segmentation depends on the focal depth: small focal depth on the top row (21km, first quartile) and greater one on the bottom row (68km, third quartile). For instance, looking at the bottom left corner of the map, which is around the Fiji archipelago, one can predict that earthquakes with small focal depth are more likely to be of large magnitude than deep earthquakes. However the converse is true in the south Atlantic area (bottom center part of the map). Moreover, the predictor achieves a circular coherence along longitude due to the periodicity of covariance. Periodicity along latitude is more difficult to observe because is covers only 180°.

Performances are evaluated using the Predictive Accuracy: it is the percentage of instances that are classified correctly. It is measured on binarized predicted membership degrees, on a 10 times 10-fold cross-validation, as in the OpenML benchmark, in order to get comparable results. Prior to that, the hyperparameters optimization has been treated separately on other train/test splits in order not to overfit the data.

Figure 8 presents from top to bottom: two results found in openML i.e. the best recorded model which is kernel logistic regression with RBFK kernel and Random Forest for reference, below are presented results of Joint Kriging models i.e. simple model without constraint, with weights summing to 1, with constraint on the prediction and weights summing to 1, affine with weights summing to 1 and affine with constrained output.

For the 10 runs, each diagram shows the Predictive Accuracy of each run (colored points), the minimum, first quartile, median, 3rd quartile and maximum, as well as the mean value materialized by a cross. Although the runs' performances stay in the range of those observed for Random Forest and Kernel Logistic Regression, the average values obtained with Joint Kriging are greater: the average is 0.556 ± 0.018 for the best model in OpenML benchmark, and 0.5661 ± 0.0038 for the best Joint Kriging model. The latter was even slightly greater, 0.5669, during hyperparameter optimisation, due to a slight overfit that has been reduced when using different train/test splits. Benchmark being based on this average value, it means that Joint Kriging has a better performance than the 69 models tested in the OpenML benchmark.

One can expect the multiplication of constraints to have an adverse effect on performance as a constrained optimization has less degree of freedom than an unconstrained one. On the other hand, injecting useful information through constraints may improve the performance. Figure 8 shows that

overall, the performance is improved, especially when adding the constraint on the output. The whisker plots show that the Joint Kriging performance is less dispersed than that of presented OpenML competitors. This is partly due to the fact that lengthscales have been optimized separately, so that the dispersion of their estimators is not taken into account: the variability rely on different train/test splitting. This prior optimisation slightly reduces the average measured performance of Joint Kriging models, due to different train/test split, but it also reduces its dispersion.

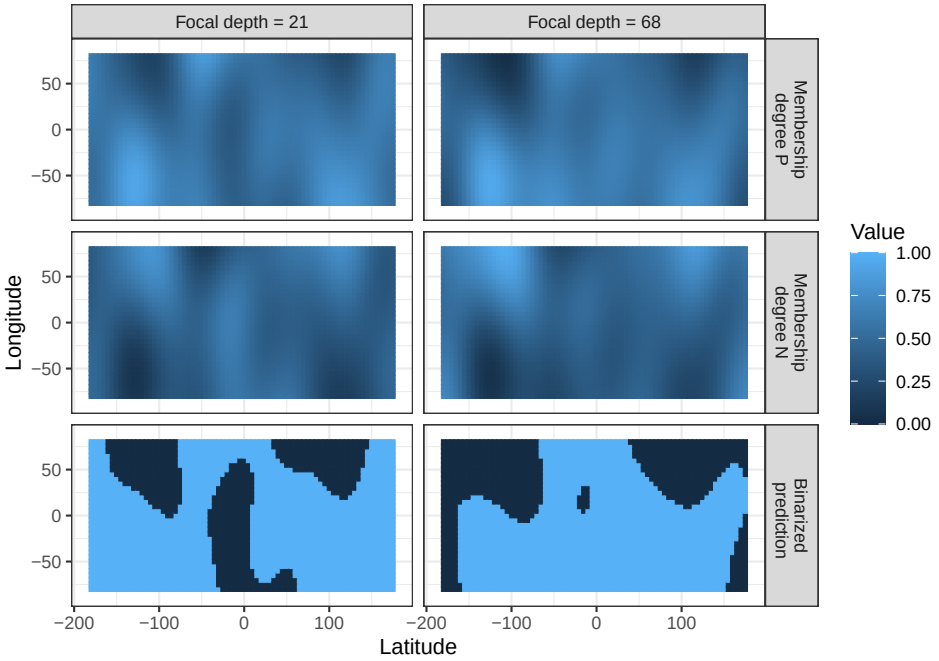


Figure 7: Joint Kriging with 2 constraints, earthquakes magnitude prediction into 2 classes. From top to bottom: membership degree of "P: magnitude is above average", membership degree of "N: magnitude is below average", binarized prediction (1 if membership degree of P is greater than 0.5). Left: focal depth of 21km. Right: focal depth of 68km.

In Figure 9, one uses the affine version of Joint Kriging with two constraints, weights summing to one and prescribed average prediction. In left panels, an adverse scenario forces the average predicted membership degrees of the first class (large magnitude events) to be equal to 65%. In the right panels, this percentage is set to the observed percentage of large magnitude events, 55%. This illustrates the usefulness of the constraint for adverse modeling.

In order to compare the results with existing benchmarks, we studied above the $p = 2$ binary classification problem. But the method can handle

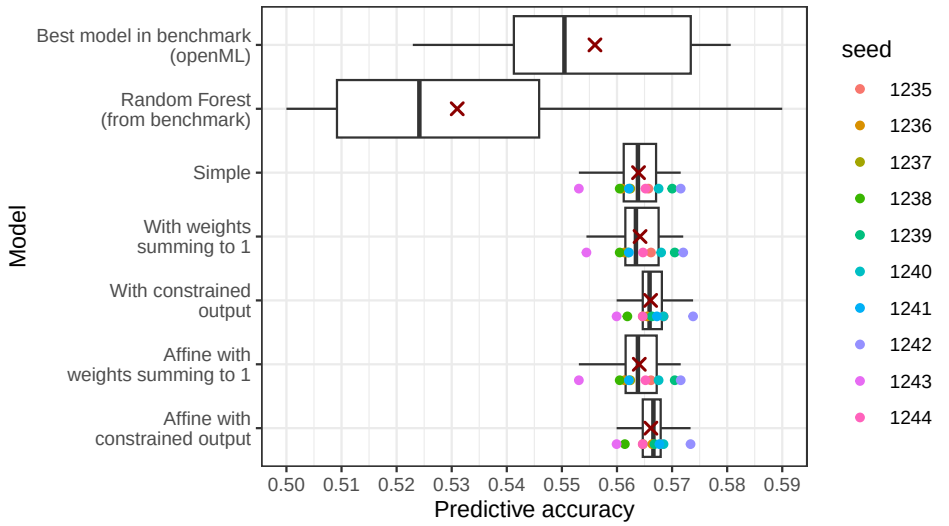


Figure 8: Distribution of performances for 10 runs of two OpenML models which are kernel logistic regression (best model in benchmark) and Random Forest, and the 5 different types of Joint Kriging model . The whisker plots give minimum, first quartile, median, third quartile and maximum. The dark red cross indicates the average predictive accuracy, the higher the better. Average is $0.566\,1 \pm 0.003\,8$ for the best Joint Kriging model, 0.556 ± 0.018 for the best model in OpenML benchmark.

more classes as well. As an example, in Figure 10, we give a prediction for $p = 4$ classes. Observations have been converted into four classes, using three Richter magnitude thresholds 5.85, 5.95, 6.15. Specific thresholds have been chosen for this illustration, in order to get enough observations in each classes (at least 17% observations), but a seismology study might focus on other thresholds. Once again, the predictor achieves a circular coherence along longitude, and one can observe complex patterns that would be difficult to catch with classification trees. The presented classification task was constructed from indicators deriving from an underlying real value, the Richter magnitude, and from thresholds, thus creating ordinal classes. But the prediction can be derived as well for observations of non ordinal class labels, without any underlying process or thresholds.

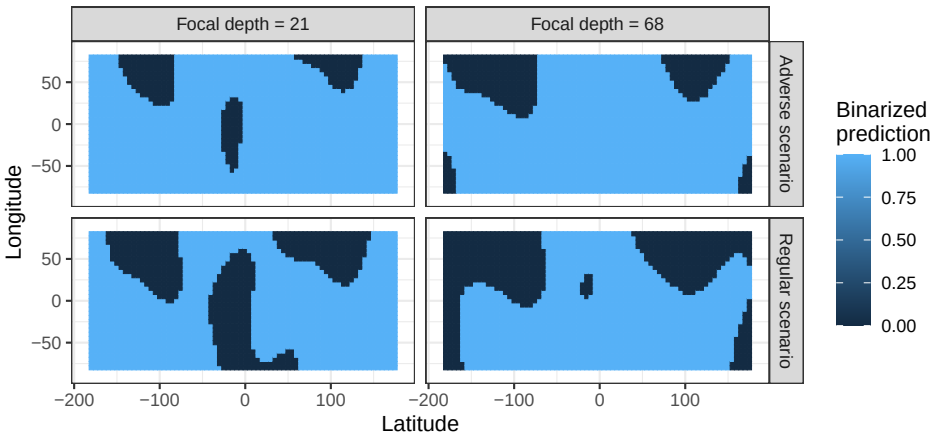


Figure 9: Adverse scenario: predicted membership degrees of earthquakes magnitude using Joint Kriging with two constraints. Top panels: adverse scenario, first class output average constrained to be 65%. Bottom panels: regular scenario, output average constrained to 55.5%. Left: focal depth of 21km. Right: focal depth of 68km.

