A criterion and incremental design construction for simultaneous kriging predictions

Helmut Waldl [a],[*], Werner G. Müller [a], Paula Camelia Trandafir [b]

[a] Department of Applied Statistics, Johannes-Kepler-University Linz, Altenberger Straße 69, Linz, A-4040, Austria
[b] Department of Statistics, Informatics and Mathematics, Public University of Navarre, Campus Arrosadia, Pamplona, 31006, Spain

A R T I C L E   I N F O

A B S T R A C T

In this paper, we further investigate the problem of selecting a set of design points for universal kriging, which is a widely used technique for spatial data analysis. Our goal is to select the design points in order to make simultaneous predictions of the random variable of interest at a finite number of unsampled locations with maximum precision. Specifically, we consider as response a correlated random field given by a linear model with an unknown parameter vector and a spatial error correlation structure. We propose a new design criterion that aims at simultaneously minimizing the variation of the prediction errors at various points. We also present various efficient techniques for incrementally building designs for that criterion scaling well for high dimensions. Thus the method is particularly suitable for big data applications in areas of spatial data analysis such as mining, hydrogeology, natural resource monitoring, and environmental sciences or equivalently for any computer simulation experiments. We have demonstrated the effectiveness of the proposed designs through two illustrative examples: one by simulation and another based on real data from Upper Austria.

1. Introduction

Given a finite number $k + m$ of locations x we are interested into making simultaneous predictions $\hat{Y}(\cdot)$ of $Y(\cdot)$ at m unsampled locations using observations $Y(x_1), \ldots, Y(x_k)$ collected at some design points $\xi = (x_1, \ldots, x_k) \subset \mathcal{X}^k$. Our objective is to select ξ (of given size k) in order to maximize the precision of the predictions $\hat{Y}(x)$ over $\mathcal{X}$. This setup is used in such diverse areas of spatial data analysis as mining, hydrogeology, natural resource monitoring and environmental sciences, see, *e.g.*, Cressie (1993), and has become the standard modeling paradigm in computer simulation experiments (cf. Fang et al. (2005), Kleijnen (2009), Rasmussen and Williams (2005), Santner et al. (2003)), known under the designations of Gaussian Process (GP) modeling and kriging analysis. For a general review in the context of spatial statistics see Wang et al. (2012).

Specifically, the model underlying our investigations is the model for universal kriging, i.e. we have a correlated scalar random field given by

$$Y(x) = \mu(x, \boldsymbol{\beta}) + \varepsilon(x) \tag{1}$$

Here, $\boldsymbol{\beta}$ is an unknown vector of parameters in $\mathbb{R}^p$, $\mu(\cdot, \cdot)$ a known function of regressors at some given locations x in a compact subset $\mathcal{X}$ of $\mathbb{R}^d$ and the random term $\varepsilon(x)$ has zero mean, variance σ^2 and a parameterized spatial error correlation structure such that $\mathbb{E}(\varepsilon(x)\varepsilon(x')) = \sigma^2 c(x, x'; \nu)$ with ν some covariance parameters. We further assume that

* Corresponding author.
E-mail addresses: helmut.waldl@jku.at (H. Waldl), Werner.Mueller@jku.at (W.G. Müller), camelia@unavarra.es (P.C. Trandafir).

- the deterministic term is linear in the parameters $\boldsymbol{\beta}$, *i.e.*, $\mu(x, \boldsymbol{\beta}) = \mathbf{f}(x)\boldsymbol{\beta}$, where $\mathbf{f}(x) = \begin{pmatrix} f_1(x) & f_2(x) & \cdots & f_p(x) \end{pmatrix}$ is a vector of known functions,
- the first two moments of the error $\varepsilon(x)$ and hence of $Y(x)$ exist and
- the variance σ^2 and the covariance parameters ν are known.

It is often assumed that the random field $\varepsilon(x)$ is Gaussian, allowing estimation of β and $\theta = \{\sigma^2, \nu\}$ by Maximum Likelihood. We do not need to assume a stationary nor an isotropic covariance structure.

Traditional optimality criteria for designs for prediction are kriging variance-minimizing, albeit not all authors understand the same by the term *kriging variance* or *kriging covariance*. Some mean the variance of the prediction $\text{Var}\left(\hat{Y}(x)\right)$, cf. Müller et al. (2015), some the variance of the prediction error $\text{Var}\left(\hat{Y}(x) - Y(x)\right)$, cf. Cressie (1993), and some even $\text{Var}\left(Y(x)|Y(\xi)\right)$, i.e. the variance of $Y(\cdot)$ at unsampled locations x given the observations on the design points ξ, cf. Chevalier and Ginsbourger (2012). We follow the second perception because trying to minimize the variation of the prediction errors seems to yield most precise predictions.

Definition 1. Let $x \in \mathcal{X}$ be an arbitrary unsampled location and $\hat{Y}(x)$ the best linear unbiased predictor (BLUP) at x. The *kriging variance* at x is

$$\sigma^2(x) := \text{Var}\left(\hat{Y}(x) - Y(x)\right)$$

i.e. the variance of the best linear predictor minus the random variable to be predicted.

Let $x' \neq x$ be a second unsampled location and $\hat{Y}(x')$ the BLUP at $x' \in \mathcal{X}$. The *kriging covariance* for x and x' is

$$\sigma(x, x') := \mathbb{C}\text{ ov}\left(\hat{Y}(x) - Y(x); \hat{Y}(x') - Y(x')\right)$$

With the above definition G-optimal designs (cf. Dasgupta et al. (2022-03-12)) try to minimize the maximum kriging variance, *i.e.*

$$\min_{\xi} \max_{x \in \mathcal{X}} \text{Var}\left(\hat{Y}(x) - Y(x)\right) . \tag{2}$$

Another popular optimality criterion tries to minimize the average prediction variance over a set of m specific points (V-optimality), *i.e.*

$$\min_{\xi} \frac{1}{m} \sum_{x_i \in \mathcal{X}} \text{Var}\left(\hat{Y}(x_i) - Y(x_i)\right) . \tag{3}$$

Note that the latter if not supported on a grid but rather covering the whole $\mathcal{X}$, expressed as an integral is often called I-optimality, see Dasgupta et al. (2022) for a recent example of the terminology.

None of these and other criteria for designs for prediction considers the *kriging covariances* or the *kriging covariance matrix* at the unsampled locations $(x_{k+1}, \ldots, x_{k+m})$

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(x_{k+1}, \ldots, x_{k+m}) = \begin{pmatrix} \sigma^2(x_{k+1}) & \sigma(x_{k+1}, x_{k+2}) & \cdots & \sigma(x_{k+1}, x_{k+m}) \\ \sigma(x_{k+1}, x_{k+2}) & \sigma^2(x_{k+2}) & \cdots & \sigma(x_{k+2}, x_{k+m}) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(x_{k+1}, x_{k+m}) & \sigma(x_{k+2}, x_{k+m}) & \cdots & \sigma^2(x_{k+m}) \end{pmatrix}$$

The popular design criteria just use the diagonal of $\boldsymbol{\Sigma}$ which means to voluntarily dispense of valuable information given by the kriging covariances.

The design criterion considered in this paper, is the *generalized variance* of the kriging covariance matrix:

$$GV(\xi) = \det \boldsymbol{\Sigma}(x_{k+1}, \ldots, x_{k+m}) = |\boldsymbol{\Sigma}| . \tag{4}$$

Definition 2. The design ξ that minimizes criterion (4) or equivalently any root of $|\boldsymbol{\Sigma}|$ is defined as the GV-optimal design, GV stands for Generalized (kriging) Variance.

As designs that simply minimize the kriging variance, GV-optimal designs are often space-filling but typically position more design points at the edge of the design region. GV-optimal designs depend more on the special covariance structure which is in contrast to G- and V-optimal designs in particular for small numbers of observations k (see Fig. 1 for an exemplary comparison). Unfortunately the minimization of the GV-criterion is computationally demanding, since the evaluation of (4) requires the evaluation of the determinant of an $(m \times m)$-matrix, being unfeasible for large m which is especially necessary in high dimensional design spaces as it is often the case for computer experiments. A remedy for this problem is provided by the use of incrementally assembled designs proposed here which turn out to be GV-optimal. This reduces the computational effort for the evaluation of (4) for arbitrary m to the evaluation of the determinant of a $(k \times k)$-matrix, where k is just the design size.

The paper is organized as follows. In Section 2 we motivate our approach, exploiting the fact that the volume of the simultaneous confidence region for the prediction errors is proportional to $\sqrt{|\boldsymbol{\Sigma}|}$. Section 3 presents the basis of the main contribution of the paper, giving an update formula for the determinant of the kriging covariance matrix which should be used if we are constructing our designs incrementally. These update formulas are also necessary to speed up the computation of the design criterion (4) and may as well be used for a more efficient computation of the design criterion V-optimal designs. Finally, Section 4 considers the efficiency of incremental GV-optimal designs and also the efficiency of GV-optimal designs with respect to G- and V-optimal designs which is demonstrated by means of a representative simulation study in Section 5, followed by a real world example in Section 6, and Section 7 draws conclusions on the efficiency and limitations of the approach and suggests topics for future work.
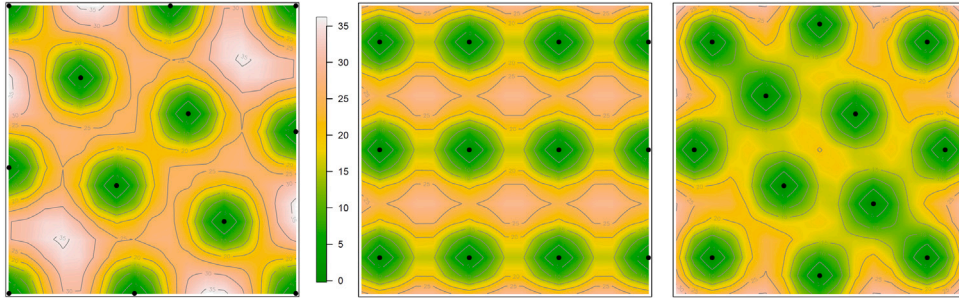
Fig. 1. Map of kriging variances of GV-optimal (left), G-optimal (middle) and V-optimal (right) 12-point design for the ordinary kriging setup on the unit square and Matern covariance function with $\kappa = 2.5$ and $\phi = 3$. The black dots are the design points. Note that in some zones of the design region the GV-optimal design has considerable larger kriging variances than the G- and V-optimal designs. This is a natural consequence of the GV-criterion, which does not aim at minimizing the point-wise kriging variances while the G- and V-criterion do.

