On the application of common principal components in biplots

Theo Pepler
Genetics Department
Stellenbosch University

1 November 2011

What are common principal components (CPCs)?

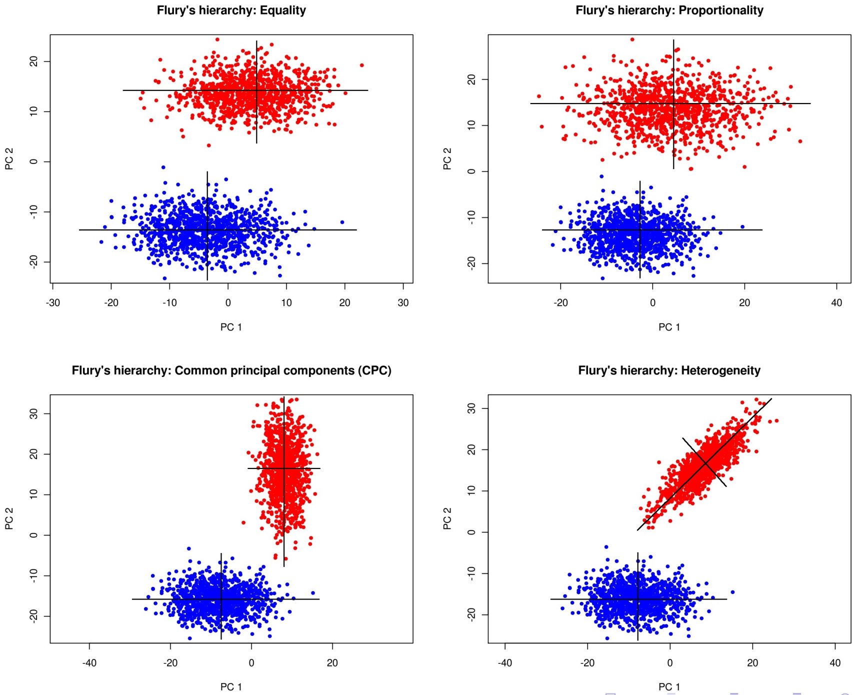
How can variance structures of two (or more) groups differ?

Univariate case:

- Homoscedastic or heteroscedastic (nothing in between)

Multivariate case:

- Number of different ways covariance matrices can differ (Flury 1988):
  1. Equality $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$
  2. Proportionality $\boldsymbol{\Sigma}_1 = \rho\boldsymbol{\Sigma}_2$
  3. Common principal components
  4. Partial common principal components
  5. Heterogeneity


Flury's hierarchy: Equality
Flury's hierarchy: Proportionality
Flury's hierarchy: Common principal components (CPC)
Flury's hierarchy: Heterogeneity
PC 1
PC 2

Principal component analysis (PCA):

$$\boldsymbol{\Sigma} = \mathbf{B}\boldsymbol{\Lambda}\mathbf{B}'$$

Common principal components (CPC):

$$\boldsymbol{\Sigma}_1 = \mathbf{B}\boldsymbol{\Lambda}_1\mathbf{B}'$$

$$\boldsymbol{\Sigma}_2 = \mathbf{B}\boldsymbol{\Lambda}_2\mathbf{B}'$$

Partial common principal components (CPC(q)):

$$\begin{array}{lll} \boldsymbol{\Sigma}_1 = \mathbf{B}_1\boldsymbol{\Lambda}_1\mathbf{B}_1' & \text{where} & \mathbf{B}_1 = [\mathbf{b}_1 \ldots \mathbf{b}_q : \mathbf{b}_{q+1(1)} \ldots \mathbf{b}_{p(1)}] \\ \boldsymbol{\Sigma}_2 = \mathbf{B}_2\boldsymbol{\Lambda}_2\mathbf{B}_2' & & \mathbf{B}_2 = [\mathbf{b}_1 \ldots \mathbf{b}_q : \mathbf{b}_{q+1(2)} \ldots \mathbf{b}_{p(2)}] \end{array}$$

Advantages the CPC model might provide:

- **more stable estimates** than when incorrectly assuming *heterogeneity* of covariance matrices
- **more accurate estimates** than when incorrectly assuming *equality* of covariance matrices

Identifying the CPCs

Table 7.9. Decomposition of X^2_{total} in Head Dimension Example ($k = 2, p = 6$)

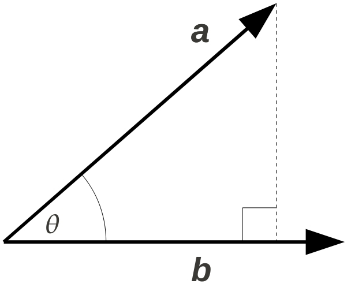
Model: Higher	Model: Lower	X^2	df	$\frac{X^2}{\text{df}}$	AIC for Higher Model
Equality	Proportionality	42.29	1	42.29	89.78
Proportionality	CPC	25.66	5	5.13	49.49
CPC	CPC(1)	15.12	10	1.51	33.82*
CPC(1)	Unrelated	6.70	5	1.34	38.70
Unrelated	---				42.0
Equality	Unrelated	89.78	21		

*Minimum AIC.

- The χ^2 statistics are *not independent* and *assume normality* of the k populations
- The AIC is *not a formal hypothesis test*

Different approach (Krzanowski 1979)

Geometrically: dot product of two unit vectors **a** and **b** = cosine of the angle between the two vectors in p-dimensional space.



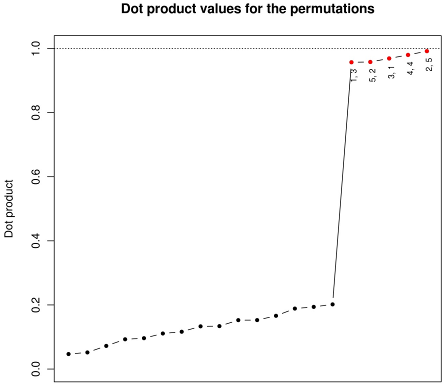
$$\mathbf{a}'\mathbf{b} = \cos\theta$$

- Do pairwise comparisons of the dot products from all combinations of the p principal components from k groups.

Simulated CPC data, $k = 2$, $p = 5$, $n = 200$

- Arbitrary cut-off point: $\cos^{-1}(0.95) = 18.2$ degrees

		Dot products
2	5	**0.99**
4	4	**0.98**
3	1	**0.97**
5	2	**0.96**
1	3	**0.96**
5	3	0.20
1	2	0.19
3	2	0.19
4	3	0.17
1	1	0.15



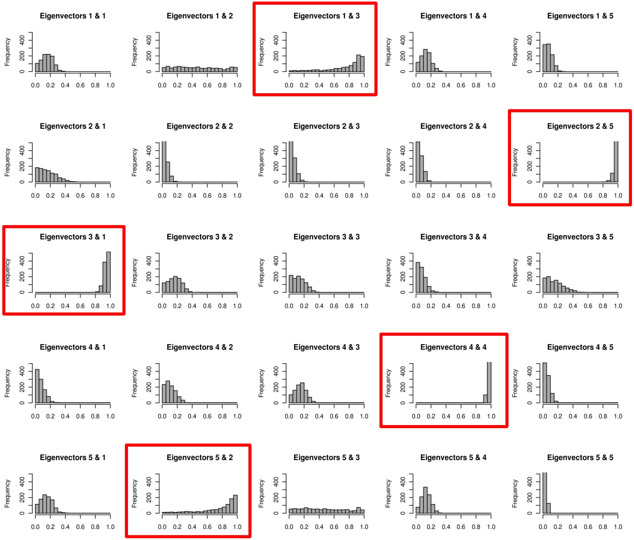
Simulated CPC data:

$k = 2$

$p = 5$

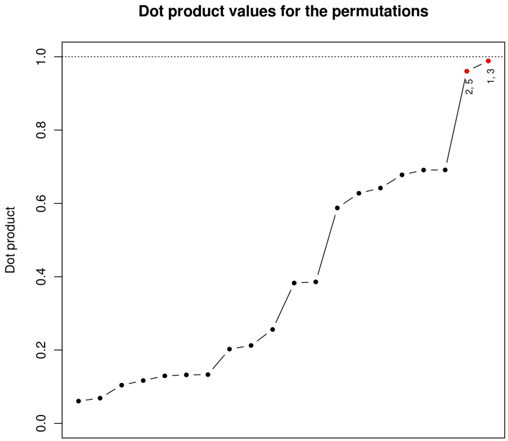
$n = 200$

bootstrap reps $= 1000$



Simulated CPC(2) data, $k = 2$, $p = 5$, $n = 200$

		Dot products
1	3	**0.99**
2	5	**0.96**
3	4	0.69
4	2	0.69
5	2	0.68
3	1	0.64
5	1	0.63
4	4	0.59
4	1	0.39
5	4	0.38



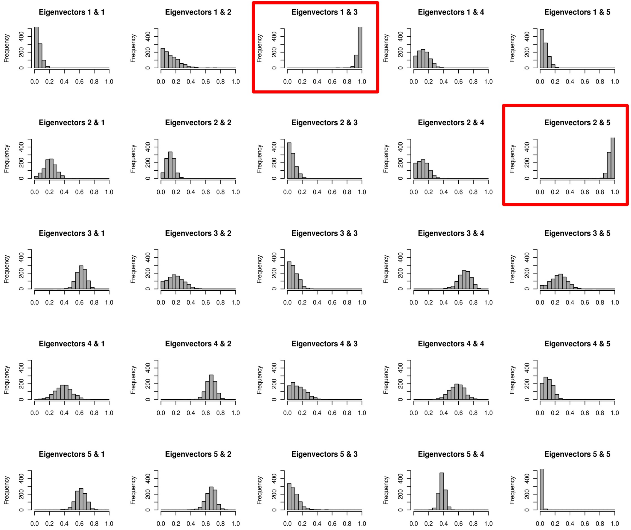
Simulated CPC(2) data:

$k = 2$

$p = 5$

$n = 200$

bootstrap reps $= 1000$



Simultaneous diagonalisation methods

- **FG algorithm** (Flury 1988)

$$\min \phi(\mathbf{\Lambda}_i) := \frac{\det(\operatorname{diag}(\mathbf{\Lambda}_i))}{\det(\mathbf{\Lambda}_i)}$$

- **Stepwise CPC** (Trendafilov 2010)
- **rjd/JADE** (Cardoso & Souloumiac 1996)

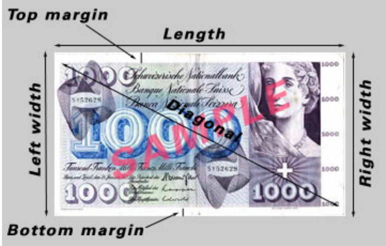
$$\min \sum_{i=1}^{p} \sum_{j>i}^{p} \lambda_{ij}^2$$

Compared these with:

- Eigenvectors of the *pooled covariance matrix*
- Eigenvectors of the covariance matrix of the *pooled data*

Application of the CPC model in biplots

Swiss bank notes data:



X_1 : Length of the bank note,

X_2 : Height of the bank note, measured on the left,

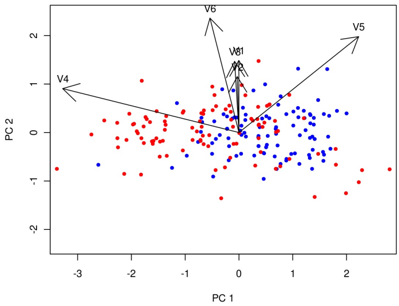
X_3 : Height of the bank note, measured on the right,

X_4 : Distance of inner frame to the lower border,

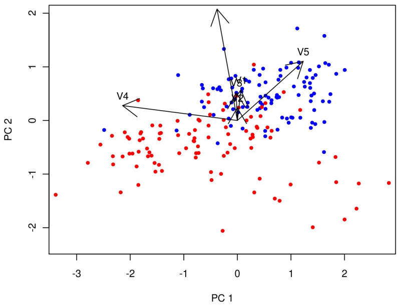
X_5 : Distance of inner frame to the upper border,

X_6 : Length of the diagonal.

