© Statistical Research and Training Center
مجلهٔ پژوهش‌های آماری ایران
J. Statist. Res. Iran **9** (2012): 1–10
دوره‌ی ۹، شماره‌ی ۱، بهار و تابستان ۱۳۹۱، صص ۱–۱۰
پژوهشکده‌ی آمار

Bayesian Optimum Design Criterion for Multi Models Discrimination

F. Z. Labbaf and H. Talebi*

University of Isfahan

Abstract. The problem of obtaining the optimum design, which is able to discriminate between several rival models has been considered in this paper. We give an optimality-criterion, using a Bayesian approach. This is an extension of the Bayesian KL-optimality to more than two models. A modification is made to deal with nested models. The proposed Bayesian optimality criterion is a weighted average, where the weights are corresponding probabilities of models to let them be true. We consider these probabilities coming from a Poisson distribution.

Keywords. Kulback-Leibler distance; discrimination; nested models; optimum design; optimality criterion.

MSC 2010: 62k05, 62k20.

1 Introduction

The problem of seeking the optimum design, which is able to discriminate the true model from a list of candidate models, has been studied by several researchers. T-optimality criterion is provided to obtain optimal designs for discriminating between two models (Atkinson and Fedorov, 1975). This criterion is given under normality and homogeneity assumptions of the random errors. A generalization, i.e. for heteroscedastic normal models, is proposed by Ucinki and Bogacka (2004). Further, Ponce de Leon and Atkinson (1991) extended T-criterion for generalized linear models.

* Corresponding author

A more general criterion so-called KL-optimality is introduced by López-Fidalgo et al. (2007). This criterion, which covers T-optimality and all its generalizations, is based on the Kullback-Leibler (KL) distance and it is useful for discriminating between non-normal and more generally nonlinear models. Tommasi (2007) generalized KL-optimality to deal with discrimination between several models.

It is noted that all of these criteria work when one assumes a certain model, with known parameters, being true. That is, T- and KL-optimal designs depend on the nominal values of the parameters of the true model and hence they are only locally optimal. In contrast, many researchers proceed to Bayesian approach. The motivation for using Bayesian criteria is that sometimes neither the true model nor the parameters in the postulated true model are known, but there is prior information through probability distributions. Although, this approach has some limitations, it reduces dependency of inferences on the nominal values. Ponce de Leon and Atkinson (1991) extended T-optimality by a Bayesian approach. Tommasi and López-Fidalgo (2010) proposed Bayesian KL-optimality criterion for discriminating between two models.

In this paper, we extend Bayesian KL-optimality criterion for discriminating between several models with any distribution. Note that KL-criterion depends on the postulated true model, however, assuming a certain model as the true one is not always possible. To overcome this problem, we attribute suitable prior probability to each model. One more problem arises when a smaller model is taken as the true one in nested models, that is, KL-criterion becomes zero. in order to solve this problem, following Atkinson and Fedorov (1975), we impose some constraints on the parameters.

In Section 2, the new criterion is introduced. A useful numerical algorithm for obtaining optimal design is given in Section 3. Finally, in Section 4 an illustrative example is provided.

2 Generalized Bayesian KL-Optimality Criterion

Suppose y is a response variable, χ is the compact design region, H is the class of all discrete probability distributions on χ and $\xi \in$ H is a design measure. Consider $f_i(y, x, \theta_i), i = 1, 2, ..., r$ as a density function, where $x \in \chi$ is an explanatory variable. Let the true model be a model among r models with a density function $f_i(y, x, \theta_i)$ with corresponding prior probability π_i. It is assumed that the parameter $\theta_i \in \Omega_i \subseteq R^{m_i}$ have a prior probability distribution $p_i(\theta_i), i = 1, 2, ..., r$.

