# A convex approach to optimum design of experiments with correlated observations*

**Andrej Pázman**

*Comenius University Bratislava, Slovakia*
*e-mail:* andrej.pazman@fmph.uniba.sk

**Markus Hainy and Werner G. Müller**

*Johannes Kepler University Linz, Austria*
*e-mail:* markus.hainy@jku.at; werner.mueller@jku.at

**Abstract:** Optimal design of experiments for correlated processes is an increasingly relevant and active research topic. Present methods have restricted possibilities to judge their quality. To fill this gap, we complement the virtual noise approach by a convex formulation leading to an equivalence theorem comparable to the uncorrelated case and to an algorithm giving an upper performance bound against which alternative design methods can be judged. Moreover, a method for generating exact designs follows naturally. We exclusively consider estimation problems on a finite design space with a fixed number of elements. A comparison on some classical examples from the literature as well as a real application is provided.

## 1. Introduction

### 1.1. An elementary description

In this paper we propose a particular approach for evaluating the performance of any $n$-point exact design for linear regression with correlated observations. This means that we consider any $n$-tuple of observation points $x_1, \ldots, x_n$ within a model

$$y(x) = f^{\mathrm{T}}(x)\theta + \varepsilon(x); \quad x \in \mathcal{X}, \tag{1}$$

where $\mathcal{X} = \{x_1, \dots, x_N\}$ is a finite discrete space containing $N$ points and $\theta$ is the unknown vector of parameters with $\dim(\theta) = p$. What is further conventionally assumed is that the random error of the observation collected at $x$ has zero expectation, i.e. $E\{\varepsilon(x)\} = 0$, and is correlated with observations at other locations, defined through the $N \times N$ matrix $C$ with elements

$$C_{ij} = \text{cov}\{\varepsilon(x_i), \varepsilon(x_j)\}; \qquad i, j = 1, \dots, N,$$

which is supposed to be known.

The number $n$ of required observations is considered fixed, and the points $x_1, \dots, x_n$ are to be taken from a design space coinciding with $\mathcal{X}$. Our ultimate aim is to look for the $n$-point exact design which allows for the most precise estimation of the parameters of the model (1).

As is well known, following [13], excellent results in this sense have been obtained when the observations are uncorrelated and consequently can be replicated independently. Then one can augment the set of possible exact designs with some "design measures" reflecting possible replications of the observations. This approach allowed the use of very powerful tools of convex optimization for computing the optimum "design measure of replications."

The situation is quite different when the observations are correlated and hence independent replications are not possible. The approach we take in our paper is to enlarge the set of possible $n$-point exact designs (fixed $n$) to a larger set of "design measures" in a way to make use of convexity considerations. This is achieved by adding some purely theoretical random variable, called virtual noise, to the real errors of observations. Therefore, we pass from model (1) to an auxiliary model and from the set of all exact $n$-point designs to a convex set of all design measures meaningful in this auxiliary model. Nevertheless, we still keep in mind our ultimate aim, namely the consideration of $n$-point exact designs in the original model (1). These $n$-point exact designs arise as special cases in our auxiliary model when the design measure is uniformly concentrated just in $n$ points. Then our model fully coincides with model (1) considered with nonreplicated observations performed just in those $n$ support points. This is because in this particular case the virtual noise in the auxiliary model automatically disappears, or put more precisely, it has zero variance and zero mean.

A nice property of our auxiliary model is that it does not coincide with model (1) if the design measure is uniformly concentrated in $k$ points, $k \neq n$. Another nice property is that in the particular case of uncorrelated observations with constant variances, our concept of design measure coincides with the design measure commonly used for uncorrelated observations with great success. The virtual noise is used here to turn the problem of optimum design in the auxiliary model into a convex problem. Consequently, we present a method for numerically computing an optimum design measure and use it heuristically to obtain nearly optimum exact $n$-point designs for model (1). Moreover, using our convex approach we can provide an upper bound for the performance of any given $n$-point design for model (1). This allows us to quantify the relative quality of various designs by comparing them to the closeness to this bound, which was not possible before.

To have an idea of possible applications consider the problem of an optimal localization of weather stations for daily rainfall observations (see our Example 3). The average rainfall over the whole territory is modeled by a regression with unknown parameters and the problem to be solved is to find $n$ best localizations of weather stations to get the best estimates of these parameters.

## *1.2. Motivation*

Due to its importance in spatio-temporal monitoring (cf. [14]) and particularly computer simulation (cf. [23]), experimental design for regression models with correlated errors has gained increasing interest. The theory is well developed for classical (non-)linear regression with uncorrelated errors (cf. [15]), where the corresponding optimum design approach has been initiated from Kiefer's concept of design measures (cf. [13]) and has emerged into textbook status (cf. [1, 10], etc.). However, for the correlated setting the literature is only scattered.

The meanwhile classic approach by [25] relied heavily on asymptotic considerations, making it valid and useful for only limited situations. In contrast, the algorithm devised by [3] is purely heuristic but applicable to and surprisingly efficient for a great variety of problems. The proposal of [8] to transform the problem into a random coefficient model allows for embedding it into standard convex design theory but requires elaborate tuning to achieve a given design size. More recently, in a remarkable series of papers started with [29], Dette, Pepelyshev and Zhigljavsky built upon and extended much of the discussions and material exposed in the pioneering monograph by [18], essentially concentrating on the ordinary least squares estimator. Latest, more promising approaches derived from the best linear unbiased estimator of the continuous-time model were proposed in [6] and [5].

An entirely different suggestion redefining the role of design measures was given in [20] and fully developed in [16]. Therein a virtual noise was introduced to influence the behavior of the designs and the corresponding measure was reflecting the amount of signal suppression. The resulting procedures were quite effective, but their broader application was hampered by the downside that the method was nonconvex and thus no Kiefer-Wolfowitz-type equivalence theorem was available. In the present paper we fill this gap by asserting convexity through a new and different variant of virtual noise and consequently provide a corresponding equivalence theorem. We will show the effectiveness of this novel modification on classical examples from the literature and compare to the above mentioned alternative techniques.

## 2. Theory

## *2.1. The introductory setup*

In the presentation we will constrict ourselves on linear regression models with the obvious generalizations to the asymptotic approach to the nonlinear case as

outlined in [21], so we assume our observations to be generated from the model (1). The Gaussian process literature additionally assumes normality, which leads to the well-known kriging equations for prediction (cf. [23]).

Let us now first consider an exact unreplicated $n$-point design $\mathcal{T} \subset \mathcal{X}$, where $n$ is the required number of observations. The information matrix of $\theta$ in the model (1) and for the design $\mathcal{T}$ is inversely related to the (asymptotic) variance-covariance matrix of the best linear unbiased estimator $\hat{\theta}$. As is well known, this information matrix is $M_{\mathcal{T}} = F(\mathcal{T})^{\mathrm{T}} C^{-1}(\mathcal{T}) F(\mathcal{T})$, where $C(\mathcal{T})$ is the $n \times n$ submatrix of $C$ corresponding to points of the set $\mathcal{T}$ and $F(\mathcal{T})$ is an $n \times p$ matrix, $F_{i,j}(\mathcal{T}) = f_j(x_i)$, with $x_i \in \mathcal{T}$ and $j = 1, \ldots, p$.

The situation is quite different when the observations are uncorrelated, so that for any $x \in \mathcal{X}$ observations $y(x)$ can be replicated. Traditionally, following [13], we then define a design as a probability measure $\xi$ on $\mathcal{X}$, $\xi(x)$ being interpreted as to be proportional to the number of replicated observations at the design point $x$.

Typically criteria functions used for uncorrelated observations, i.e. the case of a diagonal $C$, can be expressed as concave increasing functions of the information matrix. For example, for D-optimality we have $\Phi(M_\xi) = \{\det(M_\xi)\}^{1/p}$ or $\Phi(M_\xi) = \log \det(M_\xi)$, with $M_\xi = \sum_{x \in \mathcal{X}} f(x) f^{\mathrm{T}}(x) \xi(x)$.

It is the usual purpose of optimum design to find the design

$$\xi^* = \arg\max_\xi \Phi(M_\xi).$$

The maximum of $\Phi(M_\xi)$ over the set of all designs $\Xi$ can be obtained by convex methods. In case of strict concavity of $\Phi(M_\xi)$, there is only one maximum value of $M$.

It is fully justified to use in the correlated setup of model (1) the same optimality criteria functions $\Phi(M)$ as for the uncorrelated case, but now $M_{\mathcal{T}}$ should be used instead of $M_\xi$ ([19]). In the method presented in this paper, we take great advantage of the fact that all statistically meaningful criteria functions $\Phi(M)$ are concave functions on the set of all symmetric positive semidefinite matrices $M$, cf. [22]. Thus, we will be able to employ efficient convex optimization methods for the evaluation of approximate designs.

Note that this is usually not the case in the correlated setup, i.e. for a general $C$, where typically sophisticated non-convex optimization techniques, such as simulated annealing or particle swarms, have to be applied.

### 2.2. Some techniques from the literature

Such non-convex methods are for instance required for the mathematically elaborate design approaches developed by Dette, Zhigljavsky and collaborators, most recently in [6] and [5]. There they generate their results by moving from the discretely indexed model (1) to its continuously indexed counterpart and back, thus ending up with the eventual need for discretization and approximation of the best linear unbiased estimator. In [6], they make use of the fact that the

optimally-weighted signed least squares estimator has the same variance as the best linear unbiased estimator. They derive the weight function of the signed least squares estimator for the continuous model and regard this as the design measure. In [5], they propose an approximate criterion based on a discrete approximation of the best linear unbiased estimator for the continuous model. The approximate criterion needs to be optimized numerically in the same way as the original criterion, so there is no major conceptual difference when it comes to optimization with the exception of the former being computationally more efficient if $p \ll n$. Therefore, in the following we shall not consider these approximate criteria, as [5] have already investigated how good/bad this approximation works. Limitations of the methods of [6] and [5] are that they are constrained to a limited form of covariance kernels and that it may be quite difficult to generalize to higher-dimensional design spaces.

The most straightforward and pragmatic approach, however, is the greedy algorithm for exact design generation first proposed by [3], in the modification and interpretation given in [8]. There the so-called sensitivity function of the classical one-point correction algorithms for D-optimality in the uncorrelated setup is used with expressions simply replaced by their analogues from the correlated case. That is, at each step $r$ we augment the point

$$x_r = \arg\max_{\mathcal{X}} f^{\mathrm{T}}(x) M_{\mathcal{T}_{r-1}}^{-1} f(x)/\sigma_\varepsilon^2(x)$$

to the current design $\mathcal{T}_{r-1} \subset \mathcal{T}$, with $f$, $M$, and the error variance $\sigma_\varepsilon^2$ replaced accordingly, see the appendix for details. Furthermore, the appendix contains an adaption for A-optimality, which will be utilized in some of the examples. Note that despite its compellingly simple form, this algorithm has little theoretical basis other than that the independent case can be seen as a particular limit instance of the general setup.

A good overview of all these approaches with an application in spatio-temporal sensor placement, albeit mainly from the perspective of using the ordinary least squares estimator, can most recently be found in [27].

### *2.3. Virtual noise*

In this paper, we will use a technique that has first been proposed in [20], which is very different from the above. Despite a certain similarity to the standard kriging setup, in that we consider two error components, one correlated and the other uncorrelated, our additional uncorrelated error component however is not considered as an observational error but as a regulatory device without physical meaning, termed virtual noise. To make this operational, the variance of this virtual noise needs to be assumed being of a specific form. Unfortunately the specifications given in [20] as well as in [16] both yielded a nonconvex solution. In this paper, we intend to fill this gap by providing convexity and consequently an equivalence theorem that allows for quick checks of the quality of $n$-points designs in model (1) as well as an algorithm to find the optimum.

First, similarly as in [16], define a convex set of restricted probability measures, also termed design measures or simply designs,

$$\Xi = \left\{ \xi : \sum_{x \in \mathcal{X}} \xi(x) = 1, \ \forall_{x \in \mathcal{X}} \ 0 \ \leq \ \xi(x) \leq 1/n \right\},$$

on the design space $\mathcal{X}$, which is supposed to be finite. The integer number $n$ represents the desired number of points of support of an optimum exact design $\xi^*$. Now instead of (1), consider the auxiliary perturbed model

$$y(x) = f^{\mathrm{T}}(x)\theta + \varepsilon(x) + w_\xi(x); \quad x \in \mathcal{X}, \tag{2}$$

where the variance of the supplementary 'virtual noise' $w_\xi$, independent of $\varepsilon$, is fixed as

$$\mathrm{var}\left\{w_\xi(x)\right\} = \sigma_\xi^2(x) = \kappa \frac{1/n - \xi(x)}{\xi(x)},$$

whereas $\mathrm{cov}\left\{w_\xi(x), w_\xi(x')\right\} = 0$ when $x \neq x'$. The positive number $\kappa$ is a tuning parameter which must be chosen within the bounds given by Theorem 2.1 and should be taken as large as possible to emphasize the influence of the virtual noise. The experience obtained in the examples below is that a smaller choice of $\kappa$ diminishes mainly the speed of the linear programming algorithm from Section 2.4.

