Optimum designs with leverage control

Marcelo Andrade da Silva*
Interinstitutional Graduate Program in Statistics - USP/UFSCar, Sao Carlos, Brazil - marceloddr@gmail.com

Luzia Aparecida Trinca
Institute of Bioscience - Unesp, Botucatu, Brazil - ltrinca@ibb.unesp.br

Abstract

In many experiments in agriculture and biology, several factors are investigated at the same time. Limited material, work-people, physical space, equipment use, time or financial resources may restrict the number of units to be tested such that the use of full factorials are unviable. Careful planning is important in order to help the researcher obtaining the desired information. Usually the desired information is translated to a function of the information matrix in order to produce optimum designs. Composite design criteria incorporating several desired properties are very promising such that the design chosen presents good performances under several aspects. For the classic linear model, an optimum design specifies the design matrix, such that some function of interest of the information matrix or of its inverse is optimized. These functions, named "criterion functions", have the purpose of making sure the researcher meets his experimental objectives. Single property functions may produce designs that are too tight and lack robustness to missing data. Thus, we propose to include the H property, as defined below, in the expression of a compound criterion to prevent the inclusion of points in the design that are too influential in the model fitting. A fairly simple measure is based on the diagonal elements of the H matrix, sometimes also called influence matrix, projection matrix or hat matrix. These diagonal elements of the H are simple measures of the influence of each observation in the fitting of the model. Thus, we propose to minimize the variability of the diagonal elements of the H matrix. In the literature, the most familiar method for optimizing the design is the exchange algorithm. This method is a heuristic that perform swaps of treatments (point exchange) or factor levels (coordinate exchange) in an initial design until the criterion value stops improving. For experiments with large number of factors, computation is expensive and thus computational efficiency is an issue. We implemented the method in C++ language and produced designs for some examples varying the composition of the design criterion. In general, the new compound criterion that incorporates the H property produced quite attractive designs since their efficiency under usual properties are very high and their leverages are more homogeneous indicating that the designs are more robust to missing data and prevents the data from having influential observations.

Keywords: optimum factorial designs; robust designs; missing observation; compound criteria.

1. Introduction

In Biotechnological, Industry, Pharmaceutical, Agricultural, and others areas, it is often necessary obtain information on products and processes empirically. People involved with the problem in question need to plan and perform experiments, collect data and analyze them. Experiments are controlled studies with deliberate alteration that are made to solve manufacturing or production of inert or biologic material problems, decide between different products or methodologies, understand the influence os certain factors, among others. This task becomes more and more complex and must be performed very attention to the extended that intensifies the technology-base products.

The optimum designs area has been increased greatly with Kiefer (1959), which organized a design construction theory aiming at the optimization of certain properties of the estimators of the model parameters, receiving the name of Theory of Optimum Designs. The lack of computer resources made the application of this theory to be realized only after 1970.

For the classic linear model ($\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$), a optimum design specifies the design matrix, $\mathbf{X}$, so that some of interest function $\mathbf{X}'\mathbf{X}$ is optimized. These functions reflect the experiment's objective and are called "criterion function". Optimality criteria compound incorporate multiple objectives according to the needs of the experimenter in the later stages to experiment planning.

In practice, during the execution of the experiment can result in missing of observations due to unexpected situations. Missing observations may produce biased estimates, skewed results or even not be possible to estimate some parameters of the predefined model. Observations that play influential role in model fit are also not desired in statistical modeling in general. Robust designs to the missing observations are attractive because they are more reliable for the experimenter. The linear model's theory proposes measures to influence the observations. The ideal experiment in this regard shall provide homogeneity in values of such measures. Thus, we proposed an optimality criterion in order to seek designs formed by treatments that do not stand out in terms of influence.

In the literature, the best known and used search method to construct exact optimum designs is the exchange algorithm. Proposed by Fedorov (1972), this method is a heuristic optimization consisting in, starting from an initial design, make exchanges, by substituting their points for the candidate points until the improvement in the value of the criterion ceases. Usually, the exchange algorithm's implementation is performed in statistical software, such as, for example, the R software, but for experiments with many factors and high levels, it requires high computational efficiency. Hence the implementation of this method was made in language C.

