AG DANK/BCS Meeting 2013 in London
University College London, 8/9 November 2013

MODELS FOR SIMULTANEOUS CLASSIFICATION AND REDUCTION OF THREE-WAY DATA

Roberto Rocci
University "Tor Vergata", Rome

A general classification model: Gaussian Mixtures

Let $\mathbf{x}=[x_1, x_2, \ldots, x_J]'$ be a random vector of J variables. We assume

$$f(\mathbf{x}) = \sum_{g=1}^{G} p_g \phi_g(\mathbf{x}), \qquad p_q \geq 0, \sum_{g=1}^{G} p_g = 1$$ **mixture model**

where each component represents an underlying group, in our case

$$\phi_g(\mathbf{x}) = (2\pi)^{-\frac{J}{2}} \left|\mathbf{\Sigma}_g\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\mathbf{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\}$$ **Gaussian**

and each observation is assigned to a group by computing

$$p(g \mid \mathbf{x}) = \frac{p_g \phi_g(\mathbf{x})}{\sum_h p_h \phi_h(\mathbf{x})}$$ **posterior probabilities**

Given a sample of N i.i.d. observations, the parameters are estimated by maximizing

$$L(\vartheta) = \sum_n \log\left(\sum_g p_g \phi_g(\mathbf{x}_n)\right)$$ **log-likelihood**

Problems

- a very large number of parameters;
- difficult to understand which are the "discriminant" variables, i.e. the variables that describe the clustering structure.

Idea

The mixture model induces the following covariance structure

$$\mathrm{Var}(\mathbf{x}) = \overbrace{\sum_{g=1}^{G} p_g(\boldsymbol{\mu}_g - \boldsymbol{\mu})(\boldsymbol{\mu}_g - \boldsymbol{\mu})'}^{\text{Between}} + \overbrace{\sum_{g=1}^{G} p_g \boldsymbol{\Sigma}_g}^{\text{Within}}$$

$$= \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$$

variance decomposition

Model the Between covariance matrix to:

- **reduce the number of parameters;**
- **find the components (linear combinations of the variables) explaining the "largest information" about the classification.**

Reduction Model

The model is a "component analysis" of the centroid matrix.

Scalar

$$\mu_{jg} = \mu_j + \sum_{q=1}^{Q} b_{jq}\eta_{qg}\,, \quad \sum_{g=1}^{G} p_g \eta_{qg} = 0$$

where:

μ_{jg} is the mean of variable *j* in component *g*;

η_{qg} is the mean of *prototype variable q* in component *g*;

b_{jq} is the loading of variable *j* on *prototype variable q*.

Vector

$$\boldsymbol{\mu}_g = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\eta}_g, \quad \sum_{g=1}^{G} p_g \boldsymbol{\eta}_g = \mathbf{0}$$

Matrix

$$\mathbf{M} = \mathbf{N}\mathbf{B}', \qquad \mathbf{1}'\mathbf{N} = \mathbf{0}$$

where:

- $\mathbf{M} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_G]' - \mathbf{1}\boldsymbol{\mu}'$, (centred) centroid matrix;
- $\mathbf{N} = [\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_G]'$, centroid matrix on the reduced space.

The component model is not identified. In fact

$$\mathbf{B}\boldsymbol{\eta}_g = \mathbf{B}\mathbf{F}^{-1}\mathbf{F}\boldsymbol{\eta}_g = \tilde{\mathbf{B}}\tilde{\boldsymbol{\eta}}_g.$$

We exploit such rotational freedom by requiring that

$$\mathbf{B}'\boldsymbol{\Sigma}^{-1}\mathbf{B} = \mathbf{I}_Q.$$

ML Estimation (homoscedastic case): EM algorithm

Maximization of the loglikelihood

$$L(\vartheta) = \sum_{n=1}^{N} \log\left(\sum_{g=1}^{G} p_g \phi_g(\mathbf{x}_n)\right)$$ **objective**

is equivalent to the maximization of the “fuzzy” function (Hathaway, 1986)

$$l(\vartheta) = \sum_{ng} u_{ng} \log\left(p_g \phi_g(\mathbf{x}_n)\right) - \sum_{ng} u_{ng} \log\left(u_{ng}\right)$$ **fuzzy objective**

where $u_{ng} \geq 0$ and $\Sigma_g\ u_{ng}=1$. This is so because $l(\vartheta)$ reaches a maximum respect to $\mathbf{U}=[u_{ng}]$ when

$$u_{ng} = \frac{p_g \phi_g(\mathbf{x}_n)}{\sum_h p_h \phi_h(\mathbf{x}_n)}$$ **posterior probabilities**

Substituting the previous in $l(\vartheta)$ we obtain $L(\vartheta)$.