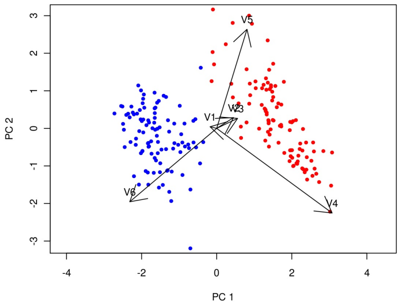
Stepwise CPC biplot: Bank notes data



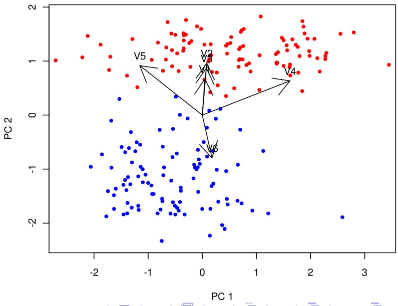
Pooled covariance matrix biplot: Bank notes data



Pooled data biplot: Bank notes data



Flury CPC biplot: Bank notes data



Biplot goodness of fit

Overall quality of the display (Gower, Lubbe & Le Roux 2011)

Letting $\mathbf{X}$ contain the data from all k groups, with the columns of $\mathbf{X}$ centred, and $||\mathbf{X}||^2 = \text{tr}(\mathbf{X}'\mathbf{X})$, the total variation in the data can be partitioned as follows:

$$||\mathbf{X}||^2 = ||\hat{\mathbf{X}}_{[r]}||^2 + ||\mathbf{X} - \hat{\mathbf{X}}_{[r]}||^2$$

$$\textit{Total} \text{ goodness of fit} = \frac{||\hat{\mathbf{X}}_{[r]}||^2}{||\mathbf{X}||^2} = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

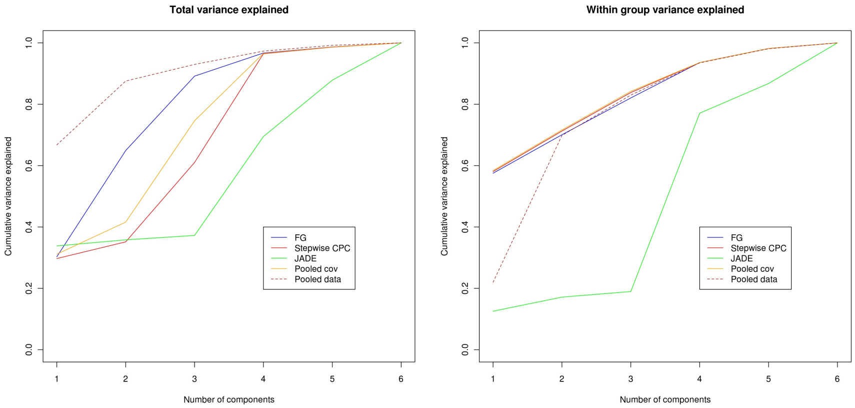
Biplot goodness of fit

Within group variation

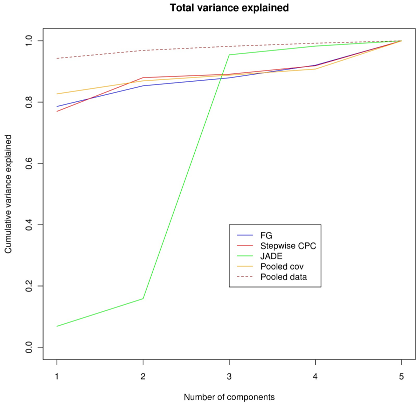
Letting $\mathbf{X}_i$ contain the data from the i^{th} group, with the columns of $\mathbf{X}_i$ centred *per group*, the quality of representation of the within group variation can be measured as follows:

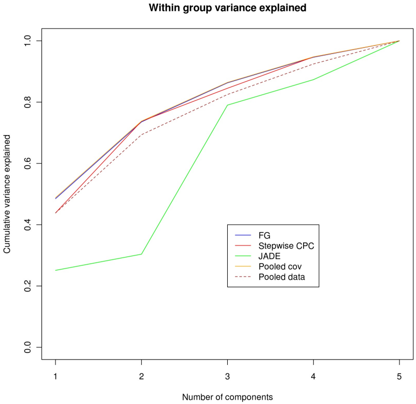
$$\textit{Within groups}\text{ goodness of fit} = \frac{\sum_{i=1}^{k} ||\hat{\mathbf{X}}_{[r]}||^2}{\sum_{i=1}^{k} ||\mathbf{X}||^2} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{r} \lambda_{ji}}{\sum_{j=1}^{k}\sum_{i=1}^{p} \lambda_{ji}}$$

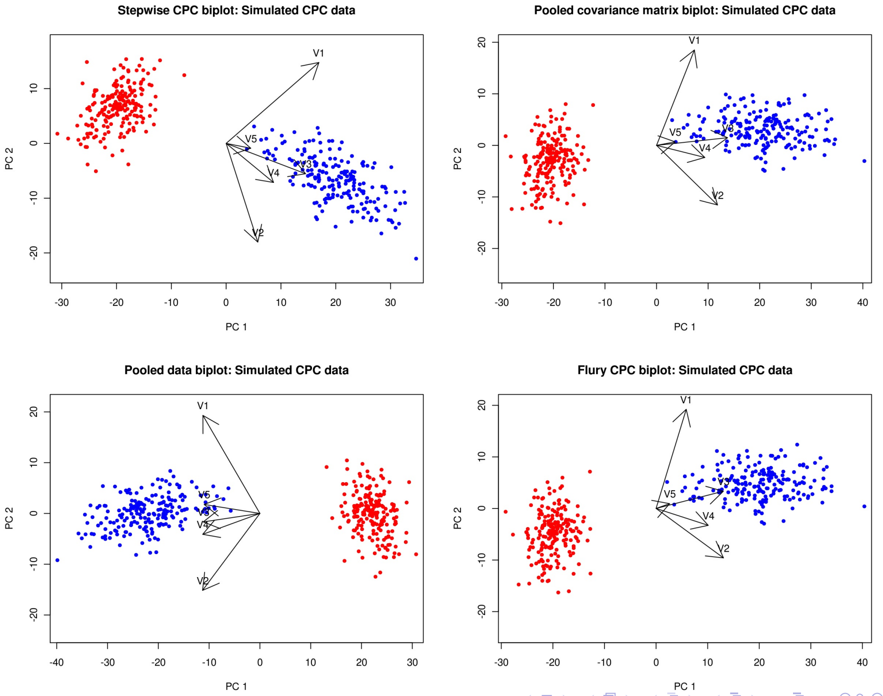
Swiss bank notes data: $k = 2$, $p = 6$, $n = 100$



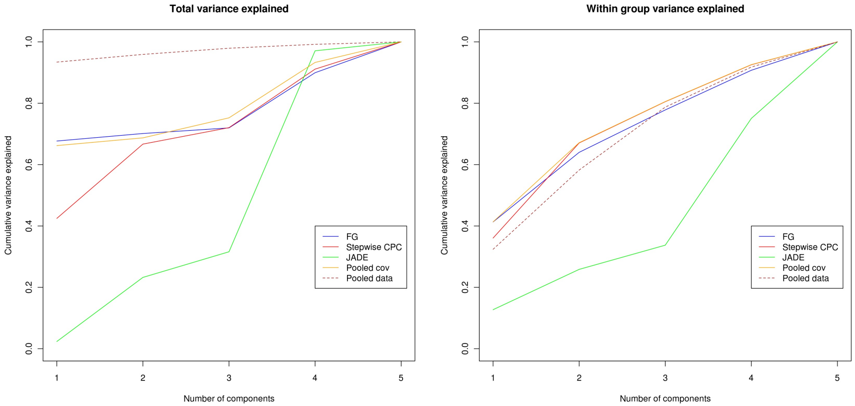
Simulated CPC data: $k = 2$, $p = 5$, $n = 200$



