WP. 2
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (i): Disclosure risk assessment

Disclosure Risk from Factor Scores in a Remote Access Environment

Prepared by Philipp Bleninger and Jörg Drechsler, Institute for Employment Research, Germany and Gerd Ronning, University of Tübingen, Germany

Disclosure Risk from Factor Scores in a Remote Access Environment

Philipp Bleninger*, Jörg Drechsler*, and Gerd Ronning**

* Institute for Employment Research, Regensburger Str. 104, 90478 Nürnberg. (philipp.bleninger@iab.de;joerg.drechsler@iab.de)

** University of Tübingen, Nauklerstr. 48, 72074 Tübingen (gerd.ronning@uni-tuebingen.de)

Abstract. Remote access is a promising tool for broadening the access to microdata without violating confidentiality requirements. In a remote access setting the user submits queries to a system provided by the statistical agency and only the results of the queries are reported back to the user. Since no direct access to the data is granted, generally no alteration of the underlying microdata is required. Still, remote access bears the risk of disclosing sensitive information even though the actual data are not directly available. Most disclosive queries are easily detected and can be suppressed by the system. However, more complex procedures such as multivariate analyses can also lead to a breach of confidentiality if applied in a sophisticated manner to exploit certain features of the data.

In this paper we illustrate how an intruder could employ commonly used factor analysis to obtain sensitive information regarding the underlying microdata. We present the general concept and evaluate the approach using a German establishment survey, the IAB Establishment Panel. We find theoretical and empirical evidence for a high risk of disclosure from factor analysis.

1 Introduction

Statistical analyses via remote access seem to offer both, preservation of confidentiality and unlimited use of the data. But even though the list of allowed queries is generally limited in a remote access setting to avoid disclosure from simple attacks like maximum queries, some attacks are harder to detect especially if these attacks are based on multivariate analysis. One of the more prominent examples is the disclosure risk from linear regression. Gomatam *et al.* (2005) describe two possible strategies that an intruder with background knowledge about some of the survey respondents can apply to obtain any sensitive information contained in the dataset regarding these survey respondents. Bleninger *et al.* (2011) further formalize these strategies and apply them to a German establishment survey. They find that very limited background information is necessary to obtain exact information on sensitive attributes in the dataset. Since the risks from linear regression are well known in the SDC community, the current implementations of remote access already take measures to ensure that these strategies cannot be applied. However, this highlights the essential dilemma of the remote access environment: Possible intruder strategies need to be known in advance to enable the implementation of counter strategies. Remote analysis servers try to circumvent this dilemma by only allowing computations that are considered save under all circumstances. But as a consequence the set of allowed queries will be very limited and many users will find this set to be too limited to answer their respective research question. Thus, for most researchers full remote access is the only viable solution. In this context, full remote access would mean that only those queries that are known to be disclosive would be prohibited. But implementing such a fully automated approach would mean that all potentially risky queries are known in advance so that the number of suppressed queries can be kept at a minimum. This is an ambitious goal and it is not clear whether this goal can ever be achieved.

In this paper we provide one example of a multivariate analysis for which the potential risk of disclosure is anything but obvious: factor analysis. Factor analysis is very popular in the social sciences serving to a wide range of explorative and confirmatory tasks and it would be a severe drawback of remote access if this kind of analysis would not be possible. On the other hand, as we will illustrate in this paper, there is a risk of disclosure if unrestricted factor analysis is allowed.

The remainder of the paper is organized as follows: Following a brief description of the method, we provide a short overview of different estimation procedures for factor scores. Section 4 demonstrates that there is a risk of disclosure for all these approaches if a single variable is uncorrelated with the other variables used in the factor scores analysis. The empirical example in Section 5 shows that such a correlation structure is not uncommon in practice and once the "appropriate" set of variables is selected, it is possible to estimate the true values for every record in the dataset very precisely for the variable of interest. The data for this empirical illustration are taken from the IAB Establishment Panel, a survey conducted by the Institute for Employment Research in Germany. The paper concludes with some final remarks.

2 Some basic facts on factor analysis

Consider a set of m random variables $\boldsymbol{\eta} = (\eta_1, \eta_2 \ \ldots, \eta_m)'$ with $E[\boldsymbol{\eta}] = \boldsymbol{\mu}_\eta$, $cov[\boldsymbol{\eta}] = \boldsymbol{\Sigma}_{\eta\eta}$ for which n observations are available leading to the $(n \times m)$ data matrix $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m)$. The factor model seeks to explain the m variables by a set of $p < m$ "common factors" $\mathbf{f} = (f_1, f_2, \ldots, f_p)'$ through the linear model

$$\boldsymbol{\eta} - \boldsymbol{\mu}_\eta = \boldsymbol{\Lambda}\,\mathbf{f} + \mathbf{u} \qquad (1)$$

where Λ is the $(m \times p)$ factor loading matrix and $\mathbf{u}$ is an m-dimensional vector of "specific factors" with

$$E[\mathbf{u}] = \mathbf{0}\ ,\ cov[\mathbf{u}] = \mathbf{\Psi} = \begin{pmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_m \end{pmatrix} \qquad .$$

Since the factors are assumed to be orthogonal with $cov[\mathbf{f}] = \mathbf{I}$ as well as independent from $\mathbf{u}$, we obtain the so-called "fundamental equation"

$$\mathbf{\Sigma}_{\eta\eta} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi} \qquad .$$

Let F be the the $(n \times p)$-matrix of realized factor scores which is related to the data matrix $\mathbf{Y}$ by the equation (McDonald and Burr, 1967, p. 384)

$$\mathbf{Y} - \mathbf{M} = \mathbf{F}\mathbf{\Lambda}' + \mathbf{U}. \tag{2}$$

which implicitly defines the $(n \times m)$ matrix $\mathbf{M}$ by

$$\mathbf{M} = \boldsymbol{\iota}_n \otimes \boldsymbol{\mu}'_\eta \quad .$$

Here $\boldsymbol{\iota}_n$ is an n-vector of ones and $\otimes$ denotes the Kronecker product. We will call (2) the "empirical factor model" whereas (1) will be called the "theoretical factor model".
If the estimated matrix $\mathbf{\Lambda}$ has a block-diagonal structure, particular factors can be related to a subset of the vector $\boldsymbol{\eta}$ which helps to interpret these factors. In addition, it is possible to rotate the factors to attach factors more clearly to certain variables and to facilitate their interpretation.

3 Estimation of factor scores

This section provides a short review of the four different approaches that are discussed in the literature for obtaining factor scores (see Ronning and Bleninger (2011) for a more detailed review that also presents the derivations for all estimators). In the following we assume that the factor loading matrix $\mathbf{\Lambda}$ is known or rather has been estimated in an earlier step. Hence the resulting estimates of $\mathbf{f}$ depend on the method by which the factor loading matrix was determined. In all cases $\mathbf{\Lambda}$, resp. its estimate $\tilde{\mathbf{\Lambda}}$, represents the original or the rotated factor loadings. We will only present the results for the empirical model (2) since this will be the relevant model for our disclosure risk evaluations in the following sections. Derivation of the results for the theoretical model (1) is straightforward.

3.1 Least squares solution

The theoretical factor model (2) can be seen as a regression model with unknown vector $\mathbf{F}$ which can be estimated by least squares. The resulting estimator is

$$\hat{\boldsymbol{F}}_{LS} = (\mathbf{Y} - \mathbf{M})\,\tilde{\boldsymbol{\Lambda}}\,(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}})^{-1} \qquad . \tag{3}$$

Horst (1965) seems to have been one of the first using this approach (McDonald and Burr, 1967, p. 386).

3.2 Bartlett's method

Considering the non-scalar structure of the covariance matrix $\boldsymbol{\Psi}$, a generalized least squares formula seems more appropriate:

$$\hat{\boldsymbol{F}}_{BA} = (\mathbf{Y} - \mathbf{M})\,\tilde{\boldsymbol{\Psi}}^{-1}\,\tilde{\boldsymbol{\Lambda}}\;\,(\tilde{\boldsymbol{\Lambda}}'\,\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}})^{-1} \qquad . \tag{4}$$

Note that in this case the matrix $\boldsymbol{\Psi}$ has to be determined in advance, too. This method has been proposed by Bartlett (1937). Fahrmeir *et al.* (1996, p. 648 and 690) remark that (4) can be regarded as a Maximum Likelihood estimator when normality for $\boldsymbol{\eta}$ is assumed. Non-normally distributed variables in $\boldsymbol{\eta}$ lead to Quasi-Maximum Likelihood estimation of loadings and scores, being still asymptotically normally distributed and consistent.

