Level-screening, proofs, correlations and code

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

Section 1 derives singular values of the two factor sawtooth with $2m$ observations and m levels of each factor, it also proves the result in Lemma 1 of the main paper on the product of non-zero eigenvalues. Section 2 derives optimality results for the dumbbell and the cross-linked dumbbell. Section 3 obtains correlation structures for the sawtooth and dumbbell. Finally in Section 4 a method for efficient computation of average variance of prediction and contrasts for the 3-factor sawtooth is given.

1 Two factor sawtooth singular values

We refer to the over-parameterized design specification, without an overall mean but with every level included. Thus with two factors each at m levels the model matrix, X, will be a $2m \times 2m$ matrix, with at most $2m-1$ non zero eigenvalues. It is easy to see that the X^TX matrix is also a $2m \times 2m$ matrix of the form:

$$\begin{pmatrix} 2I & C \\ C^T & 2I \end{pmatrix}.$$

Here I is the $m \times m$ identity matrix and C is a $m \times m$ circulant matrix, (Aitken, 1956), with C^TC being a symmetric circulant. This derived circulant matrix C^TC has three non-zero entries in each row with entries {2,1,1}, an entry 2 on the diagonal and for the sawtooth in the first row of C^TC, denoted $b = (b_1, \ldots b_m)$, ones in positions 2 and m. This derived matrix generates the eigenvalues of X^TX since those eigenvalues are given by the non-zero roots of $\det(\lambda I - X^TX) = 0$ which by a standard re-arrangement of the determinant of a partitioned matrix is identical to

$$\det[(2-\lambda)^2 - C^TC] = 0. \tag{1}$$

The eigenvalues of the symmetric circulant C^TC are given from the roots of unity as

$$\ell_i = \sum_{j=1}^{m} b_j g_{ij}, i = 1, \ldots, m, \tag{2}$$

with $g_{ij} = \cos[\frac{2\pi}{m}(i-1)(j-1)]$, (see Press, 1982, eqns 2.8.4, 8.3.20). Constructing the $m \times m$ matrix $G = (g_{ij})$ from this equation we can see that the columns of G apart from the first (a unit vector), appear in pairs with the 2nd and mth identical in the sawtooth, intuitively as a result of powers of the primitive root, equally spaced around the unit circle with $\cos(x) = \cos(2\pi - x)$. These pairings correspond to the unit elements of the b vectors and give the roots in (2) as

$$\ell_i = 2 + 2\cos[2\pi(i-1)/m], \qquad i = 1, \ldots, m.$$

Now from eqn (1) the $2m$ eigenvalues of X^TX are obtained from solving m quadratic equations giving

$$\lambda_i = 2 \pm \sqrt{\ell_i} \tag{3}$$

of which exactly one is zero. The roots appear in pairs with the two elements having a sum of 4, and product $(4-\ell_i), i = 1, \ldots, m$. The product of the $(m-1)$ pairs of products (with non-zero product) is m^2, (see Gradshteyn and Ryzhik, 1980, section 1.396 p34) so that the overall product of $(2m-1)$ non-zero singular values of X^TX is $4m^2$. It is of interest for the determinant criterion of optimality for both A and B main effects present, all other permutation designs consisting of more than one cycle being suboptimal.

2 A- and P-optimal 2-factor designs with $2m-1$ and $2m$ points

This section relies heavily on Tjur (1991) and Bailey (2007). These authors were concerned with block-treatment designs, whereas we are considering designs for two treatment factors with no blocking. However, in both cases, the assumed model is additive in the two factors. The difference is that in block-treatment designs, one is interested in comparisons only between levels of the treatment factor, and not directly in comparisons between the levels of the block factor, whereas we are equally interested in both factors.

Tjur represents a design as a two-colour graph, as in our Figure 1, and considers it as representing an electrical network, where the edges are connections of unit resistance, and lets $R(i,j)$ denote the resistance through the network between vertices i and j. These resistances can be calculated using rules for analysing electrical circuits. Theorem 1 of Tjur states that

$$\begin{aligned} \text{var}\,(\hat{\alpha}_i - \hat{\alpha}_{i'}) &= \sigma^2 R(i,i'), \\ \text{var}(\hat{\alpha}_i + \hat{\beta}_j) &= \sigma^2 R(i,j), \end{aligned}$$

and it enables the quantities V_A and V_P to be calculated directly for simple designs.