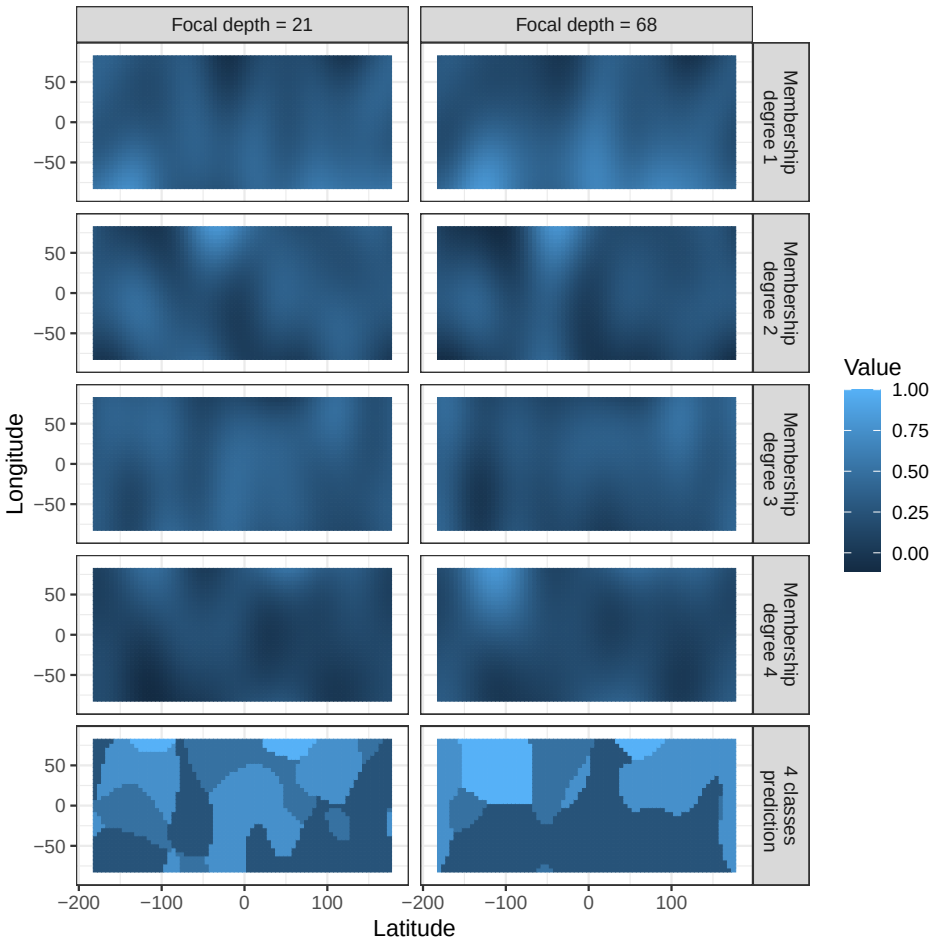


Figure 10: Affine Joint Kriging with 2 constraints, earthquakes magnitude prediction into 4 classes. From top to bottom: membership degrees of "1: magnitude is smaller than 5.85", "2: magnitude is between 5.85 and 5.95", "3: magnitude is between 5.95 and 6.15", "4: magnitude is greater than 6.15" and class of greatest membership degree in the 5th row colored by increasing magnitude from dark to light blue. From left to right: focal depth of 21km, focal depth of 68km.

In this classification example, we used a direct implementation of the model with a single covariance family. Nevertheless the average performance is the best one among the whole OpenML benchmark, fine tuning of the model would surely lead to performance improvements. The illustration above aims at demonstrating that with very basic assumptions, the method is competitive with an open benchmark that has numerous competitors, as shown in Figure 8. It also aims at showing that it can model adverse scenarios, as in Figure 9, or multiple classes, as in Figure 10.

The illustrations that have been presented in this subsection are available in the supplementary material `ApplicationClassificationQuake`.

5 Conclusion

A Joint Kriging model on multiple outputs has been presented, where at each prediction location, the same weights apply to all outputs. This simplification allows for an easy covariance modelling, with very few hyperparameters even though the number of outputs p is large. Still, the model benefits from Kriging advantages: interpretability, ability to interpolate data, uncertainty measurement associated to each prediction, specific covariance modelling. As any simplification, the model can surely be improved and may have some limitations compared to heavily parameterized models: e.g. coKriging with many cross-covariance functions might be more flexible for dealing with time series with different regularities, or models with parameterized distortions of locations might be convenient for dealing with non-stationarities. However, the limited number of hyperparameters and the simplicity of their estimation is an asset of the model, while allowing specific model characteristics as periodicity. Furthermore the model is not limited to Gaussian Processes as it only relies on the existence of moments of order one and two.

An original constraint on predicted values was also introduced. It appears to be useful for using external information, for adverse modelling, for homogenizing results or for considering fairness constraints. To handle this constraint, all weights of predicted points need to be computed at the same time, unlike usual Kriging techniques. But the resulting predictor itself is quite simple to derive since it is given by a closed formula. Some extensions using an affine term were also proposed, allowing to account for an external information, and providing more control on the behavior of the predictor far from observations.

Ultimately, an application to classification was developed. Applying a multi-outputs Kriging model on classification is feasible through the prediction of membership degrees. Even without constraints, it is in itself interesting: it allows for interpretability, modelling uncertainty estimation and interpolating data. Using Joint Kriging with the proposed constraints eas-

ily ensures that membership degrees sum to one, and allows for prescribed percentages of each predicted class. The simplified covariance model eases a lot the hyperparameters estimation. At the same time, with Joint Kriging, classification tasks benefit from the diversity of covariance kernels including periodicity. The resulting classification performs especially well in the investigated practical case: in the quake numerical example, the model competes with the best provided approaches on an open data set with numerous competitors.

Multiple extensions to the model can be imagined. For instance, the model with constrained predicted values does not guarantee continuous interpolation so that further work may fix this problem. A specific estimation procedure of the underlying joint covariance structure could also be of interest. Moreover, once applied to classification, membership degrees summing to one do not imply the combinations to be convex: some weights can still be negative or greater than 1. Ensuring the combinations to be convex could also be an improvement.

A Proofs

Proof of Proposition 1 Simple Joint Kriging weights. The proof is very similar to the geo-statistical proof of Simple Kriging model. It does not rely on any Gaussian assumption, but just on existing moments of order two. Recall that $\mathbb{W}$ is a symmetrical positive definite matrix, so that $\mathbb{W} = \mathbb{W}^\top$. Let us calculate the gradient of $\Delta(x^\star)$ with respect to $\boldsymbol{\alpha}(x^\star)$:

$$\begin{aligned}
& \nabla_{\boldsymbol{\alpha}(x^\star)} \mathrm{E}\left[\|\mathbf{M}(x^\star) - \mathbf{Y}(x^\star)\|^2_{\mathbb{W}}\right] \\
= \; & \nabla_{\boldsymbol{\alpha}(x^\star)} \mathrm{E}\left[(\mathbf{M}(x^\star) - \mathbf{Y}(x^\star))^\top \mathbb{W} (\mathbf{M}(x^\star) - \mathbf{Y}(x^\star))\right] \\
= \; & \nabla_{\boldsymbol{\alpha}(x^\star)} \mathrm{E}\left[(\mathbb{Y}\boldsymbol{\alpha}(x^\star) - \mathbf{Y}(x^\star))^\top \mathbb{W} (\mathbb{Y}\boldsymbol{\alpha}(x^\star) - \mathbf{Y}(x^\star))\right] \\
= \; & \nabla_{\boldsymbol{\alpha}(x^\star)} \mathrm{E}\left[\boldsymbol{\alpha}(x^\star)^\top \mathbb{Y}^\top \mathbb{W} \mathbb{Y} \boldsymbol{\alpha}(x^\star) - 2\boldsymbol{\alpha}(x^\star)^\top \mathbb{Y}^\top \mathbb{W} \mathbf{Y}(x^\star) + \mathbf{Y}(x^\star)^\top \mathbb{W} \mathbf{Y}(x^\star)\right] \\
= \; & 2\,\mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbb{Y}\right] \boldsymbol{\alpha}(x^\star) - 2\,\mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbf{Y}(x^\star)\right] .
\end{aligned}$$

Thus,

$$\nabla_{\boldsymbol{\alpha}(x^\star)} \Delta(x^*) = 2\mathbb{K}\boldsymbol{\alpha}(x^\star) - 2\mathbf{h}(x^\star) . \tag{28}$$