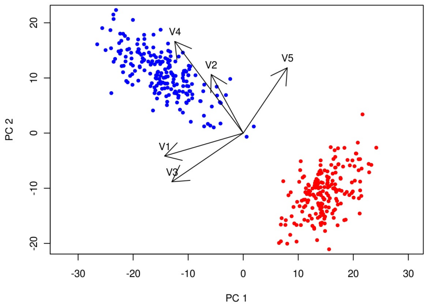



Stepwise CPC biplot: Simulated CPC data
PC 1
PC 2
V1
V2
V3
V4
V5
Pooled covariance matrix biplot: Simulated CPC data
PC 1
PC 2
V1
V2
V3
V4
V5
Pooled data biplot: Simulated CPC data
PC 1
PC 2
V1
V2
V3
V4
V5
Flury CPC biplot: Simulated CPC data
PC 1
PC 2
V1
V2
V3
V4
V5

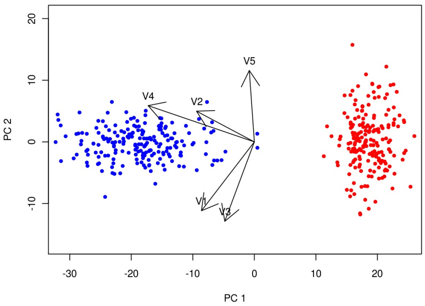
Simulated CPC(2) data: $k = 2$, $p = 5$, $n = 200$



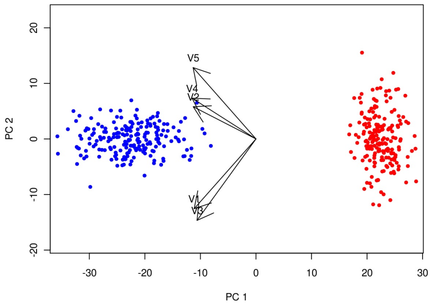
Stepwise CPC biplot: Simulated CPC(2) data



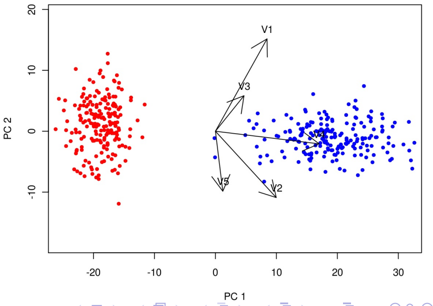
Pooled covariance matrix biplot: Simulated CPC(2) data



Pooled data biplot: Simulated CPC(2) data



Flury CPC biplot: Simulated CPC(2) data



Conclusions

- Eigenvectors of the covariance matrix of the *pooled data* provide the simplest and best quality display for grouped data in 2D or 3D biplots
- Preliminary work also indicates that the axis predictivities (quality of representation of the variables) of the pooled data biplot are higher than for CPC biplots
- Eigenvectors of the pooled covariance matrix and the CPC solutions provide similar quality biplot displays
- CPC solutions are more useful for maximising the variation *within* groups than the variation *between* groups

Sources

- J. Cardoso and A. Souloumiac. *Jacobi angles for simultaneous diagonalization*. SIAM Journal on Matrix Analysis and Applications, 17(1):pp. 161-164, 1996.
- P. Diaconis and B. Efron. *Computer-intensive methods in statistics*. Sci. Am.;(United States), 248(5):116-130, 1983.
- B. Flury. *Common principal components and related multivariate models*. Wiley series in probability and mathematical statistics. Wiley, 1988.
- J. Gower, S.G. Lubbe, and N.L. le Roux. *Understanding Biplots*. John Wiley & Sons, 2011.
- W. J. Krzanowski. *Between-groups comparison of principal components*. Journal of the American Statistical Association, 74(367):pp. 703-707, 1979.
- N. Trendafilov. *Stepwise estimation of common principal components*. Computational Statistics and Data Analysis, 54(12):pp. 3446-3457, 2010.