2.1 A- and P-optimal design with $2m-1$ points

Consider first a minimal design with $2m-1$ points. The graph then has $2m$ vertices and $2m-1$ edges and must be connected if all parameters are to be estimable. Therefore, the graph must be a tree and there is a unique path between any two vertices. The resistance along this path is just the length of the path. Thus a design is A-optimal if it minimizes the average distance between pairs of points of the same colour, and is P-optimal if it minimizes the average distance between points of opposite colour, and it is easy to see that the dumbbell design (with a single replicate of the anchor point) is optimal by both criteria.

For A-optimality, note that all points of the same colour are joined by a path of length two, and it is the minimum possible length, since points of the same colour are never joined directly. Similarly for P-optimality, there are $2m-1$ edges joining vertices of opposite colour and the remaining $(m-1)^2$ pairs of opposite-coloured vertices must be separated by a path of length at least 3. Therefore, the average variance is at least

$$\frac{3(m-1)^2+2m-1}{m^2} = \frac{3m^2-4m+2}{m^2}.$$

The dumbbell design achieves this lower bound, and is therefore P-optimal. It is not the *unique* P-optimal design, since points can be moved from one side of the dumbbell to the other without changing the value of V_P, as for example in Figure 1; however, such a design is clearly inferior to the dumbbell in terms of V_A.

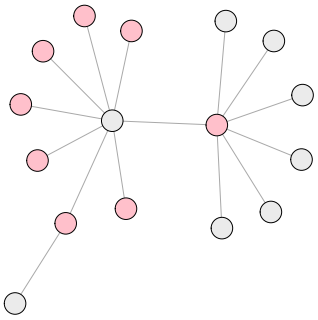

Figure 1: Example of a P-optimal design with $2m-1$ points that is not a dumbbell design

2.2 A- and P-optimal design with $2m$ points

Bailey notes that any connected graph with t vertices and t edges consists of a circuit of length s, where $2 \leq s \leq t$, and each vertex of the circuit may have a tree attached. Here, any $2m$-point design that allows all parameters to be estimated will give rise to a connected graph with $2m$ edges and $2m$ vertices, m of one colour and m of another colour, that consists of a circuit of length $2s$, where $2 \leq s \leq m$, and each vertex of the circuit may have a tree attached. In addition, every edge joins two vertices of opposite colour. Examples are shown in Figure 2 for $m = 8$.

Bailey uses a different graphical representation of block-treatment designs in which the vertices represent the treatments and there is an edge between two vertices if these two treatments occur in the same block. The relevant analysis is in Section 4.1 of Bailey. She shows that in a single colour graph, the A-efficiency of the design is improved by moving any trees that are attached to different vertices of the circuit so that they are all attached to the same vertex. A second argument of Bailey is that A-efficiency is also improved by collapsing the tree which is now joined to a single vertex, so that all vertices in the tree are joined directly to the vertex that is part of the circuit.

Similar arguments apply to the two-coloured graph, except that we can only move trees when only the vertices at which they are joined to the circuit have the same colour. Moreover, to improve P-efficiency, if there are trees attached to vertices of both colours, these vertices should be *adjacent*.

The result of applying these processes is a graph of the following canonical form

- There is a circuit consisting of $2s$ $(1 \leq s \leq m)$ vertices of alternating colour.
- One vertex on the circuit representing a level of factor A has a further $m - s$ B vertices attached to it.
- An adjacent vertex on the circuit representing a level of factor B has a further $m - s$ A vertices attached to it.

Particular instances of this canonical design are the dumbbell design with double replication of the anchor point ($s = 1$), the cross-linked dumbbell design ($s = 2$) and the sawtooth design ($s = m$).

Figure 2 gives an example for $m = 8$. The original design on the left is modified to the design on the right, which is in the canonical form above with $s = 3$. The original design has $V_P = 2.760$. It is not symmetrical in the factors A and B and for comparisons of levels of factor A, $V_A = 3.220$, whereas for comparisons of levels of factor B, $V_A = 2.738$. The design on the right is symmetrical in A and B and has $V_P = 2.141$ and, for comparison of levels of either factor, $V_A = 1.958$.

