# Comparison of some methods for the identification of common eigenvectors

Theo Pepler
Unit for Biometry
University of Stellenbosch

8 November 2012

# The CPC model

**Principal component analysis** (PCA):

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

**Common principal components** (CPC):

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

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

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

$$\begin{array}{lll} \mathbf{\Sigma}_1 = B_1\mathbf{\Lambda}_1 B_1' & \text{where} & B_1 = [\boldsymbol{b}_1 \ldots \boldsymbol{b}_q : \boldsymbol{b}_{q+1(1)} \ldots \boldsymbol{b}_{p(1)}] \\ \mathbf{\Sigma}_2 = B_2\mathbf{\Lambda}_2 B_2' & & B_2 = [\boldsymbol{b}_1 \ldots \boldsymbol{b}_q : \boldsymbol{b}_{q+1(2)} \ldots \boldsymbol{b}_{p(2)}] \end{array}$$

# Flury's (1988) methods

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

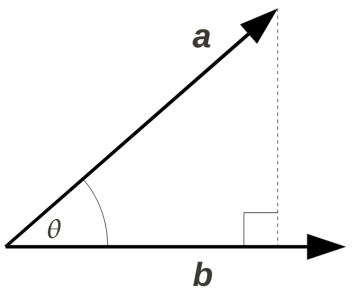
| Model: Higher | Model: Lower | $X^2$ | df | $\frac{X^2}{\text{df}}$ | AIC for Higher Model |
|---|---|---|---|---|---|
| Equality | Proportionality | 42.29 | 1 | 42.29 | 89.78 |
| Proportionality | CPC | 25.66 | 5 | 5.13 | 49.49 |
| CPC | CPC(1) | 15.12 | 10 | 1.51 | 33.82* |
| CPC(1) | Unrelated |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

*Minimum AIC.

- $\chi^2$ statistics not *independent*, and depend on *multivariate normality assumption*
- AIC not *formal hypothesis test*

# Vector correlations

**Different approach (Krzanowski 1979)**



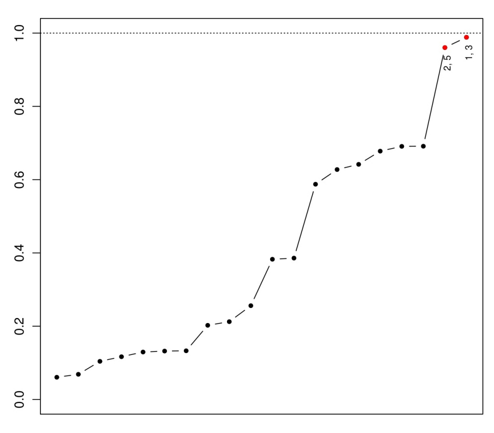

$$\boldsymbol{a}'\boldsymbol{b} = \cos\theta$$

- Inspect *vector correlations* from pairwise combinations of all $p$ eigenvectors from the $k$ groups

# Vector correlations

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

| | | Correlations |
|---|---|---|
| 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 |



