Tatra Mt. Math. Publ. **39** (2008), 135–143

TATRA MOUNTAINS
Mathematical Publications

SOME COMPLEXITIES IN OPTIMAL EXPERIMENTAL DESIGNS INTRODUCED BY REAL LIFE PROBLEMS

Sandra Garcet-Rodríguez — Jesús López-Fidalgo —
— Raúl Martín-Martín

ABSTRACT. Designing an experiment for a real life problem may involve new and complex situations. As a motivation a medical problem of finding an experimental design to predict cardiopulmonary morbidity after lung resection with standardized exercise oximetry is considered. Designing an experiment for models with a mixture of controlled and uncontrolled variables has already been considered in the literature. Another degree of complexity appears when an experimental unit can not complete the assigned experimental condition, e.g., the prescribed exercise time in the medical example. Thus, the controlled variable has to be considered as potentially censored. This paper is focused mainly on this problem for censored discrete distributions.

1. Introduction

Optimum experimental design is used in a variety of applications, engineering, operations research, economics and medicine, among other. This plethora of applications has led to the development of the classical theory due to some new complexities which may appear in real problems. Thus, A r d a n u y and L ó p e z - F i d a l g o (1992) computed optimal designs for experiments which have constraints both in its support and replications. C o o k and T h i b o d e a u (1980) considered some examples in which some of the independent variables are not subject to control of the practitioner. In particular, they considered two kinds of variables, one of them controlled and other uncontrolled, whose values are known before the experiment is performed. These designs are called marginally restricted designs. L ó p e z - F i d a l g o and G a r c e t - R o d r í g u e z (2004) considered the problem of constructing optimal experimental designs for

2000 M a t h e m a t i c s S u b j e c t C l a s s i f i c a t i o n: 62K05.
K e y w o r d s: censored variable, D-optimality, information matrix, sequential designs, uncontrolled variable.

regression models when the variable that is not under control can have unknown values before the experiment is performed. These are the so called conditionally restricted designs. In fact, they considered the mixture of both cases and provided equivalence theorems and iterative algorithms for generating approximate optimal designs. The main motivation was a real medical problem. V a r e l a, C o r d o v i l l a, J i m é n e z and N o v o a (2001) applied an exercise test to obtain more information in order to predict cardiopulmonary morbidity after lung resection with standardized exercise oximetry. The independent variables, happen to be in the model are: the expired volume of air in one second, the oxygen desaturation during the test and the exercise time in minutes.

Some other complexities appear in this real life problem. Thus, whenever a new patient arrives, his "Respiratory Function" may be measured. Then, an exercise time has to be assigned according to his specific "Respiratory Function" value. M a r t í n - M a r t í n (2006) provided a sequential algorithm to construct a marginally restricted D-optimal design in order to solve this problem.

Another complication arises when the patient, for some non-informative reason, can not complete the time of the exercise. This means that the controlled variable is potentially censored. In this paper, a known censoring probability distribution function is assumed for the controlled variable. Thus, when a particular design is tried, another different design is expected according to this distribution. An expression of the information matrix is provided in order to calculate the optimal design.

Some work has been done in optimal experimental design for potentially failing in the response. H a c k l (1995) provided a criterion based on D–optimality and obtained optimal exact uniform designs for possible missing observations in the quadratic model. I m h o f, S o n g and W o n g (2002, 2004) provided general procedures to compute approximate designs. As far as the authors know, nothing has been done for potentially censored independent variables. We deal with the problem of obtaining D-optimal approximate designs for a linear model when the values of some independent variables are potentially censored according to a known probability distribution function.

Let us consider a linear model defined by

$$\mathrm{E}[y] = \eta^T(x)\alpha,$$

where the components of $\eta^T(x) = \big(\eta_1(x), \ldots, \eta_m(x)\big)$ are m linearly independent continuous functions on some compact space χ, $\alpha^T = (\alpha_1, \ldots, \alpha_m)$ are unknown parameters to be estimated and the variance of the observations is assumed constant.

An *exact design* is a sequence of experimental conditions $x_1, \ldots, x_N$ from the design space χ. Assuming that only n of the points are different, a probability measure represents the design. If the point x_i appears N_i times in the design,

$p_i = N_i/N$ will be the probability of x_i, that is the proportion of experiments to be performed under these conditions.