The algorithm is based on the conditional maximization of $l(\vartheta)$ with respect to a subset of parameters given the others.

The fundamental steps are the following.

a) Update $\mathbf{U}=[u_{ng}]$:

$$u_{ng} = \frac{p_g \phi_g(\mathbf{x}_n)}{\sum_h p_h \phi_h(\mathbf{x}_n)}, \; n=1,2,\ldots,N; \; g=1,2,\ldots, G,$$

b) Update $\mathbf{p}=[p_g]$:

$$p_g = \frac{1}{N}\sum_n u_{ng}, \; g=1,2,\ldots,G.$$

c) Update $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{ng} u_{ng}(\mathbf{x}_n - \boldsymbol{\mu}_g)(\mathbf{x}_n - \boldsymbol{\mu}_g)'.$$

They are simply the steps of a ordinary EM algorithm.

d) Update $\boldsymbol{\mu}$:
We consider centered data, $\boldsymbol{\mu} = \mathbf{0}$.

e) Update $\mathbf{N}$ and $\mathbf{B}$:
It can be shown that the objective function can be written as

$$l(\vartheta) = -\frac{1}{2}\mathrm{tr}\left\{\mathbf{D}(\overline{\mathbf{X}} - \mathbf{N}\mathbf{B}')\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{X}} - \mathbf{N}\mathbf{B}')'\right\} + c$$

where c is a constant term (independent of $\mathbf{N}$ and $\mathbf{B}$), $\mathbf{D}$ = diag(u_{+1}, u_{+2}, ..., u_{+G}) and $\overline{\mathbf{X}} = [\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, ..., \overline{\mathbf{x}}_G]'$ is the matrix of centroids, $\overline{\mathbf{x}}_g = \frac{1}{\sum_n u_{ng}} \sum_n u_{ng} \mathbf{x}_n$, computed on the centred variables.

This algorithm can be also seen as an ECM (Meng & Rubin, 1993).

Use and interpretation of components

Step M of the EM algorithm shows that:

1) the within-standardized component loadings matrix $\widehat{\mathbf{B}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{B}$ derives from a PCA of the matrix of within-standardized centroids

$$\operatorname{tr}\left\{\mathbf{D}(\overline{\mathbf{X}} - \mathbf{N}\mathbf{B}')\mathbf{\Sigma}^{-1}(\overline{\mathbf{X}} - \mathbf{N}\mathbf{B}')'\right\} = \left\|\mathbf{D}^{\frac{1}{2}}\overline{\mathbf{X}}\mathbf{\Sigma}^{-\frac{1}{2}} - \mathbf{D}^{\frac{1}{2}}\mathbf{N}\mathbf{B}'\mathbf{\Sigma}^{-\frac{1}{2}}\right\|^2 = \left\|\mathbf{D}^{\frac{1}{2}}\overline{\mathbf{Z}} - \mathbf{D}^{\frac{1}{2}}\mathbf{N}\widehat{\mathbf{B}}'\right\|^2 \to \min_{\mathbf{N},\mathbf{B}}$$

2) the component scores

$$\mathbf{Y} = \mathbf{Z}\widehat{\mathbf{B}} = \mathbf{X}\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{B} = \mathbf{X}\mathbf{\Sigma}^{-1}\mathbf{B} = \mathbf{X}\breve{\mathbf{B}}$$

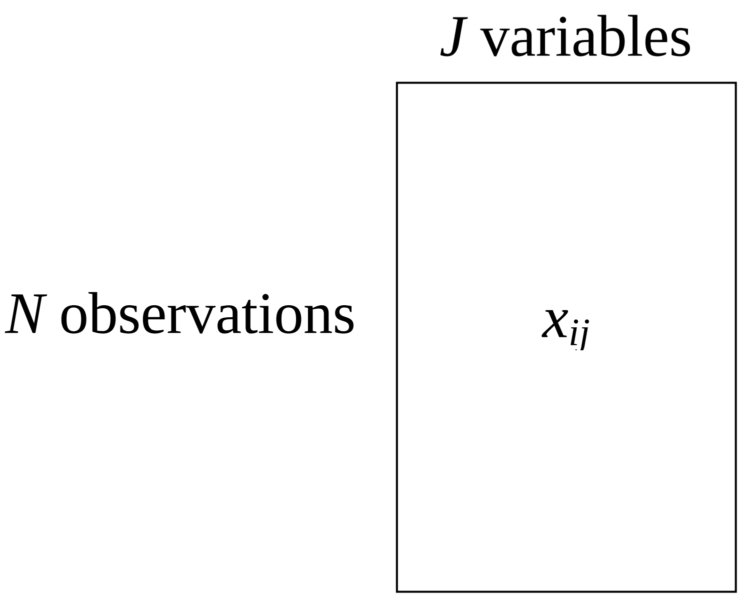
maximize the between variance subject to the constraint of unit within variance, i.e.

$$\max\ \operatorname{tr}\left(\breve{\mathbf{B}}'\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\breve{\mathbf{B}}\right)$$
$$\text{subject to } \breve{\mathbf{B}}'\mathbf{\Sigma}\breve{\mathbf{B}} = \mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{B} = \mathbf{B}'\mathbf{\Sigma}^{-1}\mathbf{B} = \mathbf{I}_Q .$$

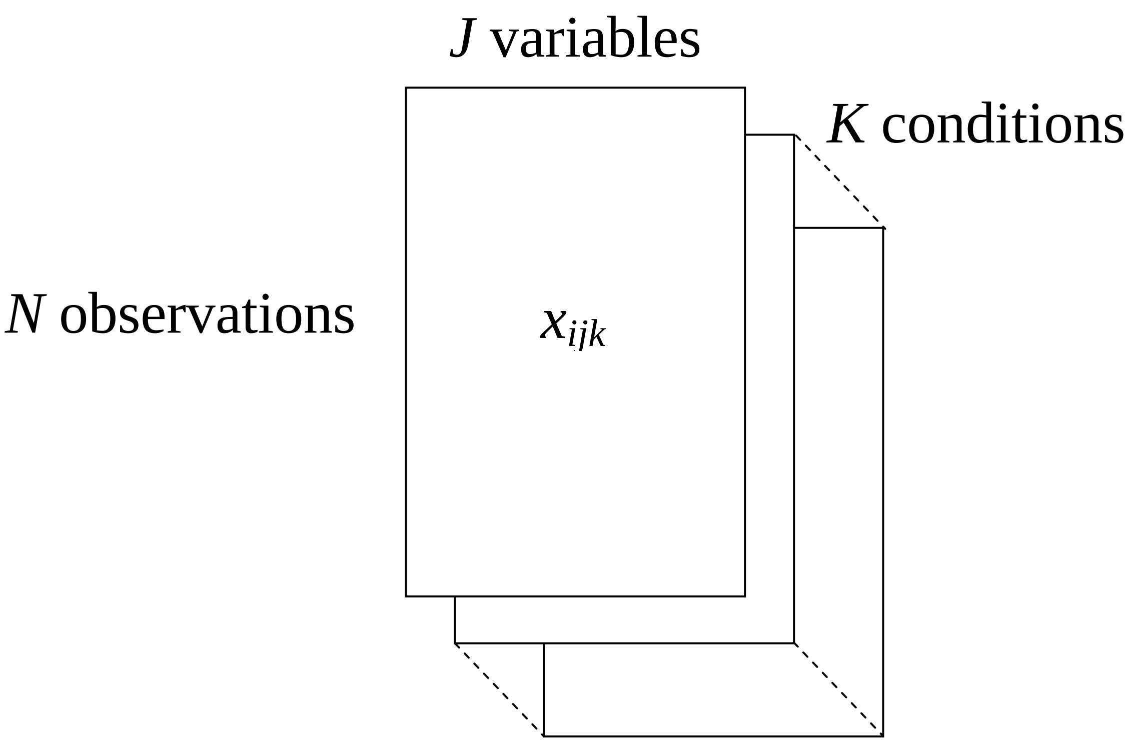
Fisher linear discriminant analysis (LDA)

Three-way Extension

Two-way sample



Three-way sample



Let $\mathbf{x} = [x_{11}, x_{21}, \ldots, x_{J1}, \ldots, x_{1K}, x_{2K}, \ldots, x_{JK}]'$ be a random vector of J variables observed under K different conditions.

General classification model

$$f(\mathbf{x}) = \sum_{g=1}^{G} p_g \phi_g(\mathbf{x})$$

mixture model

where

$$\phi_g(\mathbf{x}) = (2\pi)^{-\frac{JK}{2}} \left|\boldsymbol{\Sigma}_g\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\}$$

Gaussian components

Problems

- a very large number of parameters;
- difficult to understand which are the "discriminant" variables and/or occasions;
- difficult to distinguish the role of variables from that of occasions.

Within Covariance Structure

Model

Direct Product (Browne, 1984)

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_O \otimes \boldsymbol{\Sigma}_V = \begin{bmatrix} \sigma_{11}\boldsymbol{\Sigma}_V & \cdots & \sigma_{1K}\boldsymbol{\Sigma}_V \\ \vdots & \ddots & \vdots \\ \sigma_{K1}\boldsymbol{\Sigma}_V & \cdots & \sigma_{KK}\boldsymbol{\Sigma}_V \end{bmatrix}$$

in scalar notation

$$\sigma_{jklm} = \sigma_{jl}\sigma_{km}$$

Basford & MacLachlan (1985) proposed

$$\boldsymbol{\Sigma}_g = \mathbf{I}_K \otimes \boldsymbol{\Sigma}_{V;g}$$

Reduction Model

The model is a "Tucker 2 component analysis" of the centroid matrix.

Scalar

$$\mu_{jkg} = \mu_{jk} + \sum_{q=1}^{Q}\sum_{r=1}^{R} b_{jq} c_{kr} \eta_{qrg}, \quad \sum_{g=1}^{G} p_g \eta_{qrg} = 0$$

where:
- μ_{jkg} is the mean of variable j under condition k in component g;
- η_{qrg} is the mean of *prototype variable q* under *prototype condition r* in component g;
- $\sum_q b_{jq}\eta_{qrg}$ is the mean of variable j under *prototype condition r* in component g;
- b_{jq} is the loading of variable j on *prototype variable q*;
- $\sum_r c_{kr}\eta_{qrg}$ is the mean of *prototype variable q* under condition k in component g;
- c_{kr} is the loading of occasion k on *prototype occasion r*.

Often used in Chemistry and Psychology, see http://three-mode.leidenuniv.nl/

Vector

$$\boldsymbol{\mu}_g = \boldsymbol{\mu} + (\mathbf{C} \otimes \mathbf{B})\boldsymbol{\eta}_g, \quad \sum_{g=1}^{G} p_g \boldsymbol{\eta}_g = \mathbf{0}$$

Matrix

$$\mathbf{M} = \mathbf{N}(\mathbf{C}' \otimes \mathbf{B}'), \quad \mathbf{1}'\mathbf{N} = \mathbf{0}$$

where:

- $\mathbf{M} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_G]' - \mathbf{1}\boldsymbol{\mu}'$, (centred) centroid matrix;
- $\mathbf{N} = [\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_G]'$, centroid matrix on the reduced space.

