Model selection and parameter estimation in non-linear nested models: a sequential generalized DKL-optimum design

Caterina May Chiara Tommasi*

December 20, 2011

Abstract

This work proposes a sequential procedure which is useful to select the best model among several nested non-linear models and to estimate efficiently the parameters of the chosen model. At the first step of this procedure, a generalized DKL-optimum design is computed, which is optimal for the double goal of model selection and parameter estimation. Subsequently, at each following step, an adaptive generalized DKL-optimum design is computed on the base of the data accrued and tests previously performed. The proposed sequential scheme selects the best non-linear model with probability converging to one; moreover it estimates efficiently its parameters, since the adaptive sequential DKL-optimum designs converge to the D-optimum design for the "true" model.

AMS 2000 subject classifications: Primary 62K05, 60B10; secondary 62L05, 60B05.

Keywords: D-optimality, KL-optimality, DKL-optimality, log-likelihood ratio test, stochastic convergence, sequential design of experiments, semi-continuity, argmin processes, convexity.

1 Introduction

The classical theory of optimum design is based on the assumption that the statistical model for the data is completely specified except for some unknown parameters.

*Corresponding author: Via Conservatorio 7, 20122 Milano, Italy. E-mail address: chiara.tommasi@unimi.it

Therefore, the goal of an optimum design is to provide the best estimates of the parameters of the assumed model. However, more frequently, in real-life applications several rival models are available. Thus, the goal of an optimum design should be dual: to select the "true" model and to estimate efficiently the parameters of the identified model. Several authors have combined these two objectives in one compound criterion by averaging two design criteria, one for parameter estimation and another one for model discrimination. In the context of two nested regression models which differ by only one parameter, [7] has considered the compound criterion given by the weighted geometric mean of D_1- and D-efficiencies. This criterion has been generalized to the case of two nested regression models which differ by more than one parameter by [28] and [31], who have replaced the D_1-criterion with the D_s-one (with $s > 1$) and by [1], who has considered the T-criterion as a measure of discrimination. The D_s-criterion can be used to discriminate between any two nested models. Differently, the T-criterion can be applied to separate models but they must be homoscedastic with Gaussian errors. Another proposal, by [21], is the the KL-criterion, which can be applied in a very general context: the rival models may be nested or not, homoscedastic or heteroscedastic and with any distribution for the errors. In order to consider both the aims of model selection and parameter estimation, [27] has proposed the DKL-optimality criterion which is a weighted geometric mean of KL- and D-efficiencies. In the present paper, the DKL-criterion is suitably generalized to handle the case when more than two rival statistical models are available, with the goal to select the correct model and to estimate efficiently its parameters. This new criterion is called generalized DKL-criterion.

Only compound criteria are considered in this paper. However, let us note that there exist several ways to incorporate different goals in one design criterion. Some examples are given in [8] and [9], among others.

When the rival models are non-linear, the designs which maximize the above mentioned multi-objective criteria are only locally optimum, because the optimality criterion functions depend on the unknown parameters of the models. There are essentially three ways to solve this problem:

1. to follow a Bayesian approach. See, for instance, [19] and [5];

2. to use a max-min criterion. Some examples are [11] and [12];

3. to apply a sequential adaptive procedure. See, for instance, [6] and [16].

In this paper it is assumed that the experiments can be performed sequentially and hence the last strategy is considered. In more detail, at the first step of the proposed sequential procedure a generalized DKL-optimum design is computed. Subsequently, at each step, an adaptive optimum design is computed maximizing a generalized DKL-criterion function where the unknown parameters are replaced by

suitable estimates obtained at the previous step; in addition, some statistical tests are performed to select a model. This adaptive generalized DKL-optimum design, which is "updated" step by step, is called sequential generalized DKL-optimum design. The proposed sequential scheme simultaneously achieves asymptotically both the goals of correct model selection and efficient estimation of the parameters of the "true" model; in fact, it selects the correct non-linear model with probability that tends to one and the adaptive generalized DKL-optimum designs converge to the D-optimum design for the true non-linear model.

In [4], it has been proposed a different sequential scheme, which is applicable only in the set up of nested linear models. [10] have compared, through a simulation study, Biswas and Chaudhuri's sequential design with some non-sequential optimum designs, showing the superiority of non-sequential methods. Actually, in the context of linear models the use of a sequential procedure is not fully justified since optimality criteria do not depend on unknown parameters. [4], as well as [22], use the sequential approach essentially to update the information about the form of the unknown linear model. In this paper, instead, non-linear models are considered and hence a sequential procedure, based on different criteria, is considered as a useful device to avoid model parameter dependence.

Very recently, [30] has proposed a robust optimality criterion for model discrimination and parameter estimation and has provided both sequential and non-sequential versions of this new optimality criterion.

The outline of the paper is the following. In Section 2, the basic notation is setted and KL- and D-optimality criteria are recalled. In Section 3, a generalized DKL-criterion is proposed to discriminate among several nested statistical models and to estimate model parameters. Section 4 is devoted to describe an adaptive sequential procedure, where, at each step, a generalized DKL-optimum design is computed on the base of past data and performed tests. In Section 5, together with some important auxiliary results, two fundamental properties of the procedure are proved: as the number of steps goes to infinity,

- the sequential procedure selects the best statistical model with probability that tends to one;
- the sequential generalized DKL-optimum design converges to the D-optimum design for the true statistical model.

Finally, in Section 6, some ideas about future developements are discussed.

2 Notation setting and description of the models

On a rich enough probability space $(\Omega, \mathcal{F}, P)$, let us define the following random elements. Let an experimental condition X in $\mathcal{X}$ be generated by the experimenter

from a design ξ. More specifically, X is a random variable (or a random vector) having probability distribution equal to ξ, which has support on the experimental domain $\mathcal{X}$, a compact subset of $\mathbb{R}$ (or $\mathbb{R}^q$, $q \geq 1$). Let a random variable Y be the response corresponding to the experimental condition X, and consider that there are k rival families of distribution functions $F_j(y|X;\boldsymbol{\beta}_j)$, with $j = 1,\ldots,k$, for Y conditioned to X, each one depending on a vector of unknown parameters $\boldsymbol{\beta}_j$ in Θ_j which is an open subset of $\mathbb{R}^{d_j}$.

Models $F_j(y|X;\boldsymbol{\beta}_j)$ satisfy standard hypotheses of regularity for every $j = 1,\ldots,k$, as follows. Assume that there exist a $G : \mathbb{R} \times \mathbb{R}^q \to [0,1]$ such that $G(\cdot|x)$ is a distribution function for every x, $G(y|\cdot)$ is measurable for every y, and there is a version of the conditional Radon-Nikodym density relative to G: $f_j(y|x;\boldsymbol{\beta}_j) = F_j(dy|x;\boldsymbol{\beta}_j)/G(dy|x)$ which is measurable in (y,x) for every $\boldsymbol{\beta}_j$ in Θ_j, and it is $\mathcal{C}^2(\Theta_j)$ in $\boldsymbol{\beta}_j$ for every (y,x). Assume also that the support of $F_j(y|x;\boldsymbol{\beta}_j)$ is independent of $\boldsymbol{\beta}_j$ and that the models are identifiable, that is: if $f_j(y|x;\boldsymbol{\beta}_j) = f_j(y|x;\boldsymbol{\beta}'_j)$ a.s. $G(dy|x)$, then $\boldsymbol{\beta}_j = \boldsymbol{\beta}'_j$.

Moreover assume that, for any $j = 2,\ldots,k$,

1. $\boldsymbol{\beta}_j^T = (\boldsymbol{\beta}_{j-1}^T, \boldsymbol{\tau}_j^T)$, where $\boldsymbol{\tau}_j$ is the vector of the last $d_j - d_{j-1}$ components of $\boldsymbol{\beta}_j$;

2. assigning a specific value $\boldsymbol{\tau}_j^0$ to $\boldsymbol{\tau}_j$, then $f_j[y|x;(\boldsymbol{\beta}_{j-1}^T, \boldsymbol{\tau}_j^{0\,T})^T] = f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})$, i.e. $f_j(y|x;\boldsymbol{\beta}_j)$ and $f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})$ are nested models.

In order to choose a specific model among the k rival models, given m independent observations $(Y_1;X_1),\ldots,(Y_m;X_m)$, some statistical tests can be carried out in a stepwise manner until a specific statistical model is selected. The tests are performed for the following hypotheses

$$\begin{cases} \mathrm{H}_{0,j}: & f_{j-1}(y|x;\boldsymbol{\beta}_{j-1}) & \text{is the true model} \\ \mathrm{H}_{1,j}: & f_j(y|x;\boldsymbol{\beta}_j) & \text{is the true model} \end{cases} \tag{2.1}$$

for $j = k, k-1,\ldots,2$. Thus, it is important to choose the design ξ in order to get observations which enable us to discriminate between $f_j(y|x;\boldsymbol{\beta}_j)$ and $f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})$ in the best way.