Kullback-Leibler distance between $f_1(y, x, \theta_1)$ and $f_2(y, x, \theta_2)$, assuming that $f_1(y, x, \theta_1)$ is the true model, is (Kullback and Leibler, 1951):

$$l[f_1(y, x, \theta_1), f_2(y, x, \theta_2)] = \int f_1(y, x, \theta_1) \log \frac{f_1(y, x, \theta_1)}{f_2(y, x, \theta_2)} dy, \quad x \in \chi. \quad (1)$$

Switching $f_1(y, x, \theta_1)$ by $f_2(y, x, \theta_2)$ as the true model in (1), Kullback-Leibler distance between $f_2(y, x, \theta_2)$ and $f_1(y, x, \theta_1)$ is denoted by $l[f_2(y, x, \theta_2), f_1(y, x, \theta_1)]$ and is different from (1). Note that the quantity in (1) is parameter dependent.

Following Tommasi (2007), we generalize KL-criterion for more than two models in a Bayesian approach. KL-criterion for $f_{\bar{i}}(y, x, \theta_{\bar{i}})$, assuming $f_i(y, x, \theta_i)$ is the true model, is:

$$I_{\bar{i}i}(\xi, \theta_i) = \min_{\theta_{\bar{i}} \in \Omega_{\bar{i}}} \int_{\chi} l[f_i(y, x, \theta_i), f_{\bar{i}}(y, x, \theta_{\bar{i}})] \xi(dx), \quad (i, \bar{i} = 1, ..., r, i \neq \bar{i}). \quad (2)$$

Design $\xi^*_{\bar{i}i}$ which maximizes $I_{\bar{i}i}(\xi, \theta_i)$ is the KL-optimum design. It is noted that $\xi^*_{\bar{i}i}$ is locally optimum since it depends on the nominal values of the parameters in the true model. Define $\bar{I}_{\bar{i}i}(\xi, \theta_i)$ as:

$$\bar{I}_{\bar{i}i}(\xi, \theta_i) = \frac{1}{r-1} \sum_{\bar{i}=1, \bar{i} \neq i}^{r} I_{\bar{i}i}(\xi, \theta_i), \quad i = 1, ..., r. \quad (3)$$

To deal with the problem on not knowing the true model *a priori*, a probability weight π_i, which is the probability that the ith model is the true one, is attributed to $\bar{I}_{\bar{i}i}(\xi, \theta_i)$, given in (3). At each stage, the KL-criterion $\bar{I}_{\bar{i}i}(\xi, \theta_i)$ is calculated, assuming the ith model is true. Then for each θ_i the expectation of KL is taken over the prior distributions of corresponding parameter. Afterward the mean of all expected-KLs are obtained, using corresponding weights π_i. Now, the Bayesian KL$_r$-optimality criterion is:

$$I_r^{\mathrm{BKL}}(\xi) = \sum_{i=1}^{r} \pi_i E_{\theta_i}[\bar{I}_{\bar{i}i}(\xi, \theta_i)]. \quad (4)$$

The criterion in (4), as a function of ξ, is proposed to obtain optimal design. That is,

$$\xi^*_{\mathrm{BKL}_r} = \arg \max_{\xi \in H} I_r^{\mathrm{BKL}}(\xi),$$

where $\xi^*_{\mathrm{BKL}_r}$ is the KL_r-optimal design for discriminating between several models.
Note that $I_r^{\mathrm{BKL}}(\xi)$ is a linear function of $\bar{I}_{\bar{i}i}$. Thus, due to the linearity of $\bar{I}_{\bar{i}i}$ with respect to $I_{\bar{i}i}$ and concavity of the $I_{\bar{i}i}$ on H, $I_r^{\mathrm{BKL}}(\xi)$ is a concave function on H.

Here, we present some notations and definitions which are useful in the following. Suppose $\Omega_{\bar{i}}(\xi, \theta_i)$ is defined as:

$$\Omega_{\bar{i}}(\xi, \theta_i) = \{\hat{\theta}_{\bar{i}} : \hat{\theta}_{\bar{i}}(\xi, \theta_i) = \arg\min_{\theta_{\bar{i}} \in \Omega_{\bar{i}}} \int_i l[f_i(y, x, \theta_i), f_{\bar{i}}(y, x, \theta_{\bar{i}})]\xi(dx)\}. \quad (5)$$

It is noted that for any $\theta_i \in \Omega_i (i, \bar{i} = 1, ..., r, i \neq \bar{i})$, a design, ξ, for which $\Omega_{\bar{i}}$ has a unique member, is a regular design, otherwise is called singular.

