Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2003/05

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Observational data and optimal experimental design discriminating between more than two models

Rossella Berni

Università degli Studi
di Firenze

Statistics

Observational data and optimal experimental design discriminating between more than two models

Rossella Berni[1]
Department of Statistics ”G. Parenti”
University of Florence
E-mail: berni@ds.unifi.it

abstract

In this work we focus on observational data and on the use of large data sets to implement an experimental design without additional runs, for an efficient use of these data, exploiting theory about optimal designs. More specifically, the proposed procedure is based on several steps and is an extension of a previous work. Furthermore, the final optimal design can be obtained discriminating between two models or more than two models. In this last case the suggested algorithm must be revised in order to discriminate between several equally close models assigning a weight to each one.The final goal is an experimental design that can be assumed as optimal by the point of view of a sequential optimality.

Keywords: Experimental design, observational data, D-optimal design, Sequential designs.

1 Introduction

The aim of this paper is to improve the use of observational data directed to build an experimental design starting from existing data. Nevertheless, the difficulties of our attempt are related to the fundamental theory on which design of experiment is based. Therefore we suggest a multi-step procedure in which each step aims for improvement towards optimality of the selected trials, where the final set of trials constitutes an optimal experimental design built through sequential plans. This work is an extension of a previous work (Berni, 2002), where we pointed out the relevance of the use of observational data by a suggested procedure in which the final steps is made using the sequential design approach when the models analyzed are two. In this work the extension is about a discrimination between several models obtaining an optimal design. The paper is structured as follows: the second section is dedicated to the modified new procedure; the third section is devoted to data and results of an empirical example, based on real data.

2 The procedure

The procedure described in this section is based on the design of experiments theory (Atkinson and Fedorov, 1992). We emphasize that this is a first attempt for an efficient use of observational data and we must keep in mind that there is a substantial difference between observational and experimental data, as the latter are implemented through criteria, previously recalled. Generally, a design of experiment may be built by means of

[1] Preliminary version presented at the International Conference on Economic and Social Statistics, 17-18 December 2002, Economic College Jinan University, Guangzhou, China.

the balance, orthogonality and randomization principles or/and optimality criteria. In both cases we select a set of trials, constrained to specific rules or measures.
Randomization is one of the fundamental rule for a standard experimental design; the systematic bias reduction is its main feature and the properties of the model estimates, subsequently computed, are well known.
Optimality criteria are based on iterative algorithms intended for improvement of the design efficiency. The main difference compared to a standard design, is that, in the optimality case, the selection is evaluated through efficiency measures based on information matrix; therefore, the choice of trials depends on the hypothetic model assumed. For example, if it is easy to suppose a curvature, the theory on optimality will find an efficient design for the research and detection of this curvature.
Regarding optimality criteria, the T-optimality is strictly connected to sequential designs, applied in this paper, just because through sequential designs we obtain a T-optimal design. Sequential designs have a substantial element of distinction compared to the methods based on iterative algorithms. In the sequential procedure, the model parameters are updated with the new selected trial at each sequence, using also the corresponding known value of the response variable; while, by optimality methods based on iterative algorithms, the selection of trials depends only on the factors' levels.
Two main issues should be solved to use observational data to build an experimental design: the lack of randomization and the problem of searching trials on a set, however large, of collected data. In substance, the presence of bias. In fact, an observed variable could retain variations not controlled by the experimenter. This problem is overcome when data are referred to technological systems or processes, as in most of the cases where this procedure is applied, where observational variables are however constrained to specific values in an established range, close to operational target or operational conditions, set by an operator.
Moreover, the availability of a very large data-set about a process is a resource for an industry and an efficient use of these data could be a valid alternative to their explorative use alone. This efficient use of the data is linked to the construction of a D-optimal design. In fact, when a design is D-optimum, the determinant of the information matrix is maximized; therefore this directly influences the variances of parameter estimates, and, obviously, the volume of the ellipsoid, confidence region for the parameters, which is strictly connected to the precision of the design.
A further and not secondary reason to select a particular subset of the whole data set resides into the need of searching points within a restricted experimental region. In fact the dimension of the experimental region is controlled by the conditions given by the initial set of candidate points and, subsequently, the selection of following runs from a list of candidates facilitates avoidance of considering less informative, even though complete, experimental area, by a technical point of view. With this respect, in section (3), the area of the experimental space is controlled by the conditions given by the initial set of points.