In order to discriminate between a pair of subsequent nested models $f_j(y|x;\boldsymbol{\beta}_j)$ and $f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})$, the design ξ may be selected by maximizing the KL–optimality criterion, which is defined as

$$\begin{aligned} I_{j-1,j}(\xi;\boldsymbol{\beta}_j) &= \inf_{\boldsymbol{\beta}_{j-1}\in\Theta_{j-1}} \int_{\mathcal{X}}\int_{\mathcal{Y}} \log \frac{f_j(y|x;\boldsymbol{\beta}_j)w(x)}{f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})w(x)}\, F_j(dy|x;\boldsymbol{\beta}_j)\,\xi(dx) \\ &= \inf_{\boldsymbol{\beta}_{j-1}\in\Theta_{j-1}} \int_{\mathcal{X}}\int_{\mathcal{Y}} \log \frac{f_j(y|x;\boldsymbol{\beta}_j)}{f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})}\, F_j(dy|x;\boldsymbol{\beta}_j)\,\xi(dx), \end{aligned} \tag{2.2}$$

where $\mathcal{Y} \subseteq \mathbb{R}$ is the support of Y and $w(x) = \xi(dx)/\nu(dx)$. If the largest model is assumed to be completely known, then criterion (2.2) is the minumum Kullback-Leibler divergence between the joint statistical models $f_j(y|x;\boldsymbol{\beta}_j)w(x)$ and $f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})w(x)$. The KL-criterion (2.2) is a concave function of ξ (as proved by [26]) and a design $\xi^*_{j-1,j}$ which maximizes $I_{j-1,j}(\xi)$ for a given $\boldsymbol{\beta}_j$ is called KL–optimum.

Let

$$\mathcal{I}(x, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1}) = \int_{\mathcal{Y}} \log \frac{f_j(y|x;\boldsymbol{\beta}_j)}{f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})}\, F_j(dy|x;\boldsymbol{\beta}_j) \tag{2.3}$$

be the conditional Kullback-Leibler divergence between the statistical models $f_j(y|x;\boldsymbol{\beta}_j)$ and $f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})$. Once fixed a value of $\boldsymbol{\beta}_j$, a design for which the following set

$$\Omega_{j-1}(\xi, \boldsymbol{\beta}_j) = \left\{ \tilde{\boldsymbol{\beta}}_{j-1} :\ \tilde{\boldsymbol{\beta}}_{j-1}(\xi) = \arg \min_{\boldsymbol{\beta}_{j-1} \in \Theta_{j-1}} \int_{\mathcal{X}} \mathcal{I}(x, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})\, \xi(dx) \right\} \tag{2.4}$$

is a singleton, is called a *KL–regular* design, otherwise it is called *KL–singular* design. Assuming that $\xi^*_{j-1,j}$ is regular, [21] prove that $\xi^*_{j-1,j}$ is a KL-optimum design if and only if $\psi_{j-1,j}(x, \xi^*_{j-1,j}, \boldsymbol{\beta}_j) \leq 0$ for any $x \in \boldsymbol{\mathcal{X}}$, where

$$\psi_{j-1,j}(x, \xi, \boldsymbol{\beta}_j) = \mathcal{I}(x, \boldsymbol{\beta}_j, \tilde{\boldsymbol{\beta}}_{j-1}) - \int_{\mathcal{X}} \mathcal{I}(x, \boldsymbol{\beta}_j, \tilde{\boldsymbol{\beta}}_{j-1})\, \xi(dx) \tag{2.5}$$

is the directional derivative of the criterion function (2.2) at ξ in the direction of $\delta_{\xi_x} = \xi_x - \xi$ and ξ_x is the design which concentrates the whole mass at point x. The quantity $\tilde{\boldsymbol{\beta}}_{j-1}$ in equation (2.5) is the assumed unique element of set (2.4).

The KL-efficiency of a design ξ relative to the optimum design $\xi^*_{j-1,j}$ is

$$\text{Eff}_{j-1,j}(\xi, \boldsymbol{\beta}_j) = \frac{I_{j-1,j}(\xi, \boldsymbol{\beta}_j)}{I_{j-1,j}(\xi^*_{j-1,j}, \boldsymbol{\beta}_j)}.$$

This efficiency is a pure number in $(0, 1)$ which measures the goodness of a design ξ for discriminating purposes.

As previously established, to select a model among k rival models some statistical tests are carried out sequentially starting from H_{0k} against H_{1k} in reverse order until a null hypothesis is rejected. Suppose that H_{0j} is rejected for some $j \in \{k, ..., 2\}$, then $f_j(y|x;\boldsymbol{\beta}_j)$ is considered as the true model. Otherwise, if no null hypothesis is rejected, then $f_1(y|x;\boldsymbol{\beta}_1)$ is considered as the true model. Therefore, in any case, the parameter $\boldsymbol{\beta}_j$ of the true model has to be estimated. Hence, another important design goal is to choose the experimental conditions in order to estimate efficiently the model parameters. Among all the design criteria which are useful for parameter estimation, the D-optimality criterion is indeed the most popular. See for instance,

[13], [23] and [2]. In the general context of non-linear models (see [25]), the D-optimality criterion is defined by the following function

$$\Phi_{D_j}[\boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j)] = \begin{cases} \log|\boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j)| & \text{if } \boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j) \text{ is non-singular} \\ -\infty & \text{if } \boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j) \text{ is singular} \end{cases} \tag{2.6}$$

where, except for the constant m of proportionality, $\boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j)$ is the Fisher information matrix corresponding to the joint distribution $f_j(y|x;\boldsymbol{\beta}_j)\,w(x)$. Thus, $\boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j) = E_X[\boldsymbol{J}_j(x,\boldsymbol{\beta}_j)] = \int_{x\in\mathcal{X}} \boldsymbol{J}_j(x,\boldsymbol{\beta}_j)\,\xi(dx)$ where $\boldsymbol{J}_j(X,\boldsymbol{\beta}_j)$ is the $d_j \times d_j$ matrix whose (r,s)-th element is $E_{Y|X}[-\partial^2 \log f_j(y|x;\boldsymbol{\beta}_j)/\partial\beta_{jr}\partial\beta_{js}]$, and the expected value is taken with respect to $f_j(y|x;\boldsymbol{\beta}_j)$, $j=1,\ldots,k$.

A design $\xi^*_{D_j}$ is a D-optimum design for the parameter estimation of model $f_j(y|x;\boldsymbol{\beta}_j)$ if and only if $\psi_{D_j}(x,\xi^*_{D_j},\boldsymbol{\beta}_j) \le 0,\ x\in\mathcal{X}$, where

$$\psi_{D_j}(x,\xi,\boldsymbol{\beta}_j) = \mathrm{tr}[\boldsymbol{M}_j^{-1}(\xi,\boldsymbol{\beta}_j)\boldsymbol{J}_j(x,\boldsymbol{\beta}_j)] - d_j, \quad j=1,\ldots,k \tag{2.7}$$

is the directional derivative of the D-criterion function (2.6) at ξ in the direction of δ_{ξ_x}. The D-efficiency of a design ξ is defined by the following ratio,

$$\mathrm{Eff}_{D_j}(\xi,\boldsymbol{\beta}_j) = \frac{|\boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j)|^{1/d_j}}{|\boldsymbol{M}_j(\xi^*_{D_j},\boldsymbol{\beta}_j)|^{1/d_j}}, \quad j=1,\ldots,k.$$

3 Generalized DKL-criterion for several nested models

In [27] the DKL-optimality criterion to discriminate between two statistical models and to estimate efficiently their parameters has been proposed. This criterion is here generalized to the case of k nested models by the following weighted geometric mean of efficiencies,

$$\Phi_{DKL}(\xi,\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{j=2}^{k}\left(\frac{I_{j-1,j}(\xi,\boldsymbol{\beta}_j)}{I_{j-1,j}(\xi^*_{j-1,j},\boldsymbol{\beta}_j)}\right)^{\gamma_D} \prod_{j=1}^{k}\left(\frac{|\boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j)|}{|\boldsymbol{M}_j(\xi^*_{D_j},\boldsymbol{\beta}_j)|}\right)^{\frac{\gamma_j}{d_j}}, \tag{3.1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_k^T)^T$, while $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_k,\gamma_D)$ is a vector of fixed constants with $0\le\gamma_j\le 1$ for any $j=1,\ldots,k$, and $0\le\gamma_D\le 1$, fulfilling the linear constraint $(k-1)\gamma_D + \sum_{j=1}^k \gamma_j = 1$. Note that the coefficient γ_D reflects the importance of the discrimination goal while the coefficients γ_j, $j=1,\ldots,k$, balance the importance of the parameter estimation in the k rival models.

