Applications of the Common Principal Component model: How borrowing information about multivariate data structure can be useful

Theo Pepler
University of Glasgow

Stirling University
September 13, 2016

Outline of the talk

1. **Models and methods:**
   - The CPC and partial CPC models
   - Estimation of common eigenvectors
   - Identifying common eigenvectors

2. **Applications of the CPC model:**
   - Covariance matrix estimation
   - Discriminant analysis
   - Biplots
   - Time series

How can (co)variance structures of two groups differ?

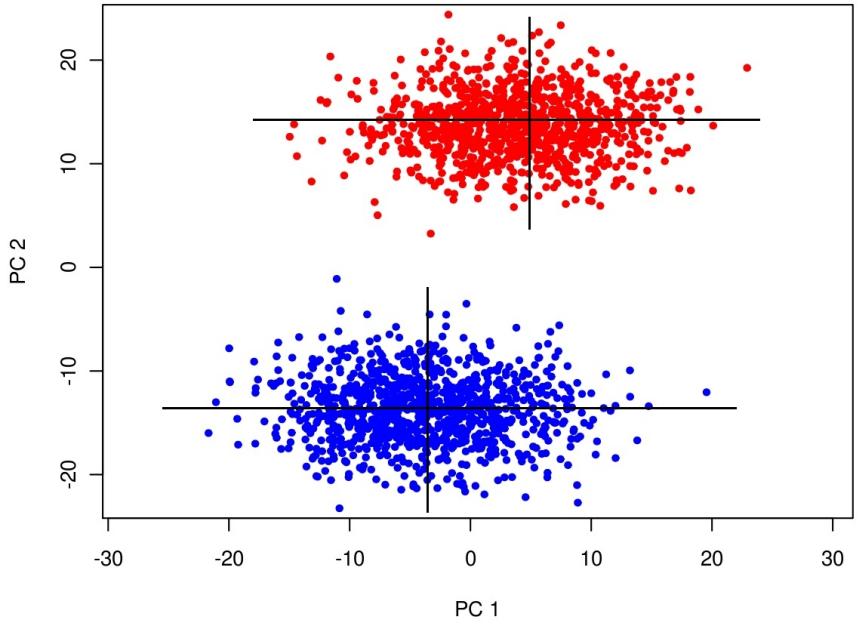
Univariate:

- Homoscedastic or heteroscedastic

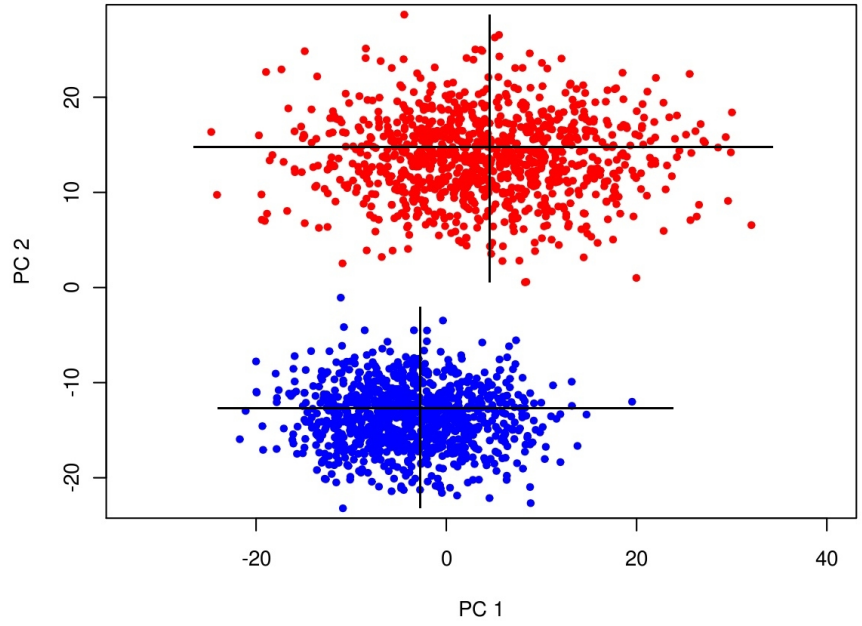
Multivariate case:

- Flury's hierarchy (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

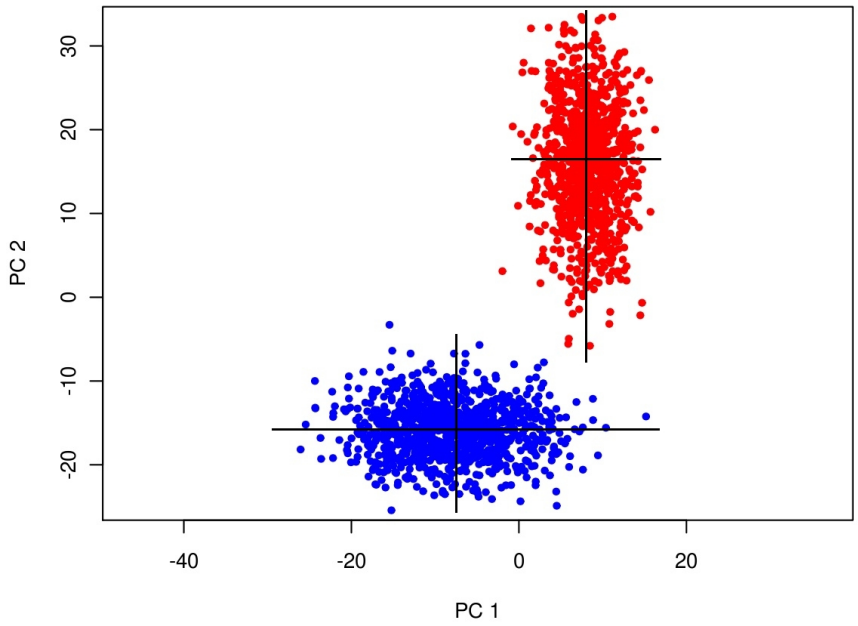
Flury's hierarchy: Equality



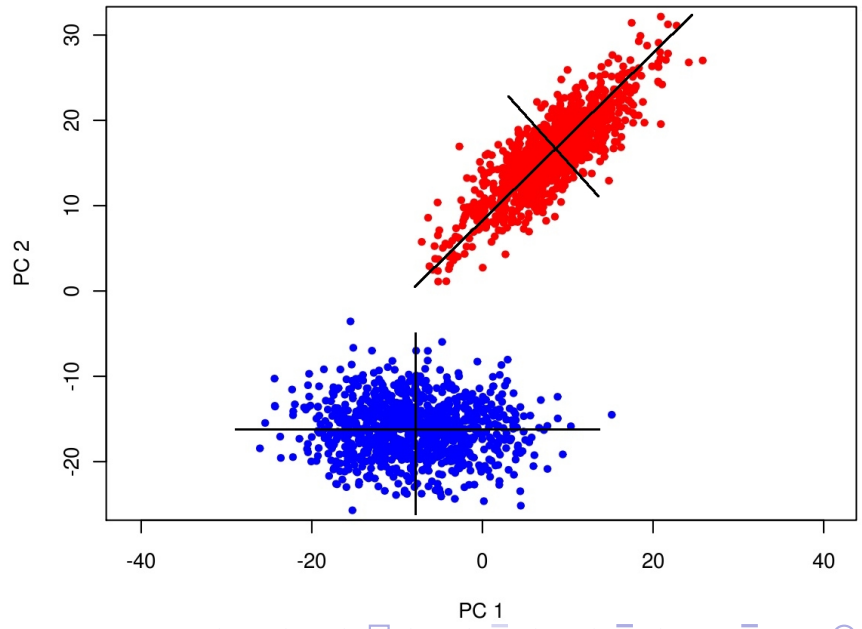
Flury's hierarchy: Proportionality



Flury's hierarchy: Common principal components (CPC)



Flury's hierarchy: Heterogeneity

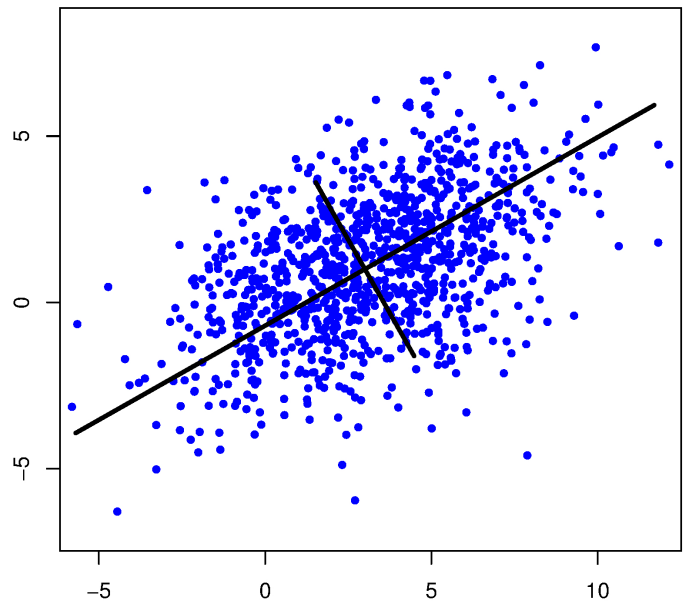


Principal component analysis (PCA)

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

Example:

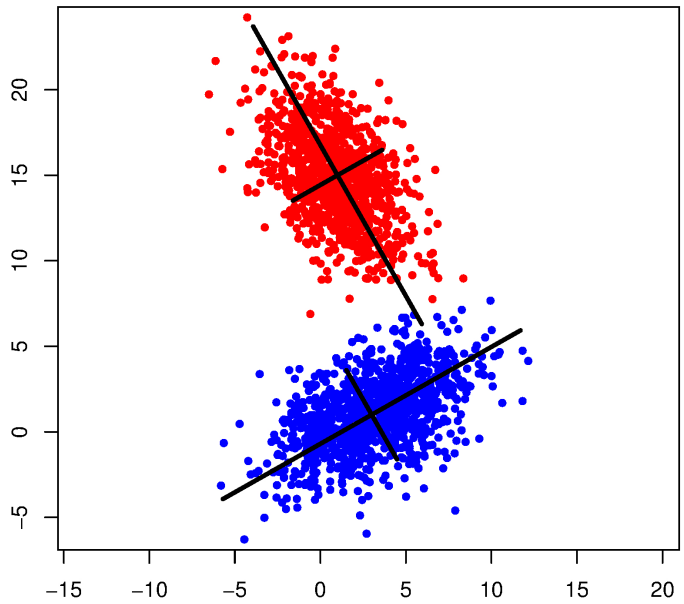
$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.87 & -0.49 \\ 0.49 & 0.87 \end{bmatrix} \begin{bmatrix} 10 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 0.87 & 0.49 \\ -0.49 & 0.87 \end{bmatrix}$$