The procedure here described involves three steps; each step is intended to prepare the final experimental design as close as possible to optimality. The third step is the new step in comparison with the previous work (Berni, 2002) and it is an extension just because allows us to compare more than two models, assigning to each model a weight.
The steps are: 1) selection of a random sample of trials n_0 from the data sets;
2) use of a procedure to build up a design with n runs that maximizes $|X'X|$;

3) starting from the n runs, building up the final experimental design using sequential technique, involving the model specification and, obviously at this step, the response values.

We will now show in details these three steps with particular attention to the second and third stage.

The first step is intended for the selection of a subset of data for starting the procedure. The choice of a random sample should avoid a further bias due to a reasoned choice; at the same time it is beyond doubt that this step can not be a solution for the initial lack of randomization, since we proceed to select from an existing data-set.

However, two points could be positively evaluated: this initial selection is only performed at the beginning of the proposed algorithm and, in addition, if the data-set is very large, the bias can be reduced by means of the large size of available observations. This last remark is connected to the experimental region on which the data are collected. Furthermore, it is not irrelevant that, at this level of the procedure, we only use the factors' values while not considering response values.

The second step is based on the sample previously extracted (first step). At this stage we refer to Dykstra (1971) for building an optimal design by means of augmentation of experimental data. The aim of this stage is to create an initial optimal design starting from a number n_0 of candidate points, while the response value are not yet involved. In fact, at this point, the optimal design is built exploiting only the factors value. The method here applied maximizes $|X'X|$ starting from the sets of candidate points. We know (section 2.2) that the $|X'X|$ maximization is equivalent to minimize the generalized variance, obtaining a D-optimal design. For more details on this approach we refer to Dykstra (1971). Here we turn our attention to those points of interest for us. First of all, as stated previously, this algorithm exploits existing data to add, at each step, a new trial maximizing $|X'X|$. In substance, given n_0, we compute $(X'X)^{-1}$. Afterwards, $Var(\hat{y})$ is calculated for each candidate point and, as next trial, we choose that candidate point for which $Var(\hat{y})$ is greatest. We point out that, given an existing data-set, in our case the random sample of the first step, the new trials are selected within an area (experimental region) based on the list of candidate points and they are evaluated by a ratio so defined:

$$\frac{|X'X|_{n_0+1}}{|X'X|_{n_0}} \tag{1}$$

A general drawback of this procedure is its dependence on specification of the model (which is also the general common run when an optimal design is built); in fact, the selection of new trials is performed according to the kind of model: linear, quadratic or cubic model. A further limit of this approach when applied in our setting, is the constrained selection onto the remaining data-set, after the first step. The n experimental runs selected through the second step are used in the following step.

2.1 The new third step

This final stage is based on the theory described in Atkinson e Fedorov (1975), concerning sequential designs and T-optimality. More precisely we base our approach on sequential designs, that should lead to an asymptotically T-optimal design. The algorithm applied at this step is lightly modified compared to the algorithm described in Atkinson e Fedorov (1975). The modification is concerned with the area where the research of the new trial

x_{n+1} is performed, for each sequence; this change is due to the observational data and therefore the subsequent added new trials are selected within the residual data-set and not on a continuous set $x \in \chi$. At this final stage, once the new trial is selected, the available response value is used to update the parameters estimate of the models, recomputed on the new set. The algorithm for the third step is described in the following. This algorithm could be useful when the discrimination is made between more than two models. The approach can be equal to the algorithm showed in Berni(2002) if the rival models are reduced to only one model by means of ranking criteria.

With this respect, when we consider several models, the research of an optimal design by a sequential designs approach starts with a ranking of the Residual Sum of Squares (RSS), computed for each model. Assuming as k the number of considered models, the strictly ranking is:

$$R_1(\xi_1) < R_2(\xi_2) < R_3(\xi_3) < .... < R_k(\xi_k)$$

where R is for RSS.

