Lecture #10 Notes Summary

Block Designs

Block designs

Block designs denotes a particular class of combinatorial optimal experimental design problems, where the effects of different *treatments* are to be compared, by testing them in several *plots* that are arranged in *blocks*. This terminology comes historically from field experiments in agriculture, which was at the origin of the development of the theory of optimal experimental designs. Here, the blocks represent a portion of a field where a some treatments can be compared. When designing the experiment, we must keep in mind that the block itself might have an effect on the observations. Indeed the conditions of sunshine and shadow, draining of rainwater, or the quality of earth might differ from block to block. Note that all plots might not be equivalent inside one block, but this is usually handled by *randomization*.

Block designs have a very wide application spectrum, as shown by the following examples, (taken from a lecture of R.A. Bailey and P. Cameron):

- *Does red wine protect against heart diseases?* To investigate this, you have 40 volunteers who participate to the experiments for a period of 30 days. Each volunteer is assigned to one of these "treatments" for the first 15 days, and then to another "treatment" for the last 15 days:
  - Drink no alcohol at all
  - Drink one glass of red wine per day
  - Drink two glasses of red wine per day
  - Drink two glasses of grape juice per day
  - Drink two standard drinks of whiskey per day

  The scientist take a blood sample of each volunteer after Day 15 and Day 30, to measure the amount of inflammatory substances.

  In this example, 5 treatments are compared, and there are 40 blocks of size 2 (the patients). #1
- *Consumer experiments* 12 housewives volunteer to test 16 new detergents. It is not possible to ask each participant to test the 16 types of detergent, so each volunteer tests only 4 detergents over 4 different washloads, and gives a note to assess the quality of each detergent, judging the cleanliness of each washload.

  Here, there are 16 treatments to compare, over 12 blocks of size 4.
- A biologist measures the rate of diffusion of proteins by cells of *Escherichia Coli*. She tests the effect of adding 0,1,2,3, or 4 green fluorescent proteins to the cells. She can prepare 10 samples a day and works on it one week (that is, 5 days from Monday to Friday).

  Here the number of treatments is 5. We can regard the days of the week as blocks, because 1) the biologist might have got better at preparing the samples as the week goes on, and 2) maybe there have been other environmental changes in the lab. So there are 5 blocks of size 10.

Notation and definitions We have t treatments to compare, over b blocks of size k. If the treatments are numbered from 1 to t, a design is represented in an array as follows, with *block in columns*. For example,

1	1	2	3	4	5
2	4	5	6	6	8
3	7	8	9	7	9

represents the design where the treatments $(1, 2, 3)$ are tested in the first block, the treatments $(1, 4, 7)$ are tested in the second block, etc.

Definition 1. The *replication* r_i of a treatment i is the total number of occurences of i in the design. A design is called *equireplicate* if there is a $r \in \mathbb{N}$ such that $r_i = r$ for all $i \in \{1, \dots, t\}$.

A design is called *binary* if no treatment occurs more than once in a block.

The *concurrence* $\lambda_{i,j}$ is the number of times that treatments i and j occur in the same block. A binary design is called *balanced* if there is a $\lambda \in \mathbb{N}$ such that $\lambda_{i,j} = \lambda$, for all $i \neq j$ in $\{1, \dots, t\}$.

When $k < t$, the blocks are called *incomplete* (because it is not possible to test all treatments in a single block). In this case, a balanced design is also called *balanced incomplete-block design* (BIBD).

BIBDs are very important; indeed, we will see in the next lecture that they are $D-$, $E-$, and $A-$optimal. The next theorem gives a simple necessary condition for the existence of a balanced design:

Theorem 1. *If a binary design is balanced with concurrence λ, then the design is equireplicate, with replication r satisfying*

$$r(k-1) = \lambda(t-1).$$

Hence, a necessary condition for a block design with parameters k, b, t to be balanced is:

$$t \text{ divides } bk \qquad \text{and} \qquad t(t-1) \text{ divides } bk(k-1).$$

Proof. Treatment i is in r_i blocks. In each of these blocks, i is in concurrence with $k-1$ other treatments. So the sum of the concurrences with treatment i is

$$\sum_{j \neq i} \lambda_{i,j} = r_i(k-1) = (t-1)\lambda.$$

This proves the formula and the equireplicatedness of the design.

The necessary condition just expresses that the replication number $r = \frac{bk}{t}$ must be an integer, as well as the concurrence number $\lambda = r\frac{k-1}{t-1} = b\frac{k(k-1)}{t(t-1)}$, □

In exercises, we will see another necessary condition, somehow more surprising: a BIBD can exist only if there are more blocks than treatments ($b \geq v$). This limits the use of BIBDs, since very often we rather have $b \leq v$ for cost reasons.

Theorem 2 (Fisher's inequality)**.** *In a BIBD, we must have $b \geq t$.*

Blocking: statistical model

We consider a block design with t treatments and b blocks of size k. In total, there are bk plots (experimental units). We assume that an observation of treatment i in block j gives a unbiased measure of the sum of the treatment effect τ_i and the block effect β_j: $\mathbb{E}[y] = \tau_i + \beta_j$; as usual $\mathrm{Var}(y) = \sigma^2$ and the different observations are uncorrelated. In matrix notation:

$$\boldsymbol{y} = X\boldsymbol{\tau} + Z\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where X is the $bk \times t$ incidence matrix between the plots and the treatments, Z is the $bk \times b$ incidence matrix between the plots and the blocks, and $\boldsymbol{\epsilon}$ is a centered noise with a diagonal variance: $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 I$.

