Discrimination between Gaussian process models: active learning and static constructions

mODa 13
July 14th 2023

Markus Hainy

Department of Applied Statistics
Johannes Kepler University Linz

joint work with *E. Yousefi, L. Pronzato, W.G. Müller & H.P. Wynn*

JOHANNES KEPLER UNIVERSITY LINZ

Contents

Introduction

- Investigating the properties of static as well as incremental/sequential design criteria for discriminating between the correlation structure of (two) Gaussian process models.
- T-optimality (Atkinson and Fedorov, 1975) not applicable since it assumes iid normal data with constant variance.
- Instead, one may use the symmetrized Kullback-Leibler (KL) divergence between the two models as criterion.
- Symmetrized KL divergence computationally expensive if many design points $\Rightarrow$ develop new criteria inspired by Fréchet distance.
- For these new design criteria it is straightforward to introduce design measures and derive necessary conditions for optimality $\Rightarrow$ possible to apply approximate design methods.

Gaussian processes/fields

- Marginals of Gaussian fields at any subset of points/locations have a multivariate normal distribution.
- Frequently used e.g. as surrogate models for computer experiments (Gramacy, 2020) or in machine learning.
- Gaussian process regression / kriging (Stein, 1999): assume Gaussian process prior and obtain distribution for “true” function at unseen locations given observed points.

Gaussian processes/fields

- Marginals of Gaussian fields at any subset of points/locations have a multivariate normal distribution.
- Frequently used e.g. as surrogate models for computer experiments (Gramacy, 2020) or in machine learning.
- Gaussian process regression / kriging (Stein, 1999): assume Gaussian process prior and obtain distribution for "true" function at unseen locations given observed points.
- General setting and notation:
  - random field Z_x, indexed by $x \in \mathscr{X} \subset \mathbb{R}^d$.
  - $Y(x)$: realization of the random field.
  - $\mathsf{E}\{Z_x\} = 0 \; \forall \; x$ and $\mathsf{E}\{Z_x Z'_x\} = K(x, x') \; \forall \; (x, x') \in \mathscr{X}^2$.
  - kernel $K(x, x') = \sigma^2 \, f_\theta \left(\|x - x'\|\right)$: isotropic, continuous and decreasing function of the distance.
  - We do not consider repeated observations (no nugget effect).

Symmetrized Kullback-Leibler divergence

- Given two probability density functions $\varphi_0(y, \theta_0)$ and $\varphi_1(y, \theta_1)$, maximise the expected power of the likelihood ratio test if model 1 is the true model (see, e.g., López-Fidalgo et al., 2007):

$$\mathrm{E}_1(L) = \int \varphi_1(y, \theta_1) \log \left\{ \frac{\varphi_1(y, \theta_1)}{\varphi_0(y, \theta_0)} \right\} \mathrm{d}y = D_{KL}(\varphi_1 \| \varphi_0)$$

- Now maximise the power of the likelihood ratio test if model 0 is the true model:

$$\mathrm{E}_0(-L) = \int \varphi_0(y, \theta_0) \log \left\{ \frac{\varphi_0(y, \theta_0)}{\varphi_1(y, \theta_1)} \right\} \mathrm{d}y = D_{KL}(\varphi_0 \| \varphi_1)$$

- The symmetrized KL divergence is the average of these two divergences (see, e.g., Pronzato et al., 2019):

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

Symmetrized KL divergence for Gaussian random field

- The two models differ through their kernel functions.
- Given the n-point design $\mathbf{X}_n = (x_1, \ldots, x_n)$, construct the kernel matrix for model i ($i = 0, 1$), $\mathbf{K}_{n,i}$, as $\{\mathbf{K}_{n,i}\}_{j,k} = K(x_j, x_k)$ for $1 \le j, k \le n$.
- For the Gaussian random field,

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

- Therefore,

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

- Disadvantages: cumbersome and unstable computation (matrix inverses), no generalisation to design measures.

