Block designs and statistics

Notes for Math 447

May 3, 2011

The main parameters of a block design are

- number of varieties v,
- block size k,
- number of blocks b.

A design is built on a set of v elements. Each element is called a *variety* or *treatment*. Fix $k \geq 2$. Consider subsets of the set of the varieties with k elements. The design is specified by specifying certain of these subsets. A *block* is one of the specified subsets. Thus the design is a multiset of k-element subsets of the set of varieties. There are b blocks in this multiset.

Each variety i occurs in r_i blocks. This number r_i is called the *replication* of i. Each pair of varieties i, j with $i \neq j$ occurs in λ_{ij} blocks. The idea of a balanced design is to attempt to take λ_{ij} to be constant. The present discussion explores optimality for parameter values when there is no balanced design. In summary, the other important parameters for a block design are

- number of blocks to which variety i belongs: r_i,
- number of blocks to which varieties $i \neq j$ belong together: λ_{ij}.

A variety i with a block to which it belongs is called an experimental *unit*. One wants to compare varieties within a block, that is, to consider each block in turn and examine what happens with the corresponding units. The block design is given by a v by b *incidence matrix* A. Here A_{im} is 1 if variety i belongs to block m, otherwise $A_{im} = 0$. So the 1 entries correspond to the experimental units.

The number of varieties in each block

$$\sum_{i=1}^{v} A_{im} = k \tag{1}$$

is constant. The number of replications of a variety is

$$\sum_{m=1}^{b} A_{im} = r_i. \tag{2}$$

The first important identity is obtained by counting the units. It says that

$$\sum_{i=1}^{v} r_i = kb. \tag{3}$$

Form the symmetric v by v *concurrence matrix* $\Lambda = AA^T$. Thus

$$\Lambda_{ij} = \sum_{m=1}^{b} A_{im}A_{jm}. \tag{4}$$

For $i \neq j$ this is the number of times λ_{ij} that the pair i, j occurs in a block. We have

$$\Lambda_{ii} = \sum_{m=1}^{b} A_{im}^2 = \sum_{m=1}^{b} A_{im} = r_i. \tag{5}$$

Also

$$\sum_{i=1}^{v} \Lambda_{ij} = kr_j \tag{6}$$

and

$$\sum_{j=1}^{v} \Lambda_{ij} = kr_i. \tag{7}$$

This leads to the second important identity

$$\sum_{j \neq i} \lambda_{ij} = (k-1)r_i. \tag{8}$$

Define the *concurrence multigraph* to be the multigraph with v vertices and with λ_{ij} edges between each $i \neq j$. The degree of vertex i is the number of edges attached to vertex i. Thus it is given by

$$d(i) = \sum_{j \neq i} \lambda_{ij} = (k-1)r_i. \tag{9}$$

For each multigraph there is a corresponding symmetric *Laplacian matrix* L. Its diagonal elements are $L_{ii} = d(i)$. Its off diagonal elements are $L_{ij} = -\lambda_{ij}$. Thus we have

$$\sum_{i=1}^{v} L_{ij} = 0. \tag{10}$$

and

$$\sum_{j=1}^{v} L_{ij} = 0. \tag{11}$$

Furthermore, the trace of L is

$$\operatorname{tr} L = (k-1)\sum_{i=1}^{v} r_i = (k-1)kb. \tag{12}$$

Sometimes it is convenient to write $L = kR - \Lambda$, where R is the diagonal matrix with diagonal entries r_i and where Λ is the concurrence matrix.

It is convenient to introduce another matrix that is a multiple of the Laplacian matrix. This is the *information matrix* C defined by $C = \frac{1}{k}L$. We have that

$$\sum_{i=1}^{v} C_{ij} = 0 \tag{13}$$

and

$$\sum_{j=1}^{v} C_{ij} = 0. \tag{14}$$

Furthermore,

$$\operatorname{tr} C = (k-1)b. \tag{15}$$

In the following we will use the normalized trace

$$\langle C \rangle = \frac{1}{v-1} \operatorname{tr} C = \frac{(k-1)b}{v-1}. \tag{16}$$

The information matrix C may be written directly in terms of the concurrence matrix Λ by the formula

$$C = R - \frac{1}{k}\Lambda. \tag{17}$$

Here the entry $\Lambda_{ii} = r_i$ is the number of times that the variety i occurs in a block, while for $i \neq j$ the entry $\Lambda_{ij} = \lambda_{ij}$ is the number of times that varieties i, j occur in the same block. It follows that the entry $C_{ii} = (k-1)r_i/k$ and for $i \neq j$ the entry $C_{ij} = -\lambda_{ij}/k$.

