Designs for factors with many levels

Philip J. Brown and Martin S. Ridout
SMSAS, University of Kent, Canterbury, Kent, CT2 7NF, UK

Abstract

We consider designs for f factors each at m levels, where f is small but m is large. Main effect designs with $m \times f$ experimental points are presented. For $f = 2$, two types of designs are considered, termed *sawtooth* and *dumbbell* designs. For three factors, cyclic sawtooth designs are considered. The designs are compared using various criteria. We compare our designs with others using a dataset arising in screening for drug discovery with $f = 2$ and $m = 50$, and find that the dumbbell design outperforms others.

KEYWORDS Connectivity; identifiability; lead optimisation in drug discovery; main effects; microarray loop designs; prediction and contrast variance; screening

In factorial experimentation, screening designs such as Plackett-Burman designs (Plackett and Burman, 1946) have been developed to identify the most important factors amongst a possibly large set of factors. These designs typically involve a small number of levels (2 or 3) of each factor. A different type of screening problem arises when there are just a few factors, but each has a large number of potential levels. One example arises in the development of new drugs, where a compound that has shown some desirable pharmaceutical activity, termed a *lead* compound, can be modified at each of two or three molecular sites. At each site there are many possible chemical modifications that can be made. In this context, the molecular sites are the factors and the modifications are the levels of the factors. The aim of the experimentation is to identify modifications to the lead compound that have improved pharmaceutical properties.

Although these are quite different screening problems, a common feature of both problems is that it is only possible to consider a small fraction of the full set of possible treatment factor combinations. Here we consider designs for the second screening problem which, like Plackett-Burman designs, are most useful when interactions are small in comparison with main effects. Specifically, we consider designs for f factors, each at m levels, where $f = 2$ or $f = 3$, but m is large, assuming an additive main-effect model.

We introduce two types of design for two factors, termed the sawtooth design and the dumbbell design. The sawtooth design uses $2m$ design points, which is almost minimal for estimating all the main effects of each factor. Each level of each factor occurs twice in the design. This design is related to loop designs used in microarray experiments (Kerr and Churchill, 2001; Vinciotti et al., 2005), and it properties can be determined using results on loop designs from Bailey (2007). The dumbbell design has $2m - 1$ points and factor levels are not equally replicated. Nonetheless, it is more efficient than the sawtooth design when the assumed additive model is correct, unless m is very small. We consider efficiency

in terms of the average variance of a difference between two main effect parameters, and in terms of the average variance of a full set of predictions for all combinations of the levels of the two factors.

Neither design provides degrees of freedom for estimating the residual variance, but this is not considered important in a preliminary screening stage where the emphasis is on assessing the relative magnitudes of the main effect parameters. However, some additional design points can be incorporated if it is considered important to estimate the residual variability.

Calculations of efficiency assume that the additive model is correct and that there are no missing values in the data. But in a real pharmaceutical setting there may be missing values for a variety of reasons; it may not be possible to synthesise some of the modified compounds, the modified compound may turn out to be inactive, or the assay used to quantify the pharmaceutical activity may fail. We consider a particular data set generated and analysed by Pickett et al. (2011) that is available at `http://pubs.acs.org`. The data are intended to be typical of data arising in the process of lead optimisation in drug development. In this example 50 possible modifications were considered at each of two molecular sites. The basic compound was an inhibitor and the aim was to synthesise all 50×50 possible modifications and measure their inhibitory strength (pIC_{50}). However, for the reasons outlined above, approximately one third of the potential data are missing.

We use simulations to compare different designs based on this data set, in terms of measures such as the mean squared prediction error for the prediction of the inhibitory strength of compounds that were not included in the experimental design. The comparisons included the dumbbell, the sawtooth and two ad hoc designs. Generally, the dumbbell design performs the best, though in occasional simulations it's performance is less satisfactory.

We also explore designs for three factors. There is a natural extension of the dumbbell design from two factors to three factors, but for the sawtooth design there are many more possibilities with three factors. We consider a range of designs with cyclic structure. The efficiency of these designs is very variable. By an exhaustive search within a large class of designs for $m \leq 12$, we identify that the optimal design always lies within a smaller class. For larger m, we restrict our search to this smaller class and devise efficient computational algorithms for computing our measures of statistical efficiency. These efficient algorithms enable us to identify optimal designs for m as large as 100 in just a few seconds of computing time.

References

Bailey, R. A. (2007). Designs for two-colour microarray experiments. *Journal of the Royal Statistical Society, Series C 56*, 365–394.

Kerr, M. K. and G. A. Churchill (2001). Experimental design for gene expression microar-

rays. *Biostatistics 2*, 183–201.

Pickett, S. D., D. V. S. Green, D. L. Hunt, D. A. Pardoe, and I. Hughes (2011). Automated lead optimisation of MMP -12 inhibitors using a genetic algorithm. *ACS Medicinal Chemistry Letters 2*, 28–33.

Plackett, R. L. and J. P. Burman (1946). Design of optimal multifactorial experiments. *Biometrika 23*, 305–325.

Vinciotti, V., R. Khanin, D. D'Alemonte, O. de Jesus, J. Rasaiyaah, C. P. Smith, P. Kellam, and E. Wit (2005). An experimental evaluation of a loop versus a reference design for two channel microarrays. *Bioinformatics 21*, 492–501.