Chemometrics and Intelligent Laboratory Systems 261 (2025) 105362

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Augmented designs to choose between constant absolute and relative errors and to estimate model parameters

Carlos de la Calle-Arroyo [a,d,*], Samantha Leorato [b], Licesio J. Rodríguez-Aragón [c], Chiara Tommasi [b]

[a] Departamento de Estadística e I.O., Universidad de Oviedo, Gijón, Asturias, Spain
[b] Department of Economics, Management and Quantitative Methods, University of Milan, Milan, Lombardia, Italy
[c] IMACI, EIIA, Universidad de Castilla-La Mancha, Toledo, Castilla-La Mancha, Spain
[d] DATAI, Universidad de Navarra, Pamplona, Navarra, Spain

ARTICLE INFO

MSC:
62K05

Keywords:
Optimal experimental designs
Error-variance specification
KL-optimality
Augmented designs
Multi-objective criteria

ABSTRACT

In experimental sciences such as chemistry, the measurement error may be homoscedastic or heteroscedastic. The data should be collected with the goal of identifying the right error-variance structure, as an incorrectly specified model would lead to wrong conclusions. A design criterion that reflects this goal is KL-optimality. Frequently, however, KL-optimum designs are wholly inefficient for other inferential purposes, such as precise estimation. In this case, the addition of some experimental points might be convenient. This work focuses on the enrichment of a design through the inclusion of some additional support points, with the goal of guaranteeing a minimum KL-efficiency to be able to optimally choose between different variance specifications. This strategy is also useful for modifying a design that is already available, for instance a D-optimal design, to manage the problem of correct error-variance specification.

1. Introduction

Very frequently, a phenomenon of interest y can be described by a regression model:

$$y = \eta(x, \boldsymbol{\beta}) + \varepsilon, \tag{1}$$

where $\eta(x, \boldsymbol{\beta})$ is the mean response at an experimental condition $x \in \mathcal{X}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_s)^T$ is a vector of unknown coefficients, and ε is a random error term that follows a Gaussian distribution with zero mean. In chemistry, as well as in several other experimental settings, ε may be either homoscedastic (with constant absolute error), or heteroscedastic (with constant relative error). This topic has been documented experimentally and within the design of experiments point of view. In vapour pressure data analysis, where temperature ranges can be of more than a hundred Celsius degrees, this heteroscedastic behaviour is sometimes observed [1]. Even though natural phenomena can show heteroscedastic behaviour, if this fact is not known beforehand, practitioners usually work with the normal homoscedastic assumption. From the design of experiments point of view wrong assumptions on the error structure make that designs that are optimal for one situation perform inefficiently when the other assumption is fulfilled [2, Section 4]. Designs of experiments has been applied for heteroscedastic behaviour, for example [3,4], but designing in order to select the adequate error assumption is an open question in the literature. To decide in favour of one of these two kinds of error, the following hypothesis test is adequate to be considered:

$$\begin{cases} \mathbf{H}_0 : \ \varepsilon \sim \mathcal{N}(0, \sigma_0^2) \\ \mathbf{H}_1 : \ \varepsilon \sim \mathcal{N}(0, \sigma_1^2\, \eta(x, \boldsymbol{\beta})^2) \end{cases} \tag{2}$$

In order to fix the experimental conditions optimally, with the goal of testing hypotheses (Eq. (2)), the KL-optimality criterion is applied; see for instance [5,6].

Experiments ξ are considered in this study to be design measures over the design space $\mathcal{X}$. The use of so-called continuous or approximate designs differs from exact designs in that there is no restriction on the number of support points, and their weights vary continuously on $[0, 1]$; see for instance [7,8].

As will be shown in Section 2, the KL-optimum design has only two support points. Whenever two support points are not enough to estimate the whole parameter vector (or to reach other inferential goals), it is necessary to enrich the design by adding new experimental points.

In addition, it is not unusual that an optimal design for a specific inferential purpose (such as precise estimation of the parameters) is already available. If it is necessary to test Eq. (2) then, even in this

* Corresponding author at: Departamento de Estadística e I.O., Universidad de Oviedo, Gijón, Asturias, Spain.
E-mail address: callecarlos@uniovi.es (C. de la Calle-Arroyo).

https://doi.org/10.1016/j.chemolab.2025.105362
Received 1 October 2024; Received in revised form 7 February 2025; Accepted 21 February 2025
Available online 7 March 2025
0169-7439/

case, the existing design should be augmented by adding new points to approach KL-optimality, as this helps to improve the performance of the hypothesis test.

These two situations justify our interest in adapting the methodology of augmenting designs (see for instance [9–11]):

(i) To enrich the KL-optimum design, with the goal of performing other analyses, such as to get a precise estimation of the model parameters ($c-$ or D_s- optimality), a function of them (L-optimality) or accurate predictions (I-optimality).
(ii) To improve an existing design, with the goal of testing the hypothesis in Eq. (2).

The rest of the paper is structured as follows. Section 2 gives the closed form expression for KL-optimality for the hypothesis test specified in Eq. (2) and the KL-optimal design. Section 3 introduces and supports the methodology of KL-augmented designs. Section 4 provides an example of the implementation of the methodology to Antoine's Equation, and compares the augmented designs with compound designs. Lastly, in Section 5 numerical results are given, and the pros and cons of the approaches presented here are discussed.

2. Constant errors: absolute versus relative

In order to design an experiment with the goal of testing whether the error is constant absolutely or relatively, the KL-criterion is used, and is here briefly recalled.

Let ξ be a design over an experimental space $\mathcal{X}$, and f_0, f_1 be two rival statistical models, considering f_1 as the true, completely known model, then the KL-optimality criterion function is

$$\mathcal{I}(\xi,\theta_1)=\min_{\theta_0}\int_{\mathcal{X}}\mathbb{E}_{f_1}\left[\log\frac{f_1(y;x,\theta_1)}{f_0(y;x,\theta_0)}\right]\,d\xi(x),\tag{3}$$

where θ_0 is the vector of unknown parameters of the distribution function f_0, θ_1 are the parameters of f_1 (assumed to be known) and $\mathbb{E}_{f_1}(\cdot)$ denotes the expected value with respect to f_1.

This study considers the heteroscedastic model as the true, completely known model, and the homoscedastic model as the competitor. The following theorem provides the expressions for the KL-optimum design $\xi^{\star}_{KL}=\arg\max_{\xi}\mathcal{I}(\xi,\theta_1)$ and for $\mathcal{I}(\xi^{\star}_{KL},\theta_1)$.

Theorem 1. *If $f_1(y;x,\theta_1)$ is the density corresponding to the regression model (1) with $\varepsilon\sim N(0;\sigma_1^2\eta(x;\beta_1))$, where $\theta_1=(\beta_1,\sigma_1^2)$ are known, and $f_0(y;x,\theta_0)$ denotes the density function of the same regression model with $\varepsilon\sim N(0;\sigma_0^2)$, where $\theta_0=(\beta_0,\sigma_0^2)$ is unknown, then*

$$\xi^{\star}_{KL}=\left\{\begin{array}{cc}\underline{x} & \overline{x}\\ \omega & 1-\omega\end{array}\right\},\quad \omega=\frac{\eta_M^2}{\eta_M^2-\eta_m^2}-\frac{1}{\log\eta_M^2-\log\eta_m^2}.\tag{4}$$

where $\eta_m^2:=\eta_m^2(\beta_1)=\inf_x\eta(x;\beta_1)^2>0$, $\eta_M^2:=\eta_M^2(\beta_1)=\sup_x\eta(x;\beta_1)^2<\infty$, and $\underline{x}\in\{x:\ \eta(x;\beta_1)^2=\eta_m^2\}$ and $\overline{x}\in\{x:\ \eta(x;\beta_1)^2=\eta_M^2\}$. In addition,

$$\mathcal{I}^{\star}(\beta_1)=\mathcal{I}(\xi^{\star}_{KL},\theta_1)=\frac{1}{2}(H-\log H-1),\tag{5}$$

where $H=\eta_m^2\cdot\dfrac{\log\eta_M^2-\log\eta_m^2}{\eta_M^2-\eta_m^2}$.

Proof. See Appendix A. □

To measure the "goodness" of a design ξ at performing the hypothesis test (Eq. (2)) the KL-efficiency is used:

$$\mathrm{eff}_{KL}(\xi,\theta_1)=\frac{\mathcal{I}(\xi,\theta_1)}{\mathcal{I}^*(\beta_1)}=\frac{\log A_\eta(\xi;\beta_1)-\log G_\eta(\xi;\beta_1)}{2\,\mathcal{I}^*(\beta_1)},\tag{6}$$

where the last expression has been obtained from Eqs. (A.6) and (5).

3. Augmenting-design methodology

Following similar ideas such as those in [9], this section describes a methodology for augmenting a design with the goals of:

(i) Improving the KL-optimum design (Eq. (4)) to be able to perform other analyses (controlling the KL-efficiency loss).
(ii) Enriching a design to provide good discrimination between the models with constant, either absolute or relative, errors (ensuring a minimum KL-efficiency).

A design ξ_k with k support points is augmented. To be general, ξ_k may be the KL-optimum design (Eq. (4)) or any other design at hand. The methodology consists in three steps:

Step 1: Let x_j, $j=1,\dots,p$ denote the p candidate points to be added, and α_j the proportion of replications to be taken at x_j. Their corresponding weights α_j with $j=1,\dots,p$ are constrained to be $\sum_{j=1}^p\alpha_j=\alpha$, where $\alpha\in(0,1)$ is the proportion of 'contamination' of ξ_k, and is fixed in advance by the researcher. The points $x_1,\dots,x_p$ must be chosen from a proper region, as described in Step 3.

Step 2: Let ξ_{x_i} denote the single-point design with the whole mass at a point x_i.
The goal is to produce an augmented design

$$\xi_{k+p}^{(x_1,\dots,x_p)}=(1-\alpha)\,\xi_k+\alpha_1\,\xi_{x_1}+\cdots+\alpha_p\,\xi_{x_p}$$

such that $\mathrm{eff}_{KL}\left(\xi_{k+p}^{(x_1,\dots,x_p)}\right)\geq\delta$, where δ is a desired threshold for the KL-efficiency.

Step 3: The next Theorem 2 ensures that this goal is achieved if the additional points are selected from the following design region:

$$\mathcal{R}(\xi_k;\alpha,\delta)=\{x:\ \xi_{k+1}^{(x)}=(1-\alpha)\xi_k+\alpha\xi_x;\ \mathrm{eff}_{KL}\big(\xi_{k+1}^{(x)}\big)\geq\delta\,\}.\tag{7}$$

In other terms, $\mathcal{R}(\xi_k;\alpha,\delta)$ is formed by the experimental conditions that guarantee at least a KL-efficiency equal to δ when they are added to ξ_k with a whole proportion of replicates equal to α.
Despite a natural preference for δ to be high, δ cannot exceed $\max_x\mathrm{eff}_{KL}\big(\xi_{k+1}^{(x)}\big)$, otherwise the region $\mathcal{R}(\xi_k;\alpha,\delta)$ would be empty.

Remark 1. $A_\eta(\xi_{k+1}^{(x)};\beta_1)=(1-\alpha)A_\eta(\xi_k;\beta_1)+\alpha\,\eta^2(x;\beta_1)$ and $G_\eta(\xi_{k+1}^{(x)};\beta_1)=G_\eta(\xi_k;\beta_1)^{(1-\alpha)}\cdot\eta^2(x;\beta_1)^{\alpha}$. Therefore, from Eq. (6), for any given ξ_k and any fixed α,

$$\begin{aligned}\mathrm{eff}_{KL}(\xi_{k+1}^{(x)},\beta_1)&=\frac{\log A_\eta(\xi_{k+1}^{(x)};\beta_1)-\log G_\eta(\xi_{k+1}^{(x)};\beta_1)}{2\,\mathcal{I}^*(\beta_1)}\\&=\frac{1}{2\,\mathcal{I}^*(\beta_1)}\Big\{\log\big[(1-\alpha)A_\eta(\xi_k;\beta_1)+\alpha\,\eta^2(x;\beta_1)\big]\\&\quad-\log G_\eta(\xi_k;\beta_1)^{(1-\alpha)}+\alpha\log\eta^2(x;\beta_1)\Big\}\end{aligned}\tag{8}$$

is a function of the design point x.

Theorem 2. *Let ξ_k be a design with k support points and $\xi_{x_1},\dots,\xi_{x_p}$ be p designs with all their mass concentrated at $x_j\in\mathcal{R}(\xi_k)$, for $j=1,\dots,p$. If $\xi_{k+p}^{(x_1,\dots,x_p)}=(1-\alpha)\,\xi_k+\sum_{j=1}^p\alpha_j\,\xi_{x_j}$, with $\alpha_j\in(0,1)$ and $\alpha=\sum_{j=1}^p\alpha_j$, then $\mathrm{eff}_{KL}\left(\xi_{k+p}^{(x_1,\dots,x_p)}\right)\geq\delta$.*

Proof. See Appendix A. □

Remark 2. Theorem 2 still holds if some of the points x_j, for $j=1,\dots,p$, are already included in the support of ξ_k (see for instance, the second case of Section 4).

Note that due to the concavity of the KL-criterion the following theorem holds:

Theorem 3. *Given a design ξ_k with k support points, a one-point design ξ_x and a weight $\alpha \in (0;1)$, if $\xi_{k+1}^{(x)} = (1-\alpha)\xi_k + \alpha\xi_x$, then $\mathit{eff}_{KL}(\xi_{k+1}^{(x)}) \geq (1-\alpha)\,\mathit{eff}_{KL}(\xi_k) + \alpha\,\mathit{eff}_{KL}(\xi_x)$.*

Using Inequality (A.8) and setting $\xi_k = \xi^*_{KL}$ in Theorem 3, we have $\mathrm{eff}_{KL}\left(\xi_{k+p}^{(x_1,\ldots,x_p)}\right) \geq \mathrm{eff}_{KL}\left(\xi_{k+1}^{(x)}\right) \geq 1-\alpha$.

Hence, when $\xi_k = \xi^*_{KL}$, δ should be chosen larger than $1-\alpha$ to have a non-trivial set $\mathcal{R}(\xi_k;\alpha,\delta)$.

4. An application to Antoine's equation

4.1. Antoine's equation

Antoine's Equation is considered as the regression model:

$$y = \eta(x,\boldsymbol{\beta}) + \varepsilon = 10^{a-\frac{b}{c+x}} + \varepsilon, \tag{9}$$

where $\boldsymbol{\beta} = (a,b,c)$ is a vector of unknown parameters. Eq. (9) relates the temperature x of a pure substance (or a mixture) to its vapour pressure y. The error ε can be either homoscedastic, i.e. $\varepsilon \sim \mathcal{N}(0,\sigma_0^2)$, or heteroscedastic with constant relative error, i.e. $\varepsilon \sim \mathcal{N}(0,\sigma_1^2\,\eta^2(x,\boldsymbol{\beta}))$. Consider the heteroscedastic model as the true model with design space $\mathcal{X} = [1,100]$, $\boldsymbol{\beta}_1 = (a_1,b_1,c_1) = (8.07131, 1730.63, 233.426)$. These parameter values correspond to water in the liquid state [12].

From Eq. (4) the design that maximises the KL-optimality criterion to discriminate between homoscedasticity and heteroscedasticity is:

$$\xi^\star_{KL} = \left\{ \begin{array}{cc} 1 & 100 \\ 0.90 & 0.10 \end{array} \right\}.$$

For this design the KL-criterion function is $\mathcal{I}^*(\boldsymbol{\beta}_1) = 3.391391$. This KL-optimum design only enables testing for the presence of heteroscedasticity; unfortunately, the parameter vector $\boldsymbol{\beta}_1$ of Antoine's equation cannot be estimated, as $\xi^\star_{KL}$ has only two support points. Thus the necessity to fix additional experimental conditions besides the support points of $\xi^\star_{KL}$.

In what follows, three different methods of enriching ξ^*_{KL} are proposed, with the goal of obtaining a new augmented design that provides an estimate of the parameters. To measure the goodness of a design ξ for precise estimation, D-efficiency is adopted:

$$\mathrm{eff}_D(\xi) = \left(\frac{|M(\xi)|}{|M(\xi^*_D)|}\right)^{1/3},$$

where

$$M(\xi) = \int_{\mathcal{X}} \frac{\partial\eta(x,\theta)}{\partial\beta}\frac{\partial\eta(x,\theta)}{\partial\beta^T}\,d\xi,$$

which is proportional to the Fisher information matrix for the Antoine model (Eq. (9)) and $\xi^*_D = \arg\max_\xi |M(\xi)|^{1/3}$ denotes the D-optimal design. The analytical expression for the D-optimum design for Antoine's Equation, both with homoscedastic and heteroscedastic variance, is known [2]. Henceforth, the focus is only on D-optimality for the heteroscedastic case, which is more common when the design region is broad [1], and the D-optimal design is:

$$\xi^*_D = \left\{ \begin{array}{ccc} x_{min} & x_2^\star & x_{max} \\ 1/3 & 1/3 & 1/3 \end{array} \right\}, \tag{10}$$

where

$$x_2^\star = \frac{c x_{max} + c x_{min} + 2 x_{max} x_{min}}{2c + x_{max} + x_{min}}.$$

Another reason for considering only the D-optimal design for the heteroscedastic model is that when the design region is narrow the D-optimal design for the homoscedastic model is similar to the D-optimal design for the heteroscedastic case [2, Eqs. (16) and (17)]. This happens when the optimal design is supported in both extremes of the design space, being the minimum of the design space greater than the smallest support point.

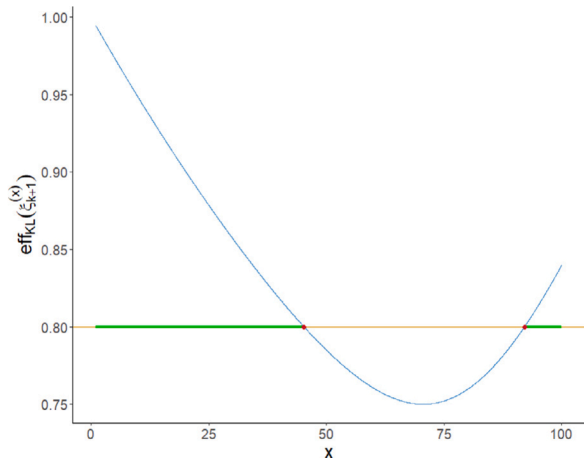

Fig. 1. KL-efficiency of $\xi_{k+1}^{(x)}$ as a function of x. The green part gives the region of candidate points $\mathcal{R}(\xi^*_{KL};\alpha=0.25,\delta=0.8)$.

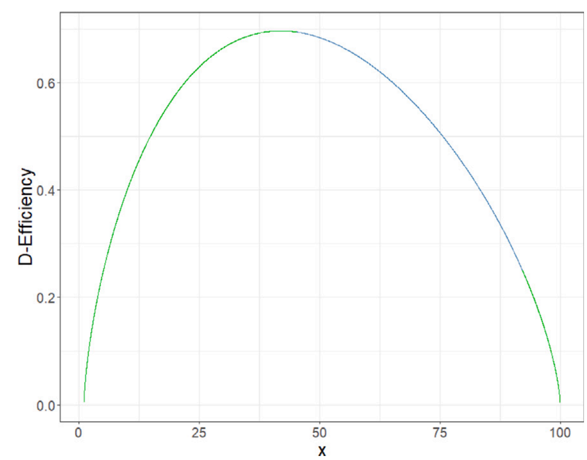

Fig. 2. D-efficiency of $\xi_{k+1}^{(x)} = 0.75\xi^*_{KL} + 0.25\xi_x$ as a function of x. The green part corresponds to $x \in \mathcal{R}(\xi^*_{KL};\alpha=0.25,\delta=0.8)$.

4.2. Augmenting the KL-optimum design

First approach

As a first approach to augmenting ξ^*_{KL}, it is suggested to add just one point, forming $\xi_{k+1}^{(x)} = (1-\alpha)\,\xi^\star_{KL} + \alpha\,\xi_x$. The KL-efficiency of $\xi_{k+1}^{(x)}$ as a function of x is given in Eq. (8) and displayed in Fig. 1 for $\alpha = 0.25$.

If a minimum efficiency is fixed of $\delta = 0.8$ and $\alpha = 0.25$, then the region from which the new experimental point should be selected is

$$\mathcal{R}(\xi^*_{KL};0.25,0.8) = [1,\ 45.23] \cup [92.12,\ 100],$$

which corresponds to the green part in Fig. 1. One possibility is to choose the point $x_D \in \mathcal{R}(\xi^*_{KL};0.25,0.8)$ that guarantees the largest D-efficiency for the augmented design, i.e. $\arg\max_{x\in\mathcal{R}(\xi^*_{KL};0.25,0.8)} \mathrm{eff}_D(\xi_{k+1}^{(x)})$. From Fig. 2, which shows $\mathrm{eff}_D(\xi_{k+1}^{(x)})$ as a function of x, it may be observed that the point to add is 41.88. The augmented design is then:

$$\xi_{k+1}^{(41.88)} = \left\{ \begin{array}{ccc} 1 & 41.88 & 100 \\ 0.68 & 0.25 & 0.07 \end{array} \right\},$$

with a D-efficiency of 69.71% and a KL-efficiency of 81.16%.

Second approach

Before describing another approach to enriching ξ^*_{KL}, define $\xi^{\mathcal{R}}_D$ to be the D-optimum design in the region $\mathcal{R}(\xi^*_{KL};\alpha=0.25,\delta=0.8)$. To compute $\xi^{\mathcal{R}}_D$ the algorithm provided by [13,14] has been slightly

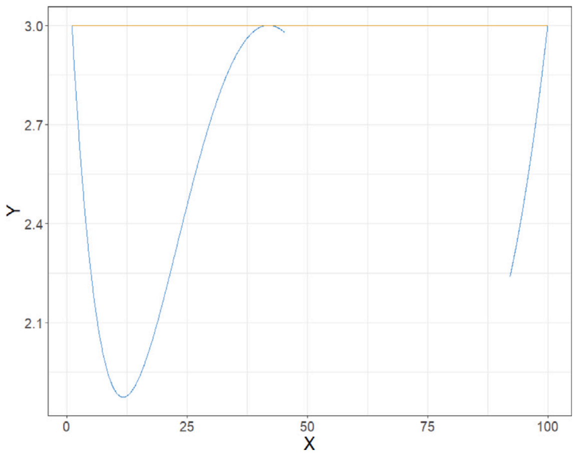

Fig. 3. Sensitivity function of the D-optimum design over $\mathcal{R}(\xi^*_{KL}; \alpha = 0.25, \delta = 0.8)$.

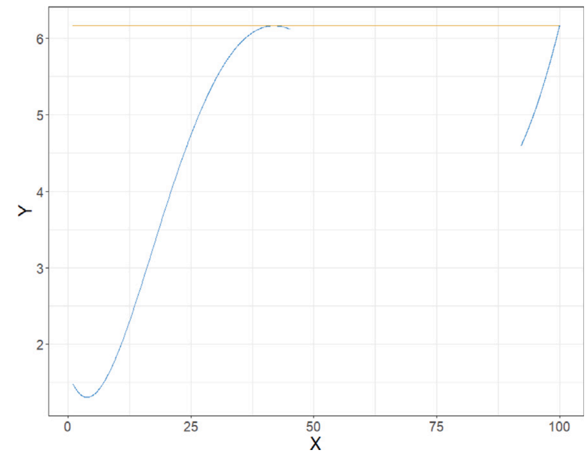

Fig. 4. Sensitivity function of the restricted D-optimum design over $\mathcal{R}(\xi^*_{KL}; \alpha = 0.25, \delta = 0.8)$.

modified to allow non-connex design spaces, obtaining:

$$\xi_D^{\mathcal{R}} = \left\{ \begin{array}{ccc} 1 & 41.88 & 100 \\ 1/3 & 1/3 & 1/3 \end{array} \right\}.$$

The optimality of $\xi_D^{\mathcal{R}}$ can be checked via the sensitivity function in Fig. 3.

The second approach consists in adding to ξ^*_{KL} the three support points of $\xi_D^{\mathcal{R}}$ (which are in $\mathcal{R}(\xi^*_{KL}; 0.25, 0.8)$ by construction), with weights $\alpha_1 = \alpha_2 = \alpha_3 = \alpha/3$, obtaining:

$$\xi_{k+3}^{(1,41.88,100)} = (1-\alpha)\xi^\star_{KL} + \alpha\xi_D^{\mathcal{R}} = \left\{ \begin{array}{ccc} 1 & 41.88 & 100 \\ 0.76 & 0.08 & 0.16 \end{array} \right\}.$$

While the D-optimum design has 3 points, since two of them were already support points of the KL-optimum design, only the weight was modified. Theorem 2 guarantees that $\xi_{k+3}^{(1,41.88,100)}$ has at least a KL-efficiency of 80%. In fact, $\mathrm{eff}_{KL}\left(\xi_{k+3}^{(1,41.88,100)}\right) = 91.98\%$ while $\mathrm{eff}_D\left(\xi_{k+3}^{(1,41.88,100)}\right) = 64.57\%$.

Third approach

Given a KL-optimum design ξ^*_{KL}, the last approach consists in finding the D-optimum design in the design region $\mathcal{R}(\xi^*_{KL}; \alpha = 0.25, \delta = 0.8)$, among all the designs with the structure $(1-\alpha)\xi^*_{KL} + \alpha\xi$:

$$\xi_{rD}^{\mathcal{R}} = \underset{\xi \text{ supported in } \mathcal{R}(\xi^*_{KL}; 0.25, 0.8)}{\arg\max} \left| M\left((1-\alpha)\xi^*_{KL} + \alpha\xi\right)\right|^{1/3}. \quad (11)$$

This problem is equivalent, in form, to the regularisation problem considered in [7, pp. 79–80]. By implementing a version of Wynn–Fedorov algorithm, $(1-\alpha)\xi^*_{KL} + \alpha\xi_{rD}^{\mathcal{R}}$ is obtained. The augmented design obtained is:

$$\xi_{k+2}^{(41.88,100)} = \left\{ \begin{array}{ccc} 1 & 41.88 & 100 \\ 0.68 & 0.16 & 0.16 \end{array} \right\},$$

with a D-efficiency of 78.28% and a KL-efficiency of 85.90%; its optimality can be checked from Fig. 4 which provides directional derivatives of the D-criterion. It is worth noting that this design has the best D-efficiency among the augmented designs for this choice of $\alpha = 0.25$ and $\delta = 0.80$.

4.3. Augmenting the D-optimum design

As another possible strategy, we might want to augment the D-optimum design in the direction of the KL-criterion. In this case the D-optimum design is $\xi^\star_D = \xi_D^{\mathcal{R}}$ in Eq. (10), and it has D-efficiency of 100% and KL-efficiency of 58.3%.

The procedure to augment $\xi^\star_D$ remains the same. To enrich the D-optimum design by just one additional point, then the augmented

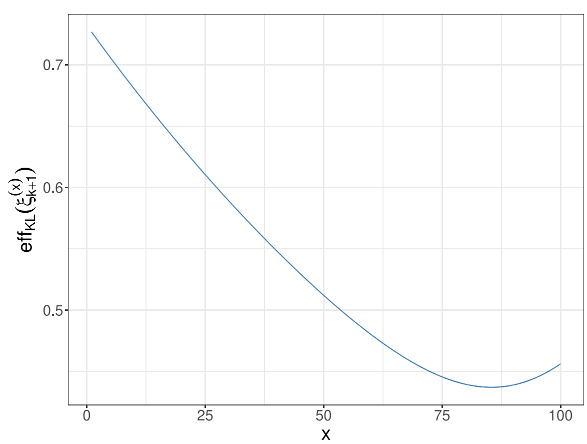

Fig. 5. KL-efficiency of the D-optimal design augmented for $\alpha = 0.25$.

design is $\xi_{k+1}^{(x)} = (1-\alpha)\xi^\star_D + \alpha\xi_x$. The KL-efficiency of $\xi_{k+1}^{(x)}$ is calculated with $\alpha = 0.25$, choosing δ and then suitably determining the point to be added to $\xi^\star_D$. Fig. 5 shows how the KL-efficiency of $\xi_{k+1}^{(x)}$ varies along $x \in \mathcal{X}$. It can immediately be observed that fixing $\delta = 0.8$, as in the previous example, would yield an empty region $\mathcal{R}$. The best KL-efficiency can be achieved by adding the point $x = 1$, obtaining the following enriched D-optimum design:

$$\xi_{k+1}^{(1)} = \left\{ \begin{array}{ccc} 1 & 41.88 & 100 \\ 1/2 & 1/4 & 1/4 \end{array} \right\},$$

which has a KL-efficiency of 72.6% and a D-efficiency of 94.5%.

In order to reach a KL-efficiency of 80%, it is necessary to increase α. By setting $\alpha = 0.4$, Fig. 6, the value of δ could be chosen up to 80.5%. The value of x that produces a KL-efficiency of 80% is $x = 1.63$, and the resulting augmented design would be:

$$\xi_{k+1}^{(1.6)} = \left\{ \begin{array}{cccc} 1 & 1.63 & 41.88 & 100 \\ 1/5 & 2/5 & 1/5 & 1/5 \end{array} \right\},$$

with $\mathrm{eff}_D\left(\xi_{k+1}^{(1.6)}\right) = 85.5\%$.

The reader should be aware that, when the augmented design is not the KL-optimum design, it is possible that no value α allows a certain desired minimum efficiency, δ. Finally, it is also worth noting that, from [9, Eq. (10)], which states that $(1-\alpha)\left(1 + \frac{\alpha \min[d(x,\xi)]}{1-\alpha}\right)^{1/m} \leq \delta$, it follows immediately that the minimum D-efficiency of the augmented design $\xi_{k+1}^{(x)} = (1-\alpha)\xi^\star_D + \alpha\xi_x$ is $1-\alpha$.

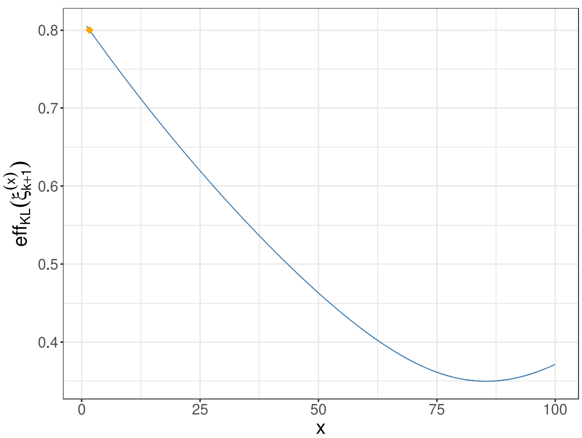

Fig. 6. KL-efficiency of the D-optimal design augmented for $\alpha = 0.40$.

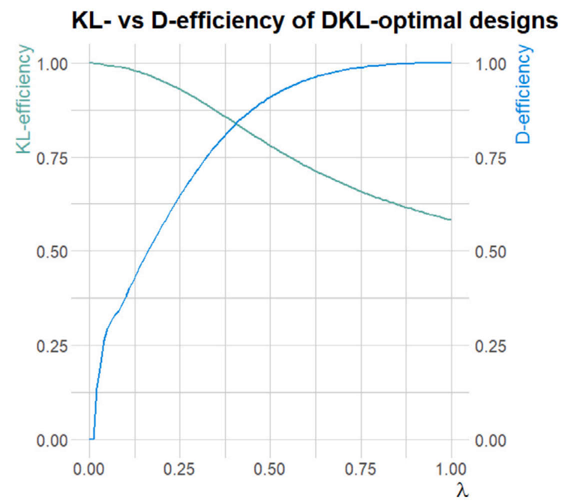

Fig. 7. D- and KL-efficiencies of the DKL-optimal designs, obtained through an evolutionary algorithm.

4.4. Compound criterion

All the previous approaches to augmenting the KL-optimum design are aimed at addressing the problem of obtaining a precise estimate of the parameters, besides performing the hypothesis test, Eq. (2), while augmenting the D-optimum design aims to improve the performance of the hypothesis test. A different and natural strategy is the compound DKL-criterion,

$$\Phi_{DKL}(\xi) = (1-\lambda)\mathrm{eff}_{KL}(\xi) + \lambda \mathrm{eff}_{D}(\xi),$$

with $\lambda \in [0,1]$. For more details, see for instance, [15,16].

As shown by [17], a compound criterion provides maximal efficiency for one criterion, given a certain level of the other (and vice versa).

Since the directional derivative for KL-optimality is constant, this study uses a genetic algorithm to find a DKL-optimal design, $\xi^{\lambda} = \arg\max_{\xi} \Phi_{DKL}(\xi)$, which depends on the weight λ. In particular, an evolutionary algorithm is applied to calculate the solutions ξ^{λ} for different values of λ. The memetic algorithm implemented is similar to the one used in [18], and its pseudocode is shown in Algorithm 1. To see the details of the algorithm, see Appendix B.

Algorithm 1 Multiobjective memetic algorithm

Generate the initial population of size *popSize*.
while Number of iterations < *MaxIterations* **do**
 Perform tournament selection to find the reproducing chromosomes.
 Group those chromosomes randomly into pairs.
 Apply the crossover operator to each pair with probability *crossover_probability*.
 Apply the mutation operator to each chromosome with probability *mutation_probability*.
 Apply heuristics.
 Apply *LSiterations* iterations of local search to every offspring generated.
 Choose chromosomes for the next generation with the replacement strategy.
end while
return The design from the best chromosome.

As the result of this algorithm, Fig. 7 shows the D- and the KL-efficiencies of the designs for the different values of λ, which can help the researcher in choosing the best option for their goal.

A frequently chosen DKL-optimum design is the one with equal (or the closest) D- and KL-efficiencies. This design is:

$$\xi^{0.40}_{DKL} = \left\{ \begin{array}{ccc} 1 & 32.77 & 100 \\ 0.61 & 0.21 & 0.18 \end{array} \right\},$$

with a D-efficiency of 83.4% and KL-efficiency of 84.1%.

For this example, the design with the KL-efficiency closest to $\delta = 80\%$ is also worthy of attention; this is:

$$\xi^{0.46}_{DKL} = \left\{ \begin{array}{ccc} 1 & 34.07 & 100 \\ 0.56 & 0.23 & 0.21 \end{array} \right\},$$

with a KL-efficiency of 80.4% and a D-efficiency of 88.3%.

4.5. Proof-of-concept example with simulated data

The proposed design strategy can be applied at the first step of the following research procedure:

1. Apply an augmented design;
2. Collect the data;
3. Select a model by testing

$$\begin{cases} \mathbf{H}_0 : \ \varepsilon \sim \mathcal{N}(0, \sigma_0^2) \\ \mathbf{H}_1 : \ \varepsilon \sim \mathcal{N}(0, \sigma_1^2\, \eta(x, \boldsymbol{\beta})^2) \end{cases} \tag{12}$$

4. Compute the maximum likelihood estimator (MLE) of the parameters of the chosen model.

To show how this methodology applies in practice, we generate $n = 75$ responses $y_1, \dots, y_n$ from Antoine's model (9) with $a = 8.07131, b = 1730.63$ and $c = 233.426$, at the following 4 different scenarios:

- Homoscedastic model with $\sigma_0 = 2$ and $\xi^{(1,41.88,100)}_{k+3}$
- Heteroscedastic model with $\sigma_1 = 0.1$ and $\xi^{(1,41.88,100)}_{k+3}$
- Homoscedastic model with $\sigma_0 = 2$ and $\xi^{(1.6)}_{k+1}$
- Heteroscedastic model with $\sigma_1 = 0.1$ and $\xi^{(1.6)}_{k+1}$

Since the two rival models in (12) are non-nested, to perform the test of hypothesis we apply the log-likelihood ratio statistic $R = \log \mathcal{L}_0(\hat{\theta}_0) - \log \mathcal{L}_1(\hat{\theta}_1)$ [19], where $\hat{\theta}_j$ and $\mathcal{L}_j(\hat{\theta}_j)$ denote the MLE and the maximum of the log-likelihood function for Antoine's model (9) under H_j with $j = 0; 1$, respectively. Unfortunately, the distribution of the test-statistic R under the null hypothesis (homoscedastic Antoine's model) is unknown. Therefore, we approximate the distribution of R (under H_0) by using

Table 1
Bootstrap p-values, maximum likelihood estimates of the chosen model and their standard errors.

Gener. model	Design	p-value	$\hat{\beta}$	sd
Homoscedastic	$\xi_{k+3}^{(1,41.88,100)}$	0.82	8.038 1715.041 232.604	0.42 252.77 21.72
	$\xi_{k+1}^{(1.6)}$	0.81	8.065 1730.571 233.809	0.41 248.03 21.48
Heteroscedastic	$\xi_{k+3}^{(1,41.88,100)}$	0	8.076 1730.624 233.533	0.60 339.03 26.77
	$\xi_{k+1}^{(1.6)}$	0	7.967 1683.078 230.422	0.49 268.89 21.49

Table 2
Efficiencies for the different designs.

	Approach	Design	$\mathrm{Eff}_{KL}(\cdot)$	$\mathrm{eff}_D(\cdot)$	Avg
KL-augmented ($\alpha = 0.25$)	I	$\xi_{k+1}^{(41.88)}$	81.2	69.7	75.45
	II	$\xi_{k+3}^{(1,41.88,100)}$	92.0	64.6	78.30
	III	$\xi_{k+2}^{(41.88,100)}$	85.9	78.3	82.10
KL-augmented ($\alpha = 0.3$)	I	$\xi_{k+1}^{(35.67)}$	80	69.9	74.95
	II	$\xi_{k+3}^{(1,35.67,100)}$	89.6	58.4	74.00
	III	$\xi_{k+1}^{(35.67)}$	80	69.9	74.95
D-augmented	$\alpha = 0.25$	$\xi_{k+1}^{(1)}$	72.6	94.5	83.55
	$\alpha = 0.4$	$\xi_{k+1}^{(1.6)}$	80	85.5	82.75
Compound designs	$\lambda = 0.40$	$\xi_{DKL}^{0.40}$	84.1	83.4	83.75
	$\lambda = 0.46$	$\xi_{DKL}^{0.46}$	80.4	88.3	84.35
Optimal designs	KL-	$\xi_{KL}^{\star}$	100	0	50
	D-	$\xi_{D}^{\star}$	58.3	100	79.15

the parametric Bootstrap technique. In more detail, in each scenario, given the sample of observations $y_1,\dots,y_n$, we generate $B = 5000$ Bootstrap samples of size n from $\mathcal{L}_0(\hat{\theta}_0)$ and we compute B Bootstrap values of R, denoted by R_b, with $b = 1,\dots,B$. The null hypothesis is rejected for low values of R, therefore the Bootstrap p-value is the proportion of times that $R_b < r$, where $r = R(y_1,\dots,y_n)$ is the "observed value" of R. After having chosen a model, we may use the same sample of data to compute the MLE of the parameters and the corresponding standard deviation. Herein, we have applied the R function mle. All the results are displayed in Table 1.

A detailed simulation study is presented in Appendix C, to show the usefulness of KL-augmented designs in terms of both model discrimination and parameter estimation.

5. Discussion

The discrimination between constant absolute error and constant relative error is, as shown throughout this work, an important matter to clarify. The KL-criterion can be applied to tackle this problem. The explicit expression for this criterion and for KL-optimal designs has been provided. Having solved the minimisation involved in the KL-criterion analytically greatly simplifies the computation of KL-efficiencies and compound designs for this discrimination issue.

Unfortunately, KL-optimum designs do not allow for parameter estimation whenever they have fewer support points than the number of parameters. To address this inconvenience, a methodology is proposed to 'enrich' the KL-optimum design by augmenting its support points. The example presented in Section 4 gives the reader an idea of how efficient the augmented designs can be, and it also highlights the drawbacks or particularities of the different choices necessary to produce this enrichment. One particular choice, key to the procedure, is the selection of α and δ. This choice highly depends on the interests of the experimenter, and no close answer can be given. However, empirical tests have suggested that a moderate value such as $\alpha \in [0.20, 0.45]$, and an informed subsequent minimum efficiency δ, depending on the experimentation goals, should in most cases yield adequate results for the practitioner.

Table 2 and Fig. 8 show the efficiencies (D-efficiency, KL-efficiency and the average of both) of the different augmented designs proposed in Section 4 (where $\alpha = 0.25$), as well as the designs obtained with another choice of $\alpha = 0.3$, always with fixed $\delta = 0.8$. When ξ_k is the KL-optimum design, for a fixed δ it can be observed that $\mathcal{R}$ grows as α decreases. A choice of $\alpha = 0.25$ instead of $\alpha = 0.3$ implies an extra flexibility that comes from a bigger region $\mathcal{R}$. By comparing $\xi_{k+3}^{(1,41.88,100)}$ vs $\xi_{k+3}^{(1,35.67,100)}$ and $\xi_{k+2}^{(41.88,100)}$ vs $\xi_{k+1}^{(35.67)}$ (the approach is the same for $\alpha = 0.25$ and $\alpha = 0.3$, respectively), it is immediately observed that this extra flexibility leads to augmented designs that are better with respect to KL-optimality (the KL-optimum design being less 'soiled'), and can also be better with respect to D-efficiency, as in this case.

The compound design cannot be dominated by definition, and here it is computed as a benchmark. The designs $\xi_{k+2}^{(41.88,100)}$ with $\alpha = 0.25$ and $\xi_{k+1}^{(1.6)}$ are the best choices among the augmented designs. The table shows that these augmented designs are able to reach competitive efficiencies with respect to the compound designs, even if by definition they perform worse than the optimal solution (see for instance that in the example the KL-augmented design with approach I is dominated by both compound designs). However, the enriched designs have a lower computational cost than compound optimal designs and their algorithmic implementation is easier. Among the augmented solutions, the restricted optimal design takes the longest time to calculate, at 8 s. Meanwhile, the memetic algorithm takes 253 s to calculate a compound design for a certain value of λ, with a population of 100 designs, 20 local search iterations, and 100 maximum iterations.

Finally, another advantage of this methodology is that by defining a whole region $\mathcal{R}$ of candidate points, it allows the experimenter to include other constraints in the selection of number and location of the new support points for the augmented design. For example, from the practitioner's point of view, sometimes replications are preferred to additional experimental setups, and at other times, certain observations have different experimental costs, which leads the practitioner to avoid/prefer certain ranges of values in the design space. The proposed region $\mathcal{R}$ of candidate points versus a given compound or restricted design, allows the practitioner to include these types of preferences in the augmented design.

CRediT authorship contribution statement

Carlos de la Calle-Arroyo: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Samantha Leorato:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis. **Licesio J. Rodríguez-Aragón:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Funding acquisition, Conceptualization. **Chiara Tommasi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

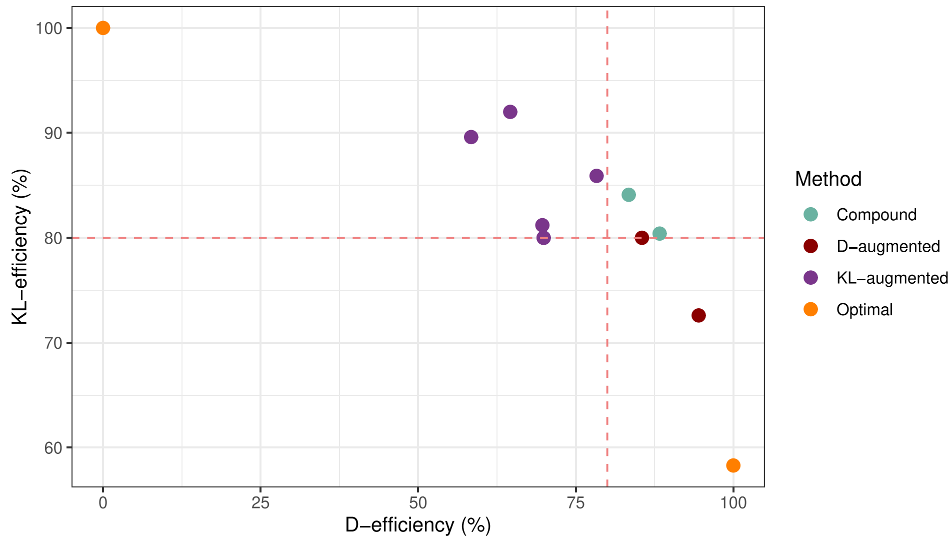

Fig. 8. Efficiencies of the different augmented and compound designs.

Acknowledgements

This work was supported by Ministerio de Ciencia e Innovación, Spain [grant number PID2020-113443RB-C21], by Junta de Comunidades de Castilla-La Mancha through Fondo Europeo de Desarrollo Regional, Spain [grant number SBPLY/21/180501/000126] and by Universidad de Castilla-La Mancha, Spain, Plan Propio co-funded by FEDER [grant number 2022-GRIN-34288]. The work was developed while de la Calle-Arroyo was visitig Leorato and Tommasi supported by Ministerio de Universidades, Spain, ayudas de movilidad José Castillejo [grant number CAS22/00161]. The work of Tommasi and Leorato was funded by the European Union - NextGenerationEU through the Italian Ministry of University and Research under the National Recovery and Resilience Plan (PNRR) - Mission 4 Education and research - Component 2 From research to business - Investment 1.1 Notice Prin2022 - DD N. 104 del 2/2/2022, title [Optimal and adaptive designs for modern medical experimentation], proposal code [2022TRB44L] - CUP [G53D23001600006].

Appendix A. Proof of theorems

Here are presented the proof of the main theoretical results:

Theorem 1. *If $f_1(y;x,\boldsymbol{\theta}_1)$ is the density corresponding to the regression model (1) with $\varepsilon \sim N(0;\sigma_1^2\eta(x;\boldsymbol{\beta}_1))$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1,\sigma_1^2)$ are known, and $f_0(y;x,\boldsymbol{\theta}_0)$ denotes the density function of the same regression model with $\varepsilon \sim N(0;\sigma_0^2)$, where $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0,\sigma_0^2)$ is unknown, then*

$$\xi_{KL}^\star = \left\{ \begin{array}{cc} \underline{x} & \overline{x} \\ \omega & 1-\omega \end{array} \right\}, \quad \omega = \frac{\eta_M^2}{\eta_M^2-\eta_m^2} - \frac{1}{\log\eta_M^2 - \log\eta_m^2}.$$

where $\eta_m^2 := \eta_m^2(\boldsymbol{\beta}_1) = \inf_x \eta(x;\boldsymbol{\beta}_1)^2 > 0$, $\eta_M^2 := \eta_M^2(\boldsymbol{\beta}_1) = \sup_x \eta(x;\boldsymbol{\beta}_1)^2 < \infty$, and $\underline{x} \in \{x : \eta(x;\boldsymbol{\beta}_1)^2 = \eta_m^2\}$ and $\overline{x} \in \{x : \eta(x;\boldsymbol{\beta}_1)^2 = \eta_M^2\}$. In addition,

$$\mathcal{I}^\star(\boldsymbol{\beta}_1) = \mathcal{I}(\xi_{KL}^\star,\boldsymbol{\theta}_1) = \frac{1}{2}(H - \log H - 1),$$

where $H = \eta_m^2 \cdot \dfrac{\log\eta_M^2 - \log\eta_m^2}{\eta_M^2-\eta_m^2}$.

Proof. Let $\mathbb{V}_{f_1}(\cdot)$ denote the variance with respect to f_1.

Since $\mathbb{E}_{f_1}[y] = \eta(x,\beta_1)$ and $\mathbb{E}_{f_1}\left[y^2\right] = \mathbb{V}_{f_1}[y] + \mathbb{E}_{f_1}[y]^2 = \sigma_1^2\eta(x,\beta_1)^2 + \eta(x,\beta_1)^2 = (1+\sigma_1^2)\eta(x,\beta_1)^2$, the KL-divergence between $f_1(y;x,\boldsymbol{\theta}_1)$ and $f_0(y;x,\boldsymbol{\theta}_0)$ is:

$$\begin{aligned}
\mathbb{E}_{f_1}\left[\log\frac{f_1(y;x,\theta_1)}{f_0(y;x,\theta_0)}\right] &= \mathbb{E}_{f_1}\left[\log\left(\frac{\sigma_0\sqrt{2\pi}}{\sigma_1\eta(x,\beta_1)\sqrt{2\pi}}\,\frac{\exp-\frac{1}{2}\left(\frac{y-\eta(x,\beta_1)}{\sigma_1\eta(x,\beta_1)}\right)^2}{\exp-\frac{1}{2}\left(\frac{y-\eta(x,\beta_0)}{\sigma_0}\right)^2}\right)\right]\\
&= \mathbb{E}_{f_1}\left[\log\sigma_0 - \log\sigma_1\eta(x,\beta_1) - \frac{1}{2}\left(\frac{y^2}{\sigma_1^2\eta(x,\beta_1)^2} - \frac{2y}{\sigma_1^2\eta(x,\beta_1)} + \frac{1}{\sigma_1^2} - \frac{y^2}{\sigma_0^2} + \frac{2y\eta(x,\beta_0)}{\sigma_0^2} - \frac{\eta(x,\beta_0)^2}{\sigma_0^2}\right)\right]\\
&= \log\sigma_0 - \log\sigma_1\eta(x,\beta_1) - \frac{1}{2}\left(\frac{(1+\sigma_1^2)\eta(x,\beta_1)^2}{\sigma_1^2\eta(x,\beta_1)^2} - \frac{2\eta(x,\beta_1)}{\sigma_1^2\eta(x,\beta_1)} + \frac{1}{\sigma_1^2} - \frac{(1+\sigma_1^2)}{\sigma_0^2}\eta(x,\beta_1)^2 + \frac{2\eta(x,\beta_1)\eta(x,\beta_0)}{\sigma_0^2} - \frac{\eta(x,\beta_0)^2}{\sigma_0^2}\right)\\
&= \log\sigma_0 - \log\sigma_1\eta(x,\beta_1) - \frac{1}{2} + \frac{(1+\sigma_1^2)}{2\sigma_0^2}\eta(x,\beta_1)^2 - \frac{\eta(x,\beta_1)\eta(x,\beta_0)}{\sigma_0^2} + \frac{\eta(x,\beta_0)^2}{2\sigma_0^2}.
\end{aligned}$$

Hence, the KL-optimality function (3) becomes:

$$\mathcal{I}(\xi,\theta_1) = -\frac{1}{2}\int_{\mathcal{X}}\log\sigma_1^2\eta(x,\beta_1)^2\,d\xi - \frac{1}{2} + \frac{1}{2}\min_{\theta_0} K(\boldsymbol{\beta}_0;\sigma_0^2), \tag{A.1}$$

where

$$\begin{aligned}
K(\boldsymbol{\beta}_0;\sigma_0^2) &= \log\sigma_0^2 + \int_{\mathcal{X}}\frac{(1+\sigma_1^2)}{\sigma_0^2}\eta(x,\beta_1)^2\,d\xi - \int_{\mathcal{X}}\frac{2\eta(x,\beta_1)\eta(x,\beta_0)}{\sigma_0^2}\,d\xi + \int_{\mathcal{X}}\frac{\eta(x,\beta_0)^2}{\sigma_0^2}\,d\xi\\
&= \log\sigma_0^2 + \frac{\sigma_1^2}{\sigma_0^2}\int_{\mathcal{X}}\eta(x,\beta_1)^2\,d\xi + \frac{1}{\sigma_0^2}\int_{\mathcal{X}}\Big(\eta(x,\beta_1) - \eta(x,\beta_0)\Big)^2\,d\xi.
\end{aligned} \tag{A.2}$$

Differentiating Eq. (A.2) with respect to σ_0^2 and making it equal to zero gives:

$$0 = \frac{\partial K}{\partial\sigma_0^2} = \frac{1}{\sigma_0^2} - \frac{(1+\sigma_1^2)\int_{\mathcal{X}}\eta(x,\beta_1)^2\,d\xi}{\sigma_0^4} + \frac{2\int_{\mathcal{X}}\eta(x,\beta_1)\eta(x,\beta_0)\,d\xi}{\sigma_0^4} - \frac{\int_{\mathcal{X}}\eta(x,\beta_0)^2\,d\xi}{\sigma_0^4}. \tag{A.3}$$

Multiplying the above equation by σ_0^4 and solving for σ_0^2 gives:

$$\hat{\sigma}_0^2 = (1+\sigma_1^2)\int_{\mathcal{X}}\eta(x,\beta_1)^2\,d\xi - 2\int_{\mathcal{X}}\eta(x,\beta_1)\eta(x,\beta_0)\,d\xi + \int_{\mathcal{X}}\eta(x,\beta_0)^2\,d\xi. \tag{A.4}$$

Plugging in the value of $\hat{\sigma}_0^2$ in Eq. (A.2) gives:

$$K(\beta_0, \hat{\sigma}_0^2) = \log\left[\int_{\mathcal{X}} (\eta(x,\beta_1) - \eta(x,\beta_0))^2\, d\xi + \sigma_1^2 \int_{\mathcal{X}} \eta(x,\beta_1)^2\, d\xi\right] + 1. \quad (A.5)$$

The value of β_0 that minimises Eq. (A.5) is $\hat{\beta}_0 = \beta_1$, and thus the KL-criterion (Eq. (A.1)), becomes:

$$\begin{aligned}
\mathcal{I}(\xi,\theta_1) &= -\frac{1}{2}\int_{\mathcal{X}} \log \sigma_1^2 \eta(x,\beta_1)^2\, d\xi - \frac{1}{2} + \frac{1}{2}\left(\log\left[\sigma_1^2 \int_{\mathcal{X}} \eta(x,\beta_1)^2\, d\xi\right] + 1\right)\\
&= \frac{1}{2}\left(\log \int_{\mathcal{X}} \eta(x,\beta_1)^2\, d\xi - \int_{\mathcal{X}} \log \eta(x,\beta_1)^2\, d\xi\right)\\
&= \frac{1}{2}\left(\log A_\eta(\xi;\beta_1) - \log G_\eta(\xi;\beta_1)\right), \qquad (A.6)
\end{aligned}$$

where $A_\eta(\xi;\beta_1) = \sum_i \eta(x_i,\beta_1)^2\, \xi(x_i)$ and $G_\eta(\xi;\beta_1) = \prod_i \eta(x_i,\beta_1)^{2\,\xi(x_i)}$ are the arithmetic and geometric means of $\eta(x,\beta_1)^2$ with respect to the design ξ on $\mathcal{X}$. Note that this KL-criterion (Eq. (A.6)) depends only on β_1, and therefore $\mathcal{I}(\xi,\theta_1) := \mathcal{I}(\xi,\beta_1)$. Since Equality (A.6) is the semi-difference between the logarithms of an arithmetic and a geometric mean, the explicit expression of the KL-optimum design given in Eq. (4) is obtained as in [6, Theorem 3]. Finally, the expression for $\mathcal{I}^\star(\beta_1)$ comes from [20, Lemma 1, Eq. (7)]. □

Theorem 2. *Let ξ_k be a design with k support points and $\xi_{x_1},\dots,\xi_{x_p}$ be p designs with all their mass concentrated at $x_j \in \mathcal{R}(\xi_k)$, for $j = 1,\dots,p$. If $\xi_{k+p}^{(x_1,\dots,x_p)} = (1-\alpha)\,\xi_k + \sum_{j=1}^{p} \alpha_j\, \xi_{x_j}$, with $\alpha_j \in (0,1)$ and $\alpha = \sum_{j=1}^{p} \alpha_j$, then $\mathit{eff}_{KL}\left(\xi_{k+p}^{(x_1,\dots,x_p)}\right) \geq \delta$.*

Proof. Note that for $p \geq 2$,

$$\begin{aligned}
\xi_{k+p}^{(x_1,\dots,x_p)} &= (1-\alpha)\,\frac{\sum_{j=1}^{p-1}\alpha_j + \alpha_p}{\alpha}\,\xi_k + \sum_{j=1}^{p-1}\alpha_j\,\xi_{x_j} + \alpha_p\,\xi_{x_p}\\
&= (1-\alpha)\frac{\sum_{j=1}^{p-1}\alpha_j\,\xi_k}{\alpha} + (1-\alpha)\frac{\alpha_p}{\alpha}\xi_k + (1-\epsilon_p)\frac{\sum_{j=1}^{p-1}\alpha_j\,\xi_{x_j}}{1-\epsilon_p} + \alpha\frac{\alpha_p}{\alpha}\,\xi_{x_p}\\
&= (1-\alpha)(1-\epsilon_p)\,\xi_k + (1-\epsilon_p)\sum_{j=1}^{p-1}\alpha'_j\,\xi_{x_j} + \epsilon_p\left[(1-\alpha)\,\xi_k + \alpha\,\xi_{x_p}\right]\\
&= (1-\epsilon_p)\left[(1-\alpha)\xi_k + \sum_{j=1}^{p-1}\alpha'_j\,\xi_{x_j}\right] + \epsilon_p\left[(1-\alpha)\xi_k + \alpha\,\xi_{x_p}\right]\\
&= (1-\epsilon_p)\,\xi_{x+p-1}^{(x_1,\dots,x_{p-1})} + \epsilon_p\,\xi_{k+1}^{(x_p)},
\end{aligned} \qquad (A.7)$$

where $\epsilon_p = \frac{\alpha_p}{\alpha}$ and $\xi_{k+p-1}^{(x_1,\dots,x_{p-1})} = (1-\alpha)\,\xi_k + \sum_{j=1}^{p-1}\alpha'_j\,\xi_{x_j}$, with $\alpha'_j = \frac{\alpha_j}{1-\epsilon_p}$ for $j = 1,\dots,p-1$ (as a consequence $\sum_{j=1}^{p-1}\alpha'_j = \alpha$).

From the concavity of the KL-criterion, by using the steps above recursively and setting $\epsilon_j = \frac{\alpha_j}{\alpha}\prod_{l=j+1}^{p}(1-\epsilon_l)$ for $j = 1,\dots,p-1$, we have that:

$$\begin{aligned}
\mathcal{I}\left(\xi_{k+p}^{(x_1,\dots,x_p)}\right) &\geq (1-\epsilon_p)\,\mathcal{I}\left(\xi_{k+p-1}^{(x_1,\dots,x_{p-1})}\right) + \epsilon_p\,\mathcal{I}(\xi_{k+1}^{(x_p)})\\
&\geq (1-\epsilon_p)(1-\epsilon_{p-1})\,\mathcal{I}\left(\xi_{k+p-2}^{(x_1,\dots,x_{p-2})}\right) + (1-\epsilon_p)\epsilon_{p-1}\,\mathcal{I}(\xi_{k+1}^{(x_{p-1})}) + \epsilon_p\,\mathcal{I}(\xi_{k+1}^{(x_p)})\\
&\geq (1-\epsilon_p)(1-\epsilon_{p-1})(1-\epsilon_{p-2})\,\mathcal{I}\left(\xi_{k+p-3}^{(x_1,\dots,x_{p-3})}\right)\\
&\quad + (1-\epsilon_p)(1-\epsilon_{p-1})\epsilon_{p-2}\,\mathcal{I}(\xi_{k+1}^{(x_{p-2})}) + (1-\epsilon_p)\epsilon_{p-1}\,\mathcal{I}(\xi_{k+1}^{(x_{p-1})}) + \epsilon_p\,\mathcal{I}(\xi_{k+1}^{(x_p)})\\
&\geq \cdots\\
&\geq \left[\epsilon_p + (1-\epsilon_p)\epsilon_{p-1} + (1-\epsilon_p)(1-\epsilon_{p-1})\epsilon_{p-2} + \cdots + (1-\epsilon_p)\dots(1-\epsilon_1)\right]\\
&\quad \times \min_{1\leq j\leq p} \mathcal{I}\left(\xi_{k+1}^{(x_j)}\right)\\
&= \min_{1\leq j\leq p} \mathcal{I}\left(\xi_{k+1}^{(x_j)}\right),
\end{aligned} \qquad (A.8)$$

where the last equality is due to the fact that

$$\epsilon_p + (1-\epsilon_p)\epsilon_{p-1} + (1-\epsilon_p)(1-\epsilon_{p-1})\epsilon_{p-2} + \cdots + (1-\epsilon_p)\dots(1-\epsilon_1) = 1.$$

From Inequality (A.8) by dividing by $\mathcal{I}(\xi_{KL}^\star,\theta_1)$, and taking into account that $x_1,\dots,x_p \in \mathcal{R}(\xi_k)$, the result is obtained. □

Appendix B. Memetic algorithm

A population of $popSize$ non-singular designs, or chromosomes, is randomly generated (otherwise the D-efficiency would be zero). Then, an iteration of the algorithm is performed. The first step is the tournament selection (pool size is $popSize$ and tournament size is 2). The tournament consists in randomly selecting (with replacement) $popSize$ pairs of designs from the population. From each pair, the best design is kept, according to the compound criterion.

Then, the crossover operator is applied to the current set of $popSize$ designs. This crossover operator consists in drawing (without replacement) $popSize/2$ pairs of designs, and then for each pair, with probability $crossover\_probability$, swapping a random number (from 1 up to the minimum number of points of the two parents minus 1) of support points between the two chromosomes. If a pair is not crossed, then it is discarded. The resulting designs form the set of offspring.

At this point the original population and the set of offspring are merged, forming the candidate population.

Then, with probability $mutation\_probability$, the mutation over each chromosome is applied in the candidate population. The mutation is performed over either a point or a weight of the design. The BGA mutation operator is used, and it works as follows: if $c_i \in [a_i, b_i]$ is a value to be mutated, the resulting value c'_i is,

$$c'_i = c_i \pm rang_i \times \sum_{k=0}^{15} \alpha_k 2^{-k},$$

where $rang_i$ defines the mutation range, and is set to 0.1 in the case of weights, and $0.1 \times (\max \mathcal{X} - \min \mathcal{X})$. The + or − sign is chosen with a probability of 1/2 and $\alpha_i \in \{0,1\}$ is randomly generated with probability $P(\alpha_i = 1) = 1/16$.

After the mutation step, $LSiterations$ of "the local search" are performed over each design of the candidate population. A single iteration of the local search consists in applying the mutation operator to a design ξ_c, and keeping the resulting updated design ξ_u, if ξ_u has a better criterion value than ξ_c.

Then, the typical heuristics, as used in classic algorithms in optimal experimental design, are applied: points with very small weights are eliminated and points that are very close to one another are merged together.

Lastly, the best $popSize$ chromosomes in the candidate population are selected, where best is defined in the same sense as in the tournament selection. These $popSize$ designs now form the initial population for the next iteration of the memetic algorithm.

Appendix C. Simulation

To show the usefulness of KL-augmented designs in terms of model discrimination and precise parameter estimation, we have developed a detailed simulation study. In this study, the heteroscedastic Antoine's Eq. (9) is considered as the true model within the range of temperatures $\mathcal{X} = (30, 60)$ and the initial best guesses for the parameters $\beta = (a,b,c) = (3.022, 535.051, 229)$, which correspond to iodoethane in liquid state [21]. We have chosen this example as the design of space is not a large range of temperatures and differences between homoscedastic and heteroscedastic assumptions are not easily detected.

The designs considered for this study were the uniform design, with 5 equally weighted support points, the D-optimal design and two augmented designs, analogous to $\xi_{k+3}^{(1,41.88,100)}$ and $\xi_{k+1}^{(1.6)}$ respectively (the latter with α=0.25 since, in this case, increasing α does not provide a better KL-efficiency). For this substance and temperature range, these designs are characterised as follows.

$$\xi_{Unif} = \left\{\begin{array}{ccccc} 30 & 37.50 & 45 & 52.5 & 60\\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{array}\right\},$$

Table C.3
Monte Carlo rejection rates, and Monte Carlo bias and standard deviation of the MLEs of the homoscedastic model, when it is the true model.

Design	$n = 25$				$n = 75$			
	Rej. rate	β	Bias	sd	Rej. rate	β	Bias	sd
		a	27e−4	71e−5		a	23e−4	32e−5
ξ_{Unif}	0.046	b	1.11	21.45	0.035	b	1.19	27.44
		c	0.15	11.71		c	0.29	5.41
		a	25e−4	59e−5		a	38e−4	29e−5
ξ_D	0.044	b	1.47	35.37	0.039	b	1.39	40.99
		c	0.38	9.50		c	0.43	5.28
		a	15e−4	43e−5		a	24e−4	22e−5
$\xi_{k+3}^{(30,39.12,60)}$	0.064	b	0.58	8.89	0.049	b	1.14	22.84
		c	0.07	6.81		c	0.25	3.50
		a	17e−4	56e−5		a	25e−4	27e−5
$\xi_{k+1}^{(30)}$	0.036	b	1.05	32.09	0.040	b	1.51	29.31
		c	0.29	9.67		c	0.42	4.25

Table C.4
Monte Carlo rejection rates, and Monte Carlo bias and standard deviation of the MLEs of the heteroscedastic model, when it is the true model.

