The regularised CPC covariance matrix estimator and its application in discriminant analysis

Pieter Theo Pepler

Unit for Biometry, Genetics Department, Stellenbosch University, South Africa

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Purpose

To investigate whether the use of the regularised common principal component (CPC) estimators (for the covariance matrices of two groups) in the quadratic discriminant function can improve the misclassification error rate, compared to ordinary quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA).

Common principal components (CPC) model (Flury, 1988)

Spectral decomposition of two covariance matrices ($\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$):

$$\boldsymbol{\Sigma}_1 = \beta\boldsymbol{\Lambda}_1\beta' \qquad \boldsymbol{\Sigma}_2 = \beta\boldsymbol{\Lambda}_2\beta'$$

The covariance matrices have the same eigenvectors (columns of β), but different eigenvalues (diagonal elements of $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$). The common eigenvector matrix can be estimated with the Flury-Gautschi (or another) algorithm.

Regularised CPC estimator

Let

- $\mathbf{S}_i$: $p \times p$ unbiased sample covariance matrix estimator for i^{th} group, $i = 1, 2$
- $\mathbf{B}$: estimator of modal matrix, β

Estimator of $\boldsymbol{\Sigma}_i$ under the CPC model (Flury, 1988):

$$\mathbf{S}_{i(CPC)} = \mathbf{B}\mathbf{L}_i^0\mathbf{B}', \tag{1}$$

where

$$\mathbf{L}_i^0 = \mathrm{diag}(\mathbf{B}'\mathbf{S}_i\mathbf{B}). \tag{2}$$

Regularised CPC estimator of $\boldsymbol{\Sigma}_i$:

$$\mathbf{S}^{\star}_{i(CPC)} = \alpha_i\mathbf{S}_i + (1-\alpha_i)\mathbf{S}_{i(CPC)}, \tag{3}$$

where $\alpha_i \in [0; 1]$ is the shrinkage intensity parameter. An appropriate value for α_i is estimated using cross-validation, by dividing the original sample for the i^{th} group r times randomly into a 70% training set and a 30% validation set, and performing the following procedure:

For $r = 1, \ldots, 100$ replications

70%	Estimate $\mathbf{S}^{(r)}_{i(TRAIN)}$ (unbiased estimator) and $\mathbf{S}^{(r)}_{i(CPC)}$ (CPC estimator)
30%	Estimate $\mathbf{S}^{(r)}_{i(VALID)}$ (unbiased estimator)

Find $\alpha_i^{(r)}$ which minimises

$$\| \left[\alpha_i^{(r)}\mathbf{S}^{(r)}_{i(TRAIN)} + (1-\alpha_i^{(r)})\mathbf{S}^{(r)}_{i(CPC)}\right] - \mathbf{S}^{(r)}_{i(VALID)} \|_{F^{\star}}, \tag{4}$$

where

$$\| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \|_{F^{\star}} = \sqrt{\sum_{j=1}^{p}\sum_{h\leq j}^{p}(\hat{\sigma}_{jh} - \sigma_{jh})^2}, \qquad j, h = 1, \ldots, p. \tag{5}$$

is a modified version of the Frobenius matrix norm.

Estimator of the shrinkage intensity parameter for the i^{th} group:

$$\hat{\alpha}_i = \frac{\sum_r \alpha_i^{(r)}}{r}. \tag{6}$$

CPC discriminant analysis ($k = 2$ populations)

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

$$-\frac{1}{2}\mathbf{x}'_{\text{new}}(\mathbf{S}^{-1}_{1(\text{CPC})} - \mathbf{S}^{-1}_{2(\text{CPC})})\mathbf{x}_{\text{new}} + (\bar{\mathbf{x}}_1'\mathbf{S}^{-1}_{1(\text{CPC})} - \bar{\mathbf{x}}_2'\mathbf{S}^{-1}_{2(\text{CPC})})\mathbf{x}_{\text{new}} \geq c, \tag{7}$$

