Measurement and Research Department Reports 2010-2

Principal Component Analysis for Exploratory Analysis of Multivariate Group Differences

Timo M. Bechger

Measurement and Research Department Reports 2010-2

Principal Component Analysis for Exploratory Analysis of Multivariate Group Differences

Timo M. Bechger

Cito
Arnhem, 2010

Abstract

Although it is relatively simple to determine whether multivariate group differences are statistically significant or not, such differences are often difficult to interpret. Especially, when there are no prior hypotheses on the relations between the variables. This paper is about multi-group principal component analysis as a tool for the investigation of multivariate group differences. As an illustration we analyze real data sets using the LISREL program.

1. Introduction

Comparative researchers are interested in differences between population means. Population data are rarely available and (population) mean differences are estimated with sample means. For a meaningful comparison it is important that the samples are chosen in such a way that the population means are estimated consistently. Furthermore, the measurement procedures must be the same across groups since otherwise the observed differences become confounded by differences in measurement procedures. Even when these conditions are met, it may be difficult to interpret the observed differences. Especially, when the variables of interest are correlated and there are no prior hypotheses about their relations. We believe that in this situation, multi-group principal component analysis (MPCA) furnishes a very useful analysis technique. MPCA is an exploratory technique that can be used, next to other methods, to suggest hypotheses for further study.

MPCA as presented here was introduced by Flury, Nel and Pienaar (1995), and Nel and Pienaar (1998). In this paper, we introduce MPCA in non-technical language and, following Dolan, Bechger and Molenaar (1999), discuss how common software for *Structural Equation Modeling (SEM)* may be used to implement MPCA. We use the LISREL notation (Jöreskog & Sörbom, 1996) because of its familiarity. Other programs for SEM, like Mx[1] (Neale, Boker, Xie & Maes, 2003), EQS (Bentler, 1995; Byrne, 2006), Mplus (Muthén & Muthén, 1998-2007), or Amos (Arbuckle, 2003; Byrne, 2009), may be used in a similar way. We analyze various real data sets to illustrate the use of MPCA in social science applications. Note that application of MPCA in the social sciences are rare. Apart from Dolan & Lubke (2001), we only know of applications in *allometry*: i.e., the study of size and shape.

[1] Now an open-source R package! consult http://www.vcu.edu/mx/

2. Takemura (1985): Hotelling's T^2 and Principal Component Analysis

Assume that two populations are being compared with respect to two or more variables. Hotelling's T^2 (Hotelling, 1931; Morrison, 1988, Chapter 4) is a generalization of the common t-test to the multivariate case (see Appendix). The null-hypothesis is that the means are equal. If the null-hypothesis is rejected, there is at least one linear combination of the variables on which the groups differ. The question is: which?

Unless we have a clear idea which linear combinations are of interest, we want *the data to suggest interesting linear combinations*. To this aim, we may consider linear combinations that optimally predict group-membership. This criterion is used with *linear discriminant analysis* (Fisher, 1936; Maddala, 1981, § 2.3 en 2.4). Takemura (1985) considers those linear combinations that have as large a variance as possible and are uncorrelated among themselves. These linear combinations are called *Principal Components (PCs)* (Pearson, 1901; Hotelling, 1933). Takemura argues that Hotelling's T^2 can be written as the sum of mean differences on the PCs and each of these differences can be tested for significance using Student's t-test (see also Basilevsky, 1994 §5.8.4).

In this paper we will bring these procedures (T^2, linear discriminant analysis, and Takemura) within a common framework for SEM. Like Dolan, et al., (1999) we believe that it is useful to formulate the problem in such a way that standard software can be used.

3. Theory

3.1. The Population Model

A PC is a linear combination of k scores: $\mathbf{Y} = (Y_1, ..., Y_k)^t$. There are k PCs $\boldsymbol{\eta} = (\eta_1, \eta_2, ...\eta_k)^t$. Each PC is a linear combination of the observed variables: $\eta_i =$

$\lambda_{1i}Y_1 + \ldots, \lambda_{ki}Y_k$, where λ_{ij} denotes the weight of the jth variable in the ith PC. In matrix notation

$$\boldsymbol{\eta} = \boldsymbol{\Lambda}\mathbf{Y}, \tag{1}$$

where $\boldsymbol{\Lambda}$ denotes a $k \times k$ matrix with the weights of k linear combinations of the observed scores. The matrix of weights is such that

$$\boldsymbol{\Lambda}\boldsymbol{\Lambda}^t = \mathbf{I} \tag{2}$$

where $\mathbf{I}$ denotes the identity matrix. Equation 2 is a restriction on the values of $\boldsymbol{\Lambda}$ called *orthonormality (ON)*. ON ensures that the PCs are uncorrelated.

We assume that $\mathbf{Y}$ is multivatiate normally distributed. The mean vector and the covariance matrix of $\mathbf{Y}$ in the gth population are

$$\boldsymbol{\mu}_g = \boldsymbol{\Lambda}^t\boldsymbol{\kappa}_g \text{ and } \boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Psi}_g\boldsymbol{\Lambda}^t. \tag{3}$$

The k-vector $\boldsymbol{\kappa}_g$ contains the means of the PCs in population g, and $\boldsymbol{\Psi}_g$ is a $k \times k$ diagonal matrix with the variances of the PCs in the gth population.

The null hypothesis is that the mean differences are equal, i.e.,

$$H_0 : \boldsymbol{\mu}_g = \boldsymbol{\tau} \quad \text{(for all groups g)} \tag{4}$$

The alternative hypothesis states that the means are not equal: That is,

$$\begin{aligned} H_1 : \boldsymbol{\mu}_1 &= \boldsymbol{\tau} \\ \boldsymbol{\mu}_g &= \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\alpha}_g \quad (g \neq 1) \end{aligned}$$

where the vector $\boldsymbol{\alpha}_g$ denotes the vector with mean differences on the PCs. That is,

$$\boldsymbol{\alpha}_g = \boldsymbol{\kappa}_g - \boldsymbol{\kappa}_1 = \boldsymbol{\Lambda}^t(\boldsymbol{\mu}_g - \boldsymbol{\mu}_1). \tag{5}$$

Under H_1, $\boldsymbol{\alpha}_1 = \mathbf{0}$ and $\boldsymbol{\alpha}_g \neq \mathbf{0}$, for $g \neq 1$.

3.2. Implementing Orthonormality

ON is a *non-linear* restriction. Specifically, (2) states that

$$\sum_j \lambda_{ij}^2 = 1 \text{ for all } i, \text{ and } \sum_{i,j:i\neq j} \sum_k \lambda_{ik}\lambda_{jk} = 0.$$

There are various ways to impose non-linear restrictions during estimation. One particular simple way is proposed by Dolan (1996).

This works as follows: The estimates are the values that minimize a *fitting function*: An increasing function of the difference between the observed means, variances and covariance and those specified under the model (Bollen, 1989, pp. 104-143). Here we use maximum likelihood estimation assuming multivatiate normality, and the fitting function is proportional to the logarithm of the likelihood function (Bollen, 1989, Appendix 4A). Each group adds a term to the fitting function, weighted by the sample size in each group. To impose ON, we specify an additional, *dummy*, group. There are no data for this group and we choose the observed covariance matrix $\mathbf{S}_d$ and mean $\bar{\mathbf{y}}_d$ ourselves. Specifically, $\mathbf{S}_d = \mathbf{I}$, and $\bar{\mathbf{y}}_d = \mathbf{0}$. As a sample size we choose any large number, e.g., $n = 5,000,000$. In the dummy group, we specify the model as $\mathbf{\Sigma}_d = \mathbf{\Lambda}\mathbf{\Lambda}^t$, and $\boldsymbol{\mu}_d = \mathbf{0}$. The estimation algorithm will try to minimize the difference $\mathbf{S}_d - \mathbf{\Sigma}_d = \mathbf{I} - \mathbf{\Lambda}\mathbf{\Lambda}^t$ and thus impose ON. In view of the large sample size, the difference will be very small.

3.3. Testing

Now that we have specified the model as a LISREL model it becomes simple to test hypotheses using the *chi-square difference* (i.e., likelihood ratio) test (Bollen, 1989, p. 292).

It is opportune to start by testing whether the weights $\mathbf{\Lambda}$ are the same in all groups. This is essential, because it allows us to connect *within-group* differences, expressed in the variances and covariances to *between-group* differences, expressed in the means (Eq. 5), and interpret the latter in terms of PCs that are common to all groups. This assumption is equivalent to measurement invariance in factor analysis (e.g., Meredith, 1993).