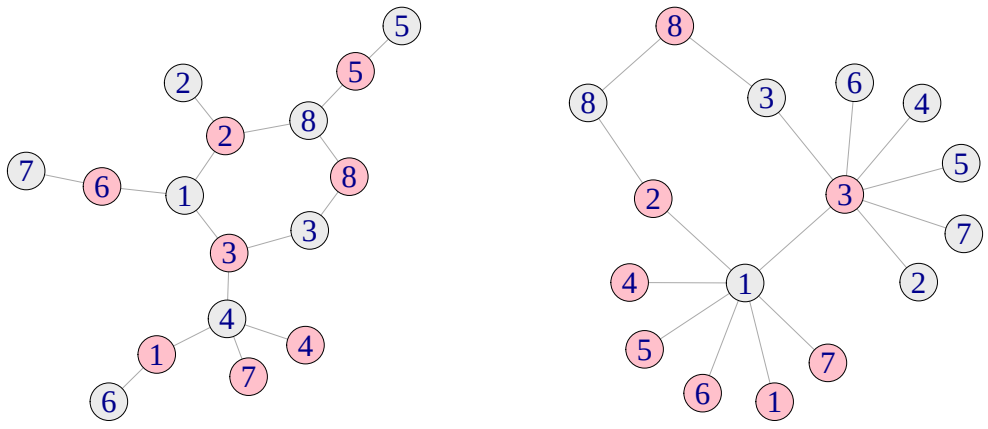

Figure 2: Example of the reduction of a design to the canonical form given in the text. Levels of factor A are indicated by grey points and levels of factor B by pink points.

Straightforward calculations, similar to Bailey, and based on the analogy with electrical

networks, give the following results for the canonical design

$$V_A = \frac{\sigma^2}{3m(m-1)} \left\{ 6m^2 - 5m + (4-6m)s + 2ms^2 - s^3 \right\}$$

and

$$V_P = \frac{\sigma^2}{6m^2 s} \left\{ 4ms^3 - 2s^4 + 18m^2 s - 24ms^2 + 6s^3 - 3m^2 + 2ms + 2s^2 \right\}$$

We can therefore obtain A- and $P-$optimal designs by choosing s to minimize these expressions.

For A-optimality, ignoring terms that do not involve s, we want to minimize the cubic

$$f(s) = (4-6m)s + 2ms^2 - s^3$$

over the range $s = 1, 2, \ldots, m$. We find $f(1) = 3 - 4m$, $f(2) = -4m$ and $f(3) = -15$, $f(m) = m(m^2 - 6m + 4)$. Thus $s = 2$ is always preferred to $s = 1$, and for $m > 3$, $f(2) < f(3)$. Also, $f(2) < f(m)$ for $m > 4$. Since $f(s)$ is a cubic, it can have at most one local minimum, and therefore for $m > 4$ the A-optimal design has $s = 2$, which is the cross-linked dumbbell design. For $m = 2, 3$ the A-optimal design is the sawtooth. For $m = 4$, the sawtooth design and the cross-linked dumbbell design give the same optimal value of V_A.

For P-optimality, the analysis is more complicated, because of the appearance of s in the denominator of the expression for V_P. For $m \leq 5$, the sawtooth design is optimal. For $m = 6, 7$, the optimal design has $s = 3$. For $m = 8$, V_P is minimized by choosing $s = 1$ or $s = 3$. However, for $m \geq 9$, V_P is minimized by choosing $s = 1$, the dumbbell design.

3 Correlations between estimators in 2-factor sawtooth and dumbbell designs

V_A-optimality focuses on the average of variance of all estimated pairwise differences between levels of the same factor, of the form $\hat{\alpha}_i - \hat{\alpha}_{i'}$. In general, these estimators will be correlated. V_P-optimality focuses on the average of variance of all estimated expected responses of the form $\hat{\alpha}_i + \hat{\beta}_j$. Typically these estimators will also be correlated. In this Section, we compare these two sets of correlations for the 2-factor dumbbell and sawtooth designs.

Because of the simple structure of the dumbbell design, only a limited set of distinct correlations can occur. These are shown below, along with their frequencies.

Correlations between estimators $\hat{\alpha}_i - \hat{\alpha}_{i'}$						
Correlation	$-\frac{1}{\sqrt{3}}$	$-\frac{1}{2}$	0	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{\sqrt{3}}$
Frequency	$\binom{m-1}{2}$	$\binom{m-1}{3}$	$3\binom{m}{4}$	$\binom{m-1}{2}$	$2\binom{m-1}{3}$	$\binom{m-1}{2}$

Correlations between estimators $\hat{\alpha}_i + \hat{\beta}_{i'}$					
Correlation	$-\sqrt{\frac{1}{5}}$	0	$\frac{1}{5}$	$\frac{3}{5}$	$\sqrt{\frac{2}{5}}$
Frequency	$(m-1)^2$	$(m-1)(2m^2-4m+3)$	$\frac{(m-1)^2(m-2)^2}{2}$	$(m-1)^2(m-2)$	$2(m-1)^2$