Except for some terms which are constant with respect to ξ, the logarithm of (3.1), provided that each matrix $\boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j)$ is not singular, is

$$\log\Phi_{DKL}(\xi,\boldsymbol{\beta},\boldsymbol{\gamma}) \approx \gamma_D \sum_{j=2}^{k}\log I_{j-1,j}(\xi,\boldsymbol{\beta}_j) + \sum_{j=1}^{k}\frac{\gamma_j}{d_j}\log|\boldsymbol{M}_j(\xi,\boldsymbol{\beta}_j)|;$$

hence, maximizing $\Phi_{DKL}(\xi, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is equivalent to maximize the following criterion function:

$$\Psi_{DKL}(\xi, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{cases} \gamma_D \sum_{j=2}^{k} \log I_{j-1,j}(\xi, \boldsymbol{\beta}_j) + \sum_{j=1}^{k} \frac{\gamma_j}{d_j} \log |\boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j)| & \text{if } |\boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j)| \neq 0, \\ & \text{for any } j = 1, ..., k \\ -\infty & \text{otherwise.} \end{cases} \tag{3.2}$$

A generalized DKL-optimum design, ξ^*_{DKL}, maximizes $\Phi_{DKL}(\xi, \boldsymbol{\beta}, \boldsymbol{\gamma})$ or equivalently $\Psi_{DKL}(\xi, \boldsymbol{\beta}, \boldsymbol{\gamma})$.

In the next Theorem 3.1, the following stronger definition of regular design will be adopted.

Definition 3.1. *A design ξ is regular for a given $\boldsymbol{\beta}$ if and only if all the sets $\Omega_{j-1}(\xi; \boldsymbol{\beta}_j)$, defined in (2.4), are singletons and all the Fisher information matrices $\boldsymbol{M}_j(\xi; \boldsymbol{\beta}_j)$ are non singular, for any $j = 1, \ldots, k$.*

Design criterion (3.2) is a concave function in the first argument since it is a convex combination of concave functions, thus the following important equivalence theorem may be stated.

Theorem 3.1. *A regular design ξ^*_{DKL} is generalized DKL-optimum if and only if it fulfils the following inequality*

$$\psi_{DKL}(x, \xi^*_{DKL}, \boldsymbol{\beta}) \leq 0, \quad x \in \mathcal{X},$$

where

$$\psi_{DKL}(x, \xi, \boldsymbol{\beta}) = \gamma_D \sum_{j=2}^{k} \frac{\psi_{j-1,j}(x, \xi, \boldsymbol{\beta}_j)}{I_{j-1,j}(\xi, \boldsymbol{\beta}_j)} + \sum_{j=1}^{k} \frac{\gamma_j}{d_j} \psi_{D_j}(x, \xi, \boldsymbol{\beta}_j)$$

is the directional derivative of criterion function (3.2) at ξ in the direction of δ_{ξ_x}.

The criterion of optimality (3.2) depends on the unknown parameter vector $\boldsymbol{\beta}$ and on the choice of the weights $\boldsymbol{\gamma}$; thus, a generalized DKL-optimum design is only locally optimal when non-linear models are considered. In order to overcome this problem an adaptive sequential design is proposed in the next section.

4 A sequential generalized DKL-optimum design

Suppose that a number of experiments can be carried out sequentially with the goal of discriminating between the k models described in Section 2 and estimating

efficiently the parameters of the true model. A generalized DKL-optimum design proposed in Section 3 may be computed to perform the experiments, but, since the models are non-linear, the optimality would be only locally reached. To overcome the problem of the dependance on the unknown parameter, let us perform the experiments in n sequential steps as follows, and denote the stage of the sequential procedure by $r = 0, 1, \ldots, n$.

At the first stage, i.e. for $r = 0$, a generalized DKL-optimum design is computed, that is a designs maximizing criterion (3.2) based a nominal value for $\boldsymbol{\beta}$ and on an arbitrary choice of values for γ_j $(j = 1, \ldots, k)$. Let $\xi^*_{DKL} = \xi^*_0$ be such generalized DKL-optimum design. Then m independent experimental conditions are generated from ξ^*_0, and denote by $\mathbf{X}_0 = (X_{0,1}, \ldots, X_{0,m})^T$ the random vector of these experimental conditions. Also, a vector of m independent observations $\mathbf{Y}_0 = (Y_{0,1}, \ldots, Y_{0,m})^T$ is obtained from these experimental conditions, and a statistic $\mathcal{T}_{0j}$ is used for testing

$$\mathrm{H}_{0,j} : \boldsymbol{\tau}_j = \boldsymbol{\tau}^0_j \quad \text{against} \quad \mathrm{H}_{1,j} : \boldsymbol{\tau}_j \neq \boldsymbol{\tau}^0_j \tag{4.1}$$

in a stepwise manner, i.e. for $j = k, k-1, \ldots, 2$, until a specific null hypothesis is rejected. Let us stress that hypotheses (4.1) are equivalent to those in (2.1) as the models are nested. Consider the -2 log-likelihood ratio statistic

$$\mathcal{T}^j_{0,m} = -2 \log \frac{L^{j-1}_0(\hat{\boldsymbol{\beta}}_{0,j-1})}{L^j_0(\hat{\boldsymbol{\beta}}_{0,j})}$$

based on the likelihood

$$\mathcal{L}_l(\mathbf{Y}_0, \mathbf{X}_0; \boldsymbol{\beta}_l) = \mathcal{L}_l(\mathbf{Y}_0|\mathbf{X}_0; \boldsymbol{\beta}_l) \cdot \mathcal{L}_l(\mathbf{X}_0) \propto \prod_{s=1}^{m} f_l(y_{0,s}|x_{0,s}; \boldsymbol{\beta}_l), \tag{4.2}$$

so that, for $l = j-1, j$, $L^l_0(\hat{\boldsymbol{\beta}}_{0,l})$ is the likelihood evaluated at its maximum $\hat{\boldsymbol{\beta}}_{0,l}$. A null hypothesis $H_{0,j}$ is rejected with level $\alpha_{0,j}$ if $\mathcal{T}^j_{0,m} > c_{0,j}$, $c_{0,j}$ being the cut-off point corresponding to the significance level $\alpha_{0,j}$.

For $r = 1, 2, \ldots, n$ (i.e. at the next stages), let us define the following random weights: for each $j = 1, \ldots, k$ let $\gamma_{r-1,j}$ to be the square of the proportion of times that model $f_j(y|x; \boldsymbol{\beta}_j)$ has been selected up to the $(r-1)$-th step, provided that such proportion is lower than 1. Otherwise, if the proportion of times that a specific model $f_{\bar{\jmath}}(y|x; \boldsymbol{\beta}_{\bar{\jmath}})$ has been selected is equal to 1, then $\gamma_{r-1,\bar{\jmath}} = 1 - 1/2r$ and $\gamma_{r-1,j} = 0$ for $j \neq \bar{\jmath}$. Finally, let

$$\gamma_{r-1,D} = \frac{1 - \sum_{j=1}^{k} \gamma_{r-1,j}}{k-1}.$$

Denoting $\hat{\boldsymbol{\beta}}_{r-1} = (\hat{\boldsymbol{\beta}}^T_{r-1,1}, \ldots, \hat{\boldsymbol{\beta}}^T_{r-1,k})^T$, an adaptive sequential DKL-optimum design ξ^*_r is found by maximizing the following random criterion function,

$$\begin{aligned}\Psi_{DKL}\left[\xi, \hat{\boldsymbol{\beta}}_{r-1}(\omega), \boldsymbol{\gamma}_{r-1}(\omega)\right] &= \gamma_{rD}(\omega) \sum_{j=2}^{k} \log I_{j-1,j}\left[\xi, \hat{\boldsymbol{\beta}}_{r-1,j}(\omega)\right] \\ &+ \sum_{j=1}^{k} \frac{\gamma_{r-1,j}(\omega)}{d_j} \log\left|\boldsymbol{M}_j\left[\xi, \hat{\boldsymbol{\beta}}_{r-1,j}(\omega)\right]\right|, \quad (4.3)\end{aligned}$$

if $\boldsymbol{M}_j\left[\xi, \hat{\boldsymbol{\beta}}_{r-1,j}(\omega)\right]$ is not singular for any $j = 1, \ldots, k$, otherwise

$$\Psi_{DKL}\left[\xi, \hat{\boldsymbol{\beta}}_{r-1}(\omega), \boldsymbol{\gamma}_{r-1}(\omega)\right] = -\infty.$$

In (4.3), if $r = 1$ $\hat{\boldsymbol{\beta}}_{r-1,j}$ is the maximum likelihood estimator for $\boldsymbol{\beta}_j$ based on (4.2), with $l = j$; from the adaptive sequential DKL-optimum design ξ^*_1, a vector $\mathbf{X}_1 = (X_{1,1}, \ldots, X_{1,m})^T$ of m experimental conditions are generated, that is a vector of conditionally independent and identically distributed random variables with respect to the past $\sigma(\mathbf{Y}_0, \mathbf{X}_0)$, having conditional distribution equal to ξ^*_1. Given $\mathbf{X}_1$, a vector of m conditionally independent responses $\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{1,m})^T$ is observed. If $r \geq 2$, $\hat{\boldsymbol{\beta}}_{r-1,j}$ is the maximum likelihood estimator for $\boldsymbol{\beta}_j$ based on the conditional likelihood of $(\mathbf{Y}_{r-1}, \mathbf{X}_{r-1})$ given all the past observations:

$$\mathcal{L}_j(\mathbf{Y}_{r-1}, \mathbf{X}_{r-1} | \mathbf{Y}_{r-2}, \mathbf{X}_{r-2}, \ldots, \mathbf{Y}_0, \mathbf{X}_0; \boldsymbol{\beta}_j); \quad (4.4)$$

from ξ^*_r, a vector of m experimental conditions $\mathbf{X}_r = (X_{r,1}, \ldots, X_{r,m})^T$, which are conditionally independent and identically distributed with respect to the past $\sigma(\mathbf{Y}_{r-1}, \mathbf{X}_{r-1}, \ldots, \mathbf{Y}_0, \mathbf{X}_0)$, is generated. Given $\mathbf{X}_r$, a vector of m conditionally independent responses $\mathbf{Y}_r = (Y_{r,1}, \ldots, Y_{r,m})^T$ is observed. Note that the response vector $\mathbf{Y}_r$ depends on the past observations only through $\mathbf{X}_r$, therefore the conditional distribution of $\mathbf{Y}_r$ given $\sigma(\mathbf{X}_r, \mathbf{Y}_{r-1}, \mathbf{X}_{r-1}, \ldots, \mathbf{Y}_0, \mathbf{X}_0)$ is equal to the conditional distribution of $\mathbf{Y}_r$ given $\sigma(\mathbf{X}_r)$. Hence (4.4) satisfies

$$\begin{aligned}& \mathcal{L}_j(\mathbf{Y}_{r-1}, \mathbf{X}_{r-1} | \mathbf{Y}_{r-2}, \mathbf{X}_{r-2}, \ldots, \mathbf{Y}_0, \mathbf{X}_0; \boldsymbol{\beta}_j) \\ = \; & \mathcal{L}_j(\mathbf{Y}_{r-1} | \mathbf{X}_{r-1}, \mathbf{Y}_{r-2}, \mathbf{X}_{r-2}, \ldots, \mathbf{Y}_0, \mathbf{X}_0; \boldsymbol{\beta}_j) \cdot \mathcal{L}_j(\mathbf{X}_{r-1} | \mathbf{Y}_{r-2}, \mathbf{X}_{r-2}, \ldots, \mathbf{Y}_0, \mathbf{X}_0) \\ \propto \; & \mathcal{L}_j(\mathbf{Y}_{r-1} | \mathbf{X}_r, \mathbf{Y}_{r-2}, \mathbf{X}_{r-2}, \ldots, \mathbf{Y}_0, \mathbf{X}_0; \boldsymbol{\beta}_j) \\ = \; & \mathcal{L}_j(\mathbf{Y}_{r-1} | \mathbf{X}_{r-1}; \boldsymbol{\beta}_j) = \prod_{s=1}^{m} f_j(y_{r-1,s} | x_{r-1,s}; \boldsymbol{\beta}_j), \quad j = 1, \ldots, k, \quad (4.5)\end{aligned}$$

In the notation of (4.3) it is stressed that the second and third arguments of $\Psi_{DKL}(\cdot, \cdot, \cdot)$ are now functions of $\omega \in \Omega$, and hence the optimal designs ξ^*_r, for any $r \geq 1$, are random distributions.

Then, hypotheses (4.1) are tested through the statistic

$$\mathcal{T}_{r,m}^{j} = \mathcal{T}_{r-1,m}^{j} + T_{r,m}^{j}, \tag{4.6}$$

for $j = k, k-1, \ldots, 2$ until a specific null hypothesis is rejected, where

$$T_{r,m}^{j} = -2 \log \frac{L_r^{j-1}(\hat{\boldsymbol{\beta}}_{r,j-1})}{L_r^{j}(\hat{\boldsymbol{\beta}}_{r,j})}, \tag{4.7}$$

is the log-likelihood ratio statistic based on the following conditional likelihood

$$\mathcal{L}_l(\mathbf{Y}_r, \mathbf{X}_r | \mathbf{Y}_{r-1}, \mathbf{X}_{r-1}, \ldots, \mathbf{Y}_0, \mathbf{X}_0; \boldsymbol{\beta}_j) \propto \prod_{s=1}^{m} f_j(y_{r,s} | x_{r,s}; \boldsymbol{\beta}_j), \quad j = 1, \ldots, k,$$

and $L_r^l(\hat{\boldsymbol{\beta}}_{r,l})$ is the corresponding conditional likelihood evaluated at its maximum point $\hat{\boldsymbol{\beta}}_{r,l}$, $l = j-1, j$. A null hypothesis $H_{0,j}$ is rejected with level $\alpha_{r,j}$ if $\mathcal{T}_{r,m}^{j} > c_{r,j}$, where $c_{r,j}$ is the cut-off point corresponding to the level $\alpha_{r,j}$.

Let us remark the following. Test statistic $\mathcal{T}_{r,m}^{j}$ is based on all the observations obtained up to the r-th step, which are dependent. Despite of this, the computational effort to determine $\mathcal{T}_{r,m}^{j}$ should be quite low. From equation (4.6), to compute $\mathcal{T}_{r,m}^{j}$ is enough to update the test statistic $\mathcal{T}_{r-1,m}^{j}$, computed at the previous step, by adding the log-likelihood ratio statistic $T_{r,m}^{j}$. In addition, $T_{r,m}^{j}$ is based on m independent observations (given the vector of experimental conditions $\mathbf{X}_r$) and for some models it is already implemented in many statistical software packages. For instance, for the most commonly used generalized linear models, the log-likelihood ratio statistics $T_{i,m}^{j}$, $i = 1, \ldots, r$ (and therefore also $\mathcal{T}_{r,m}^{j}$) can be easily computed using common statistical packages.

Note also that for easy of notation it has been considered the same number m of observations at each step; this could be straightforward generalized to the case of m_r observations at each step $r = 0, \ldots, n$, assuming that the hypotheses considered in the the rest of the paper hold for $m = \min\{m_o, \ldots, m_n\}$. Note that, after n steps, $N = \sum_{r=0}^{n} m_r$ dependent observations $(X_{r,s}, Y_{r,s})$, $s = 1, \ldots, m_r$ and $r = 0, \ldots, n$, are collected in the experiment.

5 Selection of the correct model and convergence to the corresponding D-optimal design

The main results of this section are Theorem 5.1 and Theorem 5.2 which guarantee two fundamental properties of the sequential procedure. Some methods used in [4] are extended to the different scheme proposed in this paper. Theorem 5.1

assures that the true model is asymptotically selected; Theorem 5.2 states that the sequence of generalized DKL-optimum designs converges in probability to the D-optimal design for the true model. In addition, some important auxiliary results are provided. The first one is Proposition 5.1 which gives the asymptotic distribution, under the null hypothesis, of the test statistics defined in (4.6), as the number m of observations increases to infinity.

From now on, let the true model for Y conditioned to X be $f_{j^*}(y|x;\boldsymbol{\beta}_{j^*})$, $j^* \in \{1,\ldots,k\}$, and let $\bar{\boldsymbol{\beta}}_{j^*}$ denote the true value of the parameter; this means that, whenever $j^* \geq 2$, the last components of $\bar{\boldsymbol{\beta}}_{j^*}$ verifies $\bar{\boldsymbol{\tau}}_{j^*} \neq \boldsymbol{\tau}^0_{j^*}$. Some further assumptions on the models will be required.

Assumptions 5.1. *For any design ξ such that $M_{j*}(\xi,\boldsymbol{\beta}_{j^*})$ is not singular, it holds*

5.1.1. *Second partial derivatives of $f_{j^*}(y|x;\boldsymbol{\beta}_{j^*})$ may be passed under the integral sign in $\int_{\mathcal{Y}} f_{j^*}(y|x;\boldsymbol{\beta}_{j^*})\, G(dy|x)$.*

5.1.2. *$|\partial^2 f_{j^*}(y|x;\boldsymbol{\beta}_{j^*})/\partial\beta_{j^*r}\partial\beta_{j^*s}| \leq k(y,x)$ for all $\boldsymbol{\beta}_{j^*}$ in some neighborhood of $\bar{\boldsymbol{\beta}}_{j^*}$, with*

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} k(y,x)\; F_{j^*}(dy|x;\bar{\boldsymbol{\beta}}_{j^*})\,\xi(dx) < \infty.$$

5.1.3.

