Active Discrimination Learning for Gaussian Process Models

Elham Yousefi, Luc Pronzato, Markus Hainy, Werner G. Müller, Henry P. Wynn

Active Discrimination Learning for Gaussian Process Models

Elham Yousefi[1], Luc Pronzato[2], Markus Hainy[1], Werner G. Müller[1], Henry P. Wynn[3]

[1]JKU Linz, [2]CNRS, [3]London School of Economics

Abstract

The paper covers the design and analysis of experiments to discriminate between two Gaussian process models, such as those widely used in computer experiments, kriging, sensor location and machine learning. Two frameworks are considered. First, we study sequential constructions, where successive design (observation) points are selected, either as additional points to an existing design or from the beginning of observation. The selection relies on the maximisation of the difference between the symmetric Kullback Leibler divergences for the two models, which depends on the observations, or on the mean squared error of both models, which does not. Then, we consider static criteria, such as the familiar log-likelihood ratios and the Fréchet distance between the covariance functions of the two models. Other distance-based criteria, simpler to compute than previous ones, are also introduced, for which, considering the framework of approximate design, a necessary condition for the optimality of a design measure is provided. The paper includes a study of the mathematical links between different criteria and numerical illustrations are provided.

Keywords: model discrimination; Gaussian random field; kriging

1 Introduction

The term 'active learning' (cf. Hino (2020) for a recent review) has replaced the traditional (sequential or adaptive) 'design of experiments' in the computer science literature, typically when the response is approximated by Gaussian process regression (GPR, cf. Sauer et al. (2022)). It refers to selecting the most suitable inputs to achieve the maximum of information from the outputs, usually with the aim of improving prediction accuracy. A good overview is given in Chapter 6 of Gramacy (2020).

Frequently the aim of an experiment – in the broad sense of any data acquisition exercise – may rather be the discrimination between two or more potential explanatory models. When data can be sequentially collected during the experimental process, the literature goes back to the classic procedure of Hunter and Reiner (1965) and has generated ongoing research (see e.g. Schwaab et al. (2008), Olofsson et al. (2018) and Heirung et al. (2019)). When the design needs to be fixed before the experiment and thus no intermediate data will be available, the literature is less developed. While in the classical (non)linear regression case the criterion of T-optimality (cf. Atkinson and Fedorov (1975)) and the numerous papers extending it was a major step, a similar breakthrough for Gaussian process regression is lacking.

With this paper we would like to investigate various sequential/adaptive and non-sequential design schemes for GPRs and their relative properties. When the observations associated with the already collected points are available, one may base the criterion on the predictions and

prediction errors (Section 3.1). On the one hand, one natural choice will be to put the next design point where the symmetric Kullback-Leibler divergence between those two predictive (normal) distributions differs most. On the other hand, when the associated observations are not available, the incremental construction of the designs could be based on the mean squared error (MSE) for both models, assuming in turn that either of the two models is the true one (Section 3.2). The static construction of a set of optimal designs of given size for nominal model parameters is the last mode we have considered (Section 4). Our first choice is to use the difference between the expected values of the log likelihood ratios, assuming in turn that either of the two models is the true one. This is actually a function of the symmetric Kullback-Leibler divergence, which also arises from Bayesian considerations. In a similar spirit, the Fréchet distance between two covariance matrices provides another natural criterion. Some further novel but simple approaches are considered in this paper as well. In particular we are interested whether complex likelihood-based criteria like the Kullback-Leibler-divergence can be effectively replaced by simpler ones based directly on the respective covariance kernels. The construction of optimal design measures for model discrimination (approximate design theory) is considered in Section 5.

Eventually, to compare the discriminatory power of the resulting designs from different criteria, one can compute the correct classification (hit) rates after selecting the model with the higher likelihood value. A numerical illustration is provided in Section 6 for two Matérn kernels with different smoothness.

2 Notation

One of the most popular design criteria for discriminating between rival models is T-optimality (Atkinson and Fedorov, 1975). This criterion is only applicable when the observations are independent and normally distributed with a constant variance. López-Fidalgo et al. (2007) generalised the normality assumption and developed an optimal discriminating design criterion to choose among non-normal models. The criterion is based on the log-likelihood ratio test under the assumption of independent observations. We denote by $\varphi_0(y, x, \theta_0)$ and $\varphi_1(y, x, \theta_1)$ the two rival probability density functions for one observation y at point x. The following system of hypotheses might be considered:

$$H_0 : \varphi(y, x) = \varphi_0(y, x, \theta_0)$$
$$H_1 : \varphi(y, x) = \varphi_1(y, x, \theta_1)$$

where $\varphi_1(y, x, \theta_1)$ is assumed to be the true model. A common test statistic is the log-likelihood ratio given as

$$L = -\log \frac{\varphi_0(y, x, \theta_0)}{\varphi_1(y, x, \theta_1)} = \log \frac{\varphi_1(y, x, \theta_1)}{\varphi_0(y, x, \theta_0)},$$

where the null hypothesis is rejected when $\varphi_1(y, x, \theta_1) > \varphi_0(y, x, \theta_0)$ or equivalently when $L > 0$. The power of the test refers to the expected value of the log-likelihood ratio criterion under the alternative hypothesis H_1. We have

$$\begin{aligned} \mathsf{E}_{H_1}(L) = \mathsf{E}_1(L) &= \int \varphi_1(y, x, \theta_1) \log \left\{ \frac{\varphi_1(y, x, \theta_1)}{\varphi_0(y, x, \theta_0)} \right\} \mathrm{d}y \\ &= D_{KL}(\varphi_1 \| \varphi_0), \end{aligned} \tag{1}$$

where $D_{KL}(\varphi_1\|\varphi_0)$ is the KullbackLeibler distance between the true and the alternative model (Kullback and Leibler, 1951).

Interchanging the two models in the null and the alternative hypothesis, the power of the test would be

$$\mathsf{E}_0(-L) \quad = \quad D_{KL}(\varphi_0\|\varphi_1). \tag{2}$$

If it is not clear in advance which of the two models is the true model, one might consider to search for a design optimising a convex combination of (1) and (2), most commonly using weights 1/2 for each model. This would be equivalent to maximising the symmetric Kullback-Leibler distance

$$D_{KL}(\varphi_0,\varphi_1) \quad = \quad \frac{1}{2}\left[D_{KL}(\varphi_0\|\varphi_1)+D_{KL}(\varphi_1\|\varphi_0)\right].$$

In this paper we will consider random fields, i.e. we will allow for correlated observations. When the random field is Gaussian, we might still base the design strategy on the log-likelihood ratio criterion to choose among two rival models.

For a positive definite kernel $K(x,x')$ and an n-point design $\mathbf{X}_n=(x_1,\ldots,x_n)$, $\mathbf{k}_n(x)$ is the n-dimensional vector $(K(x,x_1),\ldots,K(x,x_n))^\top$ and $\mathbf{K}_n$ is the $n\times n$ (kernel) matrix with elements $\{\mathbf{K}_n\}_{i,j}=K(x_i,x_j)$. Although x is not bold, it may correspond to a point in a (compact) set $\mathscr{X}\subset\mathbb{R}^d$. Assume that $Y(x)$ corresponds to the realisation of a random field Z_x, indexed by x in $\mathscr{X}$, with zero mean $\mathsf{E}\{Z_x\}=0$ for all x and covariance $\mathsf{E}\{Z_xZ_{x'}\}=K(x,x')$ for all $(x,x')\in\mathscr{X}^2$. Our prediction of a future observation $Y(x)$ based on observations $\mathbf{Y}_n=(Y(x_1),\ldots,Y(x_n))^\top$ corresponds to the best linear unbiased predictor (BLUP) $\widehat{\eta}_n(x)=\mathbf{k}_n^\top(x)\mathbf{K}_n^{-1}\mathbf{Y}_n$. The associated prediction error is $e_n(x)=Y(x)-\widehat{\eta}_n(x)$ and we have

$$\mathsf{E}\{e_n^2(x)\}=\rho_n^2(x)=K(x,x)-\mathbf{k}_n^\top(x)\mathbf{K}_n^{-1}\mathbf{k}_n(x).$$

The index n will often be omitted when there is no ambiguity, and in that case $\mathbf{k}_i(x)=\mathbf{k}_{n,i}(x)$, $\mathbf{K}_i=\mathbf{K}_{n,i}$, $e_i(x)=e_{n,i}(x)$, $\rho_i^2(x)=\rho_{n,i}^2(x)$ will refer instead to model i, with $i\in\{0,1\}$. We shall need to distinguish between the cases where the truth is model 0 or model 1, and following Stein (1999, p. 58) we denote by E_i the expectation computed with model i assumed to be true. We reserve the notation $\rho_i^2(x)$ to the case where the expectation is computed with the true model; i.e.,

$$\rho_i^2(x)=\mathsf{E}_i\{e_i^2(x)\}.$$

Hence we have $\rho_0^2(x)=\mathsf{E}_0\{e_0^2(x)\}=K_0(x,x)-\mathbf{k}_0^\top(x)\mathbf{K}_0^{-1}\mathbf{k}_0(x)$ and calculation gives

$$\begin{aligned}\mathsf{E}_0\{e_1^2(x)\} &= K_0(x,x)+\mathbf{k}_1^\top(x)\mathbf{K}_1^{-1}\mathbf{K}_0\mathbf{K}_1^{-1}\mathbf{k}_1(x)-2\,\mathbf{k}_1^\top(x)\mathbf{K}_1^{-1}\mathbf{k}_0(x), \qquad (3)\\ \mathsf{E}_0\{[e_1(x)-e_0(x)]^2\} &= \mathsf{E}_0\{e_1^2(x)\}-\mathsf{E}_0\{e_0^2(x)\},\end{aligned}$$

with an obvious permutation of indices 0 and 1 when assuming the model 1 is true to compute $\mathsf{E}_1\{\cdot\}$.

If model 0 is correct, the prediction error is larger when we use model 1 for prediction than if we use the BLUP (i.e., model 0). Stein (1999, p. 58) shows that the relation

$$\frac{\mathsf{E}_0\{e_1^2(x)\}}{\mathsf{E}_0\{e_0^2(x)\}}=1+\frac{\mathsf{E}_0\{[e_1(x)-e_0(x)]^2\}}{\mathsf{E}_0\{e_0^2(x)\}}$$

shown above is valid more generally for models with linear trends. Also of interest is the assumed mean squared error (MSE) $\mathsf{E}_1\{e_1^2(x)\}$ when we use model 1 for assessing the prediction error (because we think it is correct) while the truth is model 0, and in particular the ratio

$$\frac{\mathsf{E}_1\{e_1^2(x)\}}{\mathsf{E}_0\{e_1^2(x)\}} = \frac{K_1(x,x) - \mathbf{k}_1^\top(x)\mathbf{K}_1^{-1}\mathbf{k}_1(x)}{\mathsf{E}_0\{e_1^2(x)\}},$$

which may be larger or smaller than one.

Another important issue concerns the choice of covariance parameters in K_0 and K_1. Denote $K_i(x,x') = \sigma_i^2\, C_{i,\theta_i}(x,x')$, $i=0,1$, $(x,x') \in \mathscr{X}^2$, where the σ_i^2 define the variance, the θ_i may correspond to correlation lengths in a translation invariant model and are thus scalar in the isotropic case, and $C(x,x')$ defines a correlation.

3 Prediction-based discrimination

For the incremental construction of a design for model discrimination, points are added conditionally on previous design points. We can distinguish the case where the observations associated with those previous points are available and can thus be used to construct a sequence of predictions (sequential, i.e., conditional, construction) from the unconditional case where observations are not used.