Where $\mathbb{K} := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbb{Y}\right]$ is a $n \times n$ matrix and $\mathbf{h}(x^\star) := \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbf{Y}(x^\star)\right]$ is a $n \times 1$ vector, thus leading to a $n \times 1$ gradient. Hence $\boldsymbol{\alpha}(x^\star) = \mathbb{K}^{-1}\mathbf{h}(x^\star)$ in Equation (4) when the gradient is zero. The matrix expression $\mathbb{A} = \mathbb{K}^{-1}\mathbb{H}$ of Equation (5) is obtained by binding column vectors of Equation (4), for all prediction locations. Remark that, under assumption that $\mathrm{E}[\mathbf{Y}(x)] = \mathbf{0}_p$ for all $x \in \mathcal{X}$, it is clear that $\mathrm{E}[\mathbf{M}(x^\star)] = \mathrm{E}[\mathbf{Y}(x^\star)] = 0$, so that the predictor is unbiased.

In that case, the (i, j) component of the matrix $\mathbb{Y}^\top \mathbb{Y}$ is

$$(\mathrm{E}\left[\mathbb{Y}^\top \mathbb{Y}\right])_{ij} = \sum_{k=1}^{n} \mathrm{E}\left[Y_i(x_k) Y_j(x_k)\right] = \sum_{k=1}^{n} \mathrm{Cov}\left[Y_i(x_k), Y_j(x_k)\right] .$$

Hence $\mathbb{Y}^\top \mathbb{Y}$ is a symmetric positive semi-definite matrix. The same holds for $\mathbb{K}$: writing $\mathbb{K} = \left(\mathbb{W}^{1/2}\mathbb{Y}\right)^\top \left(\mathbb{W}^{1/2}\mathbb{Y}\right)$, it is clear that for any vector $\mathbf{v}$, $\mathbf{v}^\top \mathbb{K} \mathbf{v} = \widetilde{\mathbf{v}}^\top \widetilde{\mathbf{v}} \geq 0$, where the vector $\widetilde{\mathbf{v}} := \mathbb{W}^{1/2}\mathbb{Y}\mathbf{v}$. Thus $\mathbb{K}$ is a symmetric semi-definite positive matrix, i.e. a covariance matrix. □

Proof of Proposition 2 Ordinary Joint Kriging weights. Under the constraint (6), and using a Lagrange multiplier $\lambda \in \mathbb{R}$, the loss to minimize is

$$\Delta_1(x^\star) := \Delta(x^\star) - 2\lambda(x^\star) \left(\boldsymbol{\alpha}(x^\star)^\top \mathbf{1}_n - 1\right)$$

Using Equation (28), the gradient of $\Delta_1(x^\star)$ with respect to $\boldsymbol{\alpha}(x^\star)$ is

$$\nabla_{\boldsymbol{\alpha}(x^\star)}\Delta_1(x^\star) = 2\,\mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbb{Y}\right]\boldsymbol{\alpha}(x^\star) - 2\,\mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbf{Y}(x^\star)\right] - 2\lambda(x^\star)\mathbf{1}_n \quad (29)$$

Setting this $\nabla_{\boldsymbol{\alpha}(x^\star)}\Delta_1(x^\star)$ to be zero for all of its p components, we get

$$\mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbb{Y}\right]\boldsymbol{\alpha}(x^\star) = \mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbf{Y}(x^\star)\right] + \lambda(x^\star)\mathbf{1}_n\,.$$

and finally,

$$\mathbb{K}\boldsymbol{\alpha}(x^\star) = \mathbf{h}(x^\star) + \lambda(x^\star)\mathbf{1}_n\,.$$

Once $\boldsymbol{\alpha}(x^\star)$ is written as a function of $\lambda(x^\star)$, one easily gets the value of $\lambda(x^\star)$ by setting ${\mathbf{1}_n}^\top\boldsymbol{\alpha}(x^\star) = 1$. Hence the result. Matrix expressions are obtained by binding column vectors for all $x^\star$ in $\{x_1^\star, \ldots, x_q^\star\}$ □

Proof of Remark 1 Covariance matrices. Recall that $\mathbb{K} = \mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbb{Y}\right]$ and $\mathbf{h}(x^\star) = \mathrm{E}\left[\mathbb{Y}^\top\mathbb{W}\mathbf{Y}(x^\star)\right]$. Under the chosen mean assumption, both $\mathrm{E}\left[\mathbf{Y}(x^\star)\right] = \boldsymbol{\mu}$ and $\mathrm{E}\left[\mathbb{Y}\right] = \boldsymbol{\mu}{\mathbf{1}_n}^\top$. Thus, under the given constraint ${\boldsymbol{\alpha}(x^\star)}^\top\mathbf{1}_n = 1$, or when $\boldsymbol{\mu} = \mathbf{0}_p$,

$$\mathrm{E}\left[\mathbb{Y}^\top\right]\mathbb{W}\,\mathrm{E}\left[\mathbb{Y}\right]\boldsymbol{\alpha}(x^\star) = \mathrm{E}\left[\mathbb{Y}^\top\right]\mathbb{W}\,\mathrm{E}\left[\mathbf{Y}(x^\star)\right] = \mathbf{1}_n\boldsymbol{\mu}^\top\mathbb{W}\boldsymbol{\mu}\,.$$

Hence the gradient of $\Delta(x^\star)$ in Equation (28) also writes

$$\nabla_{\boldsymbol{\alpha}(x^\star)}\Delta(x^*) = 2\mathbb{K}\boldsymbol{\alpha}(x^\star) - 2\mathbf{h}(x^\star) = 2\widetilde{\mathbb{K}}\boldsymbol{\alpha}(x^\star) - 2\widetilde{\mathbf{h}}(x^\star)\,.$$