If the hypothesis of equal weights is not rejected we may subsequently test whether the means are equal. To this aim, we compare the fit of a model with $\boldsymbol{\alpha}_g = \mathbf{0}$ for all groups to an model without such restrictions. The same procedure may be used to test the validity of other restrictions. We may for example, test wheter the variances of the PCs are equal among the groups (i.e., $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$ for all groups g). Further possibilities are illustrated below.

In the Appendix we show how the correct number of degrees of freedom is calculated when ON is imposed by means of a dummy group.

3.4. Identifiability

The variances of the PCs must be different, otherwise the model cannot be estimated. The reason is that the weights are not uniquely determined in this case. To see why, it is useful to consider PCs as principal axes describing the distribution of the data with variances that are proportional to the length of the axes. When the variances are equal it means that the distribution is spherical and the location of the principal axes arbitrary (Jolliffe, 2002, §2.4).

In practice, it is unlikely that all or some of the PCs have equal variances. However, even when the variances are distinct but almost equal, one may expect the estimation algorithm to have problems converging.

3.5. T^2, Discriminant Analysis and Takemura (1985)

Hotelling's T^2 compares two groups. It assumes that both samples are drawn from normally distrubuted populations with equal variances and covariances. That is,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^t = \boldsymbol{\Sigma}_2. \tag{6}$$

This assumption is called *homoskedasticity.* If we indicate estimates with a hat,

$$\begin{aligned}
T^2 &= \frac{n_1 n_2}{n_1 + n_2}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^t \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \\
&= \frac{n_1 n_2}{n_1 + n_2}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^t (\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}}^t)^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \\
& \quad \Downarrow \text{ Using ON} \\
&= \frac{n_1 n_2}{n_1 + n_2}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^t \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}}^t(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \qquad (7) \\
& \quad \Downarrow \text{ Using Eq. 5} \\
&= \frac{n_1 n_2}{n_1 + n_2}\hat{\boldsymbol{\alpha}}_2^t \hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\alpha}}_2 \\
&= \frac{n_1 n_2}{n_1 + n_2}\sum_{j=1}^{k} \frac{\hat{\alpha}_{2j}^2}{\hat{\psi}_{jj}} \quad , \qquad (8)
\end{aligned}$$

where n_1 and n_2 are the sample sizes in the two groups, and $\hat{\alpha}_{2j}^2/\hat{\psi}_{jj}$ the contribution of the jth PC to T^2. LISREL can calculate T^2 directly. If, under the stated assumptions, we test the hypothesis that $\boldsymbol{\alpha}_2 = \mathbf{0}$ the chi-square difference test is equal to Hotelling's T^2 (Anderson, 1958, Theorem 5.2.1).

Takemura (1985) proposes to assess the significance of the contribution of each PC using a t-test. A simulation study by Jolliffe, Morgan and Young (1996) shows that this works well. MPCA extents Takemura's procedure to the case with more than two groups allowing for heteroskedasticity. To determine the significance of each $\hat{\alpha}_{2j}^2$, it is convenient to use a Wald-test (Wald, 1943; Bollen, 1989, Eq. 7.81); i.e., we look whether the value 0 is in the (asymptotic) confidence intervals.

Finally, note that the vector $(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^t \hat{\boldsymbol{\Sigma}}^{-1} = \hat{\boldsymbol{\alpha}}_2^t \hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}}^t$ (see Eqs. 5 and 7) contains the weights for the linear discriminant function (cf. Morrison, 1988, Equation 4 in §6.1).

- Surinam Children, N=93

$$\mathbf{S}_1 = \begin{bmatrix} 42.2200 & & \text{symm.} \\ 27.2810 & 29.9770 & \\ 17.9870 & 14.1870 & 32.3610 \end{bmatrix}$$

$$\bar{\mathbf{y}}_1 = (28.2260, 23.8170, 23.4840)^t$$

- Dutch Children, N=92

$$\mathbf{S}_2 = \begin{bmatrix} 77.9940 & & \text{symm.} \\ 39.7600 & 44.0680 & \\ 20.3250 & 25.2680 & 49.2630 \end{bmatrix}$$

$$\bar{\mathbf{y}}_2 = (28.3040, 23.6740, 26.4780)^t$$

4. Application

4.1. Lexical Ability of two Groups of Children

In this application we analyse samples of childen from The Republic of Suriname and the Netherlands in the age of 11 to 13 years. We wish to investigate whether the groups differ with respect to three aspects of lexical language proficiency:

1. *Power*: The number of words the child knows.
2. *Precision*: How well the words are understood.
3. *Variety*: The number of different meanings of each word the child knows.