3.3 Thomson's method

The method is attributed to both Thomson (1939) and Thurstone (1935). Thurstone (1935) has derived factor scores from the requirement that the estimated factor score $\hat{f}_j$ is as close to the "true" factor score f_j as possible for $j = 1, ..., p$. He considers the linear estimator

$$\hat{f}_j \quad = \quad \mathbf{a}_j'\,(\boldsymbol{\eta} - \boldsymbol{\mu})$$

for which the mean-squared error should be minimized with respect to the vector $\mathbf{a}_j$ (see Ronning and Bleninger (2011) for details). With this approach, the factor scores in the empirical model are given by:

$$\hat{\boldsymbol{F}}_{TH} = (\mathbf{Y} - \mathbf{M})\,\tilde{\boldsymbol{\Psi}}^{-1}\,\tilde{\boldsymbol{\Lambda}}\,\left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1} \quad , \tag{5}$$

3.4 Principal component analysis

Of course, the principal component approach can also be used to estimate the factor scores: If we consider the spectral decomposition of the covariance matrix

$$\boldsymbol{\Sigma}_{\eta\eta} \quad = \quad \mathbf{Q}\,\boldsymbol{\Theta}\,\mathbf{Q}' \qquad ,$$

the principal components $\mathbf{p}_j$, $j = 1, \ldots, m$, are given by the matrix

$$\left(\;\mathbf{p}_1, \;\; \mathbf{p}_2, \;\; \ldots, \;\; \mathbf{p}_{m-1}, \;\; \mathbf{p}_m\;\right) = \mathbf{P} = \mathbf{YQ} = \left(\;\mathbf{Yq}_1, \;\; \mathbf{Yq}_2, \;\; \ldots, \;\; \mathbf{Yq}_{m-1}, \;\; \mathbf{Yq}_m\;\right)$$

where the columns $\mathbf{q}_j$ are the characteristic vectors of the covariance matrix whereas the diagonal matrix $\mathbf{\Theta}$ contains the characteristic values. Usually, only the principal components corresponding to the largest characteristic values are used since they represent "maximal variation". The matrix $\mathbf{P}$ can be seen as the matrix of estimated factors, i.e.,

$$\hat{\mathbf{F}}_{PC} \quad = \quad \mathbf{P} \quad . \tag{6}$$

For more details see any textbook on multivariate analysis, e.g., Press (1972).

4 Disclosure risk from factor analysis

In this section we will illustrate, under which scenarios the factor scores disclose sensitive information. We show analytically that a severe risk of disclosure exists, if a set of variables can be identified in the dataset that is (nearly) uncorrelated with the variable of interest. For concreteness let us assume that η_1 is the variable in question so that the covariance matrix may have the following block diagonal structure:

$$\mathbf{\Sigma}_{\eta\eta} \quad = \quad \left(\begin{array}{c|c} \sigma_{11} & \mathbf{0}' \\ \hline \mathbf{0} & \mathbf{\Sigma}_{22} \end{array}\right) \tag{7}$$

where $\mathbf{\Sigma}_{22}$ is the $(m-1)\times(m-1)$ covariance matrix of the remaining $m-1$ variables. Clearly, this leads to a factor loading matrix with a first factor "loading" only on the first variable and the remaining $p-1$ factors having zero loading weight on this variable. Note that this implies

$$(\mathbf{\Lambda}'\mathbf{\Lambda})^{-1} \quad = \quad \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & (\mathbf{\Lambda}_2'\mathbf{\Lambda}_2)^{-1} \end{pmatrix} \quad , \tag{8}$$

where Λ_2 is the $(m-1)\times(m-1)$ loading matrix of the remaining $m-1$ variables.

Substituting (8) into (3) for the least-squares solution and into (4) for Bartlett's method, we obtain identical results regarding the uncorrelated variable (the derivations are presented in the Appendix)

$$\boldsymbol{F}_{LS} = \boldsymbol{F}_{BA} \quad = \quad \left(\begin{array}{c|ccc} 1 \cdot \begin{pmatrix} y_{11}-\mu_1 \\ \vdots \\ y_{n1}-\mu_1 \end{pmatrix} & \mathbf{f}_2 \; , & \ldots \; , & \mathbf{f}_p \end{array} \right) \quad .$$

Therefore for both, the least-squares solution and Bartlett's method the first factor $\mathbf{f}_1$ is identical (up to an additive constant) to the data vector $\mathbf{y}_1$ and it will be easy for the intruder to derive the values for $\mathbf{y}_1$ at least approximately since computing the mean of a variable will usually be allowed in a remote access environment. Note that only the first factor $\mathbf{f}_1$ is identical for the least-squares solution and for Bartlett's method. The estimated factors for $j = 2, ..., p$ will generally differ for the two methods.

For the solution of Thomson/Thurstone we obtain a slightly different result (again, derivations are presented in the Appendix):

$$\mathbf{F}_{TH} = \left(\frac{1}{1+\psi_1} \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \; \mathbf{f}_2 \, , \quad \ldots \; , \quad \mathbf{f}_p \right) .$$

The results show that in this case the estimated factor $\mathbf{f}_1$ differs not only by an additive constant; additionally the multiplicative factor $1/(1+\psi_1)$ has to be taken into account. If ψ is small or the estimate of ψ used in the computation is available, disclosure risk again is high.

Finally for the principal component approach one of the characteristic values, say θ_j, equals σ_{11}. The corresponding characteristic vector then must satisfy $\mathbf{q}_j = (1\ 0 \ldots 0)'$.Therefore, the corresponding principal component is given by

$$\mathbf{p}_j = \mathbf{Y}\mathbf{q}_j = \mathbf{y}_1$$

so that in this case the data vector $\mathbf{y}_1$ is exactly reproduced by the principal component. It should be noted however that θ_j is not necessarily the largest characteristic value (see Ronning and Bleninger (2011) for a formal proof). Since usually only the principal components corresponding to the largest characteristic values are used in practice, extracting the vector for components corresponding to small characteristic values might be suspicious and agencies might prevent some attacks based on this approach if only the components corresponding to the largest characteristic values can be retrieved.

5 Empirical Evidence

5.1 The data

The IAB Establishment Panel is a nationwide annual survey of establishments in Germany conducted by the Institute for Employment Research (IAB). It includes establishments with at least one employee covered by social security and contains business-related facts (e.g. management, business policy, innovations), a large number of employment policy-related subjects (e.g. personnel structure, recruitment, wages and salaries) and various background information (e.g. regional information, industrial sector). For further description see Fischer *et al.* (2008) and Kölling (2000).

The IAB collects the data under the pledge of confidentiality. Additionally, German law restricts the release of data from public administrations to avoid the disclosure of sensitive information. Therefore direct access to the survey for external researchers is only granted at the research data center (RDC) of the IAB. Alternatively researchers can submit queries to the RDC that are run on the original data by the

Table 1: Variables used in the factor scores model

variable	$\rho(lgturn, y_j)$
turnover from sales after taxes on the log scale (lgturn.)	1.0000
investments in EDP, information and communication technology equipment (inv.)	0.0587
total number of civil servant aspirants (asp.)	0.0082
total number of vacant positions for unskilled, low-, semi-, and skilled workers (vac.w.1)	0.0536
number of vacancies reported to employment agency for unskilled, low-, semi-, and skilled workers (vac.w.2)	0.0374
number of vacancies reported to employment agency for qualified employees (vac.em.)	0.1193
employees with wage subsidies (sub.)	0.0984
employees older than 50 with wage subsidies (sub.50)	0.0513

staff of the RDC. Only the results are reported back to the researchers after the output has been carefully checked for confidentiality violations. Currently all confidentiality checks are performed manually, but remote access is seen to be the gold standard for data providers. We therefore assume in the following that a remote access that doesn't restrict factor analysis is already implemented at the RDC.
For our empirical evaluations we use the cross-section from the year 2007 of the survey. All missing values in this dataset are replaced by single imputation and treated like observed values. See Drechsler (2011) for a description of the imputation of the missing values in the survey. The sensitive variable to be disclosed is the turnover from an establishment's sales after taxes. Thus we exclude all non-industrial organizations, regional and local authorities and administrations, financial institutions, and insurance companies. The remaining dataset includes 12,814 completely observed establishments.