For two arbitrary designs ξ and $\bar{\xi}$, direction derivative of $I_{\bar{i}i}$ with respect to ξ, in the direction of $\delta_{\bar{\xi}} = \bar{\xi} - \xi$ is:

$$\partial I_{\bar{i}i}(\xi, \bar{\xi}, \theta_i) = \lim_{\alpha \to 0^+} \frac{I_{\bar{i}i}[(1-\alpha)\xi + \alpha\bar{\xi}] - I_{\bar{i}i}(\xi)}{\alpha}.$$

Moreover, due to linearity of $I_r^{\mathrm{BKL}}(\xi)$, directional derivative of $I_r^{\mathrm{BKL}}(\xi)$ in ξ, in the direction of $\delta_{\bar{\xi}}$ is:

$$\partial I_r^{\mathrm{BKL}}(\xi, \bar{\xi}) = \sum_{i=1}^{r} \pi_i E_{\theta_i}[\partial \bar{I}_{\bar{i}i}(\xi, \bar{\xi}, \theta_i)], \quad (6)$$

where

$$\partial \bar{I}_{\bar{i}i}(\xi, \bar{\xi}, \theta_i) = \frac{1}{r-1} \sum_{\bar{i}=1, \bar{i} \neq i}^{r} \partial I_{\bar{i}i}(\xi, \bar{\xi}, \theta_i) \quad , \quad i = 1, \ldots, r.$$

Let ξ_x be a design which puts the whole mass at point x, then the directional derivative of $I_{\bar{i}i}(\xi, \bar{\xi}, \theta_i)$ at ξ in the direction of $\delta_{\xi_x} = \xi_x - \xi$ is defined as:

$$\psi_{\bar{i}i}(x, \xi, \theta_i) = l[f_i(y, x, \theta_i), f_{\bar{i}}(y, x, \hat{\theta}_{\bar{i}})] - \int_{\chi} l[f_i(y, x, \theta_i), f_{\bar{i}}(y, x, \hat{\theta}_{\bar{i}})]\xi(dx),$$

where $\hat{\theta}_{\bar{i}}$ is the unique element of $\Omega_{\bar{i}}(\xi, \theta_i)$.

If $\bar{\xi}$ is an arbitrary design and ξ is a regular design, then

$$\partial I_{\bar{i}i}(\xi, \bar{\xi}, \theta_i) = \int_{\chi} \psi_{\bar{i}i}(x, \xi, \theta_i)\bar{\xi}(dx) \quad (7)$$

(López-Fidalgo et al., 2007).
Define $\bar{\psi}_{\bar{i}i}(x, \xi, \theta_i)$ as:

$$\bar{\psi}_{\bar{i}i}(x, \xi, \theta_i) = \frac{1}{r-1} \sum_{\bar{i}=1, \bar{i}\neq i}^{r} \psi_{\bar{i}i}(x, \xi, \theta_i), \quad i = 1, \ldots, r.$$

From (7), directional derivative of (6) is:

$$\partial I_r^{\mathrm{BKL}}(\xi, \bar{\xi}) = \int_{\chi} \psi_r^{\mathrm{BKL}}(x, \xi) \bar{\xi}(dx), \tag{8}$$

where

$$\psi_r^{\mathrm{BKL}}(x, \xi) = \sum_{i=1}^{r} \pi_i E_{\theta_i}[\bar{\psi}_{\bar{i}i}(x, \xi, \theta_i)] \tag{9}$$

is the directional derivative of $I_r^{\mathrm{BKL}}(\xi)$ in ξ in the direction of δ_{ξ_x}.

Now, for a regular Bayesian KL$_r$-optimal design, $\xi^*_{\mathrm{BKL}_r}$, we state the following theorem.

***Theorem* 1.** *If $\xi^*_{\mathrm{BKL}_r}$ is a regular design, then*

(i) Design $\xi^*_{\mathrm{BKL}_r}$ is Bayesian KL$_r$-optimal if and only if

$$\psi_r^{\mathrm{BKL}}(x, \xi^*_{\mathrm{BKL}_r}) \leqslant 0, \quad x \in \chi.$$

(ii) The equality occurs at the support points of the optimum design.