The component model is not identified. In fact

$$(\mathbf{C} \otimes \mathbf{B})\boldsymbol{\eta}_g = (\mathbf{C} \otimes \mathbf{B})(\mathbf{D}^{-1} \otimes \mathbf{F}^{-1})(\mathbf{D} \otimes \mathbf{F})\boldsymbol{\eta}_g = (\mathbf{C}\mathbf{D}^{-1} \otimes \mathbf{B}\mathbf{F}^{-1})\tilde{\boldsymbol{\eta}}_g = (\tilde{\mathbf{C}} \otimes \tilde{\mathbf{B}})\tilde{\boldsymbol{\eta}}_g.$$

We exploit such rotational freedom by requiring that

$$\mathbf{B}'\boldsymbol{\Sigma}_V^{-1}\mathbf{B} = \mathbf{I}_Q, \; \mathbf{C}'\boldsymbol{\Sigma}_O^{-1}\mathbf{C} = \mathbf{I}_R.$$

ML Estimation (homoscedastic case): EM algorithm

An EM algorithm can be programmed following the analogous algorithm already seen for the two-way case.

About the update of $\mathbf{N}$, $\mathbf{B}$ and $\mathbf{C}$, it is interesting to note that the complete log-likelihood can be written as

$$l(\vartheta) = -\frac{1}{2}\operatorname{tr}\left\{\mathbf{D}[\overline{\mathbf{X}} - \mathbf{N}(\mathbf{C} \otimes \mathbf{B})'](\boldsymbol{\Sigma}_O^{-1} \otimes \boldsymbol{\Sigma}_V^{-1})[\overline{\mathbf{X}} - \mathbf{N}(\mathbf{C} \otimes \mathbf{B})']'\right\} + c$$

where c is a constant term and $\overline{\mathbf{X}}$ is the matrix of centroids computed on the centred variables.

It follows that the parameters can be updated by computing a weighted least squares approximation of the centroid matrix.

Use and interpretation of components

1) the within-standardized component loadings matrices $\widehat{\mathbf{B}} = \mathbf{\Sigma}_V^{-\frac{1}{2}}\mathbf{B}$ and $\widehat{\mathbf{C}} = \mathbf{\Sigma}_O^{-\frac{1}{2}}\mathbf{C}$ derive from a Tucker2 analysis of the matrix of within-standardized centroids

$$\left\| \mathbf{D}^{\frac{1}{2}}\overline{\mathbf{X}}(\mathbf{\Sigma}_O^{-\frac{1}{2}} \otimes \mathbf{\Sigma}_V^{-\frac{1}{2}}) - \mathbf{D}^{\frac{1}{2}}\mathbf{N}(\mathbf{C}' \otimes \mathbf{B}')(\mathbf{\Sigma}_O^{-\frac{1}{2}} \otimes \mathbf{\Sigma}_V^{-\frac{1}{2}}) \right\|^2 = \left\| \mathbf{D}^{\frac{1}{2}}\overline{\mathbf{Z}} - \mathbf{D}^{\frac{1}{2}}\mathbf{N}(\widehat{\mathbf{C}}' \otimes \widehat{\mathbf{B}}') \right\|^2 \rightarrow \min_{\mathbf{N},\mathbf{B},\mathbf{C}}$$

2) the component scores

$$\mathbf{Y} = \mathbf{Z}(\widehat{\mathbf{C}} \otimes \widehat{\mathbf{B}}) = \mathbf{X}(\mathbf{\Sigma}_O^{-1} \otimes \mathbf{\Sigma}_V^{-1})(\mathbf{C} \otimes \mathbf{B}) = \mathbf{X}(\breve{\mathbf{C}} \otimes \breve{\mathbf{B}})$$

maximize the between variance subject to the constraint of unit within variance, i.e.

$$\max\ \mathrm{tr}\left[(\breve{\mathbf{C}} \otimes \breve{\mathbf{B}})'\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}(\breve{\mathbf{C}} \otimes \breve{\mathbf{B}})\right]$$

$$\text{subject to } \breve{\mathbf{C}}'\mathbf{\Sigma}_O\breve{\mathbf{C}} = \mathbf{I}_R, \breve{\mathbf{B}}'\mathbf{\Sigma}_V\breve{\mathbf{B}} = \mathbf{I}_Q \Leftrightarrow (\breve{\mathbf{C}} \otimes \breve{\mathbf{B}})'(\mathbf{\Sigma}_O \otimes \mathbf{\Sigma}_V)(\breve{\mathbf{C}} \otimes \breve{\mathbf{B}}) = \mathbf{I}_R \otimes \mathbf{I}_Q$$