Fréchet distance and Φ_p criteria

- Consider alternatively the Fréchet distance, related to the Wasserstein distance (Dowson and Landau, 1982):

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

- This is also difficult to compute, but squaring all matrices gives

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

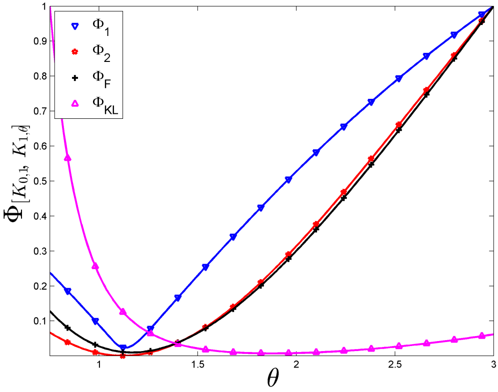
- *Idea*: introduce criteria $\Phi_{p\,[K_0,K_1]}$ defined as

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

Example setup

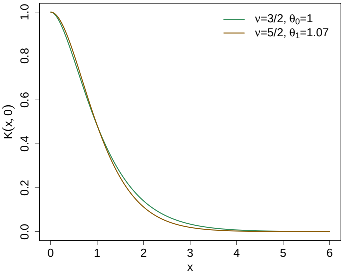
- $\mathscr{X} = [0,10]^2$, grid size: 25×25.
- Rival models: Matérn family ($\sigma^2 = 1$):
  - model 0: Matérn 3/2
  - model 1: Matérn 5/2
- We consider discrimination designs only for fixed parameters (locally optimum designs).
- Find inverse length-scales θ_0 and θ_1 where both models agree most:
  - Take $\theta_0 = 1$ in the first kernel, adjust the parameter in the second kernel minimizing Φ_1, Φ_2, Φ_F, and Φ_{KL} for the design $\mathbf{X}_{625}$.
  - Results: $\theta_1 = 1.0047,\ 1.0285,\ 1.0955,\ 1.3403$, resp.
  - We have finally used the setting $\theta = (1, 1.07)$.

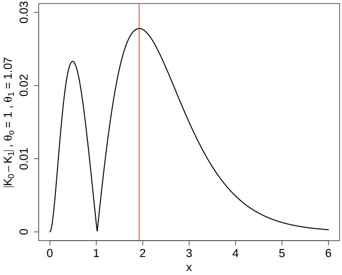
Selection of parameters



Φ_1, Φ_2, Φ_F, Φ_{KL} as functions of θ for an equally-spaced 11-point design.

Covariance functions at selected parameters

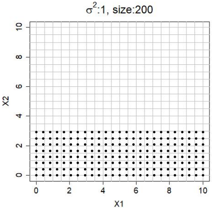




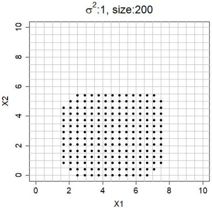
Left: Plot of the Matérn covariance functions.

Right: Absolute difference of covariance functions at distance x.

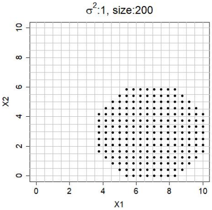
Exact designs for static criteria



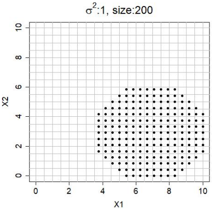
Φ_{KL}



Φ_F



Φ_1



Φ_2

Performance of criteria in simulation study

- For each design of size n, $N = 100$ independent sets of n observations are simulated from the "true" model 0.
- The hit rate is computed as the proportion of samples where the likelihood of model 0 is larger than the likelihood of model 1.
- The same is repeated for the case where model 1 is the correct model.
- The two hit rates are averaged to obtain the average hit rates.

	Average hit rate									
Design size	5	6	7	8	9	10	20	30	40	50
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**

Design measure version of Φ_p criterion

- Defining ξ_n as the empirical measure on the points in $\mathbf{X}_n$, $\xi_n = (1/n)\sum_{i=1}^{n} \delta_{x_i}$, one can write

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

where

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

Necessary condition for optimality

Theorem

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* $x \in \mathscr{X}$.

Simplified problem with explicit solution for optimum

- Let $K_i(x,x') = \Psi_i(\|x-x'\|)$, $i=0,1$, and $\psi(t) = |\Psi_1(t) - \Psi_0(t)|$, $t \in \mathbb{R}^+$.
- Then $\phi_p(\xi) = \int_{\mathscr{X}^2} \psi(\|x-x'\|)^p \, \mathrm{d}\xi(x)\mathrm{d}\xi(x')$.
- Consider the extreme case $\psi = \psi_*$ defined by

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

- Since $\psi_*(t)^p = \psi(t)^p$ for any $p > 0$, we only need to consider $p = 1$.

Simplified problem with explicit solution for optimum

- Let $K_i(x,x') = \Psi_i(\|x-x'\|)$, $i=0,1$, and $\psi(t) = |\Psi_1(t) - \Psi_0(t)|$, $t \in \mathbb{R}^+$.
- Then $\phi_p(\xi) = \int_{\mathscr{X}^2} \psi(\|x-x'\|)^p \, \mathrm{d}\xi(x)\mathrm{d}\xi(x')$.
- Consider the extreme case $\psi = \psi_*$ defined by

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

- Since $\psi_*(t)^p = \psi(t)^p$ for any $p > 0$, we only need to consider $p = 1$.

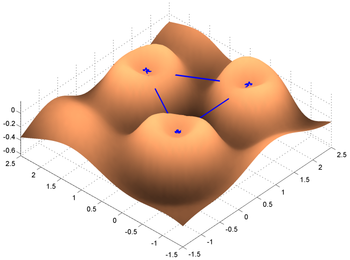
Theorem

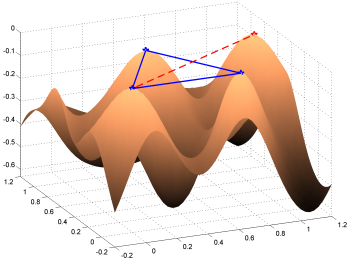
When $\psi = \psi_*$ and $\mathscr{X} \subset \mathbb{R}^d$ is large enough to contain a regular d simplex with edge length Δ, any measure ξ^* allocating weight $1/(d+1)$ at each vertex of such a simplex maximises $\phi_1(\xi)$, and $\phi_1(\xi^*) = d/(d+1)$.

Are results for $\psi = \psi_*$ generalisable?

- The results for $\psi = \psi_*$ **do not** generalise to the general function $\psi(t) = |\Psi_1(t) - \Psi_0(t)|$.
- Even if $p \to \infty$, one can show that one can always find a better design than ξ^*, given the design space is large enough.
- However, the simplex design might be close to optimal (at least for high p).

Illustration of directional derivative



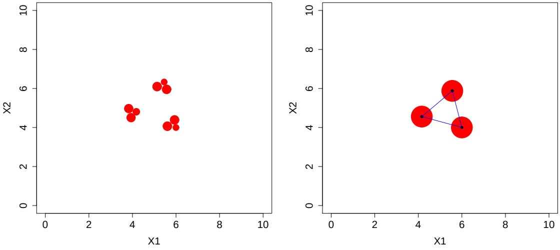


Surface plots of directional derivatives:
Left: $K_0 = K_{0,1}$, $K_1 = K_{1,1.07}$ ($\Delta \simeq 1.92$), $p = 2$.
Right: $K_0 = K_{0,1}$, $K_1 = K_{1,1}$ ($\Delta \simeq 0.7$), $p = 10$.

Numerical optimisation

Two-step approach:

1. Run Fedorov-Wynn algorithm (Fedorov, 1971; Wynn, 1970) for 1000 iterations using directional derivatives from necessary condition on a dense regular grid.
2. Run continuous optimisation algorithm for coordinates and design weights starting from design found in step 1.



Left: The optimal measure for ϕ_2. Right: The optimal measure for ϕ_{10}. $\theta_1 = 1.07$. The edge lengths of the triangles are $\Delta \simeq 1.92$.

Sequential/incremental designs

We also compared the static criteria to the following construction methods:

- Sequential design:
  put next observation(s) where symmetrized KL divergence between predictive distributions of the models differs most.

Sequential/incremental designs

We also compared the static criteria to the following construction methods:

- Sequential design:
  put next observation(s) where symmetrized KL divergence between predictive distributions of the models differs most.
- Incremental design:
  - Incrementally build design by putting next point where (normalised) differences between the prediction errors of the two models are largest.
  - Alternatively use symmetrized KL divergence (cond. on current design).
  - Theoretical investigations:
    - prediction-based and KL criteria tend to behave differently and depend on covering radius (CR) relative to correlation length.
    - KL divergence sometimes clearly better suited for discrimination (remains positive with decreasing CR).

Conclusion

- Introducing a new family of criteria which are simple to compute and allow for a formulation in terms of approximate design measures.
- In our examples, they lead to marginally worse performance than symmetrized KL divergence.
- For large p, designs with $d+1$ support points placed on the vertices of a simplex are often (close to) optimal for the new criteria, depending on the size of the design space.
- Not considered yet: behaviour under parameter uncertainty.

References

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

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

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

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

Jesús López-Fidalgo, Chiara Tommasi, and Paula C. Trandafir. 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, 2007.

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

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

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

Elham Yousefi, Luc Pronzato, Markus Hainy, Werner G. Müller, and Henry P. Wynn. Discrimination between Gaussian process models: active learning and static constructions. *Statistical Papers*, 2023. doi: 10.1007/s00362-023-01436-x.