3.1 Sequential (conditional) design

Consider stage n, where n design points $\mathbf{X}_n$ and n observations $\mathbf{Y}_n$ are available. Assuming that the random field is Gaussian, when model i is true we have $Y(x) \sim \mathscr{N}(\widehat{\eta}_{n,i}(x), \rho^2_{n,i}(x))$. A rather natural choice is to choose the next design point x_{n+1} where the symmetric Kullback-Leibler divergence between those two normal distributions differs most; that is,

$$x_{n+1} \in \operatorname{Arg}\max_{x\in\mathscr{X}} \frac{\rho^2_{n,0}(x)}{\rho^2_{n,1}(x)} + \frac{\rho^2_{n,1}(x)}{\rho^2_{n,0}(x)} + [\widehat{\eta}_{n,1}(x) - \widehat{\eta}_{n,0}(x)]^2 \left[\frac{1}{\rho^2_{n,0}(x)} + \frac{1}{\rho^2_{n,1}(x)}\right]. \tag{4}$$

Other variants could be considered as well, such as

$$\begin{aligned}
x_{n+1} &\in \operatorname{Arg}\max_{x\in\mathscr{X}} [\widehat{\eta}_{n,1}(x) - \widehat{\eta}_{n,0}(x)]^2,\\
x_{n+1} &\in \operatorname{Arg}\max_{x\in\mathscr{X}} \frac{[\widehat{\eta}_{n,1}(x) - \widehat{\eta}_{n,0}(x)]^2}{\rho^2_{n,0}(x) + \rho^2_{n,1}(x)},\\
x_{n+1} &\in \operatorname{Arg}\max_{x\in\mathscr{X}} [\widehat{\eta}_{n,1}(x) - \widehat{\eta}_{n,0}(x)]^2 \left[\frac{1}{\rho^2_{n,0}(x)} + \frac{1}{\rho^2_{n,1}(x)}\right].
\end{aligned}$$

They will not be considered in the rest of the paper.

If necessary one can use plug-in estimates $\widehat{\sigma}^2_{n,i}$ and $\widehat{\theta}_{n,i}$ of σ_i^2 and θ_i, for instance maximum likelihood (ML) or leave-one-out estimates based on $\mathbf{X}_n$ and $\mathbf{Y}_n$, when we choose x_{n+1}. Note that the value of σ^2 does not affect the BLUP $\widehat{\eta}_n(x) = \mathbf{k}_n^\top \mathbf{K}_n^{-1}\mathbf{Y}_n$. In the paper we do not address the issues related to the estimation of σ^2 or of the correlation length or smoothness parameters of the kernel; one may refer to Karvonen et al. (2020) and the recent papers Karvonen (2022); Karvonen and Oates (2022) for a detailed investigation. The connection between the notion of microergodicity, related to the consistency of the maximum-likelihood estimator, and discrimination through a KL divergence criterion is nevertheless considered in Example 1 below.

3.2 Incremental (unconditional) design

Consider stage n, where n design points $\mathbf{X}_n$ are available. We base the choice of the next point on the difference between the MSEs for both models, assuming that one or the other is true. For instance, assuming that model 0 is true, the difference between the MSEs is $\mathsf{E}_0\{e_1^2(x)\} - \mathsf{E}_0\{e_0^2(x)\} = \mathsf{E}_0\{[e_1(x) - e_0(x)]^2\} = \mathsf{E}_0\{[\widehat{\eta}_{n,1}(x) - \widehat{\eta}_{n,0}(x)]^2\}$.

A first, un-normalised, version is thus

$$\begin{aligned} \phi_A(x) &= \mathsf{E}_0\{[e_1(x) - e_0(x)]^2\} + \mathsf{E}_1\{[e_1(x) - e_0(x)]^2\} \\ &= \mathsf{E}_0\{e_1^2(x)\} - \mathsf{E}_0\{e_0^2(x)\} + \mathsf{E}_1\{e_0^2(x)\} - \mathsf{E}_1\{e_1^2(x)\}\,. \end{aligned} \tag{5}$$

A normalisation seems in order here too, such as

$$\phi_B(x) = \frac{\mathsf{E}_0\{[e_1(x) - e_0(x)]^2\}}{\rho_{n,0}^2(x)} + \frac{\mathsf{E}_1\{[e_1(x) - e_0(x)]^2\}}{\rho_{n,1}^2(x)} = \frac{\mathsf{E}_0\{e_1^2(x)\}}{\mathsf{E}_0\{e_0^2(x)\}} + \frac{\mathsf{E}_1\{e_0^2(x)\}}{\mathsf{E}_1\{e_1^2(x)\}} - 2\,. \tag{6}$$

A third criterion is based on the variation of the symmetric Kullback-Leibler divergence (10) of Section 4 when adding an $(n+1)$-th point x to $\mathbf{X}_n$. Direct calculation, using

$$\mathbf{K}_{n+1,i} = \begin{pmatrix} \mathbf{K}_{n,i} & \mathbf{k}_{n,i}(x) \\ \mathbf{k}_{n,i}^\top(x) & K_i(x,x) \end{pmatrix}\,,\ i = 0, 1\,,$$

and the expression of the inverse of a block matrix, gives

$$\Phi_{KL\,[K0,K1]}(\mathbf{X}_n \cup \{x\}) = \Phi_{KL\,[K0,K1]}(\mathbf{X}_n) + \frac{1}{2}\left[\frac{\mathsf{E}_1\{e_0^2(x)\}}{\mathsf{E}_0\{e_0^2(x)\}} + \frac{\mathsf{E}_0\{e_1^2(x)\}}{\mathsf{E}_1\{e_1^2(x)\}}\right] - 1\,.$$

We thus define

$$\phi_{KL}(x) = \frac{1}{2}\left[\frac{\mathsf{E}_1\{e_0^2(x)\}}{\mathsf{E}_0\{e_0^2(x)\}} + \frac{\mathsf{E}_0\{e_1^2(x)\}}{\mathsf{E}_1\{e_1^2(x)\}}\right] - 1\,, \tag{7}$$

to be maximised with respect to $x \in \mathscr{X}$.

Although the σ_i^2 do not affect predictions, $\mathsf{E}_i\{e_j^2(x)\}$ is proportional to σ_i^2. Unless specific information is available, it seems reasonable to assume that $\sigma_0^2 = \sigma_1^2 = 1$. Other parameters θ_i should be chosen to make the two kernels the most similar, which seems easier to consider in the approach presented in Section 4, see (11). In the rest of this section we suppose that the parameters of both kernels are fixed.

The un-normalised version $\phi_A(x)$ given by (5) could be used to derive a one-step (non-incremental) criterion, in the same spirit as those of Section 4, through integration with respect to x for a given measure μ on $\mathscr{X}$. Indeed, we have

$$\mathsf{E}_0\{[e_1(x) - e_0(x)]^2\} = \mathbf{k}_0^\top(x)\mathbf{K}_0^{-1}\mathbf{k}_0(\mathbf{x}) + \mathbf{k}_1^\top(x)\mathbf{K}_1^{-1}\mathbf{K}_0\mathbf{K}_1^{-1}\mathbf{k}_1(x) - 2\,\mathbf{k}_1^\top(x)\mathbf{K}_1^{-1}\mathbf{k}_0(x)\,,$$

so that

$$\int_{\mathscr{X}} \mathsf{E}_0\{[e_1(x) - e_0(x)]^2\}\,\mathrm{d}\mu(x) = \mathrm{trace}\left[\mathbf{K}_0^{-1}\mathbf{A}_0(\mu) + \mathbf{K}_1^{-1}\mathbf{K}_0\mathbf{K}_1^{-1}\mathbf{A}_1(\mu) - 2\,\mathbf{K}_1^{-1}\mathbf{A}_{0,1}(\mu)\right]\,,$$

where $\mathbf{A}_i(\mu) = \int_{\mathscr{X}} \mathbf{k}_i(x)\mathbf{k}_i^\top(x)\,\mathrm{d}\mu(x)$, $i = 0, 1$, and $\mathbf{A}_{0,1}(\mu) = \int_{\mathscr{X}} \mathbf{k}_0(x)\mathbf{k}_1^\top(x)\,\mathrm{d}\mu(x)$. Similarly,

$$\int_{\mathscr{X}} \mathsf{E}_1\{[e_1(x) - e_0(x)]^2\}\,\mathrm{d}\mu(x) = \mathrm{trace}\left[\mathbf{K}_1^{-1}\mathbf{A}_1(\mu) + \mathbf{K}_0^{-1}\mathbf{K}_1\mathbf{K}_0^{-1}\mathbf{A}_0(\mu) - 2\,\mathbf{K}_0^{-1}\mathbf{A}_{0,1}(\mu)\right]\,.$$

The matrices $\mathbf{A}_i(\mu)$ and $\mathbf{A}_{0,1}(\mu)$ can be calculated explicitly for some kernels and measures μ. This happens in particular when $\mathscr{X} = [0,1]^d$, the two kernels K_i are separable, i.e., products of one-dimensional kernels on $[0,1]$, and μ is uniform on $\mathscr{X}$.

Example 1: exponential covariance, no microergodic parameters. We consider Example 6 in Stein (1999, p. 74) and take $K_i(x,x') = \mathsf{e}^{-\alpha_i|x-x'|}/\alpha_i$, $i = 0, 1$. The example focuses on two difficulties: first, the two kernels only differ by their parameter values; second, the particular relation between the variance and correlation length makes the parameters α_i not microergodic and they cannot be estimated consistently from observations on a bounded interval; see Stein (1999, Chap. 6). It is interesting to investigate the behaviour of the criteria (5), (6) and (7) in this particular situation.

We suppose that n observations are made at $x_i = (i-1)/(n-1)$, $i = 1, \ldots, n \geq 2$. We denote $\delta = \delta_n = 1/[2(n-1)]$ the half-distance between two design points. The particular Markovian property of random processes with kernels K_i simplifies the analysis. The prediction and MSE at a given $x \in (0,1)$ only depend on the position of x relative to its two closest neighbouring design points; moreover, all other points have no influence. Therefore, due to the regular repartition of the x_i, we only need to consider the behaviour in one (any) interval $\mathbb{I}_i = [a_i, b_i] = [x_i, x_{i+1}]$.

We always have $\phi_A(x) \to 0$ as $x \to x_i \in \mathscr{X}_n$. Numerical calculation shows that for δ_n small enough, $\phi_A(\cdot)$ has a unique maximum in $\mathbb{I}_i$ at the centre $C_i = (x_i + x_{i+1})/2$. The next design point x_{n+1} that maximises $\phi_A(\cdot)$ is then taken at C_i for one of the $n-1$ intervals, and we get

$$\phi_A(C_i) = \frac{1}{4}\,\frac{(\alpha_1-\alpha_0)^2(\alpha_1+\alpha_0)^3}{\alpha_0\alpha_1}\,\delta_n^4 + \mathcal{O}(\delta_n^5)\,,\ n \to \infty\,.$$

Similar results apply to the case where the design $\mathbf{X}_n$ contains the endpoints 0 and 1 and its covering radius $\mathsf{CR}(\mathbf{X}_n) = \max_{x\in[0,1]} \min_{i=1,\ldots,n} |x - x_i|$ tends to zero, the points x_i being not necessarily equally spaced: C_i is then the centre of the largest interval $[x_i, x_{i+1}]$ and $\delta_n = \mathsf{CR}(\mathbf{X}_n)$.

When δ_n is large compared to the correlation lengths $1/\alpha_0$ and $1/\alpha_1$, there exist two maxima, symmetric with respect to C_i, that get closer to the extremities of $\mathbb{I}_i$ as α_1 increases, and C_i corresponds to a local minimum of $\phi_A(\cdot)$. This happens for instance when $\alpha_0\,\delta_n = 1$ and $\alpha_1\,\delta_n \gtrsim 2.600455$.