That means, similarly as in [16] in this auxiliary model (2), if $\xi(x) = 0$, i.e. $\sigma_\xi^2 = \infty$, there is no observation at the point $x$, and if $\xi(x) = 1/n$, i.e. $\sigma_\xi^2 = 0$, the observation at the point $x$ is not disturbed at all by the virtual noise.

An important notion in model (2) is the concept of a realizable $k$-point design, by which we denote a design $\xi \in \Xi$ for which the signal is not disturbed by the virtual noise in the $k$ points $x_1, \ldots, x_k \in \mathcal{X}$, i.e. $\xi(x_i) = 1/n$, and the signal is totally suppressed in all other points, i.e. $\xi(x) = 0 \ \forall x \in \mathcal{X} \backslash \{x_1, \ldots, x_k\}$. It follows trivially from the definition of $\Xi$ that all realizable $n$-point designs are in $\Xi$, but $\Xi$ contains no realizable $k$-point designs with $k \neq n$. The set $\Xi$ has also another property: it is the smallest closed convex set of probability measures on $\mathcal{X}$ which contains all realizable $n$-point designs. In the proofs we shall also require the set

$$\Xi_+ = \left\{ \xi : \sum_{x \in \mathcal{X}} \xi(x) = 1, \ \forall_{x \in \mathcal{X}} \ 0 < \xi(x) \leq 1/n \right\} = \left\{ \xi \in \Xi : \ \mathrm{supp}(\xi) = \mathcal{X} \right\}.$$

Its advantage is that $\mathrm{var}\left\{w_\xi(x)\right\} < \infty$ for every $x \in \mathcal{X}, \xi \in \Xi_+$. One can easily see that $\Xi$ is the closure of $\Xi_+$.

Denote by $W(\xi)$ the $N \times N$ diagonal matrix

$$\begin{aligned}
W_{i,j}(\xi) &= \kappa \frac{1/n - \xi(x_i)}{\xi(x_i)}; \ \text{if} \ \ i = j \\
&= 0; \qquad\qquad \text{if} \ \ i \neq j.
\end{aligned}$$

When $\xi \in \Xi_+$, i.e. $\mathrm{supp}(\xi) = \mathcal{X}$, the information matrix of $\theta$ in model (2) is given by

$$M(\xi) = F^{\mathrm{T}} \left\{ C + W\left(\xi\right) \right\}^{-1} F, \tag{3}$$

where $C$ is the $N \times N$ covariance matrix defined after (1), and $F$ is the $N \times p$ matrix $F_{i,j} = f_j(x_i)$, $x_i \in \mathcal{X}$, $i = 1, \ldots, N$, $j = 1, \ldots, p$. Evidently, $M(\xi)$ is nonsingular if $F$ has full rank $p$.

On the other hand, when $\xi \in \Xi \backslash \Xi_+$, we define

$$M(\xi) = F'^{\mathrm{T}} \left\{ C' + W'\left(\xi\right) \right\}^{-1} F', \tag{4}$$

where $C'$, $W'(\xi)$ are submatrices of $C$, $W(\xi)$ restricted to the support $\mathrm{supp}(\xi)$, and similarly are the rows of $F'$. An important property is the continuity of $M(\xi)$ on the whole set $\Xi$, i.e. $\xi_n(x) \to \xi(x)$ for every $x \in \mathcal{X}$ implies $M(\xi_n) \to M(\xi)$, see Lemma F.1 in the appendix.

Another interesting property of $M(\xi)$ is that in case of uncorrelated observations with constant variances, $C = \sigma^2 I$, with $\kappa = \sigma^2$ we obtain $M(\xi)$ proportional to $M_\xi$, the information matrix from Kiefer's design theory. However, for general $C$ the design measure can no more be interpreted as reflecting the replications of observations. In any case we can still use it to obtain heuristically exact $n$-point designs approaching the quality of an optimum exact design as will be demonstrated in the examples section.

We now have the following

**Theorem 2.1.** *If $\kappa \leq \lambda_{\min}(C)$, the minimal eigenvalue of the matrix $C$, and if $\Phi(M)$ is any optimality criterion expressed as a concave, increasing, and continuous function of the matrix $M$, then the mapping*

$$\xi \in \Xi \to \Phi\left\{M(\xi)\right\}$$

*is concave as well, with $M(\xi)$ defined in (3) and (4).*

The proof of the theorem is given in the appendix. Its consequences are that the maximum of $\Phi(M)$ over the set $\Xi$ can be obtained by convex methods, and in case of a strict concavity of $\Phi(M)$, there is only one optimal $M$.

Unfortunately however, the gradient method so popular in uncorrelated observations cannot be used here because one-point design measures do not enter into the set $\Xi$. Neither can we use the usual equivalence theorem, which requires the use of one-point designs as well. However, the methods of modified linear programming can be employed, see the next subsection, and also a certain form of the equivalence theorem can be provided, which is given in Section 2.5.

### 2.4. A design algorithm

The method used in this subsection originated as the cutting-plane method of [12], was then adapted to experimental design in Chapter 9 of [21] and applied algorithmically to classical experimental design in [4].

For computational reasons we shall restrict our attention in this section to the set

$$\Xi^\epsilon = \{\xi \in \Xi : \ \forall_{x \in \mathcal{X}} \ \xi(x) \geq \epsilon\}$$

for some small $\epsilon > 0$. Evidently,

$$\Xi_+ = \bigcup_{\epsilon > 0} \Xi^\epsilon.$$

Since the closure of $\Xi_+$ is the set $\Xi$ and since according to Lemma F.1 in the appendix the criterion function $\Phi\{M(\xi)\}$ is continuous on the whole of $\Xi$, we can approach the value of $\max_{\xi \in \Xi} \Phi\{M(\xi)\}$ arbitrarily close by computing $\Phi\{M(\xi_\epsilon^*)\}$, where $\xi_\epsilon^* \in \arg\max_{\xi \in \Xi^\epsilon} \Phi\{M(\xi)\}$.

If $\Phi(M)$ is any (concave) criterion, like $D-$ or $A$-optimality, having a gradient $\nabla_M \Phi(M)$, then for any $\xi$ in $\Xi_+ = \{\xi \in \Xi : \text{supp}(\xi) = \mathcal{X}\}$ we have the derivative

$$\frac{\partial \Phi\{M(\xi)\}}{\partial \xi(\bar{x})} = \text{tr} \left[ \nabla_M \Phi\{M(\xi)\} \ \frac{\partial M(\xi)}{\partial \xi(\bar{x})} \right],$$

where $\nabla_M \Phi\{M(\xi)\}$ denotes $\nabla_M \Phi(M)$ at $M = M(\xi)$. The linear Taylor formula for $\xi \to \Phi\{M(\xi)\}$ at the point $\mu$ gives

$$\Phi\{M(\mu)\} + \sum_{\bar{x} \in \mathcal{X}} \text{tr} \left[ \nabla_M \Phi\{M(\mu)\} \ \frac{\partial M(\mu)}{\partial \mu(\bar{x})} \right] \{\xi(\bar{x}) - \mu(\bar{x})\}, \tag{5}$$

which is linear in $\xi$. Due to the concavity of $\Phi[M(\xi)]$ its graph is always below any point of its tangent plane. The expression (5) corresponds to this tangent plane at the point $\mu$. Consequently, we have for any $\xi \in \Xi^\epsilon$

$$\Phi\{M(\xi)\} = \min_{\mu \in \Xi^\epsilon} \left( \Phi\{M(\mu)\} + \sum_{\bar{x} \in \mathcal{X}} \text{tr} \left[ \nabla_M \Phi\{M(\mu)\} \ \frac{\partial M(\mu)}{\partial \mu(\bar{x})} \right] \{\xi(\bar{x}) - \mu(\bar{x})\} \right) \tag{6}$$

and $\arg\max_\xi \Phi\{M(\xi)\}$ is the optimal design on $\Xi^\epsilon$.

We can now reformulate the design problem into a linear program. Any concave and positive criterion

$$\xi \in \Xi^\epsilon \to \Phi\{M(\xi)\}$$

can be written in a form

$$\Phi\{M(\xi)\} = \min_{\mu \in \Xi^\epsilon} \left\{ a(\mu) + \sum_{i=1}^N b_i(\mu) \, \xi(x_i) \right\}.$$

Here, $a(\mu)$ and $b_i(\mu)$ can be obtained either from the linear Taylor formula (6), or, in case that there is no gradient, from subgradients, or using some algebraic tricks as in [4].

At the first step of the linear program, we set a finite set $\Xi_1 \subset \Xi^\epsilon$, which will be increased in the later steps. Hence, at the $k$th step we start with a set $\Xi_k$.

Consider the following finite set of constraints, which are linear in the variables $t \in \mathbb{R}$ and $\xi(x)$; $x \in \mathcal{X}$,

$$t \geq 0, \qquad t \leq a(\mu) + \sum_{i=1}^{N} b_i(\mu) \, \xi(x_i), \quad \mu \in \Xi_k;$$

$$\xi(x_i) \geq \epsilon; \qquad \xi(x_i) \leq \frac{1}{n}; \qquad i = 1, \dots, N, \qquad \sum_{i=1}^{N} \xi(x_i) = 1,$$

and consider a standard linear program which maximizes $t$ under these constraints. Denote $t^{(k)}$ and $\xi^{(k)}$ those values of $t$ and of $\xi$ where the maximum of $t$ is attained. Define $\Xi_{k+1} = \Xi_k \cup \{\xi^{(k)}\}$ and repeat the linear program with $\Xi_{k+1}$ instead of $\Xi_k$.

To make this operational, we also require a stopping rule. We have that

$$t^{(k)} = \max_{\xi \in \Xi^\epsilon} \min_{\mu \in \Xi_k} \left\{ a(\mu) + \sum_{i=1}^{N} b_i(\mu) \, \xi(x_i) \right\}$$

$$\geq \max_{\xi \in \Xi^\epsilon} \min_{\mu \in \Xi^\epsilon} \left\{ a(\mu) + \sum_{i=1}^{N} b_i(\mu) \, \xi(x_i) \right\}$$

$$= \max_{\xi \in \Xi^\epsilon} \Phi\{M(\xi)\} \geq \max_{\xi \in \Xi_k} \Phi\{M(\xi)\}.$$

Hence, if we have a small number $\delta > 0$, such that $t^{(k)} - \max_{\xi \in \Xi_k} \Phi\{M(\xi)\} < \delta$, then

$$\max_{\xi \in \Xi_k} \Phi\{M(\xi)\} > \max_{\xi \in \Xi^\epsilon} \Phi\{M(\xi)\} - \delta$$

and $\arg\max_{\xi \in \Xi_k} \Phi\{M(\xi)\}$ is the nearly optimum design. Evidently, $\max_{\xi \in \Xi_k} \Phi\{M(\xi)\}$ can be simply computed iteratively,

$$\max_{\xi \in \Xi_k} \Phi\{M(\xi)\} = \max \left[ \max_{\xi \in \Xi_{k-1}} \Phi\{M(\xi)\}, \; \Phi\left\{ M\left(\xi^{(k)}\right) \right\} \right].$$

### 2.5. Equivalence theorem

In this section, we shall consider only global optimality criteria having a gradient $\nabla_M \Phi(M)$ such as the D- and A-criterion. We shall use an abbreviated notation writing

$$h(x, \xi) = \{T(\xi)\}_{x,.} \, G(\xi) \, \{T^{\mathrm{T}}(\xi)\}_{.,x},$$

where

$$G(\xi) = F\left[\nabla_M \Phi\{M(\xi)\}\right] F^{\mathrm{T}},$$

$$T(\xi) = \left[(C - \kappa I) \, \mathrm{diag}\{\xi(\cdot)\} + \frac{\kappa}{n} I\right]^{-1}.$$

**Theorem 2.2.** *Let $\Phi$ be a global, concave and nonnegative optimality criterion, zero on singular matrices, and having a gradient. A design measure $\bar{\xi} \in \Xi$ with a nonsingular $M(\bar{\xi})$ is $\Phi$-optimal in model (2) if and only if for every $n$-tuple $z_1, \ldots, z_n$ of points from $\mathcal{X}$ we have*

$$\frac{1}{n} \sum_{i=1}^{n} h(z_i, \bar{\xi}) \leq \sum_{x \in \mathcal{X}} \bar{\xi}(x) h(x, \bar{\xi}). \tag{7}$$

*Moreover, if $\bar{\xi}$ satisfies the inequality*

$$\frac{1}{n} \sum_{i=1}^{n} h(z_i, \bar{\xi}) \leq \sum_{x \in \mathcal{X}} \bar{\xi}(x) h(x, \bar{\xi}) + \delta \tag{8}$$

*for some $\delta > 0$ and any $z_1, \ldots, z_n \in \mathcal{X}$, we have that*

$$\Phi\{M(\bar{\xi})\} \geq \max_{\xi \in \Xi} \Phi\{M(\xi)\} - \frac{\kappa\delta}{n}. \tag{9}$$

The full proof is given in the appendix.