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} \log\frac{f_{j^*}(y|x;\bar{\boldsymbol{\beta}}_{j^*})}{f_{j^*-1}(y|x;\boldsymbol{\beta}_{j^*-1})}\; F_{j^*}(dy|x;\bar{\boldsymbol{\beta}}_{j^*})\,\xi(dx)$$

has a unique minimum in $\tilde{\boldsymbol{\beta}}_{j^*-1}$.

5.1.4. *$|\log f_{j^*-1}(y|x;\boldsymbol{\beta}_{j^*-1})| \leq m(y,x)$ for all $\boldsymbol{\beta}_{j^*-1}$ in some neighborhood of $\tilde{\boldsymbol{\beta}}_{j^*-1}$, with*

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} m(y,x)\; F_{j^*}(dy|x;\bar{\boldsymbol{\beta}}_{j^*})\,\xi(dx) < \infty.$$

Note that the design ξ_0^* satisfies that $M_j(\xi,\boldsymbol{\beta}_j)$ is not singular for any $j = 1,...,k$, since it maximizes (3.2); for the same reason, this property is satisfied by each ξ_r^*, $r \geq 1$, conditionally to the past.

Proposition 5.1. *Under the null hypothesis $H_{0,j}$, the test statistic $\mathcal{T}^j_{r,m}$ converges in distribution, as $m \to \infty$, to a chi-squared distributed random variable $\mathcal{T}^j_r$ having $(r+1)(d_j - d_{j-1})$ degrees of freedom, for any $r = 0,\ldots,n$.*

Proof. From Assumptions 5.1.1 and 5.1.2, $\mathcal{T}^j_{0,m}$ converges to a chi-squared distribution with $(d_j - d_{j-1})$ degrees of freedom (see [15, Theorem 22]).

For any $i = 1,\ldots,r$, the i-th term $T^j_{i,m}$ of $\mathcal{T}^j_{r,m}$ defined in equation (4.7) is a function of $(\mathbf{Y}_i,\mathbf{X}_i)$, and the response vector $\mathbf{Y}_i$ depends on the corresponding

exact design $\mathbf{X}_i$ and on all the past response vectors $\mathbf{Y}_{i-1}, \ldots, \mathbf{Y}_0$ and exact designs $\mathbf{X}_{i-1}, \ldots, \mathbf{X}_0$ only through $\mathbf{X}_i$; therefore

$$P(T^j_{i,m} \leq t_i | \mathbf{X}_i, \mathbf{Y}_{i-1}, \mathbf{X}_{i-1}, \ldots, \mathbf{Y}_0, \mathbf{X}_0) = P(T^j_{i,m} \leq t_i | \mathbf{X}_i). \tag{5.1}$$

Moreover, responses $Y_{i,1}, \ldots, Y_{i,m}$ are independent and identically distributed conditionally to the exact design $\mathbf{X}_i$, and hence, again from Assumptions 5.1.1 and 5.1.2, for $m \to \infty$

$$P(T^j_{i,m} \leq t_i | \mathbf{X}_i) \to P(T^j_i \leq t_i), \tag{5.2}$$

where T^j_i is a chi-squared distributed random variable with $(d_j - d_{j-1})$ degrees of freedom. Equations (5.1) and (5.2) imply that, for m growing to infinity, $T^j_{i,m}$ is asymptotically independent on $\sigma(\mathbf{X}_i, \mathbf{Y}_{i-1}, \mathbf{X}_{i-1}, \ldots, \mathbf{Y}_0, \mathbf{X}_0)$ and it is asymptotically distributed as a chi-squared random variable with $(d_j - d_{j-1})$ degrees of freedom. It follows that $\mathcal{T}^j_{r,m}$ is a sum of asymptotically independent, chi-squared distributed random variables, and hence

$$\mathcal{T}^j_{r,m} \xrightarrow{d} \mathcal{T}^j_r$$

as $m \to \infty$, where $\mathcal{T}^j_r = \sum_{i=1}^r T^j_i$ has a chi-squared distribution with $(r+1)(d_j - d_{j-1})$ degrees of freedom. □

From now on, let us denote by c^j_r the quantile of order $(1 - \alpha^j_r)$ of a chi-squared distribution with $(r+1)(d_j - d_{j-1})$ degrees of freedom. Then, at each stage r, the null hypothesis $H_{0,j}$ is rejected if $\mathcal{T}^j_{r,m} > c^j_r$, with an α^j_r asymptotic level of significance. Moreover, for $r = 0, \ldots, n$, and for $j = k, k-1, \ldots, 2$, let Z^j_r be the indicator of the event "the j-th model is selected at stage r", that is

$$Z^j_r = \begin{cases} 1, & \text{if } \mathcal{T}^h_{r,m} \leq c^h_r \text{ for } h = k, \ldots, j+1 \text{ and } \mathcal{T}^j_{r,m} > c^j_r \\ 0, & \text{otherwise,} \end{cases}$$

and for $j = 1$ let Z^1_r be the indicator of the event "the smaller model is selected at stage r", that is

$$Z^1_r = \begin{cases} 1, & \text{if } \mathcal{T}^h_{r,m} \leq c^h_r \text{ for } h = k, \ldots, 2 \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 5.1. *As $m \to \infty$,*

(a) $\hat{\boldsymbol{\beta}}_{0,j^*}$ and $\hat{\boldsymbol{\beta}}_{0,j^*-1}$ converge almost surely to $\bar{\boldsymbol{\beta}}_{j^*}$, and $\tilde{\boldsymbol{\beta}}_{j^*-1}$, respectively;

(b) in some neighborhoods of $\bar{\boldsymbol{\beta}}_{j^*}$, and $\tilde{\boldsymbol{\beta}}_{j^*-1}$, respectively,

$$\sup_{\boldsymbol{\beta}_{j^*}} \left| \frac{1}{m} \sum_{s=1}^{m} \log f_{j^*}(Y_{0,s}|X_{0,s};\boldsymbol{\beta}_{j^*}) - E(\log f_{j^*}(Y_{0,s}|X_{0,s};\boldsymbol{\beta}_{j^*})) \right| \to 0, \ a.s.,$$

$$\sup_{\beta_{j^*-1}} \left| \frac{1}{m} \sum_{s=1}^{m} \log f_{j^*-1}(Y_{0,s}|X_{0,s};\beta_{j^*-1}) - E(\log f_{j^*-1}(Y_{0,s}|X_{0,s};\beta_{j^*-1})) \right| \to 0, \ a.s.$$

Proof. (a) Assumptions 5.1.1 and 5.1.2 guarantees the the strong consistency of the maximum likelihood estimator of the parameter of the true model (see, for instance, [15, Theorem 18]). The convergence of the maximum likelihood estimator of the misspecified model is guaranteed by Assumptions 5.1.3 and 5.1.4, from [29, Theorem 2.2].
(b) To obtain the uniform law, apply [15, Theorem 16 (a)]. □

The next auxiliary lemma provides the "non-null" behavior of the test statistic.

Lemma 5.2. *There exists a constant $k_0 > 0$ such that, almost surely,*

$$\lim_{m\to\infty} \frac{\mathcal{T}_{0,m}^{j^*}}{m} = k_0.$$

Proof. For $i = 0$, the observations $(X_{i,s}, Y_{i,s})$, $s = 1, 2, \ldots, m$, are independent and identically distributed, therefore

$$\frac{\mathcal{T}_{0,m}^{j^*}}{m} = \frac{1}{m} \sum_{s=1}^{m} -2 \log \frac{f_{j^*-1}(Y_{0,s}|X_{0,s};\hat{\boldsymbol{\beta}}_{0,j^*-1})}{f_{j^*}(Y_{0,s}|X_{0,s};\hat{\boldsymbol{\beta}}_{0,j^*})}.$$