A similar behaviour is observed for $\phi_B(x)$ and $\phi_{KL}(x)$: for small enough δ_n they both have a unique maximum in $\mathbb{I}_i$ at C_i, with now

$$\begin{aligned}
\phi_B(C_i) &= \frac{1}{4}\,\frac{(\alpha_1-\alpha_0)^2(\alpha_1+\alpha_0)^3}{\alpha_0\alpha_1}\,\delta_n^3 + \mathcal{O}(\delta_n^4)\,,\ n \to \infty\,,\\
\phi_{KL}(C_i) &= \frac{1}{8}\,\frac{(\alpha_1-\alpha_0)^2(\alpha_1+\alpha_0)^3}{\alpha_0\alpha_1}\,\delta_n^3 + \mathcal{O}(\delta_n^4)\,,\ n \to \infty\,.
\end{aligned}$$

Also, $\phi_B(x) \to 0$ and $\phi_{KL}(x) \to 0$ as $x \to x_i \in \mathbf{X}_n$. For large values of δ_n compared to the correlation lengths $1/\alpha_0$ and $1/\alpha_1$, there exist two maxima in $\mathbb{I}_i$, symmetric with respect to C_i. When $\alpha_0\,\delta_n = 1$, this happens for instance when $\alpha_1\,\delta_n \gtrsim 2.020178$ for $\phi_B(\cdot)$ and when $\alpha_1\,\delta_n \gtrsim 7.251623$ for $\phi_{KL}(\cdot)$. However, in the second case the function is practically flat between the two maxima.

The left panel of Figure 1 presents $\phi_A(x)$, $\phi_B(x)$ and $\phi_{KL}(x)$ for $x \in [x_1, x_2] = [0, 0.1]$ when $n = 11$ ($\delta_n = 0.05$) and $\alpha_0 = 1$, $\alpha_1 = 10$. The right panel is for $\alpha_0\,\delta_n = 1$, $\alpha_1\,\delta_n = 10$.

This behaviour of $\phi_{KL}(C_i)$ for small δ_n sheds light on the fact that α is not estimable in this model. Indeed, consider a sequence of embedded n_k-point designs $\mathbf{X}_{n_k}$, initialised with the design $\mathbf{X}_n = \mathbf{X}_{n_0}$ considered above and with $n_k = 2^k\,(n_0 - 1) + 1$, all these designs having the form $x_i = (i-1)/(n_k - 1)$, $i = 1, \ldots, n_k$. Then, $\mathsf{CR}(\mathbf{X}_{n_k}) = \mathsf{CR}(\mathbf{X}_j) = \delta_j = 1/[2(n_k - 1)]$ for

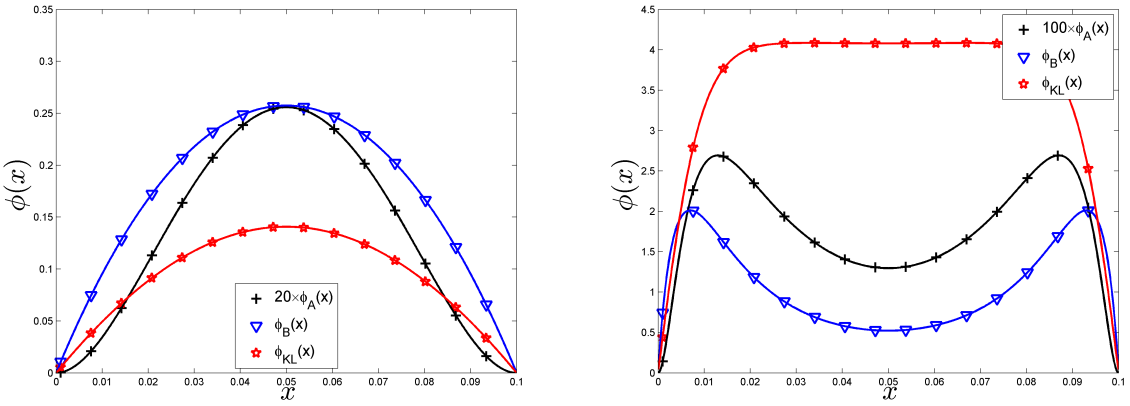

Figure 1: $\phi_A(x)$, $\phi_B(x)$ and $\phi_{KL}(x)$, $x \in [x_1, x_2]$, for $n = 11$ ($\delta_n = 0.05$) in Example 1. Left: $\alpha_0 = 1$, $\alpha_1 = 10$; Right: $\alpha_0 = 20$, $\alpha_1 = 200$.

$j = n_k, \ldots, n_{k+1} - 1 = 2\,n_k - 2$. For k large enough, the increase in Kullback-Leibler divergence (10) from $\mathbf{X}_{n_k}$ to $\mathbf{X}_{n_{k+1}}$ is thus bounded by $c/(n_k - 1)^2$ for some $c > 0$, so that the expected log-likelihood ratio $\mathsf{E}_0\{L_{n_k}\} - \mathsf{E}_1\{L_{n_k}\}$ remains bounded as $k \to \infty$.

More generally, denote by $0 \le x_1 \le x_2 \le \cdots \le x_n \le 1$ the ordered points of an n-point design $\mathbf{X}_n$ in $[0, 1]$, $n \ge 3$. Let $i^* \ge 3$ be such that $|x_{i^*-2} - x_{i^*}| = \min_{i=3,\ldots,n} |x_{i-2} - x_i|$. Then necessarily $|x_{i^*-2} - x_{i^*}| \le 1/(\lceil n/2 \rceil - 1)$. Indeed, consider the following iterative modification of $\mathbf{X}_n$ that cannot decrease $\min_{i=3,\ldots,n} |x_{i-2} - x_i|$: first, move x_1 to zero, then move x_2 to x_1; leave x_3 unchanged, but move x_4 to x_3, etc. For n even, the design $\mathbf{X}'_n$ obtained is the duplication of an $(n/2)$-points design; for n odd, only the right-most point x_n remains single. In the fist case, the minimum distance between points of $\mathbf{X}'_n$ is at most $1/(n/2 - 1)$, in the second case it is at most $1/(\lceil n/2 \rceil - 1)$. We then define $\mathbf{X}_{n-1} = \mathbf{X}_n \setminus \{x_{i^*-1}\}$. For n large enough, the increase in Kullback-Leibler divergence (10) from $\mathbf{X}_{n-1}$ to $\mathbf{X}_n$ is thus bounded by $c/(\lceil n/2 \rceil - 1)^3$ for some $c > 0$ depending on α_0 and α_1. Starting from some design $\mathbf{X}_{n_0}$, we thus have, for n_0 large enough,

$$\Phi_{KL\,[K0,K1]}(\mathbf{X}_n) - \Phi_{KL\,[K0,K1]}(\mathbf{X}_{n_0}) \le c \sum_{k=n_0+1}^{n} \frac{1}{(\lceil k/2 \rceil - 1)^3}\,,$$

which implies $\lim_{n\to\infty} \Phi_{KL\,[K0,K1]}(\mathbf{X}_n) \le B$ for some $B < \infty$. Assuming, without any loss of generality, that model 0 is correct, we have $0 \le \mathsf{E}_0\{L_n\} \le B$ (we get $0 \le \mathsf{E}_1\{-L_n\} \le B$ when we assume that model 1 is correct), implying in particular that L_n does not tend to infinity a.s. and the ML estimator of α is not strongly consistent.

Example 2: exponential covariance, microergodic parameters. Consider now two exponential covariance models with identical variances (which we take equal to one without any loss of generality): $K_i(x, x') = \mathsf{e}^{-\alpha_i |x - x'|}$, $i = 0, 1$.

Again, $\phi_A(x) \to 0$ as $x \to x_i \in \mathbf{X}_n$ and $\phi_A(\cdot)$ has a unique maximum at C_i for small enough

δ_n, with now

$$\phi_A(C_i) = \frac{1}{2}\,(\alpha_1^2 - \alpha_0^2)^2\,\delta_n^4 + \mathcal{O}(\delta_n^5)\,,\ n \to \infty\,.$$

There are two maxima for $\phi_A(\cdot)$ in $\mathbb{I}_i$, symmetric with respect to C_i for large δ_n: when $\alpha_0\,\delta_n = 1$, this happens for instance when $\alpha_1\,\delta_n \gtrsim 2.558545$. Nothing is changed for $\phi_B(\cdot)$ compared to Example 1 as the variances cancel in the ratios that define $\phi_B(\cdot)$, see (3) and (6). The situation is quite different for $\phi_{KL}(\cdot)$, with

$$\phi_{KL}(C_i) = \frac{1}{2}\,\frac{(\alpha_1 - \alpha_0)^2}{\alpha_0\alpha_1} + \mathcal{O}(\delta_n)\,,\ n \to \infty\,,$$

indicating that it is indeed possible to distinguish between the two models much more efficiently with this criterion than with the two others. Interestingly enough, the best choice for next design point is not at C_i but always as close as possible to one of the endpoints a_i or b_i, with however a criterion value similar to that in the centre C_i when δ_n is small enough, as $\lim_{x\to x_i}\phi_{KL}(x) = (\alpha_1 - \alpha_0)^2/(2\,\alpha_0\alpha_1)$. Here, the same sequence of embedded designs as in Example 1 ensures that $\mathsf{E}_0\{L_{n_k}\} - \mathsf{E}_1\{L_{n_k}\} \to \infty$ as $k \to \infty$. Figure 2 presents $\phi_A(x)$, $\phi_B(x)$ and $\phi_{KL}(x)$ in the same configuration as in Figure 1 but for the kernels $K_i(x,x') = \mathsf{e}^{-\alpha_i|x-x'|}$, $i = 0, 1$.

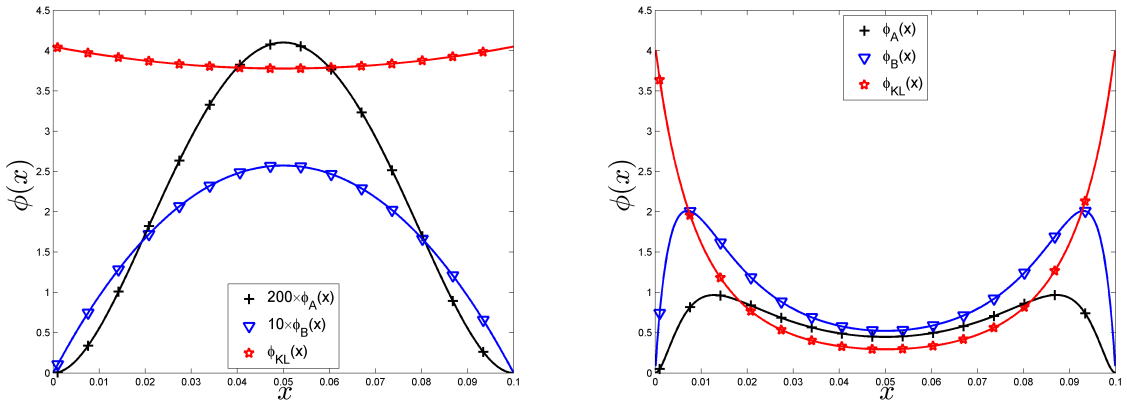

Figure 2: $\phi_A(x)$, $\phi_B(x)$ and $\phi_{KL}(x)$, $x \in [x_1, x_2]$, for $n = 11$ ($\delta_n = 0.05$) in Example 2. Left: $\alpha_0 = 1$, $\alpha_1 = 10$; Right: $\alpha_0 = 20$, $\alpha_1 = 200$.