The proof is given in Appendix A.

3 Algorithm

Due to intractability of a Bayesian KL$_r$-optimal design construction, we use a numerical approach. The following numerical algorithm is useful in obtaining the optimal design.

1- Let ξ_s is the design at step s. For every support point of $p_i(\theta_i)$ define:

$$\theta_{\bar{i},s} = \arg\min_{\theta_{\bar{i}} \in \Omega_{\bar{i}}} \int_i l[f_i(y, x, \theta_i), f_{\bar{i}}(y, x, \theta_{\bar{i}})] \xi_s(dx), \quad i, \bar{i} = 1, \ldots, r, \ i \neq \bar{i}$$

Then compute:

$$x_s = \arg\max_{x \in \chi} \psi_r^{\mathrm{BKL}}(x, \xi_s).$$

2- Suppose $\{\alpha_s\}$ is a sequence such that

$$0 \leqslant \alpha_s \leqslant 1, \quad \lim_{s\to\infty} \alpha_s = 0, \quad \sum_{s=1}^{\infty} \alpha_s = \infty, \quad \sum_{s=1}^{\infty} \alpha_s^2 < \infty.$$

Compute ξ_{s+1} such that

$$\xi_{s+1} = (1-\alpha_s)\xi_s + \alpha_s \xi_{x_s}.$$

where ξ_{x_s} is a design with the whole mass at the point x_s. In providing a stopping rule for the algorithm, consider the following efficiency criterion for discriminating:

$$0 \leqslant Eff_{\mathrm{BKL}_r}(\xi) = \frac{I_r^{\mathrm{BKL}}(\xi)}{I_r^{\mathrm{BKL}}(\xi^*_{\mathrm{BKL}_r})} \leqslant 1, \tag{10}$$

where ξ is any arbitrary design. Due to concavity of $I_r^{\mathrm{BKL}}(\xi)$, the following is true:

$$\partial I_r^{\mathrm{BKL}}(\xi, \xi^*_{\mathrm{BKL}_r}) \geqslant I_r^{\mathrm{BKL}}(\xi^*_{\mathrm{BKL}_r}) - I_r^{\mathrm{BKL}}(\xi). \tag{11}$$

Taking $\bar{\xi} = \xi^*_{\mathrm{BKL}_r}$ in (7) and replacing $\psi_r^{\mathrm{BKL}}(x,\xi)$ with $\max_{x\in\chi}$ $\psi_r^{\mathrm{BKL}}(x,\xi)$, we can write

$$\partial I_r^{\mathrm{BKL}}(\xi, \xi^*_{\mathrm{BKL}_r}) = \int_{\chi} \psi_r^{\mathrm{BKL}}(x,\xi)\xi^*_{\mathrm{BKL}_r}(dx) \leqslant \max_{x\in\chi} \psi_r^{\mathrm{BKL}}(x,\xi). \tag{12}$$

From (11) and (12), a lower bound for the efficiency can be written as:

$$I(\xi) = [1 + \frac{\max_{x\in\chi} \psi_r^{\mathrm{BKL}}(x,\xi)}{I_r^{\mathrm{BKL}}(\xi)}]^{-1} \leqslant Eff_{\mathrm{BKL}_r}(\xi) \leqslant 1.$$

Thus, the iterative procedure will finish, if $I(\xi) > \delta$, for a suitable δ (for example $\delta = 0.99$).

4 Example

Comparison of linear nested models is one of the important examples in model discrimination problems. Because of simplicity and properties of linear models, many experimenters are interested in using these models. That

Table 1. Prior probabilities of the models and parameter prior distributions.