2. Optimum designs of experiments

Optimum designs are experimental designs based on certain criteria and are optimum just for a specific statistical model. The purpose of a search for a optimum design or near-optimum is to choose n points of a set of N possible points, called candidate points (set of all possible combinations of the factor levels), so that some function of the information matrix $\mathbf{X}'\mathbf{X}$ is optimal, it means, find an optimum design means searching a combination among the experimental region χ that optimizes the criterion function. This function is defined by one or more optimality criteria.

The optimality criteria, also originally called alphabetic criteria of optimality (Kiefer 1959; Atkinson et al., 2007) are almost always established by a function of the information matrix $\mathbf{M} = \mathbf{X}'\mathbf{X}$, or its inverse $\mathbf{M}^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$, which is proportional to covariance matrix of the model's parameters. In practice, the criteria for optimality have been developed to meet the objectives of the experimenter in the later steps of experiment planning. There are cases where the experimenter will need more than one criteria to find the right design for your experiment. Thus, the compounds criteria provide flexibility and efficiency for multi objectives designs construction and may include more than one optimality criterion, each weighing reflecting the relative importance of each objective of the experiment or the experimenter.

Gilmour and Trinca (2012) detached the following procedures that are, in general, applied in the analysis of results of a response surface experimental:

1. Global test F on the treatment effects, for which we shall use $(DP)_S$-optimality;

2. Test t for individual effects, for which we shall use AP-optimality, possible on the weighted version;

3. Point Estimation of the individual effects, for which we use weighted A-optimality;

4. To verify the lack of fit of the simplified model and, if appropriate, the inclusion of some heigher order terms in the polynomial. The efficiency related to the use of experimental procedures, referred to as efficiency in terms of degrees os freedom by Daniel (1976) was used in this question.

$$
\begin{array}{cc}
|\mathbf{X}'\mathbf{Q}_0\mathbf{X}|^{\frac{1}{p-1}} & \text{criteria } D, \\
\dfrac{1}{\text{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{Q}_0\mathbf{X})^{-1}\}} & \text{criteria } A, \\
(n-d) & \text{degrees of freedom,} \\
\dfrac{|\mathbf{X}'\mathbf{Q}_0\mathbf{X}|^{\frac{1}{p-1}}}{F_{(1-\alpha_1;p-1;d)}} & \text{criteria } DP, \\
\dfrac{1}{F_{(1-\alpha_2;1;d)}\text{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{Q}_0\mathbf{X})^{-1}\}} & \text{criteria } AP.
\end{array}
$$

At this form, Gilmour and Trinca (2012) proposed a compound criteria function for the properties below: where $\mathbf{Q}_0 = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$, so that the criteria considers the model as a nuisance parameters and not estimation parameters.
Gathering the five listed properties, Gilmour and Trinca (2012) obtained the composite criterion function given by:

$$
\frac{\left|\mathbf{X}'\mathbf{Q}_0\mathbf{X}\right|^{\frac{\kappa_1+\kappa_4}{p-1}}\left(n-d\right)^{\kappa_3}}{\left[F_{(1-\alpha_1;p-1;d)}\right]^{\kappa_4}\left[F_{(1-\alpha_2;1;d)}\right]^{\kappa_5}\left[\text{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{Q}_0\mathbf{X})^{-1}\}\right]^{\kappa_2+\kappa_5}}, \tag{1}
$$

where $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5)$ is the priority weight vector of which property and d is the degrees of freedom for pure error.

3. Robustness to missing data

In the literature, there is no optimality criteria that give robustness to a experiment related to the lack of observation. Our proposal is to find an optimal design, including in the optimality criteria a property to prevent that the design includes influential points in the model fit. A reasonable measure simple of the influence is given by the diagonal elements of the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the h_{ii}'s $(i = 1, \ldots, n)$, as these elements measure the influence of each observation on the model fit. According to the properties of the matrix, $\mathbf{H}$, the ideal design, according this criteria, would present all the equal elements to p/n, since they are n elements in the diagonal and the sum of it is p. Thus, we explore to minimize $\sum_{i=1}^{n}(h_{ii} - p/n)^2$, which means to minimize the variability of the h_{ii}'s, making it the next of p/n, and so, minimizing the influence heterogeneity of each experiment observation. This criteria will be called H-optimality in reference to the matrix $\mathbf{H}$.