2. The generalized variance and optimal designs for prediction

Wilks (1932) remarked that there were "...statistical coefficients which have not been adequately generalized for samples from a multivariate population including the variance ..." and introduced the *generalized variance* which is simply the determinant of the covariance matrix of a multivariate population. Wald (1943) used the idea of minimizing the generalized variance in his criterion for optimal designs for parameter estimation (D-optimality), *i.e.*

$$\max_{\xi} |M_{\theta}(\xi)|,$$

where M_{θ} is the information matrix of the parameter estimates $\hat{\theta}$. A combination of these two ideas naturally leads to criterion (4).

Shewry and Wynn (1987) introduced maximum entropy sampling (MES) where the Shannon entropy is used as a measure of information to get optimal designs for prediction. Here the design criterion is similar to (4): in the case of Gaussian response Y, MES tries to maximize the determinant of the covariance matrix of $Y(\xi)$ which is equivalent to minimizing the determinant of the covariance matrix of $(Y(x_{k+1}), \ldots, Y(x_{k+m})|Y(x_1), \ldots, Y(x_k))$, *i.e.* the determinant of the covariance matrix of $Y(\cdot)$ at the unsampled locations $(x_{k+1}, \ldots, x_{k+m})$ conditioned on $Y(\cdot)$ at sampled locations $(x_1, \ldots, x_k)$. Though this criterion does not aim at minimizing the variation of the prediction errors, it is not even directly connected with a prediction method. MES it is rather a sampling method trying to absorb the maximum amount of variability into the sample, such that conditional on the sample the unsampled points have minimum variability. The method is suitable for observations on a finite closed system which is the main connection to the present work. Beyond that we connect the sampling method directly with the BLUP for the response Y as presented hereinafter.

Interestingly MES yields designs that are exactly GV-optimal in the simple kriging setup and also seem to be GV-optimal in the ordinary kriging setup. For universal kriging GV-optimal designs approach MES designs with increasing effective range, *i.e.* more strongly correlated response.

2.1. The best linear unbiased predictor and the corresponding kriging covariance matrix

As with MES we assume that the allowable choice of the designs ξ is from a fixed finite set of $N = k+m$ points $X = \{x_1, \ldots, x_{k+m}\}$. Given a k-point design $\xi = (x_1, \ldots, x_k)$, $0 < k < N$, the complementary design ξ_0 is the set $X \setminus \xi$ and we get the corresponding partitioning of the response vector $\mathbf{Y} = \left(\mathbf{Y}_{\xi}^T \ \mathbf{Y}_0^T \right)^T$. We now want to simultaneously predict the response from the complementary design on the basis of the response from the design ξ.

As mentioned above we are using the model of universal kriging, i.e. our linear predictor is the linear combination of a vector of deterministic functions of the locations $x \in \xi$: $\mathbf{f}(x) = (f_1(x) \ f_2(x) \ \cdots \ f_p(x))$ with the first component usually being $f_1(x) = 1$. The design matrix and corresponding vector of errors then are

$$\mathbf{F}_{\xi} = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \ldots & f_p(x_1) \\ f_1(x_2) & f_2(x_2) & \ldots & f_p(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ f_1(x_k) & f_2(x_k) & \ldots & f_p(x_k) \end{pmatrix} \qquad \boldsymbol{\varepsilon}_{\xi} = \begin{pmatrix} \varepsilon(x_1) \\ \varepsilon(x_2) \\ \vdots \\ \varepsilon(x_k) \end{pmatrix}$$

We denote the covariance matrix of $\boldsymbol{\varepsilon}_{\xi}$ with $\mathbf{C}_{\xi}$ and use analogous nomenclature for the complementary design to get

$$\mathbb{E}\begin{pmatrix} \mathbf{Y}_{\xi} \\ \mathbf{Y}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{\xi} \\ \mathbf{F}_0 \end{pmatrix} \boldsymbol{\beta} \qquad \mathbb{C}\text{ov}\begin{pmatrix} \mathbf{Y}_{\xi} \\ \mathbf{Y}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{\xi} & \mathbf{C}_{\xi 0} \\ \mathbf{C}_{\xi 0}^T & \mathbf{C}_0 \end{pmatrix} \qquad \Rightarrow$$

$$\Rightarrow \qquad \mathbb{E}\left(\mathbf{Y}_0 | \mathbf{Y}_{\xi} \right) = \mathbf{F}_0 \boldsymbol{\beta} + \mathbf{C}_{\xi 0}^T \mathbf{C}_{\xi}^{-1} \left(\mathbf{Y}_{\xi} - \mathbf{F}_{\xi} \boldsymbol{\beta} \right) \qquad \mathbb{C}\text{ov}\left(\mathbf{Y}_0 | \mathbf{Y}_{\xi} \right) = \mathbf{C}_0 - \mathbf{C}_{\xi 0}^T \mathbf{C}_{\xi}^{-1} \mathbf{C}_{\xi 0}$$

The generalized least squares (GLS) estimate of the parameter vector then is $\hat{\beta} = \left(\mathbf{F}_\xi^T \mathbf{C}_\xi^{-1} \mathbf{F}_\xi \right)^{-1} \mathbf{F}_\xi^T \mathbf{C}_\xi^{-1} \mathbf{Y}_\xi$ yielding the simultaneous kriging prediction for all complementary design points as the BLUP

$$\hat{\mathbf{Y}}_0 = \mathbf{F}_0 \hat{\beta} + \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \left(\mathbf{Y}_\xi - \mathbf{F}_\xi \hat{\beta} \right) = \left(\mathbf{F}_0 \mathbf{B} + \mathbf{C}_{\xi 0}^T \mathbf{A} \right) \mathbf{Y}_\xi = \mathbf{W} \mathbf{Y}_\xi$$

where $\mathbf{B} = \left(\mathbf{F}_\xi^T \mathbf{C}_\xi^{-1} \mathbf{F}_\xi \right)^{-1} \mathbf{F}_\xi^T \mathbf{C}_\xi^{-1}$ and $\mathbf{A} = \mathbf{C}_\xi^{-1} \left(\mathbf{I}_k - \mathbf{F}_\xi \mathbf{B} \right)$. The components of the $(m \times k)$–*weight matrix* $\mathbf{W} = \left(\omega_{ij} \right)$ give the kriging weights of $Y(x_j)$ for the prediction $\hat{Y}(x_{k+i})$. Combining the results from above the kriging prediction errors have expectation zero and the *kriging covariance matrix*

$$\mathbb{C} \operatorname{ov} \left(\hat{\mathbf{Y}}_0 - \mathbf{Y}_0 \right) = \boldsymbol{\Sigma} = \mathbf{W} \mathbf{C}_\xi \mathbf{W}^T - \mathbf{W} \mathbf{C}_{\xi 0} - \mathbf{C}_{\xi 0}^T \mathbf{W}^T + \mathbf{C}_0 \tag{5}$$

2.2. Geometrical interpretation of the GV criterion

In the case of Gaussian kriging prediction errors,

$$\left(\hat{\mathbf{Y}}_0 - \mathbf{Y}_0 \right) \sim \mathbb{N} \left(\mathbf{0} \, ; \, \boldsymbol{\Sigma} \right) \qquad \Rightarrow \qquad \left(\hat{\mathbf{Y}}_0 - \mathbf{Y}_0 \right)^T \boldsymbol{\Sigma}^{-1} \left(\hat{\mathbf{Y}}_0 - \mathbf{Y}_0 \right) \sim \chi_m^2$$

we get a simultaneous confidence region for the prediction errors with

$$\left\{ \hat{\mathbf{Y}}_0 - \mathbf{Y}_0 \, : \, \left(\hat{\mathbf{Y}}_0 - \mathbf{Y}_0 \right)^T \boldsymbol{\Sigma}^{-1} \left(\hat{\mathbf{Y}}_0 - \mathbf{Y}_0 \right) \le \chi_{m;1-\alpha}^2 \right\}$$

where $\chi_{m;1-\alpha}^2$ is the $(1 - \alpha)$-quantile of the χ_m^2 distribution.

This region is a k-dimensional ellipsoid who's volume is proportional to $\sqrt{|\boldsymbol{\Sigma}|}$, *i.e.* GV-optimal designs minimize the area of the prediction errors. It is evident that the usually used design criteria like G-optimality or V-optimality do not have this highly desirable property.

2.3. Comparison of MES and the GV-criterion

The criterion for a GV-optimal design or a GV-optimal increment (13) looks similar to the criterion of MES with multivariate normal distributed $\mathbf{Y}$, which in our notation would be to maximize $\left| \mathbb{C} \operatorname{ov} \left(\mathbf{Y}_\xi \right) \right| = \left| \mathbf{C}_\xi \right|$, which is equivalent to minimizing $\left| \mathbb{C} \operatorname{ov} \left(\mathbf{Y}_0 | \mathbf{Y}_\xi \right) \right| = \left| \mathbf{C}_0 - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \mathbf{C}_{\xi 0} \right|$. The reason for the equivalence lies in the determinant identity $\left| \begin{pmatrix} \mathbf{C}_\xi & \mathbf{C}_{\xi 0} \\ \mathbf{C}_{\xi 0}^T & \mathbf{C}_0 \end{pmatrix} \right| = \left| \mathbf{C}_\xi \right| \cdot \left| \mathbf{C}_0 - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \mathbf{C}_{\xi 0} \right|$ and the fact that the left hand side is fixed and finite (see Shewry and Wynn (1987)).

However, there are considerable differences between the GV-criterion and the MES-criterion:

- The above MES-criterion just holds in the case of Gaussian $\mathbf{Y}$. We do not assume multivariate normal distributed $\mathbf{Y}$ for our GV-criterion, *i.e.* the two criteria are different in the case of non-Gaussian $\mathbf{Y}$.
- With GV-optimal designs we minimize $\mathbb{C} \operatorname{ov} \left(\hat{\mathbf{Y}}_0 - \mathbf{Y}_0 \right)$ in contrast to $\mathbb{C} \operatorname{ov} \left(\mathbf{Y}_0 | \mathbf{Y}_\xi \right)$ which is minimized with MES-optimal designs. This difference just vanishes in the case of simple kriging which is shown later.
- The implementation of MES for a linear model (1) uses a Bayesian model which needs some prior distribution for the parameter vector β and particularly some prior covariance matrix $\mathbb{C} \operatorname{ov} (\beta) = \mathbf{R}^{-1}$. The MES criterion for a linear model then turns out to be: maximize $\left| \mathbf{C}_\xi + \mathbf{F}_\xi \mathbf{R}^{-1} \mathbf{F}_\xi^T \right|$. We have

$$\left| \mathbf{C}_\xi + \mathbf{F}_\xi \mathbf{R}^{-1} \mathbf{F}_\xi^T \right| = \left| \mathbf{C}_\xi \right| \left| \mathbf{I}_p + \mathbf{R}^{-\frac{1}{2}} \mathbf{F}_\xi^T \mathbf{C}_\xi^{-1} \mathbf{F}_\xi \mathbf{R}^{-\frac{1}{2}} \right| = \left| \mathbf{C}_\xi \right| \left| \mathbf{R}^{-1} \right| \left| \mathbf{R} + \mathbf{F}_\xi^T \mathbf{C}_\xi^{-1} \mathbf{F}_\xi \right|$$

and since $\mathbf{R}$ does not depend on the design ξ the MES criterion for a linear model can also be formulated as: Select ξ such that $\left| \mathbf{C}_\xi \right| \left| \mathbf{R} + \mathbf{F}_\xi^T \mathbf{C}_\xi^{-1} \mathbf{F}_\xi \right|$ is maximized. Interestingly maximizing $\left| \mathbf{R} + \mathbf{F}_\xi^T \mathbf{C}_\xi^{-1} \mathbf{F}_\xi \right|$ alone is known as Bayesian D-optimality (Chaloner and Verdinelli, 1995).

  We are not working in a Bayesian framework and do not assume a prior distribution or a prior covariance matrix for β, so the GV-criterion is clearly different from MES for a linear model.

Despite these differences it might be interesting to compare GV-optimal designs with MES for Gaussian response just using a constant model. Here the MES-criterion is simply: choose ξ such that $\left| \mathbf{C}_\xi \right|$ is maximized. For simple kriging this criterion is equivalent to the GV-criterion because here the kriging covariance matrix is $\mathbb{C} \operatorname{ov} \left(\hat{\mathbf{Y}}_0 - \mathbf{Y}_0 \right) = \mathbf{C}_0 - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \mathbf{C}_{\xi 0}$ and minimizing $|\boldsymbol{\Sigma}_{\text{SK}}| = \left| \mathbf{C}_0 - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \mathbf{C}_{\xi 0} \right|$ is equivalent to maximizing $\left| \mathbf{C}_\xi \right|$ (see above).

With ordinary kriging the above equivalence is not clear: The kriging weights for ordinary kriging are

$$\mathbf{W} = \frac{1}{b} \mathbb{1}_m \mathbb{1}_k^T \mathbf{C}_\xi^{-1} + \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \left(\mathbf{I}_k - \frac{1}{b} \mathbb{1}_m \mathbb{1}_k^T \mathbf{C}_\xi^{-1} \right)$$

where $b = \mathbb{1}_k \mathbf{C}_\xi^{-1} \mathbb{1}_k^T$ and $\mathbb{1}_n$ is a vector of ones with length n. The kriging covariance matrix for ordinary kriging then is

$$\boldsymbol{\Sigma}_{\text{OK}} = \mathbf{C}_0 - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \mathbf{C}_{\xi 0} + \frac{1}{b} \left(\mathbb{1}_m - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \right) \left(\mathbb{1}_m - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \right)^T$$
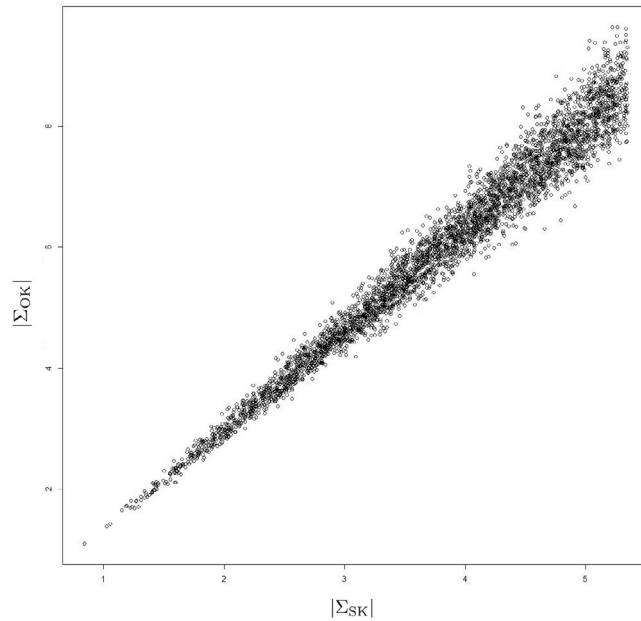
Fig. 2. GV-criterion function for simple kriging $|\Sigma_{\mathrm{SK}}|$ against GV-criterion function for ordinary kriging $|\Sigma_{\mathrm{OK}}|$ for different designs. The minima of both criteria seem to be the same.

and the GV-criterion in this case requires the minimizing of

$$\left|\boldsymbol{\Sigma}_{\mathrm{OK}}\right| = \left|\mathbf{C}_0 - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \mathbf{C}_{\xi 0}\right| \mathrm{abs}\left(1 + \frac{1}{b}\left(\mathbb{1}_m - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1}\right)^T \left(\mathbf{C}_0 - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1} \mathbf{C}_{\xi 0}\right)^{-1} \left(\mathbb{1}_m - \mathbf{C}_{\xi 0}^T \mathbf{C}_\xi^{-1}\right)\right)$$

The above criterion function is clearly different from the simple kriging setup and the following conjecture was verified by countless computations of $|\boldsymbol{\Sigma}_{\mathrm{SK}}|$ and $|\boldsymbol{\Sigma}_{\mathrm{OK}}|$ for the same randomly generated designs respectively with different covariance models, different functions $\mathbf{f}(x)$ and designs of different sizes. A formal proof eludes us.

Conjecture 1. *Let* $\left|\boldsymbol{\Sigma}_{SK}^1\right|$ *be the GV-criterion function for design* ξ_1 *in the simple kriging setup and* $\left|\boldsymbol{\Sigma}_{OK}^1\right|$ *the criterion function in the ordinary kriging setup for the same design* ξ_1*.* $\left|\boldsymbol{\Sigma}_{SK}^2\right|$ *and* $\left|\boldsymbol{\Sigma}_{OK}^2\right|$ *are defined analogously for another design* ξ_2*. Then*

$$\left|\boldsymbol{\Sigma}_{SK}^1\right| \lesseqgtr \left|\boldsymbol{\Sigma}_{SK}^2\right| \qquad \Leftrightarrow \qquad \left|\boldsymbol{\Sigma}_{OK}^1\right| \lesseqgtr \left|\boldsymbol{\Sigma}_{OK}^2\right|,$$

but the arguments of the minima are the same in both cases as can be seen in Fig. 2.

$$\arg\min_\xi \left|\boldsymbol{\Sigma}_{SK}\right| = \arg\min_\xi \left|\boldsymbol{\Sigma}_{OK}\right|$$

In the case of universal kriging the design criterium and the optimal design is clearly different from MES even though it can be observed that with increasing effective range *i.e.* with higher correlated response GV-optimal designs approach MES-designs also for universal kriging.

3. Incrementally and decrementally constructed GV-optimal designs

In many practical situations the experiment is not stopped after a fixed number of runs but say after a certain time or when a certain budget for the runs has expired. Thus at the start of the experiment the sample size is unknown and it is not clear which optimal designs of which size to use. In these situation incremental designs should be used, starting with a first design of (small) size k and supplementing it step by step with increments of size $l_i \geq 1$ until experimenting has to be stopped. Here the question arises how to find optimal increments in an efficient way which will be answered in this section. The efficiency of incrementally built designs compared to GV-optimal designs will be examined in Section 4.

The use of decrementally constructed designs is not directly motivable, it arises from the attempt of detecting GV-optimal designs of size $k + l$ in an incremental way. This may be done very efficiently with the help of Corollary 1 as will be shown below. The idea is to start with a design of minimal size k and compute an increment of size l which may be done with small computational effort. The only problem can be caused by the starting design which may contain design points which are not elements of the GV-optimal design of size $k + l$ leading to highly efficient but still suboptimal designs. In these situations an improvement can be achieved using a decrement, i.e. omitting k_1 design points followed by another incremental step. The hope in the decremental step is to get rid of inappropriate design points which in the following incremental step are substituted by design points yielding designs closer to GV-optimality.

3.1. Update formulae for kriging weights and the kriging covariance matrix of incremental designs

For this subsection we assume that we already have a k-point design and want to add l extra design points to simultaneously predict the remaining $m \gg l$ non-design points. As the calculation of the design criterion is computationally demanding, we will show that the use of update formulae for the kriging weights and consequently for the kriging covariance matrix is of great computational benefit.

Furthermore these update formulae may also be used for updating $\mathbb{C} \text{ ov} \left(\hat{Y}(x) \right)$ and $\mathbb{C} \text{ ov} \left(Y(x) | Y(\xi) \right)$ and also for incrementally built G- and V-optimal designs.

The allowable choice of the designs ξ_1 and ξ_2 with $\xi_1 \cap \xi_2 = \emptyset$ are from a fixed finite set of $N = k+l+m$ points $X = \{x_1, \dots, x_{k+l+m}\}$. $\xi_1 = \{x_1, \dots, x_k\}$ is the k-point first or starting design, $\xi_2 = \{x_{k+1}, \dots, x_{k+l}\}$ is the l-point second design (the increment), and $X \setminus \{\xi_1 \cup \xi_2\}$ is the remaining sets of non-design points with cardinality m. Thus we get the corresponding partitioning of the response vector $\mathbf{Y} = \left(\mathbf{Y}_{\xi_1}^T \, \mathbf{Y}_{\xi_2}^T \, \mathbf{Y}_0^T \right)^T$ or simpler $\mathbf{Y} = \left(\mathbf{Y}_1^T \, \mathbf{Y}_2^T \, \mathbf{Y}_0^T \right)^T$ with expectation and covariance matrix

$$\mathbb{E} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \mathbf{F}_0 \end{pmatrix} \beta \qquad \mathbb{C} \text{ ov} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_{12} & \mathbf{C}_{10} \\ \mathbf{C}_{12}^T & \mathbf{C}_2 & \mathbf{C}_{20} \\ \mathbf{C}_{10}^T & \mathbf{C}_{20}^T & \mathbf{C}_0 \end{pmatrix}. \tag{6}$$

In the first stage the design is ξ_1 and we have to predict $\left(\mathbf{Y}_2^T \, \mathbf{Y}_0^T \right)^T$ from $\mathbf{Y}_1$:

$$\begin{pmatrix} \hat{\mathbf{Y}}_2 \\ \hat{\mathbf{Y}}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{W}_{12} \\ \mathbf{W}_{10} \end{pmatrix} \mathbf{Y}_1, \tag{7}$$

where $\mathbf{W}_{12}$ are the weights of $\mathbf{Y}_1$ for the prediction of $\mathbf{Y}_2$ and analogously $\mathbf{W}_{10}$ the weights of $\mathbf{Y}_1$ for the prediction of $\mathbf{Y}_0$.

In the second stage we add the increment and the design is now $\xi_1 \cup \xi_2$ and we have to predict $\mathbf{Y}_0$ from $\left(\mathbf{Y}_1^T \, \mathbf{Y}_2^T \right)^T$ as

$$\hat{\mathbf{Y}}_0 = \begin{pmatrix} \mathbf{W}_1 & \mathbf{W}_2 \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \tag{8}$$

where $\mathbf{W}_1$ are the new weights of $\mathbf{Y}_1$ for the prediction of $\mathbf{Y}_0$ and analogously $\mathbf{W}_2$ the weights of $\mathbf{Y}_2$ for the prediction of $\mathbf{Y}_0$.

Emery (2009) showed that the kriging weights of the first stage (in our notation the components of $\mathbf{W}_{12}$ and $\mathbf{W}_{10}$) can be updated such that in our compact matrix notation

$$\mathbf{W}_1 = \mathbf{W}_{10} - \mathbf{W}_2 \mathbf{W}_{12}. \tag{9}$$

The only problem with (9) is that $\mathbf{W}_2$ is unknown if we add an increment to a smaller design. So, an "update" formula for the computation of the weights $\mathbf{W}_2$ (in fact this is not an update because weights of $\mathbf{Y}_2$ do not exist in the first stage) is essential for an efficient prediction of $\mathbf{Y}_0$ in the second stage.

Emery (2009) also presented update formulae for the kriging variances and covariances which unfortunately are wrong in the case of $l > 1$, which was shown with a simple counter example by Chevalier and Ginsbourger (2012). The presented "corrected" update formulae just have not been for the kriging covariance but for $\mathbb{C} \text{ ov} \left(\mathbf{Y}_0 | \mathbf{Y}_\xi \right)$. Eventually Chevalier et al. (2014) introduced correct update formulae for kriging variances and kriging covariances and also formulae for the new kriging weights of the second stage $\mathbf{W}_2$. In our notation these formulae may be summarized as follows.

3.1.1. (Updated) kriging weights for the second stage

Let the weights $\mathbf{W}_{12}$, $\mathbf{W}_{10}$, $\mathbf{W}_1$ and $\mathbf{W}_2$ be as defined in (7) and (8). Let further the kriging covariance matrix of the first stage in obvious notation be

$$\begin{pmatrix} \Sigma_2 & \Sigma_{20} \\ \Sigma_{20}^T & \Sigma_0 \end{pmatrix} \tag{10}$$

Then the new weights $\mathbf{W}_1$ and $\mathbf{W}_2$ can simply be computed with

$$\begin{aligned} \mathbf{W}_1 &= \mathbf{W}_{10} - \Sigma_{20}^T \Sigma_2^{-1} \mathbf{W}_{12} \\ \mathbf{W}_2 &= \Sigma_{20}^T \Sigma_2^{-1} \end{aligned} \tag{11}$$

An algebraic proof for the simultaneous computation of the $(m \times l)$-weight matrix $\mathbf{W}_2$ which is fundamentally different from the one given in Chevalier et al. (2014) can be found in Appendix.

3.1.2. Update formula for the kriging covariance matrix

Let the kriging covariance matrix of the first stage be as in (10), and the kriging covariance matrix of the second stage be Σ_0^+. Using the weights of (11) we can then update the kriging covariance matrix from the first stage to get

$$\Sigma_0^+ = \Sigma_0 - \Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} \tag{12}$$

An algebraic proof for this simultaneous update formula for the kriging covariance matrix which uses again another reasoning than Chevalier et al. (2014) can be found in Appendix.

3.2. Efficient computation of increments for GV-optimal designs

The following theorem is of fundamental importance as it allows the efficient incremental construction of GV-optimal designs.

Theorem 2. *The GV-optimal increment for the second stage given a design at the first stage then is*

$$\xi_{GV}^{inc} = \arg\max_{\xi_2} |\Sigma_2| . \tag{13}$$

Proof of Theorem 2. For a given design at the first stage the determinant of the corresponding kriging covariance matrix is fixed and finite and the determinant may be factored as:

$$\left| \begin{pmatrix} \Sigma_2 & \Sigma_{20} \\ \Sigma_{20}^T & \Sigma_0 \end{pmatrix} \right| = |\Sigma_2| \left| \Sigma_0 - \Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} \right|$$

The GV-optimal increment is the design ξ_2 that minimizes the determinant of the kriging covariance matrix of the second stage $\left| \Sigma_0^+ \right| = \left| \Sigma_0 - \Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} \right|$, which is exactly the second factor of the determinant factorization above. Obviously for a fixed kriging covariance matrix of the first stage minimizing $\left| \Sigma_0 - \Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} \right|$ is equivalent to maximizing $|\Sigma_2|$ what completes the proof. ∎

Remark 1. Theorem 2 is the reason for the great computational benefit of using incremental designs. It reduces the computation of the usual criterion function, which beside some matrix inversions demands the computation of the determinant of a $(m \times m)$-matrix, to the computation of the determinant of a $(l \times l)$-matrix with $l \ll m$. This fact enables an increase of the set of points $X = \{x_1, \ldots, x_{k+l+m}\}$ to an arbitrary size $N = k + l + m$ without raising the computational demands. Actually, the problem of finding $\arg\max_{\xi_2} |\Sigma_2|$ is known to be NP-hard, see Ko et al. (1995), the additional demand of computing the determinant of a $(m \times m)$-matrix would make it intractable already for moderate $m > 1000$.

Remark 2. Theorem 2 can even be used for an efficient computation of GV-optimal designs of a given size $k + l$: We start with some minimal preliminary design, i.e. a design of minimal necessary size $k = p$ where p is the number of deterministic functions of the locations $x \in \xi$: $\mathbf{f}(x) = \left(f_1(x)\ f_2(x)\ \cdots\ f_p(x) \right)$ used in the linear predictor. The design points can even be chosen randomly and its kriging covariance matrix is the basis for the computation of the increment ξ_2 as above. After this incremental step the design is reduced to size $k_1 \geq k$. Also the k_1 design points are chosen randomly out of the incremental design of size $k + l$. On the basis of these k_1 points we again compute the GV-optimal increment to end up with a design of size $k + l$. As the computational effort is small, these decremental and incremental steps may be repeated many times. If k and l are of moderate size we may even loop systematically through all k_1-combinations of the $k + l$ design points.

Remark 3. By applying the above incremental step several times we may also construct highly efficient sequential designs which accounts for active learning.

We can now formulate a similar procedure for efficient computation of increments for V-optimal designs.

Corollary 3. *The V-optimal increment for the second stage given a design at the first stage is*

$$\xi_V^{inc} = \arg\max_{\xi_2} tr\left(\Sigma_2 \right) + tr\left(\Sigma_2^{-1} \Sigma_{20} \Sigma_{20}^T \right) \tag{14}$$

Proof of Corollary 3. For a given design at the first stage the trace of the corresponding kriging covariance matrix is fixed and

$$\mathrm{tr} \begin{pmatrix} \Sigma_2 & \Sigma_{20} \\ \Sigma_{20}^T & \Sigma_0 \end{pmatrix} = \mathrm{tr}\left(\Sigma_2 \right) + \mathrm{tr}\left(\Sigma_0 \right)$$

The V-optimal increment is the design ξ_2 that minimizes the trace of the kriging covariance matrix of the second stage $\mathrm{tr}\left(\Sigma_0^+ \right) = \mathrm{tr}\left(\Sigma_0 \right) - \mathrm{tr}\left(\Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} \right)$. Obviously for a fixed kriging covariance matrix of the first stage minimizing $\mathrm{tr}\left(\Sigma_0 \right) - \mathrm{tr}\left(\Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} \right)$ is equivalent to maximizing $\mathrm{tr}\left(\Sigma_2 \right) + \mathrm{tr}\left(\Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} \right) = \mathrm{tr}\left(\Sigma_2 \right) + \mathrm{tr}\left(\Sigma_2^{-1} \Sigma_{20} \Sigma_{20}^T \right)$ which completes the proof. ∎

Remark 4. Corollary 3 is the reason for the computational benefit of using incremental designs. It reduces the computation of the usual criterion function which demands the matrix inversions of one $(k+l) \times (k+l)$- and one $(p \times p)$-matrix and the computation of the new kriging covariance matrix (10 matrix multiplications of which 5 involve matrices with m as one dimension) to the computation of the inverse of a $(l \times l)$-matrix and 2 matrix multiplications of which only 1 involves a $(l \times m)$-matrix. As $l \ll m$ this reduces the computational effort to roughly one third.

Just as Theorem 2, Corollary 3 can be used for an efficient computation of designs close to V-optimality of a given size $k + l$. The computational benefit of Corollary 3 cannot be compared to the improvement of Theorem 2, as we have to limit the number of incremental and decremental steps here and we thus have no guarantee to end up with the V-optimal design.

3.3. GV-optimal designs in design spaces dense in $\mathbb{R}^d$

In Section 2.1 we state that the allowable choice of the designs ξ is from a fixed finite set of $N = k + m$ points $X = \left\{ x_1, \ldots, x_{k+m} \right\}$. Usually $k \ll m$ and going into details of Corollary 1, Remark 2 we may even extend the number of non-design-points m to any (integer) size.

The reason for this remarkable possibility is that for the computation of the GV-optimal increment ξ_2 of a given design ξ_1 we do not even need the kriging covariance matrix of the first stage which would be of enormous dimension $(l + m) \times (l + m)$, we just need the $(l \times l)$-block $\boldsymbol{\Sigma}_2$ which is

$$\boldsymbol{\Sigma}_2 = \mathbf{W}_{12} \mathbf{C}_1 \mathbf{W}_{12}^T - \mathbf{W}_{12} \mathbf{C}_{12} - \mathbf{C}_{12}^T \mathbf{W}_{12}^T + \mathbf{C}_2$$

The dimensions of all of the above matrices are just k or l. $\mathbf{W}_{12}$ are the weights of $\mathbf{Y}_1$ for the prediction of $\mathbf{Y}_2$ in the first stage:

$$\mathbf{W}_{12} = \mathbf{F}_2 \mathbf{B} + \mathbf{C}_{12}^T \mathbf{A}$$

with $\mathbf{B} = \left(\mathbf{F}_1^T \mathbf{C}_1^{-1} \mathbf{F}_1 \right)^{-1} \mathbf{F}_1^T \mathbf{C}_1^{-1}$ and $\mathbf{A} = \mathbf{C}_1^{-1} \left(\mathbf{I}_k - \mathbf{F}_1 \mathbf{B} \right)$. The dimensions of all these matrices are only k, l and p, i.e. for the computation of $\boldsymbol{\Sigma}_2$ we just need small matrices independent of the number m.

The GV-criterion function is the determinant of the kriging covariance matrix, if in the first stage this determinant is D, then the criterion function of the incremental design is $\frac{D}{|\boldsymbol{\Sigma}_2|}$.

In the decremental step we remove $l_1 = k + l - k_1$ design points from the incremental design, the according $(l_1 \times l_1)$-block of the kriging covariance matrix of the first stage is $\boldsymbol{\Sigma}_2^*$ and may be computed as shown above.

The determinant of the kriging covariance matrix after the decremental step then is $\frac{D}{|\boldsymbol{\Sigma}_2|} \cdot |\boldsymbol{\Sigma}_2^*|$. So, not knowing the value of the GV-criterion during the search for the GV-optimal design is not crucial as it suffices to know the $\boldsymbol{\Sigma}_2$-blocks according to the increments and decrements to minimize the GV-criterion. Thus in principle we can make the grid as dense as desired, as long as the number of points is finite. There is reasonable hope that the described method can be generalized to continuous design spaces, which we defer to future research.

4. Efficiency of GV-optimal designs

As mentioned above it turns out that GV-optimal designs are highly efficient with other design criteria which will be discussed in Section 4.1.

Additionally Corollary 1 allows a simple, fast and computationally very efficient calculation of incrementally constructed designs that are close to be GV-optimal. The efficiency of these incremental designs will be discussed in Section 5.1.

4.1. Efficiency with respect to other design criteria

Traditional criteria for optimal designs for prediction are usually concerned with the variance of predictions, i.e. we could also title this subsection with "Efficiency with respect to variance-based criteria". Here we compare our GV-optimal designs with G- and V-optimal designs with the help of the **relative efficiency**, a very common concept in comparing designs, see eg. López-Fidalgo (2023), p.17.

Let Φ_G, Φ_V and Φ_{GV} be the criterion function for G-, V- and GV-optimality respectively and ξ_G, ξ_V and ξ_{GV} be the G-, V- and GV-optimal designs of the same size for the prediction of the same number of points. The GV-optimal design ξ_{GV} minimizes Φ_{GV}, as the other optimal designs minimize their corresponding design criteria. Then e.g. the relative G-efficiency of the V-optimal design is

$$E_G \left(\xi_G, \xi_V \right) = \frac{\Phi_G \left(\xi_G \right)}{\Phi_G \left(\xi_V \right)}.$$

We always have $E_. (\cdot, \cdot) <= 1$ and the relative efficiency of a design ξ gives the factor the criterion function of ξ may be decreased if we switch to the optimal design. These relative efficiencies are scale invariant though the effect of scaling is different for the GV-criterion function on the one and the G- and V-criterion functions on the other hand. The kriging covariance matrix of scaled responses $s \cdot \mathbf{Y}$ is $s^2 \boldsymbol{\Sigma}$ which affects G- and V-criterion functions the same: $\max \operatorname{diag} \left(s^2 \boldsymbol{\Sigma} \right) = s^2 \max \operatorname{diag} (\boldsymbol{\Sigma})$ and $\operatorname{tr} \left(s^2 \boldsymbol{\Sigma} \right) = s^2 \operatorname{tr} (\boldsymbol{\Sigma})$. The GV-criterion can be made insensitive to scaling by applying $\sqrt[k]{|\boldsymbol{\Sigma}|}$ instead without changing the designs.

The relative efficiencies are not affected by this scaling. *I.e.*, if $\boldsymbol{\Sigma}_{GV}$, $\boldsymbol{\Sigma}_G$ and $\boldsymbol{\Sigma}_V$ are the kriging covariance matrices for the originally unscaled data of the GV-, G- and V-optimal designs, then the GV-efficiency of the G- and V-optimal designs respectively for scaled data are

$$E_{GV} \left(\xi_{GV}, \xi_G \right) = \frac{s^m \sqrt{|\boldsymbol{\Sigma}_{GV}|}}{s^m \sqrt{|\boldsymbol{\Sigma}_G|}} \qquad \qquad E_{GV} \left(\xi_{GV}, \xi_V \right) = \frac{s^m \sqrt{|\boldsymbol{\Sigma}_{GV}|}}{s^m \sqrt{|\boldsymbol{\Sigma}_V|}} \quad ,$$

i.e. arbitrary scaling does not change the relative GV-efficiencies. The same is true also for relative G- and V-efficiency.

Table 1
Average relative efficiencies for linear trend over 54 different parameter combinations of Matern covariance models.

Linear	6 points			9 points			12 points		
trend	E_{GV}	E_G	E_V	E_{GV}	E_G	E_V	E_{GV}	E_G	E_V
ξ_{GV}	1	0.9050	0.9151	1	0.9707	0.9501	1	0.9385	0.9181
ξ_G	0.9328	1	0.9704	0.9315	1	0.9781	0.9639	1	0.9606
ξ_V	0.9010	0.9025	1	0.8756	0.9370	1	0.9316	0.9362	1

Table 2
Average relative efficiencies for quadratic trend over 54 different parameter combinations of Matern covariance models.

Quadratic	7 points			9 points			12 points		
trend	E_{GV}	E_G	E_V	E_{GV}	E_G	E_V	E_{GV}	E_G	E_V
ξ_{GV}	1	0.7951	0.8997	1	0.9928	0.9606	1	0.9100	0.9059
ξ_G	0.9173	1	0.9330	0.9712	1	0.9702	0.9360	1	0.9614
ξ_V	0.8175	0.7303	1	0.8739	0.8875	1	0.8966	0.9038	1

5. A representative example

Let us demonstrate the typical relative efficiencies on the basis of the following settings, computations of many other differently adjusted models and designs yield similar results.

The design space was chosen 2-dimensional on a regular grid, $\mathcal{X} = \{1, 2, \ldots, 17\}^2$, optimal designs were computed for the Matern covariance model with all combinations of range parameters $\phi \in \{0.1, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5\}$ and smoothness parameter $\kappa \in \{0.25, 0.5, 1, 1.5, 2, 2.5\}$. The variance as scaling parameter s was chosen such that for each of the 54 combinations of ϕ and κ the design criterion $|\Sigma|$ of the GV-optimal 12-point design is one. Designs of size 6 (with quadratic trend 7), 9 and 12 were computed for linear and quadratic trend functions.

As can be seen in Tables 1 and 2, GV-optimal designs are reasonably efficient with respect to the G- and V-criterion. Here for every covariance parameter combination the GV-, G- and V-optimal designs were determined and then for each optimal design the relative efficiencies with respect to the other design criteria were computed. This was here done for a linear trend function and for designs of size 6, 9 and 12 respectively. Finally, the relative efficiencies of the optimal designs were averaged over all 54 covariance parameter combinations. The lines of the tables correspond to GV-, G- and V-optimal designs, and e.g. the mean relative efficiency of 9-point GV-optimal designs with respect to the G-criterion is 0.9707 which means that the maximum kriging variance of G-optimal designs is on average only 97% of the maximum kriging variance of GV-optimal designs.

5.1. Efficiency of incrementally assembled designs

As already mentioned incrementally assembled designs are very efficient with respect to the GV-criterion. In the following discussion we will always start with some k-point design ξ_k adding a single increment of size l. The result will only be the GV-optimal design of size $(k + l)$ if we start in the first stage with a k-point design $\xi_k \subset \xi_{k+l}$ which is very improbable if we do not utilize additional knowledge. Usually we also will not end up in the GV-optimal $(k + l)$-point design if we start with the GV-optimal k-point design, but the result will be very close to GV-optimality. How close the incrementally constructed design is to GV-optimality depends on the choice of the k-point starting design. Of course we may take advantage of prior knowledge about properties of GV-optimal designs, e.g. if the design region is the unit square as in our simulation examples, we know that the GV-optimal design for a linear or quadratic trend will always have design points in the corners and the edges of the design region. Choosing such plausible points for the k-point starting design will almost always yield GV-optimal $(k + l)$-point designs that are constructed with a single incremental step.

Here we followed two ideas:

- start with a GV-optimal design of (small) size k;
- start with a plausible design of (small) size k, i.e. with design points that most likely are elements of GV-optimal designs of arbitrary size.

It turns out that both ideas yield very efficient designs especially if the increment (the number of additional design points) is not too small.

To exemplify this efficiency we again used Matern covariance models with 54 different parameter combinations and a quadratic trend. In the first simulation series we started with a plausible 6-point design, i.e. 4 design points at the corners and 2 points on opposite margins of the unit square. Then we added the optimal increment of 6 design points as described above. The mean GV efficiency of these incremental designs was 0.9911, the median efficiency was even 100%.

In the second series of simulations we started with the 7-point GV-optimal designs for each parameter combination respectively and added the optimal 5-point increment. Here the mean GV efficiency was 0.9862 and the median efficiency again 100% indicating a satisfactory performance of this simple incremental method.
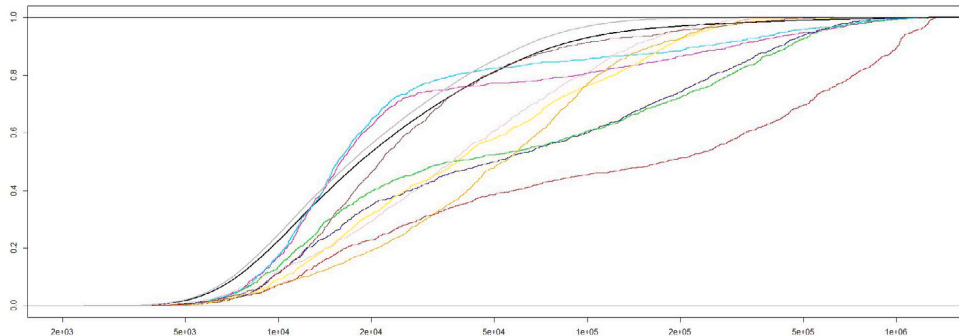
Fig. 3. Distribution of number of computations of the criterion function till convergence to the GV-optimal design including (black line) and excluding (gray line) distributions for "adverse" parameter combinations (colored lines) - mind the log-scale.

Table 3
Median number of computations of the criterion function until convergence to the GV-optimal design.

κ	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.75$	$\phi = 1$	$\phi = 1.5$	$\phi = 2$	$\phi = 3$	$\phi = 4$	$\phi = 5$
0.25	8614.5	36 704.0	15 789.0	16 425.0	18 160.0	20 516.0	22 970.5	25 423.5	27 438.0
0.5	8821.0	29 894.0	12 355.0	16 146.0	20 333.0	24 225.0	36 111.5	53 231.5	21 380.0
1	8895.0	25 630.0	11 937.5	15 876.0	21 115.0	29 033.5	33 065.0	36 066.0	14 808.5
1.5	9052.0	27 106.0	13 237.5	17 169.5	19 459.5	33 542.0	21 889.0	14 687.0	15 538.0
2	8943.0	27 852.0	15 681.0	15 581.5	20 344.0	34 206.0	21 146.5	14 710.5	179 657.0
2.5	8702.0	27 995.5	22 359.0	17 660.0	19 273.0	33 752.5	50 351.5	15 931.0	37 856.5

We applied the same concept to linear trends as well. The 4 corners of the design region were chosen as plausible starting design, and the increment of size 8 always yielded the GV-optimal 12-point design. Starting with 6-point GV-optimal designs for each parameter combination respectively and adding the optimal 6-point increment resulted in a mean GV-efficiency of 0.9881, again the median efficiency was 100%.