There are 93 Surinam children and 92 Dutch children in the sample. Summary statistics are in Table 4.1. The variables have been obtained by summing the answers to a number of dichotomous items (Kolf, 1996).

According to Hotelling's T^2, the mean differences are statistically significant (F(3,181) = 4.76, p = 0.003). However, the assumption of homoskedasticity is rejected ($\chi^2 = 28.39$, 6 df, p<0.001). Hence, we cannot interpret the value of T^2. A model with equal PCs with different variances gives a better fit ($\chi^2 = 9.00$, 3 df, p=0.03). We will continue with this model. Parameter estimates are in Table 4.1. A LISREL script is added in the Appendix.

We order the PCs according to their variance ψ_{ii}. Hence, the PC with the largest variance is the first, the one with the next largest variance the second, etc. In both groups, the first two PCs explain about 90 % of the total variance. Looking at the weights it is clear that the first component may be be interpreted as a general lexical ability.[2] The second component gives the contrast between *variety* and *power*: A child with a high value on this PC knows relatively few words but knows many different meanings of these words. The third and last PC may be interpreted as the contrast between *precision* and the other variables. To improve the interpretation of the second PC we test whether the weight, $\lambda_{2,3}$, for *precision* can be set to zero. This hypothesis could not be rejected. ($\chi^2 = 0.6$, 1 df, p = 0.43). Finally, on the basis of the estimated standard errors we conclude that the two groups differ only with respect to the second PC (z = 3.00, p = 0.002). Dutch children know more words but know less different meanings of words. Note that this illustrates that the PC which explains most variability *within* groups, need not be the one explaining *between*-group differences.

In practice, one would likely use the unweighted sum of the three variables to score the performance of the children. This can be justfied in this case because the fit of the model does not deteriorate when the weights for the general lexical proficiency PC are constrained to be equal. The unweighted sum of the variables is thus suitable as a measure of general lexical proficiency which in this study explains about 70%

[2] The largest PC will always have positive weights while the others will have negative and positive weights. This is an inherent property of PCs.

- Children from Suriname:

$$\boldsymbol{\Psi}_1 = [9.51(1.40) 75.97(11.20) 19.0738(2.81)]$$

$$\text{\% of total variance: 10\% 72\% 18\%}$$

- Children from the Netherlands:

$$\boldsymbol{\Psi}_2 = [14.17(2.10) 118.21(7.52) 38.95(5.77)]$$

$$\text{\% of total variance: 9\% 69\% 22\%}$$

- Both Groups (*=fixed at zero):

$$\boldsymbol{\tau} = (28.23(0.70), 23.82(0.56), 23.48(0.56))^t$$

$$\boldsymbol{\alpha}_2 = (-1.00(0.52), 1.27(1.46), 2.52(0.80))^t$$

$$\boldsymbol{\Lambda} = \begin{bmatrix} -0.467(0.02) & 0.717(0.03) & -0.516(0.05) \\ 0.838(0.02) & 0.545(0.02) & * \\ -0.282(0.03) & 0.433(0.04) & 0.856(0.03) \end{bmatrix}$$

of the variance. This can be interpreted as the reliability of the unweighted score. Dutch children and Surinam children score equally on the general lexical proficiency.

4.2. MPCA versus Discriminant Analysis

The data for this example are taken from Morrison (1988, Example 6.1). The data are the scores, $X_1, \ldots, X_4$, on four scales of the *Wechseler Adult Intelligence Scale (WAIS)* of 49 elderly males who had previously been diagnosed healthy or senile. Morrison only reports the means and within-group variance-covariance matrices. For this analysis, we assume that this is the common variance-covariance. We cannot test the hypothesis of var-cov homogeneity.

As one would expect, Hotelling's T^2 indicates that the means are significantly different. The following discriminant function was found:

$$Y = 0.03X_1 + 0.20X_2 + 0.01X_3 + 0.44X_4, \tag{9}$$

where Y is a binary indicator of group membership. Morrison interpreted this result as follows: “We see that the second test, similarities, and the fourth ’picture completion’, dominate the function, while the verbal subtests information and arithmetic make only a neglible contribution to it. In subsequent investigations, attention might be concentrated on the similarities and picture completion tests as indicators of the senile-factor quantity.”