Common principal components (CPC)

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

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



Example:

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.87 & -0.49 \\ 0.49 & 0.87 \end{bmatrix} \begin{bmatrix} 10 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 0.87 & 0.49 \\ -0.49 & 0.87 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.87 & -0.49 \\ 0.49 & 0.87 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} 0.87 & 0.49 \\ -0.49 & 0.87 \end{bmatrix}$$

Partial common principal components CPC(q)

If only $q(< p)$ of the eigenvectors are common:

$$\boldsymbol{\Sigma}_i = \boldsymbol{B}_i \boldsymbol{\Lambda}_i \boldsymbol{B}_i^T, \quad i = 1, \ldots, k$$

$$\boldsymbol{B}_i = \left[\boldsymbol{\beta}_1 \ \ldots \ \boldsymbol{\beta}_q \ \boldsymbol{\beta}_{q+1}^{(i)} \ \ldots \ \boldsymbol{\beta}_p^{(i)} \right]$$

Advantages of the (partial) CPC model

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

Simultaneous diagonalisation algorithms

Flury-Gautschi (FG)

$$\phi\left(\boldsymbol{L}_1, \ldots, \boldsymbol{L}_k; n_1, \ldots, n_k\right) = \prod_{i=1}^{k} \frac{\left[\det\left(\operatorname{diag}\boldsymbol{L}_i\right)\right]^{n_i}}{\left[\det\left(\boldsymbol{L}_i\right)\right]^{n_i}}$$

JADE package

$$\min\left(\sum_{i=1}^{k}\sum_{j=1}^{p}\sum_{\substack{h=1\\h\neq j}}^{p} l_{jhk}^2\right)$$

Stepwise CPC

- estimates eigenvectors sequentially
- ensures common eigenvectors have same rank order in all groups

Identifying common eigenvectors

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

Model		X^2	df	$\frac{X^2}{df}$	AIC for
Higher	Lower				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 a formal hypothesis test

The Vermont Oxford Network (VON) data

- Birth weight (kg)
- Apgar score at 1 min (0–10)
- Apgar score at 5 mins (0–10)
- Gestational age (weeks)
- Head circumference (cm)
- Temperature (°C)

Source: Wikipedia (https://en.wikipedia.org/wiki/Neonatal_intensive_care_unit)

Regions:

- South Africa ($n_1 = 2921$)
- Namibia ($n_2 = 120$)

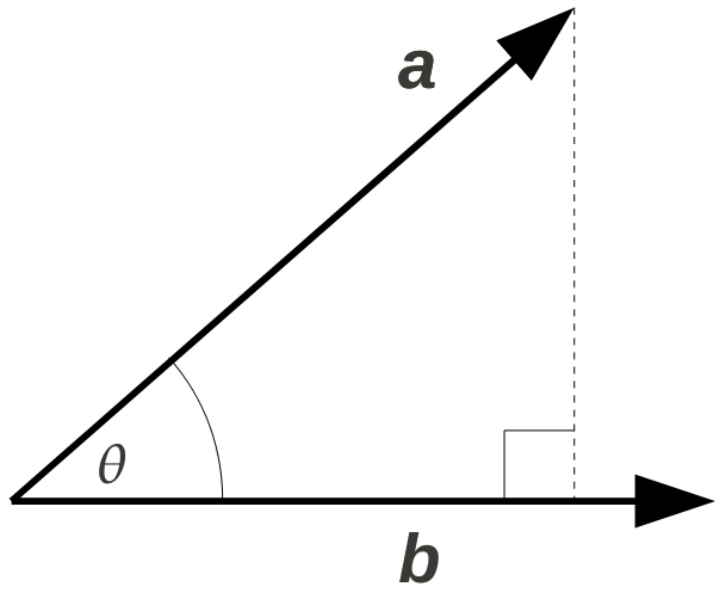
Identifying common eigenvectors

VON data: Flury's methods

Model	χ^2	**df**	$\frac{\chi^2}{\mathbf{df}}$	**AIC**
Equality	5.99	1	5.99	85.77
Proportionality	10.09	5	**2.02**	81.78
CPC	2.06	1	2.06	81.69
CPC(4)	5.27	2	2.63	81.63
CPC(3)	12.87	3	4.29	80.37
CPC(2)	34.37	4	8.59	73.50
CPC(1)	15.13	5	3.03	47.13
Heterogeneity	–	–	–	**42.00**

Identifying common eigenvectors

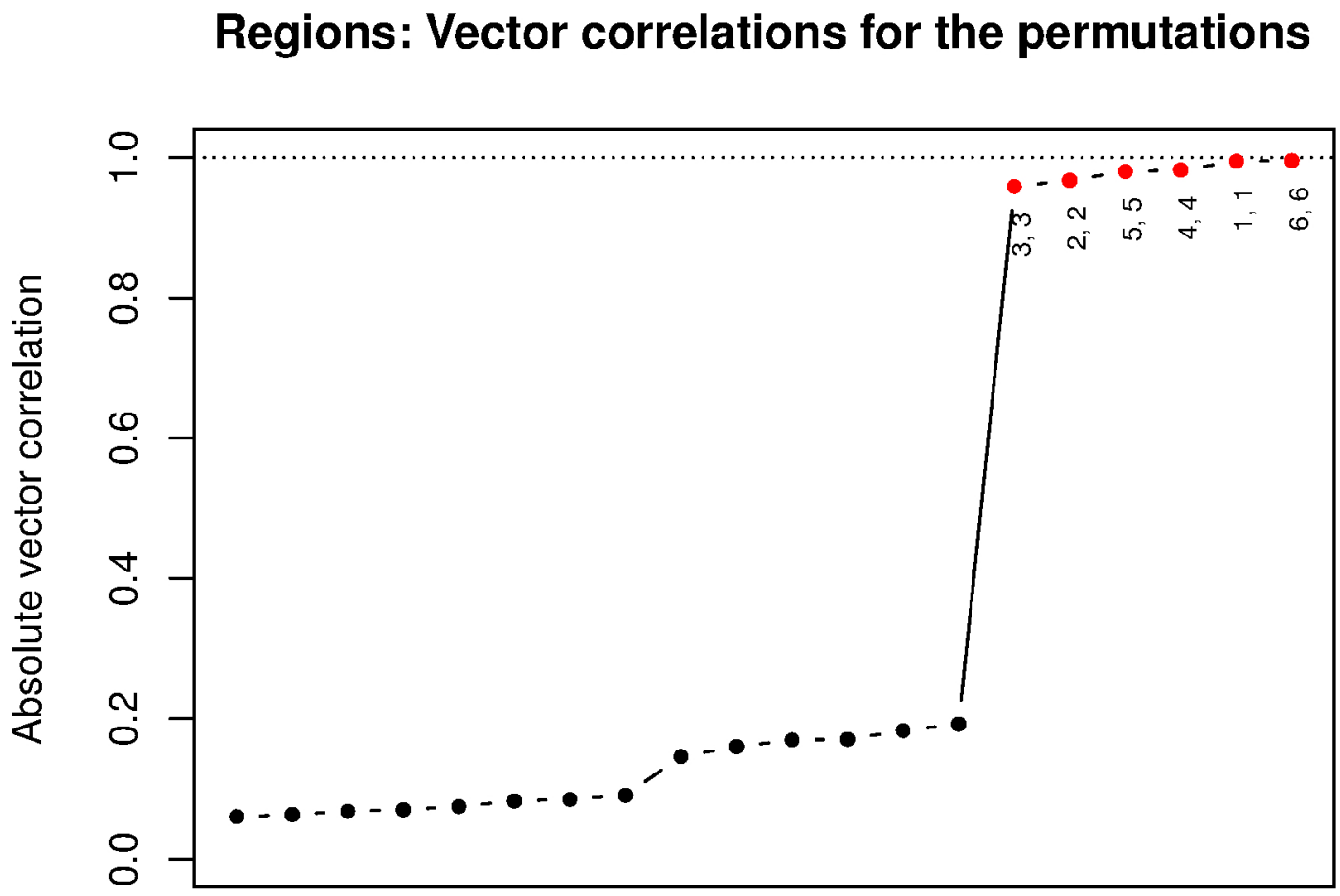
Vector correlations (Krzanowski, 1979)



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

- Inspect *vector correlations* from pairwise combinations of all p eigenvectors from the two groups.