θ_{10}	p_1	θ_{20}	θ_{21}	p_2	θ_{30}	θ_{31}	θ_{32}	p_3
4.5	0.2	-1	1	0.25	1	0.5	1.5	0.23
4	0.2	-0.5	1	0.14	0.8	0.4	1	0.14
5	0.2	0.7	2	0.11	-1	0.6	1.2	0.17
5.5	0.2	-1	0.5	0.06	-0.2	0.5	1.8	0.15
6	0.2	0.5	0.2	0.05	-0.7	0.3	2	0.12
		-1	1	0.08	0.3	0.6	1	0.19
		1	1	0.05				
		-0.8	0.2	0.12				
		-0.6	0.4	0.07				
		1	0.5	0.07				

is, the nonlinear model is approximated by a linear model, using Taylor expansion. In practice, the lower order polynomial models with a few parameters are usually of interest. In such a scenario, it is rational to assume a Poisson random variable to describe the order of polynomial model in the postulated true model. Now even, in practical cases, models with more than three parameters are improbable to be true. Thus, we can eliminate cubic and more terms in the Taylor expansion form of polynomial and attribute their probabilities to the three-parameters model. Therefore, we raise the problem to be discriminating between the nested models, $\eta_1(x, \theta_1) = \theta_{10}$, $\eta_2(x, \theta_2) = \theta_{20} + \theta_{21}x$ and $\eta_3(x, \theta_3) = \theta_{30} + \theta_{31}x + \theta_{32}x^2$, with prior probabilities 0.27, 0.35 and 0.38, respectively. That is, we use a Poisson distribution with mean 1.3.

Now, let the third model be true. It is known that $I_{13} = I_{23} = 0$. Thus, under such a situation, designing an experiment for discriminating between models is meaningless. Further, assuming the second model is true, we have $I_{12} = 0$. Thus, for nested models, $I_{12} = I_{13} = I_{23} = 0$ (López-Fidalgo et al., 2007). To overcome this problem, following Atkinson and Fedorov (1975), a constraint is imposed to each parameter. We take the parameter's constraints, $\theta_{ij} \geqslant \min(\mathsf{sample}_{ij})$, where sample_{ij} denotes a sample of size 10, taking from prior vector θ_{ij} with replacement and min means the minimum value over all samples. For the priors, given in Table 1, a Bayesian KL_r-optimal design is:

$$\xi^*_{BKL_r} = \begin{bmatrix} -1 & -0.081 & 1 \\ 0.211 & 0.403 & 0.386 \end{bmatrix}.$$

Figure 1 shows $\psi_r^{\mathrm{BKL}}(x,\xi^*_{\mathrm{BKL}_r})$ as a function of x. It can be seen that $\psi_r^{\mathrm{BKL}}(x,\xi^*_{\mathrm{BKL}_r})$ takes zeros at points of the optimal design.

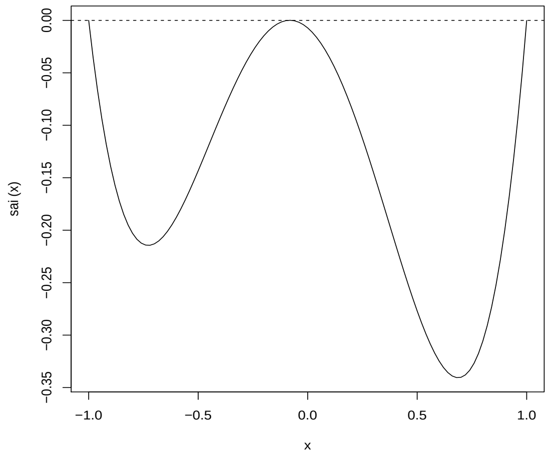

Figure 1. $\psi_r^{\mathrm{BKL}}(x,\xi^*_{\mathrm{BKL}_r})$ as a function of x.

References

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

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Statist*, **22**, 79-86.

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*, **B 69**, 231-242.

Ponce de Leon, A.C. and Atkinson, A.C. (1991). Optimal experimental design for discriminating between two rival models in the presence of prior information. *Biometrika*, **78**, 601-608.

Tommasi, C. (2007). *Optimal designs for discriminating among several non-Normal models.* In: López-Fidalgo, J., Rodríguez-Díaz, J.M., Torsney, B (Eds.), mODa 8: Advances in Model-Oriented Design and Analysis. Physica-Verlag, 213-220.