Note that the above mean and median efficiencies are just for a design built with a single incremental step. Of course we may always append a few decremental–incremental steps to guaranty GV-optimality. This approach is analyzed in the next section.

5.2. Efficiency of incrementally–decrementally assembled designs

The above described method of systematically discarding design-points after each incremental step has an efficiency of 100%. For each of the 54 combinations of the κ and ϕ parameters of the Matern covariance model (see above) we started 1000 times with a random design and ended with the GV-optimal design every time just with 26 exceptions (where a design with efficiency 99.9% was found instead of the optimal design). It turned out that these exceptions were all for designs with three special parameter combinations of κ and ϕ which seem to be adverse for finding the GV-optimal design on the chosen grid. For these three parameter combinations also the average number of required computations of the criterion function was incomparably higher then for other parameter values (see Table 3). The reason was obviously that the design points were limited to the unfavorable (17×17) grid. Changing to a (33×33) grid solved this problem. With the finer grid we started 200 times for each of the 54 parameter combinations of κ and ϕ with a random design and found the optimal designs without exception. Also the number of computations of the criterion function till convergence was distributed more uniform than with the (17×17) grid. Though there were some parameter combinations which needed clearly more function calls, for all these cases the second best design always was very efficient and the search algorithm now and then got stuck at these designs before finding the optimum.

5.2.1. Speed of convergence to GV-, G- and V-optimal designs

The speed of convergence was measured in absolute time and in the number of required computations of the criterion function.

In Table 3 we can see the median number of computations of the criterion function (in this case the determinant of the (6×6)-matrix Σ_2) needed to find the GV-optimal design. The overall median number of calls of the criterion function was 17,222, the computation time for finding 54.000 times the GV-optimal designs was 50.14 h; i.e. 3.34 s per optimal design.
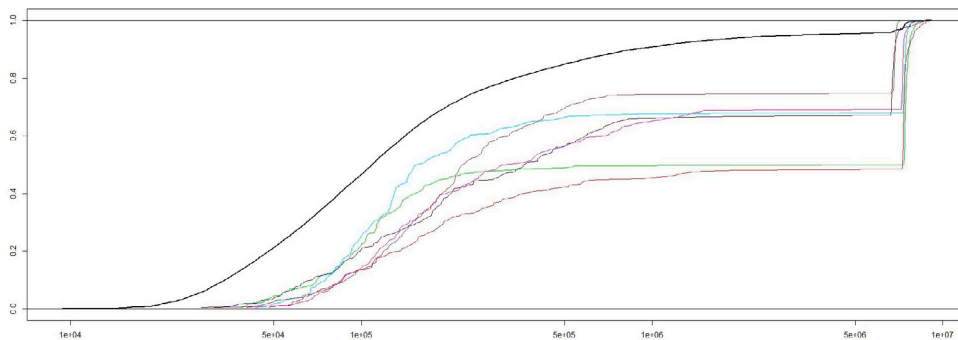
The distributions of the number of criterion function evaluations turned out to be positively skewed, the cdf of such distributions for selected parameter combinations of κ and ϕ is depicted in Fig. 3. There were 3 parameter combinations where the corresponding distributions of function calls were striking. Discarding these extreme distributions reduced the average computation time for one GV-optimal design to 2.18 s Changing to a grid twice as fine as the original (17×17) points to which the design was limited solved the problem (see above). The overall median number of calls of the criterion function increased to 24,877 with the (33×33) grid which was caused by the halved step size in the neighborhood search at the finer grid. Here continuous optimization algorithms with variable step size promise an improvement (see Section 3.3).

Table 4

Median number of computations of the criterion function until convergence to the G-optimal design.

κ	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.75$	$\phi = 1$	$\phi = 1.5$	$\phi = 2$	$\phi = 3$	$\phi = 4$	$\phi = 5$
0.25	5 257 485	5 656 141	5 054 189	1 283 471	5 187 172	5 251 948	5 164 463	4 815 720	5 856 377
0.5	5 371 069	6 300 232	4 827 949	3 486 381	3 627 394	4 942 996	5 708 568	5 169 052	1 830 513
1	4 293 287	5 535 166	5 426 545	4 163 447	5 408 354	5 452 206	3 041 665	5 309 562	7 470 061
1.5	4 465 675	4 167 031	5 375 233	4 891 069	5 037 755	5 178 237	5 707 364	4 545 278	6 490 938
2	5 457 978	5 412 340	4 255 818	2 904 401	5 235 663	5 193 218	6 097 624	6 819 590	6 976 215
2.5	4 748 402	5 251 379	2 365 235	5 010 149	6 236 549	3 721 464	6 237 881	5 627 650	6 549 851

Table 5

Median number of computations of the criterion function until convergence to the V-optimal design.

κ	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.75$	$\phi = 1$	$\phi = 1.5$	$\phi = 2$	$\phi = 3$	$\phi = 4$	$\phi = 5$
0.25	37 326.0	47 700.5	71 195.0	73 747.5	148 730.5	343 129.0	157 580.5	312 331.5	224 414.0
0.5	32 191.5	42 299.5	59 498.5	90 473.0	1662530.5	386 891.0	104 161.5	101 856.5	98 935.5
1	31 478.5	49 869.0	61 041.5	415 871.5	200 776.5	154 566.5	119 494.5	124 363.0	120 898.0
1.5	30 625.0	48 430.5	84 230.0	82 082.5	186 643.5	297 843.0	265 400.5	143 954.0	143 626.5
2	30 864.0	62 543.0	69 631.0	144 874.0	364 671.0	187 267.5	257 116.5	325 020.0	157 888.0
2.5	30 746.5	78 648.5	87 787.0	102 481.5	259 893.0	137 820.5	276 008.5	7 259 052.5	7 333 013.0

Fig. 4. Empirical cdf of the number of computations of the criterion function till convergence to the V-optimal design (black line) and for parameter combinations where the optimal design was not always found within a maximal number of tries (colored lines) - mind the log-scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The GV-optimal designs are found incomparably faster than corresponding G- and V-optimal designs even if the allowable choice of designs is from a moderate number of points.

Searching for G-optimal designs could only be tried 10 times for each of the 54 combinations of the κ and ϕ parameters and 368 of these 540 tries failed, because we had to stop the search algorithm (a combination of neighboring point exchanges and simulated annealing, same algorithm was used to find the GV-optimal designs) after 500 iterations because of the huge expenditure of time. The reason for that was partly the much larger computational effort but also a much slower convergence to the optimum, i.e. the criterion function ($\max \operatorname{diag} \left(\Sigma_0 - \Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} \right)$) had to be called much more frequently than in the search for GV-optimal designs. The mean G-efficiencies of the 540 found designs was 0.991, for κ and ϕ parameter combinations corresponding to high correlated data the mean G-efficiencies of the found designs were considerably smaller (with a minimum of 0.927 for $\kappa = 2.5$ and $\phi = 5$).

In Table 4 the median number of calls of the criterion function for each of the 54 combinations of the κ and ϕ parameters is reported. The overall median number of calls of the criterion function was 5.101.156 (296.2 times as often as for GV-optimal designs), the computation time for finding 540 times the G-optimal designs was 761.5 h, i.e. 1.41 h per optimal design which is 1520 times as long as for one GV-optimal design.

V-optimal designs are somehow found easier than G-optimal designs (because we may apply Corollary 2). We managed to search the V-optimal design 250 times for each of the 54 combinations of the κ and ϕ parameters. In Table 5 we can see the median number of computations of the criterion function (in this case $\operatorname{tr} \left(\Sigma_2 \right) + \operatorname{tr} \left(\Sigma_2^{-1} \Sigma_{20} \Sigma_{20}^T \right)$) needed to find the V-optimal design. Also here the average number of calls of the criterion function was clearly larger than for the GV-optimal design. The overall median number of calls of the criterion function was 108928.5 (6.3 times as often as for GV-optimal designs), the computation time for finding 13.500 times the V-optimal designs was 240.7 h, i.e. 64.2 s per optimal design which is 20 times as long as for one GV-optimal design.

Of the 13.500 tries to find the V-optimal design 577 failed, that is 4.3% (almost exactly 100 times more than for GV-optimal designs). This has also an impact on the cdf of the positively skewed distribution of the number of calls of the criterion functions until the optimal designs were found (Fig. 4). As with G-optimal designs we stopped the search algorithm after 500 iterations, the cdf's for parameter combinations where the optimal designs were not found within this maximal number of iterations are depicted as colored lines.
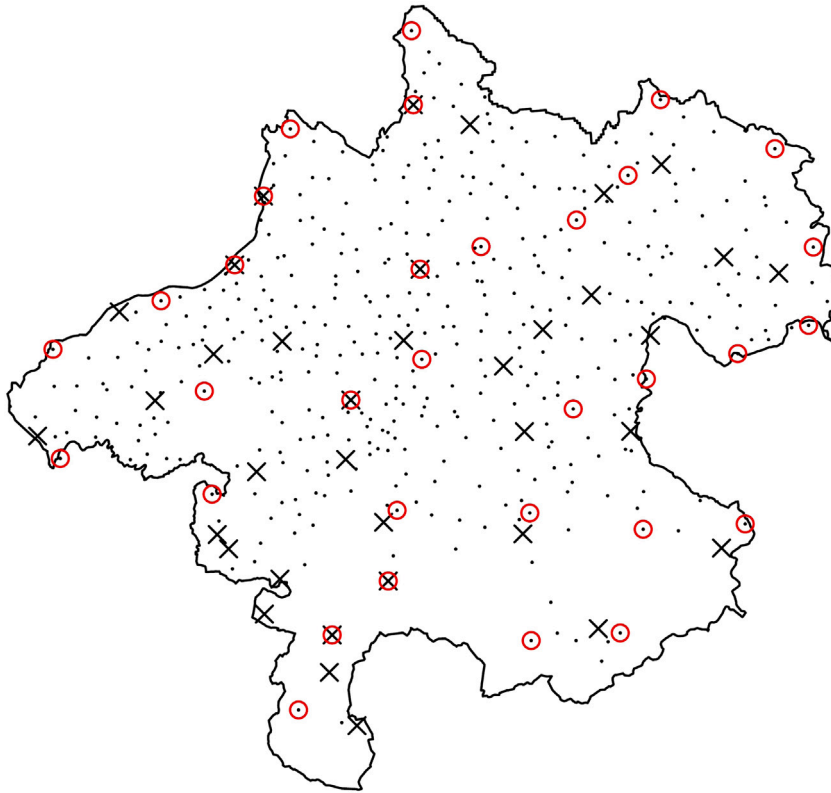
Fig. 5. Map of the 438 municipalities (dots), the current 36 meteorological stations (crosses), the locations of the 32-point GV-optimal design (red circles) for a meteorological network for simultaneous prediction of temperatures in the 438 municipalities of Upper Austria.