As a consequence, the gradient of $\Delta_1(x^\star)$ in Equation (29) is unchanged when replacing both ($\mathbb{K}$, $\mathbf{h}$) by ($\widetilde{\mathbb{K}}$, $\widetilde{\mathbf{h}}$). Thus, one can freely replace both ($\mathbb{K}$, $\mathbf{h}$) by ($\widetilde{\mathbb{K}}$, $\widetilde{\mathbf{h}}$) in the rest of the proof of Proposition 2, without changing the result. □

Proof of Proposition 3 Joint Kriging weights under predicted values constraint. Under both constraints, the quantity to minimize is

$$\Delta_2(x^\star) := \Delta(x^\star) - 2\lambda(x^\star)\left({\boldsymbol{\alpha}(x^\star)}^\top\mathbf{1}_n - 1\right) - 2{\boldsymbol{\lambda}'}^\top\left(\mathbb{Y}\mathbb{A}\boldsymbol{\pi} - \mathbf{m}\right)\,,$$

where $\boldsymbol{\lambda}'$ is a $p \times 1$ vector of Lagrange multipliers. The gradient of the last term, with respect to $\boldsymbol{\alpha}(x^\star)$ is

$$\begin{aligned}
& \nabla_{\boldsymbol{\alpha}(x^\star)}2{\boldsymbol{\lambda}'}^\top\left(\mathrm{E}\left[\mathbf{M}(X^\star)\,|\,\mathbb{Y}\right] - \mathbf{m}\right)\\
= \;& \nabla_{\boldsymbol{\alpha}(x^\star)}2{\boldsymbol{\lambda}'}^\top\left(\mathrm{P}\left[X^\star = x^\star\right]\mathrm{E}\left[\mathbf{M}(x^\star)\,|\,\mathbb{Y}\right] + \mathrm{P}\left[X^\star \neq x^\star\right]\mathrm{E}\left[\mathbf{M}(X^\star)\,|\,X^\star \neq x^\star, \mathbb{Y}\right] - \mathbf{m}\right)\\
= \;& \nabla_{\boldsymbol{\alpha}(x^\star)}2{\boldsymbol{\lambda}'}^\top\left(\mathrm{P}\left[X^\star = x^\star\right]\mathrm{E}\left[\mathbf{M}(x^\star)\,|\,\mathbb{Y}\right] - \mathbf{m}\right) + 0\\
= \;& \nabla_{\boldsymbol{\alpha}(x^\star)}2{\boldsymbol{\lambda}'}^\top\left(\mathrm{P}\left[X^\star = x^\star\right]\mathrm{E}\left[\mathbb{Y}\boldsymbol{\alpha}(x^\star) - \mathbf{m}\right)\,|\,\mathbb{Y}\right]\\
= \;& \nabla_{\boldsymbol{\alpha}(x^\star)}2{\boldsymbol{\lambda}'}^\top\left(\pi_{x^\star}\mathbb{Y}\boldsymbol{\alpha}(x^\star) - \mathbf{m}\right)\\
= \;& 2\pi_{x^\star}\mathbb{Y}^\top\boldsymbol{\lambda}'
\end{aligned}$$

Hence using the gradient of $\Delta(x^\star)$ in Equation (28), one gets

$$\nabla_{\boldsymbol{\alpha}(x^\star)}\Delta_2(x^\star) = 2\mathbb{K}\boldsymbol{\alpha}(x^\star) - 2\mathbf{h}(x^\star) - 2\lambda(x^\star)\mathbf{1}_n - 2\pi_{x^\star}\mathbb{Y}^\top\boldsymbol{\lambda}' \tag{30}$$

Setting the gradient to be equal to a $n \times 1$ vector of zeros, we get for all prediction locations $x^\star \in \{x_1^\star, \ldots, x_q^\star\}$

$$\begin{cases} \mathbb{K}\boldsymbol{\alpha}(x^\star) = \mathbf{h}(x^\star) + \lambda(x^\star)\mathbf{1}_n + \pi_{x^\star}\mathbb{Y}^\top\boldsymbol{\lambda}' \\ {\mathbf{1}_n}^\top\boldsymbol{\alpha}(x^\star) = 1 \\ \mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m} \end{cases}$$

As optimal weights are gathered in the $n \times q$ matrix $\mathbb{A} := \big[\boldsymbol{\alpha}(x_1^\star), ..., \boldsymbol{\alpha}(x_q^\star)\big]$, if one defines the $n \times q$ matrix $\mathbb{H} := \big[\mathbf{h}(x_1^\star), ..., \mathbf{h}(x_q^\star)\big]$, then the previous system can be written, by binding columns for all $x^\star \in \{x_1^\star, \ldots, x_q^\star\}$:

$$\begin{cases} \mathbb{K}\mathbb{A} = \mathbb{H} + \mathbf{1}_n\boldsymbol{\lambda}^\top + \mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top \\ {\mathbf{1}_n}^\top\mathbb{A} = {\mathbf{1}_q}^\top \\ \mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m} \end{cases}$$

with $q \times 1$ Lagrange multiplier $\boldsymbol{\lambda}$, and $p \times 1$ Lagrange multiplier $\boldsymbol{\lambda}'$.

If $\mathbb{K}$ is invertible, then the first equation writes

$$\mathbb{A} = \mathbb{K}^{-1}\mathbb{H} + \mathbb{K}^{-1}\mathbf{1}_n\boldsymbol{\lambda}^\top + \mathbb{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top$$

Injecting this value of $\mathbb{A}$ into the first constraint ${\mathbf{1}_n}^\top\mathbb{A} = {\mathbf{1}_q}^\top$, denoting $\gamma := \boldsymbol{\pi}^\top\boldsymbol{\pi} \in \mathbb{R}$ and $\delta := {\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbf{1}_n \in \mathbb{R}$ one gets:

$$\begin{aligned} {\mathbf{1}_q}^\top &= {\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbb{H} + {\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbf{1}_n\boldsymbol{\lambda}^\top + {\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top \\ {\mathbf{1}_q}^\top\boldsymbol{\pi} &= {\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbb{H}\boldsymbol{\pi} + {\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbf{1}_n\boldsymbol{\lambda}^\top\boldsymbol{\pi} + {\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top\boldsymbol{\pi} \\ 1 &= {\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbb{H}\boldsymbol{\pi} + \delta\boldsymbol{\lambda}^\top\boldsymbol{\pi} + \gamma{\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}' \\ \delta\boldsymbol{\lambda}^\top\boldsymbol{\pi} &= 1 - {\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbb{H}\boldsymbol{\pi} - \gamma{\mathbf{1}_n}^\top\mathbb{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}' \end{aligned}$$