Usually, we are not interested to estimate the block effects $\boldsymbol{\beta}$, we just want to estimate the treatment effects. However, it is not possible to estimate $\boldsymbol{\tau}$ in this model. Indeed, for any constant $c \in \mathbb{R}$ we could replace τ_i by $\tau_i + c$ and β_j by $\beta_j - c$ without changing the vector of measurements $\boldsymbol{y}$. So the best we can hope is to estimate the τ_i up to some constant. In other words, we can estimate the differences of treatment effects $\tau_i - \tau_j$. More precisely, it is possible to obtain an unbiased estimate of all quantities of the form $\boldsymbol{x}^T\boldsymbol{\tau}$, where $\boldsymbol{x} \in \mathbb{R}^t$ is a vector of *contrasts*, i.e., $\sum_{i=1}^t x_i = 0$.

Definition 2. Define $R = \mathrm{Diag}(r_1, \dots, r_t) \in \mathbb{S}^t$, and let $\Lambda \in S^t$ be the matrix with i^{th} diagonal elements r_i and (i,j) off-diagonal elements $\lambda_{i,j}$.

The matrix $C := R - \frac{1}{k}\Lambda \in \mathbb{S}^t$ is called the *information matrix* of the design.

The matrix $L = kC = kR - \Lambda$ is called the *Laplacian matrix* of the design.

The reason why C is called *information matrix* will become clear with the next theorem. The reason why L is called *Laplacian matrix* will become clear in the next lecture.

Theorem 3. *Let $\boldsymbol{x}$ be a vector of contrasts ($\sum_{i=1}^t x_i = 0$). Define $Q = I - \frac{1}{k}ZZ^T \in \mathbb{S}^{bk}$. Then, the best linear unbiased estimator of $\boldsymbol{x}^T\boldsymbol{\tau}$ is $\boldsymbol{x}^T C^\dagger X^T Q \boldsymbol{y}$, and its variance is equal to $(\boldsymbol{x}^T C^- \boldsymbol{x})\sigma^2$.*

Proof. This result is, without surprise, a consequence of Gauss Markov theorem. We skip the proof of the formula for the best linear estimator, and we focus on the expression of the optimal variance σ^*. In the model

$$\boldsymbol{y} = [X, Z] \begin{bmatrix} \boldsymbol{\tau} \\ \boldsymbol{\beta} \end{bmatrix} + \boldsymbol{\epsilon},$$

it is not difficult to check that the fact that $\boldsymbol{x}$ is a vector of contrasts implie that the estimability condition $[\boldsymbol{x}^T, \mathbf{0}^T]^T \in \mathrm{im}[X, Z]^T$ holds. So Gauss Markov theorem gives the following expression for the minimal variance:

$$\sigma^* = \sigma^2 [\boldsymbol{x}^T, \mathbf{0}^T] \begin{pmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ \end{pmatrix}^- \begin{bmatrix} \boldsymbol{x} \\ \mathbf{0} \end{bmatrix}.$$

In other words, $\sigma^* = \sigma^2\ \boldsymbol{x}^T\boldsymbol{a}$, for any vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ that solve the equation

$$\begin{pmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ \end{pmatrix} \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x} \\ \mathbf{0} \end{bmatrix}. \tag{1}$$

Now, observe that $X^TX = R$, $Z^TZ = kI$, and $N := X^TZ$ is the incidence matrix between the blocks and the treatments (that is, $N_{i,j}$ indicates how many times treatment i occurs in block j), so that the concurrence matrix is $\Lambda = NN^T$ (we will verify this in exercises). Equation (1) rewrites:

$$\begin{cases} R\boldsymbol{a} & +N\boldsymbol{b} & = \boldsymbol{x} \\ N^T\boldsymbol{a} & +k\boldsymbol{b} & = \mathbf{0} \end{cases}$$

In the second equation, we find $\boldsymbol{b} = -\frac{1}{k}N^T\boldsymbol{a}$, and substituting in the first equation yields $(R - \frac{1}{k}NN^T)\boldsymbol{a} = \boldsymbol{x}$. This shows that $\sigma^* = \sigma^2\ \boldsymbol{x}^T\boldsymbol{a} = \sigma^2\ \boldsymbol{x}^T C^- \boldsymbol{x}$. □

Exercises

1. Consider a design with plot to treatment incidence matrix X and plot to block indidence matrix Z. Let $R = \mathrm{Diag}(r_i)$ and $\Lambda = (r_i\delta_i^j + \lambda_{i,j}(1-\delta_i^j))_{1\leq i,j\leq t}$, where δ_i^j is the Kronecker product.

   Show that

   - $X^TX = R$
   - $Z^TZ = kI$
   - $N := X^TZ$ is the treatment to block incidence matrix
   - $NN^T = \Lambda$.

2. Consider a BIBD with t treatments on b blocks of size k, replication r and concurrence λ.

   - What are the elements of Λ ? What are its eigenvalues ?
   - Use the fact that $\mathrm{rank}\, N = \mathrm{rank}\, NN^T$ to prove Fisher's inequality.

3. In this exercise we show some constructions to find BIBDs

   (i) What is the smallest value of t such that a BIBD with blocks of size $k = 3$ and concurrence $\lambda = 1$ can exist ? What must be its number b of blocks? Try to construct a BIBD with these values of b and t.

   (ii) Construct a BIBD with $t = 9, b = 12, k = 3$: what must be the value of λ? Try to find appropriate blocks by arranging the numbers from 1 to 9 in a 3×3 square.

   (iii) What is the smallest value of t such that a BIBD with blocks of size $k = 4$ and concurrence $\lambda = 1$ can exist ? Extend the design of question (ii) to find a BIBD with these values of b and t.