Example 3: Matérn kernels. Take K_0 and K_1 as the 3/2 and 5/2 Matérn kernels, respectively:

$$K_{0,\theta}(x,x') = (1 + \sqrt{3}\theta\,|x - x'|)\,\exp(-\sqrt{3}\theta\,|x - x'|)\ \ \text{(Matérn 3/2)}\,, \tag{8}$$

$$K_{1,\theta}(x,x') = [1 + \sqrt{5}\theta\,|x - x'| + 5\theta^2\,|x - x'|^2/3]\,\exp(-\sqrt{5}\theta\,|x - x'|)\ \ \text{(Matérn 5/2)}\,. \tag{9}$$

We take $\theta = \theta_0 = 1$ in $K_{0,\theta}$ and adjust $\theta = \theta_1$ in $K_{1,\theta}$ to minimise $\phi_{2\,[K_{0,\theta_0},K_{1,\theta_1}]}(\mu)$ defined by Eq. (13) in Section 4 with μ the uniform measure on $[0,1]$, which gives $\theta_1 \simeq 1.1275$. The left

panel of Figure 3 shows $K_{0,\theta_0=1}(x,0)$ and $K_{1,\theta}(x,0)$ for $\theta = 1$ and $\theta = \theta_1$ when $x \in [0,1]$. The right panel presents $\phi_B(x)$ and $\phi_{KL}(x)$ for the same $n = 11$-point equally spaced design $\mathbf{X}_n$ as in Example 1 and $x \in [0,1]$ for $K_{0,1}$ and $K_{1,1.1275}$ (the value of $\phi_A(x)$ does not exceed $0.65\,10^{-4}$ and is not shown). The behaviours of $\phi_B(x)$ and $\phi_{KL}(x)$ are now different in different intervals $[x_i, x_{i+1}]$ (they remain symmetric with respect to $1/2$, however), the maximum of $\phi_{KL}(x)$ is obtained at the central point x_5. The behaviour of $\phi_{KL}(\cdot)$ could be related to the fact that discriminating between K_0 and K_1 amounts to estimating the smoothness of the realisation, which requires that some design points are close to each other.

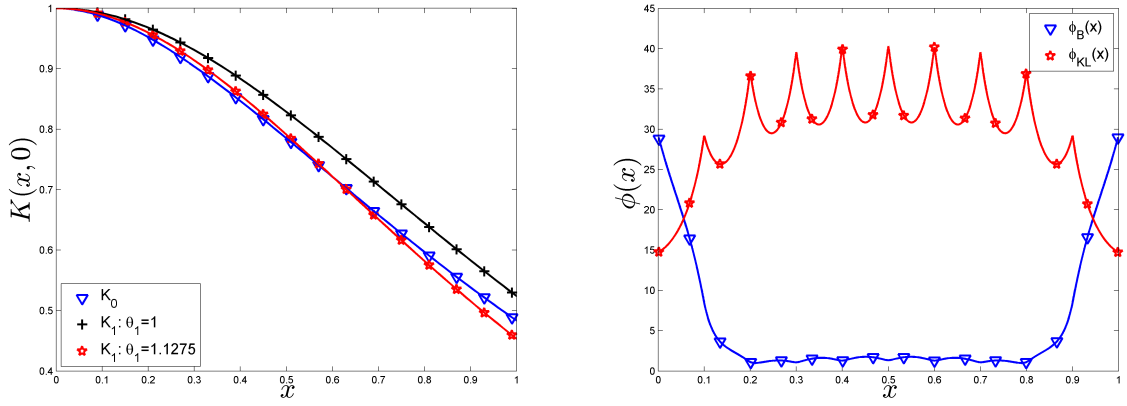

Figure 3: Left: $K_{0,1}(x,0)$, $K_{1,1}(x,0)$ and $K_{1,1.1275}(x,0)$, $x \in [0,1]$. Right: $\phi_B(x)$ and $\phi_{KL}(x)$ for $x \in [0,1]$ and the same 11-point equally spaced design $\mathbf{X}_n = \{0, 1/10, 2/10, \ldots, 1\}$ as in Example 1, with $K_{0,1}$ and $K_{1,1.1275}$.

4 Distance-based discrimination

We will now consider criteria which are directly based on the discrepancies of the covariance kernels. Ideally those should be simpler to compute and still exhibit reasonable efficiencies and some similar properties. The starting point is again the use of the log-likelihood ratio criterion to choose among the two models. Assuming that the random field is Gaussian, the probability densities of observations $\mathbf{Y}_n$ for the two models are

$$\varphi_{n,i}(\mathbf{Y}_n) = \frac{1}{(2\pi)^{n/2}\,\det^{1/2}\mathbf{K}_{n,i}}\,\exp\left[-\frac{1}{2}\,\mathbf{Y}_n^\top \mathbf{K}_{n,i}^{-1}\mathbf{Y}_n\right]\,,\ i = 0, 1\,.$$

The expected value of the log-likelihood ratio $L_n = \log\varphi(\mathbf{Y}_n|0) - \log\varphi(\mathbf{Y}_n|1)$ under model 0 is

$$\mathsf{E}_0\{L_n\} = \frac{1}{2}\log\det(\mathbf{K}_{n,1}\mathbf{K}_{n,0}^{-1}) - \frac{n}{2} + \frac{1}{2}\,\mathrm{trace}(\mathbf{K}_{n,0}\mathbf{K}_{n,1}^{-1})$$

and similarly

$$\mathsf{E}_1\{L_n\} = \frac{1}{2}\log\det(\mathbf{K}_{n,1}\mathbf{K}_{n,0}^{-1}) + \frac{n}{2} - \frac{1}{2}\,\mathrm{trace}(\mathbf{K}_{n,1}\mathbf{K}_{n,0}^{-1})\,.$$

A good discriminating design should make the difference $\mathsf{E}_0\{L_n\} - \mathsf{E}_1\{L_n\}$ as large as possible; that is, we should choose $\mathbf{X}_n$ that maximises

$$\begin{aligned}\Phi_{KL\,[K_0,K_1]}(\mathbf{X}_n) = \mathsf{E}_0\{L_n\} - \mathsf{E}_1\{L_n\} &= \frac{1}{2}\left[\mathrm{trace}(\mathbf{K}_{n,0}\mathbf{K}_{n,1}^{-1}) + \mathrm{trace}(\mathbf{K}_{n,1}\mathbf{K}_{n,0}^{-1})\right] - n \\ &= 2\,D_{KL}(\varphi_{n,0}, \varphi_{n,1})\,, \end{aligned} \tag{10}$$

i.e. twice the symmetric Kullback-Leibler divergence between the normal distributions with densities $\varphi_{n,0}$ and $\varphi_{n,1}$.

We may enforce the normalisation $\sigma_0^2 = \sigma_1^2 = 1$ and choose the θ_i to make the two kernels most similar in the sense of the criterion $\Phi(\cdot,\cdot)$ considered; that is, maximise

$$\min_{\theta_0\in\Theta_0,\,\theta_1\in\Theta_1} \Phi_{KL\,[K_0,K_1]}(\mathbf{X}_n)\,. \tag{11}$$

The choice of Θ_0 and Θ_1 is important; in particular, unconstrained minimisation over the θ_i could make both kernels completely flat or on the opposite close to Dirac distributions. It may thus be preferable to fix θ_0 and minimise over θ_1 without constraints. Also, the Kullback-Leibler distance is sensitive to kernel matrices being near singularity, which might happen if design points are very close to each other. Pronzato et al. (2019) suggest a family of criteria based on matrix distances derived from Bregman divergences between functions of covariance matrices from Kiefer's φ_p-class of functions (Kiefer, 1974). If $p \in (0,1)$, these criteria are rather insensitive to eigenvalues close or equal to zero. Alternatively, they suggest criteria computed as Bregman divergences between squared volumes of random k-dimensional simplices for $k \in \{2,\dots,d-1\}$, which have similar properties.

The index n is omitted in the following and we consider fixed parameters for both kernels. The Fréchet-distance criterion

$$\Phi_{F\,[K_0,K_1]}(\mathbf{X}_n) = \mathrm{trace}\left[\mathbf{K}_0 + \mathbf{K}_1 - 2\,(\mathbf{K}_0\mathbf{K}_1)^{1/2}\right], \tag{12}$$

related to the Kantorovich (Wasserstein) distance, seems of particular interest due to the absence of matrix inversion. The expression is puzzling since the two matrices do not necessarily commute, but the paper Dowson and Landau (1982) is illuminating.

Other matrix "entry-wise" distances will be considered, in particular the one based on the (squared) Frobenius norm,

$$\Phi_{2\,[K_0,K_1]}(\mathbf{X}_n) = \mathrm{trace}\left(\mathbf{K}_0^2 + \mathbf{K}_1^2 - 2\,\mathbf{K}_0\mathbf{K}_1\right) = \mathrm{trace}\left[(\mathbf{K}_0 - \mathbf{K}_1)^2\right],$$

which corresponds to the substitution of $\mathbf{K}_i^2$ for $\mathbf{K}_i$ in (12) for $i = 0, 1$. Denote more generally

$$\Phi_{p\,[K_0,K_1]}(\mathbf{X}_n) = \|\mathbf{K}_1 - \mathbf{K}_0\|_p^p = \sum_{i,j=1}^{n} |\{\mathbf{K}_1 - \mathbf{K}_0\}_{i,j}|^p = \mathbf{1}_n^\top |\mathbf{K}_1 - \mathbf{K}_0|^{\odot p}\mathbf{1}_n\,,\ p > 0\,,$$

where $\mathbf{1}_n$ is the n-dimensional vector with all components equal to 1, the absolute value is applied entry-wise and $^{\odot p}$ denotes power p applied entry-wise.

Figure 4 shows the values of the criteria $\Phi_{i\,[K_{0,1},K_{1,\theta}]}$, $i = 1, 2$, $\Phi_{F\,[K_{0,1},K_{1,\theta}]}$ and $\Phi_{KL\,[K_{0,1},K_{1,\theta}]}$ as functions of θ for the two kernels $K_{0,\theta}$ and $K_{1,\theta}$ given by (8) and (9) and the same regular design as in Example 1: $x_i = (i-1)/(n-1)$, $i = 1,\dots,11$. The criteria are re-scaled so that their maximum equals one on the interval considered for θ. Note the similarity between

$\Phi_{2\,[K_{0,1},K_{1,\theta}]}(\mathbf{X}_n)$ and $\Phi_{F\,[K_{0,1},K_{1,\theta}]}(\mathbf{X}_n)$ and the closeness between the distance-minimising θ for Φ_1, Φ_2 and Φ_F. Also note the good agreement with the value $\theta_1 \simeq 1.1275$ that minimises $\phi_{2\,[K_{0,1},K_{1,\theta_1}]}(\mu)$ from Eq. (13), see Example 3. The optimal θ for $\Phi_{KL\,[K_{0,1},K_{1,\theta}]}(\mathbf{X}_n)$ is much different, however, showing that the criteria do not necessarily agree between them.

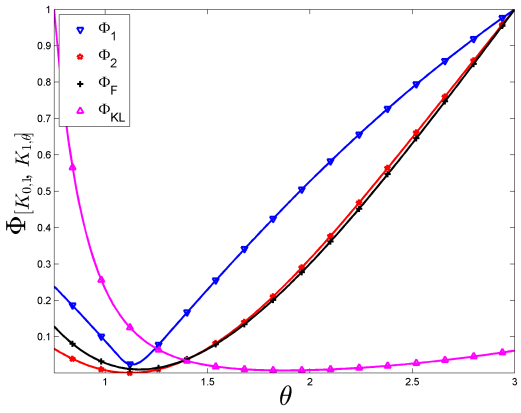

Figure 4: $\Phi_{i\,[K_{0,1},K_{1,\theta}]}(\mathbf{X}_n)$, $i = 1, 2$, $\Phi_{F\,[K_{0,1},K_{1,\theta}]}(\mathbf{X}_n)$ and $\Phi_{KL\,[K_{0,1},K_{1,\theta}]}(\mathbf{X}_n)$ as functions of $\theta \in [0.75, 3]$ for the same 11-point equally spaced design $\mathbf{X}_n$ as in Example 1 and $K_{0,\theta}$, $K_{1,\theta}$ given by (8) and (9), respectively.

