Efficiency of the CPC estimator in modelling the covariance structures of several populations

Theo Pepler
Unit for Biometry
Stellenbosch University

Principal component analysis (PCA):

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

Common principal components (CPC):

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

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

Estimate β with Flury-Gautschi (or other) algorithm.

Question: If the CPC hypothesis is tenable, can the information about the common eigenvectors be used to find improved estimates of the covariance matrices of the populations?

CPC estimator of Σ_i (Flury, 1988)

- $\boldsymbol{S}_i$: unbiased sample covariance matrix estimator
- $\boldsymbol{B}$: estimator of modal matrix, $\boldsymbol{\beta}$

$$\boldsymbol{L}_i = \boldsymbol{B}'\boldsymbol{S}_i\boldsymbol{B} \tag{1}$$

$$\boldsymbol{L}_i^0 = \text{diag}(\boldsymbol{L}_i) \tag{2}$$

$$\boldsymbol{S}_{i(CPC)} = \boldsymbol{B}\boldsymbol{L}_i^0\boldsymbol{B}' \tag{3}$$

Regularised estimator of Σ_i (Friedman, 1989)

- S_{pool} : pooled covariance matrix estimator
- α_i : shrinkage intensity parameter

$$\boldsymbol{S}_i^{\star} = \alpha_i \boldsymbol{S}_i + (1 - \alpha_i)\boldsymbol{S}_{\text{pool}} \tag{4}$$

Substitute $S_{i(CPC)}$ for S_{pool} in Equation (4) to find regularised CPC estimator:

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

Substitute $\boldsymbol{S}_i = \boldsymbol{B}\boldsymbol{L}_i\boldsymbol{B}$ and $\boldsymbol{S}_{i(CPC)} = \boldsymbol{B}\boldsymbol{L}_i^0\boldsymbol{B}$ in Equation (5):

$$
\begin{aligned}
\boldsymbol{S}^{\star}_{i(\text{CPC})} &= \alpha_i \boldsymbol{B}\boldsymbol{L}_i\boldsymbol{B}' + (1-\alpha_i)\boldsymbol{B}\boldsymbol{L}_i^0\boldsymbol{B}' \\
&= \boldsymbol{B}[\alpha_i \boldsymbol{L}_i + (1-\alpha_i)\boldsymbol{L}_i^0]\boldsymbol{B}' \\
&= \boldsymbol{B}[\alpha_i \boldsymbol{L}_i + \boldsymbol{L}_i^0 - \alpha_i\boldsymbol{L}_i^0]\boldsymbol{B}' \\
&= \boldsymbol{B}[\alpha_i (\boldsymbol{L}_i - \boldsymbol{L}_i^0) + \boldsymbol{L}_i^0]\boldsymbol{B}'
\end{aligned}
\tag{6}
$$

Estimation of α_i:

- inverse of ϕ measure (Flury, 1988):

$$\hat{\alpha}_i = 1 - \phi(\boldsymbol{L}_i)^{-1}$$

$$= 1 - \frac{\det(\boldsymbol{L}_i)}{\det\left[\mathrm{diag}(\boldsymbol{L}_i)\right]} \tag{7}$$

$$= 1 - \frac{\det(\boldsymbol{L}_i)}{\det(\boldsymbol{L}_i^0)}$$

Estimation of α_i:

- optimisation on validation data
  Group i sample

$r = 1, \ldots, 100$ replications

70%	$\rightarrow \boldsymbol{S}^{(r)}_{i(TRAIN)}, \boldsymbol{S}^{(r)}_{i(CPC)}$
30%	$\rightarrow \boldsymbol{S}^{(r)}_{i(VALID)}$ Find $\alpha_i^{(r)}$ which minimises $\|\| \left[\alpha_i^{(r)} \boldsymbol{S}^{(r)}_{i(TRAIN)} + (1 - \alpha_i^{(r)}) \boldsymbol{S}^{(r)}_{i(CPC)}\right] - \boldsymbol{S}^{(r)}_{i(VALID)} \|\|_{F^\star}$

$$\hat{\alpha}_i = \frac{\sum_r \alpha_i^{(r)}}{r} \tag{8}$$

Estimation of α_i:

- method adapted from Schäfer & Strimmer (2005):

$$\hat{\alpha}_i = 1 - \min\left(\frac{\sum_{j\neq h} \hat{\text{Var}}(l_{ijh})}{\sum_{j\neq h} l_{ijh}^2}, 1\right) \tag{9}$$

l_{ijh} : element in the j^{th} row and h^{th} column of $\boldsymbol{L}_i$

Estimate $\hat{\text{Var}}(l_{ijh})$ with bootstrap.