Combining the considered properties by Gilmour and Trinca (2012) and the criteria H proposed in this work, each one associated to a weight priority of the analysis given by the vector $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6)'$, we propose a new compound criteria

$$
\frac{\left|\mathbf{X}'\mathbf{Q}_0\mathbf{X}\right|^{\frac{\kappa_1+\kappa_4}{p-1}}\left(n-d\right)^{\kappa_3}}{\left[F_{(1-\alpha_1;p-1;d)}\right]^{\kappa_4}\left[F_{(1-\alpha_2;1;d)}\right]^{\kappa_5}\left[\text{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{Q}_0\mathbf{X})^{-1}\}\right]^{\kappa_2+\kappa_5}\left[\sum_{i=1}^{n}(h_{ii}-p/n)^2+\delta\right]^{\frac{\kappa_6}{2}}}, \tag{2}
$$

wherein $\mathbf{Q}_0 = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$, so the criteria considers the model intercept as nuisance parameter and no estimation priority and δ was fixed in 10^{-6} to avoid numerical problems in the case to find the ideal design related to the criteria H.

4. Exchange algorithm

For the design research using the others optimality criteria, the exchange algorithms of Cook and Nachtsheim (1989) and Meyer and Nachtsheim (1995) were implemented and used to search optimum designs at this work. This algorithms are modified versions of the Fedorov's exchange algorithm (1972) and are called Point-Exchange Algorithm (*point-exchange*) and Coordinated-Exchange Algorithm (*Coordinate-Exchange*), respectively.

The Point-Exchange Algorithm steps are below.

Step 1: To define the model, the number k of factors, the number n of experiment observation, the levels of each factor, the weight vector $\mathbf{W}$ (weighted criteria A) and $\boldsymbol{\kappa}$ (compound criteria) and the number v of algorithm attempts.

Step 2: To create the candidate matrix with all possible points $\mathbf{x}_i$.

Step 3: To create a random initial design (nonsingular).

Step 4: To calculate $\mathbf{M}$, $|\mathbf{M}|$, $\mathbf{M}^{-1}$ and the compound criteria value to the initial design.

Step 5: To make a exchange for point (row), it means, fixing one row of the matrix $\mathbf{X}$ and replace it for a point od the candidate set.

Step 6: To update $|\mathbf{M}|$ e $\mathbf{M}^{-1}$ and to calculate the criteria value for this design.

Step 7: If the criteria value of this new design is larger than the previous criteria value, then the exchange is done, else, returns to the previous design. To return to **Passo 5** while the exchanges producing best values in the design criteria.

Step 8: The design found is stored and a new search is done (return to **Passo 3**) to ensure that the value criteria found is not local optimum. The return to the **Passo 3** is done v times.

To the Coordinate-Exchange Algorithm version, we must to consider the **Passo 2**, because this version do not needs the candidate point matrix and the exchanges made in the **Passo 5** are performed by coordinates and not by points.

5. Results

To investigate the potential of the new optimality criterion, we consider the example presented in Ahmad and Gilmour (2010): nine designs constructed by subsets considering an experiment with 36 observations, 4 factors each with three levels and assuming the full quadratic model ($n = 36; k = 4; p = 15$).

The Table 1 contains these designs along with their respective efficiencies of the optimality criteria D, A, DP, AP and H and also their respective degrees of freedom for pure error and lack of fit.

