Statistica Sinica **24** (2014), 63-82
doi:http://dx.doi.org/10.5705/ss.2012.258

MODEL SELECTION AND PARAMETER ESTIMATION IN NON-LINEAR NESTED MODELS: A SEQUENTIAL GENERALIZED DKL-OPTIMUM DESIGN

Caterina May and Chiara Tommasi

Università del Piemonte Orientale and Università di Milano

Abstract: This work proposes a sequential procedure 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 that is optimal for the goals of model selection and parameter estimation. Subsequently, at each step, an adaptive generalized DKL-optimum design is computed from the data accrued and the tests previously performed. The proposed sequential scheme selects the best non-linear model with probability converging to one; moreover it allows efficient estimates of parameters, since the adaptive sequential DKL-optimum designs converge to the D-optimum design for the "true" model.

Key words and phrases: Argmin processes, convexity, D-optimality, KL-optimality, DKL-optimality, log-likelihood ratio test, semi-continuity, sequential design of experiments, stochastic convergence.

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; therefore, the goal of an optimum design is to estimate efficiently the parameters of the assumed model. However, more frequently, in applications several rival models are available. Thus, both the aims of model selection and parameter estimation should be achived by an optimum design. For this reason, some authors have studied compound criteria that combine criteria for parameter estimation (usually the D-criterion) and for model discrimination. In the literature there exist several optimality criteria which are useful to discriminate between models and may be applied in different settings (D_s-, T- and KL-criteria). The D_s-criterion can be used when the rival models are nested; thus, in the context of two nested regression models which differ by $s > 1$ parameters, Tsai and Zen (2004) and Zen and Tsai (2004) have considered the DD_s-compound criterion, given by a weighted geometric mean of D_s- and D-efficiencies (the case $s = 1$ is studied by Dette (1993)). Differently, the T-criterion can be applied

to discriminate nested or separate models but they must be homoscedastic with Gaussian errors; an extension of the T-criterion for non-homoscedastic errors is given in Uciński and Bogacka (2005); Atkinson (2008) has proposed the DT-criterion which is a weighted geometric mean of T- and D-efficiencies. Finally, the KL-criterion (proposed by López-Fidalgo, Tommasi, and Trandafir (2007)) can be applied when the rival models are nested or not, homoscedastic or heteroscedastic, and with any error distribution. In this general context, Tommasi (2009) 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 of selecting the correct model and estimating efficiently its parameters. This new criterion is a weighted mean of a measure of discrimination (based on the KL-criterion) and a measure of estimation (based on the D-criterion) and is called generalized DKL-criterion.

Only compound criteria are considered in this paper. However, there exist several ways to incorporate different goals in one design criterion. Some examples are given in Dette, and Franke (2000, 2001), among others.

When the rival models are non-linear, the designs which maximize multi-objective criteria are only locally optimum, because the optimality criterion functions depend on the unknown parameters of the models. To go further, one might follow a Bayesian approach; see for instance, Hill, Hunter, and Wichern (1968) and Borth (1975). Variously, one could use a max-min criterion; some examples are Dette, Melas, and Wong (2005) and Dette, and Pepelyshev (2008). In this paper it is assumed that experiments can be performed sequentially and we adopt rather a sequential adaptive approach; see, for instance, Chernoff (1975), Ford, Titterington, and Kitsos (1989) and Wiens (2009).

In more detail, at the first step of the proposed sequential procedure some nominal values for the parameters are guessed and a generalized DKL-optimum design is computed; then a bivariate sample of independent experimental conditions and responses are observed and some tests of hypotheses are performed to select a model. Subsequently, at each step of the sequential procedure, an adaptive optimum design is computed maximizing an "updated" generalized DKL-criterion with the unknown parameters estimated through the data obtained at the previous step; in addition, the weights corresponding to the discrimination and estimation measures are updated taking into account all the past statistical hypothesis tests. Then, a bivariate sample of experimental conditions and responses are observed (conditionally independent to the past) and some statistical tests are performed to select a model. The adaptive generalized DKL-optimum design that maximizes this "updated" generalized DKL-criterion is called a sequential generalized DKL-optimum design. This scheme simultaneously achieves

asymptotically 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.