Identifying common eigenvectors



Identifying common eigenvectors

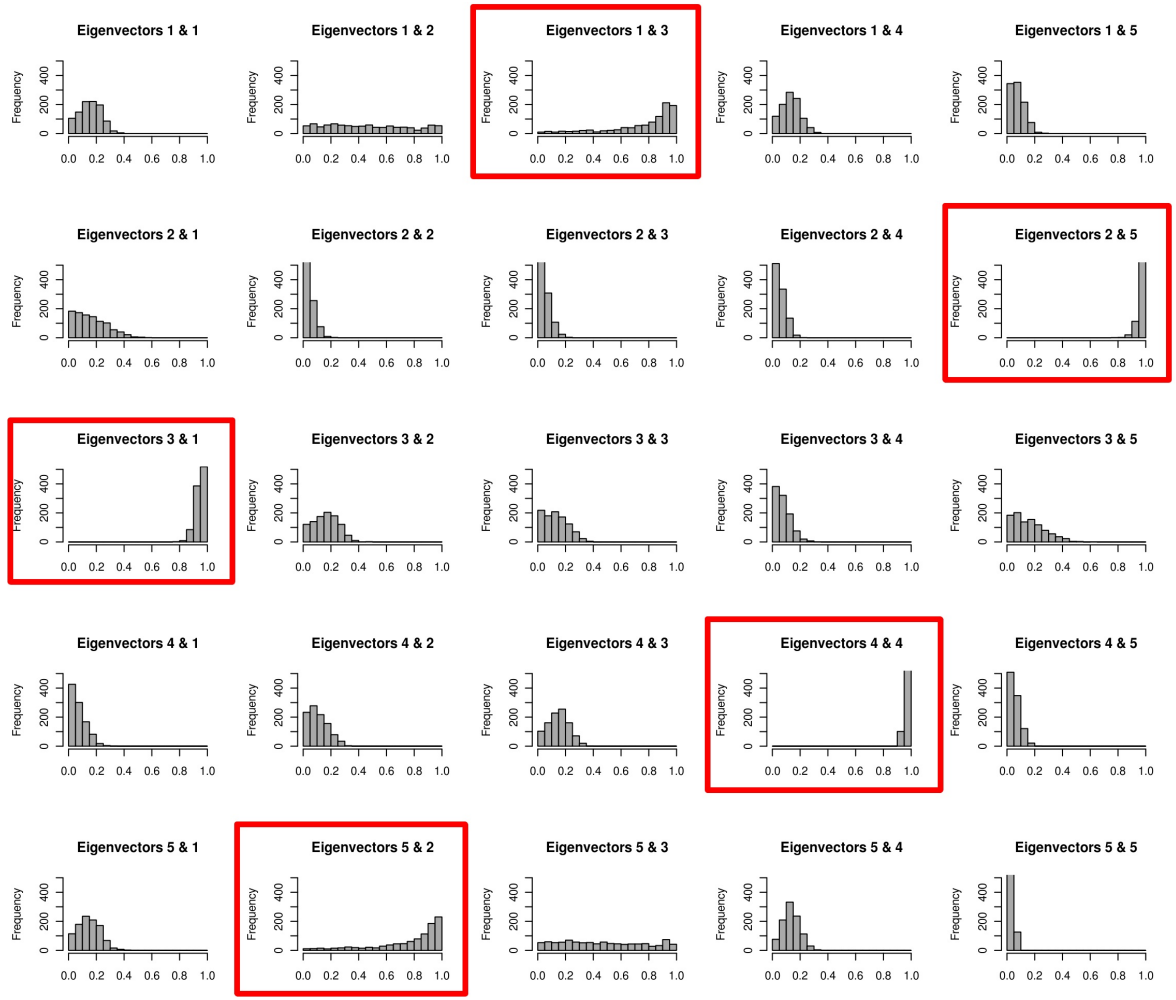
Simulated CPC(5) data:

$k = 2$ groups

$p = 5$ variables

$n_1 = n_2 = 200$

bootstrap reps $= 1000$

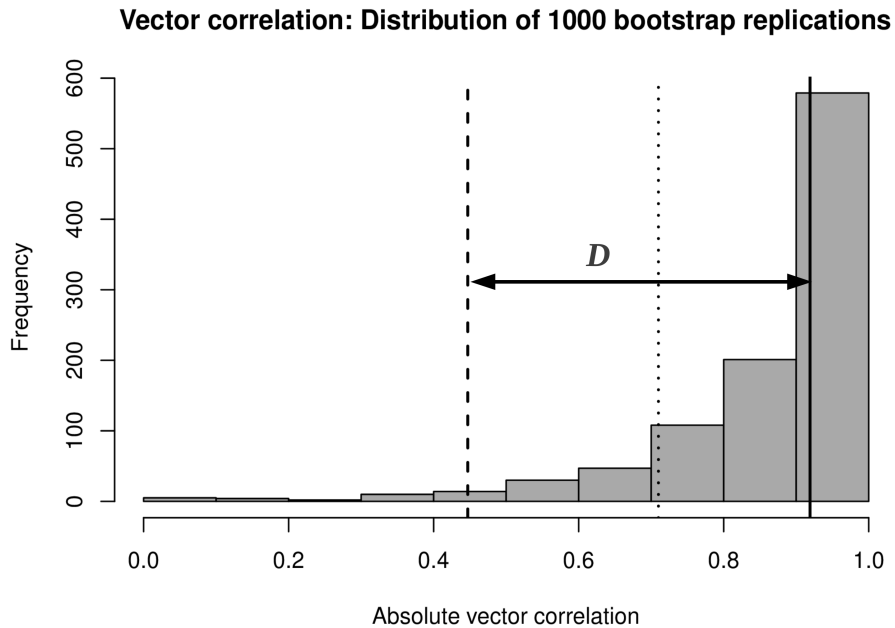


Identifying common eigenvectors

Bootstrap vector correlation distribution (BVD)

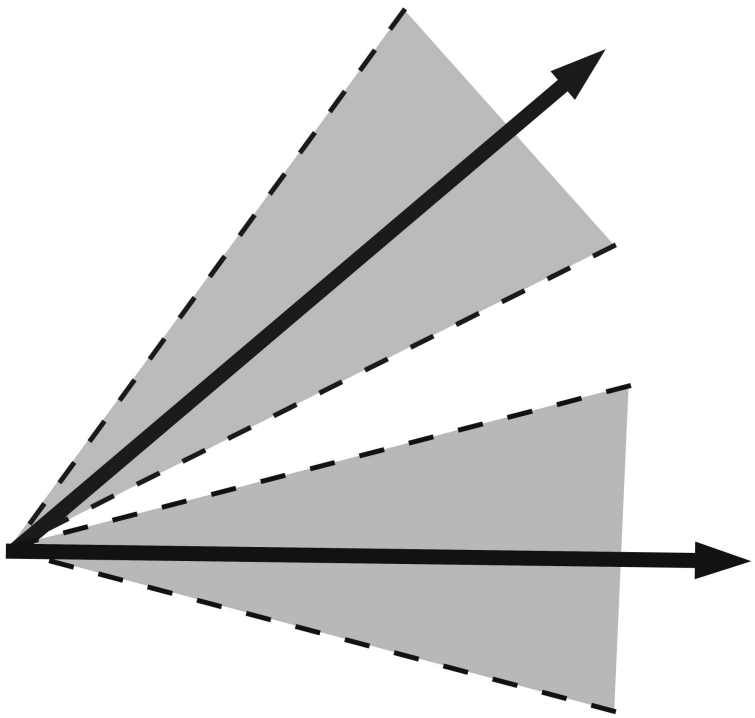
Consider two eigenvectors to be common if:

1. median > 0.71
2. median$+D \geq 1$

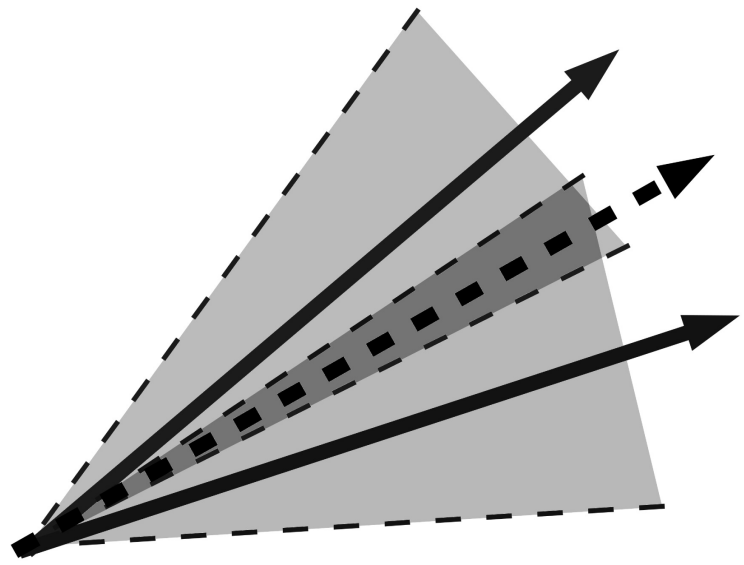


Identifying common eigenvectors

Bootstrap confidence regions (BCR)



Not common



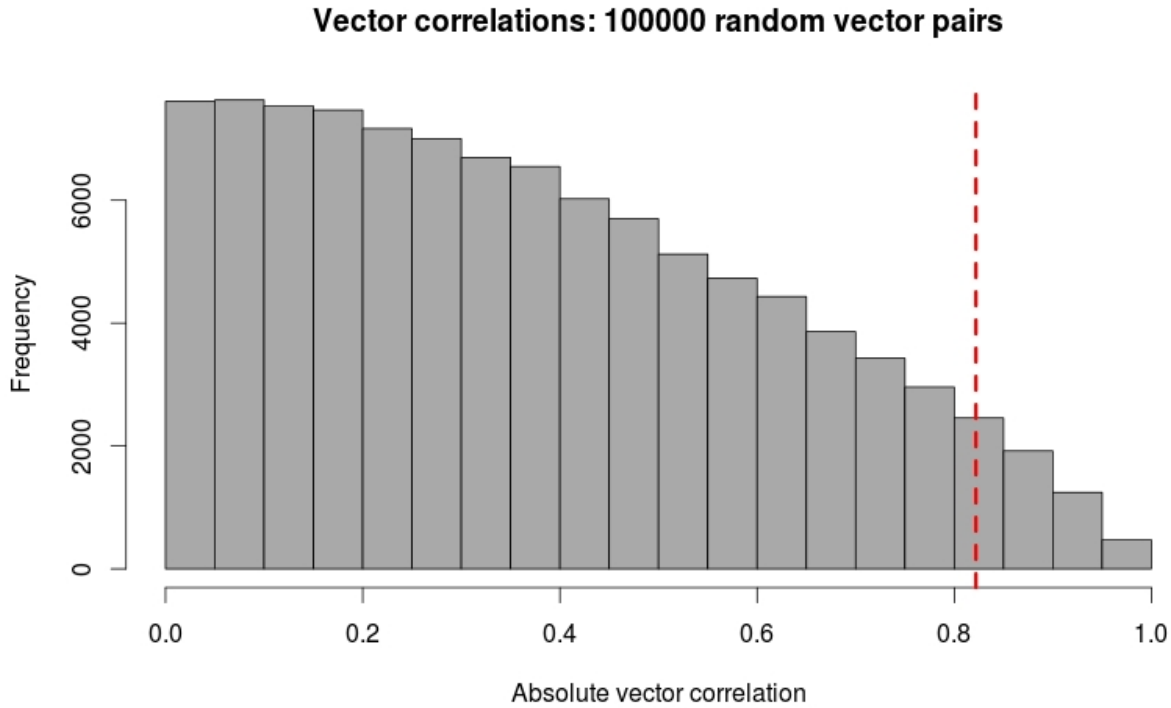
Common

Identifying common eigenvectors

Random vector correlations (RVC)

