Design and Analysis of Sensory Informed Incomplete Block Designs

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 12, 2012

Incomplete Block Design

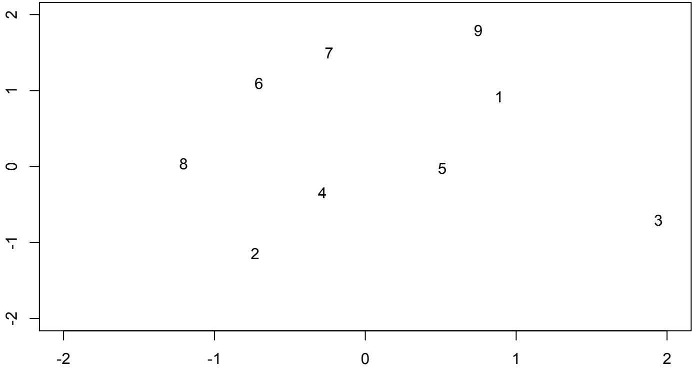
- Observe a person's response to 12 different products (randomized block design)
- However for wine and other alcohol beverages products it is difficult to obtain an individual response to several products because of intoxication, carry-over, adaption and fatigue.
- To compensate use balanced incomplete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

Sensory-Informed Design: Bread Study

Sensory Profile

- 10-13 trained panelists
- Each panelists evaluates the 12 different Bread products on 42 attributes.
- Attributes for crumb
    - Springiness
    - Firmness
    - Moistness
    - Chewiness
    - Particles

Products in the Sensory Space



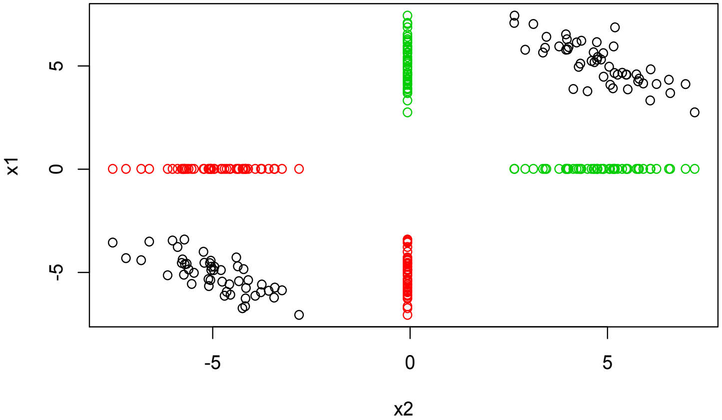
Mixture Modelling

- To find segments in the data, we assume the liking scores arise from a Gaussian mixture

$$Y \backsim \sum_{g=1}^{G} \pi_g f(y|\mu_g, \Sigma_g)$$

- If we had a complete-block design we would just apply standard methodology to this problem.
- The literature commonly suggests imputation or some variation thereof, for incomplete blocks.

Imputation using the Average



Missing Data

- However, it is possible to estimate a covariance matrix when some data are missing.
- We can do this via the expectation-maximization (EM) algorithm.
- This approach is particularly useful when the covariance matrix has a special structure.
- And even more so when

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$$

Covariance Structures

- Mclust models - (Mclust in R)

$$\mathbf{\Sigma}_g = \lambda_g \boldsymbol{D}_g \boldsymbol{A}_g \boldsymbol{D}_g^T$$

- Factor Analyzers - (pgmm package in R)

$$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g \mathbf{\Lambda}_g^T + \mathbf{\Psi}_g$$

Conditional Distribution of Missing Data

- To use EM algorithm we need to calculate the sufficient statistics for the missing data.

$$X_1|X_2 = x_2 \backsim MVN(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

- However, we have to calculate the expected sufficient statistics for the missing data in each row.
- This amounts to m choose k different matrix inverses.
- For the bread data 12 choose 6 =920 matrix inverses.
- Complete E-steps are not computationally feasible.

Incremental E-step or E-Step by column

- If start with the missing data $x_i = (x_{i1}, x_{i2}, NA, NA)$ and fill in the missing data with randomly generated observations.
- For for a particular row we say $\hat{x}_i = (x_{i1}, x_{i2}, \hat{x}_{i3}, \hat{x}_{i4})$. So, now we have a complete dataset.
- Go by column and update each estimated observation via

$$\hat{x}_{i,j} = \mu_j + \Sigma_{j,-j}\Sigma^{-1}_{-j,-j}\left(\hat{x}_{i,-j} - \mu_{-j}\right)$$

- e.g.