Bilinear discriminant analysis (BLDA)

BLDA: interpretation

Constrained LDA

$$y_{qr} = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{jk} w_{jkqr} \Leftrightarrow \mathbf{y} = (\breve{\mathbf{c}} \otimes \breve{\mathbf{b}})'\mathbf{x}$$

where

$$w_{jkqr} = \breve{b}_{jq}\breve{c}_{kr}$$

Dimensionality reduction of the variables

Dimensionality reduction of the occasions

Hierarchical LDA

$$\begin{cases} y_{qr} = \sum_{j=1}^{J} \breve{b}_{jq} f_{jr} \Leftarrow f_{jr} = \sum_{k=1}^{K} \breve{c}_{kr} x_{jk} \\ y_{qr} = \sum_{k=1}^{K} \breve{c}_{kr} h_{qk} \Leftarrow h_{qk} = \sum_{j=1}^{J} \breve{b}_{jq} x_{jk} \end{cases} \Rightarrow \begin{cases} \breve{b}_{jp} \text{ variable component weights} \\ \breve{c}_{kq} \text{ occasion component weights} \end{cases}$$

Application

Data

58 units: soybeans;

8 conditions: 4 environments (Lawes, Brookstead, Nambour, Redland Bay)
× 2 years (1970, 1971);

2 variables: yield Kg/Ha, protein.

Model selection

Model considered:

G = 2:7, Q = 1:2, R = 1:8, $\boldsymbol{\Sigma}_O$ diagonal or with non null covariances only between the same locations.

Best model selected by BIC:

$$G = 7,\ Q = 2,\ R = 2 \text{ and } \boldsymbol{\Sigma}_O \text{ diagonal.}$$

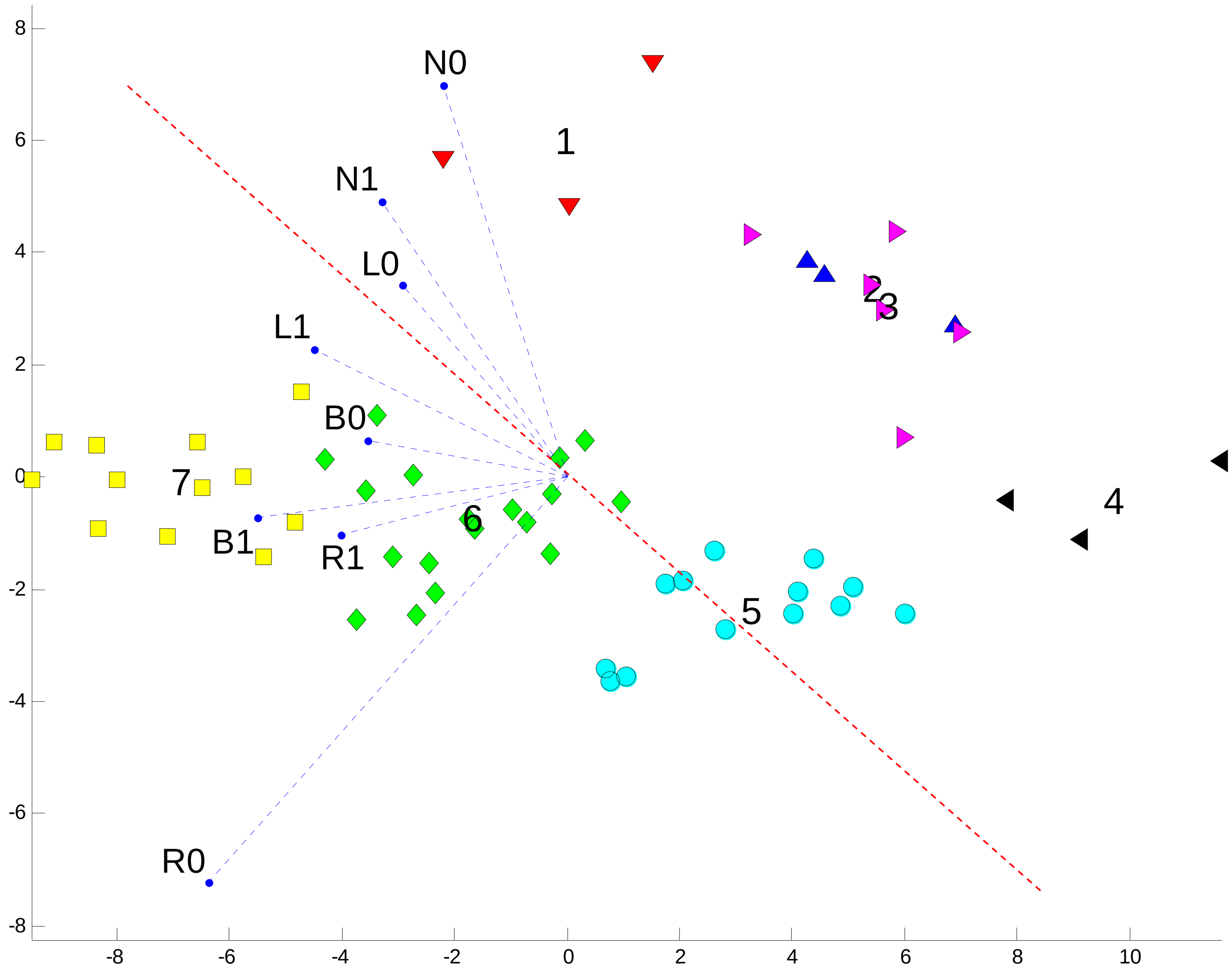
Percentage of variation accounted for by the components on the within-standardized data

	Occasions		
Variables	**1**	**2**	**Tot**
1	50.98	11.59	62.57
2	11.96	4.56	16.52
Tot	62.94	16.15	79.09

Basford & McLachlan (B&M) and our (R) classification

	R						
B&M	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	3						
2		3					
3			6	3			
4					9		
5					3	6	
6					1	8	
7						4	12

Biplot on the first latent variable at the two latent occasions



Heteroscedastic case

Reduction model

Scalar

$$\mu_{jkg} = \mu_{jk} + \sum_{q=1}^{J}\sum_{r=1}^{K} b_{jq} c_{kr} \eta_{qrg}, \quad \sum_{g=1}^{G} p_g \eta_{qrg} = 0, \ \eta_{qrg} = 0 \text{ if } q > Q \text{ and/or } r > R$$

Vector

$$\boldsymbol{\mu}_g = \boldsymbol{\mu} + (\mathbf{C} \otimes \mathbf{B})\boldsymbol{\eta}_g, \quad \sum_{g=1}^{G} p_g \boldsymbol{\eta}_g = \mathbf{0}$$

Matrix

$$\mathbf{M} = \mathbf{N}(\mathbf{C}' \otimes \mathbf{B}'), \qquad \mathbf{1}'\mathbf{N} = \mathbf{0}$$

where

- $\mathbf{C} = [\mathbf{C}_R, \mathbf{C}_{K-R}]$, square,
- $\mathbf{B} = [\mathbf{B}_Q, \mathbf{B}_{J-Q}]$, square.

Within-covariance model

$$\boldsymbol{\Sigma}_g = (\mathbf{C} \otimes \mathbf{B})\boldsymbol{\Omega}_g(\mathbf{C} \otimes \mathbf{B})'$$

where

- $\boldsymbol{\Omega}_g = \begin{bmatrix} \boldsymbol{\Omega}_{O,g} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{\Omega}_{V,g} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \boldsymbol{\Psi}$,

- $\boldsymbol{\Psi}$ diagonal.

If K =3 and R = 2, we have

$$\mathbf{\Omega}_g = \left[\begin{array}{cc|cc|cc} \omega_{11O,g}\mathbf{\Omega}_{V,g} & 0 & \omega_{12O,g}\mathbf{\Omega}_{V,g} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \omega_{21O,g}\mathbf{\Omega}_{V,g} & 0 & \omega_{22O,g}\mathbf{\Omega}_{V,g} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] + \mathbf{\Psi}$$