Table 1: Efficiency of the designs presented in Ahamad and Gilmour (2010)

Design	gl (EP; FA)	D_S-ef	A_S-ef	$(DP)_S$-ef	$(AP)_S$-ef	H-ef
$S_4 + 2S_1 + 4S_0$	(11; 10)	80.27	75.03	70.99	75.46	6.27
$S_4 + S_1 + 12S_0$	(11; 10)	68.31	59.15	60.41	59.50	4.91
$S_2 + S_1 + 4S_0$	(3; 18)	43.08	30.21	11.97	14.53	7.86
$S_3 + 4S_0$	(3; 18)	90.82	87.01	25.24	41.86	21.84
$S_4 + \frac{1}{2}S_3 + 4S_0$	(3; 18)	98.57	92.92	27.39	44.70	14.13
$S_4 + \frac{1}{2}S_4^{III} + S_1 + 4S_0$	(11; 10)	86.52	70.97	76.51	71.39	10.27
$S_4 + \frac{1}{2}S_4^{IV} + S_1 + 4S_0$	(11; 10)	85.74	70.37	75.82	70.78	9.15
$\frac{1}{2}S_4^{III} + S_2 + 4S_0$	(3; 18)	73.21	60.37	20.35	29.04	8.26
$\frac{1}{2}S_4^{IV} + S_2 + 4S_0$	(3; 18)	67.49	48.14	18.76	23.16	11.23

The ideal value of h_{ii} in this situation is $\frac{p}{n} = \frac{15}{36} = 0.417$. Obviously, the efficiency of the criterion H is high for designs that have h_{ii} neighbor than 0.417. The Table 2 presents three designs found using the criterion DP and two other compounds.

Table 2: Designs model with linear effects, quadratic effects and interactions 2 to 2 ($n = 36; k = 4; p = 15$)

	$(DP)_S$					Compound criterion: H; $\boldsymbol{\kappa} = \begin{pmatrix} D & A & GL & DP & AP & H \\ 0 & 0 & 0 & 0,5 & 0 & 0,5 \end{pmatrix}$					$\boldsymbol{\kappa} = \begin{pmatrix} D & A & GL & DP & AP & H \\ 0 & 0 & 0 & 0,8 & 0 & 0,2 \end{pmatrix}$				
	I					II					III				
	X_1	X_2	X_3	X_4	h_{ii}	X_1	X_2	X_3	X_4	h_{ii}	X_1	X_2	X_3	X_4	h_{ii}
	−1	−1	−1	0	0.413	−1	−1	−1	−1	0.423	−1	−1	−1	−1	0.426
	−1	−1	−1	0	0.413	−1	−1	−1	−1	0.423	−1	−1	−1	−1	0.426
	−1	−1	0	−1	0.659	−1	−1	1	1	0.411	−1	−1	−1	1	0.410
	−1	−1	1	1	0.449	−1	−1	1	1	0.411	−1	−1	−1	1	0.410
	−1	−1	1	1	0.449	−1	0	−1	−1	0.403	−1	−1	1	0	0.422
	−1	0	1	−1	0.398	−1	0	0	0	0.441	−1	−1	1	0	0.422
	−1	0	1	−1	0.398	−1	0	1	1	0.432	−1	−1	1	1	0.457
	−1	1	−1	−1	0.444	−1	1	−1	1	0.426	−1	0	0	0	0.436
	−1	1	−1	−1	0.444	−1	1	−1	1	0.426	−1	0	1	−1	0.410
	−1	1	−1	1	0.484	−1	1	0	1	0.420	−1	0	1	−1	0.410
	−1	1	−1	1	0.484	−1	1	1	−1	0.414	−1	1	−1	−1	0.389
	−1	1	1	0	0.403	−1	1	1	−1	0.414	−1	1	−1	−1	0.389
	−1	1	1	0	0.403	−1	1	1	0	0.401	−1	1	−1	1	0.420
	0	−1	−1	1	0.418	0	−1	−1	1	0.421	−1	1	−1	1	0.420
	0	−1	−1	1	0.418	0	−1	−1	1	0.421	−1	1	0	−1	0.427
	0	−1	1	−1	0.379	0	−1	1	−1	0.419	−1	1	1	1	0.411
	0	−1	1	−1	0.379	0	−1	1	−1	0.419	−1	1	1	1	0.411
	0	0	0	0	0.317	0	−1	1	0	0.377	0	−1	0	1	0.418
	0	0	0	0	0.317	0	0	−1	0	0.398	0	0	0	0	0.444
	0	0	0	0	0.317	0	1	−1	−1	0.422	0	0	1	1	0.397
	0	1	1	1	0.402	0	1	−1	−1	0.422	0	0	1	1	0.397
	0	1	1	1	0.402	0	1	1	1	0.429	0	1	1	−1	0.409
	1	−1	−1	−1	0.462	0	1	1	1	0.429	0	1	1	−1	0.409
	1	−1	−1	−1	0.462	1	−1	−1	−1	0.399	1	−1	−1	−1	0.440
	1	−1	0	1	0.409	1	−1	−1	−1	0.399	1	−1	−1	1	0.430
	1	−1	0	1	0.409	1	−1	−1	0	0.416	1	−1	−1	1	0.430
	1	−1	1	0	0.381	1	−1	0	1	0.447	1	−1	0	−1	0.401
	1	−1	1	0	0.381	1	−1	1	1	0.411	1	−1	1	−1	0.395
	1	0	1	1	0.400	1	−1	1	1	0.411	1	−1	1	−1	0.395
	1	0	1	1	0.400	1	0	−1	1	0.416	1	0	−1	0	0.420
	1	1	−1	−1	0.345	1	0	0	−1	0.405	1	1	−1	−1	0.419
	1	1	−1	−1	0.345	1	1	−1	1	0.420	1	1	−1	−1	0.419
	1	1	−1	1	0.444	1	1	−1	1	0.420	1	1	0	1	0.430
	1	1	−1	1	0.444	1	1	0	−1	0.403	1	1	0	1	0.430
	1	1	1	−1	0.462	1	1	1	−1	0.424	1	1	1	0	0.412
	1	1	1	−1	0.462	1	1	1	−1	0.424	1	1	1	0	0.412
gl (EP; FA)	(18; 3)					(12; 9)					(14; 7)				
D_S-ef	94.55					94.38					91.32				
A_S-ef	84.84					85.79					74.39				
$(DP)_S$-ef	100.00					86.67					89.04				
$(AP)_S$-ef	93.65					88.05					78.79				
H-ef	21.52					100.00					84.63				