From the strong consistence of estimators and the uniform laws of large numbers, guaranteed by Lemma 5.1, $\mathcal{T}_{0,m}^{j^*}/m$ converges to

$$k_0 = E\left[-2 \log \frac{f_{j^*-1}(Y|X;\tilde{\boldsymbol{\beta}}_{j^*-1})}{f_{j^*}(Y|X;\bar{\boldsymbol{\beta}}_{j^*})} \right],$$

which is greater then zero from Jensen inequality. □

Theorem 5.1. *Let α_n^j be a sequence of significance levels such that $\alpha_n^j \to 0$ as $n \to \infty$ for any $j = 2, \ldots, k$. Let also $m = m(n)$ be a non decreasing sequence of integers such that $m \to \infty$ as $n \to \infty$, and $c_n^j/m \to 0$ as $n \to \infty$.*
Then, as the number of stages n converges to infinity, the sequential procedure selects the true model with probability converging to one. That is,

$$P(Z_n^{j^*} = 1) \to 1, \quad as\ n \to \infty.$$

Proof. If $j^* \in \{k, \ldots, 2\}$ then

$$\begin{aligned} P(Z_n^{j^*} = 1) &= P(\mathcal{T}_{n,m}^k \leq c_n^k, \cdots, \mathcal{T}_{n,m}^{j^*+1} \leq c_n^{j^*+1}, \mathcal{T}_{n,m}^{j^*} > c_n^{j^*}) \\ &= 1 - P(\{\mathcal{T}_{n,m}^k > c_n^k\} \cup \cdots \{\mathcal{T}_{n,m}^{j^*+1} > c_n^{j^*+1}\} \cup \{\mathcal{T}_{n,m}^{j^*} \leq c_n^{j^*}\}) \\ &\geq 1 - \left[\sum_{j=j^*+1}^{k} P(\mathcal{T}_{n,m}^j > c_n^j) + P(\mathcal{T}_{n,m}^{j^*} \leq c_n^{j^*}) \right]. \end{aligned} \tag{5.3}$$

Under the true model $f_{j^*}(y|x; \boldsymbol{\beta}_{j^*})$, it holds $P(\mathcal{T}_{n,m}^j > c_n^j) = \alpha_n^j$ for any $j > j^*$ since the models are nested. Thus inequality (5.3) becomes

$$P(Z_n^{j^*} = 1) \geq P(\mathcal{T}_{n,m}^{j^*} > c_n^{j^*}) - \sum_{j=j^*+1}^{k} \alpha_n^j. \tag{5.4}$$

The right-hand term of inequality (5.4) converges to 1 as $n \to \infty$ by the hypotheses on the α_n^j's and since

$$\lim_{n\to\infty} P(\mathcal{T}_{n,m}^{j^*} > c_n^{j^*}) = \lim_{n\to\infty} P\left(\frac{\mathcal{T}_{n,m}^{j^*}}{m} > \frac{c_n^{j^*}}{m} \right) = 1. \tag{5.5}$$

Convergence result (5.5) follows taking into account that $\mathcal{T}_{n,m}^{j^*} > \mathcal{T}_{0,m}^{j^*}$ and that

$$\lim_{n\to\infty} P\left(\frac{\mathcal{T}_{0,m}^{j^*}}{m} > \frac{c_n^{j^*}}{m} \right) = 1, \tag{5.6}$$

as a consequence of Lemma 5.2, since $c_n^{j^*}/m \to 0$ as $n \to \infty$,

In addition, if $j^* = 1$ then

$$\begin{aligned} P(Z_n^1 = 1) &= P(\mathcal{T}_{n,m}^k \leq c_n^k, \cdots, \mathcal{T}_{n,m}^2 \leq c_n^2) \\ &= 1 - P(\{\mathcal{T}_{n,m}^k > c_n^k\} \cup \cdots \{\mathcal{T}_{n,m}^2 > c_n^2\}) \\ &\geq 1 - \left[\sum_{j=2}^{k} P(\mathcal{T}_{n,m}^j > c_n^j) \right] = 1 - \sum_{j=2}^{k} \alpha_n^j. \end{aligned} \tag{5.7}$$

The right-hand term of inequality (5.7) converges to 1 as $n \to \infty$ by the hypotheses on the α_n^j's. □

In order to prove the next Theorem 5.2, arguments of asymptotic theory for argmin of convex random functions are used. References and some general results for real-valued random functions can be found in [20]. Since stochastic criterion function (4.3) takes values in the extended real axe $\bar{\mathbb{R}} = [-\infty, +\infty)$, here the results

treated in [17] and in [24] are extended to the metric space (S, d_w), where S is the set of probability distributions ξ with support $\mathcal{X} \subset \mathbb{R}^q$ (without loss of generality, let $q = 1$) and d_w is a metric which metrizes the weak convergence on $\mathcal{X}$. For instance, take the Kantorovich-Wasserstein metric (see [18]):

$$d_w(\xi_1, \xi_2) = \inf\{E(|X_1 - X_2|) : X_1 \sim \xi_1, X_2 \sim \xi_2\}.$$

Since $\mathcal{X}$ is compact, the metric space (S, d_w), which is an infinite-dimensional space, is complete and compact (from Prokhorov).

At first, a relevant auxiliary result about continuity and semi-continuity with respect to $\xi \in S$, of D- and KL-criteria, respectively, is provided by Proposition 5.2. Let us recall that, given a topological space S, a function $h : S \to \overline{\mathbb{R}}$ is *upper semi–continuous* (or *lower semi–continuous*, respectively) at x_0 if and only if for every $\varepsilon > 0$ there exists a neighborhood U of x_0 such that $h(x) \leq h(x_0) + \varepsilon$ for all $x \in U$ (or $h(x) \geq h(x_0) - \varepsilon$, respectively), equivalently,

$$\limsup_{x \to x_0} h(x) \leq h(x_0) \qquad (\text{or } \liminf_{x \to x_0} h(x) \geq h(x_0), \text{ respectively});$$

the function h is called *upper semi–continuous* (*lower semi–continuous*) if it is upper semi–continuous (lower semi–continuous) at every point of its domain. Let us assume that, for any $j = 2, \ldots, k$, models $f_j(y|x; \boldsymbol{\beta}_j)$ and $f_{j-1}(y|x; \boldsymbol{\beta}_{j-1})$ satisfy the following condition on their conditional Kullback-Leibler divergence.

Assumption 5.2. *The Kullback-Leibler conditional divergence $\mathcal{I}(x, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})$ defined in Equation (2.3) is continuous with respect to x.*

Proposition 5.2. *Under Assumption 5.2,*

(a) the D-criterion function from (S, d_w) to $[-\infty, +\infty)$:

$$\xi \mapsto \Phi_{D_j}[\boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j)]$$

is continuous;

(b) the KL-criterion function from (S, d_w) to $[0, +\infty)$:

$$\xi \mapsto I_{j-1,j}(\xi; \boldsymbol{\beta}_j)$$

is upper semi-continuous.

Proof. (a) Let us recall that $\boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j) = \int_{x \in \mathcal{X}} \boldsymbol{J}_j(x, \boldsymbol{\beta}_j)\, d\,\xi(x)$, where $\boldsymbol{J}_j(x, \boldsymbol{\beta}_j)$ is a $d_j \times d_j$ matrix whose components are bounded continuous functions from $\mathcal{X}$ to $\mathbb{R}$. It follows that the map $\xi \mapsto \boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j)$ is continuous because d_w metrizes the

weak convergence. Since also $\boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j) \mapsto \Phi_{D_j}[\boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j)]$ is continuous as showed in [23, Proposition IV.2], this proves the thesis.
(b) Let $z(\xi, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1}) = \int_{x \in \mathcal{X}} \mathcal{I}(x, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})\, d\,\xi(x)$, where $\mathcal{I}(x, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})$ is defined in equation (2.3). The map $\xi \mapsto z(\xi, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})$ from (S, d_w) to $\mathbb{R}$ is continuous because $\mathcal{I}(x, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})$ is a continuous function from $\mathcal{X}$ to $\mathbb{R}$ from Assumption 5.2 and d_w metrizes the weak convergence. As a consequence of the continuity of $z(\xi, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})$ with respect to ξ, the KL-criterion function $I_{j-1,j}(\xi; \boldsymbol{\beta}_j) = \inf_{\boldsymbol{\beta}_{j-1} \in \Theta_{j-1}} z(\xi, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})$ (see Definition 2.2) is upper semi-continuous. □

Another auxiliary result is provided in the following lemma.

Lemma 5.3. *Let R be the set of designs ξ such that every matrix $\boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j)$, $j = 1, \ldots, k$, in (3.2) is not singular for any value of $\boldsymbol{\beta}_j$. Then R is dense in S.*

Proof. Given a design ξ and a specific value for $\boldsymbol{\beta}_k$, it is well known that if $\boldsymbol{M}_k(\xi, \boldsymbol{\beta}_k)$ is positive definite then all the principal minors are positive. Since the models are nested, if $\boldsymbol{M}_k(\xi, \boldsymbol{\beta}_k)$ is positive definite for any value of $\boldsymbol{\beta}_k \in \Theta_k$, then $|\boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j)| > 0$ for every $j = 1, \ldots, k$ and $\boldsymbol{\beta}_j \in \Theta_j$; thus $\xi \in R$. Therefore if ξ_s is a design in $S \setminus R$ then $\boldsymbol{M}_k(\xi_s, \boldsymbol{\beta}_k)$ needs to be a non-negative definite matrix at least for some values of $\boldsymbol{\beta}_k$. Let us show that there exists a sequence ξ_n of elements in R such that $\lim_{n \to \infty} d_w(\xi_n, \xi_s) = 0$.