5.2 Estimation of factor loadings

Since the very skewed distribution of the turnover variable generates some outliers among the factor scores, we transform the variable according to

$$lgturn_i = \log(turnover_i + 1)$$

before the factor analysis. The transformed variable is approximately normally distributed, leading to approximately unbiased and consistent Maximum Likelihood estimation of the corresponding loadings and scores for Bartlett's method.

In order to successfully apply the approach outlined above we need to identify variables that are (nearly) uncorrelated with this variable. It should not be difficult for an intruder to obtain this information because correlation matrices are usually not considered to provide a high risk of disclosure. Table 1 lists the eight variables that we used in the factor scores model together with their empirical correlation with the log-turnover ($\rho(lgturn)$). Of course the assumption of zero correlation underlying the results in Section 4 is unrealistic for real data settings but the correlations in Table 1 are small and we will see that the originally reported turnover can still be

Table 2: Rotated Matrix $\tilde{\mathbf{\Lambda}}$ of estimated loadings

	factor 1	factor 2	factor 3	factor 4
lgturn.	0.0202	0.0360	0.9867	0.1406
inv.	-0.0046	0.0019	0.0326	0.1888
asp.	0.0002	0.0051	0.0105	-0.0167
vac.w.1	0.9879	0.0134	0.0267	0.0487
vac.w.2	0.9325	0.0090	0.0089	0.0673
vac.em.	0.0796	0.0742	0.0853	0.2194
sub.	0.0166	0.7933	0.0719	-0.0100
sub.50	0.0041	0.9958	0.0088	0.0471

estimated almost exactly under this scenario.

Usually the number of factors p is evaluated using a screeplot (Fahrmeir *et al.*, 1996) which in our example would suggest extracting only two factors. In our application we choose $p = 4$ purposefully since only this number of factors leads to one factor solely loading on the turnover variable. We rotate the factors according to the Varimax-criterion (Kaiser, 1958) to further increase the loading on the turnover variable for this factor. Table 2 presents the loading matrix $\tilde{\mathbf{\Lambda}}$ of the rotated factors. The third factor loads primarily on turnover and thus this factor model is ideally suited for a disclosure attack.

5.3 Estimation of factor scores

In the next step, we estimate the matrix of factor scores $\hat{\mathbf{F}}$ based on the rotated loadings from Table 2. We limit our evaluation to Bartlett's (4) and Thomson's solution (5) for brevity. We note that the least squares solution and principal component analysis will provide similar results. Once we estimated the score values, we obtain the estimated values for the transformed turnover variable by adding its mean to all the factor scores based on the assumption that the mean of the (transformed) variable is available in remote access. To approximate turnover on the original scale, we transform the obtained values according to

$$\hat{turn_i} = exp\left\{\hat{lgturn_i}\right\} - 1 \qquad .$$

We note that the transformation will lead to a small bias in the estimated turnover since in general $E(\log y_i|x_i) \neq \log E(y_i|x_i)$. To evaluate how close the resulting estimate is to the reported turnover, we use the difference between reported and estimated turnover relative to the reported turnover

$$\delta_i = \frac{\hat{turn_i} - turn_i}{turn_i} \qquad , i = 1, \ldots, n\,.$$

The two leftmost panels in Figure 1 show scatter plots of these differences for Bartlett's (left panel) and Thomson's method (middle panel) respectively. In the scatter plots the establishments are ordered from largest to smallest.

Looking at the scatter plots, we find that using Bartlett's method the estimated turnover is very close to the true turnover for all establishments. The relative difference δ is close to zero for all establishments and even for the largest establishments,

Figure 1: Relative differences δ_i between the true turnover and the turnover estimated from the factor scores obtained through Barlett's method and Thompson's method with/without correction.

the estimated turnover never differs more than ± 6% from its true value. The relative difference is less than ±0.1% for more than 97% of the establishments.
For Thomson's method we find a clear trend in the relative difference between the reported turnover and the estimated turnover. The turnover derived from the factor scores overestimates the true turnover for the smallest establishments. This effect is continuously decreasing and the turnover is underestimated for large establishments. This is not surprising if we note that we obtained our estimate for turnover by adding the sample mean to the factor scores without correcting for the multiplicative factor $1/(1+\psi_1)$. Thus, assuming that y_1 is the turnover variable, the difference between the estimated and the reported turnover is given by

$$\hat{y_1} - y_1 \quad = \quad \frac{\psi_1}{1+\psi_1}\begin{pmatrix} \mu_1 - y_{11} \\ \vdots \\ \mu_1 - y_{1n} \end{pmatrix}, \tag{9}$$

which will be positive for all establishments with a turnover that is smaller than the average turnover and negative for the rest. Since turnover is highly correlated with establishment size, we observe a positive trend for the relative difference when going from the largest establishments to the smallest. We also find that there is much more variability in the relative differences.

If an estimate for the specific factor $\tilde{\psi}_1$ is available we can correct the estimator for the reported turnover. The right panel in Figure 1 presents the results based on the corrected estimate. The relative difference δ again is close to zero for almost all establishments with more than 85% of the establishments having a relative difference less than ±0.1%. In fact the estimated turnover never differs by more than ± 8% from the true turnover. Thus the risk of disclosure is comparable to the risk if Bartlett's method is applied.

6 Conclusions

There is an increasing demand from researchers for access to microdata that have been collected under the pledge of confidentiality. One promising approach to granting access without violating confidentiality guarantees is remote access. However, even though the researcher never has direct access to the underlying microdata, the approach is not free from the risk of disclosure. In our paper we illustrated this risk for a specific analysis that is commonly used in the social sciences: factor analysis. Even though factor analysis is used for information reduction and the potential risk of disclosure is anything but obvious, we showed analytically that the underlying microdata could be obtained exactly for any variable for which a set of covariables can be identified that are uncorrelated with the target variable. This result holds irrespective of which method is used to estimate the factor scores. Of course no correlation is unrealistic in practice but our empirical example illustrated that a very close approximation of the true values could be obtained even if a small correlation exists between the target variable and the other variables used in the factor model.

It is important to note at this point that the intruder will only obtain a full vector of true reported values. This will not necessarily lead to disclosure if the intruder is not able to link this information to individual units. Still, most legislations require that no individual information is released to the public no matter if a direct link is possible or not. Furthermore, it is often easy to attribute some of the obtained values to specific units, e.g., the largest turnover in the dataset.

Finally, we want to stress that it is not the aim of this paper to call for a more restrictive data access. factor analysis is a useful and widely used method that should be available for researchers in a remote access setting. We only want to raise the awareness that these kind of attacks are possible if no countermeasures are taken. Once identified, these attacks can be prevented easily for example by not reporting individual factor scores since applied analysts are usually not interested in these scores. Alternatively, if factor scores should be provided, simply checking the correlation between the factor scores and the variables in the dataset could be a useful tool for avoiding disclosure. The factor scores can be suppressed, if the bivariate correlation with any variable in the dataset is higher than an agency defined threshold, say 0.995. The aim of the paper is to illustrate that data providers granting access to sensitive data should be aware that there are many ways to obtain sensitive information without direct access to the microdata using standard analyses and not all of them are obvious.

References

Bartlett, M. (1937). The statistical conception of mental factors. *British Journal of Mathematical and Statistical Psychology* , 28, 97–104.

Bleninger, P., Drechsler, J., and Ronning, G. (2011). Remote data access and the risk of disclosure from linear regression. *SORT, Special Issue: Privacy in Statistical Databases,* 7–24.

Drechsler, J. (2011). Multiple imputation in practice – a case study using a complex german establishment survey. *Advances in Statistical Analysis* 1–26.

Fahrmeir, L., Hamerle, A., and Tutz, G. (1996). *Multivariate statistische Verfahren.* De Gruyter: Berlin, 2nd edn.

Fischer, G., Janik, F., Müller, D., and Schmucker, A. (2008). The IAB Establishment Panel-from sample to survey to projection. FDZ-Methodenreport 1.

Gomatam, S., Karr, A., Reiter, J., and Sanil, A. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statistical Science* , 20, 163–177.

Horst, P. (1965). *Factor analysis of data matrices.* Holt, Rinehart & Winston: New York.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 3, 187 – 200.

Kölling, A. (2000). The IAB-Establishment Panel. *Journal of Applied Social Science Studies* , 120, 291–300.

McDonald, R. and Burr, E. (1967). A comparison of four methods for constructing factor scores. *Psychometrika* **32**, 381–401.