Tommasi, C. and López-Fidalgo, J. (2010). Bayesian optimum designs for discriminating between models with any distribution. *Computational Statistics and Data Analysis*, **54**, 143-150.

Ucinki, D. and Bogacka, B. (2005). T-optimum designs for discrimination between two multiresponse dynamic models. *Journal of Royal Statistical Society*, **B 67**, 3-18.

Appendix A

Proof of Theorem 1.

(i) It can be easily seen that $\psi_{\bar{i}i}(x,\xi)$ is the direction derivative of $I_{\bar{i}i}(\xi,\theta_i)$ in ξ, in the direction of $\delta_{\xi_x} = \xi_x - \xi$,

$$\psi_{\bar{i}i}(x,\xi) = \partial I_{\bar{i}i}(\xi,\xi_x).$$

It is known that $I_r^{\mathrm{BKL}}(\xi)$ is a concave function of ξ, then, a necessary and sufficient condition for optimality of $\xi^*_{\mathrm{BKL}_r}$ is that for any design ξ, $\partial I_r^{\mathrm{BKL}}(\xi^*_{\mathrm{BKL}_r},\xi) \leqslant 0$. It is easy to show that

$$\max_{\xi\in H} \partial I_r^{\mathrm{BKL}}(\xi^*_{\mathrm{BKL}_r},\xi) = \max_x \psi_r^{\mathrm{BKL}}(x,\xi^*_{\mathrm{BKL}_r}).$$

Thus, a necessary and sufficient condition for optimality of $\xi^*_{\mathrm{BKL}_r}$ is:

$$\psi_r^{\mathrm{BKL}}(x,\xi^*_{\mathrm{BKL}_r}) \leqslant 0, \quad x \in \chi.$$

(ii) Assume $\chi_1 \subset \mathsf{supp}(\xi^*_{\mathrm{BKL}_r})$ and let a be a scalar such that

$$\int_{\chi_1} \psi_r^{\mathrm{BKL}}(x,\xi^*_{\mathrm{BKL}_r})\xi^*_{\mathrm{BKL}_r}(dx) \leqslant a < 0$$

and $\psi_r^{\mathrm{BKL}}(x,\xi^*_{\mathrm{BKL}_r}) = 0$, for any . Thus we have

$$\int_{\chi} \psi_r^{\mathrm{BKL}}(x,\xi^*_{\mathrm{BKL}_r})\xi^*_{\mathrm{BKL}_r}(dx) \leqslant a < 0.$$

On the other hand,

$$\begin{aligned}\int_{\chi} \psi_{\bar{i}i}(x,\xi,\theta_i)\xi(dx) &= \int_{\chi} \{l[f_i(y,x,\theta_i), f_{\bar{i}}(y,x,\hat{\theta}_{\bar{i}})] \\ &\quad - \int_{\chi} l[f_i(y,x,\theta_i), f_{\bar{i}}(y,x,\hat{\theta}_{\bar{i}})]\xi(dx)\}\xi(dx) \\ &= \int_{\chi} l[f_i(y,x,\theta_i), f_{\bar{i}}(y,x,\hat{\theta}_{\bar{i}})]\xi(dx) \\ &\quad - \int_{\chi} l[f_i(y,x,\theta_i), f_{\bar{i}}(y,x,\hat{\theta}_{\bar{i}})]\xi(dx) = 0.\end{aligned}$$

Replacing $\xi = \xi^*_{\mathrm{BKL}_r}$ in (9) we can write:

$$\int_{\chi} \psi_r^{\mathrm{BKL}}(x, \xi^*_{\mathrm{BKL}_r})\xi^*_{\mathrm{BKL}_r}(dx) = 0,$$

that contracts (8). Thus, for any $x \in \mathsf{supp}(\xi^*_{\mathrm{BKL}_r})$ we have:

$$\psi_r^{\mathrm{BKL}}(x, \xi^*_{\mathrm{BKL}_r}) = 0.$$

F. Z. Labbaf
Department of Statistics,
University of Isfahan,
Isfahan, Iran.

H. Talebi
Department of Statistics,
University of Isfahan,
Isfahan, Iran.
email: *h-talebi@sci.ui.ac.ir*