**Remark 2.1.** The numerical exploitation of this theorem is straightforward:

For a fixed design $\bar{\xi} \in \Xi$ and for any $x \in \mathcal{X}$ compute the number $h(x, \bar{\xi})$ and compute the sum $d(\bar{\xi}) = n \sum_{x \in \mathcal{X}} \bar{\xi}(x) h(x, \bar{\xi})$. Order the points of the set $\mathcal{X}$ so that $h(x_1, \bar{\xi}) \geq \cdots \geq h(x_N, \bar{\xi})$. Then (7) is equivalent to the inequality

$$\sum_{i=1}^{n} h(x_i, \bar{\xi}) \leq d(\bar{\xi}),$$

and (8) is equivalent to the inequality

$$\sum_{i=1}^{n} h(x_i, \bar{\xi}) \leq d(\bar{\xi}) + n\delta.$$

Note that the possibility to formulate an equivalence theorem due to concavity means that we can provide a rather close upper bound for the performance of all exact $n$-point designs. Thus we can calibrate the performance of all existing methods by this bound, as will be done in the next section.

## 3. Examples

### 3.1. General considerations

To check our ideas, we have computed designs for numerous examples from the literature, particularly those already presented and compared in [9]. In most of these examples, all compared methods did reasonably well, so for illustration

we have just selected four, which we consider representative and for which we give details below.

The ultimate goal in all our examples is to find an exact design in model (1) and to judge its quality. As for all measure-based design approaches, we therefore require a method to convert the calculated measure in model (2) to a discrete set of points from the design space $\mathcal{X}$. The endpoints plus quantiles of their measure are typically used in [6] in their one-dimensional examples. We will thus for comparison purposes use the same method for our approach, but there is no simple way to extend this to higher dimensions. Instead, in the spirit of the random design strategies defined in [28], one can perform random sampling according to the measure multiple times and select the sampled design with the best criterion value. We will denote this approach by R-VN (random sampling virtual noise). In our experience, this strategy has led to even more efficient designs in all cases, and we apply this approach using 100 samples whenever the design dimension is greater than one. We furthermore provide the results for the one-dimensional cases in the appendix, which also contains some additional examples.

In the following examples, we will compare our method based on the optimal measure for the virtual noise representation to the method from [6] as well as to the variant of the algorithm of [3] put forward by [8] based on approximate sensitivity functions, henceforth denoted by BKSF. An algorithmic description of the latter can be found in the appendix. We refrain from reporting results from the original algorithm which is tantamount to an exchange method based on direct comparisons of criterion values, as most of the time they gave very similar results. For comparison purposes, we also provide the median performance out of 100 randomly chosen exact designs of the respective size and denote this approach by R-UNIF. A variant of this approach always includes the endpoints and is denoted by R-UNIF+EP.

For all methods except the one from [6], the design space to search over was discretized to an $N = 101$-point design grid. For small $n$ we performed an exhaustive search of all $n$-point combinations over the design grid to obtain the true optimum exact design. We report those points in the tables as well, where we abbreviate this approach by EXS (exhaustive search). We consider two different quantile-based ways how to obtain the exact $n$-point designs for our method. The first is to select the quantiles $1/(n+1), \ldots, n/(n+1)$ of the design grid with respect to the design measure. The second way is to always include the endpoints of the design grid and select the quantiles $1/(n-1), \ldots, (n-2)/(n-1)$ of the remaining mass over the design grid. We denote these two methods by Q-VN and Q-VN+EP (quantile virtual noise plus endpoints), respectively. The second approach is similar to the approach of [6], which we will denote by Q-DPZ+EP. The design measures [6] consider usually have discrete masses at the endpoints but are continuous elsewhere, so [6] always include the endpoints and select the quantiles for the design points in between with respect to the continuous measure.

From Theorem 2.2 it follows that $\Phi\left(M_{\mathcal{T}}\right)$ for any realizable $n$-point design $\mathcal{T}$ is bounded from above by $\Phi\left\{M\left(\bar{\xi}\right)\right\}$ for the $\Phi$-optimal design measure $\bar{\xi}$,

where $M(\xi)$ is given by (3). Therefore, we report the efficiencies with respect to $\Phi\left\{M\big(\bar{\xi}\big)\right\}$, where $\bar{\xi}$ is found by applying the linear programming algorithm. Let $\Xi_k$ be the finite set of design measures determining the set of linear constraints in step $k$ of the linear programming algorithm and let

$$t^{(k)} = \max_{\xi \in \Xi^{\epsilon}} \min_{\mu \in \Xi_k} \left\{ a(\mu) + \sum_{i=1}^{N} b_i(\mu)\, \xi(x_i) \right\},$$

see Section 2.4. Denote $\Phi^{(k)} = \max_{\xi \in \Xi_k} \Phi\left\{M(\xi)\right\}$. In our programs, we stop the linear programming algorithm if $\frac{t^{(k)} - \Phi^{(k)}}{\Phi^{(k)}} \leq 10^{-4}$. We set $\epsilon = 10^{-6}$ in all our examples to avoid numerical problems. To solve the linear programming problems, we use the package `lpSolve` ([2]) from the software `R` ([26]). In all our examples we also choose $\kappa$ to be rounded down from $\lambda_{\min}(C)$ to two significant digits.

In all the figures we present, the left panel displays the discrete measure obtained by running the linear programming algorithm for the virtual noise representation for a particular $n$. The right panel shows the D-efficiencies of exact designs obtained by various methods up to $n = 20$. The methods depicted are: Q-VN (solid red line with squares), Q-VN+EP (dashed green line with large dots), Q-DPZ+EP (long-dashed blue line with triangles), BKSF (dotted black line with diamonds), and reporting the median efficiency for R-UNIF without (dot-dashed purple line with large dots) and with fixed endpoints (long-short-dashed gray line with small dots).

### 3.2. Example 1: a classic one-parameter model

This example has originally been considered by [25]. It is used by [6] to illustrate the efficiency of their method. It is a one-parameter model given by

$$
\begin{aligned}
f(x) &= 1 + 0.5\sin(2\pi x), & x \in [1,2], \\
\mathrm{cov}\left\{\varepsilon(x), \varepsilon(x')\right\} &= \begin{cases} x^2 x' & x \leq x' \\ x(x')^2 & x > x', \end{cases} & \lambda_{\min}(C) = 0.00276.
\end{aligned}
$$

For the D-criterion and $n = 4$, the selected design points and D-efficiencies for the various methods are given in Table 1. While the algorithm using the approximate sensitivity function from [8] comes close to the result of the exhaustive search and to 90% of the upper bound, the measure-based methods are considerably worse. However, when we increase the sample size (see Fig. 1), their performance improves and slowly approaches the bound.

Note that in contrast to the method by [6] our method does not require the existence of the continuous best linear unbiased estimator. To illustrate this, we modify the example by using the kernel $\mathrm{cov}\left\{\varepsilon(x), \varepsilon(x')\right\} = \min(x, x')^2$ $\{3\max(x, x') - \min(x, x')\}/6$ instead, which is the once continuously differentiable kernel for the integrated Brownian motion error process. According to [7], the continuous best linear unbiased estimator therefore has to incorporate

TABLE 1
*Optimal designs and D-efficiencies for Example 1*

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | D-eff |
|---|---|---|---|---|---|
| Q-VN | 1.11 | 1.22 | 1.37 | 1.74 | 0.8517 |
| Q-VN+EP | 1.00 | 1.20 | 1.55 | 2.00 | 0.7450 |
| Q-DPZ+EP | 1.00 | 1.28 | 1.69 | 2.00 | 0.8456 |
| R-UNIF | | | | | 0.7440 |
| R-UNIF+EP | | | | | 0.5591 |
| BKSF | 1.19 | 1.67 | 1.79 | 2.00 | 0.9076 |
| EXS | 1.22 | 1.66 | 1.79 | 2.00 | 0.9158 |
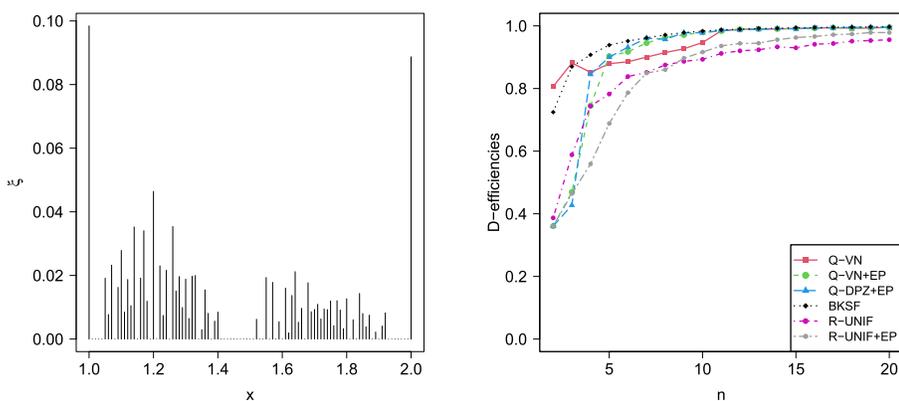


FIG 1. *Our measure (left panel) and efficiencies (based on our bound) versus sample size (right panel) for Example 1.*

TABLE 2
*Optimal designs and D-efficiencies for the modified Example 1*

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | D-eff |
|---|---|---|---|---|---|
| Q-VN | 1.00 | 1.01 | 1.39 | 1.53 | 0.4933 |
| Q-VN+EP | 1.00 | 1.22 | 1.53 | 2.00 | 0.7329 |
| R-UNIF | | | | | 0.4724 |
| R-UNIF+EP | | | | | 0.6608 |
| BKSF | 1.00 | 1.39 | 1.80 | 2.00 | 0.8042 |
| EXS | 1.00 | 1.23 | 1.75 | 2.00 | 0.9715 |

information from the first derivatives of the error process to be estimable. Since this kernel is now much smoother, the minimum eigenvalue for our design grid is rather small, namely $\lambda_{\min}(C) = 2.0854 \cdot 10^{-8}$. However, we did not observe any numerical issues due to that fact.

Our algorithm puts a high amount of mass at the two lowest points in the design grid, see Fig. 2. This can be interpreted as the algorithm trying to obtain information about the first derivative at the lower bound.

FIG 2. *Our measure (left panel) and efficiencies versus sample size (right panel) for the modified Example 1.*

### 3.3. Example 2: a multiparameter case

The next example is taken from Section 3.6 of [6]. The specifications of this four-parameter model are

$$
\begin{aligned}
f^{\mathrm{T}}(x) &= \left(1, x, x^2, x^3\right), \quad x \in [1, 2], \\
\operatorname{cov}\left\{\varepsilon(x), \varepsilon(x')\right\} &= \min\left(x, x'\right), \qquad \lambda_{\min}(C) = 0.0025.
\end{aligned}
$$

We again consider the D-criterion, so for $n = 5$, Table 3 contains the selected design points and D-efficiencies. This is one of the rare cases in which our method does even slightly better than the algorithm using the approximate sensitivity function from [8]. All methods provide considerable improvements over uniform random designs, even when the endpoints in the uniform random designs are always included.

