IAW DISCUSSION PAPERS

iaw@iaw.edu | www.iaw.edu

IAW Diskussionspapiere | Nr. 73 Juli 2011 | No. 73 July 2011

Disclosure Risk from Factor Scores

Gerd Ronning
Philipp Bleninger

Institut für Angewandte Wirtschaftsforschung e.V.
Ob dem Himmelreich 1 | 72074 Tübingen | Germany
Tel.: +49 7071 98960 | Fax: +49 7071 989699

ISSN: 1617-5654

IAW-Diskussionspapiere

Das Institut für Angewandte Wirtschaftsforschung (IAW) Tübingen ist ein unabhängiges außeruniversitäres Forschungsinstitut, das am 17. Juli 1957 auf Initiative von Professor Dr. Hans Peter gegründet wurde. Es hat die Aufgabe, Forschungsergebnisse aus dem Gebiet der Wirtschafts- und Sozialwissenschaften auf Fragen der Wirtschaft anzuwenden. Die Tätigkeit des Instituts konzentriert sich auf empirische Wirtschaftsforschung und Politikberatung.

Dieses IAW-Diskussionspapier können Sie auch von unserer IAW-Homepage als pdf-Datei herunterladen:

http://www.iaw.edu/Publikationen/IAW-Diskussionspapiere

ISSN 1617-5654

Weitere Publikationen des IAW:

- IAW-News (erscheinen 4x jährlich)
- IAW-Forschungsberichte

Möchten Sie regelmäßig eine unserer Publikationen erhalten, dann wenden Sie sich bitte an uns:

IAW Tübingen, Ob dem Himmelreich 1, 72074 Tübingen,
Telefon 07071 / 98 96-0
Fax 07071 / 98 96-99
E-Mail: iaw@iaw.edu

Aktuelle Informationen finden Sie auch im Internet unter:

http://www.iaw.edu

Disclosure Risk from Factor Scores

Gerd Ronning[1] and Philipp Bleninger[2]

Abstract

Remote access as well as remote analysis solve many problems arising from granting researchers access to sensitive data. Both allow to run analyses without actually seeing the data. Therefore none of them demand either substantively altering the data or strictly restricting the access to it. Still remote access and remote analysis bear the risk to disclose sensitive information though the actual data is not directly available. An intruder has nothing to do but to apply standard procedures in a sophisticated way to exploit certain features enabling disclosure. Even usual and unsuspicious multivariate analyses bear great potential for data snoopers.
We will illustrate how an intruder could employ commonly used factor analysis to disclose sensitive variables in a data set. We will derive the approach and evaluate it using the IAB Establishment Panel. There is theoretical and empirical evidence for the high risk for violation of confidentiality from all variants of factor analysis.

Keywords
Remote Access, Remote Analysis, Data Privacy, Disclosure Limitation, Factor Analysis, Principal Component Analysis

1 Introduction

The scientific community needs high quality data for testing single hypotheses or even whole theories empirically. But mostly data collection is expensive and laborious, so it is self-evident to go for data already surveyed appropriately by others. Public administrations, governmental agencies and other state institutions collect and produce much sought-after data. The crucial point is how to grant access to these data under legal restrictions. Most micro data sets are confidential and therefore cannot be released unrestrictedly. Statistical analyses via remote access or remote analysis seem to offer both preservation of confidentiality and unlimited use of the data. However, some precautions have to be taken in order to minimize disclosure risk. There are many sources for disclosure from statistical analysis, especially from inference.
The most obvious example is from regression where provision of both predicted values and residuals would allow to reconstruct the vector of the dependent variable. See

[1]Faculty of Economics, University of Tuebingen, Nauklerstr. 48, D-72074 Tuebingen
[2]IAB Institute for Employment Research, Regensburger Str. 104, D-90478 Nuremberg

Gomatam et al. (2005) for an overview of various scenarios involving disclosure risk. However, their paper does not include a discussion of possible disclosure from multivariate analysis, and this is typical for the entire literature related to this topic. In our short paper we demonstrate that factor analysis implies a severe risk of disclosing the micro data if factor scores are called for. These may be generated in our case either from the factor *model* or from principal components which may be seen as its *empirical counterpart*. As in the case of regression analysis we concentrate on the reconstruction of the whole data vector of a variable revealing the values of all subjects in the data set from which individual values can be extracted in a second step. The latter may be received from further background knowledge about the individual of interest.