The x_{N+1} observation is added if the following expression is verified:

$$\max_{x \in \chi} \{\eta_1(x, \hat{\theta}_{1N}) - \eta_2(x, \hat{\theta}_{2N})\}^2 \tag{2}$$

where $\hat{\theta}_{1N}$ and $\hat{\theta}_{2N}$ are OLS estimates of the model number 1 and number 2, respectively. In this simple case, we consider, after ranking, only the first two models, which have the best fit.

Unfortunately, a problem rises when two or more models are equidistant from the true model (assuming as true model the first model). This occurs because there is an experimental error that can affect the RSS, giving different values of this quantity, but, at the same time, these rival models should have the same value of the RSS in presence of a null experimental error.

The solution to this problem is performed using a new algorithm where the optimization is made by a two stage search and the best design is found for a particular combination of weights. The design selected is that design which minimizes the following:

$$\min_{p_j} \max_{x \in \chi} \sum_{j \in J(\xi_s)} p_j [\eta_1(x, \theta_1) - \eta_j(x, \theta_j(\xi_s))]^2 \tag{3}$$

with

$$\sum_{j \in J(\xi_s)} p_j = 1 \tag{4}$$

This algorithm is based on a first selection onto a set J of chosen models, where the distance between model number 2 and the others is $\leq N\delta$ where δ is equal to an arbitrary percentage of the maximum distance between the true model and the worst model belonging to J.

Therefore, the final algorithm for the third step of the procedure is:

1. Regarding N observations, obtained at the II step, we rank the models, where the true model is assumed as model number 1;

2. we found a set J of models where:

$$J_\delta = \{j : (R_2 - R_j) \leq N\delta\} \;\; j \neq 1, j = 2, ..., k \tag{5}$$

3. the x_{n+1} observation is added if:

$$\min_{p_j} \max_{x \in \chi} \sum_{j \in J(\xi_s)} p_j[\eta_1(x, \theta_1) - \eta_j(x, \theta_j(\xi_s))]^2 \tag{6}$$

If the set J_δ holds only the model number 2 then the procedure is reduced to the simple algorithm without weights. In this approach one practical problem is the choice of δ, in general it is preferable an initial high value of δ. In addition it is relevant to point out that the selection of the best design is tied to the weights and the number of the iterations to be performed.

3 An empirical example

The data applied in this section are about a foaming process for the door of a home refrigerator. The day-to-day observations regard the values of variables involved with the production of polyurethanic foam and they are collected during the first 82 days of 2000.

Independent variables:

A → Charge of cyclopentane (parts on total);
B → Calibration;
C → MDI[2] Temperature (Centigrade degree);
D → MDI Pressure (bar);
TP → Polyol temperature (Centigrade degree);
PP → Polyol pressure plus expanding material (bar);

Response variables:

Y1 → Time of thread (second);
Y2 → Free density (Kg/mc).

Table 1: Independent variables - Range and step size

Variable	min	max	range	step size
A	10.7	11.2	0.5	0.1
B	0.76	0.79	0.3	0.1
C	16	22	6	1
D	135	165	30	5
TP	16	21	5	1
PP	130	165	35	5

[2] MDI means Methyl Diphenyl Isocyanate.

The empirical example[3] here illustrated is based on the procedure of section 2. We proceed considering the 3 steps[4] but we turn our attention, above all, to the third step. We consider only two independent variables: MDI temperature (C) and Polyol temperature (TP), with only one response variable: Y_1 (table 1). The intervals of interest for this two variables are: TP 19-21; C 20-22.
The selected $n = 4$ observations for the initial experimental design are:

Table 2: Experimental design - step 2

obs	C	TP
1	20	19
2	20	20
3	22	19
4	22	20

The four chosen models are the following:

$$Y_1 = \beta_0 + \varepsilon \tag{7}$$

$$Y_1 = \beta_0 + \beta_2 TP + \varepsilon \tag{8}$$

$$Y_1 = \beta_0 + \beta_1 C + \varepsilon \tag{9}$$

and

$$Y_1 = \beta_0 + \beta_1 C + \beta_2 TP + \varepsilon \tag{10}$$

Where β denotes in general the unknown parameter for that effect. We don't know which of the four models is true, but we assume that the true model is model (10). In table (3) we show the results for the initial models, with the RSS and the OLS estimates.