TABLE 3
*Optimal designs and D-efficiencies for Example 2*

|           | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | D-eff  |
|-----------|-------|-------|-------|-------|-------|--------|
| Q-VN      | 1.00  | 1.15  | 1.51  | 1.84  | 2.00  | 0.9229 |
| Q-VN+EP   | 1.00  | 1.18  | 1.51  | 1.82  | 2.00  | 0.9297 |
| Q-DPZ+EP  | 1.00  | 1.14  | 1.33  | 1.60  | 2.00  | 0.8554 |
| R-UNIF    |       |       |       |       |       | 0.2910 |
| R-UNIF+EP |       |       |       |       |       | 0.7813 |
| BKSF      | 1.00  | 1.16  | 1.46  | 1.83  | 2.00  | 0.9270 |
| EXS       | 1.00  | 1.21  | 1.61  | 1.84  | 2.00  | 0.9308 |

### 3.4. Example 3: a real application

We consider a real-world application inspired by the example used throughout [14]. There a dataset with temperature and rainfall measurements gathered

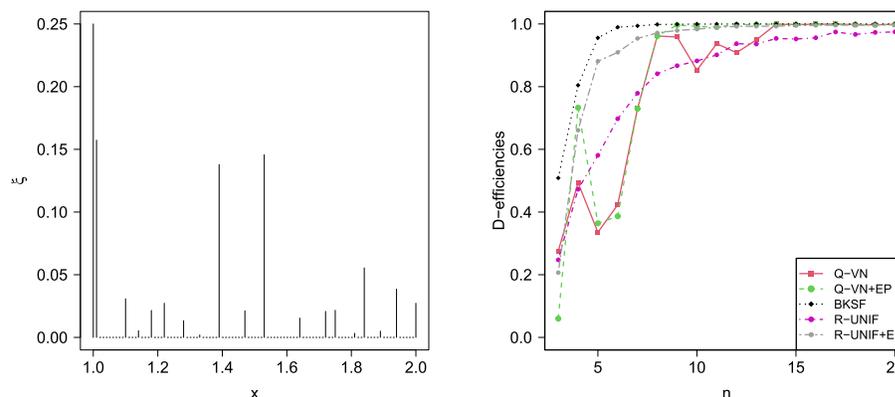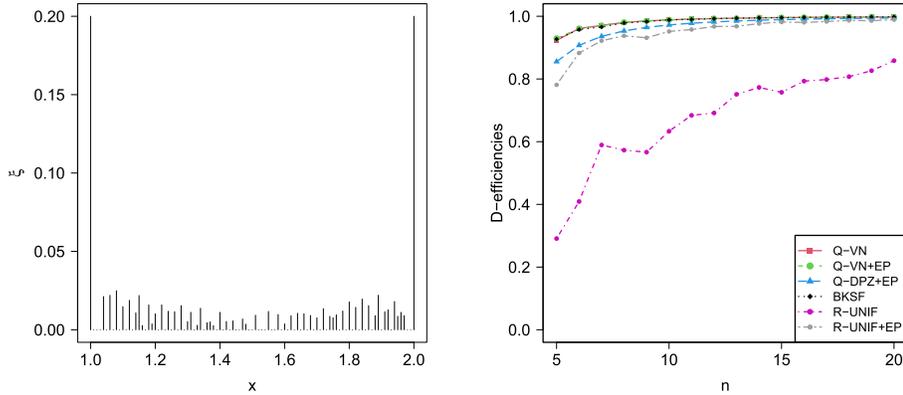FIG 3. *Our measure (left panel) and efficiencies versus sample size (right panel) for Example 2.*

from 37 weather stations in the Austrian state of Upper Austria during the years 1994 – 2009 was considered. Given this information, the goal was to find optimal designs for adding or reorganizing stations to optimize various aspects of future data collection.

As for Fig. 1.4 of [14] we employ the rainfall data from July 1994 and obtain the kriging estimates for an exponential kernel. The response function is assumed to be a plane, that is, it is linear in the coordinates. For our example, we assume that the parameters of the kernel function are known and set them to the kriging estimates. The model we consider is therefore

$$
\begin{aligned}
f^{\mathrm{T}}(x) &= (1, x_1, x_2), \qquad x = (x_1, x_2)^{\mathrm{T}}, \\
\mathrm{cov}\left\{\varepsilon(x), \varepsilon(x')\right\} &= 1756.65 \cdot \exp\left(\|x - x'\|_2 \big/ 40792.35\right).
\end{aligned}
$$

The design set is composed of the centroids of the $N = 445$ Upper Austrian municipalities, for which $\lambda_{\min}(C) = 42.15$. We set $n = 36$ because this was the number of active weather stations in July 1994. Our hypothetical objective is therefore to reorganize the whole weather station network in order to most efficiently estimate the response function parameters given the assumed kernel function. We use the D-criterion as in the previous examples.

The left panel of Fig. 4 displays the optimal measure obtained for the virtual noise representation when $n = 36$. The majority of the mass is concentrated on the borders. The right panel shows the optimal 36-point design we obtain by selecting the design with the highest criterion value out of 100 random designs sampled according to the measure depicted in the left panel.

The D-efficiencies for different methods used to construct the optimal 36-point exact design are given in Table 4. In addition to the previously mentioned approaches, we also consider a coffee-house design ([17]), which is a simple incremental design strategy aimed at obtaining designs close to space-filling maximin designs. We also compute the efficiency of the actual weather station network.
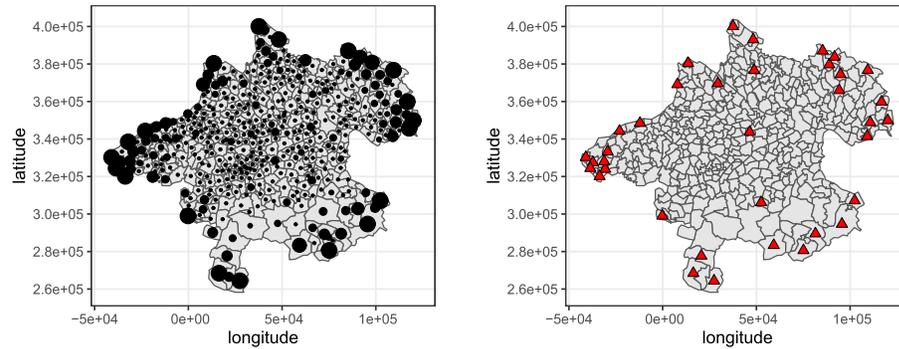
FIG 4. *Our measure (left panel) and exact 36-point design found by taking the optimal design from 100 random draws from the virtual noise measure (right panel) for Example 3.*

As in the other examples, all the efficiencies are reported with respect to our bound. The appendix contains the efficiencies for additional values of $n$ from $n = 4$ to $n = 40$ as well as maps depicting the coffee-house design for $n = 36$ and the actual weather station network.

TABLE 4
*D-efficiencies for Example 3*

|  | D-eff |
|---|---|
| R-VN (highest efficiency) | 0.9915 |
| R-VN (median efficiency) | 0.9702 |
| R-UNIF (highest efficiency) | 0.8405 |
| R-UNIF (median efficiency) | 0.6689 |
| BKSF | 0.9965 |
| Coffee-house | 0.9463 |
| Actual stations | 0.7754 |

## 4. Discussion

The present paper complements and in a certain sense completes the research in [16]. The virtual noise approach now exhibits the convexity property it previously lacked. The formulated equivalence theorem allows for calculating a general upper bound for designs in Gaussian process regression, a key methodology in many application fields. This upper bound for the first time offers the possibility to calibrate and scale other design methods proposed in the literature.

Furthermore, our "importance measures" can be used to directly produce exact designs, be it by taking quantiles or by randomly sampling from them. However, typically for all the examples considered, also those not reported here, the adaptation of [3] using the approximate sensitivity function from [8] performed slightly better than taking the quantiles from the measure given by the

virtual noise representation, which in turn was better than the method suggested by [6]. Therefore, for practical purposes a combined approach using the approximate sensitivity function for finding designs and the upper bound derived from the virtual noise representation for evaluating them seems to be the most suitable avenue. Even when one prefers not to use an optimal design, but a classical one such as fractional factorial, orthogonal array, latin hypercube, etc. our bound can be used to judge how much efficiency would be sacrificed.

## Appendix A: Proof of Theorem 2.1 on concavity

### A.1. Preliminaries

For any $u \in \mathbb{R}^N$ define the function

$$\xi \in \Xi \to \gamma_u(\xi) = \begin{cases} u^{\mathrm{T}} \{C + W(\xi)\}^{-1} u & \text{if } \mathrm{supp}(\xi) = \mathcal{X} \\ u'^{\mathrm{T}} \{C' + W'(\xi)\}^{-1} u' & \text{if } \mathrm{supp}(\xi) \neq \mathcal{X} \end{cases},$$

where $C', W'(\xi), u'$ are submatrices of $C, W(\xi), u$, respectively with rows and/or columns corresponding to the points of $\mathrm{supp}(\xi)$.

We prove the following lemma.

**Lemma A.1.** *The function $\xi \in \Xi \to \gamma_u(\xi)$ is concave.*

*Proof.* We start by proving the concavity on $\Xi_+ = \{\xi \in \Xi : \mathrm{supp}(\xi) = \mathcal{X}\}$. Consider an auxiliary function

$$\psi_u(\alpha) : \alpha \in (0, 1) \to \gamma_u(\xi_\alpha),$$

with $\xi_\alpha(x) = (1-\alpha)\xi(x) + \alpha\mu(x); \quad \xi, \mu \in \Xi_+, \alpha \in (0, 1)$. The concavity of $\gamma_u(\xi)$ on $\Xi_+$ will be proven if we show that

$$\frac{d^2\psi_u(\alpha)}{d\alpha^2} \leq 0$$

for any $\xi, \mu \in \Xi_+$ and any $\alpha \in (0, 1)$.

We can write

$$\psi_u(\alpha) = u^{\mathrm{T}} \{H(\xi_\alpha)\}^{-1} u,$$

with

$$H(\xi_\alpha) = \left[ (C - \kappa I) + \frac{\kappa}{n} \operatorname{diag} \left\{ \frac{1}{\xi_\alpha(\cdot)} \right\} \right].$$

Notice that the matrix $(C - \kappa I)$ is positive definite, hence also $H(\xi_\alpha)$ is positive definite.

From the rule for the derivative of an inverse matrix we obtain

$$\begin{aligned} \frac{d\psi_u(\alpha)}{d\alpha} &= -\frac{\kappa}{n} u^{\mathrm{T}} H^{-1}(\xi_\alpha) \left\{ \frac{d}{d\alpha} \operatorname{diag} \left( \frac{1}{\xi_\alpha(\cdot)} \right) \right\} H^{-1}(\xi_\alpha) u \\ &= \frac{\kappa}{n} u^{\mathrm{T}} H^{-1}(\xi_\alpha) \operatorname{diag} \left\{ \frac{1}{\xi_\alpha^2(\cdot)} \right\} \operatorname{diag} \{\mu(\cdot) - \xi(\cdot)\} H^{-1}(\xi_\alpha) u \end{aligned}$$

$$= \frac{\kappa}{n} u^{\mathrm{T}} T^{\mathrm{T}}(\xi_\alpha) \operatorname{diag}\{\mu(\cdot) - \xi(\cdot)\} \, T(\xi_\alpha) u,$$

where

$$T(\xi_\alpha) = \left[ (C - \kappa I) \operatorname{diag}\{\xi_\alpha(\cdot)\} + \frac{\kappa}{n} I \right]^{-1}.$$

Hence

$$
\begin{aligned}
&\frac{d^2 \psi_u(\alpha)}{d\alpha^2} \\
={} &-\frac{\kappa}{n} u^{\mathrm{T}} T^{\mathrm{T}}(\xi_\alpha) \operatorname{diag}\{\mu(\cdot) - \xi(\cdot)\} \, (C - \kappa I) \, T^{\mathrm{T}}(\xi_\alpha) \\
&\times \operatorname{diag}\{\mu(\cdot) - \xi(\cdot)\} \, T(\xi_\alpha) u \\
&-\frac{\kappa}{n} u^{\mathrm{T}} T^{\mathrm{T}}(\xi_\alpha) \operatorname{diag}\{\mu(\cdot) - \xi(\cdot)\} \\
&\times T(\xi_\alpha) \, (C - \kappa I) \operatorname{diag}\{\mu(\cdot) - \xi(\cdot)\} \, T(\xi_\alpha) u,
\end{aligned}
$$

which is nonpositive since the matrix

$$
\begin{aligned}
(C - \kappa I) \, T^{\mathrm{T}}(\xi_\alpha) &= (C - \kappa I) \left[ \operatorname{diag}\{\xi_\alpha(\cdot)\} (C - \kappa I) + \frac{\kappa}{n} I \right]^{-1} \\
&= \left[ \operatorname{diag}\{\xi_\alpha(\cdot)\} + \frac{\kappa}{n} (C - \kappa I)^{-1} \right]^{-1}
\end{aligned}
$$

is symmetric positive definite.

To finish the proof take arbitrary $\xi, \mu \in \Xi$ and sequences $\{\xi_k\}, \{\mu_k\}$ of elements of $\Xi_+$ converging to $\xi$ and $\mu$. From the inequality

$$\gamma_u\{(1 - \alpha)\xi_k + \alpha\mu_k\} \geq (1 - \alpha)\gamma_u(\xi_k) + \alpha\gamma_u(\mu_k),$$

valid for each $k$, we obtain by the continuity of $\xi \in \Xi \to \gamma_u(\xi)$, see Lemma F.1, that the same inequality also holds for $\xi$ and $\mu$. ☐

### A.2.  Proof of Theorem 2.1

Since for any $t \in \mathbb{R}^p$. we have

$$t^{\mathrm{T}} M(\xi) t = \gamma_{Ft}(\xi),$$

From Lemma A.1 it follows that for any $\xi, \mu \in \Xi$, $\alpha \in (0, 1)$, $\kappa < \lambda_{\min}(C)$

$$t^{\mathrm{T}} M\{(1 - \alpha)\xi + \alpha\mu\} t \geq (1 - \alpha) t^{\mathrm{T}} M(\xi) t + \alpha t^{\mathrm{T}} M(\mu) t. \tag{10}$$

It is easy to see that the analytical expressions defining $M(\xi)$ in (3) and (4) are formally meaningful for any $\kappa \in (0, \infty)$. They are even continuous in $\kappa$ on this interval. Indeed, by standard matrix manipulations we can even obtain the derivative of $M(\xi)$ with respect to $\kappa$, hence also the required continuity.

