Common principal components

Theo Pepler
Genetics Department
Stellenbosch University

26 August 2011

Overview

1) What are common principal components (CPCs)?

2) Identifying the CPCs

3) Simultaneous diagonalisation methods

4) Applications of the CPC model

1) What are CPCs?

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

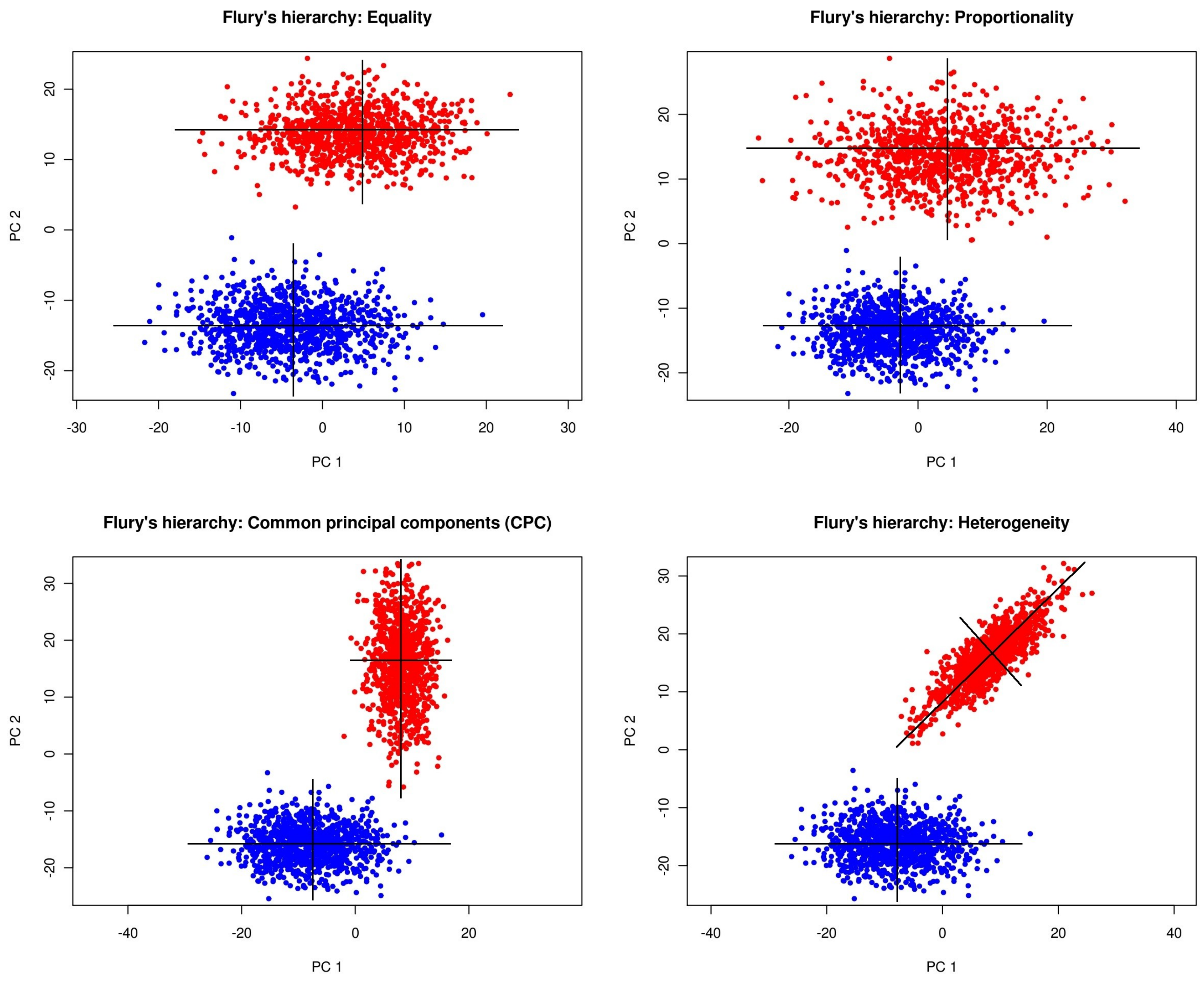
Univariate case:

- 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) Heteroscedasticity $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$


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):

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

Common principal components (CPC):

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

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

Partial common principal components (CPC(*q*)):

$$\boldsymbol{\Sigma}_1 = \boldsymbol{B}_1\boldsymbol{\Lambda}_1\boldsymbol{B}_1' \quad \text{where} \quad \boldsymbol{B}_1 = \left[\boldsymbol{b}_1 \ldots \boldsymbol{b}_q : \boldsymbol{b}_{q+1(1)} \ldots \boldsymbol{b}_{p(1)}\right]$$

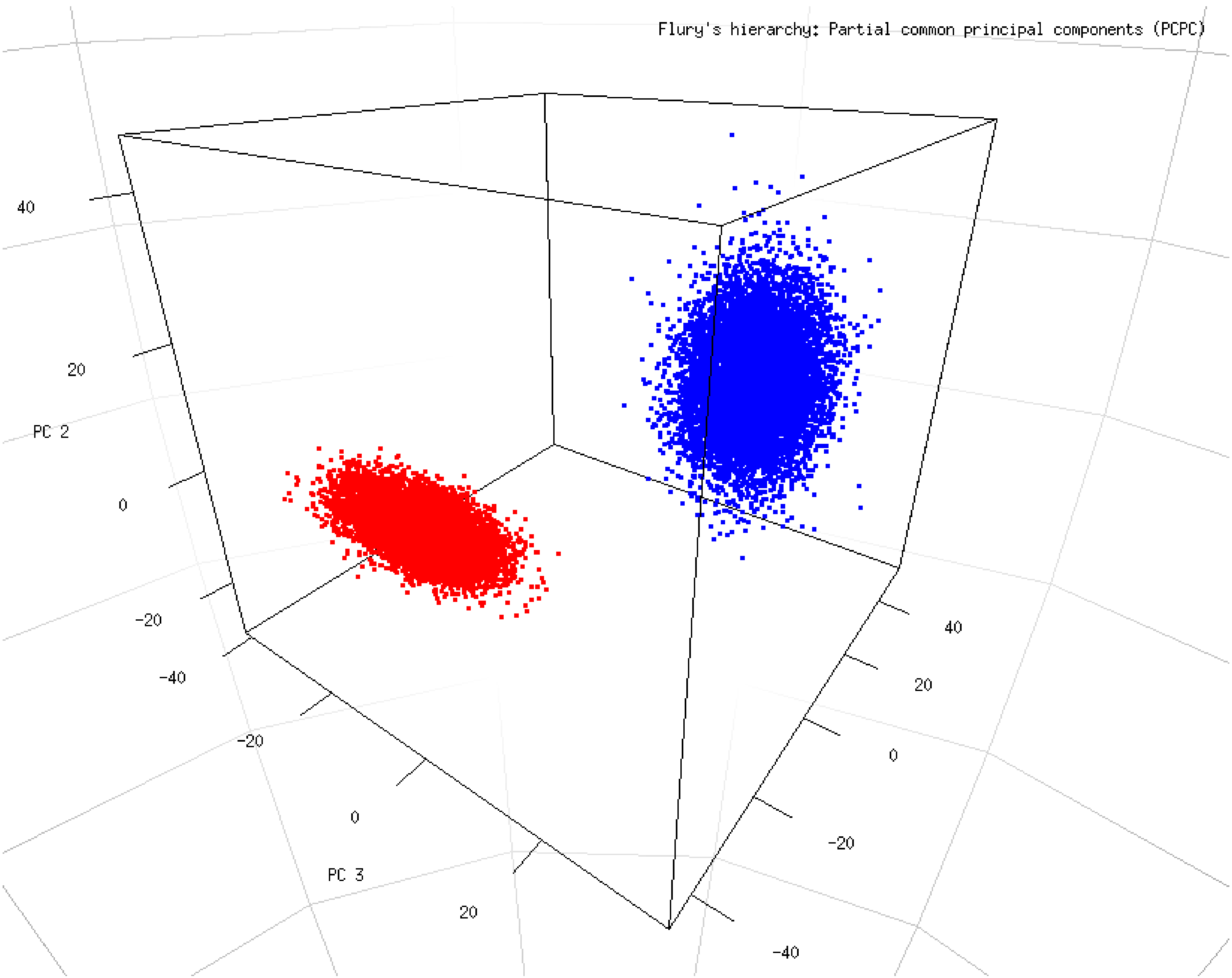
$$\boldsymbol{\Sigma}_2 = \boldsymbol{B}_2\boldsymbol{\Lambda}_2\boldsymbol{B}_2' \qquad \boldsymbol{B}_2 = \left[\boldsymbol{b}_1 \ldots \boldsymbol{b}_q : \boldsymbol{b}_{q+1(2)} \ldots \boldsymbol{b}_{p(2)}\right]$$

- CPC(*p*-1) implies CPC(*p*) due to orthogonality of components

- CPC(*q*) only possible when p > 2

- Moving down in Flury's hierarchy

  --> more parameters to estimate


Flury's hierarchy: Partial common principal components (PCPC)
40
20
PC 2
0
-20
-40
-20
0
PC 3
20
40
20
0
-20
-40

2) Identifying the CPCs

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

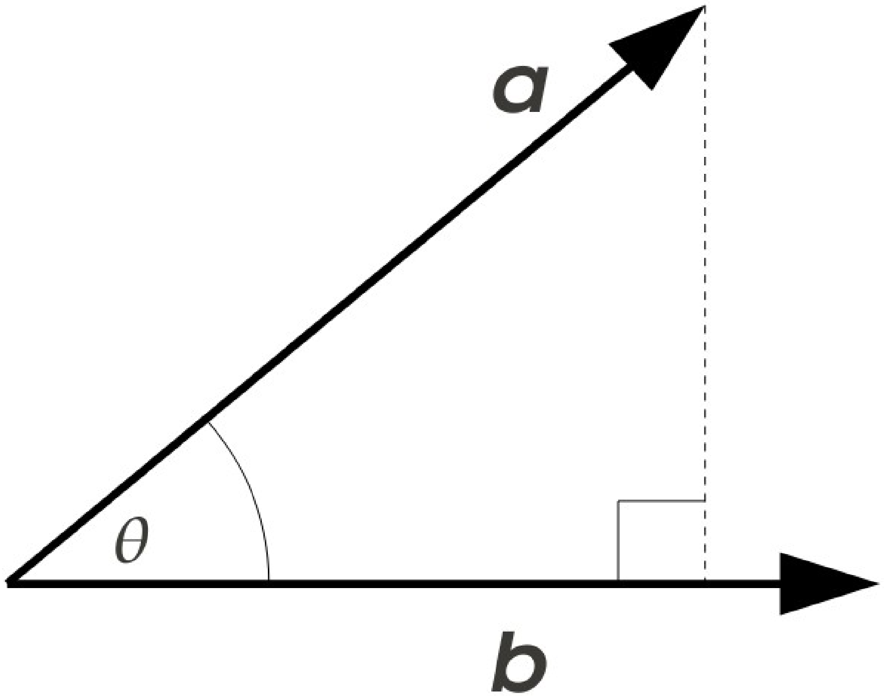
Model: Higher	Model: Lower	X^2	df	$\frac{X^2}{\text{df}}$	AIC for Higher Model
Equality	Proportionality	42.29	1	42.29	89.78
Proportionality	CPC	25.66	5	5.13	49.49
CPC	CPC(1)	15.12	10	1.51	33.82*
CPC(1)	Unrelated	6.70	5	1.34	38.70
Unrelated	---				42.0
Equality	Unrelated	89.78	21		

*Minimum AIC.

- The χ^2 statistics are ***not independent***, and ***assume normality*** of the *k* populations (Flury 1988)
- **AIC not a formal hypothesis test** (Flury 1988)
- Similar criticism also raised by Phillips & Arnold 1999, and Waldmann & Anderson 2000

Different approach: Krzanowski (1979)

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

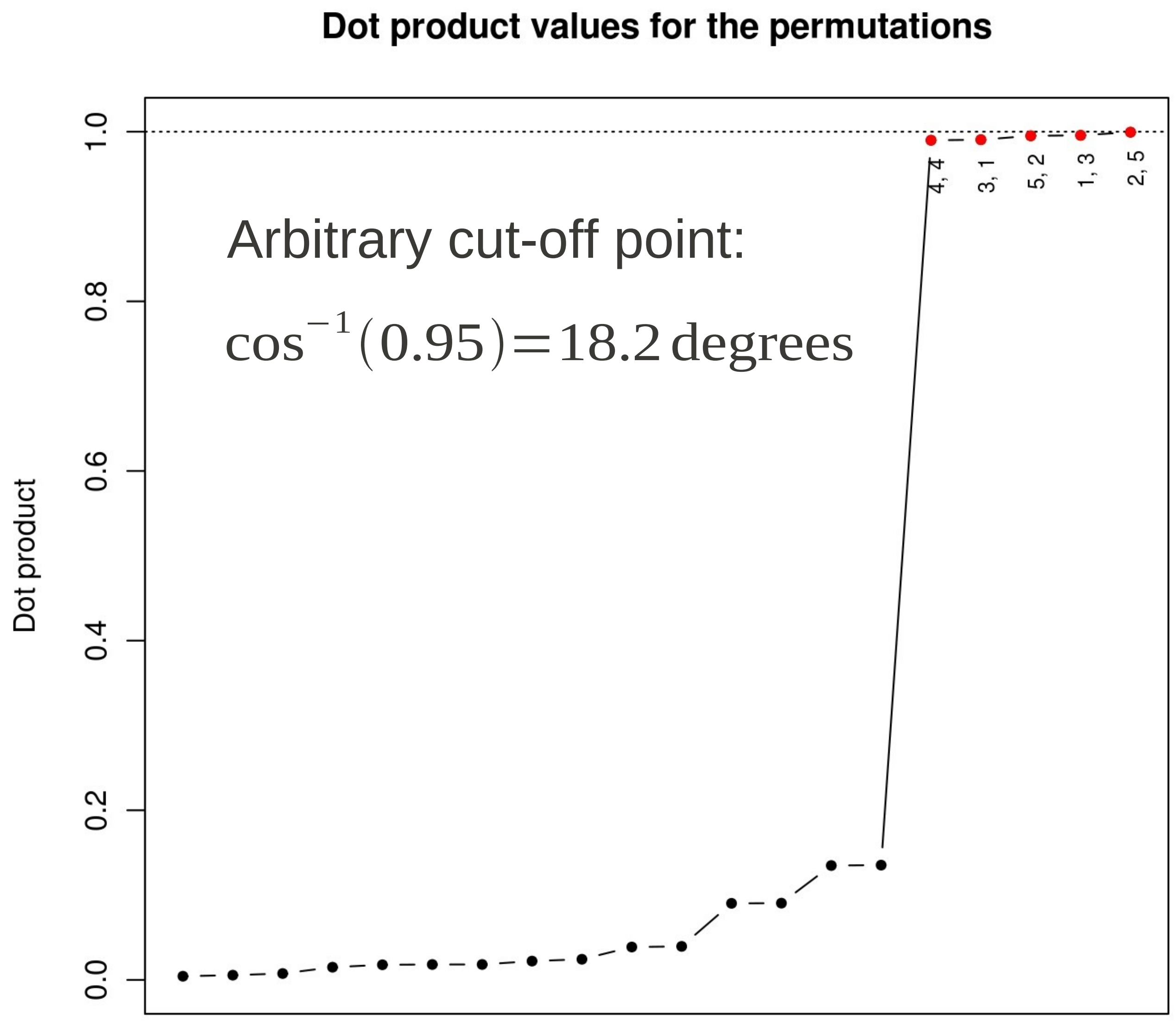


$$\boldsymbol{a}'\boldsymbol{b} = \cos\theta$$

Do pairwise comparison of the dot products from all combinations of the *p* principal components (i.e. the eigenvectors) from *k* groups.

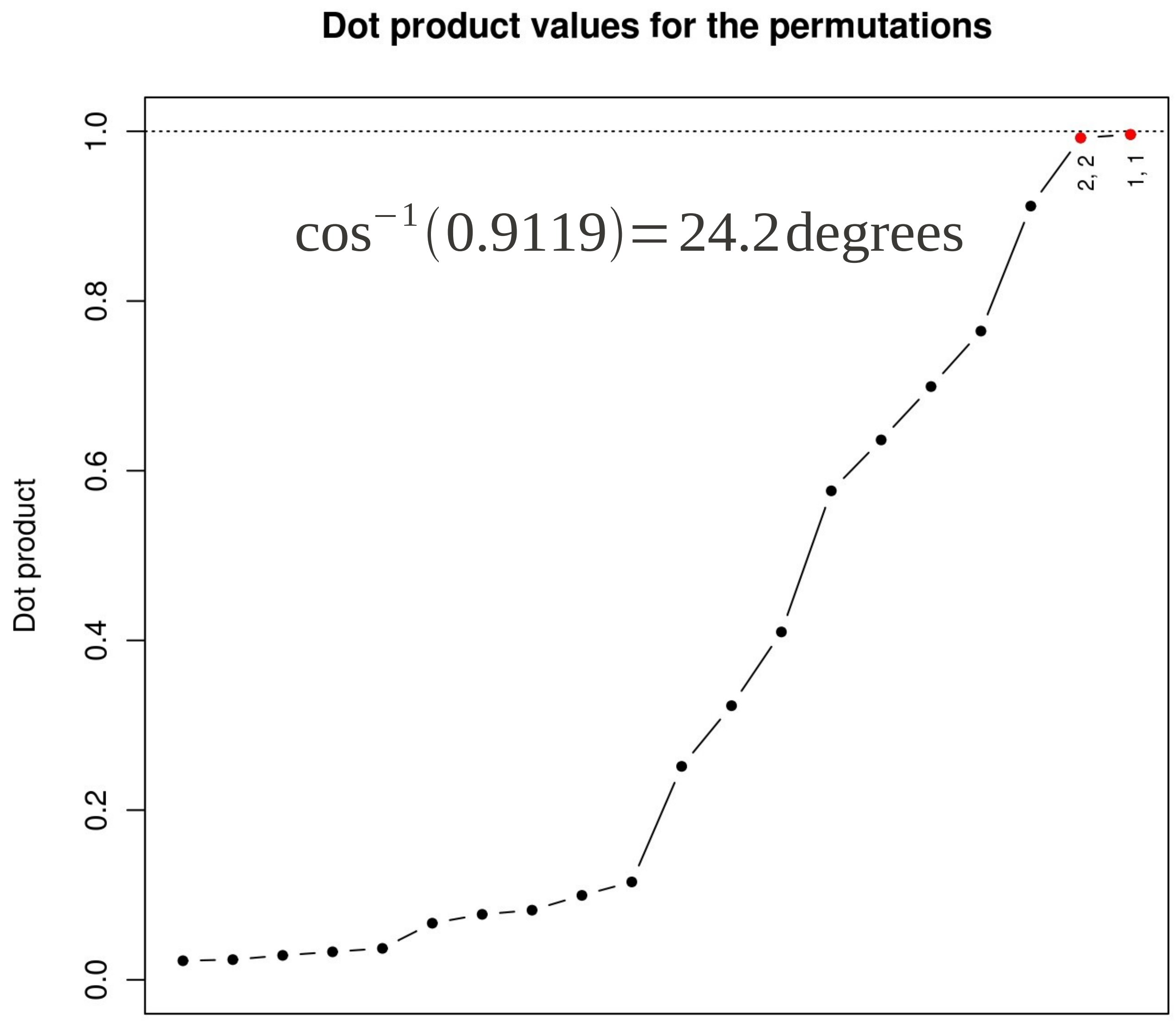
Simulated CPC data, k = 2, p = 5, n = 1000

	Dot products
2 5	**0.999**
1 3	**0.996**
5 2	**0.995**
3 1	**0.991**
4 4	**0.990**
3 4	0.135
4 1	0.135
5 3	0.090
1 2	0.090
4 2	0.039
5 4	0.039
2 1	0.024
3 5	0.022
4 5	0.018
2 3	0.018
1 5	0.018
2 4	0.015
1 4	0.008
1 1	0.006
4 3	0.004