Our methodology is partly anticipated in the sequential scheme proposed by Biswas and Chaudhuri (2002) that is applicable only in the set up of nested linear models. However, the proposed approach is different from that of Biswas and Chaudhuri: in order to find an adaptive optimum design they do not update an optimality criterion; in addition, they do not consider an adaptive design which is also "good" for model discrimination. Biswas and Chaudhuri (2002), as well as Montepiedra and Yeh (1998), use the sequential approach essentially to update the information about the form of the unknown linear model. In addition, Dette and Kwiecien (2004) have compared, through a simulation study, Biswas and Chaudhuri's sequential design with some non-sequential optimum designs. They show the superiority of non-sequential methods. In this paper non-linear models are considered, and a sequential procedure is considered a useful device to avoid model parameter dependence.

The outline of the paper is as follows. In Section 2, the basic notation is established, 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 an adaptive sequential procedure, where, at each step, a generalized DKL-optimum design is computed on the basis of past data and performed tests. In Section 5, together with some important auxiliary results, 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. In Section 6, some ideas about future developements are discussed.

2. Notation Setting and Description of the Models

Let an experimental condition X be generated by the experimenter from a design ξ: X is a random variable (or a random vector) having probability distribution ξ that 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 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 conditional on X, each one depending on a vector of unknown parameters $\boldsymbol{\beta}_j$ in Θ_j, an open subset of $\mathbb{R}^{d_j}$.

Models $F_j(y|X;\boldsymbol{\beta}_j)$ satisfy standard hypotheses of regularity as follows. For each $j=1,\dots,k$, Y has a conditional density $f_j(y|x;\boldsymbol{\beta}_j)$ that is twice differentiable in $\boldsymbol{\beta}_j$ and is supported independently of $\boldsymbol{\beta}_j$. We assume that the models are identifiable.

Moreover assume that, for any $j=2,\dots,k$, $\boldsymbol{\beta}_j^T=(\boldsymbol{\beta}_{j-1}^T,\boldsymbol{\tau}_j^T)$, where $\boldsymbol{\tau}_j$ is the vector of the last d_j-d_{j-1} components of $\boldsymbol{\beta}_j$, and that assigning a specific value $\boldsymbol{\tau}_j^0$ to $\boldsymbol{\tau}_j$ yields $f_j[y|x;(\boldsymbol{\beta}_{j-1}^T,\boldsymbol{\tau}_j^{0\,T})^T]=f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})$, 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 $(Y_1;X_1),\dots,(Y_m;X_m)$, some statistical tests can be carried out in a stepwise manner until a specific statistical model is selected. The tests are performed for the hypotheses

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

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

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 can be selected by maximizing the KL–optimality criterion,

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

where $\mathcal{Y}\subseteq\mathbb{R}$ is the support of Y and $w(x)=\xi(dx)/\nu(dx)$. If the larger 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 Tommasi (2007)) and a design $\xi_{j-1,j}^*$ that 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 $f_{j-1}(y|x;\boldsymbol{\beta}_{j-1})$. For a fixed value of $\boldsymbol{\beta}_j$, a design for which the 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 a *KL–singular* design. Assuming that $\xi^*_{j-1,j}$ is regular, López-Fidalgo, Tommasi, and Trandafir (2007) 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 $x \in \mathcal{X}$, where

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

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 (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

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

This efficiency is in $(0, 1)$ and 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. If H_{0j} is rejected for some $j \in \{k, \ldots, 2\}$, 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, Fedorov (1972), Pázman (1986) and Atkinson, Donev, and Tobias (2007). In the context of non-linear models (see Silvey (1980)), the D-optimality criterion is defined as

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

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

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

3. Generalized DKL-criterion for Several Nested Models

The DKL-optimality criterion to discriminate between two statistical models and to estimate efficiently their parameters has been proposed in Tommasi (2009). This criterion is here generalized to the case of k nested models by the weighted geometric mean of efficiencies

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

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 $0 \leq \gamma_j \leq 1$ for any $j = 1, \ldots, k$, and $0 \leq \gamma_D \leq 1$, 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 that 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 maximizing the criterion function

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

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 Theorem 3.1, a stronger definition of regular design is adopted.

Definition 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 an equivalence theorem can be stated.