An interesting feature of the family of criteria $\Phi_{p\,[K_0,K_1]}(\cdot)$, $p > 0$, is that they extend straightforwardly to a design measure version. Indeed, defining ξ_n as the empirical measure on the points in $\mathbf{X}_n$, $\xi_n = (1/n)\sum_{i=1}^n \delta_{x_i}$, we can write

$$\Phi_{p\,[K_0,K_1]}(\mathbf{X}_n) = n^2\,\phi_{p\,[K_0,K_1]}(\xi_n)\,,$$

where we define, for any design (probability) measure on $\mathscr{X}$,

$$\phi_p(\xi) = \phi_{p\,[K_0,K_1]}(\xi) \quad = \quad \int_{\mathscr{X}^2} |K_1(x,x') - K_0(x,x')|^p \,\mathrm{d}\xi(x)\mathrm{d}\xi(x')\,. \tag{13}$$

Denote by $F_{p\,[K_0,K_1]}(\xi;\nu)$ the directional derivative of $\phi_{p\,[K_0,K_1]}(\cdot)$ at ξ in the direction ν,

$$F_{p\,[K_0,K_1]}(\xi;\nu) = \lim_{\alpha\to 0^+} \frac{\phi_{p\,[K_0,K_1]}[(1-\alpha)\xi + \alpha\nu] - \phi_{p\,[K_0,K_1]}(\xi)}{\alpha}\,.$$

Direct calculation gives

$$F_{p\,[K_0,K_1]}(\xi;\nu) = 2\left[\int_{\mathscr{X}^2} |K_1(x,x') - K_0(x,x')|^p \,\mathrm{d}\nu(x)\mathrm{d}\xi(x') - \phi_{p\,[K_0,K_1]}(\xi)\right]\,,$$

and thus in particular

$$F_{p\,[K_0,K_1]}(\xi;\delta_x) = 2\left[\int_{\mathscr{X}} |K_1(x,x') - K_0(x,x')|^p \,\mathrm{d}\xi(x') - \phi_{p\,[K_0,K_1]}(\xi)\right]\,.$$

One can easily check that the criterion is neither concave nor convex in general (as the matrix $|\mathbf{K}_1 - \mathbf{K}_0|^{\odot p}$ can have both positive and negative eigenvalues), but we nevertheless have a necessary condition for optimality.

Theorem 1. *If the probability measure ξ^* on $\mathscr{X}$ maximises $\phi_{p\,[K_0,K_1]}(\xi)$, then*

$$\forall x \in \mathscr{X}, \ \int_{\mathscr{X}} |K_1(x,x') - K_0(x,x')|^p \, \mathrm{d}\xi^*(x') \leq \phi_{p\,[K_0,K_1]}(\xi^*) \, .$$

Moreover, $\int_{\mathscr{X}} |K_1(x,x') - K_0(x,x')|^p \, \mathrm{d}\xi^*(x') = \phi_{p\,[K_0,K_1]}(\xi^*)$ for ξ^*-almost every $x \in \mathscr{X}$.

This suggests the following simple incremental construction: at iteration n, with $\mathbf{X}_n$ the current design and ξ_n the associated empirical measure, choose $x_{n+1} \in \operatorname{Arg\,max}_{x\in\mathscr{X}} F_{p\,[K_0,K_1]}(\xi_n;\delta_x) = \mathbf{1}_n^\top |\mathbf{k}_{n,0}(x) - \mathbf{k}_{n,1}(x)|^{\odot p}$. It will be used in the numerical example of Section 6.2.

5 Optimal design measures

In this section we explain why the determination of optimal design measures maximising $\phi_p(\xi)$ is generally difficult, even when limiting ourselves to the satisfaction of the necessary condition in Theorem 1. At the same time, we can characterise measures that are approximately optimal for large p.

We assume that the two kernels are isotropic, i.e., such that $K_i(x,x') = \Psi_i(\|x - x'\|)$, $i = 0, 1$, and that the functions Ψ_i are differentiable except possibly at 0 where they only admit a right derivative. We define $\psi(t) = |\Psi_1(t) - \Psi_0(t)|$, $t \in \mathbb{R}^+$, and assume that the kernels have been normalised so that $K_0(x,x) = K_1(x,x)$; that is, $\psi(0) = 0$. Also, we only consider the case where the function $\psi(\cdot)$ has a unique global maximum on $\mathbb{R}^+$. This assumption is not very restrictive. Consider again the two Matérn kernels (8) and (9). Figure 5 shows the evolution of $\psi^2(t)$ for $K_0 = K_{0,1}$ and $K_1 = K_{1,\theta_1}$ with two different values of θ_1: $\theta_1 = 1$ and $\theta_1 \simeq 1.1275$; the latter minimises $\phi_{2\,[K_{0,1},K_{1,\theta}]}(\mu)$ for μ being the uniform measure on $[0,1]$.

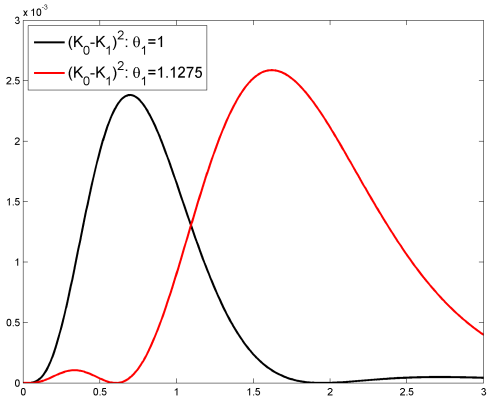

Figure 5: $\psi^2(t)$ for $K_0 = K_{0,1}$ and $K_1 = K_{1,\theta_1}$ with two different values of θ_1.

In the following, we shall consider normalised functions $\psi(\cdot)$, such that $\max_{t\in\mathbb{R}^+} \psi(t) = 1$. We denote by Δ the (unique) value such that $\psi(\Delta) = 1$. On Figure 5, $\Delta \simeq 0.7$ when $K_1 = K_{1,1}$.

5.1 A simplified problem with an explicit optimal solution

Consider the extreme case where $\psi = \psi_*$ defined by

$$\psi_*(t) = \begin{cases} 1 & \text{if } t = \Delta, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Note that $\psi_*^p(t) = \psi_*(t)$ for any $p > 0$; we can thus restrict our attention to $p = 1$ for the maximisation of $\phi_p(\xi)$ defined by (13); that is, we consider

$$\phi_1(\xi) = \int_{\mathscr{X}^2} \psi_*(\|x - x'\|) \, \mathrm{d}\xi(x) \mathrm{d}\xi(x') \, .$$

Theorem 2. *When $\psi = \psi_*$ and $\mathscr{X} \subset \mathbb{R}^d$ is large enough to contain a regular d simplex with edge length Δ, any measure ξ^* allocating weight $1/(d+1)$ at each vertex of such a simplex maximises $\phi_1(\xi)$, and $\phi_1(\xi^*) = d/(d+1)$.*

Proof. Since $\phi_1(\xi) = 0$ when ξ is continuous with respect to the Lebesgue measure on $\mathscr{X}$, we can restrict our attention to measures without any continuous component. Assume that $\xi = \sum_{i=1}^n w_i \delta_{x_i}$, with $w_i \geq 0$ for all i and $\sum_{i=1}^n w_i = 1$, $n \in \mathbb{N}$. Consider the graph $\mathcal{G}(\xi)$ having the x_i as vertices, with an edge (i, j) connecting x_i and x_j if and only if $\|x_i - x_j\| = \Delta$. We have

$$\phi_1(\xi) = \sum_{(i,j) \in \mathcal{G}(\xi)} w_i w_j,$$

and Theorem 1 of Motzkin and Straus (1965) implies that $\phi_1(\xi)$ is maximum when ξ is uniform on the maximal complete subgraph of $\mathcal{G}(\xi)$. The maximal achievable order is $d+1$, obtained when the x_i are the vertices of a regular simplex in $\mathscr{X}$ with edge length Δ. Motzkin and Straus (1965) also indicate in their Theorem 1 that $\phi_1(\xi^*) = 1 - 1/(d+1)$. This is easily recovered knowing that $\mathcal{G}(\xi^*)$ is fully connected with order $d+1$. Indeed, we then have

$$\phi_1(\xi) = \sum_{i=1}^{d+1} w_i \sum_{\substack{j=1 \\ j \neq i}}^{d+1} w_j = \sum_{\substack{i,j=1 \\ j \neq i}}^{d+1} w_i w_j = 1 - \sum_{i=1}^{d+1} w_i^2,$$

which is maximum when all w_i equal $1/(d+1)$. □

5.2 Optimal designs for $\psi(t) = |\Psi_1(t) - \Psi_0(t)|$

The optimal designs of Theorem 2 are natural candidates for being optimal when we return to the case of interest $\psi(t) = |\Psi_1(t) - \Psi_0(t)|$. In the light of Theorem 1, for a given probability measure ξ on $\mathscr{X}$, we consider the function

$$\delta_\xi(x) = \int_{\mathscr{X}} \psi^p(\|x - x'\|) \, \mathrm{d}\xi(x') - \phi_p(\xi),$$

which must satisfy $\delta_\xi(x) \leq 0$ for all $x \in \mathscr{X}$ when ξ is optimal. For an optimal measure ξ^* as in Theorem 2, with support $x_1, \ldots, x_{d+1}$ forming a regular d-simplex, we have

$$\delta_{\xi^*}(x) = \frac{1}{d+1} \left[\sum_{i=1}^{d+1} \psi^p(\|x - x_i\|) - d \right].$$

One can readily check that $\delta_{\xi^*}(x_i) = 0$ for all i (as $\psi(\|x_i - x_j\|) = \psi(\Delta) = 1$ for $i \neq j$ and $\psi(0) = 0$). Moreover, since $\psi(\cdot)$ is differentiable everywhere except possibly at zero, when $p > 1$ the gradient of $\delta_{\xi^*}(x)$ equals zero at each x_i. However, these $d+1$ stationary points may sometimes correspond to local minima — a situation when of course ξ^* is not optimal. The left panel of Figure 6 shows an illustration ($d = 2$) for $p = 1.5$, $K_0(x, x') = \exp(-\|x - x'\|)$ and K_1 being the Matérn 5/2 kernel $K_{1,1}$. The measure ξ^* is supported at the vertices of the equilateral triangle $(0,0), (\Delta, 0), (\Delta/2, \sqrt{(}3)\Delta/2)$ (indicated in blue on the figure), with $\Delta \simeq 0.53$ (the value where $\psi(\cdot)$ is maximum). Here the x_i correspond to local minima of $\delta_{\xi^*}(x)$, $\psi(\cdot)$ is not differentiable at zero but $p > 1$ so that $\delta_{\xi^*}(\cdot)$ is differentiable.

When $p \to \infty$, $\psi^p(\cdot)$ approaches the (discontinuous) function $\psi_*(\cdot)$, suggesting that ξ^* may become close to being optimal for ϕ_p when p is large enough. However, when $\mathscr{X}$ is large, ξ^* is never truly optimal, no matter how large p is. Indeed, suppose that $\mathscr{X}$ contains a point x_* corresponding to the symmetric of a vertex x_k of the simplex defining the support of ξ^* with respect to the opposite face of that simplex. Direct calculation gives

$$L = \|x_k - x_*\| = 2\,\Delta\,\left(\frac{d+1}{2\,d}\right)^{1/2}.$$

The right panel of Figure 6 shows an illustration for K_0 and K_1 being the Matérn 3/2 and Matérn 5/2 kernels $K_{0,1}$ and $K_{1,1}$, respectively. The measure ξ^* is supported at the vertices of the equilateral triangle with vertices $(0,0), (\Delta, 0), (\Delta/2, \sqrt{(}3)\Delta/2)$ with now $\Delta \simeq 0.7$. At the point x_*, symmetric to x_k, indicated in red on the figure, we have

$$\begin{aligned} \delta_{\xi^*}(x_*) &= \frac{1}{d+1}\left[\sum_{\substack{i=1\\ i\neq k}}^{d+1} \psi^p(\|x_* - x_i\|) + \psi^p(\|x_* - x_k\|) - d\right] \\ &= \frac{1}{d+1}\,\psi^p(L) > 0\,, \end{aligned} \tag{15}$$

where the second equality follows from $\|x_* - x_i\| = \Delta$ for all $i \neq k$, implying that ξ^* is not optimal. Another, more direct, proof of the non-optimality of ξ^* is to consider the measure $\widehat{\xi}$ that sets weights $1/(d+1)$ at all $x_i \neq x_k$ and weights $1/[2(d+1)]$ at x_k and its symmetric x_*. Direct calculation gives

$$\phi_p(\widehat{\xi}) = \frac{d}{d+1}\,\left(1 - \frac{1}{d+1}\right) + \frac{2}{2\,(d+1)}\left[\frac{d}{d+1} + \frac{1}{2\,(d+1)}\,\psi^p(L)\right].$$

The first term on the right-hand side comes from the d vertices x_i, $i \neq k$, each one having weight $1/(d+1)$ and being at distance Δ of all other vertices, those having total weight $1-1/(d+1)$. The second term comes from the two symmetric points x_k and x_*, each one with weight $1/[2(d+1)]$. Each of these two points is at distance Δ from d vertices with weights $1/(d+1)$ and at distance L of the other opposite point with weight $1/[2(d+1)]$. We get after simplification

$$\phi_p(\widehat{\xi}) = \frac{d}{d+1} + \frac{\psi^p(L)}{2\,(d+1)^2} > \phi_p(\xi^*) = \frac{d}{d+1},$$

showing that ξ^* is not optimal. Note that, for symmetry reasons, the design $\widehat{\xi}$ is not optimal for large enough $\mathscr{X}$. The determination of a truly optimal design seems very difficult. In the

simplified problem of Section 5.1, where the criterion is based on the function ψ_* defined by (14), the measures ξ^* and $\widehat{\xi}$ supported on $d+1$ and $d+2$ points, respectively, have the same criterion value $\phi_p(\xi^*) = \phi_p(\widehat{\xi}) = d/(d+1)$ for all $p > 0$.