6. Real illustrative example: temperature prediction in upper Austrian municipalities

The province of Upper Austria is partitioned in 438 rural and urban municipalities with considerable topographical differences ranging from lowlands in the center and hill country in the north to high-altitude mountains in the south. As temperature and its spatial variation is strongly influenced by the topography the simultaneous prediction of temperatures at all principle locations of the 438 municipalities is challenging.

Currently there exist 36 meteorological stations in Upper Austria that may be taken as data source for temperature prediction in the 438 municipalities. A natural question is whether the current network can be improved by relocation of the station and/or we can even reduce the size of the network without loss of accuracy.

We model the expected monthly mean temperatures t_{ij} with the elevation of the measurement location el_i as external drift which is in line with Hudson and Wackernagel (1994):

$$\mu(t_{ij}) = \beta_{0j} + \beta_{1j} \cdot el_i,$$

where j indicates the month and i the location of the measurements. We further use an anisotropic Matérn covariance model to describe the spatial interdependencies of temperatures measured in the same time period. The covariance parameters have been estimated with the `likfit` function of the R package `GeoR` (Ribeiro and Diggle, 2001).

The learning data for parameter estimation were the daily mean temperatures of all meteorological stations in Upper Austria in the period from 2000-01-01 until 2023-10-25. The data are publicly available at the GeoSphere Austria Data Hub (2023). The coordinates and elevations of the 438 municipalities are also publicly available at the DORIS webOffice (https://www.doris.at/)

The parameter estimates confirm the environmental temperature lapse rate of $\sim$ 6.5 °C/km (International Civil Aviation Organization, 1993; Thompson, 1998) and are similar for all months except the winter period when the phenomenon of temperature inversion (National Oceanic and Atmospheric Administration's, 2023) may be observed frequently.

Here the showcase results for the month June are presented, other months are comparable. The design region is the set of 438 locations corresponding to the principle localities of the 438 Upper Austrian administrative municipalities, actually 36 of these locations are the base of meteorological stations (https://bitly.ws/ZFpC). We want to evaluate the prediction quality of this actual meteorological network by means of the GV-criterion and compare it with a virtual network positioned at the locations of the GV-optimal design of the same or a reduced size. I.e., initially we have $k = 36$ and $m = 438 - 36 = 402$ locations for simultaneous prediction.

We then further reduced the size step by step down to $k = 32$ until which we yielded the same accuracy according to the GV-criterion as from the initial network. This resulting GV-optimal design as well as the actual meteorological network and all 438 locations are displayed in Fig. 5. Remarkably 7 locations of the optimal design are already base of a meteorological station and other 7 locations of the optimal design are within a range of 5 km from an actual station although to our knowledge no statistical analysis was involved in determining the current network. While the reduction of just 4 stations from 36 down to 32 may seem as a disappointment one must not forget that the number of prediction locations had to increase from 402 to 406 making the task more difficult.

7. Conclusions

With GV-optimality we have introduced a novel design criterion for simultaneous kriging prediction, which considers the whole prediction covariance matrix. As was shown by the real-world example there are indeed practical problems requiring simultaneous rather than individual prediction. In such situations, the presented new criterion is a natural answer and more adequate and useful.

In terms of robustness, it has been demonstrated that GV-optimal designs exhibit considerable efficiency compared to designs optimized based on other criteria. The criterion function is notably smoother compared to other criteria, meaning that slight modifications to the design do not lead to significant alterations in the criterion function. Interestingly, this might also be the main reason that GV-optimal designs are found much faster than designs optimal with respect to other criteria (which is subject to actual and future research).

Furthermore we have shown that efficient incremental construction methods are available, which makes the criterion particularly attractive for big data and higher dimensional contexts. For instance it lends itself naturally combinable with local kriging techniques such as Gramacy and Haaland (2016). These and other extensions will be subject of future research.

Appendix. Proofs

A.1. Proof of the formula for the kriging weights of the second stage $\mathbf{W}_2$

Proof for simple kriging. We first compute kriging weights and the kriging covariance matrix for the first stage and apply the update formulae (11) and then compare the results to the kriging weights directly computed for the second stage which are the same what completes the proof.

The kriging weights for the first stage are $\mathbf{W}_{12} = \mathbf{C}_{12}^T \mathbf{C}_1^{-1}$ and $\mathbf{W}_{10} = \mathbf{C}_{10}^T \mathbf{C}_1^{-1}$. Plugging this weights into (5) gives the kriging covariance matrix for the first stage:

$$
\begin{pmatrix} \boldsymbol{\Sigma}_2 & \boldsymbol{\Sigma}_{20} \\ \boldsymbol{\Sigma}_{20}^T & \boldsymbol{\Sigma}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{C}_2 - \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} & \mathbf{C}_{20} - \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \mathbf{C}_{10} \\ \mathbf{C}_{20}^T - \mathbf{C}_{10}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} & \mathbf{C}_0 - \mathbf{C}_{10}^T \mathbf{C}_1^{-1} \mathbf{C}_{10} \end{pmatrix} \tag{A.1}
$$

This is plugged into the update formula (11) to get $\mathbf{W}_2 = \left(\mathbf{C}_{20}^T - \mathbf{C}_{10}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} \right) \left(\mathbf{C}_2 - \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} \right)^{-1}$ and $\mathbf{W}_1 = \mathbf{C}_{10}^T \left(\mathbf{C}_1^{-1} + \mathbf{C}_1^{-1} \mathbf{C}_{12} \right.$ $\left. \left(\mathbf{C}_2 - \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} \right)^{-1} \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \right) - \mathbf{C}_{20}^T \left(\mathbf{C}_2 - \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} \right)^{-1} \mathbf{C}_{12}^T \mathbf{C}_1^{-1}$ which should be the kriging weights for the second stage.

Now we compute the weights for the second stage directly:

$$
\mathbf{W} = \begin{pmatrix} \mathbf{C}_{10}^T & \mathbf{C}_{20}^T \end{pmatrix} \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_2 \end{pmatrix}^{-1}
$$

Using the identity

$$
\begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_2 \end{pmatrix}^{-1} =
$$
$$
= \begin{pmatrix} \mathbf{C}_1^{-1} + \mathbf{C}_1^{-1} \mathbf{C}_{12} \left(\mathbf{C}_2 - \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} \right)^{-1} \mathbf{C}_{12}^T \mathbf{C}_1^{-1} & -\mathbf{C}_1^{-1} \mathbf{C}_{12} \left(\mathbf{C}_2 - \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} \right)^{-1} \\ - \left(\mathbf{C}_2 - \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} \right)^{-1} \mathbf{C}_{12}^T \mathbf{C}_1^{-1} & \left(\mathbf{C}_2 - \mathbf{C}_{12}^T \mathbf{C}_1^{-1} \mathbf{C}_{12} \right)^{-1} \end{pmatrix} \tag{A.2}
$$

confirms $\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{W}_2 \end{pmatrix}$ which completes the proof. ■

Proof for universal kriging. We first compute the kriging weights for the second stage $\mathbf{W}_2$ directly and again use the identity (A.2):

$$\mathbf{W}_2 = \left[\left[\mathbf{F}_0 - \begin{pmatrix} \mathbf{C}_{10} \\ \mathbf{C}_{20} \end{pmatrix}^T \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{pmatrix}\right]\left[\begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{pmatrix}^T \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{pmatrix}\right]^{-1} \cdot \right.$$

$$\left. \cdot \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{pmatrix}^T + \begin{pmatrix} \mathbf{C}_{10} \\ \mathbf{C}_{20} \end{pmatrix}^T\right]\left[\begin{matrix} -\mathbf{C}_1^{-1}\mathbf{C}_{12}\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} \\ \left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} \end{matrix}\right]$$

after some matrix manipulations we get

$$\begin{aligned} \mathbf{W}_2 =\ & \left(\mathbf{C}_{20}^T - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} + \\ & + \left(\left(\mathbf{F}_0 - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right) - \left(\mathbf{C}_{20}^T - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\right) \cdot \\ & \cdot \left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1 + \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\right)^{-1} \cdot \\ & \cdot \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} \end{aligned} \tag{A.3}$$

Now we compute kriging weights and the kriging covariance matrix for the first stage: Using the notation of (6) the kriging weights of the first stage are

$$\mathbf{W}_{12} = \left(\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)^{-1}\mathbf{F}_1^T + \mathbf{C}_{12}^T\right)\mathbf{C}_1^{-1}$$

$$\mathbf{W}_{10} = \left(\left(\mathbf{F}_0 - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)^{-1}\mathbf{F}_1^T + \mathbf{C}_{10}^T\right)\mathbf{C}_1^{-1}$$

which gives the blocks of the kriging covariance matrix of the first stage (10)

$$\boldsymbol{\Sigma}_2 = \left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)^{-1}\left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right) + \mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}$$

$$\boldsymbol{\Sigma}_{20}^T = \left(\mathbf{F}_0 - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)^{-1}\left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right) + \mathbf{C}_{20} - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}$$

Now we apply the update formulae (11) to get

$$\begin{aligned} \mathbf{W}_2 =\ & \left(\mathbf{C}_{20} - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{C}_{12} + \left(\mathbf{F}_0 - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)^{-1}\left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\right) \cdot \\ & \cdot \left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12} + \left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)^{-1}\left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\right)^{-1} \end{aligned} \tag{A.4}$$

The inverse in the above equation

$$\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12} + \left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)^{-1}\left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\right)^{-1} \tag{A.5}$$

is the lower right column block of

$$\begin{pmatrix} -\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1 & \mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12} \\ \mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1 & \mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12} \end{pmatrix}^{-1}$$

which may also be computed as

$$\begin{aligned} & \left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} - \left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right) \cdot \\ & \cdot \left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1 + \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\right)^{-1} \cdot \\ & \cdot \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} \end{aligned} \tag{A.6}$$

In (A.4) we now substitute (A.6) for (A.5) to get

$$\begin{aligned} \mathbf{W}_2 =\ & \left(\mathbf{C}_{20}^T - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} + \\ & + \left(\mathbf{F}_0 - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)^{-1} \clubsuit \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} - \\ & - \left(\mathbf{F}_0 - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)^{-1}\left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right) \cdot \\ & \cdot \left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1 + \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\right)^{-1} \cdot \\ & \cdot \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} - \\ & - \left(\mathbf{C}_{20}^T - \mathbf{C}_{10}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right) \cdot \\ & \cdot \left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1 + \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\right)^{-1} \cdot \\ & \cdot \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1} \end{aligned}$$

At the position ♣ we now multiply with the identity

$$\begin{aligned} \mathbf{I} =\ & \left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1 + \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\right) \cdot \\ & \cdot \left(\mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{F}_1 + \left(\mathbf{F}_2^T - \mathbf{F}_1^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)\left(\mathbf{C}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{C}_{12}\right)^{-1}\left(\mathbf{F}_2 - \mathbf{C}_{12}^T\mathbf{C}_1^{-1}\mathbf{F}_1\right)\right)^{-1} \end{aligned} \tag{A.7}$$

to get the same W_2 as with direct computation (A.3). ∎

A.2. Proof of the update formula for the kriging covariance matrix of the second stage Σ_0^+

Proof. We compute the kriging covariance matrix for the second stage and plug in the update formulae for the kriging weights (11) to get the update formula for the kriging covariance matrix (12). The proof is given for the most general setup of universal kriging.