Correlations are more complicated to analyse in the sawtooth design. However, one explicit result concerns the largest positive and negative correlations. For $m \geq 3$, the largest positive and negative correlations between estimators of pairwise differences are

$$\begin{aligned} &\pm\sqrt{\frac{m-2}{m+2}} && m \text{ even} \\ &\pm\frac{m-1}{m+1} && m \text{ odd.} \end{aligned}$$

For the largest positive and negative correlations between estimators of estimated expected responses, the parity is reversed, to give, again for $m \geq 3$,

$$\begin{aligned} &\pm\sqrt{\frac{m-2}{m+2}} && m \text{ odd} \\ &\pm\frac{m-1}{m+1} && m \text{ even.} \end{aligned}$$

Thus, unlike for the dumbbell design, for the sawtooth design the largest positive and negative correlations approach ± 1 as $m \to \infty$ for both types of estimator. However, the mean absolute correlation decreases as m increases for both types of estimator and for both designs, as shown in Figure 3, which also shows results for the cross-linked dumbbell design. For the cross-linked dumbbell design, the correlations range from $-5/7$ to $\sqrt{2/7}$ for estimates of pairwise differences and from $-\sqrt{3/11}$ to $\sqrt{1/2}$ for estimates of expected responses.

4 Efficient computation for the 3-factor sawtooth design

Consider the saturated model

$$Y_{ijk\ell} = \alpha_i + \beta_j + \gamma_k + \xi_\ell + \varepsilon_{ijk\ell}, \qquad (1 \leq i, j, k \leq m, 1 \leq \ell \leq 3), \tag{4}$$

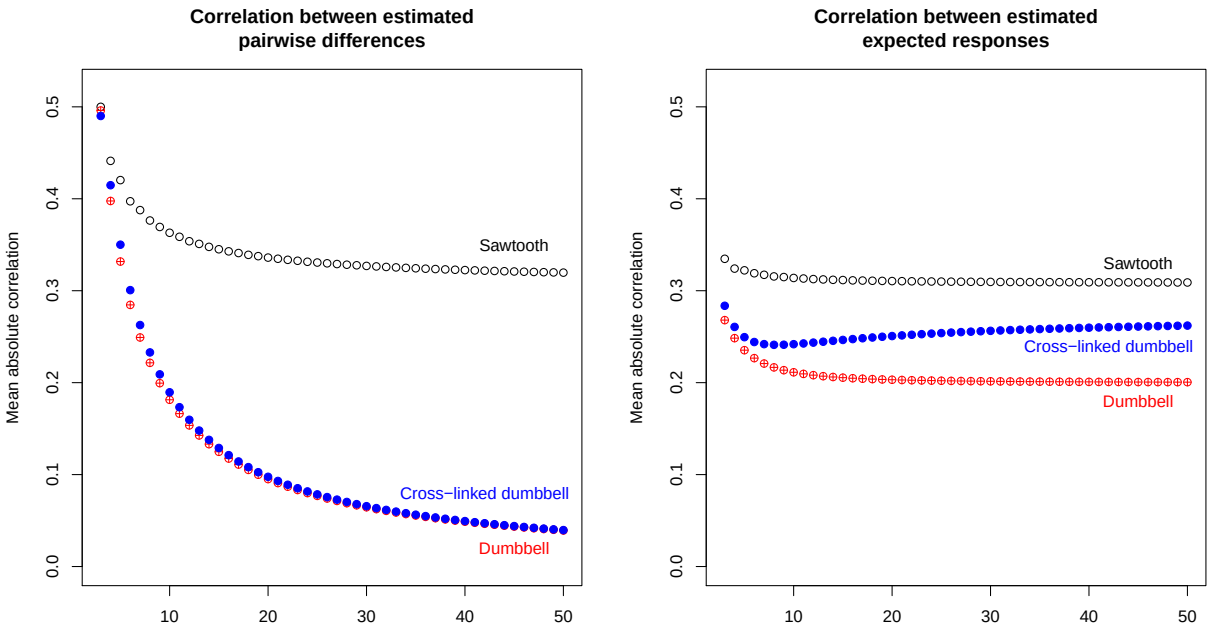

Figure 3: Correlations between estimators of pairwise differences and estimated expected responses for the sawtooth and dumbbell designs in relation to the number of levels of the two factors, m.

with the constraints $\beta_1 = 0$, $\gamma_1 = 0$ and $\xi_1 = 0$. As in the main text, we assume $\text{var}(\varepsilon_{ijk\ell}) = 1$; if instead $\text{var}(\varepsilon_{ijk\ell}) = \sigma^2$, where $\sigma^2 \neq 1$, then V_A and V_P need to be scaled by σ^2.