# Vector correlations

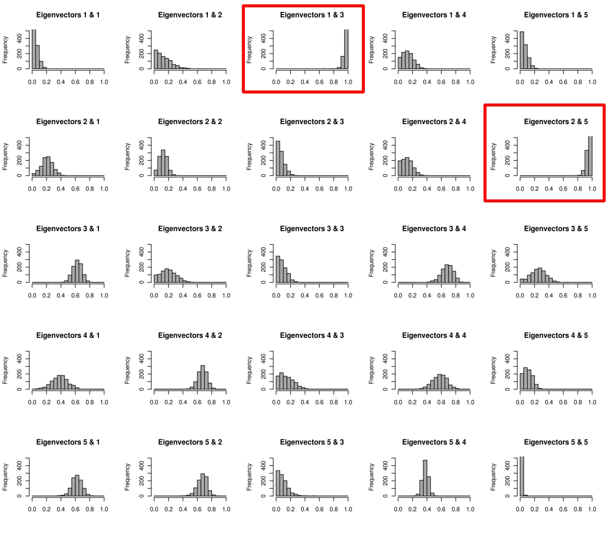
**Simulated CPC(2) data:**

$k = 2$

$p = 5$

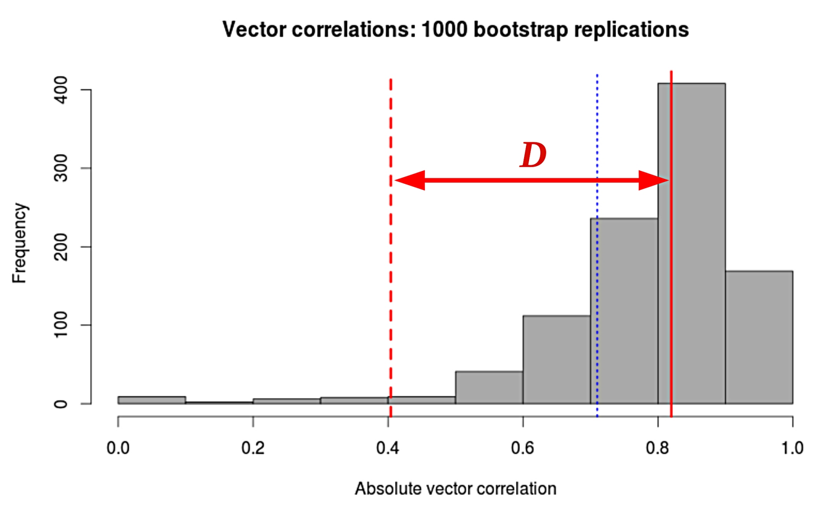
$n = 200$

bootstrap reps $= 1000$



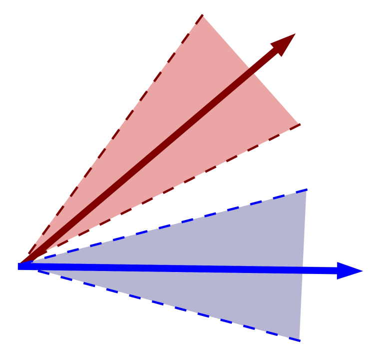
# BVD method

## Bootstrap vector correlation distribution (BVD) method

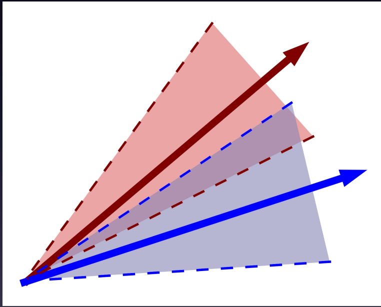


# BCR method

**Bootstrap confidence regions (BCR) method**



(a) Not common

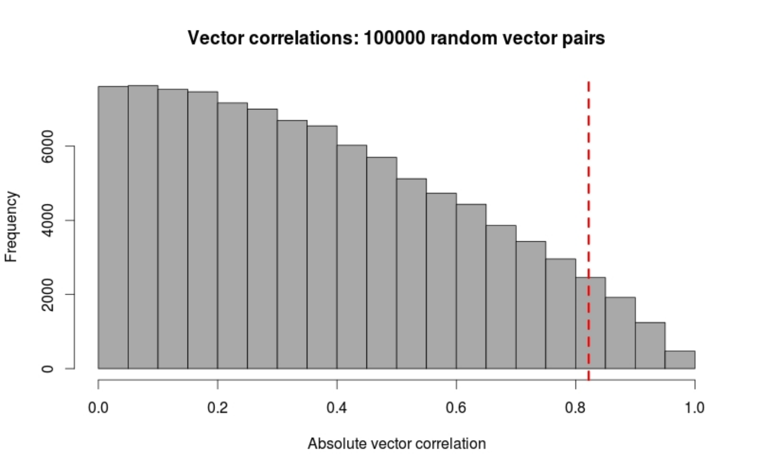


(b) Common

# Klingenberg & McIntyre (1998)

**Random vector correlations (RVC) method**

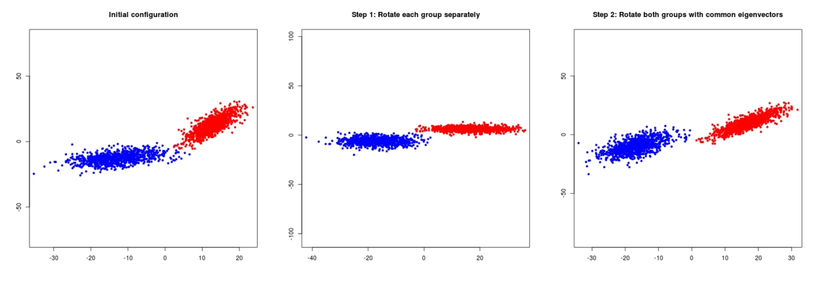
$H_0$ : pair of eigenvectors are *not* common



# Klingenberg (1996, 1998)

**Bootstrap hypothesis test (BootTest)**

$$H_0: \text{ pair of eigenvectors are common}$$



# Ensemble method

**Ensemble method**

*Majority vote* on the number of common eigenvectors from

- Flury's AIC
- BVD method
- BCR method
- Klingenberg's RVC method

**Ties:** choose *higher* model in Flury's hierarchy

# Simulation study

**Simulation study (12000 runs)**

- Groups: $k = 2$
- Variables: $p = 5$
- Sample sizes: $n = 50, 100, 200, 500, 1000$
- Eigenvalues: poorly / moderately / well separated
- Normality: multivariate normal / non-normal
- Covariance structures: CPC, CPC(3), CPC(1), heterogeneous