$$\hat{x}_3 = \mu_3 + \Sigma_{3,-3}\Sigma^{-1}_{-3,-3}\left(\hat{x}_{-3} - \mu_{-3}\right)$$

  where $\hat{x}_{-3} = (x_{i1}, x_{i2}, \hat{x}_{i4})$
- If we perform this iteratively then

$$(\hat{x}_3, \hat{x}_4) \rightarrow \mu_{(3,4)} + \Sigma_{(3,4),(1,2)}\Sigma^{-1}_{(1,2),(1,2)}\left(x_{(1,2)} - \mu_{(1,2)}\right)$$

EM by column

- If we have the inverse matrix of Σ

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{bmatrix} \quad \text{and} \quad \Sigma^{-1} = \Theta = \begin{bmatrix} \theta_{1,1} & \Theta_{1,-1} \\ \Theta_{-1,1} & \Theta_{-1,-1} \end{bmatrix}$$

$$\frac{1}{\theta_{1,1}} \Theta_{1,-1} = \Sigma_{j,-j} \Sigma_{-j,-j}^{-1}$$

- This result is possible due to a relationship between the Matrix Inverse and Schur Complement of a matrix
- We now have an incremental E-step for the 1^{st} moment.
- We can obtain a similar result for the 2^{nd} moment.

EM

- From Neal and Hinton (1998) the EM can be viewed as minimizing

$$F(N_{\mathbf{z}}, \mathbf{x}_i, \theta) = logL(\mathbf{x}_i|\theta) - D_{\text{KL}}\left(N_{\mathbf{z}}||N_{\mathbf{z}.\mathbf{x}_i}\right)$$

- E-step can be viewed as minimizing the Kullback-Leibler (KL) divergence between the missing data distribution and the conditional distribution of the missing data given the observed data.

Application

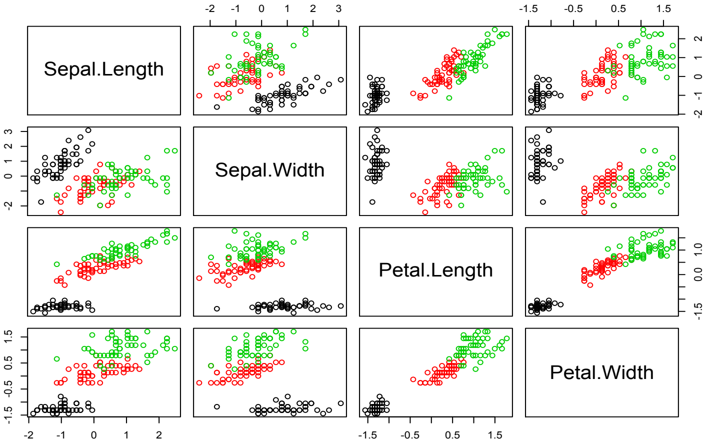
- Iris Dataset.
- Bread Dataset.

Iris Dataset

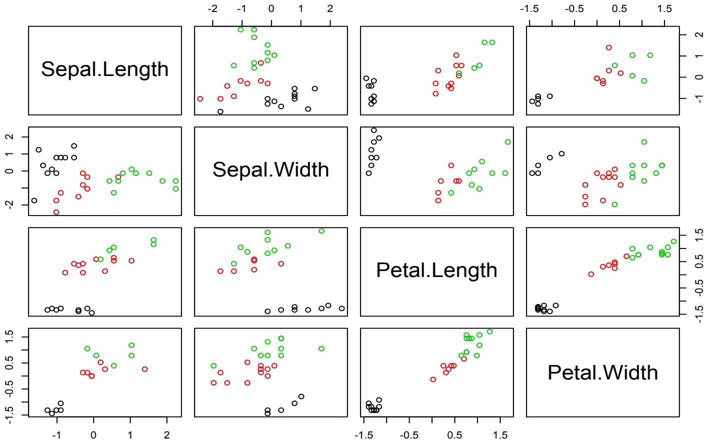
- One of the most famous data sets in statistics.
- Four measurements on three types of flowers.
- We standardized the data and for each observation we randomly removed two measurements.

SepalLength	*Sepal.Width*	*Petal.Length*	*Petal.Width*
		−1.34	−1.31
−1.14		−1.34	
	0.33		−1.31
−1.50	0.01		
		−1.34	−1.31
−0.54			−1.05
⋮	⋮	⋮	⋮