Theorem 1. *A regular design ξ^*_{DKL} is generalized DKL-optimum if and only if*

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

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 while efficiently estimating the parameters of the true model. A generalized DKL-optimum design as proposed in Section 3 can be computed to provide experiments to be performed, but, since the models are non-linear, the optimality would be reached only locally. 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 $r = 0$, a design maximizing criterion (3.2) based on a nominal value for $\boldsymbol{\beta}$ and on an arbitrary choice of values for γ_j $(j = 1, \ldots, k)$ is computed. Let $\xi^*_{DKL} = \xi^*_0$ be such a generalized DKL-optimum design. Then m independent experimental conditions are generated from ξ^*_0; 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

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

for $j = k, k-1, \ldots, 2$, until a specific null hypothesis is rejected. Here (4.1) is equivalent to (2.1) as the models are nested. Consider the -2 log-likelihood ratio statistic

$$\mathcal{T}^j_{0,m} = -2 \log \frac{L_0^{j-1}(\hat{\boldsymbol{\beta}}_{0,j-1})}{L_0^j(\hat{\boldsymbol{\beta}}_{0,j})}$$

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 $l = j-1, j$, $L_0^l(\hat{\boldsymbol{\beta}}_{0,l})$ 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 $\mathcal{T}_{0,m}^j > c_{0,j}$, $c_{0,j}$ being the cut-off point corresponding to the significance level $\alpha_{0,j}$.

For $r = 1, \ldots, n$, we define, for instance, 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; if the proportion for $f_{\bar{\jmath}}(y|x; \boldsymbol{\beta}_{\bar{\jmath}})$ is 1, then $\gamma_{r-1,\bar{\jmath}} = 1 - 1/2r$ and $\gamma_{r-1,j} = 0$ for $j \neq \bar{\jmath}$. Finally, let

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

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

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

if $\boldsymbol{M}_j[\xi, \hat{\boldsymbol{\beta}}_{r-1,j}(\omega)]$ 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 conditionally independent and identically distributed random variables with respect to the past $\sigma(\mathbf{Y}_0, \mathbf{X}_0)$, having conditional distribution ξ_1^* is generated. 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

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

A vector of m experimental conditions $\mathbf{X}_r = (X_{r,1}, \ldots, X_{r,m})^T$, 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 from ξ_r^*. 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 $\mathbf{X}_r$. Hence (4.4) satisfies

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

In the notation of (4.3) it is stressed that the second and third arguments of $\Psi_{DKL}(\cdot, \cdot, \cdot)$ are random, and hence the optimal designs ξ_r^*, for an $r \geq 1$, are stochastic distributions.

Then, hypotheses (4.1) are tested through the statistic

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

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

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

is the log-likelihood ratio statistic based on the 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 $\alpha_{r,j}$.

Note that for ease of notation we have considered the same number, m, of observations at each step; this could be straightforwardly 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 $r = 0, \ldots, n$, have been 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 2 and Theorem 3 which guarantee two basic properties of the sequential procedure. Some methods used in Biswas and Chaudhuri (2002) are extended to the different scheme proposed in this paper. Theorem 2 assures that the true model is asymptotically selected; Theorem 3 states that the sequence of generalized DKL-optimum designs converges in probability to the D-optimal design for the true model. In addition, some auxiliary results are provided. The first one is Proposition 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 conditional on X be $f_{j^*}(y|x;\boldsymbol{\beta}_{j^*})$, $j^* \in \{1,\dots,k\}$, and let $\bar{\boldsymbol{\beta}}_{j^*}$ denote the true value of the parameter; this means that, whenever $j^* \geq 2$, the last components of $\bar{\boldsymbol{\beta}}_{j^*}$ satisfy $\bar{\boldsymbol{\tau}}_{j^*} \neq \boldsymbol{\tau}^0_{j^*}$. Further assumptions on the models are required.

Assumption 1. *For any design ξ such that $M_{j*}(\xi,\boldsymbol{\beta}_{j^*})$ is not singular:*

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

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*

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

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 $\tilde{\boldsymbol{\beta}}_{j^*-1}$;

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

Proposition 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-square, $\mathcal{T}^j_r$, having $(r+1)(d_j - d_{j-1})$ degrees of freedom, for any $r = 0,\dots,n$.*

Proof. From Assumptions 1.1 and 1.2, $\mathcal{T}^j_{0,m}$ converges to a chi-squared distribution with $(d_j - d_{j-1})$ degrees of freedom (see Ferguson (1996, Thm. 22)).

For any $i = 1, \dots, r$, the ith term $T^j_{i,m}$ of $\mathcal{T}^j_{r,m}$, defined at (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 $\mathbf{Y}_{i-1}, \dots, \mathbf{Y}_0$ and exact designs $\mathbf{X}_{i-1}, \dots, \mathbf{X}_0$ only through $\mathbf{X}_i$; therefore

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

Moreover, responses $Y_{i,1}, \dots, Y_{i,m}$ are independent and identically distributed conditionally to the exact design $\mathbf{X}_i$, and hence, again from Assumptions 1.1 and 1.2, for $m \to \infty$,

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

where T^j_i is chi-square 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 of $\sigma(\mathbf{X}_i, \mathbf{Y}_{i-1}, \mathbf{X}_{i-1}, \dots, \mathbf{Y}_0, \mathbf{X}_0)$ and it is asymptotically distributed as chi-square with $(d_j - d_{j-1})$ degrees of freedom. It follows that $\mathcal{T}^j_{r,m}$ is a sum of asymptotically independent, chi-squares, and hence

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

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

From now on, we 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 $r = 0, \dots, n$, and for $j = k, k-1, \dots, 2$, let Z^j_r be the indicator of the event "the jth model is selected at stage r", that is

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

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

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

Lemma 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 1.1 and 1.2 guarantees the the strong consistency of the maximum likelihood estimator of the parameter of the true model (see for instance Ferguson (1996, Thm. 18)). The convergence of the maximum likelihood estimator of the misspecified model is guaranteed by Assumptions 1.3 and 1.4, see White (1982, Thm. 2.2).

(b) To obtain the uniform law, apply Ferguson (1996, Thm. 16 (a)).

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

Lemma 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 $(X_{i,s}, Y_{i,s})$, $s = 1, \ldots, m$, 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 consistency of estimators and the Uniform Law of Large Numbers, guaranteed by Lemma 1, $\mathcal{T}_{0,m}^{j^*}/m$ converges to

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

which is greater than zero by Jensen's inequality.

Theorem 2. *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* $m = m(n)$ *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* $P(Z_n^{j^*} = 1) \to 1$, *as* $n \to \infty$.

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

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

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

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

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

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

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

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

as a consequence of Lemma 2, since $c_n^{j^*}/m \to 0$ as $n \to \infty$.

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

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

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

In order to prove Theorem 3, 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 Kato (2009). Since the stochastic criterion function (4.3) takes values in the extended real axis $\bar{\mathbb{R}} = [-\infty, +\infty)$, the results treated in Geyer (1996) and in Rockafellar (1970) are extended to the metric space (S, d_w), where S is the set of probability distributions ξ with

support $\mathcal{X}$ and d_w is a metric which metrizes the weak convergence on $\mathcal{X}$. For instance, take the Kantorovich-Wasserstein metric (see Gibbs and Su (2002)):

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

Since $\mathcal{X}$ is compact, the metric space (S, d_w), which is an infinite-dimensional space, is complete and compact (as a consequence of Prokhorov's theorem).

At first, an auxiliary result about continuity and semi-continuity with respect to $\xi \in S$, of D- and KL-criteria, respectively, is provided by Proposition 2. Given a topological space S, a function $h : S \to \bar{\mathbb{R}}$ is *upper semi–continuous* 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 $x \in U$, equivalently, $\limsup_{x \to x_0} h(x) \leq h(x_0)$. The function h is called *upper semi–continuous* if it is upper semi–continuous at every point of its domain, with a similar definition for *lower semi–continuous*.

Assumption 2. *The Kullback-Leibler conditional divergence $\mathcal{I}(x, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j-1})$ at* (2.3) *is continuous with respect to x, $j = 2, \ldots, k$.*

Proposition 2. *Under Assumption* 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) 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 shown in Pázman (1986, Proposition IV.2), this proves the result.

(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 at (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 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 equation (2.2)) is upper semi-continuous.

Lemma 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 $\lim_{n\to\infty} d_w(\xi_n, \xi_s) = 0$.

For this purpose, 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 result is proved.

Assumption 3. *The equality $\psi_{D_{j^*}}(x, \xi^*_{D_{j^*}}, \bar{\boldsymbol{\beta}}_{j^*}) = 0$ has exactly 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 1. Assumption 3 implies the uniqueness of the D-optimum design for model $f_{j^*}(y|x; \boldsymbol{\beta}_{j^*})$ by the Equivalence Theorem for the D-optimality criterion. For more details see Fedorov, and Hackl (1997, Thm. 2.4.1).

Theorem 3. *If the hypotheses of Theorem* 2 *hold and $\sum_n \alpha_n^j < \infty$, then, for the sequence of designs ξ^*_n,*

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

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

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

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

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

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

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

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

$$P\left(Z_n^{j^*}=1, \text{for all } n \geq N\right) = P\Big(\bigcap_{n\geq N} \Big\{\mathcal{T}_{n,m}^k \leq c_n^k, \ldots, \mathcal{T}_{n,m}^{j^*+1} \leq c_n^{j^*+1}, \mathcal{T}_{n,m}^{j^*} > c_n^{j^*}\Big\}\Big)$$

$$= 1 - P\Big(\bigcup_{n\geq N} \Big\{ \{\mathcal{T}_{n,m}^k > c_n^k\} \cup \cdots \{\mathcal{T}_{n,m}^{j^*+1} > c_n^{j^*+1}\} \cup \{\mathcal{T}_{n,m}^{j^*} \leq c_n^{j^*}\}\Big\}\Big)$$

$$\geq 1 - \Big[\sum_{n\geq N} \sum_{j=j^*+1}^{k} P\left(\mathcal{T}_{n,m}^j > c_n^j\right) + P\Big(\bigcup_{n\geq N} \Big\{\mathcal{T}_{n,m}^{j^*} \leq c_n^{j^*}\Big\}\Big)\Big]$$

$$= P\Big(\bigcap_{n\geq N} \Big\{\mathcal{T}_{n,m}^{j^*} > c_n^{j^*}\Big\}\Big) - \sum_{n\geq N} \sum_{j=j^*+1}^{k} \alpha_n^j. \tag{5.11}$$

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

If $j^* = 1$, then

$$P\left(Z_n^1 = 1, \text{for all } n \geq N\right) = P\Big(\bigcap_{n\geq N} \Big\{\mathcal{T}_{n,m}^k \leq c_n^k, \cdots, \mathcal{T}_{n,m}^2 \leq c_n^2\Big\}\Big)$$

$$= 1 - P\Big(\bigcup_{n\geq N} \Big\{\{\mathcal{T}_{n,m}^k > c_n^k\} \cup \cdots \{\mathcal{T}_{n,m}^2 > c_n^2\}\Big\}\Big)$$

$$\geq 1 - \sum_{n\geq N} \sum_{j=2}^{k} P\left(\mathcal{T}_{n,m}^j > c_n^j\right) = 1 - \sum_{n\geq N} \sum_{j=2}^{k} \alpha_n^j > 1 - (k - j^* + 1)\,\varepsilon$$

and this proves (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 (see Williams (1991, p.116)), $\lim_{n\to\infty} \sum_{i=1}^n Z_i^{j^*}/n = 1$, 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 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 does not depend on the past; if $n \to \infty$ then also $m \to \infty$, as assumed in the hypotheses of Theorem 2; hence we have, as in Lemma 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), assures 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

$$g_n(\xi)(\omega) = -\Psi_{DKL}\Big[\xi, \hat{\boldsymbol{\beta}}_n(\omega), \boldsymbol{\gamma}_n(\omega)\Big],$$
$$g(\xi) = -\frac{1}{d_{j^*}} \log |\boldsymbol{M}_{j^*}(\xi, \bar{\boldsymbol{\beta}}_{j^*})|.$$

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 3. Now $g_n(\xi)(\omega)$, for any $n \geq 0$, and the limit $g(\xi)$ are convex functions with respect to ξ, as shown in Section 3. Moreover $g_n(\cdot)(\omega)$ is lower semi-continuous because, from Proposition 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 3, the infimum of $g(\xi)$ is achieved at a unique point $\xi^*_{D_{j^*}}$. From Lemma 3.1 and Theorem 3.2 in Geyer (1996) 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 Billingsley (1999)), and this proves the result.

6. Conclusion and Further Developments

The DKL-criterion of optimality, proposed by Tommasi (2009), is useful to choose experimental conditions which are "good" for discriminating between two rival models, as well as to estimate efficiently the parameters of the selected model. We tackle the problem when there are more than two rival models to be considered. To handle the case of several nested non-linear models, a modification of the DKL-criterion is given, termed the generalized DKL-criterion. We prove 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, the generalized DKL-criterion is also upper semi-continuous.

The generalized DKL-criterion depends on the values of the model parameters both for non-linear models and some linear models (when two subsequent linear models differ by more than one coefficient). To overcome the problem that

the true values of the parameters are unknown, we propose a sequential procedure. Our sequential procedure 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. In order to investigate after how many steps the sequential procedure could be stopped, a simulation study will be developed in future work. In addition, the asymptotic distribution of the test statistic (used at each step of the sequential scheme) will be compared with its Monte-Carlo distribution, to determine how many observations should be taken at each stage.

The choices of the weights in Section 4 and of the cut-off points in Section 5 are at the moment based on a simple rule, and they are very general. It is of interest to develop some theory leading to optimum choices in order to improve as much as possible the speed of convergence to the D-optimum design for the true model.

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 is another 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 Tsai and Zen (2004) 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 to the sequential procedure proposed here. The comparison between the performances of these two sequential adaptive designs will be a matter of future investigation.

Differently from the D_s-criterion, the KL-criterion can be used to discriminate between separate models. Thus, a possible extension of our sequential procedure to the situation of several non-nested models would be of great interest. Such generalization is however not straightforward. For instance, testing separate models requires that all the models need to be compared in pairs; for each comparison two tests have to be performed- interchanging the role of the null and the alternative hypothesis- and multiple answers may be reached. This extension will be object of the future research, as well.

Acknowledgements

The authors are grateful to Professor Giacomo Aletti for his suggestions that contributed to improving 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. Many thanks also to an anonymous referee for his comments.

References

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

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*, Oxford University Press, Oxford.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, New York

Biswas, A. and Chaudhuri, P. (2002). An efficient design for model discrimination and parameter estimation in linear models. *Biometrika* **89**, 709-718.

Borth, D. M. (1975). A total entropy criterion for the dual problem of model discrimination and parameter estimation. *J. Roy. Statist. Soc. Ser. B* **37**, 77-87.

Chernoff, H. (1975). 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), 67-90, North-Holland, Amsterdam.

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

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

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

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

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

Dette, H. and Pepelyshev, A. (2008). Efficient experimental designs for sigmoidal growth models. *J. Statist. Plann. Inference* **138**, 2-17.

Fedorov, V. V. (1972). *Theory of Optimal Experiments*. (Translated from the Russian and edited by W. J. Studden and E. M. Klimko), Academic Press, New York.

Fedorov, V. V. and Hackl, P. (1997). *Model-Oriented Design of Experiments*, volume 125 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

Ford, I., Titterington, D. M. and Kitsos, C. P. (1989). Recent advances in nonlinear experimental design. *Technometrics* **31**, 49-60.

Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization, unpublished manuscript.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *Internat. Statist. Rev.* **70**, 419-435.

Hill, W. J., Hunter, W. G. and Wichern, D. W. (1968). A joint design criterion for the dual problem of model discrimination and parameter estimation. *Technometrics* **10**, 145-160.

Kato, K. (2009). Asymptotics for argmin processes: convexity arguments. *J. Multivariate Anal.* **100**, 1816-1829.

López-Fidalgo, J., Tommasi, C. and Trandafir, P. C. (2007). An optimal experimental design criterion for discriminating between non-normal models. *J. Roy. Statist. Soc. Ser. B* **69**, 231-242.

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

Pázman, A. (1986). *Foundations of Optimum Experimental Design*, (Translated from the Czech), D. Reidel Publishing Co., Dordrecht.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, N.J.

Silvey, S. D. (1980). *Optimal Design*. Chapman & Hall, London.

Tommasi, C. (2007). Optimal designs for discriminating among several non-normal models. In *mODa 8—Advances in Model-oriented Design and Analysis*, 213-220, Physica-Verlag/Springer, Heidelberg.

Tommasi, C. (2009). Optimal designs for both model discrimination and parameter estimation. *J. Statist. Plann. Inference* **139**, 4123-4132.

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

Uciński, D. and Bogacka, B. (2005). T-optimum designs for discrimination between two multiresponse dynamic models. *J. Roy. Statist. Soc. Ser. B* **67**, 3-18.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.

Wiens, D. P. (2009). Robust discrimination designs. *J. Roy. Statist. Soc. Ser. B* **71**, 805-829.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press, Cambridge.

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

Università del Piemonte Orientale, Via Perrone 18, 28100 Novara, Italy.

E-mail: caterina.may@unipmn.it

Università degli Studi di Milano, Via Conservatorio 7, 20122 Milano, Italy.

E-mail: chiara.tommasi@unimi.it

(Received August 2012; accepted November 2012)