- adapted from Klingenberg and McIntyre (1998)

H_0 : pair of eigenvectors are *not* common

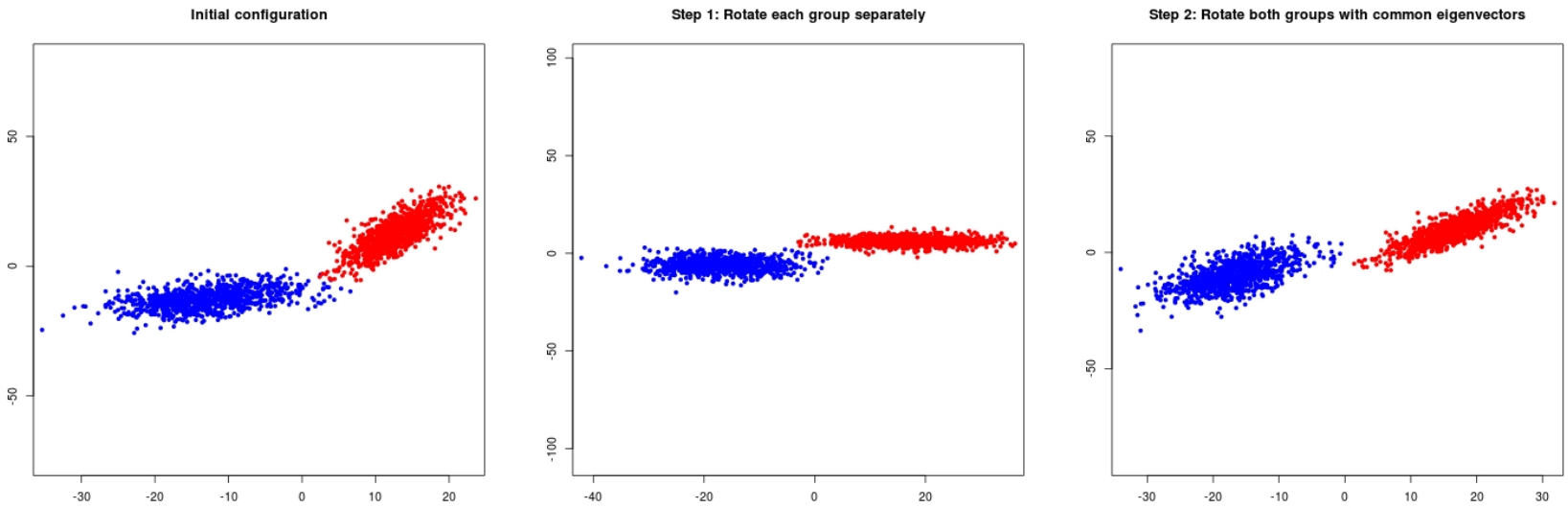


Identifying common eigenvectors

Bootstrap hypothesis test (BootTest)

- adapted from Klingenberg (1996)

H_0 : pair of eigenvectors are common



Twice rotated data for the i^{th} group:

$$\boldsymbol{X}_i^{\star} = \boldsymbol{X}_i \boldsymbol{E}_i \boldsymbol{B}', \qquad i = 1, 2.$$

Identifying common eigenvectors

Ensemble test

Eigenvector pair considered common if majority vote of

- AIC
- BVD
- BCR
- RVC
- BootTest

indicates it to be so.

Simulation results ($p = 5$ variables)

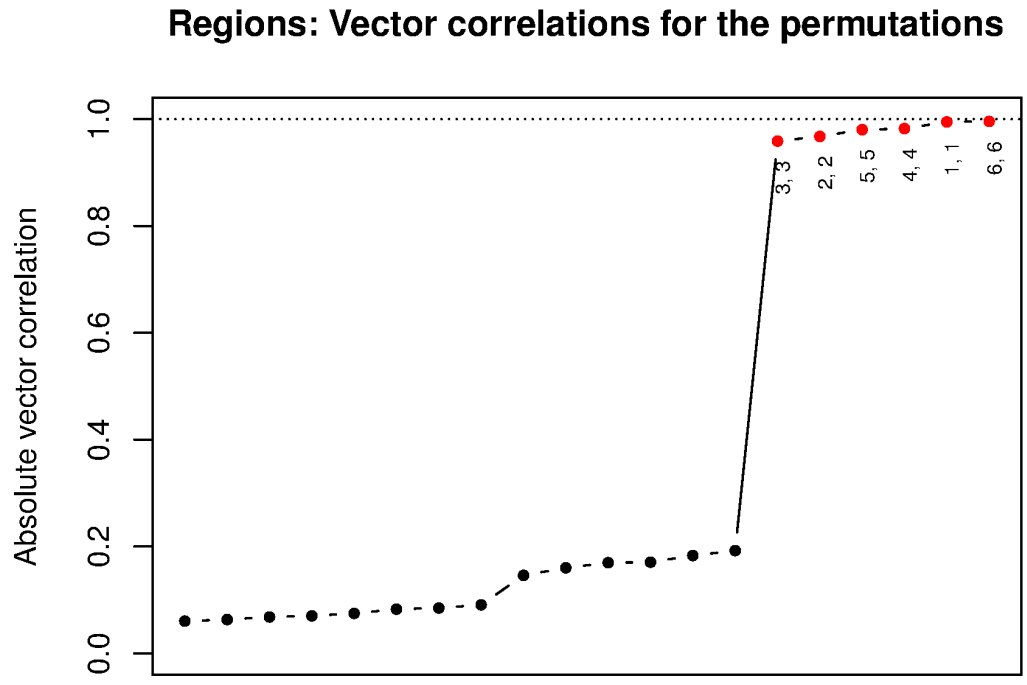
Correct number of common eigenvectors identified (% of runs)

	AIC	Chi^2	BootTest	RVC	BVD	BCR	Ensemble
Sample size							
$n = 50$	33.1	27.0	26.1	30.0	**33.9**	25.6	32.5
$n = 100$	34.2	30.7	26.4	32.2	**36.1**	29.4	35.0
$n = 200$	43.1	28.1	33.1	**47.2**	44.4	35.3	46.1
$n = 500$	43.3	34.8	46.1	53.3	**56.4**	49.4	54.2
$n = 1000$	45.8	34.1	57.2	62.5	62.8	58.1	**63.1**
Distribution							
Normal	51.5	32.4	49.3	58.2	**62.5**	51.5	59.3
Chi-squared	43.5	34.2	39.5	52.0	51.0	42.7	**52.5**
Multivariate t	24.7	26.2	24.5	25.0	**26.7**	24.5	**26.7**
Overall	39.9	31.0	37.8	45.1	**46.7**	39.6	46.2

Identifying common eigenvectors

Application to the VON data (regions)

Ensemble test: 6 common eigenvectors



Covariance matrix estimation

Covariance matrix estimators under the CPC model can be:

- less biased than when incorrectly assuming equality of the population covariance matrices, and
- more precise than when incorrectly assuming that the population covariance matrices are unrelated.

Covariance matrix estimation

CPC estimator (Flury, 1988)

- $\boldsymbol{S}_i$: unbiased sample covariance matrix estimator for i^{th} group
- $\boldsymbol{B}$: estimator of common eigenvector matrix

Estimator for $\boldsymbol{\Sigma}_i$ under the CPC model:

$$\boldsymbol{S}_{i(CPC)} = \boldsymbol{B}\boldsymbol{L}_i^0\boldsymbol{B}',$$

where

$$\boldsymbol{L}_i^0 = \text{diag}(\boldsymbol{B}'\boldsymbol{S}_i\boldsymbol{B}).$$

Regularised CPC estimator

$$\boldsymbol{S}^{\star}_{i(CPC)} = \alpha_i \boldsymbol{S}_i + (1 - \alpha_i)\boldsymbol{S}_{i(CPC)},$$

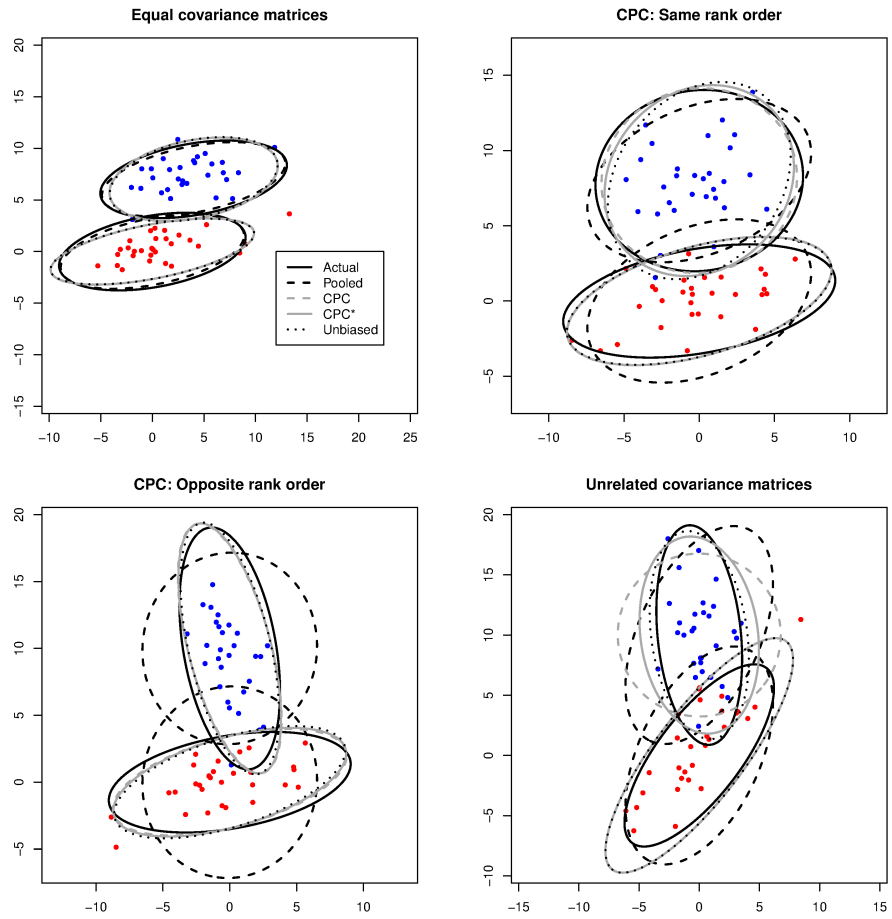
where $\alpha_i \in [0; 1]$ is the shrinkage intensity parameter.