The matrices L and $C = \frac{1}{k}L = R - \frac{1}{k}\Lambda$ are both symmetric matrices whose row and column sums are zero. This implies there is a constant eigenvector with eigenvalue 0. In the following we suppose that the concurrence multigraph is connected, so that this eigenvector has multiplicity one. The matrix P with constant entries $1/v$ is the orthogonal projection onto this space of constant vectors. It follows that $I - P$ is the orthogonal projection onto the space of vectors whose entries sum to zero.

Let C^- be the symmetric matrix with row and column sums that sum to zero, and such that

$$C^- C = C C^- = I - P. \tag{18}$$

Thus C^- is the inverse to C on the $v-1$ dimensional space of vectors whose entries sum to zero, which is the range of $I - P$. We shall see that it is reasonable to call C^- the *covariance matrix*.

The model is that for each unit with variety i in block m there is an associated number

$$\mu_{im} = \tau_i + \beta_m. \tag{19}$$

These parameters are the *treatment effect* τ_i and the *block effect* β_m. They are unknown. The goal of the experiment is to use experimental observations to

estimate quantities associated with the treatment effect. Let x be a *contrast vector* with $x_1 + x_2 + \cdots + x_v = 0$. The quantity to estimate is the *contrast* $x^T\tau = x_1\tau_1 + x_2\tau_2 + \cdots + x_v\tau_v$. The block effect is of no interest; the only reason for introducing the blocks is to eliminate systematic error in distinguishing the effect of different treatments.

Assume that for each unit with variety i in block m there is a random variable Y_{im} with mean $\mu_{im} = \tau_i + \beta_m$ and variance σ^2. These random variables are independent. These are the experimental observations. Each time an experiment is done to measure values of the random variables the results obtained are different. The goal is to estimate the contrast $x^T\tau$ by some function of the random variables Y_{im} associated with the units.

A naive approach would be to estimate the effect of the ith treatment by the sum $\sum_{\{m:i\in m\}} Y_{im}$ divided by r_i. This is the sum over blocks containing i of the observations, divided by the number of such blocks. The mean of this is $\tau_i + (1/r_i)\sum_{\{m:i\in m\}} \beta_m$. If the block effects are big, then this gives a terrible idea of the parameter τ_i. The other possibility is that the block effects are small, but then one would not even bother with a block design.

Here is a better way to approach the problem. For each variety i define the deviation sum

$$D_i = \sum_{\{m:i\in m\}} \left(Y_{im} - \frac{1}{k} \sum_{\{j:j\in m\}} Y_{jm} \right) = \sum_m A_{im} Y_{im} - \frac{1}{k} \sum_m \sum_j A_{im} A_{jm} Y_{jm}. \tag{20}$$

This is the sum over blocks containing i of the deviation of the observation in the block for variety i from the block sample mean. The goal is to subtract out the block effects, but this attempt involves other varieties. The result is that the mean of D_i is

$$\sum_m A_{im}(\tau_i+\beta_m) - \frac{1}{k} \sum_m \sum_j A_{im} A_{jm} (\tau_j+\beta_m) = \sum_m A_{im}\tau_i - \frac{1}{k} \sum_m \sum_j A_{im} A_{jm} \tau_j. \tag{21}$$

It is independent of the block parameters! The price is that it involves all the treatment parameters, but this difficulty will be overcome. The mean of D_i simplifies to

$$r_i\tau_i - \sum_j \frac{1}{k} \Lambda_{ij} \tau_j = \sum_j C_{ij} \tau_j. \tag{22}$$

In other words, the random vector D has mean $C\tau$. It follows that the mean of the random variable $x^T C^- D$ is the contrast

$$x^T C^- C \tau = x^T \tau. \tag{23}$$

It is also not difficult to compute the covariance of D_i with D_j. A routine calculation gives the result to be

$$\left(\sum_m A_{im}\delta_{ij} - \frac{1}{k} \sum_m A_{im} A_{jm} \right) \sigma^2 = \left(r_i \delta_{ij} - \frac{1}{k} \Lambda_{ij} \right) \sigma^2 = C_{ij} \sigma^2. \tag{24}$$

Here $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. This is the Kronecker delta symbol for the identity matrix. It follows that for a contrast vector z the variance of $z^T D = \sum_i z_i D_i$ is

$$(z^T C z)\sigma^2 = \left(\sum_{i=1}^{v} \sum_{j=1}^{v} z_i C_{ij} z_j \right) \sigma^2. \tag{25}$$

Take the contrast vector x with $Cx = z$, or $x = C^- z$. The final result is that the estimator $x^T C^- D$ has mean $x^T \tau$ and variance

$$(x^T C^- x)\sigma^2 = \left(\sum_{i=1}^{v} \sum_{j=1}^{v} x_i C^-_{ij} x_j \right) \sigma^2. \tag{26}$$

The striking thing is that the result for the variance is so simple.