The Design I was constructed using the criterion DP. The values of h_{ii} are between 0.317 and 0.659, which makes the design with low efficiency criterion for H, only 21.52%. For the other criteria, the efficiency of this design is high.

The Design II was found composing the criterion DP and H, each with a weight equal to 0.5. This design is also H-optimal. The values of h_{ii} range from 0.377 to 0.447.

The Design III, construction with the composition of the criteria DP and H with weights 0.8 and 0.2, respectively, have high efficiency for all criteria separately. The values of h_{ii} range from 0.389 to 0.457.

6. Conclusions

The criterion H increased the alternatives of compounds criteria, adding the property of robustness to influence or miss of observations, since the use of other criteria without the criterion H to produce efficient design criteria, but can not provide the robustness miss or influence of observations.

In general, the new composite criterion produce more attractive designs, with values leverages more homogeneous and therefore more robust to the miss of observations, however in some cases the use of this criterion alone undermines the efficiency of other criteria. Thus, we consider important the criterion H composed with other criteria, according to the objective of the researcher, because in practice the probability of missing observations is reasonable.

References

Ahmad, T.; Gilmour, S. G. (2010). Robustness of subset response surface designs to missing observations. Journal of Statistical Planning and Inference, v.140, p.92-103.

ATKINSON, A. C.; DONEV, A. N.; TOBIAS, R. D. (2007). Optimum experimental designs, with SAS. Oxford: Oxford University Press, 511p.

Cook, R. D.; Nachtsheim, C. J. (1989). Computer-aided blocking of factorial and response-surface designs. Technometrics, v.31, p.339-346.

Daniel, C. (1976). Applications of statistics to industrial experimentation. New York: Wiley, 564p.

Fedorov, V. V. (1972). Theory of optimal experiments. Academic Press.

Gilmour, S. G.; Trinca, L. A. (2012). Optimum design of experiments for statistical inference. Journal of the Royal Statistical Society, v.61, n.3, p.345-401.

Kiefer, J. (1959). Optimal experimental designs (with discussion). Journal of the Royal Statistical Society, v.21, n.2, p.272-319

Meyer, R. K.; Nachtsheim, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. Technometrics, v.37, n.1, p.60-69.