KL-optimum designs to discriminate models with different variance function

Disegni KL-ottimi per discriminare tra modelli con diversa funzione di varianza

Alessandro Lanteri, Samantha Leorato and Chiara Tommasi

Abstract In many applications, researchers are interested in models that can be defined by interpretable statistics, such as mean and variance. Kullback-Leibler criterion is one of the best known optimum criteria to select designs to discriminate between two competing models. We provide a simple closed form formula to obtain the optimal KL-design to discriminate between regression models with different variance structures and common response mean and we conduct numerical experiments to compare its performance with other benchmark designs in terms of statistical power.

Abstract *In molte applicazioni, i ricercatori sono interessati a modelli che possono essere definiti tramite statistiche interpretabili, come media e varianza. Il criterio di Kullback-Leibler e uno dei migliori criteri di ottimalità per selezionere disegni atti a discriminare tra due modelli alternativi. In questo lavoro, forniamo l'espressione in forma chiusa del disegno KL-ottimo per discriminare tra modelli di regressione con diversa struttura di varianza e stessa funzione media. Inoltre, attraverso uno studio di simulazione, abbiamo confrontato il disegno KL-ottimo con altri disegni di riferimento in termini di potenza statistica.*

Key words: Heretoschedasticity, KL-divergence, Optimal Design

1 Introduction

In many modern sciences, despite the development of new technologies, to gather information and empirical evidence about specific hypotheses can be very expensive, not only from a strictly economical standpoint. The time cost of the data-

Alessandro Lanteri[1] - e-mail: `alessandro.lanteri@unimi.it`
Samantha Leorato[2] - e-mail: `samantha.leorato@unimi.it`
Chiara Tommasi[3] - e-mail: `chiara.tommasi@unimi.it`
[123]Università degli studi di Milano, DEMM.

collection process could render the information gathered obsolete while ethical costs could overwhelm the scientific benefits of an experiment. For these and many other reasons, optimal or efficient experimental design is important in the scientific research. The T-criterion [1] is one of the most widely used methods to obtain optimal design when the goal is to discriminate between two rival regression models with homoschedastic Gaussian errors. T-criterion has been generalized with weaker assumptions in subsequent works [2, 5, 8]. A general criterion for discriminating between models, based on the Kullback-Leibler (KL) divergence, has been introduced in [3] and extended in [6]. The problem of discriminating between homoschedastic and heteroschedastic regression models with the same regression function has not been less considered in the literature. To handle this problem we consider the KL-criterion that generalizes the T-criterion (as it discriminates between any two rival statistical models). A KL-optimum design, as well as T-designs, depends on the nominal values of the parameters of the true model, that in the case of nested models is the larger model. When the values of the parameters are unknown, but it is available a prior information about such parameters, it has been proposed a Bayesian approach, for the T-criterion [4] and for the KL-criterion [7]. Therefore, beside the KL-optimal design, we compute the Bayesian KL-optimal design to handle the problem of dependence on unknown parameters. In particular, in Section 2 we state a theorem which provides a closed form for a KL-optimal design for discriminating between homoschedastic and non-homoschedastic Gaussian models. In Section 3 we conduct a numerical experiment to analyze how good are different designs in terms of statistical power when we use the log-likelihood ratio test to discriminate between the two nested models.

2 Discriminating between different variance functions

One way to discriminate between two models is with the use of the KL-optimality criterion [3] which is based on the well known Kullback Leibler divergence. In real data applications, practitioners are often interested in models where the distribution of the response $y \in \mathscr{Y}$ can be defined by some interpretable statistics somehow linked with the covariates $x \in \chi \subseteq \mathbb{R}^p$. Let us recall that a continuous design with K design points is denoted as

$$\xi = \left\{ \begin{array}{c} x_1, \; \dots, \; x_K \\ \omega_1, \dots, \omega_K \end{array} \right\}; \qquad 0 \leq \omega_k \leq 1; \qquad \sum_{k=1}^{K} \omega_k = 1$$