Table 3: Initial results for the III step

Mod.	RSS	$\hat{\beta}_0$	$\hat{TP}$	$\hat{C}$
(7)	8.75 (3)	43.75	-	-
(8)	2.50 (2)	-5.00	2.50	-
(9)	6.50 (2)	59.5	-	-0.75
(10)	0.25 (1)	10.75	2.50	-0.75

The model ranking is:

$$R_{10}(\xi_{10}) < R_9(\xi_9) < R_8(\xi_8) < R_7(\xi_7)$$

[3] The application is computed using the Fortran language and the Statistical Analysis System (SAS) software.

[4] random sample selection, evaluation of n observations that maximize $|X'X|$ and, finally, sequential designs approach

Table 4: Iterations of the III step by weights and observation added (record number)

Iter.	weights (8);(9)	obs.no
1	0.9;0.1	1
2	0.1;0.9	76
3	0.1;0.9	45
4	0.1;0.9	75
5	0.9;0.1	62
6	0.9;0.1	24
7	0.9;0.1	2

Table 5: Final results for models- III step

Iter.	$\hat{\beta}_0$	$\hat{TP}$	RSS_8	$\hat{\beta}_0$	$\hat{C}$	RSS_9	$\hat{\beta}_0$	$\hat{TP}$	$\hat{C}$	RSS_{10}	SS
1	50.81	-0.35	11.46	54.2	-0.50	6.80	45.53	1.21	-1.21	2.91	9781
2	52.73	-0.44	13.29	51.19	-0.33	12.53	52.60	-0.15	-0.26	12.41	11897
3	50.33	-0.31	13.73	51.27	-0.33	12.60	52.26	-0.12	-0.27	12.42	13833
4	51.34	-0.36	14.96	49.54	-0.24	16.17	52.55	-0.30	-0.11	14.71	15949
5	52.33	-0.41	16.32	50.76	-0.30	17.03	53.76	-0.30	-0.17	15.57	17798
6	51.94	-0.39	16.41	50.75	-0.30	17.18	53.65	-0.28	-0.18	15.59	19734
7	52.46	**-0.41**	16.48	51.41	-0.34	17.29	53.57	-0.28	-0.17	15.59	21850

and $(R_{10} - R_7)$ is equal to 8.50; regarding a percentage of a 20%, we obtain a value of $\delta = 1.7$ and therefore $N\delta = 6.8$. This choice implies the elimination of model (7) from the set J_δ, which holds model (9) and model (8).

In table (4) we show the results of the third step of the procedure, according to the combination of weights for the two models and the observation added (record number).

Observing table (4) we can see that the final design have 11 observations.

In table (5) we show the final results[5] obtained during each iteration, with number of iteration, parameter estimates, residual sum of squares (RSS) value for each model and the total sum of squares .

At each iteration, starting from the initial experimental plan, we have updated the parameter estimates for the three models, after adding the selected point (table 4) maximizing the (3).The final results for this example are satisfactory. As we can see in tables (3-5), we obtained, after only seven iterations, good results and we must stress that the estimates may be considered steady right from the iteration no.5; nevertheless the seven iteration allows us to obtain a significative value for estimate effect related to TP variable, even though only at level of $\alpha = 0.10$. In addition, we must observe that the final combination of weight gives more importance towards model (8), confirming the relevance of variable TP.

A final remark is about the choice of weights. Undoubtedly a further extension of this work must consider weights not arbitrarily defined, for example based on the likelihood

[5] in bold type the significative values by t-student test (P-value < 0.10).

of the related model.

References

- Atkinson A.C., Fedorov V.V., (1975), *The design of experiments for discriminating between several models,* Biometrika, vol.62, pagg.: 289-303.

- Berni R., (2002), *The use of observational data to implement an optimal experimental design,* Proceedings of the II Annual ENBIS Conference (European Network for Business and Industrial Statistics), 23-24 settembre 2002, Rimini, Italy, pagg.:1-8. Accepted, after review, on *Quality and Reliability Engineering International Journal,* Wiley Ed.

- Dikstra O. Jr., (1971), *The augmentation of experimental data to maximize* $|X'X|$, Technometrics, Vol.13, No.3, pagg.: 683-688.

- Nguyen N.K., Miller A.J., (1992), *A review of some exchange algorithms for constructing discete D-optimal designs,* Computational Statistical & Data Analysis, Vol.14, 489-498.