Consider now a sequence $\{\kappa_n\}_{n=1}^\infty$ such that $\kappa_n \in (0, \lambda_{\min}(C))$ and $\lim_{n \to \infty} \kappa_n = \lambda_{\min}(C)$. From (10) valid for any $\kappa_n$ and from the above continuity it follows that (10) holds also for $\kappa = \lambda_{\min}(C)$.

Summarizing, for any $\kappa \leq \lambda_{\min}(C)$, any $\xi, \mu \in \Xi$ and any $\alpha \in (0,1)$ we have the inequality

$$M\{(1-\alpha)\xi + \alpha\mu\} \geq (1-\alpha)M(\xi) + \alpha M(\mu)$$

in the Löwner ordering. From the monotonicity of $\Phi$ we have

$$\Phi\left[M\{(1-\alpha)\xi + \alpha\mu\}\right] \geq \Phi\left\{(1-\alpha)M(\xi) + \alpha M(\mu)\right\}$$

and from the concavity of $\Phi$ we have

$$\Phi\left\{(1-\alpha)M(\xi) + \alpha M(\mu)\right\} \geq (1-\alpha)\Phi\{M(\xi)\} + \alpha\Phi\{M(\mu)\}. \quad \square$$

## Appendix B: Proof of the equivalence theorem (Theorem 2.2)

Since the function

$$\xi \in \Xi \to \Phi\{M(\xi)\}$$

is concave, the design $\bar{\xi}$ is $\Phi$-optimal if and only if for every $\mu \in \Xi$ we have, see Lemma F.2,

$$\pi\left(\bar{\xi}, \mu\right) = \lim_{\alpha \to 0^+} \frac{\Phi\left\{M\left(\xi_\alpha^\mu\right)\right\} - \Phi\left\{M\left(\bar{\xi}\right)\right\}}{\alpha} \leq 0,$$

where we used the notation

$$\xi_\alpha^\mu(x) = (1-\alpha)\bar{\xi}(x) + \alpha\mu(x).$$

We have

$$
\begin{aligned}
\lim_{\alpha \to 0^+} \frac{\Phi\left\{M\left(\xi_\alpha^\mu\right)\right\} - \Phi\left\{M\left(\bar{\xi}\right)\right\}}{\alpha} &= \sum_{i,j=1}^{p} \lim_{\alpha \to 0^+} \frac{\Phi\left\{M\left(\xi_\alpha^\mu\right)\right\} - \Phi\left\{M\left(\bar{\xi}\right)\right\}}{M_{ij}\left(\xi_\alpha^\mu\right) - M_{ij}\left(\bar{\xi}\right)} \times \\
&\qquad \times \lim_{\alpha \to 0^+} \frac{M_{ij}\left(\xi_\alpha^\mu\right) - M_{ij}\left(\bar{\xi}\right)}{\alpha} \\
&= \sum_{i,j=1}^{p} \lim_{\xi \to \bar{\xi}} \frac{\Phi\left\{M\left(\xi\right)\right\} - \Phi\left\{M\left(\bar{\xi}\right)\right\}}{M_{ij}\left(\xi\right) - M_{ij}\left(\bar{\xi}\right)} \times \\
&\qquad \times \lim_{\alpha \to 0^+} \frac{M_{ij}\left(\xi_\alpha^\mu\right) - M_{ij}\left(\bar{\xi}\right)}{\alpha},
\end{aligned}
$$

since $\lim_{\alpha \to 0^+} \xi_\alpha^\mu = \bar{\xi}$. Therefore, we obtain

$$
\begin{aligned}
&\lim_{\alpha \to 0^+} \frac{\Phi\left\{M\left(\xi_\alpha^\mu\right)\right\} - \Phi\left\{M\left(\bar{\xi}\right)\right\}}{\alpha} \\
&\qquad = \operatorname{tr}\left[\nabla\Phi\left(M\right)\big|_{M=M\left(\bar{\xi}\right)} \lim_{\alpha \to 0^+} \frac{M\left(\xi_\alpha^\mu\right) - M\left(\bar{\xi}\right)}{\alpha}\right]
\end{aligned}
$$

$$= \text{tr} \left[ \nabla \Phi\left(M\right)\big|_{M=M\left(\bar{\xi}\right)} \frac{\kappa}{n} F^{\mathrm{T}} T^{\mathrm{T}}\left(\bar{\xi}\right) \text{diag}\left(\mu - \bar{\xi}\right) T\left(\bar{\xi}\right) F \right]$$

$$= \frac{\kappa}{n} \sum_{x \in \mathcal{X}} \left\{ \mu(x) - \bar{\xi}(x) \right\} h(x, \bar{\xi}),$$

as follows from (19) and (20) in Lemma F.3 and from the definition of $h(x, \bar{\xi})$ at the beginning of Section 2.5, just before Theorem 2.2. Hence we finally obtain

$$\pi\left(\bar{\xi}, \mu\right) = \frac{\kappa}{n} \sum_{x \in \mathcal{X}} \left\{ \mu(x) - \bar{\xi}(x) \right\} h(x, \bar{\xi}). \tag{11}$$

Therefore, $\bar{\xi} \in \Xi$ is $\Phi$-optimal if and only if for every $\mu \in \Xi$ we have

$$\sum_{x \in \mathcal{X}} \mu(x) h(x, \bar{\xi}) \leq \sum_{x \in \mathcal{X}} \bar{\xi}(x) h(x, \bar{\xi}). \tag{12}$$

Since this expression is linear in $\mu$ and since the convex hull of the set of all realizable $n$-point designs is the convex set $\Xi$, we do not need to consider all designs $\mu$ from $\Xi$, but just the realizable $n$-point designs. Hence the validity of the inequality (7) in Theorem 2.2 for all such $n$-point measures is equivalent to the validity of (12) for all $\mu \in \Xi$.

From (11) it follows that the inequalities (8) in Theorem 2.2 imply that for every $\mu \in \Xi$ we have

$$\pi\left(\bar{\xi}, \mu\right) \leq \frac{\kappa \delta}{n}.$$

According to Lemma F.2 it follows that the inequality (9) holds as well. □

## Appendix C: Adaptation of the algorithm from [8] for A-optimality

### *C.1. Sensitivity function for the D-criterion*

Consider the linear regression model with i.i.d. errors,

$$y(x) = f(x)^{\mathrm{T}} \theta + \varepsilon(x),$$

where $E\{\varepsilon(x)\} = 0$, $\text{var}\{\varepsilon(x)\} = \sigma_\varepsilon^2$, and $E\{\varepsilon(x)\varepsilon(x')\} = 0 \ \forall x, x' \in \mathcal{X}$.

The information matrix for a design $\xi \in \Xi$ is

$$M_\xi = \frac{1}{\sigma_\varepsilon^2} \sum_{x \in \mathcal{X}} \xi(x) f(x) f(x)^{\mathrm{T}}.$$

We also consider a one-point design at $\bar{x} \in \mathcal{X}$ denoted by $\xi_{\bar{x}}$. Its information matrix is

$$M_{\xi_{\bar{x}}} = \frac{1}{\sigma_\varepsilon^2} f(\bar{x}) f(\bar{x})^{\mathrm{T}}.$$

Let the information matrix of the convex combination $\xi_{\bar{x}}^\alpha = (1 - \alpha)\xi + \alpha\xi_{\bar{x}}$ of those two designs be denoted by

$$M_\alpha = (1 - \alpha) M_\xi + \alpha M_{\xi_{\bar{x}}}.$$

For some concave criterion function $\Phi(M)$, the directional derivative of $\Phi(\cdot)$ at the design $\xi$ in the direction of $\xi_{\bar{x}}$, the so-called sensitivity function, is

$$
\begin{aligned}
\phi(\bar{x}; \xi) &= \lim_{\alpha \to 0^+} \frac{\Phi(M_\alpha) - \Phi(M_\xi)}{\alpha} \\
&= \lim_{\alpha \to 0^+} \operatorname{tr}\left\{ \nabla_M \Phi(M_\alpha) \frac{\partial M_\alpha}{\partial \alpha} \right\} \\
&= \lim_{\alpha \to 0^+} \operatorname{tr}\left\{ \nabla_M \Phi(M_\alpha) (M_{\xi_{\bar{x}}} - M_\xi) \right\} \\
&= \operatorname{tr}\left\{ \nabla_M \Phi(M_\xi) (M_{\xi_{\bar{x}}} - M_\xi) \right\} \\
&= \frac{1}{\sigma_\varepsilon^2} f(\bar{x})^{\mathrm{T}} \nabla_M \Phi(M_\xi) f(\bar{x}) - \operatorname{tr}\left\{ \nabla_M \Phi(M_\xi) M_\xi \right\}.
\end{aligned}
\tag{13}
$$

In the case of **D-optimality**, we prefer here the criterion function $\Phi_D(M) = \log \det(M)$, so $\nabla_M \Phi_D(M) = M^{-1}$. Plugging this into (13) yields

$$
\begin{aligned}
\phi_D(\bar{x}; \xi) &= \frac{1}{\sigma_\varepsilon^2} f(\bar{x})^{\mathrm{T}} M_\xi^{-1} f(\bar{x}) - \operatorname{tr}\left( M_\xi^{-1} M_\xi \right) \\
&= \frac{1}{\sigma_\varepsilon^2} f(\bar{x})^{\mathrm{T}} M_\xi^{-1} f(\bar{x}) - \dim(\theta).
\end{aligned}
\tag{14}
$$

Up to a constant, expression (14) is equal to

$$
\frac{E\left[ \{y(\bar{x}) - \hat{y}(\bar{x})\}^2 \right]}{\sigma_\varepsilon^2} = \frac{\sigma_\varepsilon^2 + f(\bar{x})^{\mathrm{T}} M_\xi^{-1} f(\bar{x})}{\sigma_\varepsilon^2} = 1 + \frac{f(\bar{x})^{\mathrm{T}} M_\xi^{-1} f(\bar{x})}{\sigma_\varepsilon^2}.
\tag{15}
$$

In the correlated case with no replications, [8] suggests to replace the numerator in (15) by the unconditional and the denominator in (15) by the conditional variance. Let $\mathcal{T} = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ denote an exact n-point design, let $C(\mathcal{T})$ be the $n \times n$ submatrix of the covariance matrix $C$ for the set of points in $\mathcal{T}$, and let $F(\mathcal{T})$ be the $n \times p$ design matrix with elements $F_{i,j}(\mathcal{T}) = f_j(x_i)$. Let $k(x, x') = \operatorname{cov}\{\varepsilon(x), \varepsilon(x')\}$ and for a point $\bar{x} \in \mathcal{X} \backslash \mathcal{T}$, let $k(\bar{x}, \mathcal{T}) = (k(\bar{x}, x_1), \ldots, k(\bar{x}, x_n))^{\mathrm{T}}$. In the correlated case, [8] therefore replaces expression (15) with

$$
\frac{E\left[ \{y(\bar{x}) - \hat{y}(\bar{x})\}^2 \right]}{\tilde{\sigma}^2(\bar{x})} = \frac{\tilde{\sigma}^2(\bar{x}) + \tilde{f}(\bar{x})^{\mathrm{T}} M_{\mathcal{T}}^{-1} \tilde{f}(\bar{x})}{\tilde{\sigma}^2(\bar{x})},
$$

where

$$
\tilde{\sigma}^2(\bar{x}) = k(\bar{x}, \bar{x}) - k(\bar{x}, \mathcal{T})^{\mathrm{T}} C^{-1}(\mathcal{T}) k(\bar{x}, \mathcal{T}),
$$

is the conditional variance, $\tilde{f}(\bar{x})$ is

$$
\tilde{f}(\bar{x}) = f(\bar{x}) - F(\mathcal{T})^{\mathrm{T}} C^{-1}(\mathcal{T}) k(\bar{x}, \mathcal{T}),
$$

and the information matrix is

$$
M_{\mathcal{T}} = F(\mathcal{T})^{\mathrm{T}} C^{-1}(\mathcal{T}) F(\mathcal{T}).
$$

It can be shown that for D-optimality this approximate sensitivity function is equal to the factor by which the determinant of the information matrix is increased by adding design point $\bar{x}$.