In conclusion, the model for the experiment to compare treatments has three parameters:

- treatment effect vector τ,
- block effect vector β,
- variance of individual observation σ^2.

The deviation vector D is built from the individual observations. The covariance matrix C^- is computed from $C = R - \frac{1}{k}\Lambda$. If x is a contrast vector, then

- $x^T C^- D$ estimates the contrast $x^T \tau$;
- $x^T C^- D$ has variance $(x^T C^- x)\sigma^2$.

Small variance is good; it means that the quantity $x^T C^- D$ computed from the observations is likely to be close to the contrast $x^T \tau$ that one is trying to estimate.

Fix two different varieties i, j to be compared. The corresponding pairwise contrast vector is the vector x with $x_i = 1$ and $x_j = -1$ and all other components zero. Then $x^T \tau = \tau_i - \tau_j$ is the contrast between the mean responses from the two corresponding treatments. The variance for the corresponding estimator $x^T C^- D = \sum_p C^-_{ip} D_p - \sum_q C^-_{jq} D_q$ is

$$V_{ij} = \left(C^-_{ii} + C^-_{jj} - 2C^-_{ij} \right) \sigma^2. \tag{27}$$

We can compare each variety i to each of the other j. The total variance is

$$\sum_{j \neq i} V_{ij} = \left((v-1)C^-_{ii} + (\operatorname{tr} C^- - C^-_{ii}) + 2C^-_{ii} \right) \sigma^2 = (\operatorname{tr} C^- + v C^-_{ii})\sigma^2. \tag{28}$$

The average value of these pairwise variances is

$$\bar{V} = \frac{1}{v(v-1)} \sum_i \sum_{j \neq i} V_{ij} = 2\sigma^2 \frac{\operatorname{tr} C^-}{v-1}. \tag{29}$$

A design is good if this *average pairwise covariance* is small.

In summary, to get a good design, compute the Laplacian matrix L and the corresponding information matrix $C = \frac{1}{k}L = R - \frac{1}{k}\Lambda$. Then compute the covariance matrix C^-. One way to do this is to use the formula

$$C^- = (C + P)^{-1} - P. \tag{30}$$

Finally, compute the normalized trace

$$\langle C^- \rangle = \frac{1}{v-1} \operatorname{tr} C^-. \tag{31}$$

Try to choose the design so that this is small.

What is small? Here it is useful to use the inequality of geometric and harmonic mean, applied to the covariance matrix C^-. It says that

$$\frac{1}{\langle C \rangle} \leq \langle C^- \rangle. \tag{32}$$

In other words,

$$\frac{v-1}{(k-1)b} \leq \langle C^- \rangle. \tag{33}$$

Furthermore, the strong form of the inequality says that there is equality only when C is a constant multiple of $I - P$. In other words,

$$C = \frac{(k-1)b}{v-1}(I - P). \tag{34}$$

This is the balanced case. The matrix $I - P$ has diagonal entries $1 - 1/v = (v-1)/v$ and off-diagonal entries $-1/v$. So C has diagonal entries $(k-1)b/v$ and off-diagonal entries $-(b/v)(k-1)/(v-1)$. Thus $L = kC$ has diagonal entries $(kb/v)(k-1)$ and off-diagonal entries $-(kb/v)(k-1)/(v-1)$. Each replication is $r = kb/v$. In the concurrence multigraph each vertex is of degree $(k-1)r$ and the number of edges between two vertices is $\lambda = r(k-1)/(v-1)$.

From this we can compute the covariance matrix in the balanced case

$$C^- = \frac{v-1}{(k-1)b}(I - P). \tag{35}$$

The optimum value of the averaged pairwise variance is

$$\bar{V} = 2\sigma^2 \frac{v-1}{(k-1)b}. \tag{36}$$

It is difficult to estimate when you have lots of varieties. However, it is good to have big blocks, and lots of them.

In general one wants to take a design that is optimal in the following sense. Suppose the design has the property that a pair i, j with $i \neq j$ occurs in λ_{ij} blocks. Define the Laplacian matrix to be the symmetric matrix with off-diagonal entries $-\lambda_{ij}$ for $i \neq j$ and with row and column sums equal to zero. Let

$C = \frac{1}{k}L$ be the corresponding information matrix. Let C^- be the covariance matrix that inverts the information matrix C on the subspace of vectors that sum to zero. For this design the average pairwise variance is

$$\bar{V} = 2\sigma^2 \frac{\operatorname{tr} C^-}{v-1}, \tag{37}$$

Choose the design that minimizes this quantity.

Here is an example to illustrate these ideas. Take the cyclic block design with $v = 7$ varieties. Use addition modulo 7. First take a starter block with elements 0, 1, 3. Keep adding 1,1,1 to form $b = 7$ blocks each of size $k = 3$. This gives a balanced design. The matrix L has diagonal entries 6 and off-diagonal entries -1. It is easy to compute the information matrix $C = \frac{1}{3}L = \frac{7}{3}(I - P)$ and then the covariance matrix $C^- = \frac{3}{7}(I - P)$. In this case $I - P$ has diagonal elements $6/7$ and off-diagonal elements $-1/7$; in particular its normalized trace is 1. The normalized trace of the covariance matrix C^- is thus $\frac{3}{7} = 0.4286$. This is the optimum design.

Instead take a starter block with elements 0, 1, 4. Keep adding 1,1,1 to form $b = 7$ blocks each of size $k = 3$. This is not a balanced design. In fact, its Laplacian matrix is given by the following:

$$L = \begin{bmatrix} 6 & -1 & 0 & -2 & -2 & 0 & -1 \\ -1 & 6 & -1 & 0 & -2 & -2 & 0 \\ 0 & -1 & 6 & -1 & 0 & -2 & -2 \\ -2 & 0 & -1 & 6 & -1 & 0 & -2 \\ -2 & -2 & 0 & -1 & 6 & -1 & 0 \\ 0 & -2 & -2 & 0 & -1 & 6 & -1 \\ -1 & 0 & -2 & -2 & 0 & -1 & 6 \end{bmatrix} \tag{38}$$

Again define the information matrix $C = \frac{1}{3}L$. With a computer it is not hard to find that the covariance matrix C^- is given by the following:

$$C^- = \begin{bmatrix} 0.4181 & -0.0732 & -0.1568 & 0.0209 & 0.0209 & -0.1568 & -0.0732 \\ -0.0732 & 0.4181 & -0.0732 & -0.1568 & 0.0209 & 0.0209 & -0.1568 \\ -0.1568 & -0.0732 & 0.4181 & -0.0732 & -0.1568 & 0.0209 & 0.0209 \\ 0.0209 & -0.1568 & -0.0732 & 0.4181 & -0.0732 & -0.1568 & 0.0209 \\ 0.0209 & 0.0209 & -0.1568 & -0.0732 & 0.4181 & -0.0732 & -0.1568 \\ -0.1568 & 0.0209 & 0.0209 & -0.1568 & -0.0732 & 0.4181 & -0.0732 \\ -0.0732 & -0.1568 & 0.0209 & 0.0209 & -0.1568 & -0.0732 & 0.4181 \end{bmatrix} \tag{39}$$

The normalized trace of the covariance matrix is then the sum of the entries, which is 7 times 0.4181, divided by 6. This gives 0.4878. This is indeed larger than the number 0.4286 for the normalized trace from the optimal design.

In the above example the correct procedure is to use the balanced block design, since it has the smallest average pairwise covariance of any design with the same values of v, b, k. However for most values of v, b, k there will not be a balanced design. This is because if there were a balanced design with $\lambda_{ij} = \lambda$,

then we would have $(v-1)\lambda = (k-1)r_i$. In particular, it would follow that $r_i = r$ is independent of i. Also, we would have $vr = kb$. However there is no particular reason to hope that r and λ will be integers. So in general one must somehow try out various block designs that are not balanced, until one finds one with the smallest average pairwise covariance.

This treatment above is a simplification (perhaps an oversimplification) of the discussion in a long paper by R. A. Bailey and Peter J. Cameron called "Combinatorics of optimal designs." Bailey and Cameron use a more general framework where the entries in the incidence matrix are natural numbers. This means that the blocks are multisets rather than sets. For simplicity this is not done in the present account.

Also, these authors explain that there may be other quantities that one may try to minimize. Here the emphasis has been to choose the design to minimize the average variance of the pairwise contrasts, that is, to minimize the average pairwise variance $2\sigma^2\langle C^- \rangle$. One alternative, for instance, would be to choose the design to minimize the largest possible value of $\sigma^2 x^T C^- x$ over all contrast vectors x with normalization $x^T x = 1$. This would be the approach of a pessimist who worries about the worst case contrast vector that one could encounter in practice. In the situation when there is a balanced design it is still the best design, but in other cases this could give a different optimal design.