The Effect of an Incomplete Block Design on Consumer Segmentation

Ryan Browne[1], Paul McNicholas[1],
John Castura[2], Chris Findlay[2]

1 University of Guelph, Guelph, Ontario, Canada
2 Compusense, Guelph, Ontario, Canada

July 20, 2010

An Experiment

- Observe a person's response to 12 different products (randomized block design)
- However for wine and other alcohol beverages products it is difficult to obtain an individual response to several products because of intoxication, carry-over, adaption and fatigue.
- To compensate use balanced incomplete block designs.
- The goal is to determine if there is any clusters or grouping within the data.

Complete Block Design

- 2 blocks and 3 treatments

$$C = \left(\begin{array}{cc|ccc} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array}\right) = \left(\begin{array}{c|c} \mathbf{I}_2 \bigotimes \mathbf{1}_3 & \mathbf{1}_2 \bigotimes \mathbf{I}_2 \end{array}\right)$$

- k blocks and t treatments

$$C = \left(\begin{array}{c|c} \mathbf{I}_k \bigotimes \mathbf{1}_t & \mathbf{1}_k \bigotimes \mathbf{I}_t \end{array}\right)$$

Incomplete Block Design

- 3 treatments and 2 treatments per block

$$D = \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array} \right) = \left(\begin{array}{c|c} \mathbf{I}_3 \bigotimes \mathbf{1}_2 & B \end{array} \right)$$

- t treatments, s treatments per block, k repetitions of the design and let $n = \binom{t}{s}$.

$$D = \left(\begin{array}{c|c} \mathbf{I}_k \bigotimes \mathbf{I}_n \bigotimes \mathbf{1}_s & \mathbf{1}_k \bigotimes B \end{array} \right)$$

Clustering

We examine the effect of an incomplete block design on clustering,

- Block effects, the average response to the products and
- Treatments effects, the vectors of responses.

Block Effects

- In a complete block design, we take the average response from an individual i, (assuming normality) then average response would have

$$\overline{Y}_{i.} \backsim N(\delta_i + \overline{\mu}, \sigma_e^2)$$

- In an incomplete block design, we take the estimated quantities

$$\hat{\delta}_i + \frac{1}{t}\left(\hat{\mu}_1 + \ldots + \hat{\mu}_t\right)$$

- What are the distributional properties of these block estimators?

Distributional Properties of the Block Estimators

- Assuming normality, the distribution of the regression coefficients is

$$\widetilde{\beta} = N\left(\beta, \sigma_e^2(\mathbf{X}^t\mathbf{X})^{-1}\right)$$

- For the incomplete block design, to obtain the estimators $\widetilde{\delta}_i + \frac{1}{t}\sum_{i=1}^{t}\widetilde{\mu}_i$ we need to multiply D by A^t.

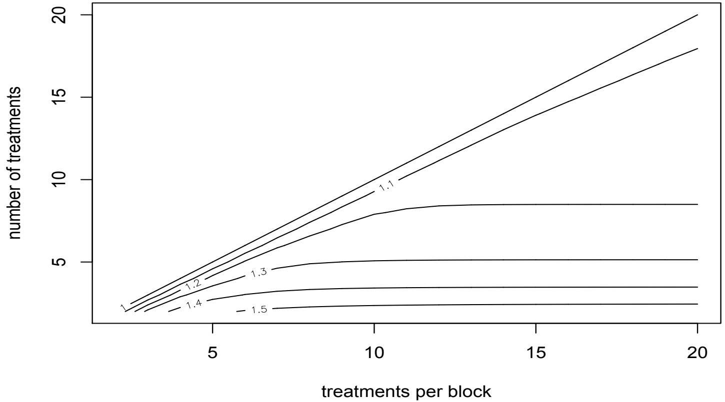
$$A = \left(\begin{array}{c|c} \mathbf{I}_n & \mathbf{J}_t/t \end{array}\right)$$

where $\mathbf{I}_n$ is n dimensional identity matrix and $\mathbf{J}_t$ is an $t \times t$ matrix of ones.

Distributional Properties of the Block Estimators

$$
\begin{aligned}
A &= \left[\ \mathbf{I}_n \mid \mathbf{J}_t/t\ \right] \\
D &= \left[\ \mathbf{I}_k \bigotimes \mathbf{I}_n \bigotimes \mathbf{1}_s \mid \mathbf{1}_k \bigotimes B\ \right] \\
\mathbf{BA}^t &= \left(\mathbf{I}_k \bigotimes \mathbf{I}_n + \mathbf{J}_k \bigotimes \mathbf{J}_n/t\right) \bigotimes \mathbf{1}_k \\
(\mathbf{BA}^t)^t\mathbf{BA}^t &= \left(\mathbf{I}_k \bigotimes \mathbf{I}_n + \mathbf{J}_k \bigotimes \mathbf{J}_n \left[2/t + \binom{t}{s}/t^2\right]\right)/s \\
\left((\mathbf{BA}^t)^t\mathbf{BA}^t\right)^{-1} &= \left(\frac{1}{s}\right)\mathbf{I}_k \bigotimes \mathbf{I}_n + \frac{2t + \binom{t}{s}}{s\left[t^2 s^2 + 2t\binom{t}{s} + \binom{t}{s}^2\right]}\mathbf{J}_k \bigotimes \mathbf{J}_n
\end{aligned}
$$

Ratio of variances from complete and incomplete



Treatment Effects

- In a complete block design, we have the full response vector from each individual i
- In an incomplete block design, we need to deal with the missing values
  - Fill in the missing observations using the fitted values.
- Compare the clustering from incomplete and block design using the Adjusted Rand Index.

Signal to noise

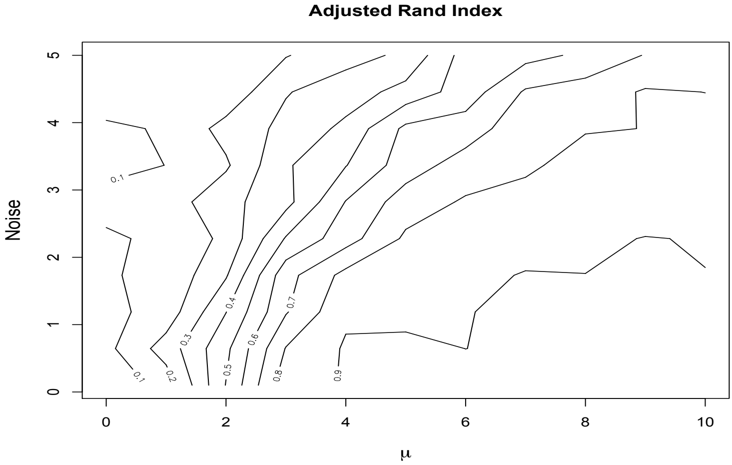
To examine the effects of noise and signal on an incomplete block design having 4 treatments and allowing 2 treatments per block

- Generate 90 observations from 2 clusters with

$$\overline{Y}_1 \backsim N(-\mu \mathbf{1}_4, \sigma^2 \mathbf{I}_4) \quad \overline{Y}_2 \backsim N(\mu \mathbf{1}_4, \sigma^2 \mathbf{I}_4)$$

- $\mu = 0, 1, \ldots, 10$
- $\sigma = 0.1, 0.5, 1, \ldots, 5$
- Cluster using hierarchical clustering with an average linkage. Compare the clustering from incomplete and block design using the Adjusted Rand Index.

Signal to noise



Cluster versus Block Size

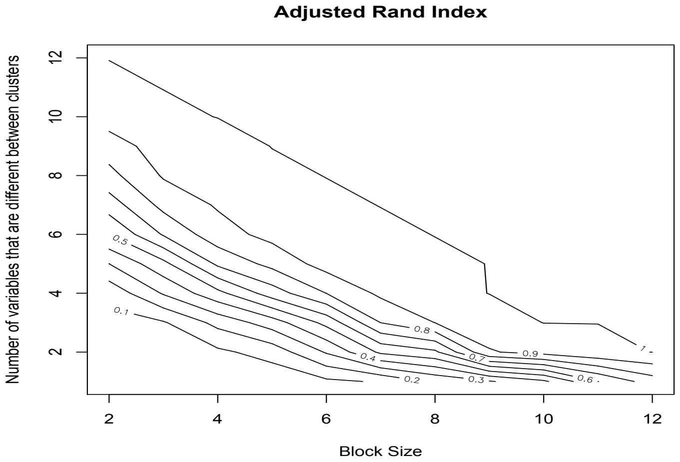
To examine the effect of block size (the number of treatments per block) on the cluster when we have $t = 12$ treatments in total

- Generate 250 observations from 2 clusters with

$$\overline{Y}_1 \backsim N(-\mu \begin{pmatrix} \mathbf{1}_r \\ \mathbf{0}_{12-r} \end{pmatrix}, \sigma^2 \mathbf{I}_{12}) \quad \overline{Y}_2 \backsim N(\mu \begin{pmatrix} \mathbf{1}_r \\ \mathbf{0}_{12-r} \end{pmatrix}, \sigma^2 \mathbf{I}_{12})$$

- Number of variables that differ between clusters $r = 1, \ldots, 12$
- Block size $s = 2, \ldots, 12$
- set $\mu = 5$ and $\sigma^2 = 1$

Cluster versus Block Size



Conclusions

- An incomplete block design is an effective tool to collect data.
- When clustering block effects, the variance is increased.
- When clustering treatment effects, we can only detect clusters which have differences on more than $t - s$ variables.

The end

Thank you.