### C.2.  Sensitivity function for the A-criterion

In the case of **A-optimality**, we use the criterion $\Phi_A(M) = -\mathrm{tr}\left(M^{-1}\right)$. We therefore have $\nabla_M \Phi_A(M) = -\left(-M^{-2}\right)^{\mathrm{T}} = M^{-2}$ and the directional derivative is

$$\phi_A(\bar{x}; \xi) = \frac{1}{\sigma_\varepsilon^2} f(\bar{x})^{\mathrm{T}} M_\xi^{-2} f(\bar{x}) - \mathrm{tr}\left(M_\xi^{-1}\right).$$

In accordance with the case of D-optimality outlined above, we obtain the approximate sensitivity function by analogous replacements. Therefore, the sensitivity function we eventually use in some of our examples is

$$\tilde{\phi}_A(\bar{x}; \mathcal{T}) = \frac{1}{\tilde{\sigma}^2(\bar{x})} \tilde{f}(\bar{x})^{\mathrm{T}} M_{\mathcal{T}}^{-2} \tilde{f}(\bar{x}) - \mathrm{tr}\left(M_{\mathcal{T}}^{-1}\right).$$

## Appendix D:  Properties of the LP-algorithm for D-optimality

To see the details of the algorithm given in Section 2.4, let us take a look at the particular specification for D-optimality. Take $\Phi(M) = \{\det(M)\}^{1/p}$. Then for any nonsingular $M$

$$\nabla_M \Phi(M) = M^{-1} \cdot p^{-1} \{\det(M)\}^{1/p}.$$

According to (3), supposing that $F$ has full rank we have that $M^{-1}(\mu)$ exists for any $\mu \in \Xi_+$, and we also have

$$\frac{\partial M(\mu)}{\partial \mu(\bar{x})} = \frac{\kappa}{n} F^{\mathrm{T}} H^{-1}(\mu) \Gamma \left\{\mu^{-2}(\bar{x})\right\} H^{-1}(\mu) F,$$

with

$$H(\mu) = \left[(C - \kappa I) + \frac{\kappa}{n} \mathrm{diag}\left\{\mu^{-1}(\cdot)\right\}\right],$$

where $\mathrm{diag}\left\{\mu^{-1}(\cdot)\right\}$ is an $N \times N$ diagonal matrix with $\mu^{-1}(x)$ as diagonal elements, and $\Gamma\left\{\mu^{-2}(\bar{x})\right\}$ is an $N \times N$ matrix having zero elements everywhere up to the $\bar{x}$th diagonal position, where we have $\mu^{-2}(\bar{x})$. Thus, the expression (5) becomes

$$\left[\det\left\{F^{\mathrm{T}} H^{-1}(\mu) F\right\}\right]^{1/p} \times$$

$$\times \left[1 + \frac{\kappa}{np} \sum_{\bar{x} \in \mathcal{X}} \left\{H^{-1}(\mu)\right\}_{\bar{x},.} F \left\{F^{\mathrm{T}} H^{-1}(\mu) F\right\}^{-1} F^{\mathrm{T}} \left\{H^{-1}(\mu)\right\}_{.,\bar{x}} \frac{\xi(\bar{x}) - \mu(\bar{x})}{\mu^2(\bar{x})}\right].$$

**Appendix E:  Outline of algorithm from [8]**

Algorithm 2 contains the description of our implementation for a finite design grid of the BKSF-algorithm proposed in [8], which is used in our examples.

---

**Algorithm 2:** BKSF-algorithm for a finite design grid

---

**Input:** Design grid $\mathcal{X}$ of possible design points; design size $n$; initial
        design $\mathcal{T}_0$; (approximate) sensitivity function $\tilde{\phi}(x; \mathcal{T})$.
**Output:** Exact design $\mathcal{T}$ found by algorithm.

**1** abort = **false**;
**2** $\mathcal{T} = \mathcal{T}_0$;
**3** **while not** *abort* **do**
**4**    **for** $i = 1$ **to** $n$ **do**
**5**       Let $x_i$ be the $i^{\text{th}}$ design point in $\mathcal{T}$;
**6**       $\mathcal{T}_{-i} = \mathcal{T} \backslash \{x_i\}$;
**7**       $\texttt{s[i]} = \tilde{\phi}(x_i; \mathcal{T}_{-i})$;
**8**    **end for**
**9**    sens_drop = $\min(\mathbf{s})$;
**10**   Let $\texttt{k}$ be the index where $\texttt{s[k]} = $ sens_drop;
**11**   Let $\mathcal{C} = \mathcal{X} \backslash \mathcal{T}_{-k}$;
**12**   $m = \text{card}(\mathcal{C})$;
**13**   **for** $i = 1$ **to** $m$ **do**
**14**      Let $\bar{x}_i$ be the $i^{\text{th}}$ design point in the candidate set $\mathcal{C}$;
**15**      $\texttt{t[i]} = \tilde{\phi}(\bar{x}_i; \mathcal{T}_{-k})$;
**16**   **end for**
**17**   sens_add = $\max(\mathbf{t})$;
**18**   Let $\texttt{l}$ be the index where $\texttt{t[l]} = $ sens_add;
**19**   $\mathcal{T} = \mathcal{T}_{-k} \cup \bar{x}_l$;
**20**   **if** sens_add $-$ sens_drop $\leq 0$ **then**
**21**      abort = **true**;
**22**   **end if**
**23** **end while**

---

**Appendix F:  Miscellanea**

***F.1.  The continuity lemma***

**Lemma F.1.** *The function $\gamma_u(\xi)$ defined in Lemma A.1 is continuous on $\Xi$ and $M(\xi)$ is continuous on $\Xi$ as well.*

    *Proof.* Take $\xi_k \in \Xi$ such that $\lim_{k \to \infty} \xi_k = \xi$. The equality $\lim_{k \to \infty} \gamma_u(\xi_k) = \gamma_u(\xi)$ is evident if $\text{supp}(\xi) = \mathcal{X}$, or more generally if $\bigcup_{k_0=1}^{\infty} \bigcap_{k \geq k_0} \text{supp}(\xi_k) = \text{supp}(\xi)$. To clarify how to do the proof in the opposite case we suppose that $\text{supp}(\xi) = \mathcal{X}$, but $\bigcup_{k_0=1}^{\infty} \bigcap_{k \geq k_0} \text{supp}(\xi_k)$ has one point less than the set $\text{supp}(\xi)$,

say the point $x_1$. By a basic property of inverse matrices we have

$$\left[\{C + W(\xi_k)\}^{-1}\right]_{ij} = \frac{(-1)^{i+j} \det \{C + W(\xi_k)\}^{i,j}}{\det \{C + W(\xi_k)\}},$$
(16)

where $\{C + W(\xi_k)\}^{i,j}$ is the submatrix of $\{C + W(\xi_k)\}$ after omitting the $i$-th row and $j$-th column. For $k \to \infty$ (16) tends to zero for $i = 1$ or $j = 1$, but to a nonzero number when $i \neq 1$ and $j \neq 1$. This can be seen when using the definition of a determinant as a sum of products of the elements of the matrix. When $k \to \infty$, some terms of these sums in the numerator and the denominator converge to $\infty$ much slower than the others, so we can neglect them. We proceed similarly when $\bigcup_{k_0=1}^{\infty} \bigcap_{k \geq k_0} \operatorname{supp}(\xi_k) \cup \{x_1, x_2\} = \operatorname{supp}(\xi)$, etc.. $\qquad \square$

### F.2. Optimality lemma

For the reader's convenience we prove a statement from convex theory, see e.g. [24].

**Lemma F.2.** *If a function $\phi : \xi \in \Xi \to \phi(\xi) \in \mathbb{R}$ is concave, then for any $\bar{\xi}, \mu \in \Xi$ the function*

$$\alpha \in (0, 1) \to \frac{\phi \{(1 - \alpha)\bar{\xi} + \alpha\mu\} - \phi(\bar{\xi})}{\alpha}$$

*is nonincreasing. Hence for any $\bar{\xi} \in \Xi$ and any $\mu \in \Xi$ there exists the limit*

$$\lim_{\alpha \to 0^+} \frac{\phi \{(1 - \alpha)\bar{\xi} + \alpha\mu\} - \phi(\bar{\xi})}{\alpha} \equiv \pi \left(\bar{\xi}, \mu\right),$$

*and we have*

$$\bar{\xi} \in \arg \max_{\xi \in \Xi} \phi(\xi)$$

*if and only if*

$$\forall \mu \in \Xi \quad \pi \left(\bar{\xi}, \mu\right) \leq 0.$$
(17)

*If for some $\delta > 0$ and for every $\mu \in \Xi$ we have*

$$\pi \left(\bar{\xi}, \mu\right) \leq \delta,$$

*then $\phi \left(\bar{\xi}\right) \geq \max_{\xi \in \Xi} \phi(\xi) - \delta$.*

*Proof.* Take $0 < \alpha_1 < \alpha_2 < 1$. From the concavity of $\phi(\xi)$ we obtain

$$
\begin{aligned}
&\phi \{(1 - \alpha_1)\bar{\xi} + \alpha_1\mu\} - \phi(\bar{\xi}) \\
={}& \phi \left[\frac{\alpha_1}{\alpha_2} \{(1 - \alpha_2)\bar{\xi} + \alpha_2\mu\} + \left(1 - \frac{\alpha_1}{\alpha_2}\right) \bar{\xi}\right] - \phi(\bar{\xi}) \\
\geq{}& \frac{\alpha_1}{\alpha_2} \phi \{(1 - \alpha_2)\bar{\xi} + \alpha_2\mu\} + \left(1 - \frac{\alpha_1}{\alpha_2}\right) \phi(\bar{\xi}) - \phi(\bar{\xi})
\end{aligned}
$$

$$= \frac{\alpha_1}{\alpha_2} \left[ \phi \left\{ (1 - \alpha_2)\bar{\xi} + \alpha_2\mu \right\} - \phi(\bar{\xi}) \right].$$

We multiply this inequality by $1/\alpha_1$ to prove the first statement of the lemma.

If $\bar{\xi} \in \arg\max_{\xi \in \Xi} \phi(\xi)$, then for any $\mu$ and $\alpha$ we evidently have

$$\frac{\phi \left\{ (1 - \alpha)\bar{\xi} + \alpha\mu \right\} - \phi(\bar{\xi})}{\alpha} \leq 0$$

and this inequality does also hold for $\alpha \to 0^+$.

On the other hand, suppose that there is $\mu \in \Xi$ such that $\phi(\mu) > \phi(\bar{\xi})$. Then for every $\alpha \in (0, 1)$ we have

$$\frac{\phi \left\{ (1 - \alpha)\bar{\xi} + \alpha\mu \right\} - \phi(\bar{\xi})}{\alpha}$$
$$\geq \frac{(1 - \alpha)\phi(\bar{\xi}) + \alpha\phi(\mu) - \phi(\bar{\xi})}{\alpha}$$
$$= \phi(\mu) - \phi(\bar{\xi}) > 0,$$

and by taking the limit for $\alpha \to 0$ we see that (17) does not hold.

If

$$\forall_{\mu \in \mathcal{X}} \, \pi \left( \bar{\xi}, \mu \right) \leq \delta,$$

then from the concavity of $\phi(\xi)$ we obtain for every $\mu \in \Xi$

$$\phi(\mu) - \phi(\bar{\xi}) = \frac{(1 - \alpha)\phi(\bar{\xi}) + \alpha\phi(\mu) - \phi(\bar{\xi})}{\alpha}$$
$$\leq \frac{\phi\{(1 - \alpha)\bar{\xi} + \alpha\mu\} - \phi(\bar{\xi})}{\alpha} \leq \pi \left( \bar{\xi}, \mu \right) \leq \delta.$$

Hence $\phi(\bar{\xi}) \geq \max_{\mu \in \Xi} \phi(\mu) - \delta$. $\qquad\qquad\square$

## F.3. Information matrix derivative lemma

**Lemma F.3.** *The matrix*

$$L(\xi) \equiv F^{\mathrm{T}} \left[ diag \left\{ \xi \left( \cdot \right) \right\} (C - \kappa I) + \frac{\kappa}{n}I \right]^{-1} diag \left\{ \xi \left( \cdot \right) \right\} F \qquad (18)$$

*is well defined and continuous on $\Xi$.*

*For every $\xi \in \Xi$ we have $L(\xi) = M(\xi)$ with $M(\xi)$ defined in (3) and (4).*

*For every $\bar{\xi} \in \Xi$ and $\mu \in \Xi$ the right-hand limit*

$$\lim_{\alpha \to 0^+} \frac{L \left\{ (1 - \alpha)\bar{\xi} + \alpha\mu \right\} - L(\bar{\xi})}{\alpha}$$

*is well defined and it is a continuous function of $\bar{\xi}$ on the whole set $\Xi$. On the set $\Xi_+$ it is equal to the derivative $\frac{\partial L(\xi_\alpha^\mu)}{\partial \alpha}\big|_{\alpha=0} = \frac{\partial M(\xi_\alpha^\mu)}{\partial \alpha}\big|_{\alpha=0}$.*

*Proof.* We denote $Z\left(\xi\right) = \left[\mathrm{diag}\left\{\xi\left(\cdot\right)\right\}\left(C - \kappa I\right) + \frac{\kappa}{n}I\right]$. To see the correctness of the definition and the continuity of $L\left(\xi\right)$, we must prove the existence of the matrix $Z\left(\xi\right)^{-1}$ used in (18). Without loss of generality we can suppose that $\xi\left(x_i\right) > 0$ for $i = 1, \ldots, k$ and $\xi\left(x_i\right) = 0$ for $i = k+1, ..., N$, where $k \leq N$. Let us write $Z\left(\xi\right) = \begin{pmatrix} V & R \\ P & T \end{pmatrix}$, where $V$ is a $k \times k$ matrix and $T$ is a $(N-k) \times (N-k)$ matrix. Evidently $P = 0$. According to [11], Lemma 8.5.4, if the inverse matrices $V^{-1}$ and $T^{-1}$ exist, then the inverse of the partitioned matrix exists as well, and

$$\begin{pmatrix} V & R \\ 0 & T \end{pmatrix}^{-1} = \begin{pmatrix} V^{-1} & -V^{-1}RT^{-1} \\ 0 & T^{-1} \end{pmatrix}.$$