Although ξ^* is not optimal, since $\psi(\|x_* - x_k\|) < 1$ (as $\psi(t)$ takes its maximum value 1 for $t = \Delta$), (15) suggests that ξ^* may be only marginally suboptimal when p is large enough. Moreover, as the right panel of Figure 6 illustrates, a design ξ^* supported on a regular simplex is optimal provided that $\mathscr{X}$ is small enough and p is large enough to make $\delta_{\xi^*}(x)$ concave at each x_i (for symmetry reasons, we only need to check concavity at one vertex). In fact, $p > 2$ is sufficient. Indeed, assuming that $p > 2$ and that $\psi(\cdot)$ is twice differentiable everywhere, with second-order derivative $\psi''(\cdot)$, except possibly at zero, direct calculation gives

$$\left.\frac{d^2\delta_{\xi^*}(x)}{dxdx^\top}\right|_{x=x_1} = \frac{1}{d+1}\,\frac{p\,\psi^{p-1}(\Delta)\psi''(\Delta)}{\Delta^2}\sum_{i=2}^{d+1}(x_1 - x_i)(x_1 - x_i)^\top,$$

which is negative-definite (since $\psi''(\Delta) < 0$, $\psi(\cdot)$ being maximal at Δ). The right panel of Figure 6 gives an illustration. Note that $p < 2$ on the left panel, and the x_i correspond to local minimas of $\delta_{\xi^*}(\cdot)$. Figure 7 shows a plot of $\delta_{\xi^*}(x)$ for $p = 2$ and K_0 and K_1 being the Matérn 3/2 and Matérn 5/2 kernels $K_{0,1}$ and $K_{1,1.07}$, respectively, suggesting that the form of optimal designs may be in general quite complicated.

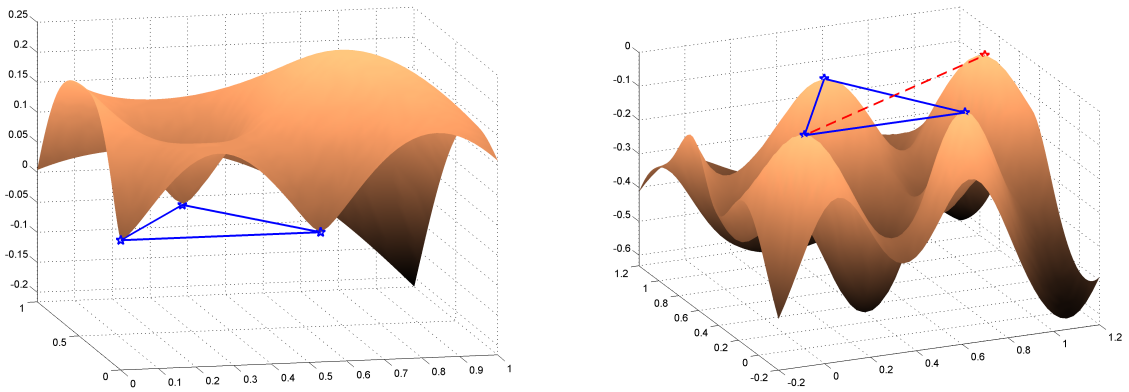

Figure 6: Surface plot of $\delta_{\xi^*}(x)$ ($x \in \mathbb{R}^2$), the support of ξ^* corresponds to the vertices of the equilateral triangle in blue. Left: $K_0(x, x') = \exp(-\|x - x'\|)$ and $K_1 = K_{1,1}$ ($\Delta \simeq 0.53$), $p = 1.5$; Right: $K_0 = K_{0,1}$, $K_1 = K_{1,1}$ ($\Delta \simeq 0.7$), $p = 10$; the red point x_* is the symmetric of the origin $(0, 0)$ with respect to the opposite side of the triangle.

6 A numerical example

6.1 Exact designs

In this section, we consider numerical evaluations of designs resulting from the prediction-based and distance-based criteria. Here, the rival models are the isotropic versions of the covariance kernels used in Example 3 (Section 3.2) for the design space $\mathscr{X} = [0, 10]^2$, discretised at $n = 25$

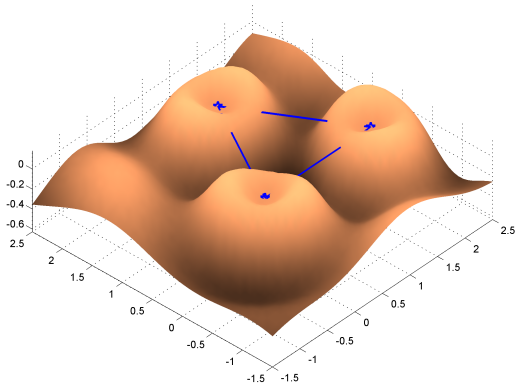

Figure 7: Surface plot of $\delta_{\xi^*}(x)$ ($x \in \mathbb{R}^2$), the support of ξ^* corresponds to the vertices of the equilateral triangle in blue: $K_0 = K_{0,1}$, $K_1 = K_{1,1.07}$ ($\Delta \simeq 1.92$), $p = 2$.

equally spaced points in each dimension. For an agreement on the setting of correlation lengths in both kernels, we have applied a minimisation procedure. Specifically, we have taken $\theta = \theta_0 = 1$ in $K_{0,\theta}(x, x')$ and adjusted the parameter in the second kernel minimising each of the distance-based criteria for the design $\mathbf{X}_{625}$ corresponding to the full grid. This resulted in $\theta_1 = 1.0047$, 1.0285, 1.0955 and 1.3403, respectively, for Φ_F, Φ_1, Φ_2 and Φ_{KL}. We have finally chosen $\theta_1 = 1.07$, which seems to be compatible with the above values.

The left panel in Figure 8 shows the plot of the two Matérn covariance functions at the assumed parameter values. This plot illustrates the similarity of the kernels which we aim to discriminate. The right panel in the figure refers to the plot of the absolute difference between the covariance kernels. The red line corresponds to the distance where the absolute difference between them is maximal. This is denoted by Δ, which is equal to $\Delta = 1.92$ in this case.

The sequential approach is the only case where the observations $\mathbf{Y}_n$ corresponding to the previous design points $\mathbf{X}_n$ are used in the design construction. We use this information to estimate the parameter setting at each step. The (box)plots of the maximum likelihood (ML) estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ of the inverse correlation lengths θ_0 and θ_1 of $K_{0,\theta}(x, x')$ and $K_{1,\theta}(x, x')$, respectively, are presented in Figure 9. This refers to the case where the first kernel, Matérn 3/2, is the data generator. The $\hat{\theta}_0$ estimates converge to their null value, $\theta_0 = 1$, drawn as a red dashed line in the left panel of Figure 9, as expected due to the consistency of the ML estimator in this case. For the second kernel to be similar to the first one (i.e., less smooth), the $\hat{\theta}_1$ estimates have increased (see the right panel). The decrease of the correlation length causes the covariance kernel to drop faster as a function of distance. We defer from presenting the opposite case (where the Matérn 5/2 is the data generator), which is similar.

Apart from the methods applied in Section 4, we have considered some other static approaches for discrimination. D_s-optimal design is a natural candidate that can be applied in the distance-based fashion. For D_s-optimality, we require the general form of the Matérn covariance kernel, which is based on the modified Bessel function of the second kind (denoted by C_ν). It is

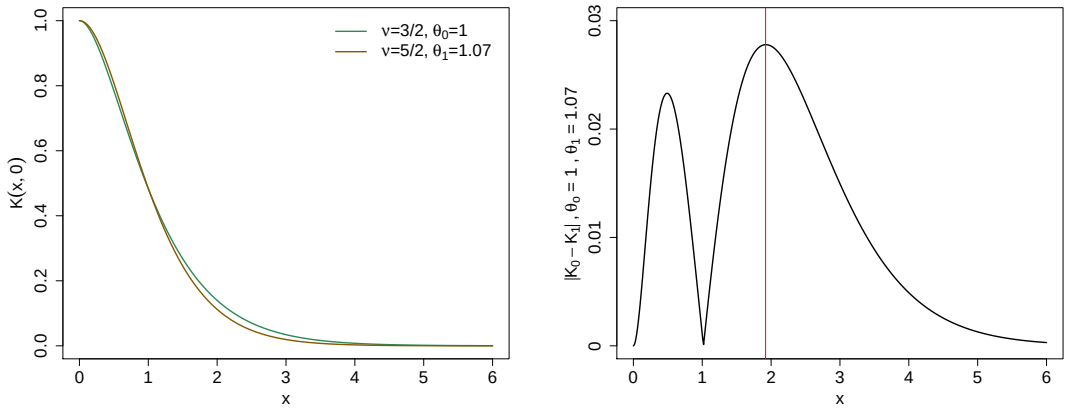

Figure 8: Left: Plot of the Matérn covariance functions at the assumed parameter setting. Right: $\psi(t) = |K_{0,\theta_0}(t,0) - K_{0,\theta_1}(t,0)|, (\theta_0 = 1, \theta_1 = 1.07)$.

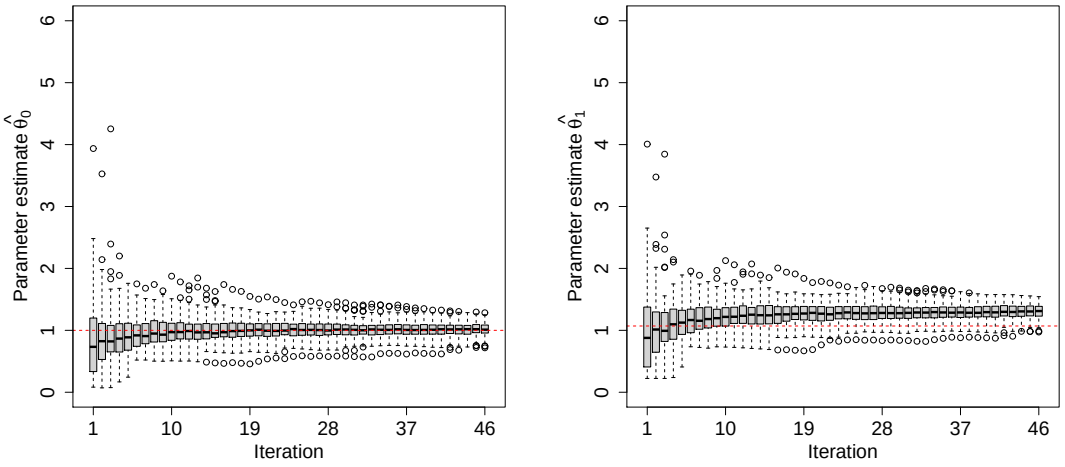

Figure 9: Maximum likelihood estimates of the correlation lengths in Matérn kernels.

given by

$$\mathbf{K}_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\ r\theta\right)^\nu C_\nu \left(\sqrt{2\nu}\ r\theta\right). \tag{16}$$

Smoothness, ν, is considered as the parameter of interest, while the correlation length θ is assumed as nuisance. The first off-diagonal element in the 2×2 information matrix, associated with the estimation of parameters $\boldsymbol{\theta} = (\theta, \nu)$, is

$$M(\mathbf{X}_n, \boldsymbol{\theta})_{12} = \frac{1}{2}\text{trace}\left\{\mathbf{K}_\nu^{-1}\frac{\partial \mathbf{K}_\nu}{\partial \theta}\mathbf{K}_\nu^{-1}\frac{\partial \mathbf{K}_\nu}{\partial \nu}\right\}, \tag{17}$$

see, e.g., Eq. (6.19) in Müller (2007). The other elements in the information matrix are calculated similarly. We have used the supplementary material of Lee et al. (2018) to compute the partial derivatives of the Matérn covariance kernel. Finally, the D_s-criterion is

$$\Phi_{D_s} = |M(\mathbf{X}_n, \boldsymbol{\theta})|/|M(\mathbf{X}_n, \boldsymbol{\theta})_{11}|, \tag{18}$$

where $M(\mathbf{X}_n, \boldsymbol{\theta})_{11}$ is the element of the information matrix corresponding to the nuisance parameter (i.e., in $M(\mathbf{X}_n, \boldsymbol{\theta})_{11}$ both partial derivatives are calculated with respect to θ). In the examples to follow we consider local D_s-optimal design; that is, the parameters θ and ν are set at given values.

From a Bayesian perspective, models can be discriminated optimally when the difference between the expected entropies of the prior and the posterior model probabilities is maximised. This criterion underlies a famous sequential procedure put forward by Box and Hill (1967) and Hill and Hunter (1969). Since such criteria typically cannot be computed analytically, several bounds were derived. The upper bound proposed by Box and Hill (1967) is equivalent to the symmetric Kullback-Leibler divergence Φ_{KL}. Hoffmann (2017) derives a lower bound based on a lower bound for the Kullback-Leibler divergence between a mixture of two normals, which is given by Eq. (21) and is denoted by Φ_Γ. Here, we assume equal prior probabilities. A more detailed account of Bayesian design criteria and their bounds is given in Appendix A.

Table 1 collects simulation results for the given example. We have included the sequential procedure (4) as a benchmark for orientation. For all other approaches the true parameter values are used in the covariance kernels. Concerning static (distance-based) designs based on maximisation of $\Phi_F, \Phi_1, \Phi_2, \Phi_{KL}, \Phi_\Gamma, \Phi_{D_s}$, for each design size considered we first built a an incremental design and then used a classical exchange-type algorithm to improve it. These designs are thus not necessarily nested, i.e., $\mathbf{X}_n \not\subset \mathbf{X}_{n'}$ for $n < n'$.

Each design of size n was then evaluated by generating $N = 100$ independent sets of n observations generated with the assumed true model, evaluating the likelihood functions for these sets of observations for both models, and then deciding for each set of observations which model has the higher likelihood value. The hit rate is the fraction of sets of observations where the assumed true model has the higher likelihood value. The procedure was repeated by assuming the other model to be the true one. The two hit rates are then averaged and stated in Table 1, which contains the results for all the criteria and design sizes we considered. For the special case of the sequential construction (4), the design path depends on the observations generated at the previously selected design points; that is, unlike for the other criteria, for a given design size n each random run produces a different design. To compute the hit rates for a particular n we used $N = 100$ independent runs of the experiment.

Table 1: Comparison of average hit rates in different methods for the first numerical example.

	Average hit rate									
Design size	5	6	7	8	9	10	20	30	40	50
Sequential (4)	0.500	0.535	0.540	0.595	0.570	0.640	0.695	0.715	0.740	0.770
ϕ_A	0.505	0.500	0.530	0.525	0.505	0.510	0.520	0.535	0.585	0.635
ϕ_B	0.520	0.545	0.575	0.585	0.615	0.650	0.785	0.875	0.900	0.910
ϕ_{KL}	0.520	0.545	0.575	0.585	0.615	0.650	0.785	0.870	0.915	0.925
Φ_F	0.580	**0.625**	**0.620**	0.625	0.670	**0.715**	0.795	**0.900**	0.925	0.950
Φ_1	0.525	0.520	0.555	0.540	0.550	0.610	0.725	0.890	0.910	0.920
Φ_2	0.525	0.520	0.555	0.540	0.550	0.610	0.715	0.860	0.890	0.910
Φ_{KL}	0.580	**0.625**	**0.620**	0.625	0.670	**0.715**	0.795	0.895	0.925	**0.955**
Φ_Γ	**0.595**	**0.625**	0.610	**0.645**	**0.675**	0.700	0.795	0.895	**0.935**	0.940
Φ_{D_s}	0.540	0.575	0.590	0.620	0.650	0.675	**0.805**	0.850	0.855	0.925

The hit rates reported in Table 1 reflect the discriminatory power of the corresponding designs. One can observe that Φ_F and as expected Φ_{KL} are outperforming in terms of hit rates. The Bayesian lower bound criterion Φ_Γ is similar to the symmetric Φ_{KL}. The sequential design strategy (4) does not behave as well as the outperforming ones. It is, however, the realistic scenario that one might consider in applications as it does not assume knowledge of the kernel parameters. The effect of this knowledge can thus be partially calibrated for by comparing the first line against the other criteria.

6.2 Optimal design measure for ϕ_p

Theorem 1 also allows the use of approximate designs as it presents a necessary condition for optimality of the family of criteria $\phi_p,\ p > 0$. This is more extensively discussed in the previous section. Here we present the numerical results for two specific cases of $p = 2$ and $p = 10$. To reach a design which might be numerically optimal (or at least nearly optimal), we have applied the Fedorov-Wynn algorithm (Fedorov, 1971; Wynn, 1970) on a dense regular grid of candidate points.

Numerical results show that for very small p (e.g., $p = 1$) explicit optimal measures are hard to derive. The left panel in Figure 10 presents the measure ξ_2^* obtained for ϕ_2. To construct ξ_2^*, we have first calculated an optimal design on a dense grid by applying 1000 iterations of the Fedorov-Wynn algorithm (see the comment following Theorem 1); the design measure obtained is supported on 9 grid points. We then applied a continuous optimisation algorithm (library NLopt (Johnson, 2021) through its R-interface `nloptr`) initialised at this 9-point design. The 9 support points of the resulting design measure ξ_2^* are independent of the grid size; they receive unequal weights, proportional to the disk areas on Figure 10-left. Any translation or rotation of ξ_2^* yields the same value of ϕ_2.

As the order p increases, we eventually reach an optimal measure with only three support points and equal weights. The right panel in Figure 10 corresponds to the optimal design measure computed for ϕ_{10}. This has, similarly as before, resulted from application of a continuous optimisation initialised at an optimal 3-point design calculated with the Fedorov-Wynn algorithm on a grid. This optimal design measure ξ^*_{10} has three support points, drawn as blue dots, with equal weights $1/3$ represented by the areas of the red disks. The blue line segments between every two locations have length $\Delta \simeq 1.92$, reflecting the ideal interpoint distance (see the right panel of Figure 8), in agreement with corresponding discussions in Section 5. Also here the optimal designs are rotationally and translationally invariant, and thus any design of such type is optimal as long as the design region is large enough to fit it.

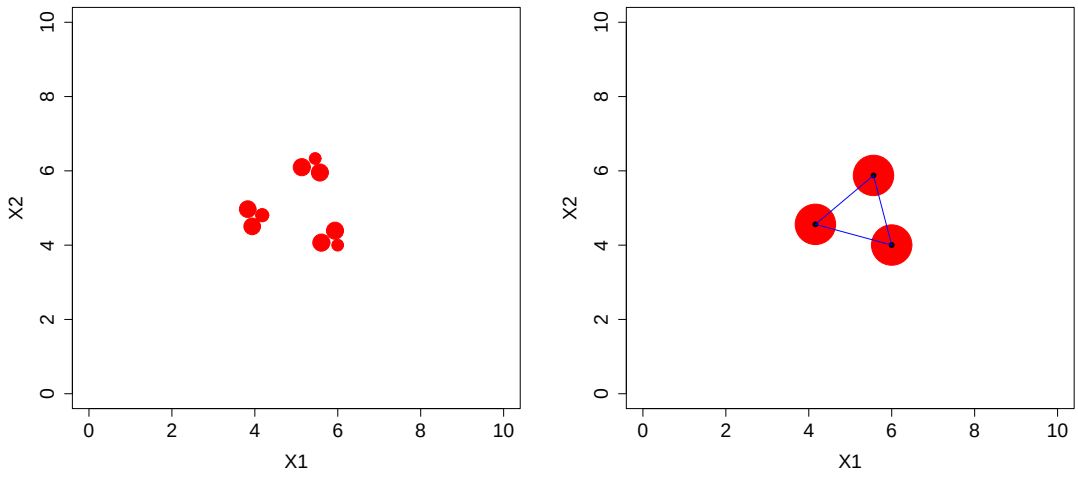

Figure 10: Left: The optimal measure for ϕ_2. Right: The optimal measure for ϕ_{10}.

7 Conclusions

In this paper we have considered the design problem for the discrimination of Gaussian process regression models. This problem differs considerably from the well-treated one in standard regression models and thus offers a multitude of challenges. While the KL-divergence is a straightforward criterion, it comes with the price of being computationally demanding and lacking convenient simplifications such as design measures. We have therefore introduced a family of criteria that allow such a simplification at least in special cases and have investigated its properties. We have also compared the performance of these and other potential criteria on several examples and see that KL-divergence can be effectively replaced by simpler criteria without much loss in efficiency. In particular designs based on the Fréchet-distance between covariance kernels seem to be competitive. Results from the approximate design computations indicate that for classical isotropic kernels, designs with $d + 1$ support points placed at the vertices of a simplex of suitable size are optimal for distance-based criteria ϕ_p when p is large enough.

Acknowledgments

This work was partly supported by project INDEX (INcremental Design of EXperiments) ANR-18-CE91-0007 of the French National Research Agency (ANR) and I3903-N32 of the Austrian Science Fund (FWF).

References

Atkinson, A. C. and Fedorov, V. V. (1975). The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70.

Box, G. E. P. and Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, 9(1):57–71.

Dowson, D. C. and Landau, B. V. (1982). The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455.

Fedorov, V. V. (1971). The design of experiments in the multiresponse case. *Theory of Probability & Its Applications*, 16(2):323–332.

Gramacy, R. B. (2020). *Surrogates : Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Chapman and Hall/CRC.

Heirung, T. A. N., Santos, T. L. M., and Mesbah, A. (2019). Model predictive control with active learning for stochastic systems with structural model uncertainty: Online model discrimination. *Computers & Chemical Engineering*, 128:128–140.

Hershey, J. R. and Olsen, P. A. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP '07*, volume 4, pages IV–317–IV–320.

Hill, W. J. and Hunter, W. G. (1969). A Note on Designs for Model Discrimination: Variance Unknown Case. *Technometrics*, 11(2):396–400.

Hino, H. (2020). Active learning: Problem settings and recent developments. Technical report, arXiv:2012.04225.

Hoffmann, C. (2017). *Numerical Aspects of Uncertainty in the Design of Optimal Experiments for Model Discrimination*. PhD thesis, Ruprecht-Karls-Universität Heidelberg.

Hunter, W. and Reiner, A. (1965). Designs for discriminating between two rival models. *Technometrics*, 7(3):307–323.

Johnson, S. G. (2021). The NLopt nonlinear-optimization package, http://github.com/stevengj/nlopt.