Monte Carlo simulation comparing estimators:

1. Unbiased sample covariance matrix
2. CPC estimator
3. Regularised CPC estimator
   - inverse ϕ
   - optimisation with validation data
   - Schäfer & Strimmer method with bootstrap

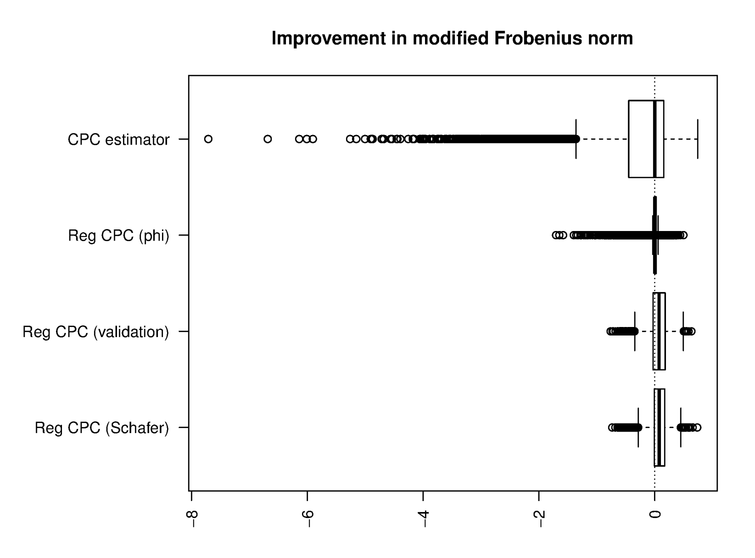
Simulation parameters

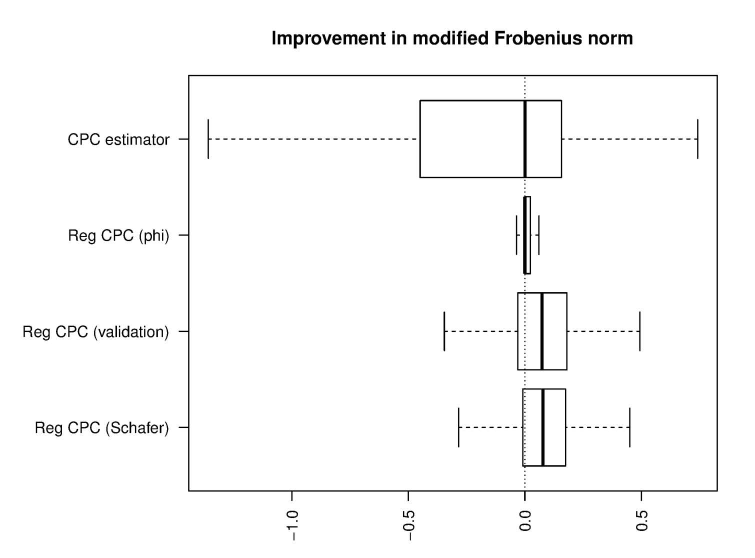
- $k = 2$ groups
- $p = 5, 10, 20$ variables
- Common eigenvector rank orders: same, similar, opposite
- Group 1 sample size: $n_1 = 200, 500, 1000$
- Group 2 sample size: $n_2 = 30, 50, 100, 200$
- Multivariate distributions: normal, chi-square (2 df), t (1 df)

Modified Frobenius norm

Error of estimation when comparing $\hat{\mathbf{\Sigma}}_i$ to $\mathbf{\Sigma}_i$:

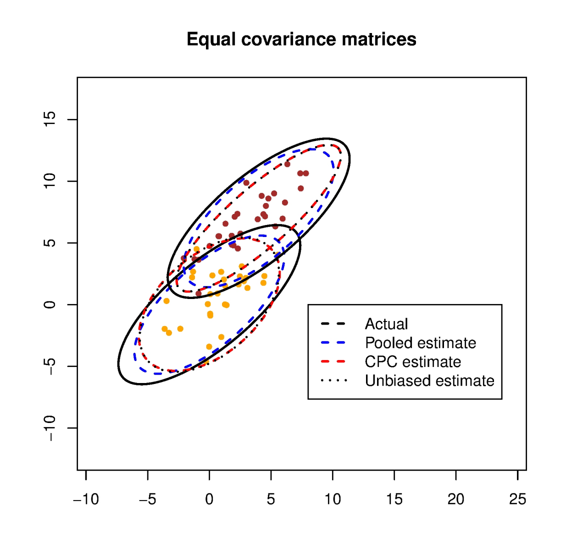
$$||\hat{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i||_{F^\star} = \sqrt{\sum_{j=1}^{p}\sum_{h\geq j}^{p}(\hat{\sigma}_{ijh} - \sigma_{ijh})^2} \tag{10}$$

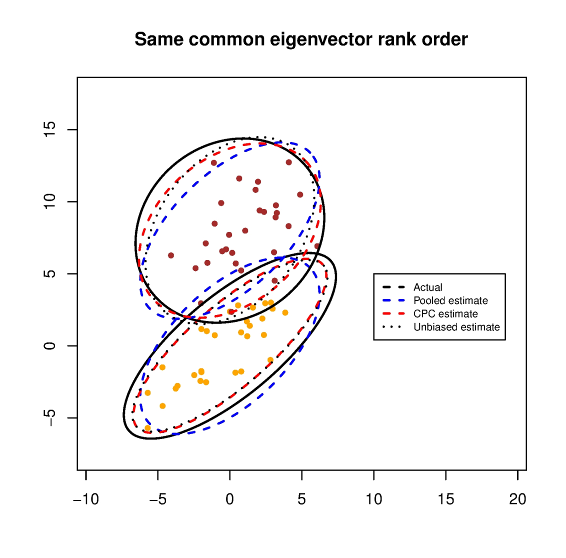

Improvement in modified Frobenius norm
CPC estimator
Reg CPC (phi)
Reg CPC (validation)
Reg CPC (Schafer)
−8
−6
−4
−2
0

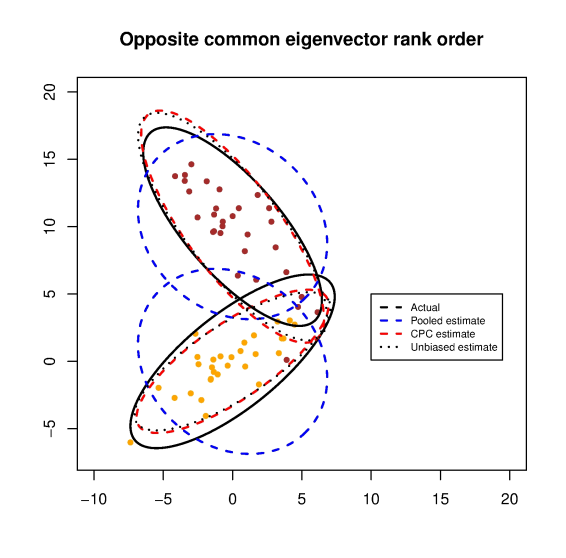

Improvement in modified Frobenius norm
CPC estimator
Reg CPC (phi)
Reg CPC (validation)
Reg CPC (Schafer)
−1.0
−0.5
0.0
0.5

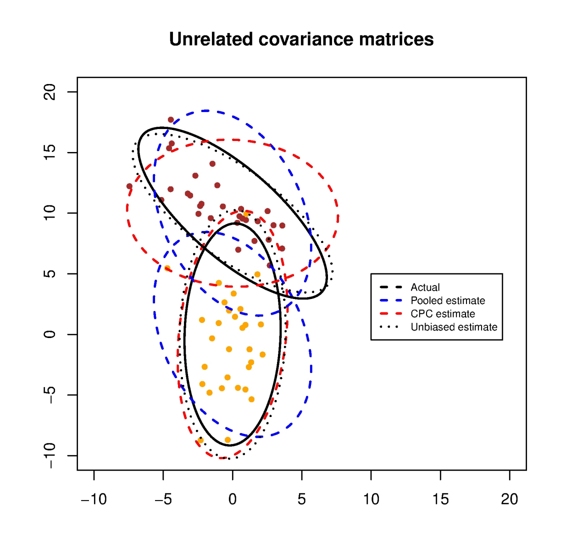
Conclusions

- Regularised CPC estimator (α_i found with validation method) performs best
- Median improvement in modified Frobenius norm of up to 27% (Multivariate t data, $n_1 = 1000$, $n_2 = 30$)
- Using CPC model in covariance matrix estimation provides greatest benefit for groups with *small* $n_i : p$ *ratios*
- Regularised CPC estimators offered greater improvement (over unbiased estimator) with *non-normal data*


Equal covariance matrices
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20
25
Actual
Pooled estimate
CPC estimate
Unbiased estimate


Same common eigenvector rank order
Actual
Pooled estimate
CPC estimate
Unbiased estimate
15
10
5
0
−5
−10
−5
0
5
10
15
20


Opposite common eigenvector rank order
Actual
Pooled estimate
CPC estimate
Unbiased estimate
−10
−5
0
5
10
15
20
−5
0
5
10
15
20


Unrelated covariance matrices
Actual
Pooled estimate
CPC estimate
Unbiased estimate
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20

References

Flury, B. (1988). *Common principal components and related multivariate models*.

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

Schäfer, J., Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, **4:1**, 1175-1189.