Press, S. (1972). *Applied Multivariate Analysis.* Holt Rinehart and Winston: New York.

Ronning, G. and Bleninger, P. (2011). Disclosure risk from factor scores. *IAW Discussion Papers* **73**.

Thomson, G. (1939). *The Factorial Analysis of Human Ability.* University of London: London.

Thurstone, L. (1935). *The vectors of mind.* University of Chicago Press: Chicago.

A Derivations of the factor scores if one variable is uncorrelated with the other variables in the model

A.1 The least squares solution

$$\begin{aligned}
\boldsymbol{F}_{LS} &= (\mathbf{Y}-\mathbf{M})\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & (\boldsymbol{\Lambda}_2'\boldsymbol{\Lambda}_2)^{-1} \end{pmatrix} \\
&= (\mathbf{Y}-\mathbf{M})\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2\,(\boldsymbol{\Lambda}_2'\boldsymbol{\Lambda}_2)^{-1} \end{pmatrix} . \\
&= \left(1\cdot\begin{pmatrix} y_{11}-\mu_1 \\ \vdots \\ y_{n1}-\mu_1 \end{pmatrix} \middle| \; \mathbf{f}_2\,, \quad \dots \quad , \quad \mathbf{f}_p \right)
\end{aligned}$$

A.2 Barlett's method

$$\begin{aligned}
\hat{\boldsymbol{F}}_{BA} &= (\mathbf{Y}-\mathbf{M})\,\boldsymbol{\Psi}^{-1}\,\boldsymbol{\Lambda}\;(\boldsymbol{\Lambda}'\,\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1} \\
&= (\mathbf{Y}-\mathbf{M})\begin{pmatrix} \psi_1^{-1} & \\ & \boldsymbol{\Psi}_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}\left(\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}'\begin{pmatrix} \psi_1^{-1} & \\ & \boldsymbol{\Psi}_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}\right)^{-1} \\
&= (\mathbf{Y}-\mathbf{M})\begin{pmatrix} \psi_1^{-1} & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Psi}_2^{-1}\boldsymbol{\Lambda}_2 \end{pmatrix}\left(\begin{pmatrix} \psi_1^{-1} & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2'\boldsymbol{\Psi}_2^{-1}\boldsymbol{\Lambda}_2 \end{pmatrix}\right)^{-1} \\
&= (\mathbf{Y}-\mathbf{M})\left(\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Psi}_2^{-1}\boldsymbol{\Lambda}_2\,(\boldsymbol{\Lambda}_2'\boldsymbol{\Psi}_2^{-1}\boldsymbol{\Lambda}_2)^{-1} \end{pmatrix}\right) \\
&= \left(1\cdot\begin{pmatrix} y_{11}-\mu_1 \\ \vdots \\ y_{n1}-\mu_1 \end{pmatrix} \middle| \; \mathbf{f}_2\,, \quad \dots \quad , \quad \mathbf{f}_p \right)
\end{aligned}$$

A.3 The solution of Thomson/Thurstone

$$
\begin{aligned}
\mathbf{F}_{TH} &= (\mathbf{Y}-\mathbf{M})\left(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'+\hat{\boldsymbol{\Psi}}\right)^{-1}\hat{\boldsymbol{\Lambda}} \\
&= (\mathbf{Y}-\mathbf{M})\left(\begin{pmatrix}1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2'\end{pmatrix}+\begin{pmatrix}\psi_1 & \\ & \boldsymbol{\Psi}_2\end{pmatrix}\right)^{-1}\begin{pmatrix}1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2\end{pmatrix} \\
&= (\mathbf{Y}-\mathbf{M})\begin{pmatrix}(1+\psi_1)^{-1} & \mathbf{0}' \\ \mathbf{0} & (\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2'+\boldsymbol{\Psi}_2)^{-1}\boldsymbol{\Lambda}_2\end{pmatrix} \\
&= \left(\left.\frac{1}{1+\psi_1}\cdot\begin{pmatrix}y_{11}-\mu_1 \\ \vdots \\ y_{n1}-\mu_1\end{pmatrix}\right|\ \mathbf{f}_2\,,\quad \ldots\quad,\ \mathbf{f}_p\right)
\end{aligned}
$$