To this aim, let ξ_r be a design in R and let α_n be a sequence of real constants in $(0, 1)$ such that $\alpha_n \to 0$ as $n \to \infty$. The sequence of designs $\xi_n = (1 - \alpha_n)\xi_s + \alpha_n \xi_r$ belongs to R, because $\boldsymbol{M}_k(\xi_n, \boldsymbol{\beta}_k) = (1 - \alpha_n)\boldsymbol{M}_k(\xi_s, \boldsymbol{\beta}_k) + \alpha_n \boldsymbol{M}_k(\xi_r, \boldsymbol{\beta}_k)$ is positive definite. Moreover ξ_n converges to ξ_s weakly as $n \to \infty$, and hence the thesis is proved. □

Assumption 5.3. *The following equality*

$$\psi_{D_{j^*}}\left(x, \xi^*_{D_{j^*}}, \bar{\boldsymbol{\beta}}_{j^*}\right) = 0$$

has esactly d_{j^*} solutions, where $\psi_{D_{j^*}}(x, \xi^*_{D_{j^*}}, \bar{\boldsymbol{\beta}}_{j^*})$ is the directional derivative (2.7) evaluated at the D-optimum design for the true distribution $f_{j^*}(y|x; \bar{\boldsymbol{\beta}}_{j^*})$.

Remark 5.1. *Assumption 5.3 implies the uniqueness of the D-optimum design for model $f_{j^*}(y|x; \boldsymbol{\beta}_{j^*})$ from the Equivalence Theorem for the D-optimality criterion. For more details see [14], Theorem 2.4.1.*

Theorem 5.2. *If the Hypotheses of Theorem 5.1 mantain and $\sum_n \alpha_n^j < \infty$, then the sequence of designs ξ^*_n converges in probability to $\xi^*_{D_{j^*}}$, that is,*

$$P\left(d_w\Big[\xi^*_n(\omega), \xi^*_{D_{j^*}}\Big] < \varepsilon\right) \to 1,$$

for any $\varepsilon > 0$, as n grows to infinity.

Proof. First, let us prove that, whenever $\sum_n \alpha_n^j < \infty$,

$$P(Z_n^{j^*} = 1,\ ev.) = 1. \tag{5.8}$$

Let $j^* \in \{k, \dots, 2\}$. From Lemma 5.2 and from the hypothesis that $c_n^{j^*}/m \to 0$, it follows that

$$P\left(\frac{\mathcal{T}_{0,m}^{j^*}}{m} > \frac{c_n^{j^*}}{m},\ ev.\right) = 1, \text{ and, a fortiori, } P\left(\frac{\mathcal{T}_{n,m}^{j^*}}{m} > \frac{c_n^{j^*}}{m},\ ev.\right) = 1,$$

since $\mathcal{T}_{n,m}^{j^*} > \mathcal{T}_{0,m}^{j^*}$. In other words, for any $\varepsilon > 0$ there exists $N_1 = N_1(\varepsilon)$ such that

$$P\left(\frac{\mathcal{T}_{n,m}^{j^*}}{m} > \frac{c_n^{j^*}}{m},\ \text{for all } n \geq N_1\right) \geq 1 - \varepsilon. \tag{5.9}$$

Since $\sum_n \alpha_n^j < \infty$, there exists also $N_2 = N_2(\varepsilon)$ such that

$$\sum_{n \geq N_2} \sum_{j=j^*+1}^{k} \alpha_n^j < (k - j^* + 1)\,\varepsilon. \tag{5.10}$$

Let now $N = \max(N_1, N_2)$; with analogous calculations of (5.3),

$$\begin{aligned}
P\left(Z_n^{j^*} = 1, \text{for all } n \geq N\right) &= P\left(\bigcap_{n \geq N} \left\{\mathcal{T}_{n,m}^{k} \leq c_n^k, \cdots, \mathcal{T}_{n,m}^{j^*+1} \leq c_n^{j^*+1}, \mathcal{T}_{n,m}^{j^*} > c_n^{j^*}\right\}\right) \\
&= 1 - P\left(\bigcup_{n \geq N} \left\{\left\{\mathcal{T}_{n,m}^{k} > c_n^k\right\} \cup \cdots \left\{\mathcal{T}_{n,m}^{j^*+1} > c_n^{j^*+1}\right\} \cup \left\{\mathcal{T}_{n,m}^{j^*} \leq c_n^{j^*}\right\}\right\}\right) \\
&\geq 1 - \left[\sum_{n \geq N} \sum_{j=j^*+1}^{k} P\left(\mathcal{T}_{n,m}^{j} > c_n^j\right) + P\left(\bigcup_{n \geq N} \left\{\mathcal{T}_{n,m}^{j^*} \leq c_n^{j^*}\right\}\right)\right] \\
&= P\left(\bigcap_{n \geq N} \left\{\mathcal{T}_{n,m}^{j^*} > c_n^{j^*}\right\}\right) - \sum_{n \geq N} \sum_{j=j^*+1}^{k} \alpha_n^j.
\end{aligned} \tag{5.11}$$

From (5.9) and (5.10), the last term of the (5.11) is greater then $1 - (k - j^* + 2)\,\varepsilon$, and this proves result (5.8) for $j^* \in \{k, \dots, 2\}$.

If $j^* = 1$, then

$$\begin{aligned}
P\left(Z_n^{1} = 1, \text{for all } n \geq N\right) &= P\left(\bigcap_{n \geq N} \left\{\mathcal{T}_{n,m}^{k} \leq c_n^k, \cdots, \mathcal{T}_{n,m}^{2} \leq c_n^{2}\right\}\right) \\
&= 1 - P\left(\bigcup_{n \geq N} \left\{\left\{\mathcal{T}_{n,m}^{k} > c_n^k\right\} \cup \cdots \left\{\mathcal{T}_{n,m}^{2} > c_n^{2}\right\}\right\}\right) \\
&\geq 1 - \sum_{n \geq N} \sum_{j=2}^{k} P\left(\mathcal{T}_{n,m}^{j} > c_n^j\right) = 1 - \sum_{n \geq N} \sum_{j=2}^{k} \alpha_n^j > 1 - (k - j^* + 1)\,\varepsilon
\end{aligned}$$

and this proves result (5.8) for $j^* = 1$.

Equation (5.8) implies that $\lim_{n\to\infty} Z_n^{j^*} = 1$, almost surely, and then, from Cesaro's lemma, $\lim_{n\to\infty} \sum_{i=1}^n Z_i^{j^*}/n = 1$, almost surely. Hence

$$\lim_{n\to\infty} \gamma_{nj^*} = \lim_{n\to\infty} \left(\frac{\sum_{i=1}^n Z_i^{j^*}}{n} \right)^2 = 1, \tag{5.12}$$

almost surely. Moreover, since $Z_n^{j^*} = 1 - \sum_{j\neq j^*} Z_n^j$, it also follows obviously that

$$\lim_{n\to\infty} \gamma_{nj} = 0, \quad \text{a.s.}, \quad \text{for any } j \neq j^*, \quad \text{and} \quad \lim_{n\to\infty} \gamma_{nD} = 0, \quad \text{a.s.} \tag{5.13}$$

From (4.5), the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{n,j^*}$ for the true parameter of the true model is obtained from a proper likelihood function which doesn't depend on the past; if $n \to \infty$ also $m \to \infty$, then we have, as in Lemma 5.1(a),

$$\hat{\boldsymbol{\beta}}_{n,j^*} \to \bar{\boldsymbol{\beta}}_{j^*}, \tag{5.14}$$

a.s. Since $\Phi_{D_{j^*}}[\boldsymbol{M}_{j^*}(\xi, \boldsymbol{\beta}_{j^*})]$ is continuous with respect to the second argument, the continuous mapping theorem together with the (5.12) and (5.13) assure that, for any ξ such that every matrix $\boldsymbol{M}_j(\xi, \boldsymbol{\beta}_j)$, $j = 1, \ldots, k$, in (3.2) is not singular and for $n \to \infty$,

$$\Psi_{DKL}(\xi, \hat{\boldsymbol{\beta}}_n, \boldsymbol{\gamma}_n) \to \frac{1}{d_{j^*}} \log |\boldsymbol{M}_{j^*}(\xi, \bar{\boldsymbol{\beta}}_{j^*})|, \tag{5.15}$$

in probability. The limit in (5.15) is proportional to the D-optimality criterion function for the true model $f_{j^*}(y|x; \boldsymbol{\beta}_{j^*})$. Let