where

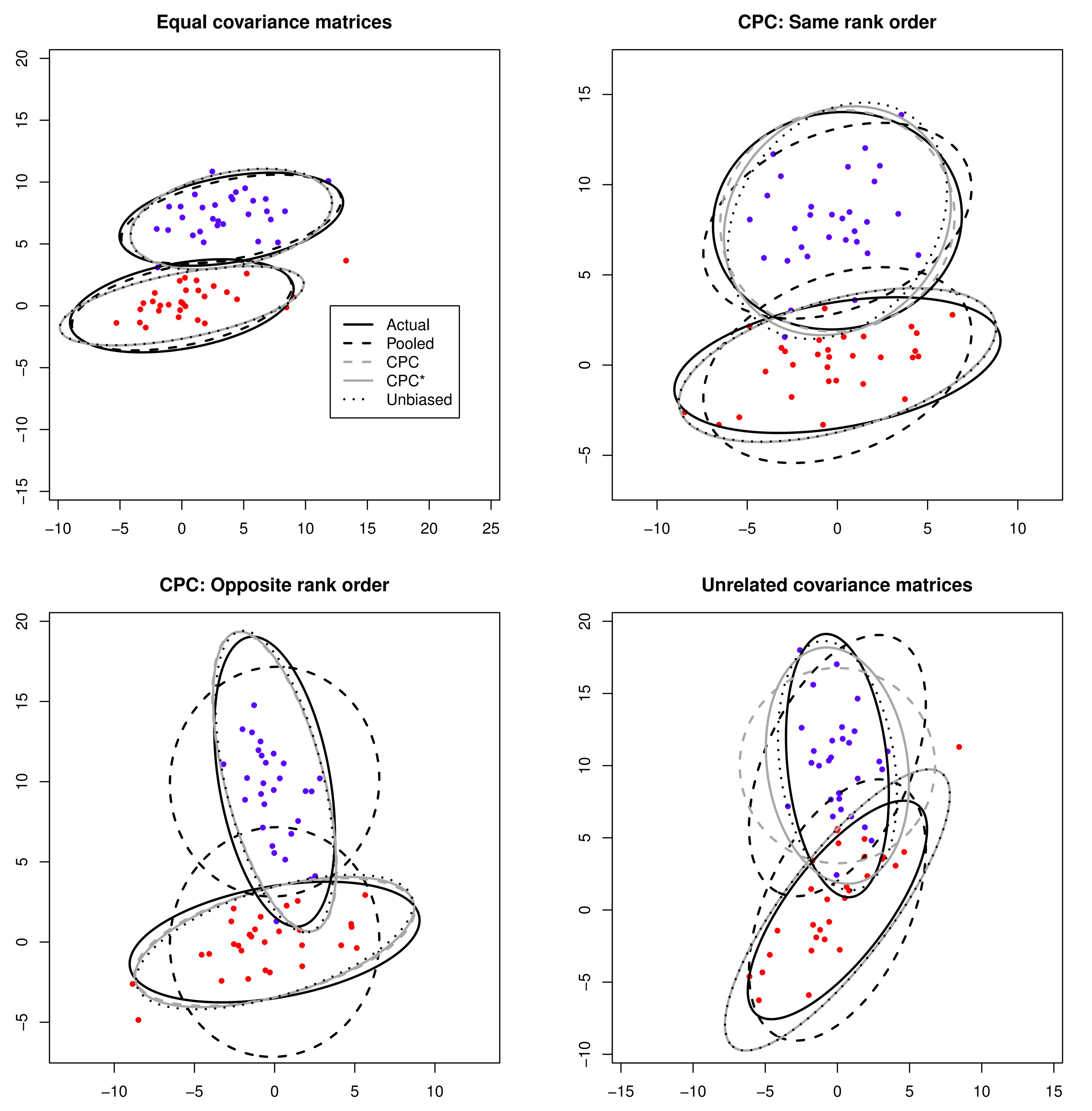
$$c = \frac{1}{2}\ln\left(\frac{|\mathbf{S}_{1(\text{CPC})}|}{|\mathbf{S}_{2(\text{CPC})}|}\right) + \frac{1}{2}(\bar{\mathbf{x}}_1'\mathbf{S}^{-1}_{1(\text{CPC})}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2'\mathbf{S}^{-1}_{2(\text{CPC})}\bar{\mathbf{x}}_2), \tag{8}$$

otherwise allocate it to the second group. $\mathbf{S}_{i(\text{CPC})}$ is the CPC estimator for the covariance matrix of the i^{th} group as defined in (1). To perform *regularised* CPC discriminant analysis, replace $\mathbf{S}_{i(\text{CPC})}$ in (7) and (8) with the regularised estimator, $\mathbf{S}^{\star}_{i(CPC)}$, defined in (3).

Ordinary CPC discrimination and regularised CPC discrimination are referred to as *CPC* and *CPC**, respectively, in the presentation of the simulation results (see the box labelled "Simulation study").

Covariance matrix shapes (95% confidence ellipses)

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



Simulation study

Simulation results for samples of equal sizes drawn from $k = 2$ multivariate normally distributed populations with $p = 10$ variables. Each of the values in the table were calculated from 1000 simulation runs.

		Misclassification error (%)			
Structure	n_i	QDA	CPC	CPC*	LDA
Equal population	50	37.79	31.65	31.67	**30.68**
covariance matrices	100	34.01	29.25	29.53	**28.44**
	200	31.27	28.25	28.35	**27.70**
Common eigenvectors with	50	22.96	**16.50**	16.81	30.43
Similar rank orders in the	100	18.12	**14.93**	15.08	28.80
two covariance matrices	200	15.89	**14.13**	14.26	27.49
Common eigenvectors with	50	3.31	**2.15**	2.22	21.55
Opposite rank orders in the	100	2.41	**1.95**	1.97	18.31
two covariance matrices	200	1.99	**1.84**	1.85	16.56
Unrelated population	50	8.66	8.14	**6.94**	32.94
covariance matrices	100	5.85	7.15	**5.57**	30.80
	200	**4.89**	6.95	4.92	29.76

Conclusion

Both ordinary and regularised CPC discrimination outperform QDA and LDA when there are common eigenvectors in two population covariance matrices. The improvement in misclassification error rate is most pronounced when the common eigenvectors have opposite rank orders in the covariance matrices.

References

Flury, B. (1988). Common principal components and related multivariate models. *Wiley series in probability and mathematical statistics*. Wiley.

Flury, B., 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.

Hastie, T.J., Tibshirani, R.J. and Friedman, J.J.H. (2009). The elements of statistical learning. *Springer series in statistics*. Springer-Verlag.

Johnson, R.A. and Wichern, D.W. (2002). Applied multivariate statistical analysis. Prentice Hall.

Correspondence to be directed to tpepler@sun.ac.za