For comparison we analyze the same data with MPCA. We assume that the PC with the largest variance represents general intelligence. We found that general intelligence explains about 70 % of the variance in either group. As judged from the estimated standard errors, the groups differ only on this PC ($\alpha_2(1) = -6.951(1.87)$). Thus, discriminant analysis and MPCA provide very different pictures of the same data. We think that this is useful. For explorative purposes, it is useful to apply different techniques to the same data and discuss the findings with content experts.

5. Discussion

We have discussed a somewhat neglected statistical technique of which we believe that it can be instrumental in explorative comparative research. Compared to *Common PCA* (CPCA: Flury, 1988; Jolliffe, 2002, §13.5), or *simultaneous PCA (SCA)* (Milsap & Meredith, 1988; Kiers & Ten Berge, 1994), the method presented here is attractive because it includes a model for the means. Note that MPCA has been presented here as an extension of Hotelling’s T^2. It can also be considered as a form of multi-group confirmatory factor analysis (Sörbom, 1974) albeit with non-standard constraints.

Like Dolan, et al., (1999) we have taken the view that it is useful to formulate the problem in such a way that standard software can be used. This is, of course, not necessary. In fact, the statistical problems involved are not difficult and, in general, it pays-off to develop special purpose software. However, being able to use standard software helps in two ways: First, it means that one can quickly try-out MPCA to see if it is worthwhile to make the effort to develop dedicated software/statistical tests. Second, the flexibility of standard software makes it simple to extent the models. Dolan, et al. (1999) discuss several extensions, including the use of covariates, and present a wide range of related models. They discuss, for example, the analysis of patterned correlation matrices which could be useful for the (exploratory) analysis of multi-trait, multi-method type of data. The main reason to write this article was that we think that it is a pity that this work has gone largely unnoticed. Note that not all possibilities offered by general software packages will prove sensible in the current context.

In closing, we mention three caveats. First, models that include PCs have the property that PCs are dependent upon the unit of measurement of the observed variables. That is, the interpretation of the PCs may change radically when we choose a different scale (see Flury & Riedwyl, 1988; Flury, 1988, p. 158; Jolliffe, 2002, §2.3). This means that the results pertain exclusively to the measures used in a particular study and one must believe that the units of measurement chosen for each variable in $\mathbf{Y}$ makes sense. Note that this scale dependence is not present in SCA or confirmatory factor analysis. Second, the method breaks down when (some of the) PCs have equal variance. Last but not least, as with any exploratory method there is no guarantee that it will give interpretable results.

6. References

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis.* New-York: Wiley.

Arbuckle, J. L. (2003). *Amos 5.0: Update to the Amos user's guide.* Chicago, Smallwaters Corporation.

Basilevsky, A. (1994). *Statistical factor analysis and related methods: Theory and practice.* New-York: Wiley.

Bentler, P. M. (1995). *EQS Structural Equations Program Manual.* Encino, CA: Multivariate Software, Inc.

Byrne, B. (2006). *Structural equation modeling with EQS: Basic concepts, Applications, and programming.* Second Edition, Lawrence Erlbaum Associates: New-York.

Byrne, B. (2009). *Structural equation modeling with Amos: Basic concepts, Applications, and programming.* Second Edition, Lawrence Erlbaum Associates: New-York.

Dolan, C. V. (1996). Principal components analysis using LISREL 8. *Structural Equation Modeling: A multidisciplinary Journal*, *3*, 307-322.

Dolan, C. V. , Bechger, T. M. , & Molenaar, P. C. M. (1999). Using structural equation modeling to fit models incorporating principal components. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(3), 233-261.

Dolan, C. V., & Lubke, G. H. (2001). Viewing Spearman's hypothesis from the prespective of multigroup PCA: A comment on Schönemann's criticism. *Intelligence*, *29*, 231-245.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenetics*, *7*, 179-188.

Flury, B. D. (1988). *Common principal components and related multivariate* *models*. New-York: Wiley and Sons.

Flury, B. D. & Riedwyl, H. (1988). *Multivariate statistics. A practical approach.* London: Chapman and Hall.