# Simulation study

### Number of common eigenvectors correctly identified (%)

| | AIC | $\chi^2$ | BVD | BCR | RVC | BootTest | Ensemble |
|---|---|---|---|---|---|---|---|
| **Sample size** | | | | | | | |
| $n = 50$ | 36 | 28 | 35 | 26 | 35 | 13 | **40** |
| $n = 100$ | 41 | 27 | 38 | 27 | 43 | 10 | **47** |
| $n = 200$ | 47 | 31 | 48 | 42 | 56 | 7 | **58** |
| $n = 500$ | 50 | 32 | 64 | 65 | 69 | 6 | **71** |
| $n = 1000$ | 51 | 34 | 74 | 74 | 74 | 4 | **76** |
| **Data** | | | | | | | |
| Normal | 51 | 33 | 55 | 49 | 59 | 7 | **62** |
| Non-normal | 39 | 28 | 48 | 44 | 52 | 9 | **55** |
| **Total** | 45 | 31 | 51 | 47 | 56 | 8 | **59** |

All methods fared worse with non-normal data than with normal data.

# Simulation study

## Number of common eigenvectors correctly identified (%)

| | AIC | $\chi^2$ | BVD | BCR | RVC | BootTest | Ensemble |
|---|---|---|---|---|---|---|---|
| **Eigenvalue separation** | | | | | | | |
| Poor | 25 | 26 | 26 | 25 | 23 | 18 | **27** |
| Moderate | 51 | 31 | 57 | 49 | 63 | 5 | **67** |
| Good | 59 | 35 | 72 | 66 | 80 | 1 | **82** |
| **Covariance structure** | | | | | | | |
| CPC | 45 | 26 | 52 | **98** | 55 | 13 | 86 |
| CPC(3) | 34 | 21 | 28 | 29 | **44** | 17 | 37 |
| CPC(1) | 44 | 45 | 46 | 29 | **61** | 2 | 49 |
| Heterogeneous | 57 | – | **80** | 30 | 63 | 0 | 61 |
| **Total** | 45 | 31 | 51 | 47 | 56 | 8 | **59** |

# Swiss heads data

**Swiss heads data:** $k = 2$, $p = 6$

**Sample sizes:** $n_1 = 200$, $n_2 = 59$

**Eigenvalues:**

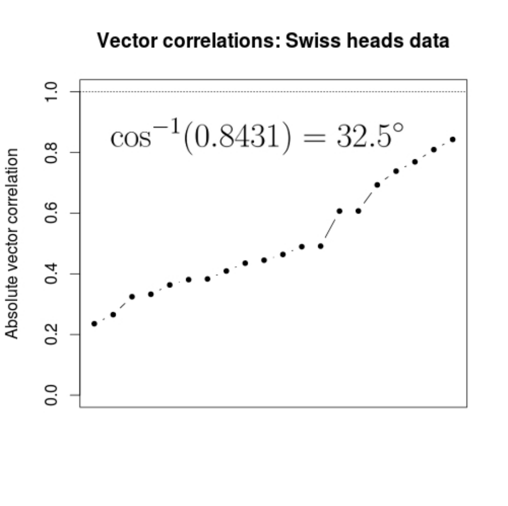
- Males: 66.3, 34.4, 19.6, 14.3, 13.0, 6.8
- Females: 73.5, 59.6, 42.0, 28.0, 15.6, 10.9
  (well separated in both groups)

**Normality:**

- Shapiro-Wilk test: Males ($p = 0.0003$), Females ($p = 0.0008$)

# Swiss heads data

| | | Correlations |
|---|---|---|
| 5 | 6 | 0.84 |
| 6 | 5 | 0.81 |
| 2 | 3 | 0.77 |
| 1 | 2 | 0.74 |
| 3 | 4 | 0.69 |
| 4 | 4 | 0.61 |
| 1 | 1 | 0.61 |
| 2 | 2 | 0.49 |
| 3 | 1 | 0.49 |
| 4 | 1 | 0.46 |



# Swiss heads data

**Verdict on the number of common eigenvectors?**

- AIC: 4
- BVD: 0
- BCR: 6
- RVC: 3
- Ensemble: 6

# Conclusions

- increased accuracy with the non-parametric methods
- Flury's $\chi^2$ and Klingenberg's BootTest perform poorly—should rather not be used
- using an ensemble of the best methods gives best performance
- larger sample sizes needed to estimate eigenvectors accurately

# Sources


- P. Diaconis and B. Efron. *Computer-intensive methods in statistics*. Sci. Am.; (United States), 248(5): 116–130, 1983.
- B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall, 1993.
- B. Flury. *Common principal components and related multivariate models*. Wiley series in probability and mathematical statistics. Wiley, 1988.
- C. P. Klingenberg. *Multivariate allometry*. NATO ASI SERIES A LIFE SCIENCES, 284: pp. 23–50, 1996.
- C. P. Klingenberg and G. S. McIntyre. *Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with Procrustes methods*. Evolution, 52(5): pp. 1363–1375, 1998.
- W. J. Krzanowski. *Between-groups comparison of principal components*. Journal of the American Statistical Association, 74(367): pp. 703–707, 1979.