The kriging covariance matrix for the second stage is:

$$
\begin{aligned}
\Sigma_0^+ = \begin{pmatrix} \mathbf{W}_1 & \mathbf{W}_2 \end{pmatrix} \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_2 \end{pmatrix} \begin{pmatrix} \mathbf{W}_1^T \\ \mathbf{W}_2^T \end{pmatrix} - \begin{pmatrix} \mathbf{W}_1 & \mathbf{W}_2 \end{pmatrix} \begin{pmatrix} \mathbf{C}_{10} \\ \mathbf{C}_{20} \end{pmatrix} - \\
- \begin{pmatrix} \mathbf{C}_{10}^T & \mathbf{C}_{20}^T \end{pmatrix} \begin{pmatrix} \mathbf{W}_1^T \\ \mathbf{W}_2^T \end{pmatrix} + \mathbf{C}_0 = \\
= \mathbf{W}_1 \mathbf{C}_1 \mathbf{W}_1^T + \mathbf{W}_1 \mathbf{C}_{12} \mathbf{W}_2^T + \mathbf{W}_2 \mathbf{C}_{12}^T \mathbf{W}_1^T + \mathbf{W}_2 \mathbf{C}_2 \mathbf{W}_2^T - \\
- \mathbf{W}_1 \mathbf{C}_{10} - \mathbf{W}_2 \mathbf{C}_{20} - \mathbf{C}_{10}^T \mathbf{W}_1^T - \mathbf{C}_{20}^T \mathbf{W}_2^T + \mathbf{C}_0
\end{aligned}
$$

Now we plug in the update formula for the kriging weight (9) to get:

$$
\begin{aligned}
\Sigma_0^+ = \ & \mathbf{W}_{10} \mathbf{C}_1 \mathbf{W}_{10}^T - \mathbf{W}_{10} \mathbf{C}_{10} - \mathbf{C}_{10}^T \mathbf{W}_{10}^T + \mathbf{C}_0 - \\
& - \mathbf{W}_2 \left(\mathbf{W}_{12} \mathbf{C}_1 \mathbf{W}_{10}^T - \mathbf{W}_{12} \mathbf{C}_{10} - \mathbf{C}_{12}^T \mathbf{W}_{10}^T + \mathbf{C}_{20} \right) + \\
& + \mathbf{W}_2 \left(\mathbf{W}_{12} \mathbf{C}_1 \mathbf{W}_{12}^T - \mathbf{W}_{12} \mathbf{C}_{12} - \mathbf{C}_{12}^T \mathbf{W}_{12}^T + \mathbf{C}_2 \right) \mathbf{W}_2^T - \\
& - \left(\mathbf{W}_{10} \mathbf{C}_1 \mathbf{W}_{12}^T - \mathbf{W}_{10} \mathbf{C}_{12} - \mathbf{C}_{10}^T \mathbf{W}_{12}^T + \mathbf{C}_{20}^T \right) \mathbf{W}_2^T
\end{aligned}
$$

Using the kriging covariance matrix for the first stage in the universal kriging setup as in (A.1)

$$
\begin{aligned}
& \begin{pmatrix} \Sigma_2 & \Sigma_{20} \\ \Sigma_{20}^T & \Sigma_0 \end{pmatrix} = \\
& = \begin{pmatrix} \mathbf{W}_{12} \mathbf{C}_1 \mathbf{W}_{12}^T - \mathbf{W}_{12} \mathbf{C}_{12} - \mathbf{C}_{12}^T \mathbf{W}_{12}^T + \mathbf{C}_2 & \mathbf{W}_{12} \mathbf{C}_1 \mathbf{W}_{10}^T - \mathbf{W}_{12} \mathbf{C}_{10} - \mathbf{C}_{12}^T \mathbf{W}_{10}^T + \mathbf{C}_{20} \\ \mathbf{W}_{10} \mathbf{C}_1 \mathbf{W}_{12}^T - \mathbf{W}_{10} \mathbf{C}_{12} - \mathbf{C}_{10}^T \mathbf{W}_{12}^T + \mathbf{C}_{20}^T & \mathbf{W}_{10} \mathbf{C}_1 \mathbf{W}_{10}^T - \mathbf{W}_{10} \mathbf{C}_{10} - \mathbf{C}_{10}^T \mathbf{W}_{10}^T + \mathbf{C}_0 \end{pmatrix}
\end{aligned}
$$

and $\mathbf{W}_2 = \Sigma_{20}^T \Sigma_2^{-1}$ from (11) we get

$$
\begin{aligned}
\Sigma_0^+ = \ & \Sigma_0 - \Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} + \Sigma_{20}^T \Sigma_2^{-1} \Sigma_2 \Sigma_2^{-1} \Sigma_{20} - \Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20} = \\
= \ & \Sigma_0 - \Sigma_{20}^T \Sigma_2^{-1} \Sigma_{20}
\end{aligned}
$$

which completes the proof. ∎

References

Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: A review. Statist. Sci. 10 (3), 273–304. http://dx.doi.org/10.1214/ss/1177009939.

Chevalier, C., Ginsbourger, D., 2012. Corrected Kriging update formulae for batch-sequential data assimilation. http://dx.doi.org/10.48550/ARXIV.1203.6452, URL: https://arxiv.org/abs/1203.6452.

Chevalier, C., Ginsbourger, D., Emery, X., 2014. Corrected Kriging update formulae for batch-sequential data assimilation. In: Pardo-Igúzquiza, E., Guardiola-Albert, C., Heredia, J., Moreno-Merino, L., Durán, J.J., Vargas-Guzmán, J.A. (Eds.), Mathematics of Planet Earth. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 119–122.

Cressie, N., 1993. Statistics for Spatial Data, Revised ed. In: Wiley Series in Probability and Statistics, Wiley-Interscience, URL: http://www.worldcat.org/isbn/0471002550.

Dasgupta, S., Mukhopadhyay, S., Keith, J., 2022. Optimal designs for some bivariate cokriging models. J. Statist. Plann. Inference 221, 9–28. http://dx.doi.org/10.1016/j.jspi.2022.02.004, URL: https://www.sciencedirect.com/science/article/pii/S0378375822000167.

Dasgupta, S., Mukhopadhyay, S., Keith, J., 2022-03-12. G-optimal grid designs for Kriging models. arXiv:2111.06632, http://arxiv.org/abs/2111.06632.

Emery, X., 2009. The Kriging update equations and their application to the selection of neighboring data. Comput. Geosci. 13, 269–280. http://dx.doi.org/10.1007/s10596-008-9116-8, URL: https://link.springer.com/content/pdf/10.1007/s10596-008-9116-8.pdf.

Fang, K.-T., Li, R., Sudjianto, A., 2005. Design and Modeling for Computer Experiments (Chapman & Hall/CRC Computer Science & Data Analysis). Chapman and Hall/CRC, URL: http://www.worldcat.org/isbn/1584885467.

GeoSphere Austria Data Hub, 2023. https://bitly.ws/ZFbS. (Accessed: 23 November 2023).

Gramacy, R.B., Haaland, B., 2016. Speeding up neighborhood search in Local Gaussian process prediction. Technometrics 58 (3), 294–303. http://dx.doi.org/10.1080/00401706.2015.1027067.

Hudson, G., Wackernagel, H., 1994. Mapping temperature using Kriging with external drift: Theory and an example from Scotland. Int. J. Climatol. 14 (1), 77–91. http://dx.doi.org/10.1002/joc.3370140107.

International Civil Aviation Organization, 1993. Manual of the ICAO Standard Atmosphere: Extended to 80 Kilometres (262 500 Feet). In: Doc (International Civil Aviation Organization), International Civil Aviation Organization, URL: https://bitly.ws/ZFNU.

Kleijnen, J.P.C., 2009. Design and Analysis of Simulation Experiments. Springer US, URL: http://www.worldcat.org/isbn/144194415X.

Ko, C., Lee, J., Queyranne, M., 1995. An exact algorithm for maximum entropy sampling. Oper. Res. 43, 684–691.

López-Fidalgo, J., 2023. Optimal Experimental Design: a Concise Introduction for Researchers. Lecture Notes in Statistics, vol. 226, Springer Nature Switzerland, Cham, http://dx.doi.org/10.1007/978-3-031-35918-7.

Müller, W.G., Pronzato, L., Rendas, J., Waldl, H., 2015. Efficient prediction designs for random fields. Appl. Stoch. Models Bus. Ind. 31 (2), 178–194. http://dx.doi.org/10.1002/asmb.2084, URL: https://onlinelibrary.wiley.com/doi/10.1002/asmb.2084.

National Oceanic and Atmospheric Administration's, 2023. National weather service. https://w1.weather.gov/glossary/index.php?word=inversion. (Accessed: 07 November 2023).

Rasmussen, C.E., Williams, C.K.I., 2005. Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning Series, The MIT Press, URL: http://www.worldcat.org/isbn/026218253X.

Ribeiro, Jr., P., Diggle, P., 2001. geoR: A package for geostatistical analysis. R-NEWS 1 (2), 15–18, URL: http://cran.R-project.org/doc/Rnews.

Santner, T.J., Williams, B.J., Notz, W., 2003. The Design and Analysis of Computer Experiments. Springer Series in Statistics, Springer, URL: http://www.worldcat.org/isbn/0387954201.

Shewry, M.C., Wynn, H.P., 1987. Maximum entropy sampling. J. Appl. Stat. 14 (2), 165–170. http://dx.doi.org/10.1080/02664768700000020.

Thompson, R., 1998. Atmospheric Processes and Systems. In: Routledge Introductions to Environment: Environmental Science Series, Routledge, URL: https://www.routledge.com/Atmospheric-Processes-and-Systems/Thompson/p/book/9780415171465.

Wald, A., 1943. On the efficient design of statistical investigations. Ann. Math. Stat. 4, 134–140.

Wang, J.-F., Stein, A., Gao, B.-B., Ge, Y., 2012. A review of spatial sampling. Spatial Stat. 2, 1–14. http://dx.doi.org/10.1016/j.spasta.2012.08.001.

Wilks, S.S., 1932. Certain generalizations in the analysis of variance. Biometrika 24 (3/4), 471–494, URL: https://www.jstor.org/stable/2331979.