On the identification of common principal components

Theo Pepler
Genetics Department
University of Stellenbosch

10 July 2012

What are common principal components (CPCs)?

How can covariance structures of two groups differ?

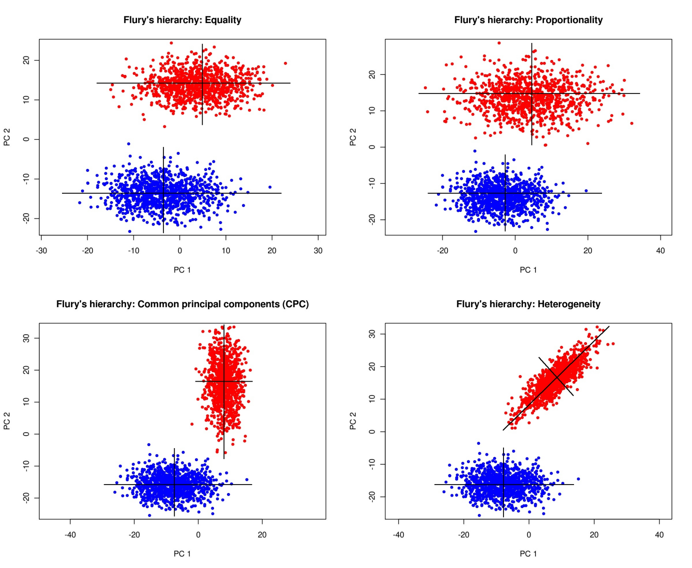
Univariate:

- Homoscedastic or heteroscedastic (nothing in between)

Multivariate case:

- Flury's hierarchy (1988):
  1. Equality $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$
  2. Proportionality $\mathbf{\Sigma}_1 = \rho\mathbf{\Sigma}_2$
  3. Common principal components
  4. Partial common principal components
  5. Heterogeneity

What are CPCs?



What are CPCs?

Principal component analysis (PCA):

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

Common principal components (CPC):

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

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

Partial common principal components (CPC(q)):

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

Identifying the CPCs

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

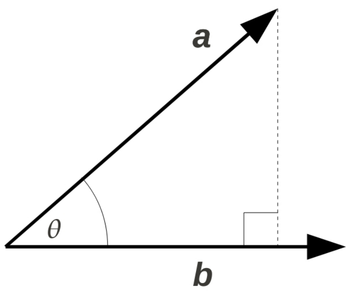
Model: Higher	Model: Lower	X^2	df	$\frac{X^2}{\text{df}}$	AIC for Higher Model
Equality	Proportionality	42.29	1	42.29	89.78
Proportionality	CPC	25.66	5	5.13	49.49
CPC	CPC(1)	15.12	10	1.51	33.82*
CPC(1)	Unrelated	6.70	5	1.34	38.70
Unrelated	---				42.0
Equality	Unrelated	89.78	21		

*Minimum AIC.

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

Identifying the CPCs

Different approach (Krzanowski 1979)



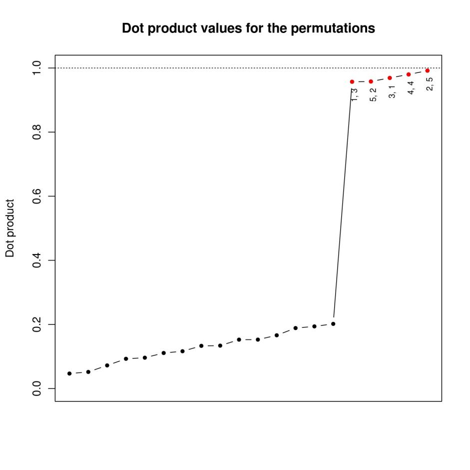
$\boldsymbol{a}'\boldsymbol{b} = \cos\theta$

- Inspect dot products from pairwise combinations of all p eigenvectors from the k groups