Now injecting the value of $\mathbb{A}$ into the second constraint $\mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m}$, and using the last equation, denoting the $p \times 1$ vector $\mathbf{u} := \mathbb{Y}\mathbb{K}^{-1}\mathbf{1}_n$, one gets

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

and finally, the vector $\boldsymbol{\lambda}'$ must satisfies

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

Hence, provided the matrix factor is invertible,

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

Once $\boldsymbol{\lambda}'$ computed, one gets for $\boldsymbol{\lambda}$

$$\begin{array}{rcl}\mathbf{1}_q{}^\top & = & \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbb{H} + \delta\boldsymbol{\lambda}^\top + \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top \\ \delta\boldsymbol{\lambda}^\top & = & -\mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbb{H} - \mathbf{1}_n{}^\top\mathbb{K}^{-1}\mathbb{Y}^\top\boldsymbol{\lambda}'\boldsymbol{\pi}^\top + \mathbf{1}_q{}^\top\end{array}$$

And finally, using $\mathbf{u} = \mathbb{Y}\mathbb{K}^{-1}\mathbf{1}_n$,

$$\boldsymbol{\lambda} = \delta^{-1}\left(\mathbf{1}_q - \mathbb{H}^\top\mathbb{K}^{-1}\mathbf{1}_n - \boldsymbol{\pi}\boldsymbol{\lambda}'{}^\top\mathbf{u}\right)$$

□

Proof of Remark 3 Covariance matrices with two constraints. The proof is similar to the one of Remark 1 and uses the fact that, under chosen assumptions and for any prediction point $x^\star$,

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

Hence the gradient of $\Delta_2(x^\star)$ in Equation (30) is unchanged when replacing $\mathbb{K}$ and $\mathbf{h}(x^\star)$ by $\widetilde{\mathbb{K}}$ and $\widetilde{\mathbf{h}}(x^\star)$, and all further expressions follows the same way in the proof of Proposition 3. □

Proof of Proposition 5 Joint Kriging variance with arbitrary weights. The first equation is a simple vector rewriting of Equation (1). For the prediction error, one simply write, whatever the weights $\boldsymbol{\alpha}(x^\star)$,

$$\begin{array}{rcl}\Delta(x^\star) & = & \mathrm{E}\left[\|\mathbf{M}(x^\star) - \mathbf{Y}(x^\star)\|^2_{\mathbb{W}}\right] \\ & = & \mathrm{E}\left[\left(\mathbf{M}(x^\star) - \mathbf{Y}(x^\star)\right)^\top\mathbb{W}\left(\mathbf{M}(x^\star) - \mathbf{Y}(x^\star)\right)\right] \\ & = & \mathrm{E}\left[\left(\mathbb{Y}\boldsymbol{\alpha}(x^\star) - \mathbf{Y}(x^\star)\right)^\top\mathbb{W}\left(\mathbb{Y}\boldsymbol{\alpha}(x^\star) - \mathbf{Y}(x^\star)\right)\right] \\ & = & \mathrm{E}\left[\boldsymbol{\alpha}(x^\star)^\top\mathbb{Y}^\top\mathbb{W}\mathbb{Y}\boldsymbol{\alpha}(x^\star) - 2\boldsymbol{\alpha}(x^\star)^\top\mathbb{Y}^\top\mathbb{W}\mathbf{Y}(x^\star) + \mathbf{Y}(x^\star)^\top\mathbb{W}\mathbf{Y}(x^\star)\right]\,.\end{array}$$

Hence the result. □

Proof of Remark 5 Covariance matrices in Joint Kriging mean and variance. The case where $\boldsymbol{\mu} = \mathbf{0}_p$ is straightforward, as in that case $\widetilde{\mathbb{K}} = \mathbb{K}$, $\widetilde{\mathbf{h}}(x^\star) = \mathbf{h}(x^\star)$ and $\widetilde{v}(x^\star) = v(x^\star)$, whatever the weights $\boldsymbol{\alpha}(x^\star)$. It remains the case where weights are summing to one. As in previous remarks, under chosen assumptions one gets

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

and moreover one can show that

$$-\alpha(x^\star)^\top \mathbf{h}(x^\star) + v(x^\star) = -\alpha(x^\star)^\top \widetilde{\mathbf{h}}(x^\star) + \widetilde{v}(x^\star)\,.$$

Hence the result. □

Proof of Proposition 6 Variance sharing. The difficulty here is to derive the cross-covariance $k_{ij}(x,x') = \mathrm{Cov}\,[Y_i(x), Y_j(x')]$ from the expression of $k(x,x')$ that is detailed in Remark 2

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

Denoting $\widetilde{\mathbf{Y}}(x) := \mathbb{W}^{1/2}\mathbf{Y}(x)$, $x \in \mathcal{X}$, this scalar covariance writes

$$k(x,x') = \mathrm{E}\left[\widetilde{\mathbf{Y}}(x)^\top \widetilde{\mathbf{Y}}(x')\right] - \mathrm{E}\left[\widetilde{\mathbf{Y}}(x)^\top\right] \mathrm{E}\left[\widetilde{\mathbf{Y}}(x')\right] \tag{31}$$