We can write

$$\mathrm{diag}\left\{\xi\left(\cdot\right)\right\} = \begin{pmatrix} \mathrm{diag}'\left\{\xi\left(\cdot\right)\right\} & 0 \\ 0 & 0 \end{pmatrix},$$

where the $k \times k$ matrix $\mathrm{diag}'\left\{\xi\left(\cdot\right)\right\}$ has the $k$ positive elements of $\xi$ on its diagonal. Consequently

$$\begin{aligned} V &= \mathrm{diag}'\left\{\xi\left(\cdot\right)\right\}\left(C' - \kappa I'\right) + \frac{\kappa}{n}I' \\ &= \mathrm{diag}'\left\{\xi\left(\cdot\right)\right\}\left[\left(C' - \kappa I'\right) + \frac{\kappa}{n}\mathrm{diag}'\left\{\xi^{-1}\left(\cdot\right)\right\}\right], \end{aligned}$$

where the prime denotes the $k \times k$ submatrix of the corresponding matrix. The symmetric matrix

$$\left[\left(C' - \kappa I'\right) + \frac{\kappa}{n}\mathrm{diag}'\left\{\xi^{-1}\left(\cdot\right)\right\}\right]$$

is positive definite, hence its inverse exists and $V^{-1}$ exists as well. Finally $T = \frac{\kappa}{n}I^*$, where the star denotes the $(N-k) \times (N-k)$ submatrix, so $T^{-1} = \frac{n}{\kappa}I^*$.

The continuity of $L\left(\xi\right)$ on $\Xi$ is now evident from (18).

To prove $L\left(\xi\right) = M\left(\xi\right)$, let us denote the upper $k \times p$ and lower $(N-k) \times p$ submatrices of $F$ by $F'$ and $F^*$, respectively. We have

$$\begin{aligned} L\left(\xi\right) &= \begin{pmatrix} F' \\ F^* \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} V^{-1} & -V^{-1}RT^{-1} \\ 0 & T^{-1} \end{pmatrix} \begin{pmatrix} \mathrm{diag}'\left\{\xi\left(\cdot\right)\right\}F' \\ 0 \end{pmatrix} \\ &= \left(F'\right)^{\mathrm{T}}V^{-1}\mathrm{diag}'\left\{\xi\left(\cdot\right)\right\}F' \\ &= \left(F'\right)^{\mathrm{T}}\left[\left(C' - \kappa I'\right) + \frac{\kappa}{n}\mathrm{diag}'\left\{\xi^{-1}\left(\cdot\right)\right\}\right]^{-1}F' = M\left(\xi\right), \end{aligned}$$

as follows from Eq. (4).

Let us suppose first that $\bar{\xi} \in \Xi_+$. Using the standard rules for the derivatives of matrices, we obtain from (18)

$$\begin{aligned} \left.\frac{\partial L\left(\xi_\alpha^\mu\right)}{\partial \alpha}\right|_{\alpha=0} &= -F^{\mathrm{T}}Z^{-1}\left(\bar{\xi}\right)\mathrm{diag}\left(\mu - \bar{\xi}\right)\left(C - \kappa I\right)Z^{-1}\left(\bar{\xi}\right)\mathrm{diag}\left(\bar{\xi}\right)F \\ &\quad + F^{\mathrm{T}}Z^{-1}\left(\bar{\xi}\right)\mathrm{diag}\left(\mu - \bar{\xi}\right)F \\ &= F^{\mathrm{T}}Z^{-1}\left(\bar{\xi}\right)\mathrm{diag}\left(\mu - \bar{\xi}\right)\left[I - \left(C - \kappa I\right)Z^{-1}\left(\bar{\xi}\right)\mathrm{diag}\left(\bar{\xi}\right)\right]F \end{aligned}$$

$$= \quad F^{\mathsf{T}} Z^{-1} \left( \bar{\xi} \right) \operatorname{diag} \left( \mu - \bar{\xi} \right) \frac{\kappa}{n} \left\{ Z^{\mathsf{T}} \left( \bar{\xi} \right) \right\}^{-1} F$$

$$= \quad \frac{\kappa}{n} F^{\mathsf{T}} T^{\mathsf{T}} \left( \bar{\xi} \right) \operatorname{diag} \left( \mu - \bar{\xi} \right) T \left( \bar{\xi} \right) F, \tag{19}$$

which is well defined and continuous on $\Xi_+$. The equality

$$\frac{\kappa}{n} \left\{ Z^{\mathsf{T}} \left( \bar{\xi} \right) \right\}^{-1} = I - (C - \kappa I) Z^{-1} \left( \bar{\xi} \right) \operatorname{diag} \left( \bar{\xi} \right)$$

for $Z^{\mathsf{T}} \left( \bar{\xi} \right) = (C - \kappa I) \operatorname{diag} \left( \bar{\xi} \right) + \frac{\kappa}{n} I$ is obtained by applying the identity $(UV + Y)^{-1} = Y^{-1} - Y^{-1} U \left( VY^{-1}U + I \right)^{-1} VY^{-1}$ with matrices $U = C - \kappa I$, $V = \operatorname{diag} \left( \bar{\xi} \right)$, and $Y = \frac{\kappa}{n} I$. From the definition of $T(\xi)$ in Section 2.5 it immediately follows that $\left\{ Z^{\mathsf{T}} \left( \bar{\xi} \right) \right\}^{-1} = T \left( \bar{\xi} \right)$ and $Z^{-1} \left( \bar{\xi} \right) = T^{\mathsf{T}} \left( \bar{\xi} \right)$.

Suppose now that $\bar{\xi} \in \Xi \backslash \Xi_+$. To obtain (19) from (18), we used two standard rules for the derivatives of matrix functions $A(\alpha)$ and $B(\alpha)$, namely $\frac{\partial}{\partial \alpha} \left[ A(\alpha) B(\alpha) \right] = \frac{\partial A(\alpha)}{\partial \alpha} B(\alpha) + A(\alpha) \frac{\partial B(\alpha)}{\partial \alpha}$ and $\frac{\partial A^{-1}(\alpha)}{\partial \alpha} = -A^{-1}(\alpha) \frac{\partial A(\alpha)}{\partial \alpha} A^{-1}(\alpha)$. The same rules hold if instead of the symbol of the derivative $\frac{\partial}{\partial \alpha}$ we use the symbol of the limit $\lim\limits_{\alpha \to 0^+}$, abbreviated below by the symbol $\Lambda$. Indeed,

$$\begin{aligned} \Lambda \left[ A \left( \alpha \right) B \left( \alpha \right) \right] &= \lim_{\alpha \to 0^+} \left[ \frac{A \left( \alpha \right) - A \left( 0 \right)}{\alpha} B \left( \alpha \right) + A \left( 0 \right) \frac{B \left( \alpha \right) - B \left( 0 \right)}{\alpha} \right] \\ &= \Lambda \left[ A \left( \alpha \right) \right] B \left( 0 \right) + A \left( 0 \right) \Lambda \left[ B \left( \alpha \right) \right], \end{aligned}$$

and from $0 = \Lambda \left[ A(\alpha) A^{-1}(\alpha) \right] = A(0) \Lambda \left[ A^{-1}(\alpha) \right] + \Lambda \left[ A(\alpha) \right] A^{-1}(0)$ we also have

$$\Lambda \left[ A^{-1} \left( \alpha \right) \right] = -A^{-1} \left( 0 \right) \Lambda \left[ A \left( \alpha \right) \right] A^{-1} \left( 0 \right).$$

Evidently $\Lambda \left[ \operatorname{diag} \left\{ (1 - \alpha) \bar{\xi} + \alpha \mu \right\} \right] = \operatorname{diag} \left( \mu - \bar{\xi} \right)$. Hence we finally obtain that

$$\lim_{\alpha \to 0^+} \frac{M \left\{ (1 - \alpha) \bar{\xi} + \alpha \mu \right\} - M \left( \bar{\xi} \right)}{\alpha} = \lim_{\alpha \to 0^+} \frac{L \left\{ (1 - \alpha) \bar{\xi} + \alpha \mu \right\} - L \left( \bar{\xi} \right)}{\alpha} \tag{20}$$

is expressed again by the right-hand side of (19), which is evidently a continuous function of $\bar{\xi}$ on the whole set $\Xi$. $\qquad \square$

## Appendix G: More examples and additional results

### *G.1. Random sampling results for Examples 1 and 2*

Tables 5, 6, and 7 contain further results for Example 1, the modified Example 1, and Example 2 when using the random sampling approaches to obtain exact designs. In addition to the approaches introduced in Section 3, we will also consider exact designs found by random sampling from the measure provided by [6] and denote this approach by R-DPZ+EP.

TABLE 5
*Further optimal designs and D-efficiencies for Example 1*

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | D-eff |
|---|---|---|---|---|---|
| R-UNIF (median efficiency) | | | | | 0.7440 |
| R-UNIF (highest efficiency) | 1.26 | 1.65 | 1.76 | 1.99 | 0.8907 |
| R-UNIF+EP (median efficiency) | | | | | 0.5591 |
| R-UNIF+EP (highest efficiency) | 1.00 | 1.20 | 1.76 | 2.00 | 0.8725 |
| R-VN (median efficiency) | | | | | 0.7957 |
| R-VN (highest efficiency) | 1.20 | 1.64 | 1.74 | 2.00 | 0.9062 |
| R-DPZ+EP (median efficiency) | | | | | 0.5554 |
| R-DPZ+EP (highest efficiency) | 1.00 | 1.22 | 1.72 | 2.00 | 0.8838 |

TABLE 6
*Further optimal designs and D-efficiencies for the modified Example 1*

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | D-eff |
|---|---|---|---|---|---|
| R-UNIF (median efficiency) | | | | | 0.4724 |
| R-UNIF (highest efficiency) | 1.05 | 1.24 | 1.70 | 1.99 | 0.9207 |
| R-UNIF+EP (median efficiency) | | | | | 0.6608 |
| R-UNIF+EP (highest efficiency) | 1.00 | 1.19 | 1.73 | 2.00 | 0.9582 |
| R-VN (median efficiency) | | | | | 0.4933 |
| R-VN (highest efficiency) | 1.01 | 1.39 | 1.75 | 2.00 | 0.8341 |

### *G.2. Example 4: absolute exponential kernel*

Similar conclusions as for the other examples hold for the next example taken from Section 4.2 of [5]. This four-parameter model is characterized by

$$
\begin{aligned}
f^{\mathrm{T}}(x) &= (\sin x, \cos x, \sin 2x, \cos 2x), \\
\mathrm{cov}\left\{\varepsilon(x), \varepsilon(x')\right\} &= \exp\left(-\left|x - x'\right|\right) \\
x &\in [1, 2], \\
\lambda_{\min}(C) &= 0.005.
\end{aligned}
$$

As [5], we consider the A-criterion for this example. Since the linear programming algorithm requires the criterion to be positive, we select the criterion to be

$$
\Phi(M) = \left\{\mathrm{tr}\left(M^{-1}\right)\right\}^{-1},
$$

for which the derivative is

$$
\nabla_M \Phi(M) = M^{-2} \cdot \left\{\mathrm{tr}\left(M^{-1}\right)\right\}^{-2},
$$

both of which we can plug into the linear Taylor formula to obtain the set of linear constraints.

The selected design points and A-efficiencies for all methods are given in Table 8.