Design	$n = 25$				$n = 75$			
	Rej. rate	β	Bias	sd	Rej. rate	β	Bias	sd
		a	−34e−4	66e−5		a	−19e−4	22e−5
ξ_{Unif}	0.455	b	−1.97	70.62	0.925	b	−0.99	17.98
		c	−0.54	11.65		c	−0.24	3.23
		a	−18e−4	58e−5		a	−32e−4	24e−5
ξ_D	0.479	b	−1.45	91.68	0.953	b	−1.39	29.89
		c	−0.50	11.14		c	−0.26	3.09
		a	−14e−4	42e−5		a	−9e−4	14e−5
$\xi_{k+3}^{(30,39.12,60)}$	0.708	b	−1.0	39.44	0.999	b	−0.51	7.52
		c	−0.32	6.71		c	−0.14	1.85
		a	−31e−4	70e−5		a	−7e−4	16e−5
$\xi_{k+1}^{(30)}$	0.510	b	−2.01	111.74	0.994	b	−0.65	12.47
		c	−0.58	11.94		c	−0.23	2.08

with a D-efficiency of 83.90% and a KL-efficiency of 49.6%.

$$\xi_D = \left\{ \begin{array}{ccc} 30 & 44.17 & 60 \\ 1/3 & 1/3 & 1/3 \end{array} \right\},$$

with a D-efficiency of 100% and a KL-efficiency of 66.2%.

$$\xi_{k+3}^{(30,39.12,60)} = \left\{ \begin{array}{ccc} 30 & 37.71 & 60 \\ 0.44 & 0.25 & 0.31 \end{array} \right\},$$

with a D-efficiency of 84.2% and a KL-efficiency of 80%.

$$\xi_{k+1}^{(30)} = \left\{ \begin{array}{ccc} 30 & 44.17 & 60 \\ 0.5 & 0.25 & 0.25 \end{array} \right\},$$

with a D-efficiency of 94.5% and a KL-efficiency of 72.8%.

We recall that we are interested in testing

$$\begin{cases} \mathbf{H}_0 \;:\; \varepsilon \sim \mathcal{N}(0, \sigma_0^2) \\ \mathbf{H}_1 \;:\; \varepsilon \sim \mathcal{N}(0, \sigma_1^2\, \eta(x, \boldsymbol{\beta})^2) \end{cases}$$

and we apply the likelihood ratio test statistic (LRT). The distribution of LRT under the null hypothesis has been generated, for each design, through 1000 simulations; herein, this is called Monte Carlo reference distribution. To inspect how sample size affects both the precision of the estimates and power of the test, two sample sizes have been used, $n = 25$ and $n = 75$. By using the 1000 samples generated from the homoscedastic model and by generating additional 1000 samples from the heteroscedastic model, the simulation results consist in:

- the rejection rates, computed as the proportion of times in which the null hypothesis has been rejected, on the basis of the Monte Carlo reference distribution;
- the Monte Carlo bias and standard deviation of the parameter estimates (MLEs).

The results obtained in the homoscedastic case are presented in Table C.3. The rejection rate gives the actual significance level. When $n = 75$, the actual significance level with the best agreement to $\alpha = 0.05$ is reached by $\xi_{k+3}^{(30,39.12,60)}$, while for $n = 25$ it is obtained by applying the uniform design. The standard deviations of the estimates are of comparable magnitude among the different designs; the smallest values correspond to the design $\xi_{k+3}^{(30,39.12,60)}$.

Of higher importance are the results obtained when the data are generated from the heteroscedastic model. In this case, displayed in Table C.4, the rejection rate is the power. It can be noted that the largest power is reached when $\xi_{k+3}^{(30,39.12,60)}$ is applied. This result is not surprising as this design has the largest KL-efficiency (80%). It discriminates almost perfectly when $n = 75$, and guarantees a power over 70% when $n = 25$. For what concerns bias and standard deviation of the MLEs, all the designs have a similar behaviour, and the best results correspond to $\xi_{k+3}^{(30,39.12,60)}$. Surprisingly, the D-optimal design does not lead to the smallest standard deviations; this can be explained through the fact that the D-optimal design minimises the determinant of the Fisher information matrix, and hence takes into account also the covariances of the MLEs. Tables C.3 and C.4 show that, when the model is correctly chosen, the MLEs have good performances, with relatively small bias and variances. When instead the incorrect model is estimated, the MSEs of the MLEs are affected. In particular, when computing bias and standard deviation of the MLEs under the incorrect model, we found a dramatic rise of the standard deviation for some of the designs, that was not mitigated by increasing the sample size. Noticeably, the worst results are observed for the augmented designs when the homoscedastic model is incorrectly assumed and $n = 75$: in that case, both b and c are not significant due to the excessive variance of their MLEs. However, it is important to highlight that this type of error (type II error) is extremely unlikely for $n = 75$, especially for $\xi_{k+3}^{(30,39.12,60)}$ (see Table C.4). These results are available on request and can be replicated with the code, Appendix D.

Appendix D. R code

The code used to compute the designs in this work is hosted in the GitHub repository https://github.com/Kezrael/KL-augment/. In there you can find the auxiliary functions as well as the code to compute the augmented designs, the compound designs and the simulations. Should there be any inquiries about it or help needed to adapt it please contact the corresponding author.

Data availability

No data was used for the research described in the article.

References

[1] A. Brozena, C.E. Davidson, B. Schlindler, D.E. Tevault, Vapor Pressure Data Analysis and Statistics, Technical Report ECBC-TR-1422, Edgewood Chemical Biological Center, U.S. Army RDECOM, Aberdeen Proving Ground, MD, 2016, URL: https://apps.dtic.mil/sti/tr/pdf/AD1022530.pdf.

[2] C. de la Calle-Arroyo, J. López-Fidalgo, L.J. Rodríguez-Aragón, Optimal designs for Antoine Equation, Chemometr. Intell. Lab. Syst. 214 (2021) 104334, http://dx.doi.org/10.1016/j.chemolab.2021.104334.

[3] Holger Dette, Weng Kee Wong, Optimal designs when the variance is a function of the mean, Biometrics 55 (3) (2004) 925–929, http://dx.doi.org/10.1111/j.0006-341X.1999.00925.x.

[4] Jim Burridge, Paola Sebastiani, D-optimal designs for generalised linear models with variance proportional to the square of the mean, Biometrika 81 (2) (1994) 295–304, http://dx.doi.org/10.2307/2336960.

[5] J. Fidalgo, C. Tommasi, P.C. Trandafir, An optimal experimental design criterion for discriminating between non-normal models, J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (2) (2007) 231–242, http://www.jstor.org/stable/4623265.

[6] A. Lanteri, S. Leorato, J. López-Fidalgo, Ch. Tommasi, Designing to detect heteroscedasticity in a regression model, J. R. Stat. Soc. Ser. B Stat. Methodol. 85 (2) (2023) 315–326, http://dx.doi.org/10.1093/jrsssb/qkad004.

[7] V.V. Fedorov, S.L. Leonov, Optimal design for nonlinear response models, in: Chapman & Hall/CRC Biostatistics Series, Taylor & Francis, 2013, http://dx.doi.org/10.1201/b15054.

[8] J. López-Fidalgo, Optimal experimental design: A concise introduction for researchers, in: Lecture Notes in Statistics, Springer Nature Switzerland, 2023, http://dx.doi.org/10.1007/978-3-031-35918-7.

[9] C. de la Calle-Arroyo, M. Amo-Salas, J. López-Fidalgo, L.J. Rodríguez-Aragón, W.K. Wong, A methodology to D-augment experimental designs, Chemometr. Intell. Lab. Syst. 237 (2023) 104822, http://dx.doi.org/10.1016/j.chemolab.2023.104822.

[10] S. Argumedo-Galván, V.I. López-Ríos, Metodología para incrementar el número de puntos experimentales en un diseño D−óptimo, Ingeniería y Cienc. 10 (2014) 181–201, https://www.redalyc.org/articulo.oa?id=83531311011.

[11] Timothy E. O'Brien, Optimal design and lack of fit in nonlinear regression models, in: Gilg U.H. Seeber, Brian J. Francis, Reinhold Hatzinger, Gabriele Steckel-Berger (Eds.), Statistical Modelling, Springer New York, New York, NY, 1995, pp. 201–206, http://dx.doi.org/10.1007/978-1-4612-0789-4_25.

[12] DDB, Dortmund data bank, 2024, www.ddbst.com, (Accessed 25 September 2024).

[13] C. de la Calle-Arroyo, J. López-Fidalgo, L.J. Rodríguez-Aragón, Optedr, 2021, https://cran.r-project.org/package=optedr. Version: 2.2.0.

[14] Y. Yu, D−optimal designs via a cocktail algorithm, Stat. Comput. 21 (2011) 475–481, http://dx.doi.org/10.1007/s11222-010-9183-2.

[15] Ch. Tommasi, Optimal designs for both model discrimination and parameter estimation, J. Statist. Plann. Inference 139 (12) (2009) 4123–4132, http://dx.doi.org/10.1016/j.jspi.2009.05.042.

[16] A.C. Atkinson, DT-optimum designs for model discrimination and parameter estimation, J. Statist. Plann. Inference 138 (1) (2008) 56–64, http://dx.doi.org/10.1016/j.jspi.2007.05.024.

[17] R.D. Cook, W.K. Wong, On the equivalence of constrained and compound optimal designs, J. Amer. Statist. Assoc. 89 (426) (1994) 687–692, http://dx.doi.org/10.1080/01621459.1994.10476794.

[18] C. de la Calle-Arroyo, M.A. González-Fernández, L.J. Rodríguez-Aragón, Optimal designs for Antoine's equation: Compound criteria and multi-objective designs via genetic algorithms, Mathematics 11 (3) (2023) 693, http://dx.doi.org/10.3390/math11030693.

[19] D.R. Cox, Tests of separate families of hypotheses, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, in: Contributions to the Theory of Statistics, vol. 1, 1961, pp. 105–123.

[20] A. Lanteri, S. Leorato, J. López-Fidalgo, Ch. Tommasi, Supplement to Designing to detect heteroscedasticity in a regression mode, J. R. Stat. Soc. Ser. B Stat. Methodol. 85 (2) (2023) 315–326, http://dx.doi.org/10.1093/jrsssb/qkad004.

[21] C.L. Yaws, H.C. Yang, To estimate vapor pressure easily, Hydrocarb. Process. (USA) 68 (10) (1989).