Use cross-validation to find the value for α_i minimising a modified version of the Frobenius matrix norm on the training and validation samples.

Covariance matrix estimation

Covariance matrix shapes (95% confidence ellipses)

$k = 2$ populations, $p = 2$ variables



Simulation results

Mean standardised modified Frobenius values (smaller is better):

	Unbiased	CPC	CPC*	Pooled
Full CPC	0.269	0.372	**0.192**	0.792
Half of eigenvectors common	0.271	0.337	**0.194**	0.789
Few common eigenvectors	0.262	0.318	**0.196**	0.794
Unrelated covariance matrices	0.259	0.294	**0.195**	0.798

Covariance matrix estimation

VON data: Namibia ($n_2 = 120$)

$$\boldsymbol{S}_2 = \begin{bmatrix} 0.87 & 0.62 & 0.45 & 3.02 & 3.39 & \mathbf{0.04} \\ 0.62 & 4.48 & 2.30 & 3.31 & 2.88 & \mathbf{-0.08} \\ 0.45 & 2.30 & 2.18 & 2.75 & 2.21 & \mathbf{-0.04} \\ 3.02 & 3.31 & 2.75 & 15.37 & 13.45 & \mathbf{-0.31} \\ 3.39 & 2.88 & 2.21 & 13.45 & 15.70 & \mathbf{0.05} \\ 0.04 & -0.08 & -0.04 & -0.31 & 0.05 & 0.50 \end{bmatrix}$$

$$\boldsymbol{S}^{\star}_{2(\mathrm{CPC})} = \begin{bmatrix} 0.87 & 0.57 & 0.44 & 3.16 & 3.30 & \mathbf{0.13} \\ 0.57 & 4.15 & 2.25 & 2.92 & 2.58 & \mathbf{0.07} \\ 0.44 & 2.25 & 2.31 & 2.45 & 2.02 & \mathbf{0.09} \\ 3.16 & 2.92 & 2.45 & 15.98 & 13.46 & \mathbf{0.21} \\ 3.30 & 2.58 & 2.02 & 13.46 & 15.22 & \mathbf{0.40} \\ 0.13 & 0.07 & 0.09 & 0.21 & 0.40 & 0.58 \end{bmatrix}$$

CPC discriminant analysis

Allocate a new observation, $\boldsymbol{x}_{\text{new}}$, to the first group if

$$-\frac{1}{2}\boldsymbol{x}'_{\text{new}}(\boldsymbol{S}^{-1}_{1(\text{CPC})} - \boldsymbol{S}^{-1}_{2(\text{CPC})})\boldsymbol{x}_{\text{new}} + (\bar{\boldsymbol{x}}'_1\boldsymbol{S}^{-1}_{1(\text{CPC})} - \bar{\boldsymbol{x}}'_2\boldsymbol{S}^{-1}_{2(\text{CPC})})\boldsymbol{x}_{\text{new}} \geq c,$$

where

$$c = \frac{1}{2}\ln\left(\frac{|\boldsymbol{S}_{1(\text{CPC})}|}{|\boldsymbol{S}_{2(\text{CPC})}|}\right) + \frac{1}{2}(\bar{\boldsymbol{x}}'_1\boldsymbol{S}^{-1}_{1(\text{CPC})}\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}'_2\boldsymbol{S}^{-1}_{2(\text{CPC})}\bar{\boldsymbol{x}}_2),$$

otherwise allocate it to the second group.

Simulation results

$n_1 = n_2$, $k = 2$ multivariate normal populations, $p = 10$ variables

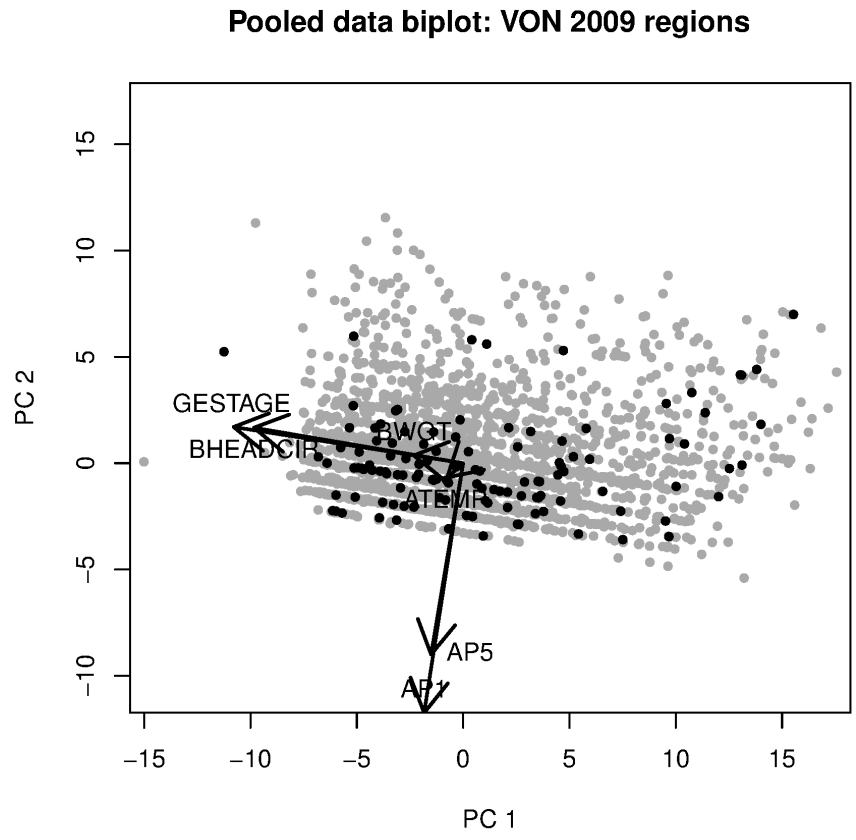
		Misclassification error (%)			
Structure	n_i	QDA	CPC	CPC*	LDA
$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$	30	42.06	33.88	34.48	**32.72**
	100	34.01	29.25	29.53	**28.44**
	200	31.27	28.25	28.35	**27.70**
CPC	30	28.58	**18.12**	18.77	33.52
(similar	100	18.12	**14.93**	15.08	28.80
rank orders)	200	15.89	**14.13**	14.26	27.49
CPC	30	5.20	**2.28**	2.46	24.73
(Opposite	100	2.41	**1.95**	1.97	18.31
rank orders)	200	1.99	**1.84**	1.85	16.56
Unrelated	30	13.78	8.94	**8.47**	34.93
covariance	100	5.85	7.15	**5.57**	30.80
matrices	200	**4.89**	6.95	4.92	29.76

CPC discriminant analysis

VON data: Regions (6 common eigenvectors)

Misclassification errors:

- QDA $= 25.2\%$
- LDA $= 25.4\%$
- CPC $= 21.2\%$
- CPC* $= 22.9\%$

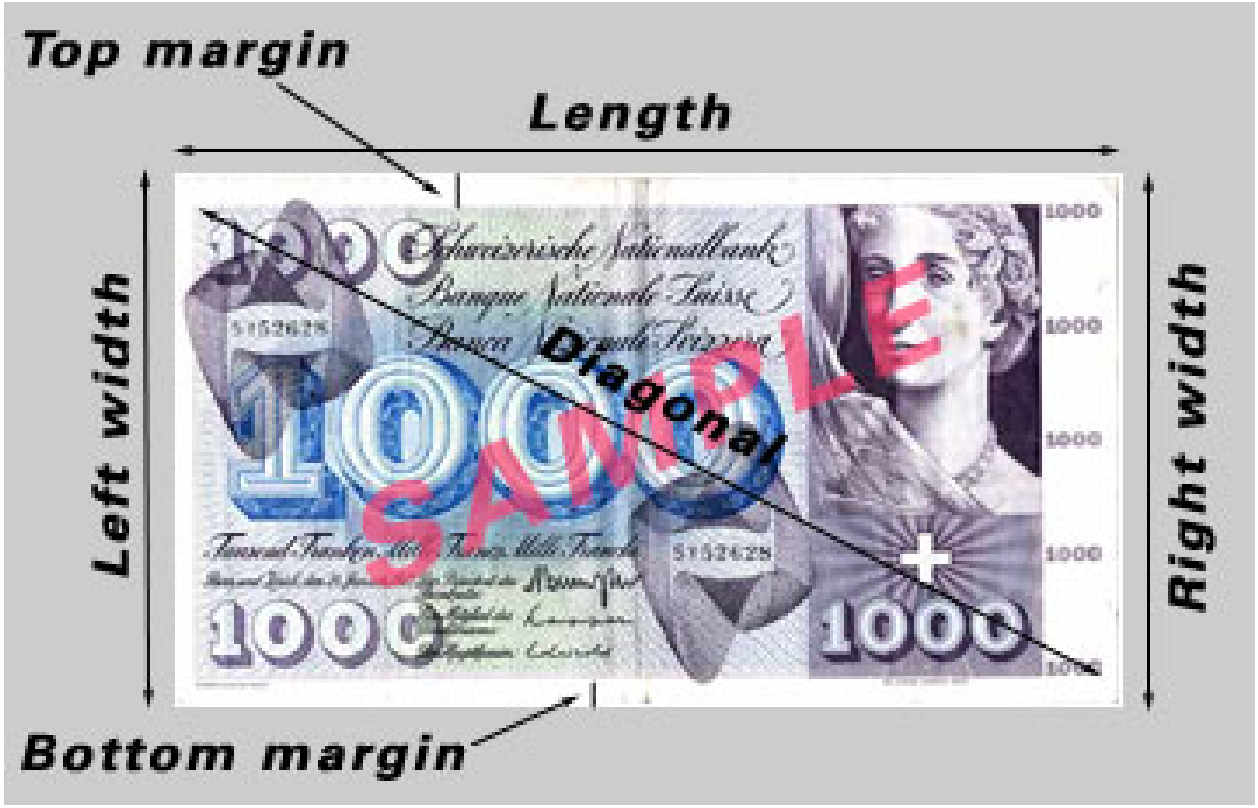


Biplots for grouped data

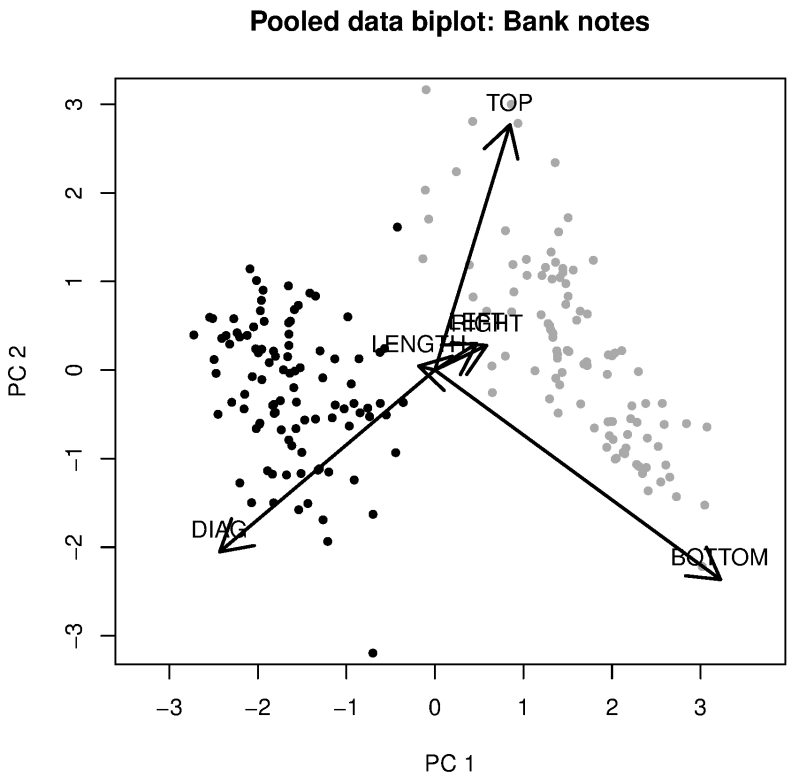
- overall quality of display
- *between-group* variation
- *within-group* variation
- representation of *variables*
  - adequacy
  - mean standard predictive error (MSPE), (Rui Alves, 2012)
- representation of *observations*
  - sample predictivities (Gower et al., 2011)

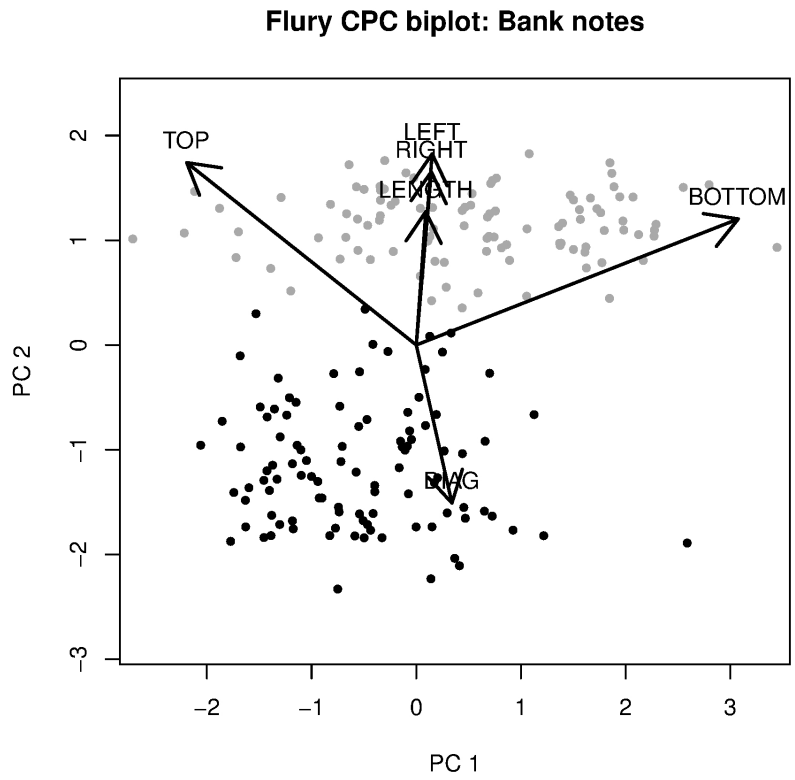
Swiss bank notes (Flury, 1988)

Genuine notes ($n_1 = 100$), Forged notes ($n_2 = 100$)



CPC biplots


Pooled data biplot: Bank notes
TOP
LENGTH
DIAG
BOTTOM
PC 1
PC 2


Flury CPC biplot: Bank notes
TOP
LEFT
RIGHT
BOTTOM
PC 1
PC 2

Quality measures for 2D biplot of Bank Notes data

	Overall	Within	Between	MSPE	Sample predictivities
Pooled S	0.42	**0.72**	0.21	0.80	0.35
Pooled data	**0.88**	0.70	**1.00**	**0.44**	**0.85**
CPC (Flury)	0.65	0.70	0.61	0.75	0.62
CPC (Stepwise CPC)	0.35	0.71	0.10	0.75	0.31
CPC (JADE)	0.44	0.71	0.26	0.79	0.38

Time series

Bovine tuberculosis (bTB), caused by the *Mycobacterium bovis* pathogen, is an important disease in cattle.

VetNet/SAM database (APHA):

- January 2003 – December 2012
  - Test/Analysis: 2003–2010
  - Forecasting: 2011–2012
- Monthly confirmed herd breakdowns (OTFW)
  - England, Wales and Scotland

Time series

1. Can we detect regular signals in the time series?

2. Are the signals from England, Wales and Scotland similar?

3. Can any similarities improve our knowledge of bTB breakdowns in Scotland?

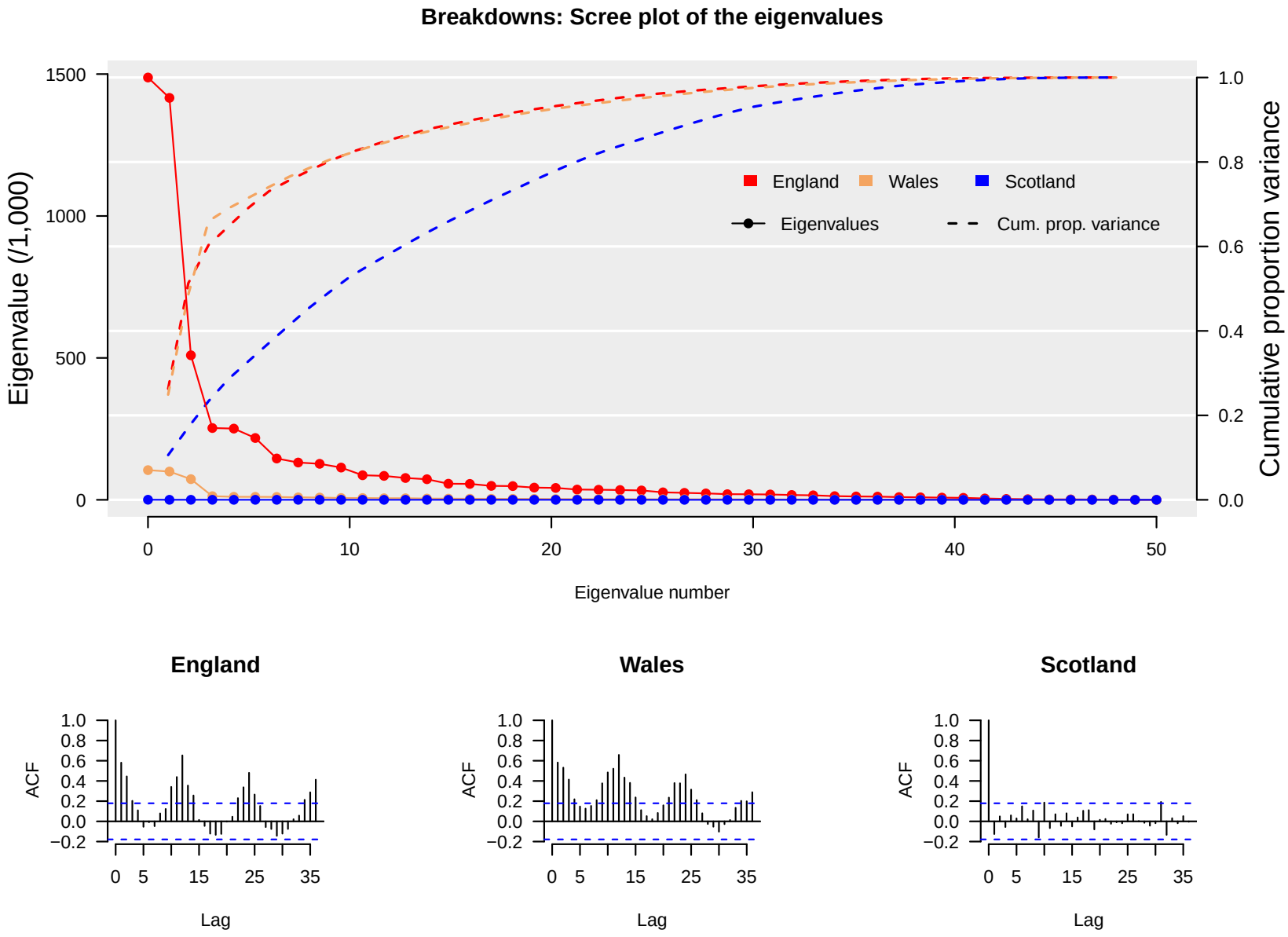
Singular Spectrum Analysis (SSA)

Trajectory matrix (Hankel structure)

$$\boldsymbol{X} = \begin{bmatrix} x_1 & x_2 & \dots & x_{n-L+1} \\ x_2 & x_3 & \dots & x_{n-L+2} \\ \vdots & \vdots & & \vdots \\ x_L & x_{L+1} & \dots & x_n \end{bmatrix}$$

SSA involves spectral decomposition of $\boldsymbol{X}\boldsymbol{X}^T$.