where the domain χ of any experimental point x is assumed to be compact. A proportion of ω_k of responses are observed at the experimental point x_k, $k = 1, \dots, K$. Let $f_1\left[y, \mu_1(x, \beta_1), \sigma_1^2(x, \theta_1)\right]$ and $f_2\left[y, \mu_2(x, \beta_2), \sigma_2^2(x, \theta_2)\right]$ be two competing statistical models, where μ_j and σ_j^2, $j = 1, 2$, are the mean and the variance functions, respectively. Let the two models be nested and let the first be the "true" and completely known largest model. The KL-optimality criterion function is

$$I_{21}(\xi) = I_{21}(\xi;\theta_1,\beta_1) = \min_{(\theta_2,\beta_2)\in\Omega_2} \int_\chi \int_{\mathscr{Y}} f_1\left[y,\mu_1(x,\beta_1),\sigma_1^2(x,\theta_1)\right] \log\left\{\frac{f_1\left[y,\mu_1(x,\beta_1),\sigma_1^2(x,\theta_1)\right]}{f_2\left[y,\mu_2(x,\beta_2),\sigma_2^2(x,\theta_2)\right]}\right\} dy\, \xi(dx)$$

and thus, a design $\xi^{KL}_{\theta_1,\beta_1}$ which maximize $I_{21}(\xi)$ is called KL-optimal. The subscripts θ_1,β_1 underline that in general a KL-optimum design depends on the assumed value for the parameters of the true model.

Let us assume that we are interested in discriminating between two rival Gaussian models with the same regression function and different variance structures so that $y_i = \mu(x_i;\beta_j) + \varepsilon_i$ with $\varepsilon_i \sim N(0,\sigma_j^2(x_i))$ for $i = 1,\ldots,n$ and $j = 1,2$. Consider the specific case where $\sigma_1^2(x_i) = \varsigma_1 h(x_i;\theta_1)$ and $\sigma_2^2(x_i) = \varsigma_2$, where $h : \mathbb{R} \to \mathbb{R}_+$ is a continuous positive function in χ. Let also $\tilde\theta$ be a specific value for θ_1 such that $h(x;\tilde\theta) = 1$, this implies that model 2 is nested in model 1. Then the following theorem, which allows to compute analytically the KL-optimal design, can be proved.

Theorem 1. *Let $\underline{h} = \inf_x h(x) > 0$ and $\bar h = \sup_x h(x) < \infty$. Let $\chi_l = \{x : h(x) = \underline{h}\}$ and $\chi_u = \left\{x : h(x) = \bar h\right\}$. Then*

$$\xi^* = \left\{\begin{matrix} x_l, & x_u \\ \omega, & 1-\omega \end{matrix}\right\}, \quad \textit{with} \quad \omega = \left(\frac{\bar h}{\bar h - \underline{h}} - \frac{1}{\log \bar h - \log \underline{h}}\right)$$

is a KL-optimal design, where $x_l \in \chi_l$ and $x_u \in \chi_u$.

Theorem 1 is quite interesting because it is uncommon to find KL-optimum design in a closed form. Note that if χ_l or χ_u contain more than one point, then any design with more support points in χ_l or χ_u is KL-optimal, provided that the sum of the weights corresponding to the points in χ_l or χ_u is ω and $1-\omega$, respectively. From this theorem we can deduce that since the design points are the ones which provide the most extreme values of $h(x)$, then, when the variance function is strictly monotone in the compact $\chi \subseteq \mathbb{R}$, the design points are necessarily the edge points of χ. This implies that, in this setting, KL-optimal designs will have the same design points, independently on the values of the true parameters, and only the design weights will vary. On the other hand, when $h(x)$ is non-monotone, KL-optimal designs might differ greatly also with regards to the design points for different values of the parameters. To discriminate between the homoschedastic and the heteroschedastic model, and thus when the hypotheses are

$$\begin{cases} H_0 : \sigma^2 = \varsigma_2 \\ H_1 : \sigma^2 = \varsigma_1 h(x;\theta_1) \end{cases} \quad \text{or equivalently} \quad \begin{cases} H_0 : \theta_1 = \tilde\theta \\ H_1 : \theta_1 \neq \tilde\theta \end{cases},$$

we use the log-likelihood ratio test.

So far we assumed that all the parameters of the true model are known, but in real applications they might be unknown. It is less stringent to assume that an approximate range of possible values of the parameters is available and it is possible to build a prior probability distribution π on the unknown parameters. Taking the expected value of KL-optimality criterion over the prior distribution of the parameters we can define the partially Bayesian KL-optimality (PBKL-optimality)

[7] as $I_{21}^{PB}(\xi,\pi) = \mathbb{E}_\pi[I_{21}(\xi)]$ and, consequently, the design ξ_π^{PB} which minimizes $I_{21}^{PB}(\xi,\pi)$ is called PBKL-optimum design under the distribution π.

3 Numerical Experiment

In this section we conduce some numerical experiments in order to compare the performance of the proposed optimal design criterion with other designs. Consider two normal models, $f_1\left[\mu_1(x),\sigma_1^2(x)\right]$ and $f_2\left[\mu_2(x),\sigma_2^2(x)\right]$, with $\sigma_1^2(x) = \varsigma_1 h(x;\theta_1)$, $\sigma_2^2 = \varsigma_2$ and the same mean structure $\mu_1(x) = \alpha_1 + \beta_1 x$ and $\mu_2(x) = \alpha_2 + \beta_2 x$. In each scenario we vary the sample size n and the variance function parameter θ_1. In the study we consider two different variance functions against homoschedasticity $h_1(x;\theta_1) = \frac{1}{1+\theta_1 x}$ and $h_2(x;\theta_1) = 1+\sin(9\theta_1 x)+\theta_1 x$. Note that the first is monotone while the second is not. From Figure 3 we can appreciate how, for different values of θ_1, h_1 reaches its minimum and maximum value always at the extremes of χ while h_2 reaches its extreme values in scattered locations. Note that, with both choices

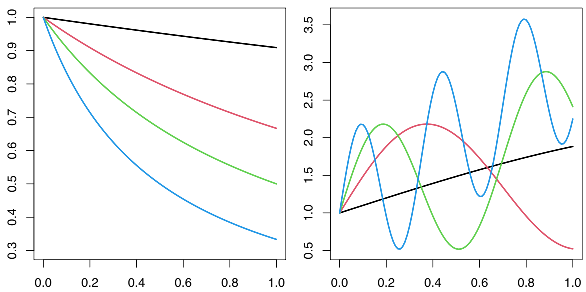

Fig. 1: Variance functions in $\chi = [0,1]$. $h_1(x;\theta_1)$, left plot, and $h_2(x;\theta_1)$, right plot, for $\theta_1 = 0.1, 0.5, 1, 2$, represented by black, red, green and blue curves, respectively.

of variance functions, the two rival models (homoschedastic and heteroschedastic) become more and more similar as θ_1 goes to $\tilde{\theta} = 0$. We let $\chi = [0,1]$ and set the nominal parameter $\alpha_1 = \beta_1 = \varsigma_1 = 1$. Table 1 displays the estimated power of the likelihood ratio test for different designs, in different experimental scenarios and using the variance function $h_1(x;\theta_1)$. Each table entry is the average over 10000 repetition of the experiment in the same setting. The notation $\xi_{\theta_1}^{KL}$ represents the KL-optimal design for a specific value of θ_1, such designs can be easily obtained using Theorem 1. For $\theta_1 = 0.5, 1, 2$ we obtain:

$$\xi_{0.5}^{KL} = \begin{pmatrix} 0 & 1 \\ 0.44 & 0.56 \end{pmatrix}; \; \xi_1^{KL} = \begin{pmatrix} 0 & 1 \\ 0.47 & 0.53 \end{pmatrix}; \; \xi_2^{KL} = \begin{pmatrix} 0 & 1 \\ 0.41 & 0.59 \end{pmatrix}.$$

We compare the performance of KL-optimal designs with two uniform designs, which consist of a fixed number of equidistant and equally weighted design points that cover all the domain χ. We denote $U3$ and $U4$ the uniform designs with three and four points, respectively.