$$g_n(\xi)(\omega) = -\Psi_{DKL}\left[\xi, \hat{\boldsymbol{\beta}}_n(\omega), \boldsymbol{\gamma}_n(\omega)\right]$$

and

$$g(\xi) = -\frac{1}{d_{j^*}} \log |\boldsymbol{M}_{j^*}(\xi, \bar{\boldsymbol{\beta}}_{j^*})|,$$

hence the sequence of random functions $g_n(\xi)(\omega)$ converges in probability, and then also in distribution, to the function $g(\xi)$ for any $\xi \in R$, which is a dense subset of S by Lemma 5.3. Let us recall that $g_n(\xi)(\omega)$, for any $n \geq 0$, and the limit $g(\xi)$ are convex functions with respect to ξ, as showed in Section 3. Moreover $g_n(\cdot)(\omega)$ is lower semi-continuous because, from Proposition 5.2, it is a linear combination of lower semi-continuous functions on $(-\infty, +\infty]$, while $g(\cdot)$ is continuous. As a consequence of compactness and convexity of the space S and of the continuity of the D-criterion, $g_n(\xi)(\omega)$ and $g(\xi)$ are finite on some open set. Finally, from Assumption 5.3, the infimum of $g(\xi)$ is achieved at a unique point $\xi^*_{D_{j^*}}$. From Lemma 3.1 and Theorem 3.2 in [17] it follows that $\xi_n^*(\omega)$ converges in distribution to $\xi^*_{D_{j^*}}$. Since this limit is not random, this is equivalent to convergence in probability [see 3], and this proves the thesis. □

6 Conclusion and further developments

The DKL-criterion of optimality, proposed by [27], is useful to choose experimental conditions which are "good" to discriminate between two rival models as well to estimate efficiently the parameters of the selected model. This paper deals again with the dual problem of model selection and parameter estimation, but more than two rival models are considered. Hence, to handle the case of several nested non-linear models, a modification of the DKL-criterion is herein given. This new criterion is called generalized DKL-criterion. An interesting theoretical result proved in this paper is the continuity and the upper semi-continuity, with respect to the design ξ, of the D- and the KL-criterion functions, respectively. As a consequence, also the generalized DKL-criterion is upper semi-continuous.

The generalized DKL-criterion depends on the values of the model parameters because of the non-linearity of the models. To overcome the problem that the true values of the parameters are unknown, a sequential procedure is then proposed. At each step of this sequential scheme, a generalized DKL-optimum design is computed using the maximum likelihood estimates obtained at the previous step (this is called sequential generalized DKL-optimum design). Then m experimental conditions are generated from such sequential adaptive DKL-optimum design and the corresponding responses are observed. Finally some statistical tests are performed in a stepwise manner until a specific model is selected. The sequential procedure here proposed selects the true model with probability that tends to one; moreover, the sequential generalized DKL-optimum design converges in probability to the D-optimum design for the true model, as the number of stages increases to infinity.

Let us observe that, since the rival models considered in this paper are nested and the D_s-criterion is useful to discriminate between nested models, a weighted geometric mean of D- and D_s-efficiencies could be another possible criterion of optimality instead of the generalized DKL-criterion. Let us call generalized DD_s-criterion this possible combination of efficiencies. In this way, the criterion proposed by [28] would be extended to the case of k models. In addition, a sequential adaptive DD_s-optimum design could be performed in a similar way than the sequential procedure proposed in this paper. The comparison between the performances of these two sequential adaptive designs will be a matter of future investigation.

Let us finally remark that, differently from the D_s-criterion, the KL-criterion can be used furthermore to discriminate between separate models. Thus, a generalization of the herein proposed sequential procedure to the case of several non-nested models will be studied in future, as well.

Acknowledgments The authors are very grateful to Professor Giacomo Aletti for his useful suggestions and comments which contributed to this work. They are also grateful to the organizers of the Design and Analysis of Experiment programme at

the Isaac Newton Institute for Mathematical Science for the great support and warm hospitality.

References

[1] A. C. Atkinson. DT-optimum designs for model discrimination and parameter estimation. *J. Statist. Plann. Inference*, 138(1):56–64, 2008.

[2] A. C. Atkinson, A. N. Donev, and R. D. Tobias. *Optimum experimental designs, with SAS*, volume 34 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, 2007.

[3] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[4] A. Biswas and P. Chaudhuri. An efficient design for model discrimination and parameter estimation in linear models. *Biometrika*, 89(3):709–718, 2002.

[5] D. M. Borth. A total entropy criterion for the dual problem of model discrimination and parameter estimation. *J. Roy. Statist. Soc. Ser. B*, 37:77–87, 1975.

[6] H. Chernoff. Approaches in sequential design of experiments. In *A survey of statistical design and linear models (Proc. Internat. Sympos., Colorado State Univ., Ft. Collins, Colo., 1973)*, pages 67–90. North-Holland, Amsterdam, 1975.

[7] H. Dette. On a mixture of the D- and D_1-optimality criterion in polynomial regression. *J. Statist. Plann. Inference*, 35(2):233–249, 1993.

[8] H. Dette and T. Franke. Constrained D- and D_1-optimal designs for polynomial regression. *Ann. Statist.*, 28(6):1702–1727, 2000.

[9] H. Dette and T. Franke. Robust designs for polynomial regression by maximizing a minimum of D- and D_1-efficiencies. *Ann. Statist.*, 29(4):1024–1049, 2001.

[10] H. Dette and R. Kwiecien. A comparison of sequential and non-sequential designs for discrimination between nested regression models. *Biometrika*, 91(1):165–176, 2004.

[11] H. Dette, V. B. Melas, and W. K. Wong. Optimal design for goodness-of-fit of the Michaelis-Menten enzyme kinetic function. *J. Amer. Statist. Assoc.*, 100(472):1370–1381, 2005.

[12] H. Dette and A. Pepelyshev. Efficient experimental designs for sigmoidal growth models. *J. Statist. Plann. Inference*, 138(1):2–17, 2008.

[13] V. V. Fedorov. *Theory of optimal experiments*. Academic Press, New York, 1972. Translated from the Russian and edited by W. J. Studden and E. M. Klimko, Probability and Mathematical Statistics, No. 12.

[14] V. V. Fedorov and P. Hackl. *Model-oriented design of experiments*, volume 125 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1997.

[15] T. S. Ferguson. *A course in large sample theory*. Texts in Statistical Science Series. Chapman & Hall, London, 1996.

[16] I. Ford, D. M. Titterington, and C. P. Kitsos. Recent advances in nonlinear experimental design. *Technometrics*, 31(1):49–60, 1989.

[17] C. J. Geyer. On the asymptotics of convex stochastic optimization. Unpublished manuscript, Available on the web, 1996.

[18] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *Internat. Statist. Rev.*, 70(3):419–435, 2002.

[19] W. J. Hill, W. G. Hunter, and D. W. Wichern. A joint design criterion for the dual problem of model discrimination and parameter estimation. *Technometrics*, 10:145–160, 1968.

[20] K. Kato. Asymptotics for argmin processes: convexity arguments. *J. Multivariate Anal.*, 100(8):1816–1829, 2009.

[21] J. López-Fidalgo, C. Tommasi, and P. C. Trandafir. An optimal experimental design criterion for discriminating between non-normal models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(2):231–242, 2007.

[22] G. Montepiedra and A. B. Yeh. A two-stage strategy for the construction of D-optimal experimental designs. *Comm. Statist. Simulation Comput.*, 27(2):377–401, 1998.

[23] A. Pázman. *Foundations of optimum experimental design*, volume 14 of *Mathematics and its Applications (East European Series)*. D. Reidel Publishing Co., Dordrecht, 1986. Translated from the Czech.

[24] R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[25] S. D. Silvey. *Optimal design*. Chapman & Hall, London, 1980. An introduction to the theory for parameter estimation, Monographs on Applied Probability and Statistics.

[26] C. Tommasi. Optimal designs for discriminating among several non-normal models. In *mODa 8—Advances in model-oriented design and analysis*, Contrib. Statist., pages 213–220. Physica-Verlag/Springer, Heidelberg, 2007.

[27] C. Tommasi. Optimal designs for both model discrimination and parameter estimation. *J. Statist. Plann. Inference*, 139(12):4123–4132, 2009.

[28] M.-H. Tsai and M.-M. Zen. Criterion-robust optimal designs for model discrimination and parameter estimation: multivariate polynomial regression case. *Statist. Sinica*, 14(2):591–601, 2004.

[29] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[30] D. P. Wiens. Robust discrimination designs. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(4):805–829, 2009.

[31] M.-M. Zen and M.-H. Tsai. Criterion-robust optimal designs for model discrimination and parameter estimation in Fourier regression models. *J. Statist. Plann. Inference*, 124(2):475–487, 2004.