One would like to compute the $p \times p$ cross-covariance matrix between $\mathbf{Y}(x)$ and $\mathbf{Y}(x')$, using $\mathbf{Y}(x) = \mathbb{W}^{-1/2}\widetilde{\mathbf{Y}}(x)$, $x \in \mathcal{X}$:

$$\begin{aligned}
\mathbb{K}_Y(x,x') &:= \mathrm{E}\left[\mathbf{Y}(x)\mathbf{Y}(x')^\top\right] - \mathrm{E}\left[\mathbf{Y}(x)\right] \mathrm{E}\left[\mathbf{Y}(x')^\top\right] \\
&= \mathbb{W}^{-1/2}\left(\mathrm{E}\left[\widetilde{\mathbf{Y}}(x)\widetilde{\mathbf{Y}}(x')^\top\right] - \mathrm{E}\left[\widetilde{\mathbf{Y}}(x)\right] \mathrm{E}\left[\widetilde{\mathbf{Y}}(x')^\top\right]\right) {\mathbb{W}^{-1/2}}^\top \\
&= \mathbb{W}^{-1/2}\mathbb{K}_{\widetilde{Y}}(x,x'){\mathbb{W}^{-1/2}}^\top
\end{aligned} \tag{32}$$

where one defines $\mathbb{K}_{\widetilde{Y}}(x,x') := \mathrm{E}\left[\widetilde{\mathbf{Y}}(x)\widetilde{\mathbf{Y}}(x')^\top\right] - \mathrm{E}\left[\widetilde{\mathbf{Y}}(x)\right] \mathrm{E}\left[\widetilde{\mathbf{Y}}(x')^\top\right]$.

Now assume that:

$$\mathrm{Cov}\left[\widetilde{Y}_i(x), \widetilde{Y}_j(x')\right] = 0 \quad \text{whenever } i \neq j, x, x' \in \mathcal{X}\ .$$

This implies that $\mathbb{W}^{1/2}$ is proportional to a whitening transformation, so that all components of $\widetilde{Y}_1(x), \ldots, \widetilde{Y}_p(x)$ are uncorrelated.

Assume furthermore that:

$$\mathrm{Cov}\left[\widetilde{Y}_1(x), \widetilde{Y}_1(x')\right] = \ldots = \mathrm{Cov}\left[\widetilde{Y}_p(x), \widetilde{Y}_p(x')\right]\ , \quad x, x' \in \mathcal{X}\,.$$

Then one easily sees from Equation (31) that the scalar $k(x,x')$ satisfies

$$k(x,x') = \sum_{i=1}^{p} \mathrm{Cov}\left[\widetilde{Y}_i(x), \widetilde{Y}_i(x')\right] = p\,\mathrm{Cov}\left[\widetilde{Y}_j(x), \widetilde{Y}_j(x')\right], \quad j = 1,\dots,p$$

Hence under these assumptions, denoting $\mathbb{I}_p$ the $p \times p$ identity matrix,

$$\mathbb{K}_{\widetilde{Y}}(x,x') = \frac{1}{p} k(x,x') \mathbb{I}_p \,.$$

As a consequence, from Equation (32),

$$\mathbb{K}_Y(x,x') := \mathrm{E}\left[\mathbf{Y}(x)\mathbf{Y}(x')^\top\right] - \mathrm{E}\left[\mathbf{Y}(x)\right]\mathrm{E}\left[\mathbf{Y}(x')^\top\right] = \frac{1}{p} k(x,x') \mathbb{W}^{-1} \quad (33)$$

$$\mathrm{Cov}\left[Y_i(x), Y_j(x')\right] = \frac{1}{p} k(x,x') (\mathbb{W}^{-1})_{ij} \quad (34)$$

Now from this, one can derive the local cross errors

$$\delta_{ij}(x,x') := \mathrm{E}\left[\left(M_i(x) - Y_i(x)\right)\left(M_j(x') - Y_j(x')\right)\right]$$

Let us denote by $\mathbb{Y}_{i.}$ the ith row vector of the matrix $\mathbb{Y}$. We get

$$\begin{aligned}
\delta_{ij}(x,x') &= \mathrm{E}\left[\left(\mathbb{Y}_{i.}\boldsymbol{\alpha}(x) - Y_i(x)\right)\left(\mathbb{Y}_{j.}\boldsymbol{\alpha}(x') - Y_j(x')\right)\right] \\
&= \boldsymbol{\alpha}(x)^\top \mathrm{E}\left[\mathbb{Y}_{i.}{}^\top \mathbb{Y}_{j.}\right] \boldsymbol{\alpha}(x') - \boldsymbol{\alpha}(x)^\top \mathrm{E}\left[\mathbb{Y}_{i.}{}^\top Y_j(x')\right] \\
&\quad - \mathrm{E}\left[Y_i(x)^\top \mathbb{Y}_{j.}\right] \boldsymbol{\alpha}(x') + \mathrm{E}\left[Y_i(x) Y_j(x')\right]
\end{aligned}$$

Now assume $\mathrm{E}\left[\mathbf{Y}(x)\right] = \boldsymbol{\mu}$ for all $x \in \mathcal{X}$. Then from Equation (34),

$$\mathrm{E}\left[Y_i(x)^\top Y_j(x')\right] - \mu_i \mu_j = \frac{1}{p} k(x,x') (\mathbb{W}^{-1})_{ij}$$

which implies, using the matrix $\widetilde{\mathbb{K}}$ defined in Equation (10):

$$\mathrm{E}\left[\mathbb{Y}_{i.}{}^\top \mathbb{Y}_{j.}\right] - \mu_i \mu_j \mathbf{1}_n \mathbf{1}_n{}^\top = \frac{1}{p} (\mathbb{W}^{-1})_{ij} \widetilde{\mathbb{K}} \quad \text{and} \quad \mathrm{E}\left[\mathbb{Y}_{i.}{}^\top Y_j(x')\right] - \mu_i \mu_j \mathbf{1}_n = \frac{1}{p} (\mathbb{W}^{-1})_{ij} \widetilde{\mathbf{h}}(x') \,.$$

Furthermore, assume that either weights sum to one or $\boldsymbol{\mu} = \mathbf{0}_p$, then terms in $\mu_i \mu_j$ vanish and one gets:

$$\delta_{ij}(x,x') = \frac{1}{p} (\mathbb{W}^{-1})_{ij} \left(\boldsymbol{\alpha}(x)^\top \widetilde{\mathbb{K}} \boldsymbol{\alpha}(x') - \boldsymbol{\alpha}(x)^\top \widetilde{\mathbf{h}}(x') - \widetilde{\mathbf{h}}^\top(x) \boldsymbol{\alpha}(x') + k(x,x')\right) .$$