Identifying the CPCs

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

		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



Identifying the CPCs

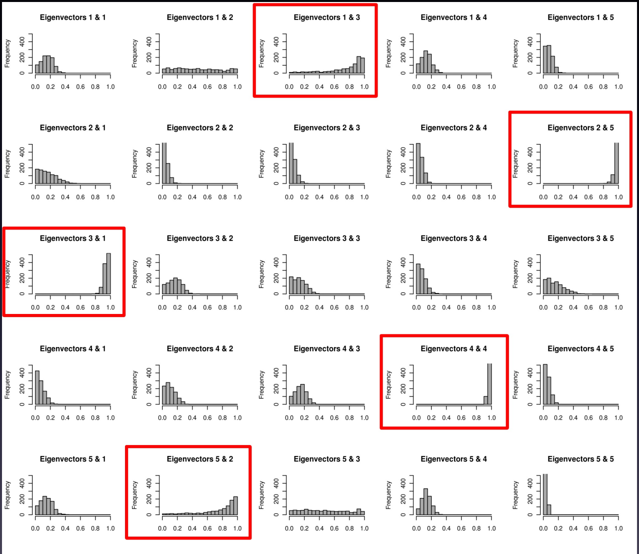
Simulated CPC data:

$k = 2$

$p = 5$

$n = 200$

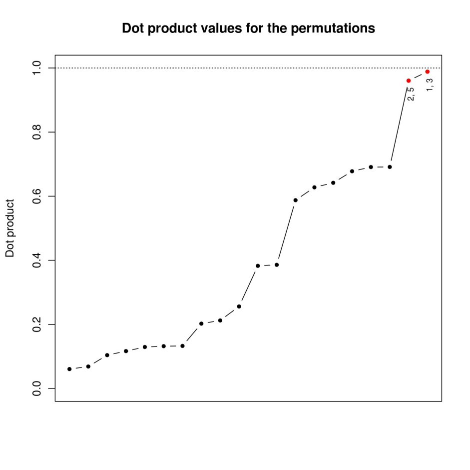
bootstrap reps $= 1000$



Identifying the CPCs

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



Identifying the CPCs

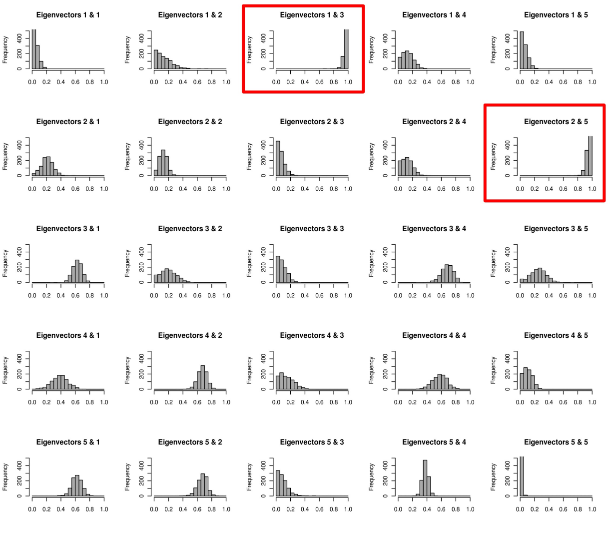
Simulated CPC(2) data:

$k = 2$

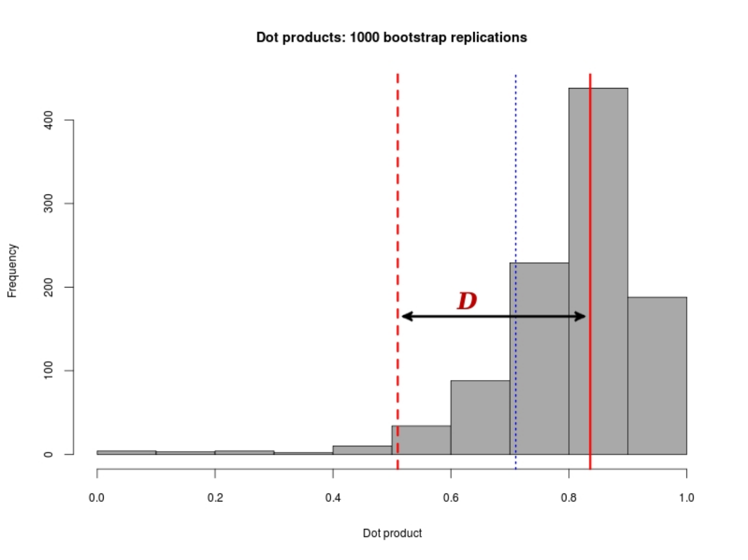
$p = 5$

$n = 200$

bootstrap reps $= 1000$



Identifying the CPCs



Identifying the CPCs

Bootstrap method:

1. Find median and 2.5^{th} percentile of bootstrap distribution
2. $D =$ median $- 2.5^{th}$ percentile
3. Then if
   - median > 0.71

     AND

   - median $+ D \geq 1$

   the two eigenvectors are deemed to be common

Simulation study

Simulation study

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

Simulation study

Number of components correctly identified (%)

	AIC	χ^2	Bootstrap
Sample size			
$n = 50$	32	27	31
$n = 100$	41	29	36
$n = 200$	46	33	47
$n = 500$	50	32	61
$n = 1000$	50	35	74
Data			
Normal	50	34	54
Non-normal	38	28	46
Total	**44**	**31**	**50**

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

Simulation study

Number of components correctly identified (%)

	AIC	χ^2	Bootstrap
Eigenvalue separation			
Poor (10%)	25	26	26
Moderate (50%)	49	32	53
Good (90%)	57	35	71
Covariance structure			
CPC	45	28	51
CPC(3)	34	20	28
CPC(1)	43	45	48
Heterogeneous	54	–	74
Total	**44**	**31**	**50**

Simulation study

Flury's AIC method:

- too greedy, especially for CPC(q)
- variability sometimes quite large
- best for small sample sizes
- performs well for full CPC with poorly separated eigenvalues

Simulation study

Flury's χ^2 method:

- poor performance overall
- large variability
- performed surprisingly well for CPC(1) with poorly separated eigenvalues

Simulation study

Bootstrap method:

- best for large sample sizes
- low variability
- does not perform well with poorly separated eigenvalues
- best for well separated eigenvalues, especially for non-normal data

Swiss heads data

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

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

Eigenvalues:

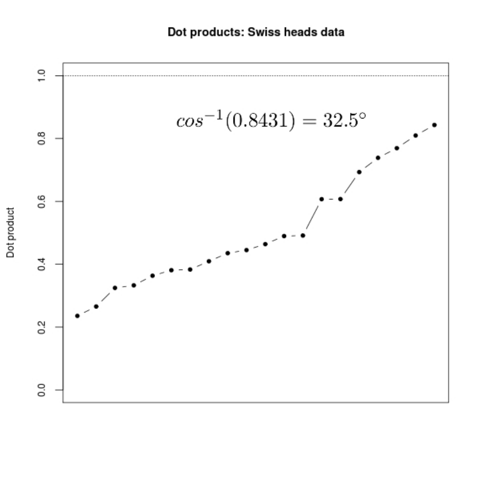
- Males: 66.3, 34.4, 19.6, 14.3, 13.0, 6.8
- Females: 73.5, 59.6, 42.0, 28.0, 15.6, 10.9
  (well separated in both groups)

Normality:

- Box's M test: $p < 0.0001$ (*not* multivariate normal)

Swiss heads data

		Dot products
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?

- Flury's AIC: 4
- Flury's χ^2: 3
- Bootstrap method: 0

Conclusions

- For smaller sample sizes and/or poorly separated eigenvalues
  $\rightarrow$ use Flury's AIC
- For larger sample sizes and well separated eigenvalues
  $\rightarrow$ use Bootstrap method
- Do not use Flury's χ^2 method

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.
- W. J. Krzanowski. *Between-groups comparison of principal components*. Journal of the American Statistical Association, 74(367): pp. 703-707, 1979.