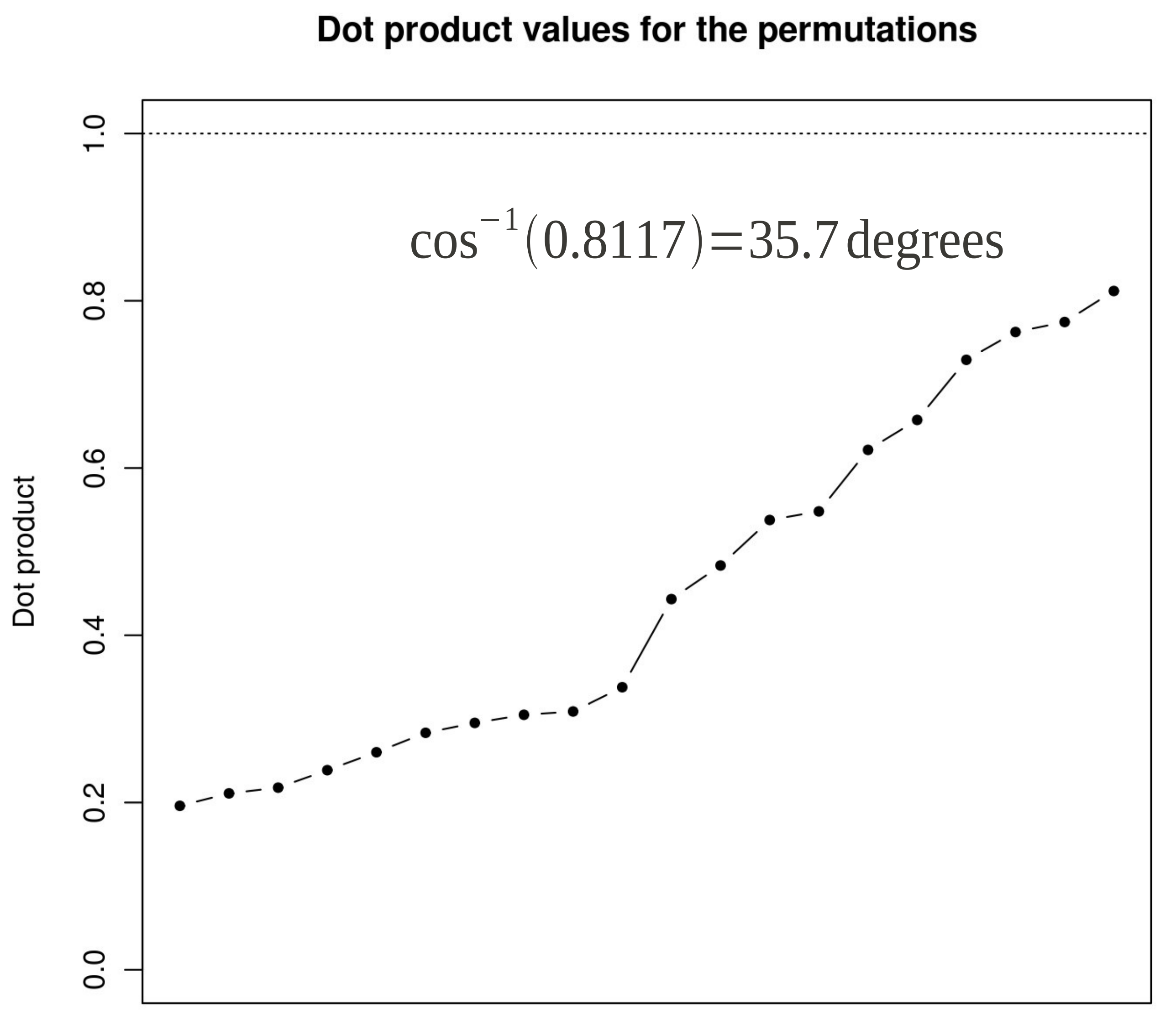
Simulated CPC(2) data, k = 2, p = 5, n = 1000

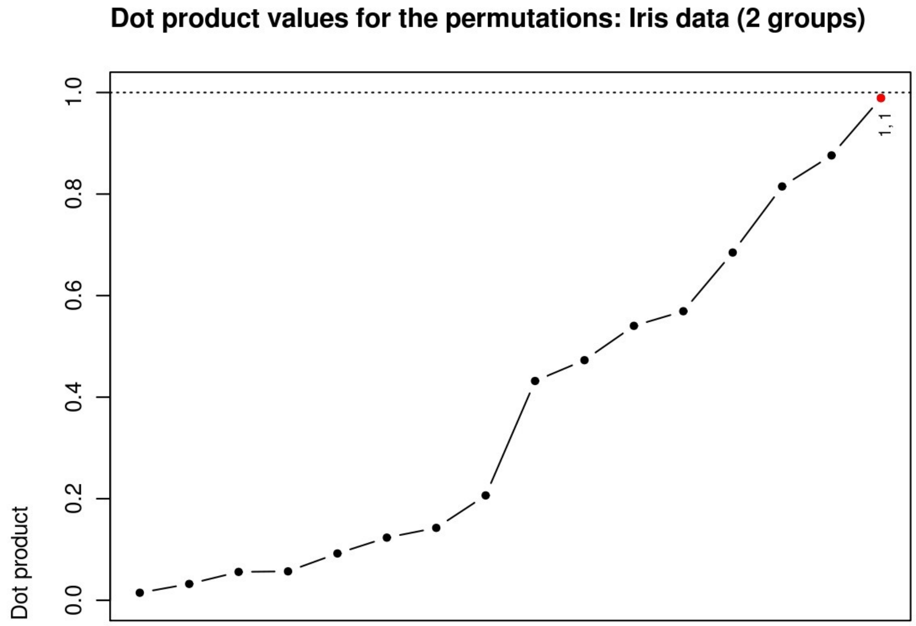
	Dot products
1 1	**0.996**
2 2	**0.992**
5 5	0.912
3 4	0.765
4 3	0.699
4 4	0.636
3 3	0.576
5 3	0.410
4 5	0.323
3 5	0.252
3 2	0.115
2 3	0.100
3 1	0.082
1 4	0.077
2 4	0.067
4 2	0.037
1 3	0.033
2 5	0.029
4 1	0.024
1 2	0.022

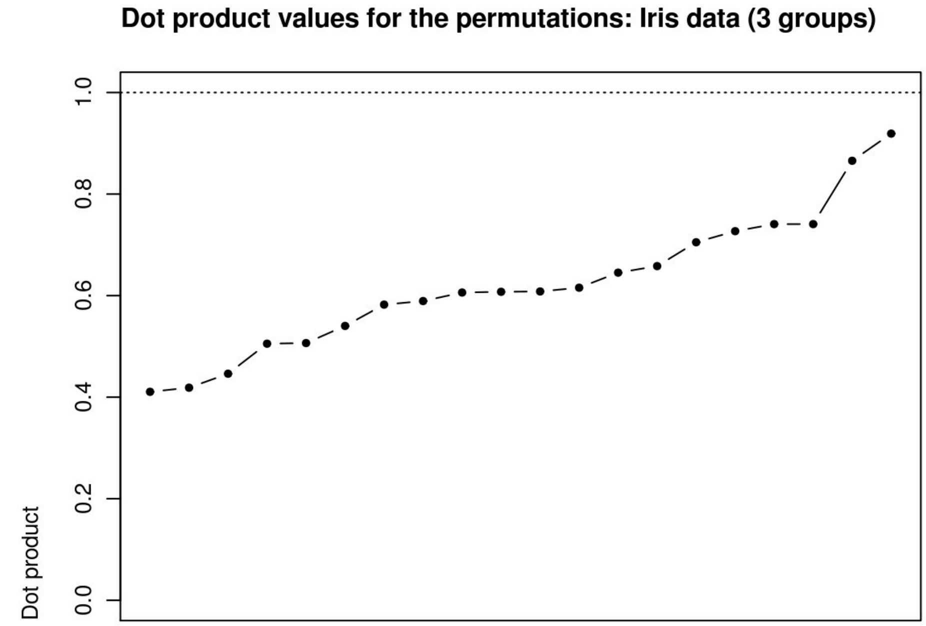


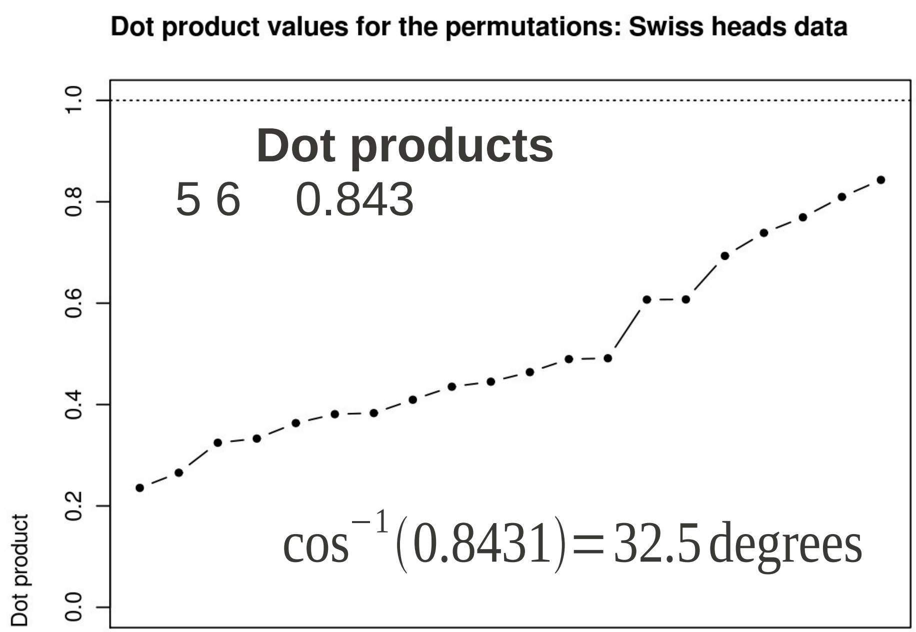
Simulated heterogeneous data, k = 2, p = 5, n = 1000

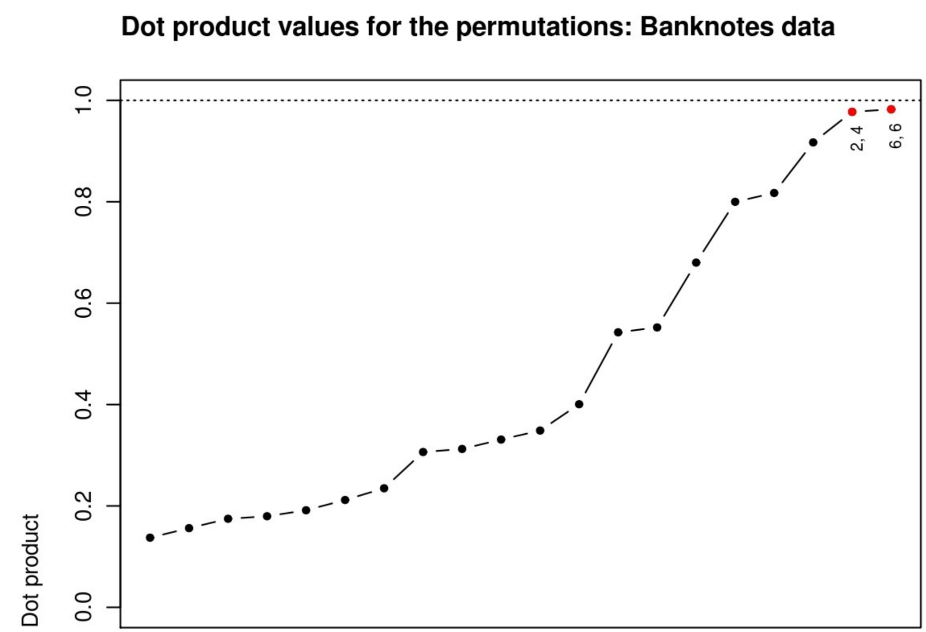
Dot products	
4 1	0.812
3 2	0.775
5 5	0.763
2 3	0.729
1 4	0.657
1 2	0.622
5 3	0.548
2 4	0.538
4 5	0.483
3 4	0.443
2 1	0.338
1 1	0.309
4 3	0.305
3 1	0.295
3 5	0.283
5 4	0.260
2 5	0.239
1 5	0.218
5 1	0.211
1 3	0.196