In particular from Proposition 5, using Remark 2 and Remark 5,

$$\delta_i(x^\star) = \frac{1}{p} (\mathbb{W}^{-1})_{ii} \Delta(x^\star) \,. \quad (35)$$

From Equation (33), when $k(x,x) = \sigma^2$ for all x, one can write

$$\mathbb{K}_Y(x,x) = \frac{1}{p}\sigma^2 \mathbb{W}^{-1}\,.$$

Using $\sigma_i^2 := \mathrm{Var}\,[Y_i(x)]$, assumed to be constant over x,

$$\frac{1}{p}(\mathbb{W}^{-1})_{ii} = \frac{(\mathbb{K}_Y(x,x))_{ii}}{\sigma^2} = \frac{\sigma_i^2}{\sigma^2}\,.$$

Hence from Equation (35),

$$\delta_i(x^\star) = \frac{\sigma_i^2}{\sigma^2}\Delta(x^\star)\,.$$

□

Proof of Remark 6 Constraints impact. The result is a very straightforward rewriting and interpretation of constraints (6) and (12). From $\mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{m}$ one derives $\mathbf{1}_p{}^\top \mathbf{m} = \mathbf{1}_p{}^\top \mathbb{Y}\mathbb{A}\boldsymbol{\pi} = \mathbf{1}_n{}^\top \mathbb{A}\boldsymbol{\pi} = \mathbf{1}_q{}^\top \boldsymbol{\pi} = 1$, hence the natural constraint on prescribed average membership degrees in $\mathbf{m}$, that must sum to one. □

B Notations

locations
$\mathcal{X}$ set of locations (inputs/design points).
n, q number of observed locations, of prediction locations.
x any location. $x_1, \ldots, x_n$ are all observed locations.
$x^\star$ any prediction location. $x_1^\star, \ldots, x_q^\star$ are all prediction locations.
$X^\star$ a random variable over prediction locations.
$\boldsymbol{\pi} = (\pi_{x_1^\star}, \ldots, \pi_{x_q^\star})$ the $q \times 1$ distribution of $X^\star$ over prediction locations.
$\gamma = \boldsymbol{\pi}^\top \boldsymbol{\pi}$ an intermediate real value used in calculations.

targets
p number of targets (i.e. number of outputs).
$\mathbf{Y}(x)$ the $p \times 1$ vector of targets at location x.
$\boldsymbol{\mu} = \mathrm{E}\left[\mathbf{Y}(x)\right]$ the $p \times 1$ mean of $\mathbf{Y}(x)$, when constant over x.
$\mathbb{Y} = \left[\mathbf{Y}(x_1), \ldots, \mathbf{Y}(x_n)\right]$ all the $p \times n$ values of observed targets.
$\mathbb{Y}^\star = \left[\mathbf{Y}(x_1^\star), \ldots, \mathbf{Y}(x_q^\star)\right]$ all $p \times q$ unknown targets at prediction locations.

prediction
$\mathbf{M}(x^\star)$ a $p \times 1$ predictor of $\mathbf{Y}(x)$
$\mathbb{M} = \left[\mathbf{M}(x_1^\star), \ldots, \mathbf{M}(x_q^\star)\right]$ the $p \times q$ matrix of all predictions.
$\boldsymbol{\alpha}(x^\star)$ the $n \times 1$ linear weights for the prediction in $x^\star$.
$\mathbb{A} = \left[(\boldsymbol{\alpha}(x_1^\star), \ldots, \boldsymbol{\alpha}(x_q^\star)\right]$ the $n \times q$ matrix of weights for all predictions.
$\mathbf{m}$ a given constant $p \times 1$ vector of prescribed mean predicted values.
$\Delta(x^\star), \Delta_1(x^\star), \Delta_2(x^\star)$ losses to be minimized for finding $\mathbf{M}(x^\star)$.
$\boldsymbol{\lambda}$ 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
$\mathbb{W}$ a given symmetric positive definite matrix for computing norms.
$\mathbf{h}(x^\star) = \mathrm{E}\left[\mathbb{Y}^\top \mathbb{W} \mathbf{Y}(x^\star)\right]$ a $n \times 1$ covariance vector.
$\mathbb{H} = (\mathbf{h}(x_1^\star, \ldots, \mathbf{h}(x_q^\star)))$ 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.
$\widetilde{\mathbb{K}}, \widetilde{\mathbf{h}}(x^\star), \widetilde{\mathbb{H}}$ 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 $\mathbf{Y}(x_i)$
$\mathbf{Q}$ additional $q \times 1$ covariance vector between $\mathbf{Z}$ and $\mathbf{Y}(x_j^\star)$

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.

References

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

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.

Banerjee, A., Dunson, D. B., and Tokdar, S. T. (2013). Efficient gaussian process regression for large datasets. *Biometrika*, 100(1):75–89.

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.

Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., Mantovani, R. G., van Rijn, J. N., and Vanschoren, J. (2021). OpenML benchmarking suites. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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

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.

Cressie, N. (1988). Spatial prediction and ordinary kriging. *Mathematical geology*, 20:405–421.

Cressie, N. and Johannesson, G. (2008). Fixed Rank Kriging for Very Large Spatial Data Sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):209–226.

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

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

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.

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

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.

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

Grossouvre, M. and Rullière, D. (2023). Supplementary material to: A Joint Kriging Model with Application to Constrained Classification. `https://gitlab.emse.fr/marc.grossouvre/jointkrigingsupplementary/`.

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.

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

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.

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

Rasmussen, C. and Ghahramani, Z. (2000). Occam's razor. *Advances in neural information processing systems*, 13.

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

Rullière, D., Durrande, N., Bachoc, F., and Chevalier, C. (2018). Nested kriging predictions for datasets with a large number of observations. *Statistics and Computing*, 28:849–867.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag.

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

Vito, S. (2016). Air Quality. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C59K5F.

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

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2(3). MIT press Cambridge, MA.

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.

Contents