On the application of the CPC model in discriminant analysis

Theo Pepler
Unit for Biometry
Stellenbosch University

29 October 2014

Quadratic discriminant analysis

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

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

where

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

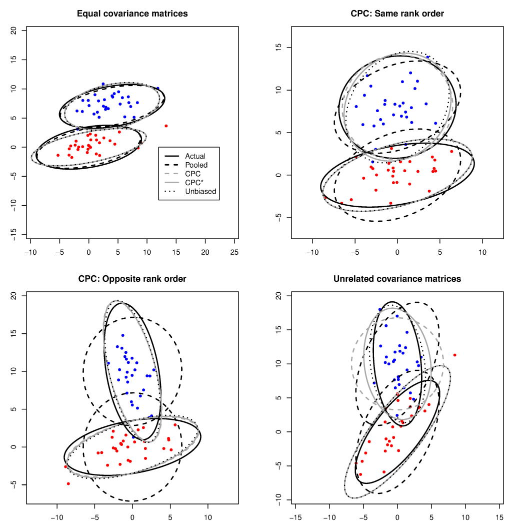
otherwise allocate it to the second group.

Covariance matrix estimators

95% confidence ellipses

$k = 2$ populations

$p = 2$ variables



Common principal components (CPC)

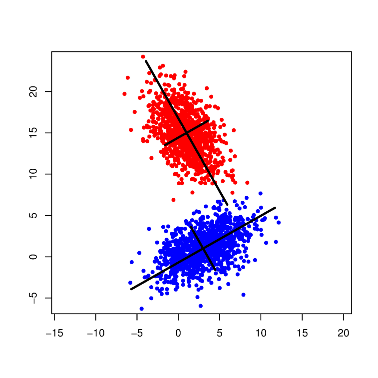
$$\boldsymbol{\Sigma}_1 = \mathbf{B}\Lambda_1\mathbf{B}'$$

$$\boldsymbol{\Sigma}_2 = \mathbf{B}\Lambda_2\mathbf{B}'$$

Example:

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

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



Purpose of the study

Can (more accurate) estimators of Σ_i under the CPC model be used to improve misclassification error rates in discriminant analysis?

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:

$$\boldsymbol{L}_i^0 = \text{diag}(\boldsymbol{B}'\boldsymbol{S}_i\boldsymbol{B}) \quad (3)$$

$$\boldsymbol{S}_{i(CPC)} = \boldsymbol{B}\boldsymbol{L}_i^0\boldsymbol{B}' \quad (4)$$

Regularised CPC estimator

$$\boldsymbol{S}^{\star}_{i(CPC)} = \alpha_i \boldsymbol{S}_i + (1 - \alpha_i)\boldsymbol{S}_{i(CPC)}, \tag{5}$$

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.

CPC discriminant analysis

Plug the CPC covariance matrix estimators into the quadratic discriminant rule:

$$-\frac{1}{2}\boldsymbol{x}'_{\text{new}}(\boldsymbol{S}^{-1}_{1(\text{CPC})}-\boldsymbol{S}^{-1}_{2(\text{CPC})})\boldsymbol{x}_{\text{new}}+(\bar{\boldsymbol{x}}'_1\boldsymbol{S}^{-1}_{1(\text{CPC})}-\bar{\boldsymbol{x}}'_2\boldsymbol{S}^{-1}_{2(\text{CPC})})\boldsymbol{x}_{\text{new}}\geq c, \tag{6}$$

where

$$c=\frac{1}{2}\ln\left(\frac{|\boldsymbol{S}_{1(\text{CPC})}|}{|\boldsymbol{S}_{2(\text{CPC})}|}\right)+\frac{1}{2}(\bar{\boldsymbol{x}}'_1\boldsymbol{S}^{-1}_{1(\text{CPC})}\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}'_2\boldsymbol{S}^{-1}_{2(\text{CPC})}\bar{\boldsymbol{x}}_2). \tag{7}$$

Simulation study

- Sample size: $n_1 = n_2 = 30, 100$ or 200
- $k = 2$ multivariate normal populations
- $p = 10$
- Different covariance matrix structures: Equal, CPC, Unrelated
- Misclassification error rates:
  - Quadratic discriminant analysis (QDA)
  - CPC discriminant analysis (CPC)
  - Regularised CPC discriminant analysis (CPC*)
  - Linear discriminant analysis (LDA)

Simulation results

		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 (similar rank orders)	30	28.58	**18.12**	18.77	33.52
	100	18.12	**14.93**	15.08	28.80
	200	15.89	**14.13**	14.26	27.49
CPC (Opposite rank orders)	30	5.20	**2.28**	2.46	24.73
	100	2.41	**1.95**	1.97	18.31
	200	1.99	**1.84**	1.85	16.56
Unrelated covariance matrices	30	13.78	8.94	**8.47**	34.93
	100	5.85	7.15	**5.57**	30.80
	200	**4.89**	6.95	4.92	29.76

Vermont Oxford Network data

Variables:

- 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)

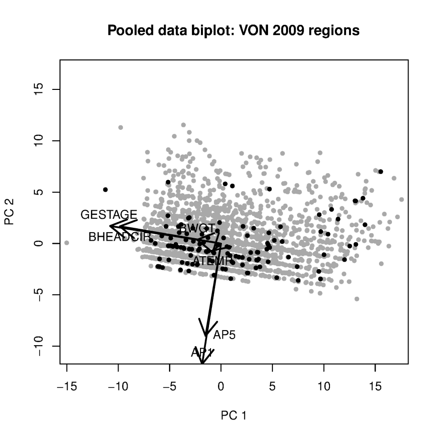
Regions:

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

Vermont Oxford Network data

Misclassification error rates:

- QDA $= 25.2\%$
- LDA $= 25.4\%$
- CPC $= 21.2\%$
- CPC$^{\star} = 22.9\%$



Conclusions

- When CPC model is appropriate: CPC discriminant analysis outperforms QDA and LDA
- CPC* offers a flexible solution, between CPC and QDA
- For small sample sizes: More parsimonious (even theoretically incorrect) model can outperform the more complex models

References

Flury, B. (1988). *Common Principal Components and Related Multivariate Models*. Wiley, 1988.

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.

Flury, B.W. and Schmid, M.J. (1992). Quadratic discriminant analysis functions with constraints on the covariance matrices: Some asymptotic results. *Journal of Multivariate Analysis*, 40:244–261.

Friedman, J.H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.

Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall.