Making optimum design of experiments more useful in practice

Luzia Trinca*
Unesp, Botucatu, Brazil - ltrinca@ibb.unesp.br

Marcelo Andrade da Silva
Inter-institutional Graduation in Statistics, USP/UFSCar, São Carlos, Brazil - silva.marcelo@usp.br

Steven Gilmour
University of Southampton, Southampton, UK - S.Gilmour@soton.ac.uk

Abstract

For practical purposes an experimental design should present several good properties as the list highlighted by Box and Draper in 1975, further emphasised and enlarged by many other authors. Optimum design theory allows the construction of very efficient and economical designs of experiments. However such designs are usually optimal for the specific property optimised and under the correctness of the statistical model supposed at the planning phase. Advances are possible by using multiple criteria or composite criteria of the type proposed by Gilmour and Trinca in 2012. In this talk a range of properties is explored in order to construct efficient and flexible factorial designs for response surface studies. The approach provides efficient designs for parameter estimation, under practical experimental restrictions, that allow inferences to be carried out while preserving good performances on several desired aspects. Illustrations are motivated from practical problems.
Keywords: compound design criteria; factorial; response surface; randomization restrictions; robustness.

1. Introduction

Researchers in many areas perform experimentation varying the levels of several factors in order to gain insight on the behavior of one (or more) response of interest. For continuous responses and factors (or a mixture of continuous and qualitative factors) the relations are usually approximated by low-order polynomials. Important experimental design principles follow from the theory of factorials, response surfaces (RS) and optimum designs. However, after more than 50 years of the introduction of optimal design theory (Kiefer, 1959) many experimenters still prefer to use classical designs, e.g. full factorials, central composite, Box-Behenken, because they are simple and present some nice properties. Several important design properties were stated by Box and Draper (1975) and re-stated by many other authors, some of which are of course conflicting, and in order to fulfill a few of them using classical design ideas, a lot of resources might be expended or just a limited number of factors investigated. Application of optimum design theory allows the construction of efficient experimental plans and are, in particular, more useful when there are many factors, the resources are restricted, the blocks, if necessary, are small, or there exist other practical restrictions that make the simple randomisation layout inviable.

Restricting the search on the full factorial treatment combinations, an optimum factorial design, in the context discussed in this paper, is a set of points from the full factorial that optimises some desired property. Popular properties are functions of the information matrix, for example, the determinant of the information matrix (the larger the better), or the trace of its inverse (the smaller the better). These are the popular D and A criteria and the definition of several other criteria can be found in Atkinson, Donev and Tobias (2007). We refer to these criteria as usual criteria. They are frequently used as single criterion in an optimisation procedure. Although very efficient under the aspect being optimised, single criterion optimum designs lack acceptability in practice since they may perform badly with respect to other important properties. Besides, criteria D and A are interpreted as having inferential justifications but, as discussed in Gilmour and Trinca (2012), that is true only if the experimenter has an external error variance estimate. In most practical situations that is not the case, and care should be taken to the design such that experimental data contain information for obtaining valid error estimate and valid treatment effects inferences.

Gilmour and Trinca (2012) showed that the use of usual criteria for small or reasonable sized experiments often results in designs that do not have error degrees of freedom (df) and thus do not provide information for valid inferences. They also proposed adjustments so that this drawback is overcome. Another strong criticism against the use of single usual criteria is that the resulting designs are strongly dependent on the statistical model assumed for their construction and thus they lack model-robustness. Lu, Anderson-Cook and Robinson (2011) deal with some of the drawbacks by using multiple criteria and Pareto front approaches for exploring properties of the designs. Gilmour and Trinca (2012) proposed the use of a compound criteria involving four important properties, each of them related to an objective of the experiment. Smucker and Drew (2015) were concerned with finding efficient model-robust designs. The use of these type of approaches means the design constructed is not necessarily optimal for any of the properties but performs reasonably well for those considered important.

In this paper we show that by using a compound criteria it is possible to bundle in one optimisation function many desired properties. Each property is attached to a priority weight. The pattern of weights can be varied and interesting designs may be discovered. In sections 2 and 3 design properties and randomisation restrictions are mentioned, respectively, and in section 4 applications of the methodology are considered. The last example illustrates the flexibility of the approach in an experiment with restrictions on the randomisation of treatments to units.

2. Design properties and criteria

A RS experiment usually has several goals. Therefore constructing an economical and efficient design may be a challenge. Besides precise estimation of treatment effects, Box and Draper (1975) listed other 14 properties a design should have. Here we highlight the following: efficiency to estimate model parameters, internal estimate of error, detectability of lack of fit (LoF) and insensitiveness to wild observations.

Efficiency to estimate model parameters: Precise estimation of treatment effects are related to the usual D and A criteria. For randomised treatments in n units, the D-optimum design for the classical linear model with p parameters in $\boldsymbol{\beta}$, $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, is represented by the $\mathbf{X}$ matrix that maximises $|\mathbf{X}'\mathbf{X}|$. Given σ^2, the error variance, $\mathbf{X}$ is the design that minimises the volume of the confidence region for $\boldsymbol{\beta}$. The L criterion, a class that includes A, minimises the trace of $\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1}$ and is related to point estimation of the parameters or of linear functions of them. In this paper $\mathbf{W}$ is a diagonal matrix that *corrects* the scale of the terms of the second-order model (makes them similar). For known σ^2 this criterion would minimises the averaged squared length of individual confidence intervals. These design criteria can also be defined in terms of interest in a subset of parameters, e.g. D_S, A_S and so on (see Atkinson *et al.*, 2007), useful when some parameters in the model are nuisances.

Valid inferences (internal estimate of error): For unknown σ^2, Gilmour and Trinca (2012) corrected the D criterion by $|\mathbf{X}'\mathbf{X}| \times (F_{\alpha;p;d})^{-p}$ where $F_{\alpha;p;d}$ is the $(1-\alpha)$ quantile of the F distribution and d is the number of *pure* error df allowed for by $\mathbf{X}$. For randomised designs the number of df d comes from treatment replications and thus, the use of this criterion, denoted $DP(\alpha)$, forces the design to have replications of treatments. An modification focusing inferences for individual parameters was also proposed, the $LP(\alpha)$ criterion, minimising the trace of $\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1} \times F_{\alpha;1;d}$. For notation simplification we will use $\alpha = 0.05$ and drop α from the criterion name. Definitions based on subset parameters, e.g. DP_S, AP_S, follow straight. As illustrated in Gilmour and Trinca (2012), designs obtained by the pure error adjusted criteria are very efficient and useful when the goals of the experiment are inferences on the parameters of the model used to construct the design. But, again, using a single property may lead to designs with severe drawbacks, such as no df for LoF checking and no robustness against influential or missing observations, a point raised by discussants of that paper.

Lack of fit: For formal evaluation of a design with respect to its capacity for checking for LoF of the initial model it would be required the specification of the set of extra terms (and their sizes) that would be in the model. Research on this has been developed by Atkinson (1972); Jones and Mitchell (1978); Goos, Kobilinsky, O'Brien and Vandebroek (2005); Lu *et al.* (2011) and Smucker and Drew (2015). Here we take the simple approach of Gilmour and Trinca (2012) and use the df efficiency, the ratio between the number of treatments and n.

Insensitiveness to wild observations or robustness to missing observations: Robustness to missing design points were considered by several authors but they did not consider optimising the design with respect to such property. Box and Draper (1975) studied the effect of changing the number of center points and values of axial points in Central Composite designs on the variance of leverages, the elements of the diagonal of $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Herzberg and Andrews (1976) and Andrews and Herzberg (1979) defined measures to compare designs taking into account given probabilities of design breakdown due to missing observations. Ahmad and Gilmour (2010) studied the losses of subset RS designs under design point dropouts. A measure of robustness to missing points based on the contribution of leverages to the Cook's distance, given by $n(\sum_{i=1}^{n} \frac{h_i}{(1-h_1)^2})^{-1}$ (the larger the better), is being proposed in Silva; Gilmour and Trinca (2015). This measured is referred as the H criterion and will be explored in section 4.

Compound criteria (CC): Optimising a function that composes several properties, each with a priority weight (κ), is quite promising. We propose an extension of the CC of Gilmour and Trinca (2012), including the H criterion. For E_D, E_{DP}, E_A, E_{AP}, E_{LoF} and E_H denoting the efficiencies with respect to the properties discussed above, our method finds $\mathbf{X}$ that maximises

$$(E_D)^{\kappa_D} \times (E_{DP})^{\kappa_{DP}} \times (E_A)^{\kappa_A} \times (E_{AP})^{\kappa_{AP}} \times (E_{LoF})^{\kappa_{LoF}} \times (E_H)^{\kappa_H},$$

where the powers are the priority weights. The priority weights add to the unity.

3. Treatment randomisation restrictions

The formulae for design criteria in section 3 refer to completely randomised designs. In practice it is frequent some more complex situations, for example the necessity of blocking because material or experimental condition heterogeneity or because hard to set factors. In case of blocking, the design will depend on the nature of blocking effects. If they are random we have a mixed model and the optimum design depends on the value of the variance ratio ($\eta = \frac{\sigma_b^2}{\sigma^2}$). For $\eta \rightarrow \infty$, a well known result is that the variance matrix of the estimator of treatment effects converges to that for the fixed effects model. Since for η very large, information on treatment effects is critical, we think it makes sense to choose the design under the fixed effects model and treating block effects as nuisance in the formulae for D, A and so on. For standard criteria the formulae for blocked designs are presented in, for example, Atkinson *et al.* (2007) while those for pure error adjusted criteria can be found in Gilmour and Trinca (2012).

The case of hard to set factors arises when the levels of one factor or the combinations of levels of a group of factors are difficult to set while the other factors can be easily handled. This leads to restrictions on the randomisation process that generates the class of multistratum designs (Trinca and Gilmour, 2001, 2015a,b). Each level of hardness to set factor defines a stratum of information and the associated model is a mixed model. There are many papers on constructing designs for such situations, most of them for two strata and considering given values of variance component ratios (Goos and Vandebroek, 2003; Jones and Goos, 2007, 2009 and others). However, as shown in Trinca and Gilmour (2015b), the resulting designs do not allow pure error estimation. After performing the experiment and getting the data it is also frequent to end up with some variance component estimates equal to zero, a result that may be due to lack of df for error.

Taking the insight that multiple strata of information introduces nested blocking schemes in which some factor effects are completely confounded with the block effects (for example in a split-plot experiment the effects of whole plot factors are confounded with whole plot effects), Trinca and Gilmour (2001) proposed a stratum-by-stratum approach that builds the multistratum design sequentially from the highest stratum to the lowest. The approach was improved in Trinca and Gilmour (2015a) taking benefits from exchange algorithms and using usual single criterion. Because in each step the design can be viewed as a blocked design, apart from the first step, which can be a completely randomised or a blocked design, adjusted pure error

Table 1: Properties of alternative designs for Example 1, completely randomised scheme

Design	Criterion	df(PE; LoF)	D	DP	A	AP	H	$tr(\mathbf{AA}')$	$tr(\mathbf{R}'\mathbf{R})$
1	$LAR^{\dagger}$	0; 4	100.00	0.00	99.05	0.00	69.67	50.25	0.00
2	D_S	0; 4	100.00	0.00	99.05	0.00	69.67	59.45	0.00
3	DP_S	4; 0	82.32	100.00	71.17	98.68	0.00	20.22	100.00
4	A_S	0; 4	99.69	0.00	100.00	0.00	73.98	53.56	0.00
5	AP_S	4; 0	82.32	100.00	72.12	100.00	0.00	20.22	72.85
6	H	0; 4	83.36	0.00	61.23	0.00	100.00	27.82	0.00
7	CC^1	2; 2	84.61	31.81	67.50	38.97	57.82	27.46	0.00
8	CC^2	4; 0	82.32	100.00	70.75	98.10	0.00	20.94	67.35
9	CC^3	3; 1	86.55	71.57	75.76	79.96	1.78	22.42	0.00
10	CC^4	2; 2	91.38	34.35	83.41	48.16	0.00	28.42	0.00
11	CC^5	2; 2	84.61	31.81	64.90	37.47	57.82	28.42	0.00
12	CC^6	3; 1	78.34	64.78	56.60	59.73	27.26	12.34	0.00
13	CC^7	3; 1	78.34	64.78	66.03	69.69	27.26	18.21	0.00
14	CC^8	3; 1	85.72	70.89	74.29	78.40	27.26	24.61	0.00

† Design 1 from Lu *et al.* (2011).
CC^1 : $\kappa_{DP} = \kappa_H = .5$; CC^2 : $\kappa_{DP} = \kappa_{LoF} = .5$; CC^3 : $\kappa_{DP} = .2; \kappa_{LoF} = .8$; CC^4 : $\kappa_{DP} = .1; \kappa_{LoF} = .9$;
CC^5 : $\kappa_{DP} = \kappa_{LoF} = \kappa_H = 1/3$; CC^6 : as CC^5 except DP has twice the weight for the others;
CC^7 : same κ for all 6 properties; CC^8 : as CC^7 except property H has half of the weights for the others.

criteria can also be used. This is explored in detail in Trinca and Gilmour (2015b). The issue of robustness of multistratum designs for missing observation is more complicated and the idea of using leverages in each step is, perhaps, not immediately applicable. More research is needed on this subject.

4. Examples

We illustrate the methods in two examples of completely randomised designs and one design with two strata. The designs are optimised by the standard point exchange algorithm. It starts with a random initial design that is improved sequentially by making changes between design points and points from a specified candidate set. Usual candidate set is the full factorial that allows the intended model to be fitted. Several tries (different initial designs) are run to increase the chance of getting the optimum design. However there is no guaranty that the optimum is found.

Example 1: This example considers a screening experiment described in Lu *et al.* (2011). There are five factors, each at two levels, to be run in 14 units. The initial model included linear effects and four out of ten two-factor interactions. The authors presented multiple properties of 17 possible designs. For their construction they consider D-efficiency, model coefficient misspecification and model error variance misspecification, and evaluated their designs through the use of Pareto frontier. Model misspecification was evaluated by considering the model with all two-factor interactions. Let $\mathbf{X}_1$ be the model matrix for the initial model and $\mathbf{X}_2$ the columns for the six extra two-factor interactions.
Table 1 shows the properties of designs constructed using different weight patterns for D, DP, H and LoF df. We did not consider model misspecification in our compound criteria but we calculated the efficiency of our designs in terms of $trace(\mathbf{AA}')$, for $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$, and $trace(\mathbf{R}'\mathbf{R})$ for $\mathbf{R} = (\mathbf{H}_1 - \mathbf{I})\mathbf{X}_2$ where $\mathbf{H}_1$ is the hat matrix for the initial model. For efficiency with respect to $trace(\mathbf{AA}')$ we used the best design Lu *et al.* (2011) found (their design 9) as the reference.

Note that designs constructed by usual criteria do not present pure error (PE) df while single modified criteria designs do not present LoF df. Our D_S-optimum design is equivalent to design 1 from Lu *et al.* in a few criteria but is better in terms of model misspecification. Designs with zero H efficiency has at least one leverage value equal to one and will breakdown in case some of these points are missing. But the use of single H criterion is not recommended because of low efficiencies on other properties. Designs 7, 10 and 11 have the same pattern of PE/LoF dfs but they are quite different in terms of other properties. Note that 10 could

Table 2: Properties of alternative designs for Example 2, completely randomised scheme

Design	Criterion	df(PE; LOF)	D	DP	A	AP	H
1	D_S	0; 8	100.00	0.00	96.10	0.00	93.05
2	DP_S	8; 0	84.67	100.00	67.09	88.16	0.00
3	A_S	1; 7	98.71	1.55	100.00	4.33	76.00
4	AP_S	7; 1	89.82	97.38	80.02	100.00	0.00
5	H	0; 8	46.71	0.00	12.27	0.00	100.00
6	CC^1	4; 4	81.85	53.48	61.38	55.64	70.71
7	CC^2	7; 1	89.82	97.38	77.59	96.96	0.00
8	CC^3	4; 4	88.72	57.97	75.53	68.46	67.68

$CC^1: \ \kappa_{DP} = \kappa_H = .5$; $CC^2: \ \kappa_{DP} = \kappa_{LoF} = .5$; $CC^3: \ \kappa_{DP} = \kappa_{LoF} = \kappa_H = 1/3$.

Table 3: Properties of alternative designs for Example 2, restricted randomisation scheme

		df(PE; LoF)		D			A		
Design	Criterion	Stratum 1	Stratum 2	$\eta = 1$	$\eta = 10$	$\eta = 100$	$\eta = 1$	$\eta = 10$	$\eta = 100$
1	D_S	0; 0	0;10	100.00	100.00	100.00	100.00	100.00	100.00
2	DP_S	2; 0	7; 1	89.77	89.56	89.53	82.02	94.71	99.35
3	A_S	0; 0	0;10	99.98	99.94	99.94	100.75	100.17	100.02
4	AP_S	2; 0	7; 1	87.00	86.18	86.07	88.15	96.55	99.58
5	CC^1	2; 0	5; 3	94.17	94.08	94.07	91.68	97.74	99.73

$CC^1: \ \kappa_{DP} = .2$; $\kappa_{LoF} = .8$.

be preferred because higher efficiencies on parameter estimation, however, even allowing for 2 df for LoF it has leverage equal to 1 and thus it is quite risk under missing data. Designs 9, 12, 13 and 14 also have the same df pattern, with the last 3 all being 27.26% while the first is only 1.78% H-efficient. We also note the error variance misspecification measure ($trace(\mathbf{R}'\mathbf{R})$) is related to PE df but not equivalent.

Example 2: This example was motivated by the experiment for metal removal from wastewater presented in Sofu, Sayilgan and Guney (2015). They run three separated experiments, one for each type of bacteria, to investigate metal removal from dairy wastewater. In each experiment ($n = 11$), three factors were varied: X_1 (bacteria biomass); X_2 (pH) and X_3 (temperature). A 2^3 factorial plus 3 center points were used for each bacteria type. They wanted to estimate the optimum level combination for maximum removal but their design did not allow proper estimation of curvature effects. By their conclusions, we suppose they also wanted to compare bacteria types but separated experiments and data analyses were performed. To illustrate that more efficient designs can be constructed, we will consider an experiment with two types of bacteria only, however, for three types the idea is the same except that there are two dummy variables in the model instead of one to account for bacteria type.

We start with randomised designs for a second-order model for the three quantitative factors plus the main effect for bacteria type and the interactions among all factors. Then, we investigate a more complex problem considering temperature as hard to set factor since we understand that in the actual experiment, samples with specified biomass and pH level adjusted were put to mix in a temperature-controlled orbital shaker together. Table 2 shows the properties of several designs for the completely randomised scheme with $n = 22$ runs. Again we highlight the danger of using single design criteria because the resulting designs may be very poor with respect to other aspects. There are groups of designs that exhibit the same df patterns but perform differently with respect to other properties. Although it seems reasonable to expect that by requiring LoF df the design will be pushed against having high leverage that is not necessarily the case.

Table 3 shows the results for the two-stratum designs where temperature is taken as hard to set and its levels will be randomised to 6 whole plots. Then the levels of the other three factors will be randomised to 4 units within each of the whole plots. The design problem in the first stratum is a completely randomised scheme for a single three-level factor. It is obvious that under several criteria the optimum is each level replicated twice. So we used this design in the first stratum and found the design in the second-stratum under a few criteria. Note that we present the properties of the whole design under the mixed model formulation. That depends on the variance component ratio and thus we varied it in order to inspect the design performances. The reference design to calculate the efficiencies is the D_S-optimum design obtained from the stratum-by-stratum construction approach. Note that standard single criteria do not allow for PE df in any of the strata. The efficiencies related to D are quite stable for all designs indicating this measure is not very sensitive to η. That is not the case for A and the efficiencies increase with η. Again, by using a compound criteria we built a compromise design that does allow PE and some LoF dfs. Note that because it is allowed only three distinct levels for temperature, there is no LoF df in the first stratum for any of the designs.

5. Conclusions

A range of properties was explored in order to construct efficient and flexible factorial designs for RS studies. Through the use of compound criteria we can construct efficient designs for parameter estimation, under practical experimental restrictions, that allow inferences to be carried out while preserving good performances on several desired aspects.

Acknowledgment

We acknowledge grant#2014/01818-0 São Paulo Research Foundation (FAPESP).

References

Andrews, D.F. & Herzberg, A.M. (1979). The robustness and optimality of response surface designs. *Journal of Statistical Planning and Inference*, **3**, 249–257.

Atkinson, A.C. (1972). Planning experiments to detect inadequate regression models. *Biometrika*, **59**, 275–293.

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

Box, G.E.P. & Draper, N.R. (1975). Robust designs. *Biometrika*, **62**, 347–352.

Gilmour, S.G. & Trinca, L.A. (2012). Optimum design of experiments for statistical inference (with discussion). *Applied Statistics*, **61**, 345–401.

Goos, P. & Vandebroek, M. (2003). D-optimal split-plot designs with given numbers and sizes of whole plots. *Technometrics*, **45**, 235–245.

Goos, P.; Kobilinsky, A.; O'Brien, T.E. & Vandebroek, M. (2005). Model-robust and model-sensitive designs. *Computl Statist. Data Anal.*, **49**, 201–216.

Herzberg, A.M.& Andrews, D.F. (1976). Some considerations in the optimal design of experiments in non-optimal situations. *Journal of the Royal Statistical Society, Series B*, **38**, 284–289.

Jones, B. & Goos, P. (2007). A candidate-set-free algorithm for generating D-optimal split-plot designs. *Applied Statistics*, **56**, 347–364.

Jones, B. & Goos, P. (2009). D-optimal design of split-split-plot experiments. *Biometrika*, **96**, 67–82.

Jones, E.R. & Mitchell, T.J. (1978). Design criteria for detecting model inadequacy. *Biometrika*, **65**, 541–551.

Kiefer, J. (1959). Optimum experimental designs (with discussion). *J. Roy. Statist. Soc., Series B*, **21**, 272-319.

Lu, L.; Anderson-Cook, C. & Robinson, T.J. (2011). Optimization of designed experiments based on multiple criteria utilizing a Pareto frontier. Technometrics, 54, 353–365.

Smucker, B.J. & Drew, N.D. (2015). Approximate Model Spaces for Model-Robust Experiment Design. *Technometrics*, **57**, 54–63.

Silva, M.A.; Gilmour, S.G. & Trinca, L.A. (2015). Missing observation robust designs. In preparation.

Sofu, A.; Sayilgan, E. & Guney, G. (2015). Experimental design for removal of Fe(II) and Zn(II) ions by different lactic acid bacteria biomasses. *Int . J. Environ. Res.*, **9**, 93–100.

Trinca, L.A. & Gilmour, S.G. (2001). Multi-stratum response surface designs.*Technometrics*, **43**, 25–33.

Trinca, L.A. & Gilmour, S.G. (2015a). Improved split-plot and multistratum response surface designs. *Technometrics*, in press.

Trinca, L.A. & Gilmour, S.G. (2015b). Split-plot and multistratum designs for statistical inference. In preparation.