Dot product values for the permutations: Iris data (2 groups)
Dot product
0.0
0.2
0.4
0.6
0.8
1.0
1,1


Dot product values for the permutations: Iris data (3 groups)
Dot product
0.0
0.2
0.4
0.6
0.8
1.0


Dot product values for the permutations: Swiss heads data
Dot product
0.0
0.2
0.4
0.6
0.8
1.0
Dot products
5 6 0.843
$\cos^{-1}(0.8431) = 32.5\,\text{degrees}$


Dot product values for the permutations: Banknotes data
Dot product
0.0
0.2
0.4
0.6
0.8
1.0
2,4
6,6

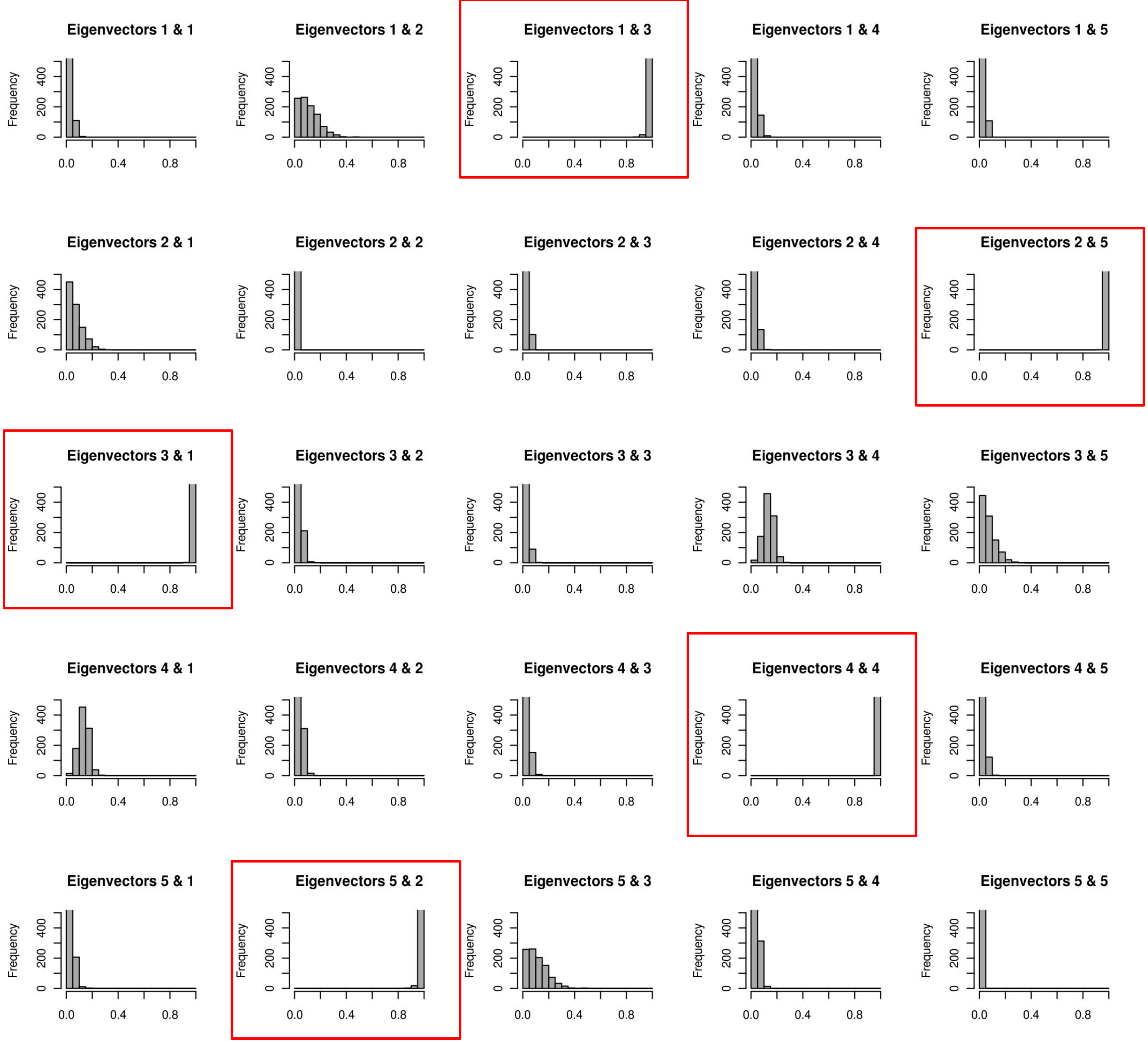
Simulated CPC data:

$k = 2$

$p = 5$

$n = 1000$

bootstrap reps = 1000



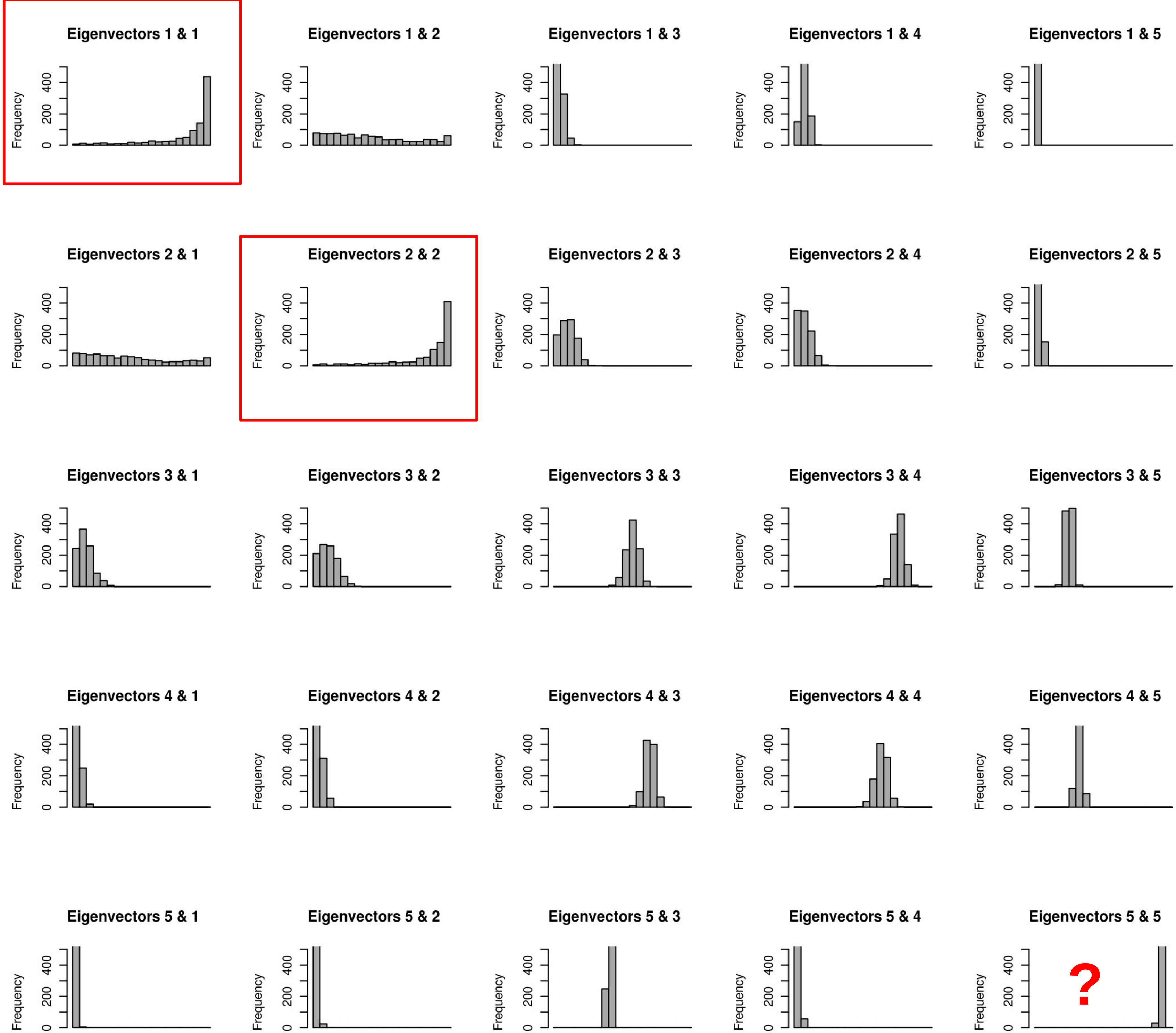
Simulated CPC(2) data:

$k = 2$

$p = 5$

$n = 1000$

bootstrap reps = 1000



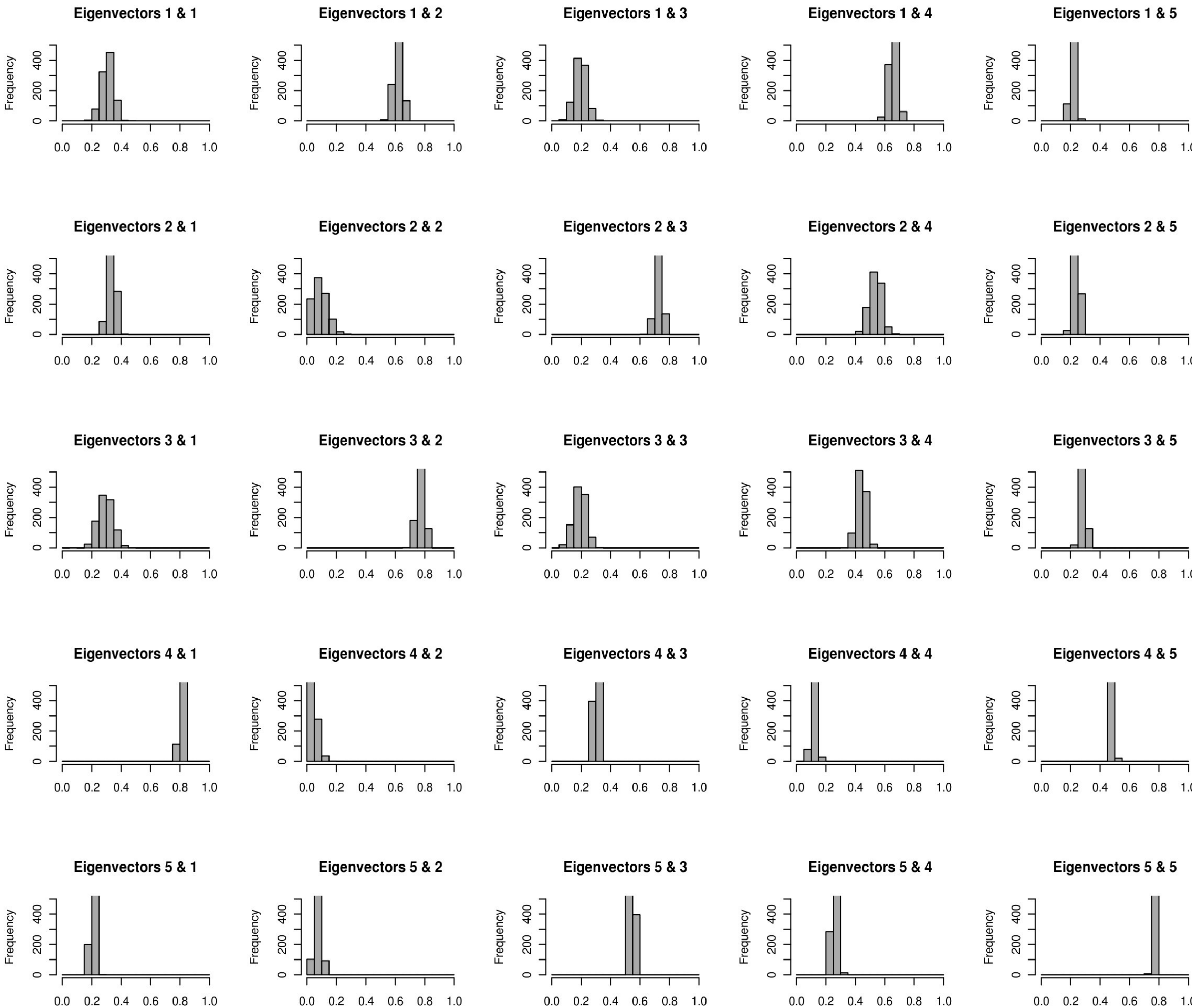
Simulated heterogeneous data:

$k = 2$

$p = 5$

$n = 1000$

bootstrap reps = 1000



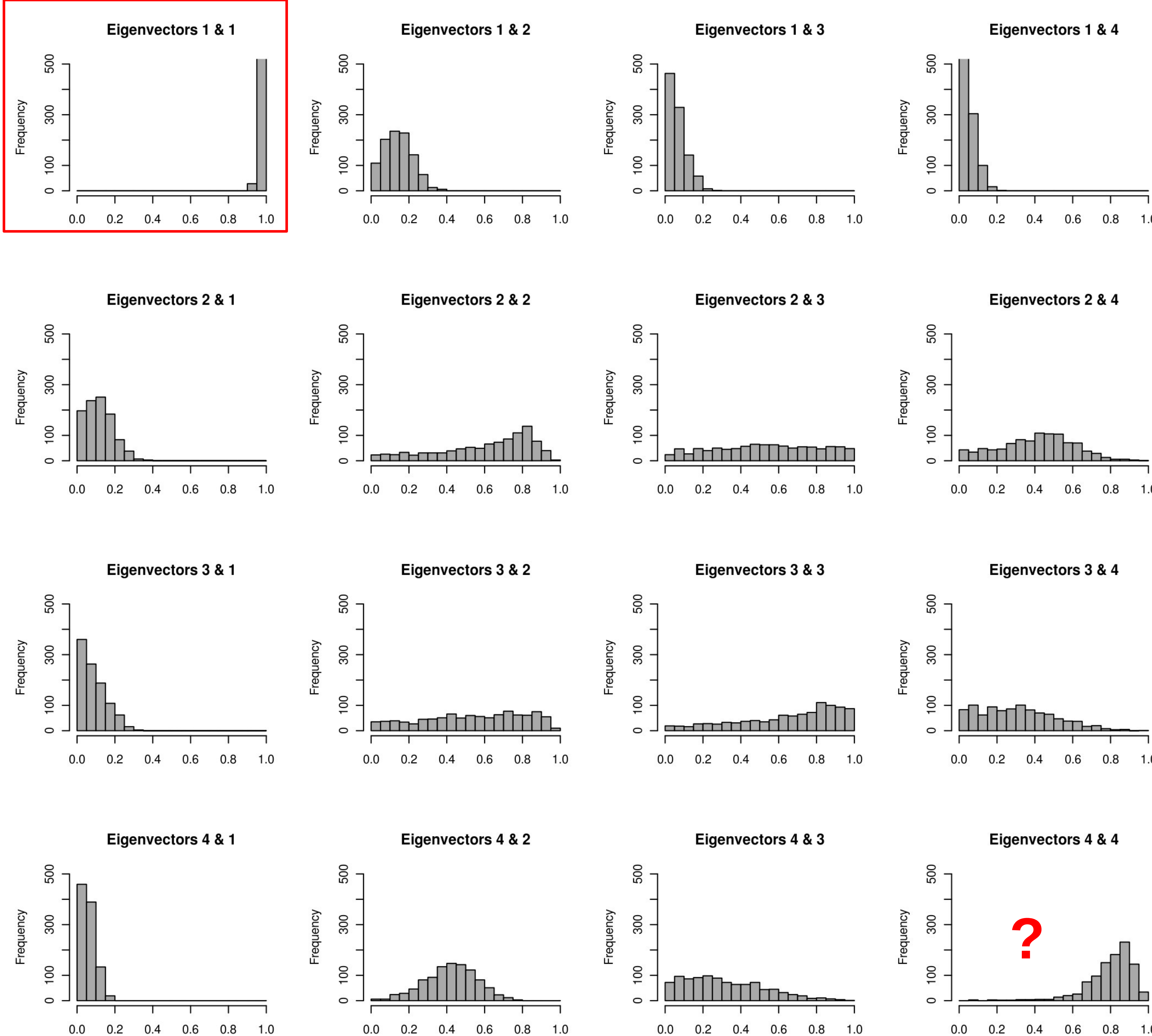
Iris data (two groups):

$k = 2$

$p = 4$

$n = 50$

bootstrap reps = 1000



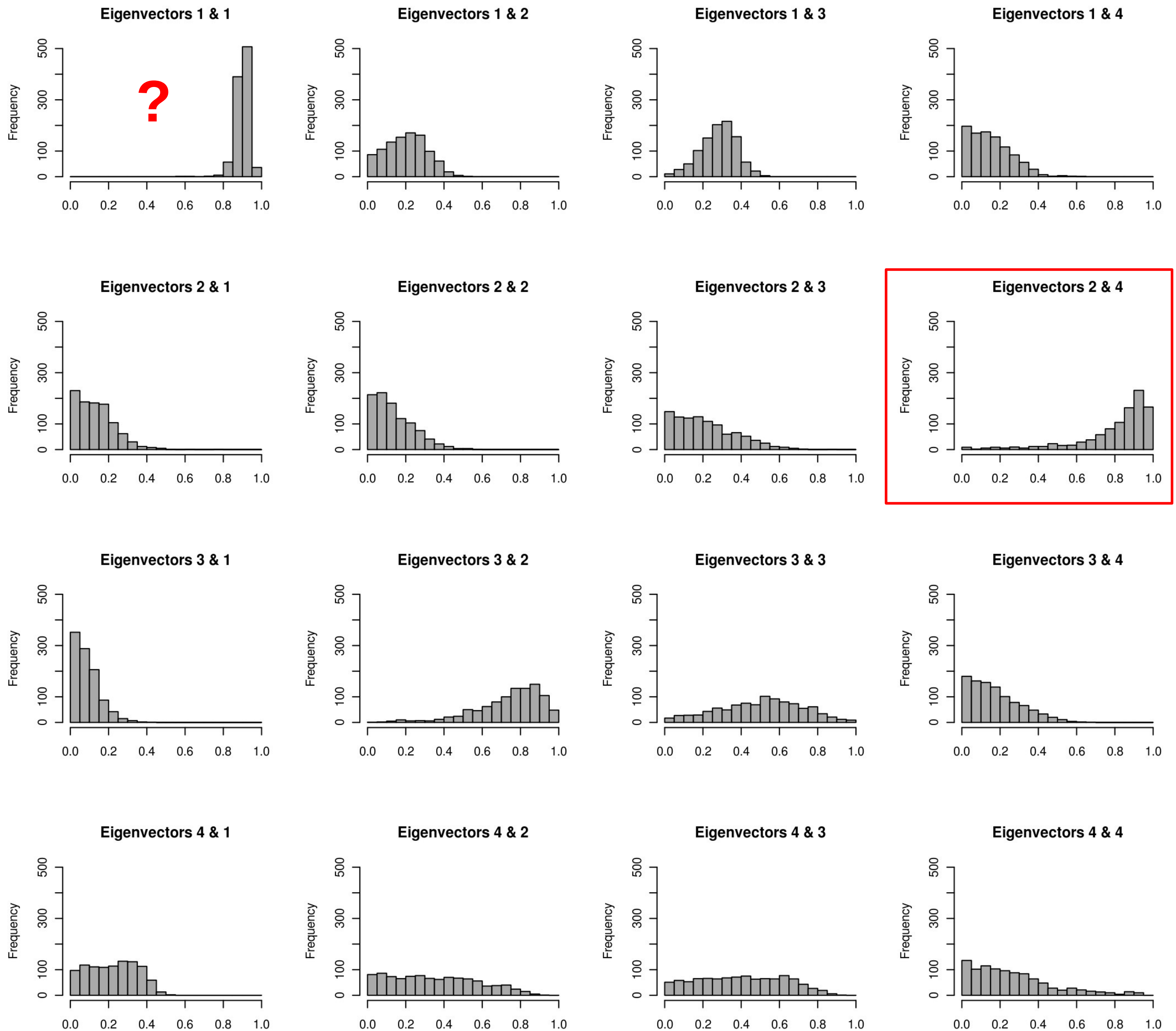
Banknotes data:

$k = 2$

$p = 6$ (only 4 shown)

$n = 100$

bootstrap reps = 1000



3) Simultaneous diagonalisation methods

- **FG algorithm** (Flury 1988)

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

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

  --> implemented in JADE package in R

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

Compared these with:

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

4) Applications of the CPC model

Advantages the CPC model might provide:

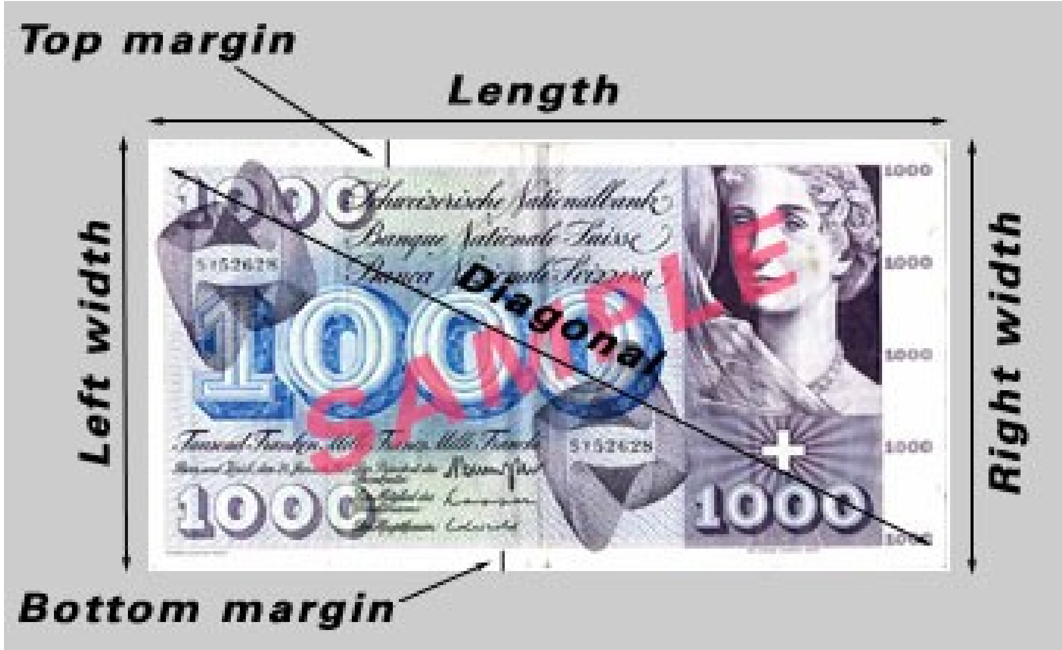
- ***more stable estimates*** than when incorrectly assuming heterogeneity of covariance matrices
- ***more accurate estimates*** than when incorrectly assuming equality of covariance matrices

Possible applications of the CPC model:

1) Biplots
2) Regression
3) Better estimator of $\boldsymbol{\Sigma}$

$$\hat{\boldsymbol{\Sigma}} = \alpha \boldsymbol{S} + (1-\alpha) \boldsymbol{S}_{\mathbf{CPC}}$$

4) Discriminant analysis



X_1: Length of the bank note,

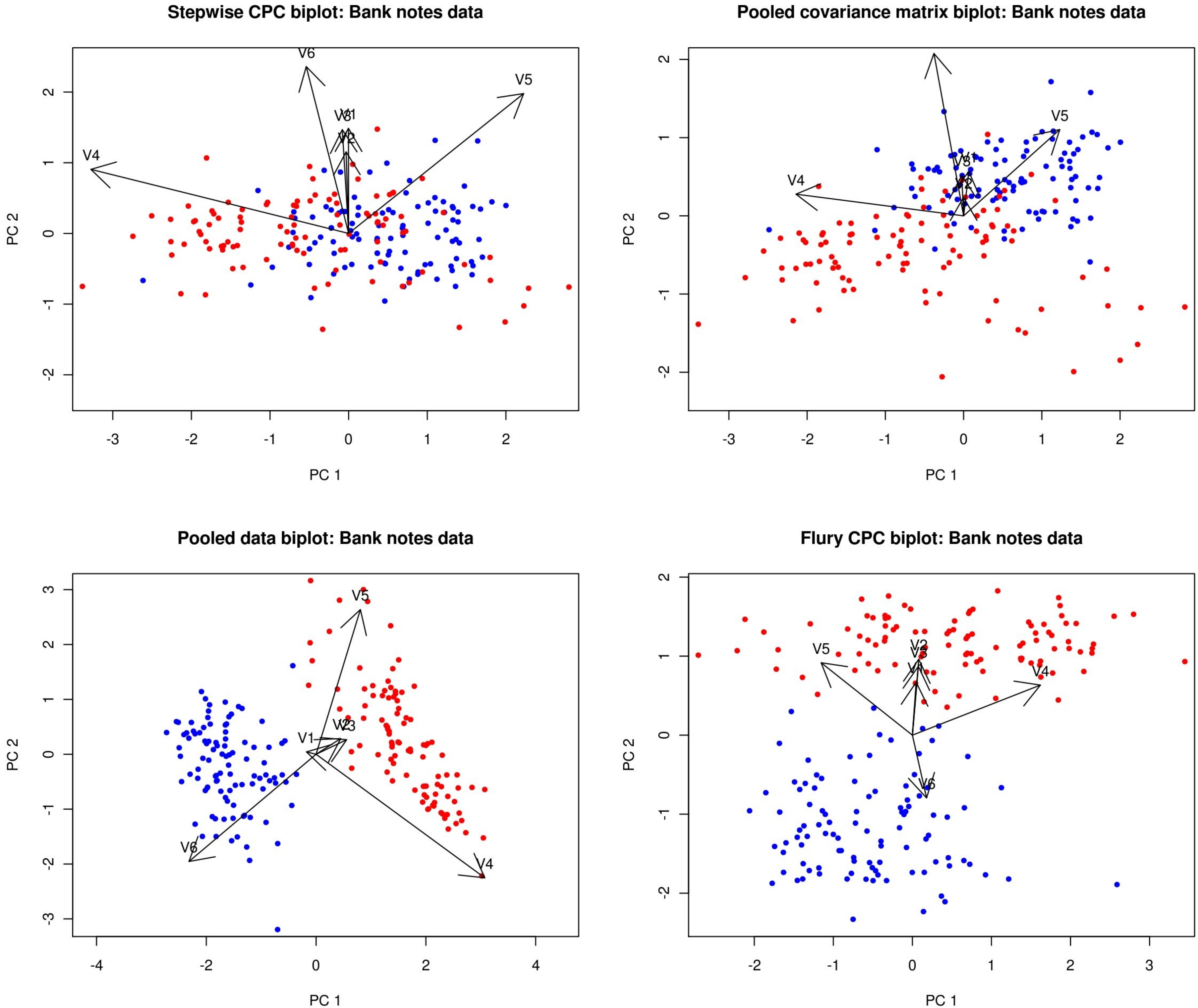
X_2: Height of the bank note, measured on the left,

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
PC 1
PC 2
V1
V2
V3
V4
V5
V6

Biplot goodness of fit

Overall quality of the display (Gower, Le Roux & Lubbe 2010)

Let $\boldsymbol{X}$ contain the data from all *k* groups, with the columns of $\boldsymbol{X}$ centred to have zero means, and letting $\|\boldsymbol{X}\|^2 = tr(\boldsymbol{X}'\boldsymbol{X})$, the total variation in the data can be partitioned as follows:

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

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

Biplot goodness of fit

Within group variation

Letting $\boldsymbol{X}_{\boldsymbol{i}}$ contain the data from the *i*th group, with the columns of $\boldsymbol{X}_{\boldsymbol{i}}$ centred to have zero mean (**for the *i*th group**), the quality of representation of the *within group variation* can be measured as follows:

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

Biplot goodness of fit

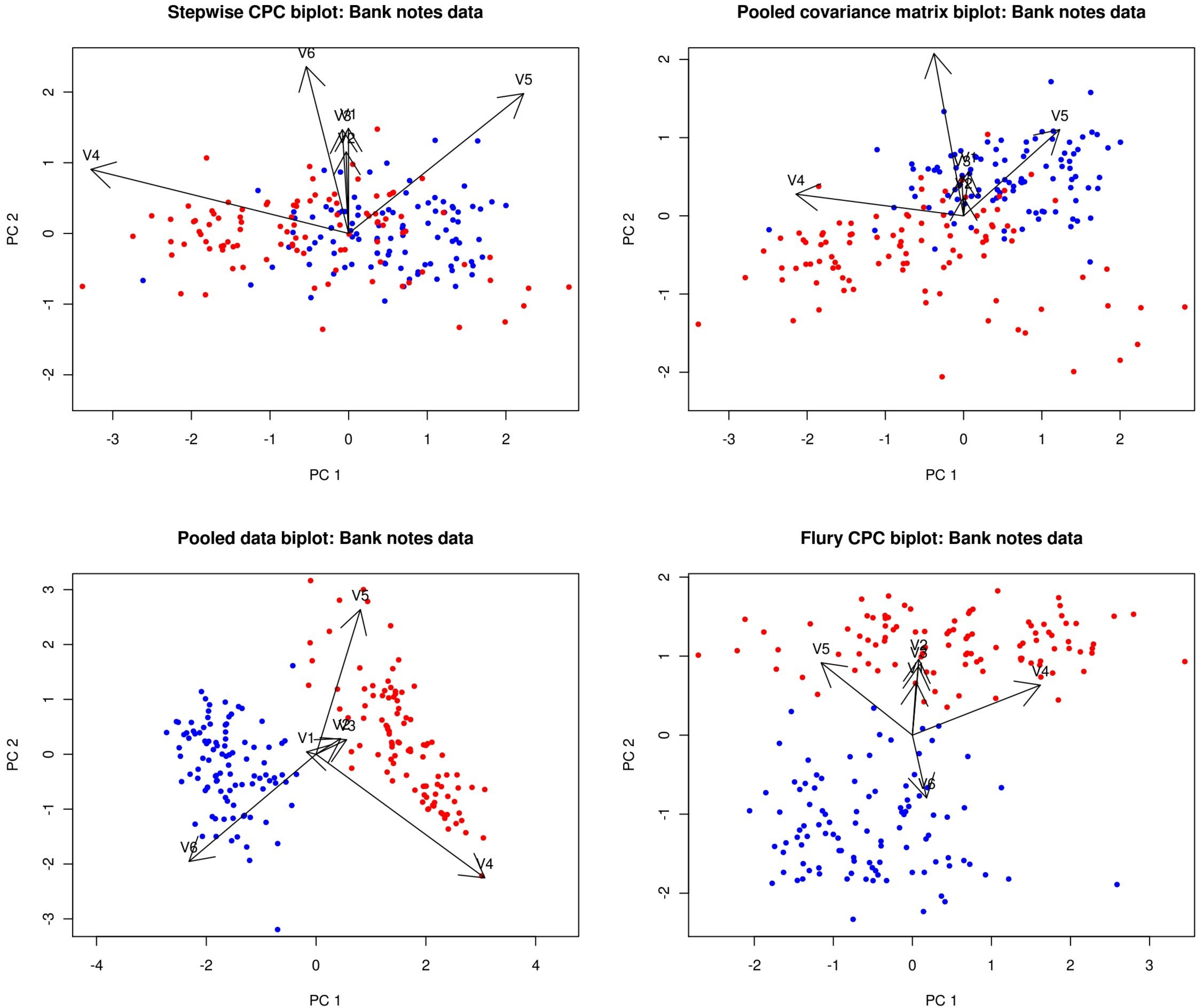
Adequacy of the variables (Gower, Le Roux & Lubbe 2010)