TABLE 7

*Further optimal designs and D-efficiencies for Example 2*

|  | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | D-eff |
|---|---|---|---|---|---|---|
| R-UNIF (median efficiency) |  |  |  |  |  | 0.2910 |
| R-UNIF (highest efficiency) | 1.05 | 1.26 | 1.32 | 1.71 | 1.99 | 0.7684 |
| R-UNIF+EP (median efficiency) |  |  |  |  |  | 0.7813 |
| R-UNIF+EP (highest efficiency) | 1.00 | 1.17 | 1.37 | 1.76 | 2.00 | 0.9282 |
| R-VN (median efficiency) |  |  |  |  |  | 0.6712 |
| R-VN (highest efficiency) | 1.00 | 1.14 | 1.34 | 1.82 | 2.00 | 0.9258 |
| R-DPZ+EP (median efficiency) |  |  |  |  |  | 0.7809 |
| R-DPZ+EP (highest efficiency) | 1.00 | 1.25 | 1.57 | 1.85 | 2.00 | 0.9237 |

TABLE 8

*Optimal designs and A-efficiencies for Example 4*

|  | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | A-eff |
|---|---|---|---|---|---|---|
| Q-VN | 1.00 | 1.14 | 1.57 | 1.84 | 2.00 | 0.7757 |
| Q-VN+EP | 1.00 | 1.16 | 1.57 | 1.82 | 2.00 | 0.7971 |
| Q-DPZ+EP | 1.00 | 1.25 | 1.50 | 1.75 | 2.00 | 0.7478 |
| R-UNIF (median efficiency) |  |  |  |  |  | 0.0448 |
| R-UNIF (highest efficiency) | 1.05 | 1.26 | 1.32 | 1.71 | 1.99 | 0.5314 |
| R-UNIF+EP (median efficiency) |  |  |  |  |  | 0.4515 |
| R-UNIF+EP (highest efficiency) | 1.00 | 1.20 | 1.76 | 1.87 | 2.00 | 0.8581 |
| R-VN (median efficiency) |  |  |  |  |  | 0.3681 |
| R-VN (highest efficiency) | 1.00 | 1.16 | 1.24 | 1.80 | 2.00 | 0.8470 |
| R-DPZ+EP (median efficiency) |  |  |  |  |  | 0.4916 |
| R-DPZ+EP (highest efficiency) | 1.00 | 1.06 | 1.24 | 1.80 | 2.00 | 0.8431 |
| BKSF | 1.00 | 1.16 | 1.27 | 1.83 | 2.00 | 0.8381 |
| EXS | 1.00 | 1.20 | 1.76 | 1.89 | 2.00 | 0.9072 |
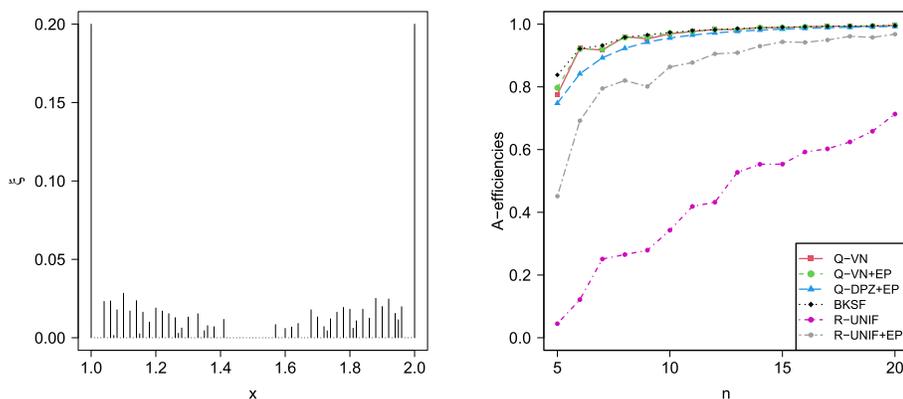


FIG 5. *Our measure (left panel) and efficiencies versus sample size (right panel) for the modified Example 4.*

## Example 5: a bivariate case

The final example is a multivariate extension of Example 4 to demonstrate once more that our methodology can principally be extended to design dimensions
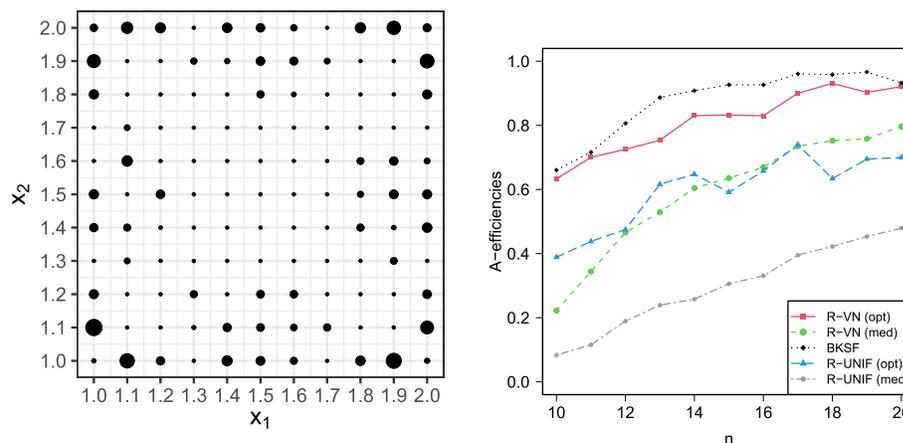
FIG 6. *Graphs for Example 5. Left: discrete measure obtained by running linear programming algorithm for virtual noise representation for $n = 10$. Right: A-efficiencies of exact desig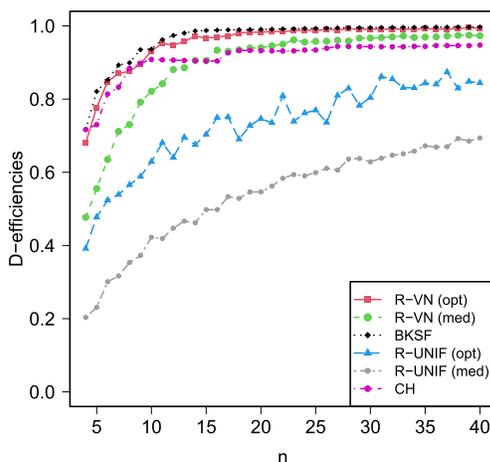ns with respect to optimal measure of virtual noise representation obtained by various methods for $n = 10$ to $n = 20$. The methods depicted are: R-VN (optimum: solid red line with squares, median: dashed green line with large dots), R-UNIF (optimum: long-dashed blue line with triangles, median: long-short-dashed gray line with small dots), BKSF (dotted black line with diamonds).*



FIG 7. *D-efficiencies for Example 3 for $n = 4$ to $n = 40$. The methods depicted are: R-VN (optimum: solid red line with squares, median: dashed green line with large dots), R-UNIF (optimum: long-dashed blue line with triangles, median: long-short-dashed gray line with small dots), BKSF (dotted black line with diamonds), coffee-house design (dot-dashed purple line with large dots).*
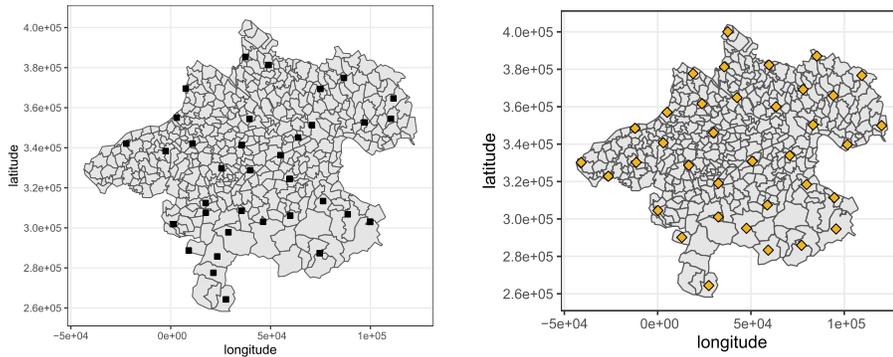
greater than one.

$$x = (x_1, x_2)^{\mathrm{T}},$$

Fɪɢ 8. *Actual weather station network (left panel) and coffee-house design (right panel) for Example 3.*

$$
\begin{aligned}
f^{\mathrm{T}}(x) &= \left(\sin x_1, \cos x_1, \sin 2x_1, \cos 2x_1, \sin x_2, \cos x_2, \sin 2x_2, \cos 2x_2\right), \\
k(x, x') &= \exp\left(-|x - x'|\right), \text{ where } |x - x'| = |x_1 - x_1'| + |x_2 - x_2'| \\
x &\in [1, 2] \times [1, 2], \\
\text{discretized: } x &\in \{1, 1.1, \ldots, 1.9, 2\} \times \{1, 1.1, \ldots, 1.9, 2\}, \\
\lambda_{\min}(C) &= 0.002599.
\end{aligned}
$$

To obtain exact designs from our design measure on a two-dimensional grid, we used the random sampling approach. That is, we sampled 100 $n$-point designs according to our measure. In Fig. 6, the best as well as the median A-efficiencies among the sampled designs are plotted. We also sampled 100 designs uniformly on the grid and computed the best and median efficiencies among those designs. Using the best among the sampled designs leads to reasonably efficient designs compared to the algorithm using the approximate sensitivity function proposed by [8].

### Additional results for Example 3

Fig. 7 displays the D-efficiencies with respect to $\Phi\left\{M(\bar{\xi})\right\}$ for Example 3 for $n = 4$ to $n = 40$. The actual weather station network used to estimate the model parameters as well as the coffee-house design are displayed in Fig. 8.

### Acknowledgments

## References

[1] Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS. Oxford Statistical Science Series* **34**. Oxford University Press, Oxford. MR2323647

[2] Berkelaar, M. et al. (2020). Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs.

[3] Brimkulov, U., Krug, G. and Savanov, V. (1980). Numerical construction of exact experimental designs when the measurements are correlated. *Zavodskaya Laboratoria* **36** 435–442.

[4] Burclová, K. and Pázman, A. (2016). Optimal design of experiments via linear programming. *Statistical Papers* **57** 893–910. MR3571179

[5] Dette, H., Konstantinou, M. and Zhigljavsky, A. (2017). A new approach to optimal designs for correlated observations. *Annals of Statistics* **45** 1579–1608. MR3670189

[6] Dette, H., Pepelyshev, A. and Zhigljavsky, A. (2016). Optimal designs in regression with correlated errors. *The Annals of Statistics* **44** 113–152. MR3449764

[7] Dette, H., Pepelyshev, A. and Zhigljavsky, A. (2019). The BLUE in continuous-time regression models with correlated errors. *The Annals of Statistics* **47** 1928–1959. MR3953440

[8] Fedorov, V. (1996). Design of spatial experiments: Model fitting and prediction. In *Handbook of Statistics*, (S. Ghosh and C. Rao, eds.) **13** 515–553. Elsevier. MR1492578

[9] Glatzer, E. and Müller, W. G. (1999). A comparison of optimum design algorithms for regressions with correlated observations. In *Probastat '98* (A. Pázman and G. Wimmer, eds.) **17** 149–156. Tatra Mountains Mathematical Publications. MR1737702

[10] Goos, P. and Jones, B. (2011). *Optimal Design of Experiments: A Case Study Approach.* Wiley. MR2849558

[11] Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective.* Springer. MR1467237

[12] Kelley, J. E. (1960). The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics* **8** 703–712. MR0118538

[13] Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)* **21** 272–304. MR0113263

[14] Mateu, J. and Müller, W. G. (2012). *Spatio-temporal Design: Advances in Efficient Data Acquisition.* Wiley. MR3289517

[15] Morris, M. (2010). *Design of Experiments: An Introduction Based on Linear Models.* Chapman and Hall/CRC. MR2682730

[16] Müller, W. and Pázman, A. (2003). Measures for designs in experiments with correlated errors. *Biometrika* **90** 423–434. MR1986657

[17] Müller, W. G. (2001). Coffee-house designs. In *Optimum Design 2000* (A. Atkinson, B. Bogacka and A. Zhigljavsky, eds.) 241–248. Springer US, Boston, MA. MR1822367

[18] NÄTHER, W. (1985). *Effective Observation of Random Fields.* Teubner. MR0863287

[19] PÁZMAN, A. (2007). Criteria for optimal design of small-sample experiments with correlated observations. *Kybernetika* **43** 453–462. MR2377923

[20] PÁZMAN, A. and MÜLLER, W. G. (1998). A new interpretation of design measures. In *MODA 5 — Advances in Model-Oriented Data Analysis and Experimental Design* (A. Atkinson, L. Pronzato and H. Wynn, eds.) 239–246. Physica-Verlag HD. MR1652228

[21] PRONZATO, L. and PÁZMAN, A. (2013). *Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties. Lecture Notes in Statistics.* Springer. MR3058804

[22] PUKELSHEIM, F. (2006). *Optimal Design of Experiments.* SIAM/Society for Industrial and Applied Mathematics. MR2224698

[23] RASMUSSEN, C. and WILLIAMS, C. (2005). *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning series.* The MIT Press. MR2514435

[24] ROCKAFELLAR, R. T. (1970). *Convex Analysis.* Princeton University Press. MR0274683

[25] SACKS, J. and YLVISAKER, D. (1966). Designs for regression problems with correlated errors. *The Annals of Mathematical Statistics* **37** 66–89. MR0192601

[26] R CORE TEAM (2020). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

[27] UCIŃSKI, D. (2020). D-optimal sensor selection in the presence of correlated measurement noise. *Measurement* **164** 107873.

[28] WAITE, T. W. and WOODS, D. C. (2022). Minimax efficient random experimental design strategies with application to model-robust design for prediction. *Journal of the American Statistical Association* **117** 1452–1465. MR4480724

[29] ZHIGLJAVSKY, A., DETTE, H. and PEPELYSHEV, A. (2010). A new approach to optimal design for linear models with correlated observations. *Journal of the American Statistical Association* **105** 1093–1103. MR2752605