Using this idea K i e f e r and W o l f o w i t z (1959) gave a more general definition of a design (*approximate design*) as any probability distribution, ξ. The information matrix is defined as

$$M(\xi) = \int_{\chi} \eta(x)\eta^T(x)\xi(\mathrm{d}t).$$

The set of the information matrices, $\mathcal{M}$, is convex and compact. Carathèodory's theorem says that given an information matrix, there always exists a design with the same information matrix and no more than $m(m+1)/2+1$ points in its support. Therefore we may restrict to the search of designs with finite support,

$$\xi = \left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{array} \right\},$$

where $\xi(x_i) = p_i$ is the proportion of experiments to be performed at the experimental condition x_i. Let $\Im$ be the convex set of the approximate designs.

The inverse of the information matrix is proportional to the covariance matrix of the least square estimates. In this paper we will focus on D-optimality, that is a criterion based on maximizing the determinant of the information matrix.

The D-efficiency will assess the goodness of a particular design ξ with respect to a D-optimal design ξ^*,

$$\mathrm{eff}_D(\xi) = \left(\frac{\det M(\xi)}{\det M(\xi^*)} \right)^{\left(\frac{1}{m}\right)}.$$

2. Potentially censored designs

A new complexity in the real case considered in this paper is analyzed in this Section. Here, $x \equiv t$ is the time that is going to be a variable potentially censored in a design space χ. The censoring distribution will be assumed known through a random variable T that measures the time, a chosen experimental unit is going to stop given no prior limitation in time. In the real case mentioned above, T would be the time a generic patient stops if he starts to ride the bicycle without any time limit imposed in advance. Assume the censored time T has a probability distribution on a set, which includes the whole design space. Let $f(t)$ and $F(t)$ be the probability distribution function and the cumulative distribution function,

respectively. A particular, but typical case, may be a distribution on $[0, \infty)$ with a design space contained in it.

Let $\hat{\xi}$ be the approximate design with finite support that is intended to be applied in practice. Then, another design ξ is expected to result at the end of the experimentation. Therefore, a design $\hat{\xi}$ should be found such that the expected design ξ will be optimal. We will call an optimal design with this restriction censoring restricted (CER) optimal design. Sometimes it is possible to find $\hat{\xi}$ such that the expected design ξ will be optimum according to the criterion without censoring. But frequently, this is not the case and a restricted search has to be performed. This happens mainly when there is an optimal time at the highest possible time value. As a matter of fact if the censoring distribution is continuous, this value will never be reached.

Let a discrete design space be

$$\chi = \{t_1, t_2, \ldots, t_n\}, \quad \text{where} \quad t_1 < \ldots < t_k < \ldots < t_n.$$

The censoring distribution will be considered as a discrete distribution on χ. The cumulative distribution function will be

$$F(t) = \sum_{i=1}^{k} f(t_i), \quad t_k \leq t < t_{k+1}, \quad k = 1, \ldots, n-1,$$

where

$$f(t_i) = P(T = t_i), \quad i = 1, \ldots, n.$$

This case corresponds to an experiment where n stages have to be completed and an experimental unit may stop the experiment at any of these stages. Thus, $f(t_i)$ will be the probability to stop exactly at time t_i, that is completing all stages until i and then stop. When a design $\hat{\xi}$ is tried in practice, an expected censored design ξ will be actually performed following the rule:

(1) All the tries at t_1 will succeed, thus all the weight given to t_1, $\hat{\xi}(t_1)$ will remain for ξ.

(2) The number of the tries at time t_2 which will not succeed, that is that will stop at time t_1, will be proportional to $f(t_1)$ and therefore the number of tries succeeding will be proportional to $1 - f(t_1)$. Thus, a proportion $f(t_1)\hat{\xi}(t_2)$ of the sample size will actually stop at t_1 and the rest $\big[1 - f(t_1)\big]\hat{\xi}(t_2)$ will reach the time challenge t_2.

(3) Following the same reasoning there will be a proportion of tries at time t_3 that is expected to succeed, $\big[f(t_3) + \cdots + f(t_n)\big]\hat{\xi}(t_3)$; a proportion that will stop at time t_2, $f(t_2)\hat{\xi}(t_3)$; and a proportion that will stop at time t_1, $f(t_1)\hat{\xi}(t_3)$.

A similar argument is used for the rest of the times. Therefore,

$$\xi(t_k) = \big[1 - F(t_{k-1})\big]\hat{\xi}(t_k) + \sum_{i=k+1}^{n} \hat{\xi}(t_i) P(T = t_k) \tag{1}$$

$$= [1 - F_{k-1}]\hat{\xi}(t_k) + f(t_k)[1 - \hat{\Xi}_k], \quad k = 1, \ldots, n-1, \tag{2}$$

and

$$\xi(t_n) = f(t_n)\hat{\xi}(t_n) = [1 - F_{n-1}]\hat{\xi}(t_n), \tag{3}$$

where

$$F_k \equiv F(t_k), \quad \hat{\Xi}_k \equiv \sum_{i=1}^{k} \hat{\xi}(t_i), \quad k = 1, \ldots, n-1; \quad F_0 \equiv 0 \quad \text{and} \quad \hat{\Xi}_0 \equiv 0.$$

By definition ξ is a probability measure and so it may be considered as a design. From the expressions above $\hat{\xi}$ may be worked out in function of ξ as described in what follows,

$$\begin{pmatrix} \hat{\xi}(t_1) \\ \hat{\xi}(t_2) \\ \cdots \\ \hat{\xi}(t_n) \end{pmatrix} = \begin{pmatrix} 1 - F_0 & f(t_1) & f(t_1) & \cdots & f(t_1) \\ 0 & 1 - F_1 & f(t_2) & \cdots & f(t_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & f(t_n) \end{pmatrix}^{-1} \begin{pmatrix} \xi(t_1) \\ \xi(t_2) \\ \cdots \\ \xi(t_n) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{1-F_0} & \frac{1}{1-F_0} - \frac{1}{1-F_1} & \cdots & \frac{1}{1-F_0} - \frac{1}{1-F_1} & \frac{1}{1-F_0} - \frac{1}{1-F_1} \\ 0 & \frac{1}{1-F_1} & \cdots & \frac{1}{1-F_1} - \frac{1}{1-F_2} & \frac{1}{1-F_1} - \frac{1}{1-F_2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{1-F_{n-2}} & \frac{1}{1-F_{n-2}} - \frac{1}{1-F_{n-1}} \\ 0 & 0 & \cdots & 0 & \frac{1}{1-F_{n-1}} \end{pmatrix} \begin{pmatrix} \xi(t_1) \\ \xi(t_2) \\ \cdots \\ \xi(t_{n-1}) \\ \xi(t_n) \end{pmatrix} \tag{4}$$

$$= \begin{pmatrix} \frac{1-\Xi_0}{1-F_0} - \frac{1-\Xi_1}{1-F_1} \\ \frac{1-\Xi_1}{1-F_1} - \frac{1-\Xi_2}{1-F_2} \\ \cdots \\ \frac{1-\Xi_{n-2}}{1-F_{n-2}} - \frac{1-\Xi_{n-1}}{1-F_{n-1}} \\ \frac{\xi(t_n)}{f(t_n)} \end{pmatrix},$$

where

$$\Xi_k \equiv \sum_{i=1}^{k} \xi(t_i), \quad k = 1, \ldots, n-1 \quad \text{and} \quad \Xi_0 \equiv 0.$$

The measure constructed in this way may not be a probability measure and therefore it can not always be considered as an experimental design. It sums up to one, but the values are non negative if and only if,

$$\frac{1 - \Xi_{i-1}}{1 - F_{i-1}} \geq \frac{1 - \Xi_i}{1 - F_i}, \qquad i = 1, \ldots, n-1. \tag{5}$$

that is, the ratio $r_i = (1-\Xi_i)/(1-F_i)$ is non increasing for $i = 1, \ldots, n-1$. Taking into account that the weights obtained sum up to 1, if this is satisfied, the values have to be no greater than 1. Therefore, this is the condition for $\hat{\xi}$ to be an experimental design. Let $\Im_F$ be the set of these designs,

$$\Im_F = \big\{\xi \mid \xi \text{ satisfies (5)}\big\}.$$

Theorem 1. *The set $\Im_F$ is convex.*

Proof. Let $\xi^{(1)}, \xi^{(2)} \in \Im_F$, $\alpha \in (0,1)$ and $\xi = (1-\alpha)\xi^{(1)} + \alpha\xi^{(2)}$. With the notation used above,

$$\begin{aligned} 1-\Xi_i =& \xi_{i+1} + \cdots + \xi_n = (1-\alpha)\Big(\xi_{i+1}^{(1)} + \cdots + \xi_n^{(1)}\Big) + \alpha\Big(\xi_{i+1}^{(2)} + \cdots + \xi_n^{(2)}\Big) \\ =& (1-\alpha)\Big(1-\Xi_i^{(1)}\Big) + \alpha\Big(1-\Xi_i^{(2)}\Big). \end{aligned}$$