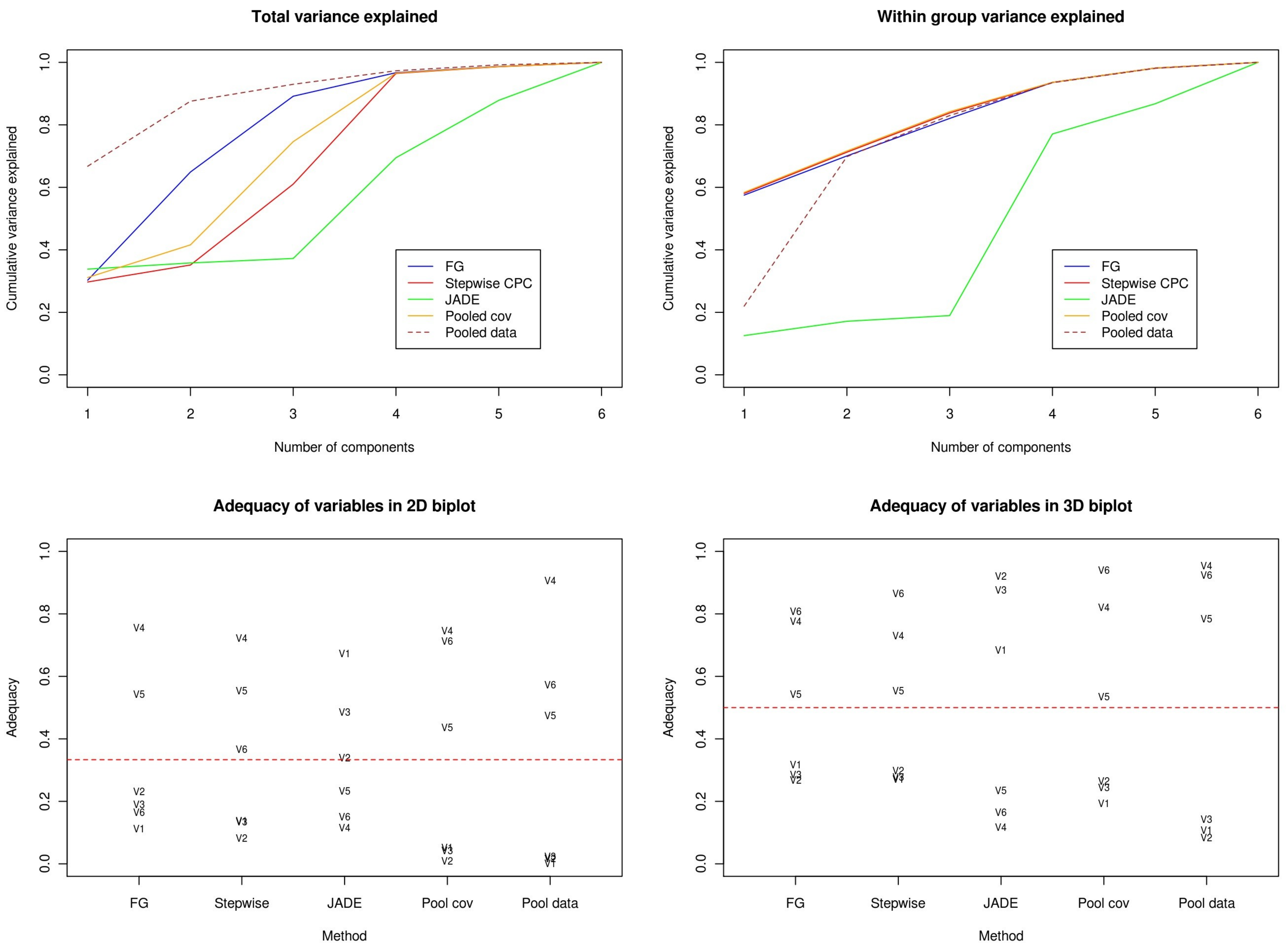
- Quality of representation of the variables in the biplot

Letting $\boldsymbol{B}_{[r]}$ contain the first *r* columns of orthogonal projection matrix $\boldsymbol{B}$ (with unit length row vectors), the adequacy of the *p* variables in the *r*-dimensional subspace will be given by

$$diag\left(\boldsymbol{B}_{[r]}\boldsymbol{B}_{[r]}'\right)$$

Mean adequacy over all *p* variables = $\dfrac{r}{p}$


Stepwise CPC biplot: Bank notes data
V6
V5
V3
V1
V2
V4
PC 1
PC 2
Pooled covariance matrix biplot: Bank notes data
V5
V4
V3
V1
V2
PC 1
PC 2
Pooled data biplot: Bank notes data
V5
V1
V2
V3
V6
V4
PC 1
PC 2
Flury CPC biplot: Bank notes data
V5
V2
V3
V1
V4
V6
PC 1
PC 2


Total variance explained
Cumulative variance explained
Number of components
FG
Stepwise CPC
JADE
Pooled cov
Pooled data
Within group variance explained
Cumulative variance explained
Number of components
FG
Stepwise CPC
JADE
Pooled cov
Pooled data
Adequacy of variables in 2D biplot
Adequacy
Method
FG
Stepwise
JADE
Pool cov
Pool data
Adequacy of variables in 3D biplot
Adequacy
Method
FG
Stepwise
JADE
Pool cov
Pool data