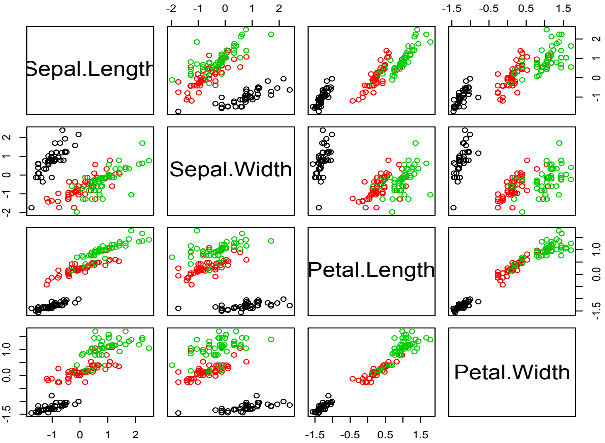
Original Iris Dataset



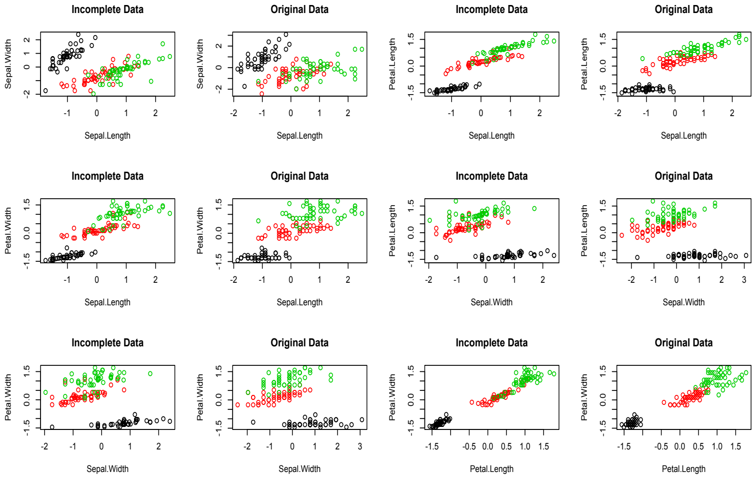
Incomplete Iris Data



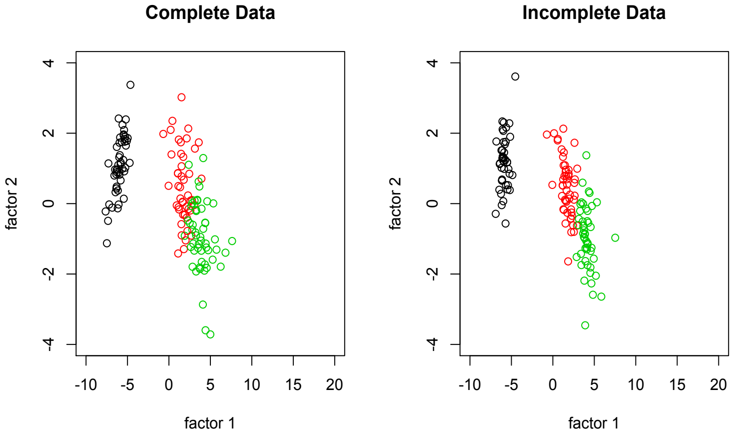
Incomplete Iris Data with Imputed Values



Comparison - Imputed Incomplete and Original Data



Comparison of the Latent Space - Iris Data



Clustering Comparision

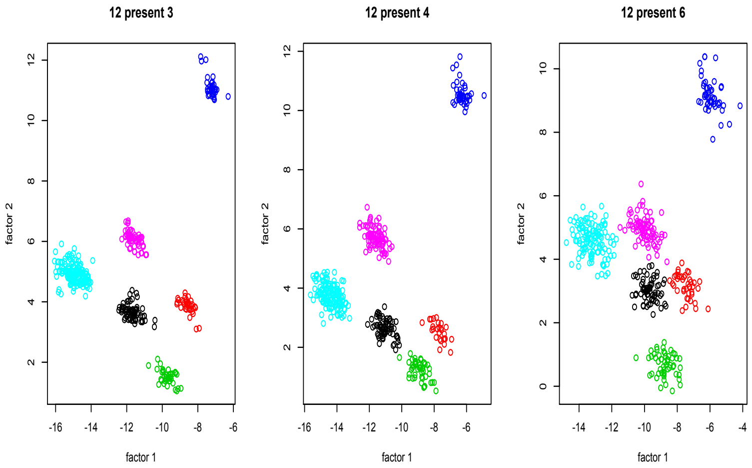
Comparison of clustering results from using the incomplete iris data and the iris data.

	1	2	3
1	50	0	0
2	0	47	0
3	0	3	50

Data

- 420 consumers.
- 12 white breads.
- Each individual evaluated 6 breads within a sensory informed incomplete block design.
- Present-3 and present-4 designs were nested within the present-6 design.
- Six groups and two factors were chosen using the Bayesian Information Criterion (BIC).

Latent Space - Bread Liking Scores



Conclusions

- We can find MLEs using incremental EM.
- We can obtain a reasonable estimate of the latent space using only incomplete data.
- This methodology can be used for imputation.

The end

Thank you.