Therefore,

$$\begin{aligned} (1-\Xi_i)(1-F_{i-1}) &= \left[(1-\alpha)\Big(1-\Xi_i^{(1)}\Big) + \alpha\Big(1-\Xi_i^{(2)}\Big)\right](1-F_{i-1}) \\ &\geq \left[(1-\alpha)\Big(1-\Xi_{i-1}^{(1)}\Big) + \alpha\Big(1-\Xi_{i-1}^{(2)}\Big)\right](1-F_i) \\ &= (1-\Xi_{i-1})(1-F_i). \end{aligned}$$

□

From Theorem 1 of López-Fidalgo and Garcet-Rodríguez (2004) an equivalence theorem may be stated. For that, the definition of a directional derivative of Φ at M in the direction of N is needed,

$$\partial\Phi(M,N) = \lim_{\epsilon\to 0} \frac{\Phi\big[(1-\epsilon)M + \epsilon N\big] - \Phi(M)}{\epsilon}.$$

Theorem 2. *If Φ is a convex function, then the following statements are equivalent:*

(1) $\Phi\big[M(\xi^\star)\big] = \inf_{\xi\in\Im_F} \Phi\big[M(\xi)\big]$,
where $\xi^\star$ is the CER Φ-optimal design.

(2) $\inf_{N\in\mathbf{M}_F} \partial\Phi\big[M(\xi^\star), N\big] = \sup_{\xi\in\Im_F^+} \inf_{N\in\mathbf{M}_F} \partial\Phi\big[M(\xi), N\big]$,
where $\mathbf{M}_F = \{M(\xi) \mid \xi \in \Im_F\}$ and $\Im_F^+$ is the set of the designs with nonsingular information matrix.

(3) $\inf_{N\in\mathbf{M}_F} \partial\Phi\big[M(\xi^\star), N\big] = 0$.

The procedure to compute Φ-optimal designs under this restriction is as follows:

(1) Compute the Φ–optimal design without any restriction, say ξ^*. If r_i^* is increasing in $i = 1, \ldots, n-1$, then compute $\hat{\xi}$ using equation (4) and the problem is solved. Otherwise go to step 2.

(2) An optimal expected design subject to the restriction (5) must be found. The information matrix associated to a generic expected design ξ, obtained with equations (2) and (3), is

$$M(\xi) = \sum_{k=1}^{n} \hat{\xi}(t_k) \left[\sum_{i=1}^{k-1} f(t_i)\eta(t_i)\eta^T(t_i) + (1 - F_{k-1})\eta(t_k)\eta^T(t_k) \right]$$

The objective is then to find,

$$\xi_F^* = \arg\min \Big\{ \Phi\big[M(\xi)\big] \mid \xi \in \Im_F \Big\}.$$

(3) In any of the two cases the design to be used in practice has to be computed from the optimal expected design using formula (4).

Remark. A one–point design, say ξ_{t_k}, is in $\Im_F$ if and only if $F_{k-1} = 0$. A two–point design, say ξ_{t_k,t_j}, $k < j$, is in $\Im_F$ if and only if $F_{k-1} = 0, F_{j-1} = f_k\ \xi(t_k) \leq \frac{1-F_k}{1-F_{k-1}}$. There is not a way to find a simple rule for the rest of the cases. Thus, it is not possible to find simple generators of the set of CER designs.

EXAMPLE. Let a model be,

$$\mathrm{E}(y) = \alpha_1 + \alpha_2 t,\ \mathrm{Var}(y) = \sigma^2,\ t \in \chi = \{0, 1, 2, 3\}.$$

Assume there is a binomial censoring distribution on χ, $\mathrm{Bi}(3, 1/3)$,

$$f \equiv \left\{ \begin{array}{cccc} 0 & 1 & 2 & 3 \\ 1/27 & 2/9 & 4/9 & 8/27 \end{array} \right\}.$$

A general design

$$\xi \equiv \left\{ \begin{array}{cccc} 0 & 1 & 2 & 3 \\ 1-p-q-r & p & q & r \end{array} \right\}$$

satisfies (5) if the sequence

$$\frac{1-\Xi}{1-F} \equiv \left\{ \frac{27(p+q+r)}{26}, \frac{27(q+r)}{20}, \frac{27r}{8} \right\},$$

is non increasing, that is $10p - 3q - 3r \geq 0$ and $2q - 3r \geq 0$. The information matrix for this design is

$$M(\xi) = \begin{pmatrix} 1 & p+2q+3r \\ p+2q+3r & p+4q+9r \end{pmatrix}.$$

Maximizing the determinant subject to those restrictions, the CER D-optimal design will be,

$$\xi \equiv \begin{Bmatrix} 0 & 1 & 2 & 3 \\ 71/162 & 7/54 & 7/27 & 14/81 \end{Bmatrix}$$

with determinant 49/36. It is well known that the unrestricted D-optimal design for this model gives half of the weight to each extreme point 0 and 3. Its determinant is 9/4 and the efficiency of the restricted optimal with respect to the unrestricted optimal is then $\frac{49/36}{9/4} = 7/9$, that is 77.8%.