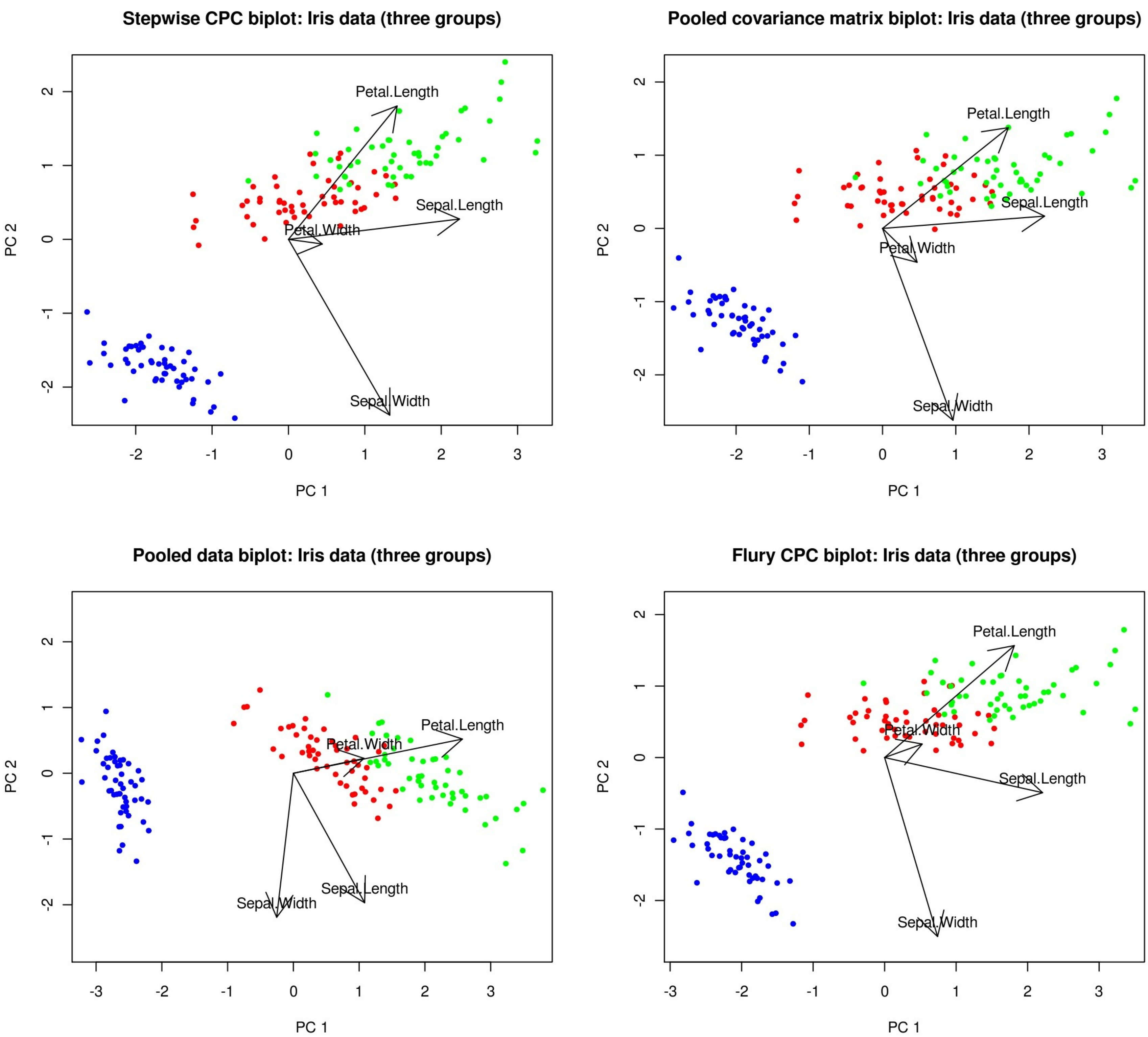
Setosa

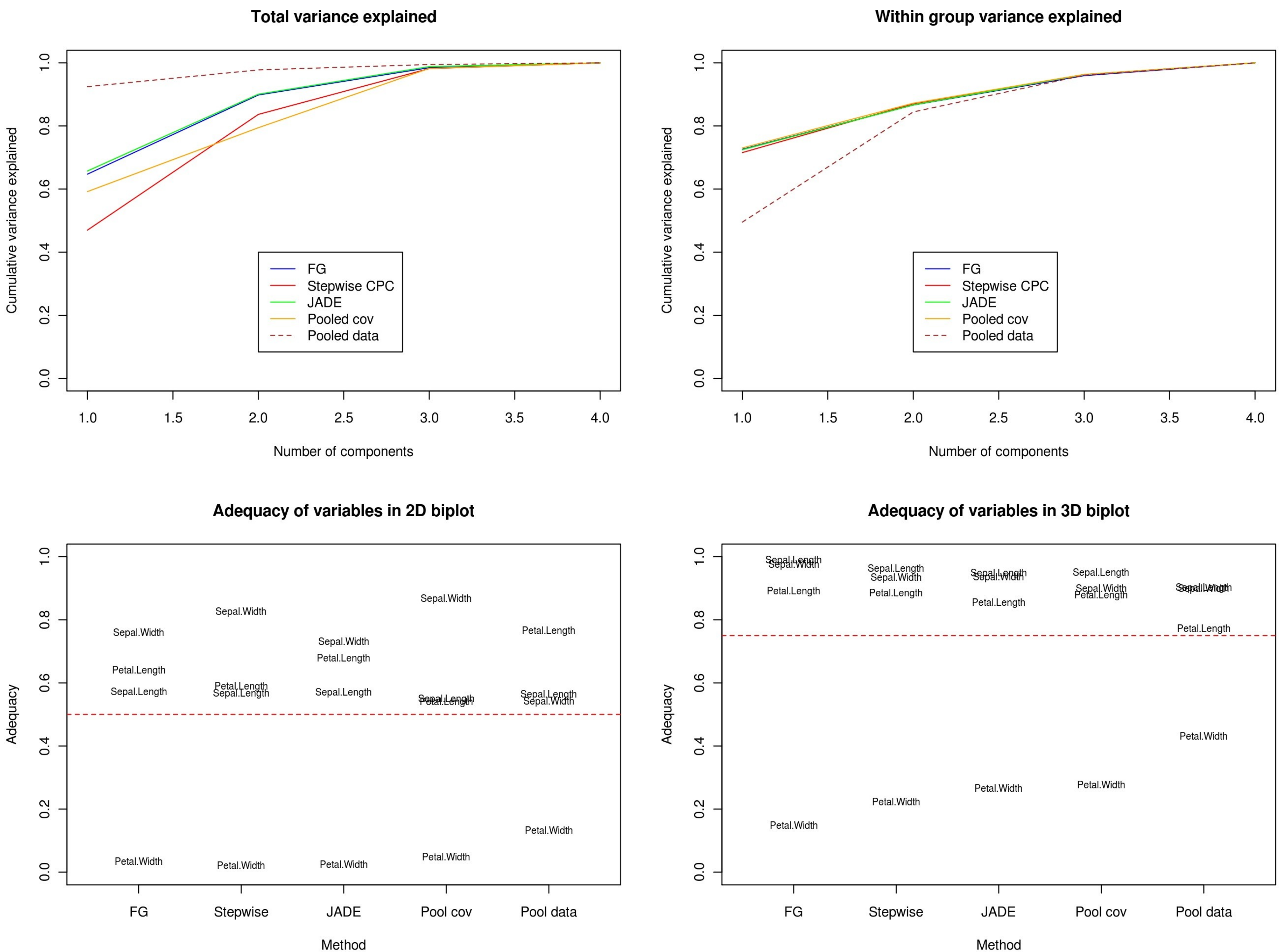
Versicolor

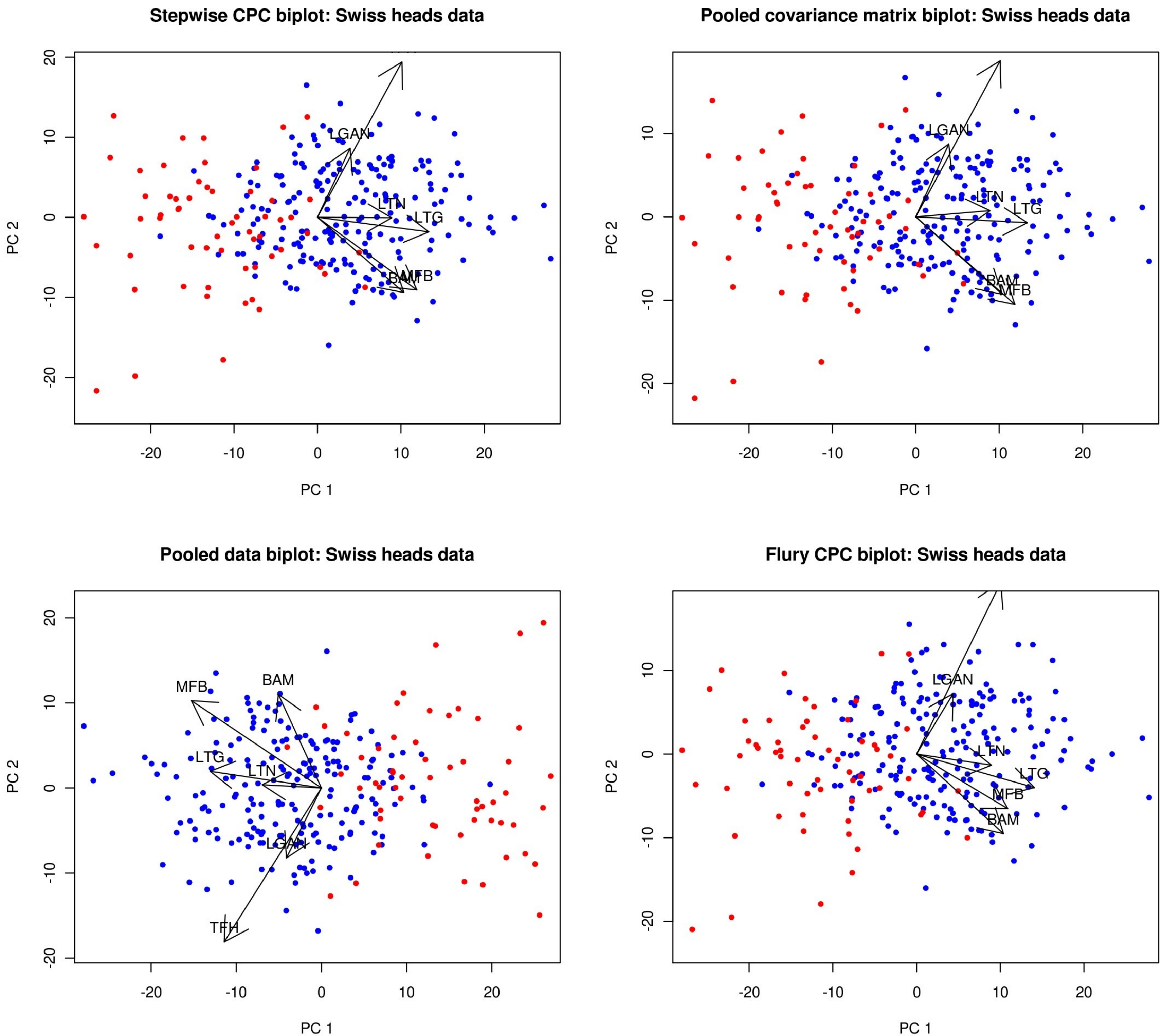
Virginica

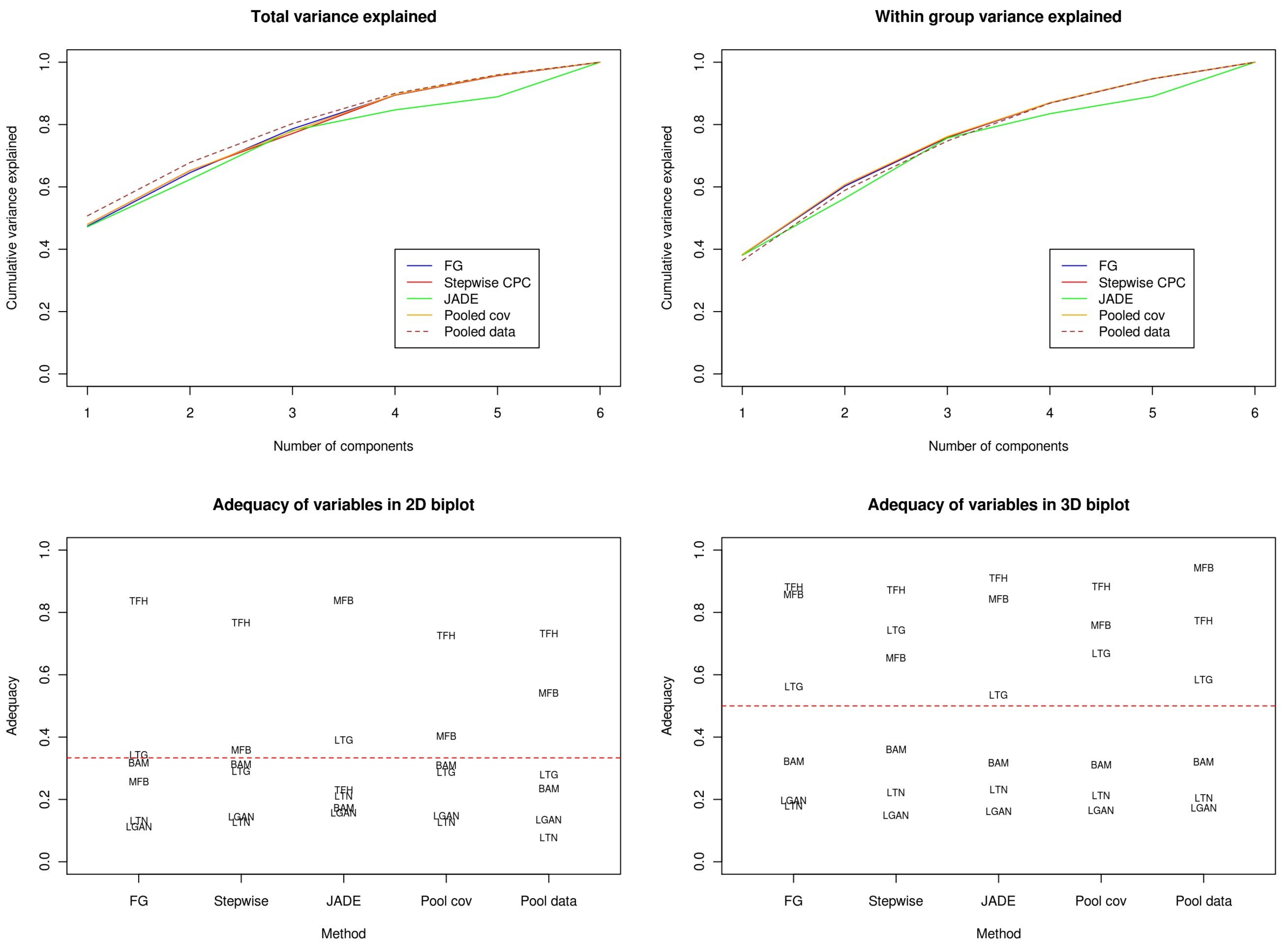
<u>Source</u>: Wikimedia Commons

(Anderson's iris data)


Stepwise CPC biplot: Iris data (three groups)
Pooled covariance matrix biplot: Iris data (three groups)
Pooled data biplot: Iris data (three groups)
Flury CPC biplot: Iris data (three groups)
Petal.Length
Sepal.Length
Petal.Width
Sepal.Width
PC 1
PC 2


Total variance explained
Cumulative variance explained
Number of components
FG
Stepwise CPC
JADE
Pooled cov
Pooled data
Within group variance explained
Cumulative variance explained
Number of components
FG
Stepwise CPC
JADE
Pooled cov
Pooled data
Adequacy of variables in 2D biplot
Adequacy
Method
FG
Stepwise
JADE
Pool cov
Pool data
Adequacy of variables in 3D biplot
Adequacy
Method
FG
Stepwise
JADE
Pool cov
Pool data


Stepwise CPC biplot: Swiss heads data
Pooled covariance matrix biplot: Swiss heads data
Pooled data biplot: Swiss heads data
Flury CPC biplot: Swiss heads data
PC 1
PC 2
LGAN
LTN
LTG
BAM
MFB
TFH


Total variance explained
Cumulative variance explained
Number of components
FG
Stepwise CPC
JADE
Pooled cov
Pooled data
Within group variance explained
Cumulative variance explained
Number of components
FG
Stepwise CPC
JADE
Pooled cov
Pooled data
Adequacy of variables in 2D biplot
Adequacy
Method
FG
Stepwise
JADE
Pool cov
Pool data
TFH
MFB
LTG
BAM
LTN
LGAN
Adequacy of variables in 3D biplot
Adequacy
Method
FG
Stepwise
JADE
Pool cov
Pool data
TFH
MFB
LTG
BAM
LTN
LGAN

Questions?