Table 1: Estimated power obtained in different designs for different scenarios using the variance function $h_1(x;\theta_1)$

θ_1	n	$\xi^{KL}_{0.5}$	ξ^{KL}_{1}	ξ^{KL}_{2}	$U3$	$U4$
	30	0.1309	0.1175	0.1203	0.1013	0.0511
0.5	50	0.1721	0.1757	0.1682	0.1117	0.1173
	100	0.2962	0.2895	0.2863	0.2175	0.1824
	30	0.2676	0.2579	0.2581	0.1889	0.0735
1	50	0.4062	0.4030	0.3989	0.2014	0.2573
	100	0.6814	0.6756	0.6563	0.5177	0.4342
	30	0.5460	0.5266	0.5232	0.4032	0.1362
2	50	0.7560	0.7613	0.7481	0.4512	0.5431
	100	0.9698	0.9631	0.9598	0.8838	0.8195

From Table 1 we can appreciate, as expected, how the power increases with the sample size n and decreases as θ_1 gets smaller, that is because for smaller values of θ_1 the two models become more and more similar an thus it is more difficult to discriminate between the two. From this numerical experiment, we can see how the KL-optimal designs outperform the uniform designs in all settings, even when they are assuming a wrong θ_1. We also notice that the design obtained from $U3$ provides better results than $U4$, this is because $U3$ is incidentally more similar to the KL-optimal designs than $U4$.

We perform a similar experiment using the non-monotone variance function $h_2(x;\theta_1)$. The KL-optimal designs, obtained with the application of Theorem 1 for $\theta_1 = 0.5, 1, 2$, are:

$$\xi^{KL}_{0.5} = \begin{pmatrix} 0.37 & 1.00 \\ 0.38 & 0.62 \end{pmatrix};\ \xi^{KL}_{1} = \begin{pmatrix} 0.51 & 0.88 \\ 0.64 & 0.36 \end{pmatrix};\ \xi^{KL}_{2} = \begin{pmatrix} 0.26 & 0.79 \\ 0.65 & 0.35 \end{pmatrix}.$$

Differently from the case with monotone variance, here the designs are very different from each other.

In Table 2 we show, for each value of $\theta_1 = 0.5, 1, 2$, the KL-efficiency of a design ξ, $\mathrm{Eff}_{KL}(\xi) = I_{21}(\xi)/I_{21}(\xi^{KL}_{\theta_1})$, which is a measure of the goodness of ξ with respect to $\xi^{KL}_{\theta_1}$ for discrimination purposes. From Table 2 we can appreciate how the difference between KL-optimum designs determines a poor KL-efficiency when a KL-optimum design with a wrong value of θ_1 is used. Uniform designs are more robust but far from been efficient.

Table 2: KL-Efficiency of different designs with respect to $\xi^{KL}_{\theta_1}$ for different values of θ_1 and variance function $h_2(x;\theta_1)$

	$\mathrm{Eff}_{KL}(\xi^{KL}_{0.5})$	$\mathrm{Eff}_{KL}(\xi^{KL}_{1})$	$\mathrm{Eff}_{KL}(\xi^{KL}_{2})$	$\mathrm{Eff}_{KL}(U3)$	$\mathrm{Eff}_{KL}(U4)$
$\xi^{KL}_{0.5}$	1.0000	0.4550	0.2201	0.6009	0.5118
ξ^{KL}_{1}	0.1689	1.0000	0.0184	0.5482	0.1442
ξ^{KL}_{2}	0.0003	0.0043	1.0000	0.1604	0.0994

In order to obtain more efficient and robust designs we rely on the PBKL-optimality criterion. We use two different prior distribution describing two different type of prior knowledge. The first prior distribution, π_1, assigns uniform weights to the values of θ_1 that we might consider to be true, in our experiment $\theta_1 = 0.5, 1, 2$. To represent the case where the candidate values of θ_1 are not known, but it is available a range of possible values, say $\theta_1 \in [0.5, 2]$, we use a second prior distribution, π_2, which is a discrete uniform distribution with several equidistant points with maximum distance between each other. The PBKL-optimal designs, that we have obtained computationally using a first order algorithm, are:

$$\xi^{PB}_{\pi_1} = \begin{pmatrix} 0.000 & 0.255 & 0.486 & 0.828 & 1.000 \\ 0.0002 & 0.3848 & 0.3118 & 0.0586 & 0.2446 \end{pmatrix}; \ \xi^{PB}_{\pi_2} = \begin{pmatrix} 0.000 & 0.293 & 0.571 & 0.842 & 1.000 \\ 0.0001 & 0.4839 & 0.2182 & 0.2125 & 0.0853 \end{pmatrix}.$$

From Table 3 we can appreciate that the KL-optimal designs provide the best results in terms of statistical power when they are used for the correct value of θ_1, although they can be very inefficient when they are improperly adopted. As we commented before, uniform designs seem to be more robust than KL-optimal designs but generally provide a low power. On the other hand, PBKL-designs seem to combine the qualities of the other two kinds of designs, providing a high power in most settings.

Table 3: Estimated power obtained in different designs for different scenarios using the variance function $h_2(x;\theta_1)$

θ_1	n	$\xi^{KL}_{0.5}$	ξ^{KL}_{1}	ξ^{KL}_{2}	$U3$	$U4$	$\xi^{PBKL}_{\pi_1}$	$\xi^{PBKL}_{\pi_2}$
	30	0.7349	0.5014	0.2753	0.6013	0.4612	0.5984	0.4237
0.5	50	0.9117	0.7029	0.4066	0.7340	0.7715	0.8285	0.7122
	100	0.9975	0.9517	0.6850	0.9662	0.9451	0.9732	0.9376
	30	0.2402	0.8461	0.0373	0.6587	0.2896	0.7034	0.5478
1	50	0.4268	0.9721	0.0615	0.8603	0.4382	0.9052	0.7424
	100	0.7164	0.9999	0.1173	0.9893	0.6690	0.9932	0.9242
	30	0.0275	0.0232	0.9038	0.4084	0.3111	0.8524	0.6356
2	50	0.0324	0.0187	0.9878	0.5861	0.4012	0.9561	0.8280
	100	0.0345	0.0246	1.0000	0.8010	0.5993	0.9969	0.9620

References

1. AC Atkinson and VV Fedorov. The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70, 1975.
2. AC Atkinson and VV Fedorov. Optimal design: Experiments for discriminating between several models. *Biometrika*, 62(2):289–303, 1975.
3. J López-Fidalgo, C Tommasi, and PC 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.
4. AC Ponce de Leon and AC Atkinson. Optimum experimental design for discriminating between two rival models in the presence of prior information. *Biometrika*, 78(3):601–608, 1991.
5. AC Ponce de Leon and AC Atkinson. The design of experiments to discriminate between two rival generalized linear models. In *Advances in GLIM and Statistical Modelling*, pages 159–164. Springer, 1992.
6. C Tommasi. Optimal designs for discriminating among several non-normal models. In *mODa 8-Advances in Model-Oriented Design and Analysis*, pages 213–220. Springer, 2007.
7. C Tommasi and J López-Fidalgo. Bayesian optimum designs for discriminating between models with any distribution. *Computational Statistics & Data Analysis*, 54(1):143–150, 2010.
8. D Ucinski and B Bogacka. T-optimum designs for multiresponse dynamic heteroscedastic models. In *mODa 7—Advances in Model-Oriented Design and Analysis*, pages 191–199. Springer, 2004.