Flury, B. D. , Nel, D. G. , & Pienaar, I. (1995). Simultaneous detection of shift in mean and variances. *Journal of the American Statistical Association. 90*, 1474-1481.

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, *2*, 360-378.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417-441, 498-520.

Jöreskog, K. G. , & Sörbom, D. (1993). *Lisrel 8: Structural Equation modeling with the Simplis command Language.* Chicago: Scientific Software International.

Jolliffe, I. T. (2002). *Principal component analysis.*Second Edition, Springer: New-York.

Jolliffe, I. T., Morgan, B. J. T., & Young, P. J. (1996). A simulation study of the use of principal components in linear discriminant analysis. *J. Statist. Comput. Simul.*, *55*, 353-366.

Kiers, H. A. L., & Ten Berge, J. M. F. (1994). Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure. *British Journal of Mathematical and Statistical Psychology*, *47*, 109-126.

Kolf, E. (1996). *De lexicale vaardigheid van Surinaamse en Nederlandse kinderen*[transl.: Lexical ability of Children from Suriname and the Netherlands]. Unpubished Master Thesis. Faculty POW, University of Amsterdam.

Maddala, G. S. (1983). *Limited-dependent variable and qualitative variables in Econometrics.* Econometric Society Monographs nr. 3. Cambrigde University Press.

Meredith, W. (1993). Measurement invariance, Factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543.

Millsap, R. E. , & Meredith, W. (1988). Component analysis in cross-sectional and longitudinal data. *Psychometrika*, *53*, 123-134.

Morrison, D. F. (1988). *Multivariate statistical methods*, 2de editie. New-York: McGraw-Hill.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus User's guide.* Fifth Edition. Los Angeles, CA: Muthén & Muthén.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling.* VCU Box 900126, Richmond, Va 23298: Department of Psychiatry, 6th Edition.

Nel, D. G., & Pienaar, I. (1998). The decomposition of the Behrens-Fisher statistic in q-dimensional common principal component submodels. *Ann. Inst. Statist. Math.*, *50*, 241-252.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Phil. Mag.*m *2*,559-572.

Schuurs, U. (1998). Verschillen in leesvaardigheid in eerste en tweede taal. [transl: Differences in reading ability in the first and second language.]*Spiegel*, *16*(2), 27-42.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*,229-239.

Takemura, A. (1985). *A principal decomposition of Hotelling's T^2 statistic.* In Multivariate Analysis VI, P. R. Krishnaiah, pp. 538-597. New-York: Elsevier Science Publishers.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 389-407.

7. Appendix

7.1. LISREL setup for the first application

The LISREL model used is:

$$\boldsymbol{\mu}_g = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{\alpha}_g \quad \boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Psi}_g \boldsymbol{\Lambda}_g^t + \boldsymbol{\Theta}_g$$

where:

1. $\boldsymbol{\Lambda}_g$ is invariant across the groups (ly=in. In the first group ly=fu,fr)
2. $\boldsymbol{\Theta}_g = \mathbf{0}$ (te=ze)
3. $\boldsymbol{\Psi}_g$ is the diagonal matrix of PC variances (ps=di,fr when the variances are allowed to differ)
4. $\boldsymbol{\alpha}_g$ is the vector of differences in PC means (al=fu,fr. In the first group al=ze).
5. $\boldsymbol{\tau}_g$ denotes the means in the first group (ty=in. In the first group ty=fu,fr and initialized on the observed means).

Note further that ne=k and ny=k. In the dummy group: $\boldsymbol{\alpha}_g = \boldsymbol{\tau}_g = \mathbf{0}$.

```
titel: group 1
Da no=93 ni=3 ng=3
Cm fi=groep1.cov
me fi=groep1.me
mo ny=3 ne=3 te=ze ly=fu,fr al=ze ty=fu,fr ps=di,fr
ma ly
1 0 0
0 1 0
0 0 1
fi ly 2 3
ma ps
```

```
70 30 3
ma ty
28.226 23.8170 23.4840
ou nd=4 it=500

Titel group 2
DA no=92 ni=3
CM fi=groep2.cov
me fi=groep2.me
mo ly=in ps=di,fr te=ze al=fu,fr ty=in
ma ps
30 10 1
ma al
0 0 0
ou

title dummy group
da ni=3 no=5000000
cm sy
1
0 1
0 0 1
me
0 0 0 0
mo ly=in ps=di,fi te=ze al=fu,fi ty=fu,fi
ma ps
```

```
1 1 1
ma ty
0 0 0
ma al
0 0 0
ou add=off
```

7.2. The Correct number of Degrees of Freedom

To determine the number of degrees of freedom, we must count the number of free parameters, and we must count the number of independent variance, covariances and means. The difference between these counts is the number of degrees of freedom.