Time series: SSA results



Common Singular Spectrum Analysis (CSSA)

$$\boldsymbol{X}_i \boldsymbol{X}_i^T = \boldsymbol{B} \boldsymbol{\Lambda}_i \boldsymbol{B}^T, \quad i = 1, \ldots, k$$

The k centred trajectory matrix cross products are simultaneously diagonalised by the same eigenvector matrix, $\boldsymbol{B}$.

Simultaneous diagonalisation: Stepwise CPC

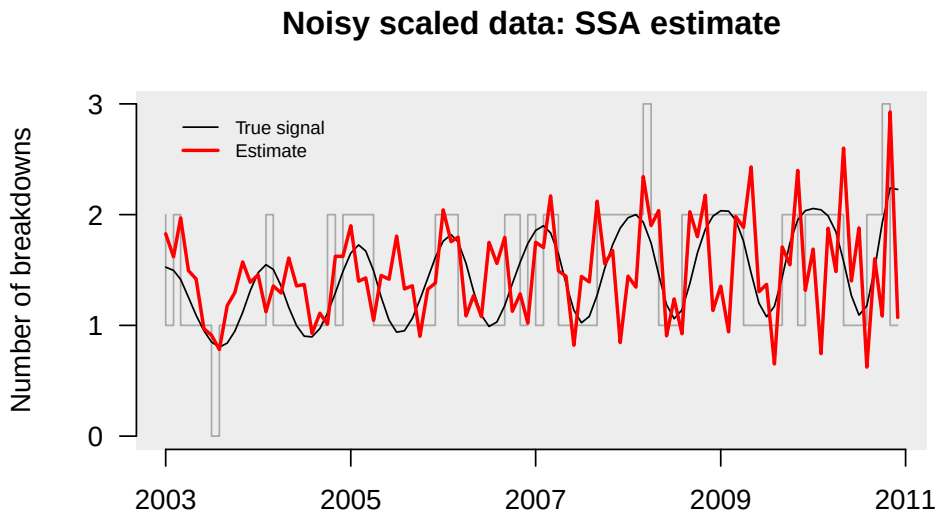
Time series: CSSA results

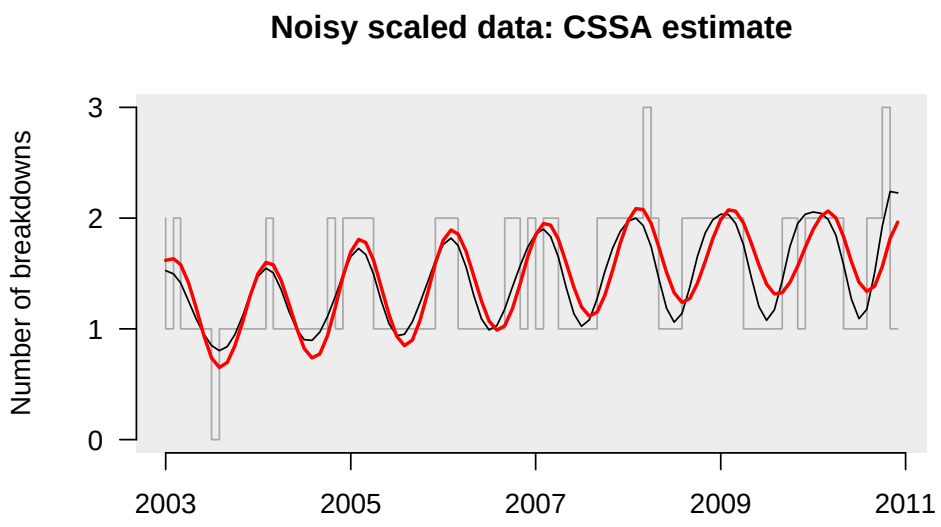
- England and Wales:

Eigenvectors		Absolute vector
England	Wales	correlation
1	2	**0.97**
2	1	**0.92**
3	3	**0.91**
18	28	0.73
9	33	0.72
⋮	⋮	⋮

- Scotland and England/Wales: No common eigenvectors

Simulation experiment: Weak noisy signal



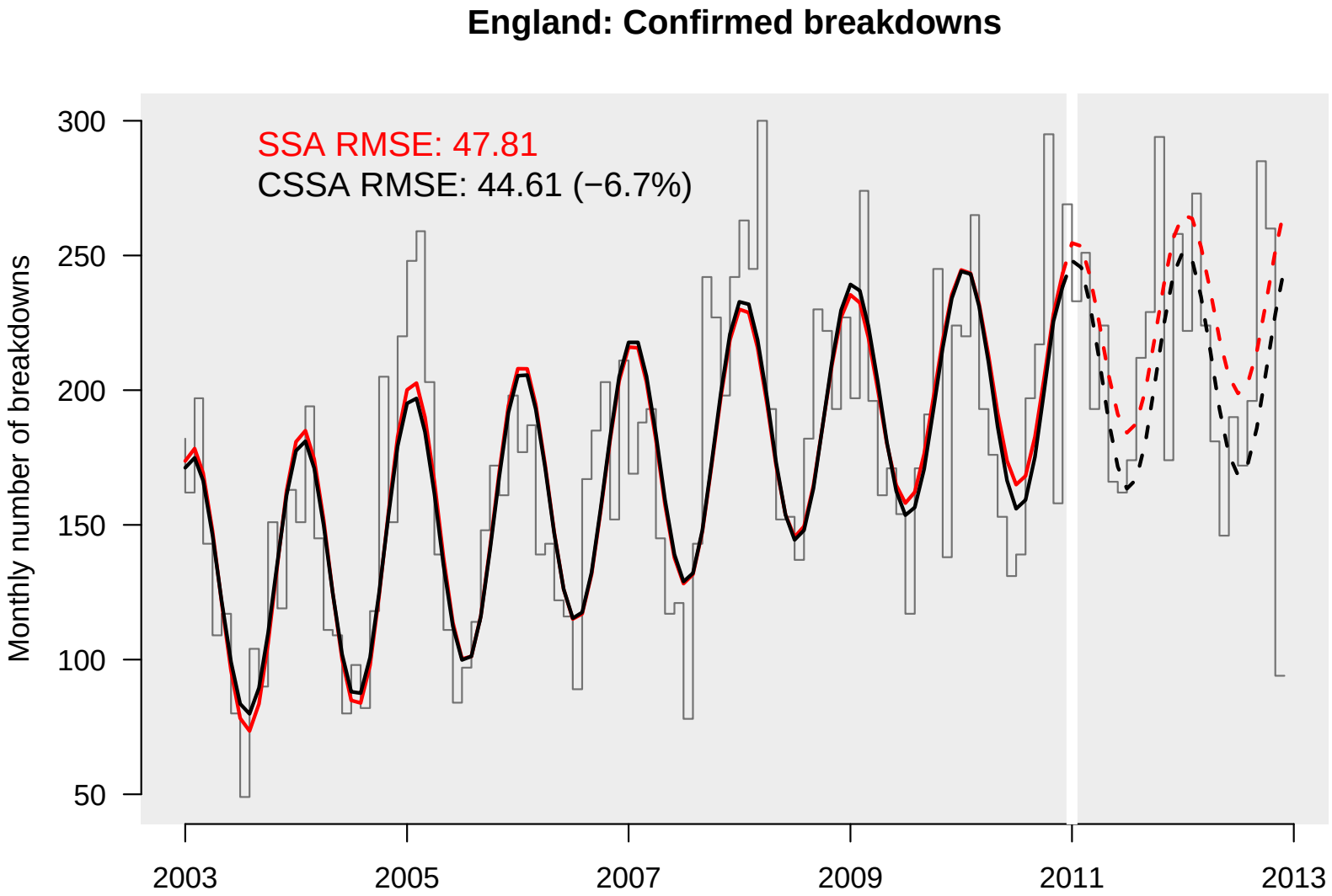


Time series: Forecasting

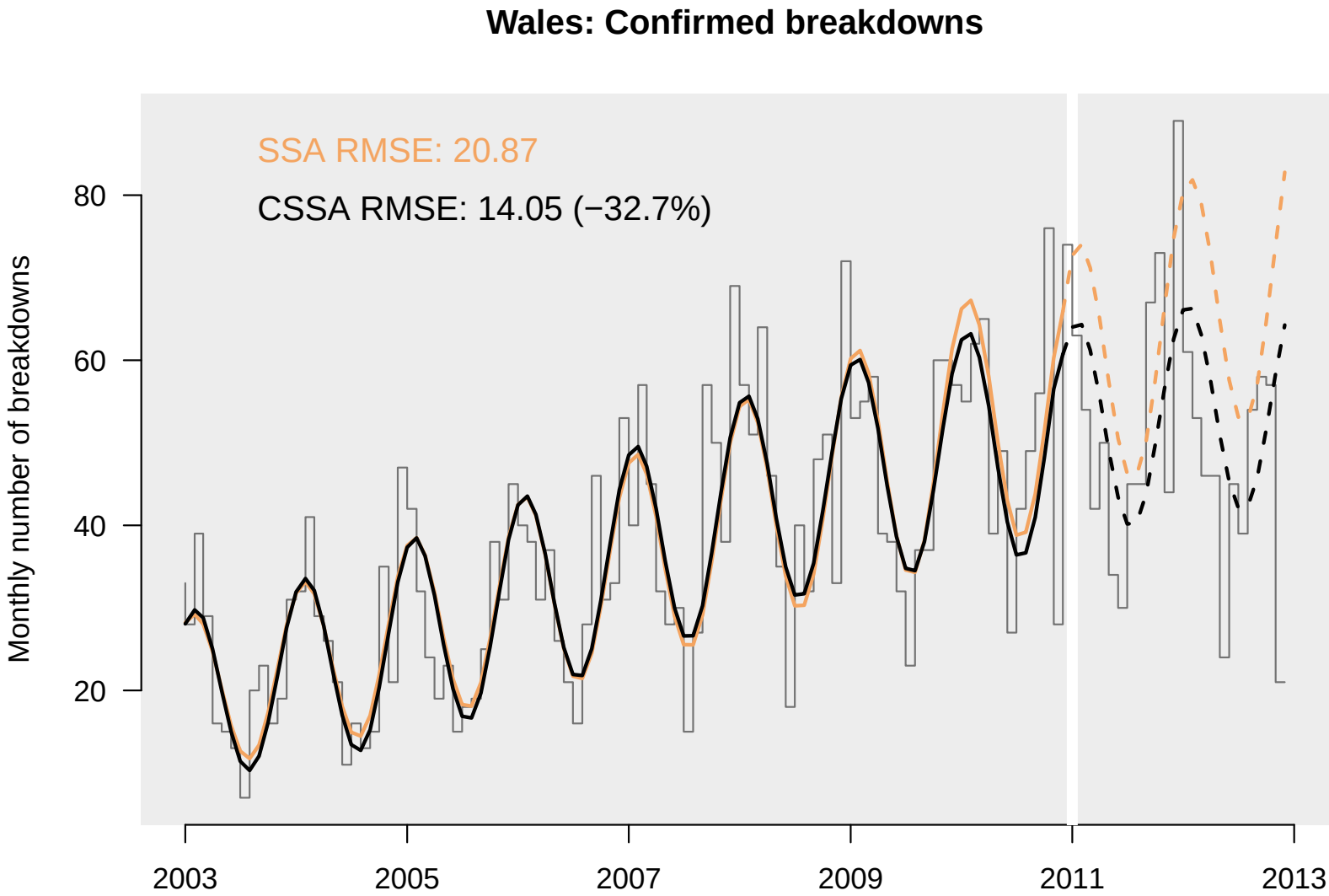
Recurrent forecasting for the period 2011–2012, using either

- first three eigenvectors (SSA)
- estimated common signal in England and Wales (CSSA)

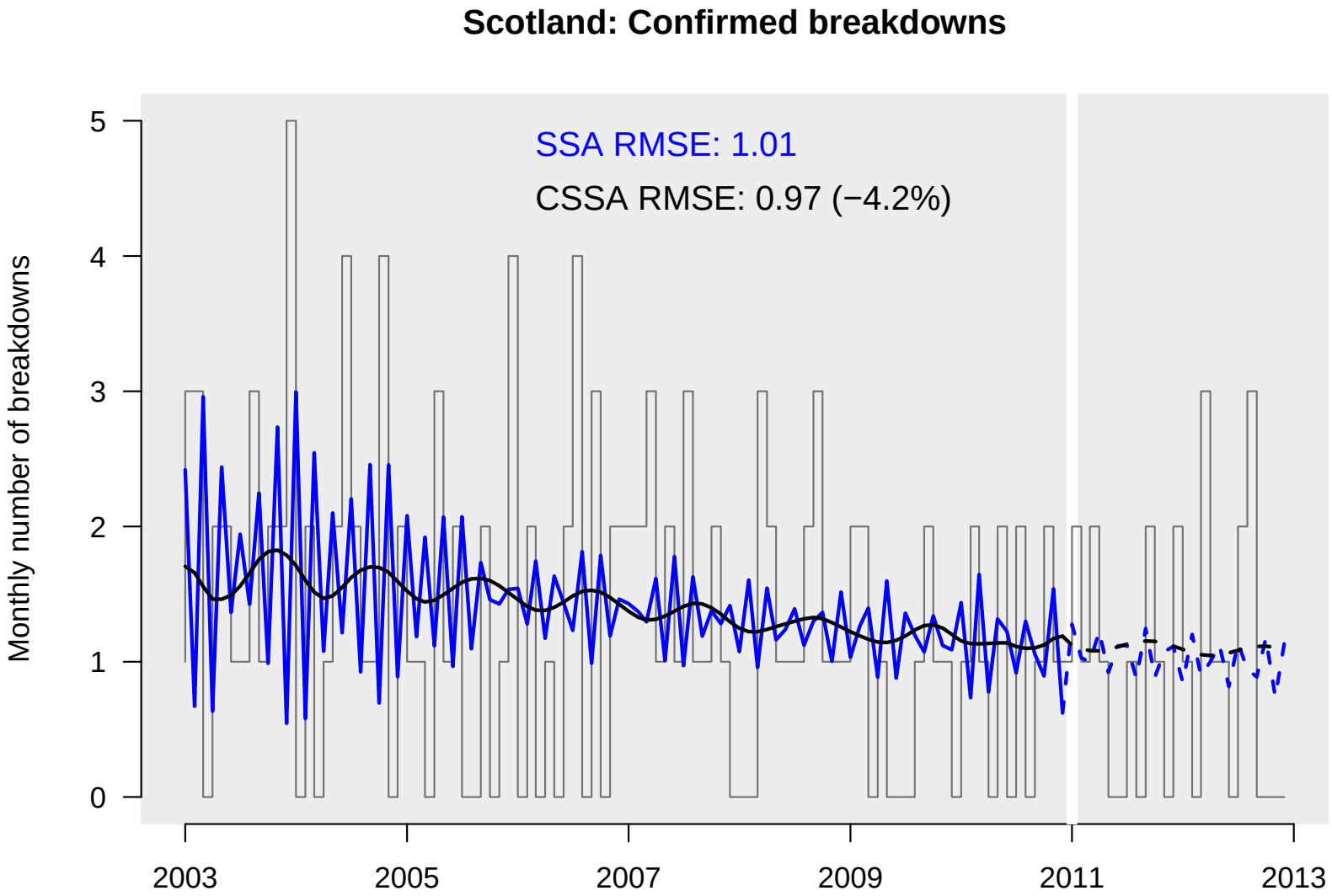
Forecasts: SSA vs. CSSA



Forecasts: SSA vs. CSSA



Forecasts: SSA vs. CSSA



Time series analysis: Conclusions

- First three singular spectra in the trajectory matrices of England and Wales constitute common signal
- Estimated signal appears to be seasonal (annual)
- Absence of any strong signal in Scottish data
- Forecasting improved using estimated common signal, in all three regions

References (page 1 of 3)

- Cardoso, J.-F. and Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164.
- Diaconis, P. and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 248(5):116–130.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.
- Flury, B. (1988). *Common Principal Components and Related Multivariate Models*. Wiley.
- Flury, B.N. and Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184.
- Friedman, J.H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.
- Golyandina, N. and Zhigljavsky, A. (2013). *Singular Spectrum Analysis for Time Series*. Springer.
- Gower, J.C., Gardner-Lubbe, S. and Roux, N.L. (2011). *Understanding Biplots*. Wiley.
- Hastie, T.J., Tibshirani, R.J. and Friedman, J.J.H. (2009). *The elements of statistical learning*. Springer-Verlag.
- Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall.

References (page 2 of 3)

- Klingenberg, C.P. (1996). Multivariate allometry. *NATO ASI SERIES A LIFE SCIENCES*, 284:23–50.
- Klingenberg, C.P. and McIntyre, G.S. (1998). Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with Procrustes methods. *Evolution*, 52(5):1363–1375.
- Krzanowski, W.J. (1979). Between-groups comparison of principal components. *Journal of the American Statistical Association*, 74(367):703–707.
- Pepler, P.T., Uys, D.W. and Nel, D.G. (2016). A comparison of some methods for the selection of a common eigenvector model for the covariance matrices of two groups. *Communications in Statistics - Simulation and Computation*, **45**(8): 2917–2936.
- Pepler, P.T., Uys, D.W. and Nel, D.G. (Accepted). Discriminant analysis under the common principal components model. *Communications in Statistics - Simulation and Computation*, In press.
- Pepler, P.T., Uys, D.W. and Nel, D.G. (Accepted). Regularised covariance matrix estimation under the common principal components model. *Communications in Statistics - Simulation and Computation*, In press.
- Rui Alves, M. (2012). Evaluation of the predictive power of biplot axes to automate the construction and layout of biplots based on the accuracy of direct readings from common outputs of multivariate analyses: 1. Application to principal component analysis. *Journal of Chemometrics*, 26(5):180–190.

References (page 3 of 3)

- Trendafilov, N.T. (2010). Stepwise estimation of common principal components. *Computational Statistics & Data Analysis*, **54**(12): 3446–3457.
- Viljoen, H. and Nel, D.G. (2010). Common singular spectrum analysis of several time series. *Journal of Statistical Planning and Inference*, **140**: 260–267.
- Viljoen, H. and Steel, S.J. (2013). Identifying secondary series for stepwise common singular spectrum analysis. *ORiON*, **29**(2): 155–167.

Thank you!

- Danie Uys, Daan Nel, Helena Viljoen (*Stellenbosch University*)
- Scottish Government Rural and Environment Science and Analytical Services Division (RESAS)
- Centre of Expertise on Animal Disease Outbreaks (EPIC)

University *of* Glasgow

More information: theo.pepler@glasgow.ac.uk