Let $\theta = (\alpha_1, \ldots, \alpha_m, \beta_2, \ldots, \beta_m, \gamma_2, \ldots, \gamma_m, \xi_2, \xi_3)^T$ denote the full vector of parameters. Since the model is saturated, the least squares estimator of θ is given by

$$\hat{\theta} = X^{-1}Y,$$

where X is the model matrix and Y is the vector of observations.

Let $Z = X^{-1}$. The least squares estimator of the treatment difference $\alpha_i - \alpha_j$ is

$$\hat{\alpha}_i - \hat{\alpha}_j = (Z_{i,\cdot} - Z_{j,\cdot})\, Y,$$

with variance

$$\|Z_{i,\cdot} - Z_{j,\cdot}\|^2,$$

where $Z_{i,\cdot}$ denotes the ith row of Z. Thus V_A can be efficiently calculated as the mean of the $m(m-1)/2$ squared Euclidean distances between the first m rows of Z.

For the average variance of predictions, we first note that due to the cyclic structure of the design, it is sufficient to average the variances of the predictions with the level of one factor held constant, since the pattern of prediction variances is repeated at each level of this factor. We therefore fix factor C at level 1. Because of the constraint $\gamma_1 = 0$, the

predicted mean response for level i of factor A and level j of factor B is

$$\hat{\mu}_{ij} = \hat{\alpha}_i + \hat{\beta}_j + \left(\hat{\xi_2} + \hat{\xi_3}\right) / 3,$$

where the last term averages across sets. Let W denote the $m \times 3m$ matrix with entries

$$w_{i,j} = \begin{cases} -\frac{1}{3}\left[3z_{i+m,j} + z_{3m-1,3m} + z_{3m,3m}\right] & 1 \le i \le m-1, \\ -\frac{1}{3}\left[z_{3m-1,3m} + z_{3m,3m}\right] & i = m. \end{cases}$$

Then

$$\hat{\mu}_{ij} = \left(Z_{i,.} - W_{j,.}\right) Y,$$

with variance

$$\left\| Z_{i,.} - W_{j,.} \right\|^2.$$

Thus the average variance of prediction, V_P, can be calculated as the mean of the m^2 squared Euclidean distances between the first m rows of Z and the m rows of W.

Moreover, V_P can be obtained without explicitly calculating all m^2 distances. Note that the mean of the squared distances between rows of W is again equal to V_A. If c_Z denotes the centroid of the first m rows of Z and c_W denotes the centroid of the m rows of W, then an analysis of variance decomposition gives

$$V_P = \frac{m-1}{m} V_A + \|c_Z - c_W\|^2.$$

The following R function generates the design with generators $(1,1,1)$, $(1,2,k+1)$ and $(1,k+1,k)$, where $1 \le k \le m$, and calculates V_A and V_P.

```
mk.design <- function(m, k)
{
    # Set up design factors
  aa = rep(1:m, each=3)
  bb <- (aa + rep(c(0,1,k), m)) %% m
  cc <- (aa + rep(c(0,k,k-1), m)) %% m
  a <- factor(aa)
  b <- factor(bb + m * (bb==0))
  c <- factor(cc + m * (cc==0))
  set <- factor(rep(c(1:3), m))

    # Obtain the inverse matrix Z
  Z <- solve(model.matrix( ~ -1 + a + b + c + set))

    # Centroids
  cz <-  colMeans(Z[1:m,])
  cw <- -colMeans(rbind(Z[c((m+1):(2*m-1)),], rep(0,3*m))) -
```

```
        2/3 * colMeans(Z[c((3*m-1):(3*m)),])

    # Calculate VA and VP
  VA <- mean(dist((Z[1:m,]))^2)
  VP <- (m-1)/m * VA + sum((cz-cw)^2)
  list(a=a, b=b, c=c, set=set, VA=VA, VP=VP)
}
```

References

Aitken, A. C. (1956). *Determinants and Matrices* (9th ed.). Oliver and Boyd.

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

Gradshteyn, I. S. and I. M. Ryzhik (1980). *Tables of Integrals, Series and Products: Corrected and Enlarged Edition.* Academic Press.

Press, S. J. (1982). *Applied multivariate analysis: using Bayesian and frequentist methods of inference* (2nd ed.). Krieger Publishing Company, Malabar, Florida.

Tjur, T. (1991). Block designs and electrical networks. *Annals of Statistics 19*, 1010–1027.