Let G denote the number of groups in the analysis. We assume that $\mathbf{\Lambda}$ is invariant across the groups. The ON restriction (2) implies that $\frac{k^2+k}{2}$ entries in $\mathbf{\Lambda}$ can be determined from the remaining $\frac{k(k-1)}{2}$ entries. This means that there are $\frac{k(k-1)}{2}$ free parameters in $\mathbf{\Lambda}$. Furthermore, we have Gk variance on the diagonal of $\boldsymbol{\Psi}$, $(G-1)k$ parameters in $\boldsymbol{\kappa}$ en k parameters in $\boldsymbol{\tau}$. In total, we have:

$$\frac{k(k-1)}{2} + Gk + (G-1)k + k = \frac{1}{2}k\left(k - 1 + 4G\right) \tag{10}$$

free parameters in the model with invariant weights but no restrictions on the means and variances of the PCs. For each additional restriction we must subtract the number of restricted parameters. For instance, if we assume equal means we subtract $(G-1)k$ since there are $(G-1)k$ free parameters α_{gj}. If we assume that $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$, for all $g = 1, \ldots, G$, there are $\frac{1}{2}k\left(k + 1 + 2G\right)$ free parameters.

With G groups, there are $G\frac{k(k+1)}{2}$ independent variances and covariance, and the number of observed means is Gk. The correct number of degrees of freedom is therefore

$$G\frac{k(k+1)}{2} + Gk - \frac{1}{2}k\,(k-1+4G) = \frac{1}{2}k\,(G-1)\,(k-1) \tag{11}$$

which simplifies to $\frac{1}{2}k\,(k-1)$ when $G = 2$, and to $\frac{1}{2}k\,(G-1)\,(k+1)$ under the assumption of equal variances and covariances.

7.3. Hotelling's T^2 as an extension of Student's t-test

In the text we remark that Hotelling's T^2 can be considered as an extension of the t-test. We will now explain this. For starters, we have:

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \Leftrightarrow \boldsymbol{\lambda}^t\boldsymbol{\mu}_1 = \boldsymbol{\lambda}^t\boldsymbol{\mu}_2$$

where $\boldsymbol{\lambda}^t\boldsymbol{\mu}_g$ is a linear combination of the means in group g. The statistic for Student's t-test is:

$$t(\boldsymbol{\lambda}) = \frac{\boldsymbol{\lambda}^t(\bar{\mathbf{x}}_2 - \bar{\boldsymbol{x}}_1)}{\sqrt{\boldsymbol{\lambda}^t\boldsymbol{S}\boldsymbol{\lambda}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

such that

$$t^2(\boldsymbol{\lambda}) = \frac{n_1 n_2}{n_1+n_2}\frac{\boldsymbol{\lambda}^t(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1)(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1)^t\boldsymbol{\lambda}}{\boldsymbol{\lambda}^t\boldsymbol{S}\boldsymbol{\lambda}}.$$

where n_g denotes the sample size in the sample from the gth group.

Now, we determine the maximum of t^2 (and t) across all possible weights λ. Let $e_{max}[.]$ denote the largest eigenvalue of the matrix within brackets.

$$\begin{aligned}
\sup_{\lambda\neq\mathbf{0}} t^2(\lambda) &= \frac{n_1 n_2}{n_1+n_2}\sup_{\lambda\neq\mathbf{0}}\frac{\lambda^t(\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)^t\lambda}{\lambda^t\mathbf{S}\lambda}\\
&= \frac{n_1 n_2}{n_1+n_2}e_{max}[\mathbf{S}^{-1}(\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)^t]\\
&= \frac{n_1 n_2}{n_1+n_2}e_{max}[(\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)^t\mathbf{S}^{-1}(\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)]\\
&= \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)^t\mathbf{S}^{-1}(\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1)\\
&= T^2
\end{aligned}$$

This shows that the null-hypothesis tested by T^2 is that there is no linear combination of the variables on which the groups differ. An alternative (slightly more complex) proof can be found in Morrison (1988).