Once the optimal expected design ξ is computed, there is the way back to compute the design to by tried in practice $\hat{\xi}$,

$$\hat{\xi} = \begin{Bmatrix} 0 & 1 & 2 & 3 \\ \frac{1-\Xi_0}{1-F_0} - \frac{1-\Xi_1}{1-F_1} & \frac{1-\Xi_1}{1-F_1} - \frac{1-\Xi_2}{1-F_2} & \frac{1-\Xi_2}{1-F_2} - \frac{1-\Xi_3}{1-F_3} & \frac{\xi(t_4)}{f(t_4)} \end{Bmatrix}$$

$$= \begin{Bmatrix} 0 & 1 & 2 & 3 \\ 1 - \frac{p+q+r}{26/27} & \frac{p+q+r}{26/27} - \frac{q+r}{20/27} & \frac{q+r}{20/27} - \frac{r}{8/27} & \frac{r}{8/27} \end{Bmatrix} = \begin{Bmatrix} 0 & 3 \\ 5/12 & 7/12 \end{Bmatrix}.$$

This result is quite logical since this is the safest way to get something as similar as possible to the unrestricted optimal design.

Acknowledgements. This paper has been written while the first author was doing a postdoc in the University of Salamanca and it was partially sponsored by Ministerio de Educación y Ciencia CMICYT MTM2004-06641-C02-01 and Junta de Castilla y León SA125/04.

REFERENCES

[AR92] ARDANUY, R.—LÓPEZ FIDALGO, J.: *Optimal design with constraint support*, Rev. Mat. Estatíst. **10** (1992), 193–205.

[CO80] COOK, R. D.—THIBODEAU, L. A.: *Marginally restricted D–optimal designs*, J. Amer. Statist. Assoc. **75** (1980), 366–371.

[ha95] HACKL, P.: *Optimal Designs for experiments with potentially failing trials.* In: MODA 4–Advances in model-oriented data analysis. (Ch. P. Kitsos et al., eds.), Proceedings of the 4th International Workshop in Spetses, Greece, June 5–9, 1995, Physica Verlag, Heidelberg, 1995, pp. 117–124.

[im02] IMHOF, L.—SONG, D.—WONG, W. K.: *Optimal design of experiments with possibly failing trials*, Statist. Sinica **12** (2002), 1145–1155.

[im04] IMHOF, L.—SONG, D.—WONG, W. K.: *Optimal design of experiments with anticipated pattern of missing observations*, J. Theoret. Biol. **228** (2004), 251–260.

[ki59] KIEFER, J.—WOLFOWITZ, J.: *Optimum design in regression problems*, Ann. Math. Statist. **30** (1959), 271–294.

[LO04] LÓPEZ FIDALGO, J.—GARCET-RODRÍGUEZ, S.: *Optimal experimental designs when some independent variables are not subject to control*, J. Amer. Statist. Assoc. **99** (2004), 1190–1199.

[MA06] MARTÍN-MARTÍN, R.: *Construction of Optimal Designs for Models with Uncontrolled Variables*, Thesis, 2006.

[Na89] NACHTSHEIM, C. J.: *On the design of experiments in the presence of fix covariates*, J. Statist. Plann. Inference **22** (1989), 203–212.

[Va01] VARELA, G.—CORDOVILLA, R.—JIMÉNEZ, M. F.—NOVOA, N.: *Utility of standarized exercise oximetry to predict cardiopulmonary morbidity after lung resection*, Eur. J. Cardiothorac. Surg. **19** (2001), 351–354.

Received September 26, 2006

Sandra Garcet-Rodríguez
Department of Statistics
Faculty of Sciences
Plaza de los Caídos
Universidad de Salamanca
37008-Salamanca
SPAIN
E-mail: sandra_garcet@usal.es

Jesús López-Fidalgo
Department of Mathematics
School of Mechanics Engineering
Avda. Camilo José Cela 3
Universidad de Castilla-La Mancha
13071-Ciudad Real
SPAIN
E-mail: jesus.lopezfidalgo@uclm.es

Raúl Martín-Martín
Department of Mathematics
School of Computer Sciences
University of Castilla-La Mancha
Avda. Camilo José Cela 3
13071-Ciudad Real
SPAIN
E-mail: raul.mmartin@uclm.es