Karvonen, T. (2022). Asymptotic bounds for smoothness parameter estimates in Gaussian process interpolation. *arXiv preprint arXiv:2203.05400*.

Karvonen, T. and Oates, C. (2022). Maximum likelihood estimation in Gaussian process regression is ill-posed. *arXiv preprint arXiv:2203.09179*.

Karvonen, T., Wynne, G., Tronarp, F., Oates, C., and Särkkä, S. (2020). Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):926–958.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 2(5):849–879.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lee, X. J., Hainy, M., McKeone, J. P., Drovandi, C. C., and Pettitt, A. N. (2018). ABC model selection for spatial extremes models applied to South Australian maximum temperature data. *Computational Statistics & Data Analysis*, 128:128–144.

López-Fidalgo, J., Tommasi, C., and Trandafir, P. C. (2007). An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):231–242.

Motzkin, T. S. and Straus, E. G. (1965). Maxima for graphs and a new proof of a theorem of Turán. *Canadian Journal of Mathematics*, 17:533–540.

Müller, W. G. (2007). *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*. Springer Berlin, Heidelberg, 3 edition.

Olofsson, S., Deisenroth, M. P., and Misener, R. (2018). Design of experiments for model discrimination using Gaussian process surrogate models. In Eden, M. R., Ierapetritou, M. G., and Towler, G. P., editors, *13th International Symposium on Process Systems Engineering (PSE 2018)*, volume 44 of *Computer Aided Chemical Engineering*, pages 847–852. Elsevier.

Pronzato, L., Wynn, H. P., and Zhigljavsky, A. (2019). Bregman divergences based on optimal design criteria and simplicial measures of dispersion. *Statistical Papers*, 60(2):545–564.

Sauer, A., Gramacy, R. B., and Higdon, D. (2022). Active learning for deep Gaussian process surrogates. *Technometrics*, 0(0):1–15.

Schwaab, M., Luiz Monteiro, J., and Carlos Pinto, J. (2008). Sequential experimental design for model discrimination: Taking into account the posterior covariance matrix of differences between model predictions. *Chemical Engineering Science*, 63(9):2408–2419.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer, Heidelberg.

Wynn, H. P. (1970). The sequential generation of D-optimum experimental designs. *The Annals of Mathematical Statistics*, 41(5):1655–1664.

Appendix

A Notes on Box-Hill-Hunter Bayesian criteria for model discrimination between Gaussian random fields

Chapter 5 of Hoffmann (2017) contains an overview of Bayesian design criteria for model discrimination and some useful bounds on them. We assume there are M models $m_0, \dots, m_{M-1}$. The most common Bayesian design criterion for model discrimination has the following form:

$$\Phi_\Lambda(\mathbf{X}_k) = -\sum_{i=0}^{M-1} p(m_i)\log(p(m_i)) + \int_{\mathbf{Y}_k\in\mathcal{Y}} p(\mathbf{Y}_k)\sum_{i=0}^{M-1} p(m_i|\mathbf{Y}_k)\log(p(m_i|\mathbf{Y}_k))\,\mathrm{d}\mathbf{Y}_k, \tag{19}$$

where the data $\mathbf{Y}_k = (Y_1(x_1), \dots, Y_k(x_k))^\top$ are observed at the design $\mathbf{X}_k = (x_1, \dots, x_k)$, $p(m_i)$ denotes the prior and $p(m_i|\mathbf{Y}_k)$ the posterior model probability of model m_i and $p(\mathbf{Y}_k)$ is the marginal distribution of $\mathbf{Y}_k$ with respect to the models. Hence, this criterion is the (expected) difference of the model entropy and the conditional model entropy (conditional on the observations). The posterior model probability $p(m_i|\mathbf{Y}_k)$ is defined by

$$p(m_i|\mathbf{Y}_k) \propto p(\mathbf{Y}_k|m_i)p(m_i),$$

where $p(\mathbf{Y}_k|m_i)$ is the likelihood of model m_i (marginalised over the parameters), and $p(\mathbf{Y}_k)$ is given by

$$p(\mathbf{Y}_k) = \sum_{i=0}^{M-1} p(\mathbf{Y}_k|m_i)p(m_i).$$

The first term in (19) does not depend on the design and can therefore be ignored.

A common alternative formulation of criterion (19) is the one adopted by Box and Hill (1967) and Hill and Hunter (1969), which will henceforth be called Box-Hill-Hunter (BHH) criterion:

$$\Phi_\Lambda(\mathbf{X}_k) = \sum_{i=0}^{M-1} p(m_i)\int_{\mathbf{Y}_k\in\mathcal{Y}} p(\mathbf{Y}_k|m_i)\log\left(\frac{p(\mathbf{Y}_k|m_i)}{p(\mathbf{Y}_k)}\right)\mathrm{d}\mathbf{Y}_k. \tag{20}$$

In our case, if we assume point priors for the kernel parameters, we have

$$p(\mathbf{Y}_k|m_i) = \varphi(\mathbf{Y}_k|\boldsymbol{\eta}_{k,i}, \mathbf{K}_{k,i}),$$

where $\boldsymbol{\eta}_{k,i} = (\eta_{1,i}(x_1), \dots, \eta_{k,i}(x_k))^\top$ is the mean vector of model i at design $\mathbf{X}_k$, $\mathbf{K}_{k,i}$ is the $k \times k$ kernel matrix of model i with elements given by $\{\mathbf{K}_{k,i}\}_{j,l} = K_i(x_j, x_l)$, and $\varphi(\cdot|\boldsymbol{\eta}, \mathbf{K})$ is the normal pdf with mean vector $\boldsymbol{\eta}$ and variance-covariance matrix $\mathbf{K}$.

For example, for a static design involving n design points, we set $k = n$ and assume that $\boldsymbol{\eta}_{n,i} = \mathbf{0}$ for each design $\mathbf{X}_n$. The model probabilities $p(m_i)$ would just be the prior model probabilities before having collected any observations.

In a sequential design setting, where n observations $\mathbf{Y}_n$ have already been observed at locations $\mathbf{X}_n$ and we want to find the optimal design point x where to collect our next observation, we have $k = 1$ and set $\boldsymbol{\eta}_{k,i}$ to the conditional mean $\widehat{\eta}_{n,i}(x) = \mathbf{k}_{n,i}(x)^\top \mathbf{K}_{n,i}^{-1}\mathbf{Y}_n$ and $\mathbf{K}_{k,i}$ to the conditional variance $\rho^2_{n,i}(x) = K_i(x,x) - \mathbf{k}_{n,i}(x)^\top \mathbf{K}_{n,i}^{-1}\mathbf{k}_{n,i}(x)$, where $\mathbf{k}_{n,i}(x)^\top =$

$(K_i(x, x_1), \ldots, K_i(x, x_n))$, see Section 3.1. The prior model probabilities would have to be set to the posterior model probabilities given the already observed data:

$$p(m_i) = p(m_i|\mathbf{Y}_n) \propto \varphi(\mathbf{Y}_n|\,\mathbf{0}, \mathbf{K}_{n,i})\, p(m_i).$$

It follows that $p(\mathbf{Y}_k)$ is a mixture of normal distributions. The criterion representations (19) and (20) cannot be computed directly. However, several bounds have been developed for the criterion, the most famous being the classic upper bound derived by Box and Hill (1967).

A.1 Upper bound

The upper bound has the following form (see also Hoffmann (2017, Thm. 5.2, p. 168)):

$$\Phi_U(\mathbf{X}_k) = \frac{1}{2} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} p(m_i)p(m_j) \left\{ \left\|\boldsymbol{\eta}_{k,i} - \boldsymbol{\eta}_{k,j}\right\|^2_{\mathbf{K}^{-1}_{k,j}} + \operatorname{trace}\left(\mathbf{K}_{k,i}\mathbf{K}^{-1}_{k,j}\right) - n \right\}.$$

For $M = 2$, the formula simplifies to

$$\begin{aligned} \Phi_U(\mathbf{X}_k) \;=\; & \frac{1}{2} p(m_0)p(m_1) \Big\{ \left\|\boldsymbol{\eta}_{k,0} - \boldsymbol{\eta}_{k,1}\right\|^2_{\mathbf{K}^{-1}_{k,0}} + \left\|\boldsymbol{\eta}_{k,0} - \boldsymbol{\eta}_{k,1}\right\|^2_{\mathbf{K}^{-1}_{k,1}} \\ & + \operatorname{trace}\left(\mathbf{K}_{k,0}\mathbf{K}^{-1}_{k,1}\right) + \operatorname{trace}\left(\mathbf{K}_{k,1}\mathbf{K}^{-1}_{k,0}\right) - 2n \Big\}. \end{aligned}$$

This is equivalent to the symmetric Kullback-Leibler divergence that we use as the criterion Φ_{KL} (with $p(m_0) = p(m_1) = 1/2$ and $\boldsymbol{\eta}_{k,0} = \boldsymbol{\eta}_{k,1} = \mathbf{0}$) .

A.2 Lower bound

Hershey and Olsen (2007, Sec. 7) derive a lower bound for the Kullback-Leibler divergence between a mixture of two normals, see also Hoffmann (2017, Thm. 5.4 and Cor. 5.5, pp. 173–174). This result is then used by Hoffmann (2017) to find a lower bound for the BHH criterion $\Phi_\Lambda(\mathbf{X}_k)$ (Hoffmann, 2017, Thm. 5.9, p. 178). This lower bound is given by

$$\Phi_\Gamma(\mathbf{X}_k) = -\sum_{i=0}^{M-1} p(m_i) \log \left\{ \sum_{j=0}^{M-1} p(m_j) \exp\left(-\frac{1}{2}\boldsymbol{\Gamma}(\mathbf{X}_k)_{ij}\right) \right\},$$

where

$$\boldsymbol{\Gamma}(\mathbf{X}_k)_{ij} = \left\|\boldsymbol{\eta}_{k,i} - \boldsymbol{\eta}_{k,j}\right\|^2_{\mathbf{K}^{-1}_{k,j}} + \operatorname{trace}\left(\mathbf{K}_{k,i}\mathbf{K}^{-1}_{k,j}\right) - \log\det\left(\mathbf{K}_{k,i}\mathbf{K}^{-1}_{k,j}\right) - n.$$

For $M = 2$, as is the relevant case for our setup we get

$$\begin{aligned}
\Phi_\Gamma(\mathbf{X}_k) \;=\; & -p(m_0)\log\Big\{p(m_0) \\
& \quad + p(m_1)\exp\Big(-\frac{1}{2}\big[\big\|\boldsymbol{\eta}_{k,0}-\boldsymbol{\eta}_{k,1}\big\|^2_{\mathbf{K}_{k,1}^{-1}} + \operatorname{trace}\left(\mathbf{K}_{k,0}\mathbf{K}_{k,1}^{-1}\right) \\
& \qquad - \log\det\left(\mathbf{K}_{k,0}\mathbf{K}_{k,1}^{-1}\right) - n\big]\Big)\Big\} \\
& -p(m_1)\log\Big\{p(m_1) \\
& \quad + p(m_0)\exp\Big(-\frac{1}{2}\big[\big\|\boldsymbol{\eta}_{k,0}-\boldsymbol{\eta}_{k,1}\big\|^2_{\mathbf{K}_{k,0}^{-1}} + \operatorname{trace}\left(\mathbf{K}_{k,1}\mathbf{K}_{k,0}^{-1}\right) \\
& \qquad - \log\det\left(\mathbf{K}_{k,1}\mathbf{K}_{k,0}^{-1}\right) - n\big]\Big)\Big\} \qquad (21)
\end{aligned}$$

where $\varphi_i(\cdot) = \varphi(\cdot|\boldsymbol{\eta}_{k,i}, \mathbf{K}_{k,i})$, which we are also using to compute designs in Section 6.1 (again with $p(m_0) = p(m_1) = 1/2$ and $\boldsymbol{\eta}_{k,0} = \boldsymbol{\eta}_{k,1} = \mathbf{0}$).