Factor analysis is very popular in the social sciences serving to a wide range of explorative and confirmatory tasks. It might also be worth pointing out that factor scores played an important role in econometrics already for a long time. Kendall (1957) and McCallum (1970) suggested that factor scores (generated from principal components) should be used in regression analysis in order to alleviate the problem of multicollinearity in regression when the number of regressors is large. Modern time series econometrics have resumed this idea after having established desirable stochastic properties in case of time series data (Stock and Watson, 2002).

However, since the aim was to forecast macroeconomic variables, no micro data were involved in these studies. Moreover, disclosure risk from factor scores occurs (as will be demonstrated below) if some variable is uncorrelated with all others which may not be typical for macroeconomic data sets. Most recently, however, Buch et al. (2010) have used the micro data from a set of 1512 banks in their factor-augmented vector-autoregression (FAVAR) approach.

After a brief description of the respective method, we give a short overview on different estimation procedures for factor scores. Section 4 demonstrates that for all these approaches a severe disclosure risk exists if a single variable is uncorrelated with all others. The empirical example in section 5 shows that such correlation structure can be generated by selecting the ”appropriate” set of variables and that on this basis it is possible to disclose a data vector. The data are taken from the IAB Establishment Panel, a survey collected by the Institute for Employment Research of the Federal Employment Agency.

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

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1m} \\ y_{21} & y_{22} & \dots & y_{2m} \\ y_{31} & y_{32} & \dots & y_{3m} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nm} \end{pmatrix} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_m) \qquad .$$

The factor model seeks to explain the m variables by a set of $p < m$ "common factors"

$$\mathbf{f} = \left(f_1, f_2, \dots, f_p\right)'$$

by the linear model

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

where the $(m \times p)$ factor loading matrix is given by

$$\boldsymbol{\Lambda} \quad = \quad \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1p} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2p} \\ \vdots & \vdots & \dots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \dots & \lambda_{mp} \end{pmatrix}$$

and $\mathbf{u}$ is an m-dimensional vector of "specific factors" with

$$E[\mathbf{u}] = \mathbf{0}\ ,\ cov[\mathbf{u}] = \boldsymbol{\Psi} = \begin{pmatrix} \psi_1 & & & & \\ & \psi_2 & & & \\ & & \ddots & & \\ & & & \psi_{m-1} & \\ & & & & \psi_m \end{pmatrix} \qquad .$$

Often ψ_j is also called "uniqueness" as it measures the degree to which the variable j is unique in the sense of not being part of a common factor. 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"

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

Note that for each observation i we have a p-dimensional vector $\mathbf{f}_i$ of "factor scores" from (1) so that for all n observations we arrive at the $(n \times p)$-matrix

$$\boldsymbol{F} \quad = \quad \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1p} \\ f_{21} & f_{22} & \dots & f_{2p} \\ f_{31} & f_{32} & \dots & f_{3p} \\ \vdots & \vdots & & \vdots \\ f_{n1} & f_{n2} & \dots & f_{np} \end{pmatrix}$$

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

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

or more explicitly by

$$y_{ij} - \mu_j = \sum_{k=1}^{p} \lambda_{jk} f_{ik} + u_{ij} \quad , i = 1, \ldots, n\,;\, j = 1, \ldots, m\,,$$

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".
Usually only model (1) is presented and discussed since the estimated matrix of factor loadings is the only measure of interest. If the estimated matrix $\boldsymbol{\Lambda}$ 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 with regards to contents.

3 Estimation of factor scores

In the following we assume that the factor loading matrix $\boldsymbol{\Lambda}$ is known or rather has been estimated. The $\tilde{\cdot}$ indicates that the matrix $\boldsymbol{\Lambda}$ 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 $\boldsymbol{\Lambda}$, resp. its estimate $\tilde{\boldsymbol{\Lambda}}$, represents the original or the rotated factor loadings. There are four approaches to determine the matrix $\mathbf{F}$ of factor scores which will be described in the following subsections.

3.1 Least squares solution

If one considers the theoretical factor model (1) under the above assumptions, it can be seen as a regression model with unknown vector $\mathbf{f}$ which should be estimated by least squares. The resulting estimator is

$$\hat{\mathbf{f}}_{LS} = (\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}})^{-1}\tilde{\boldsymbol{\Lambda}}\,(\boldsymbol{\eta} - \boldsymbol{\mu}_\eta)\,. \tag{3}$$

Horst (1965) seems to have been one of the first using this formula (McDonald and Burr, 1967, p. 386). If we apply the same estimation principle to the empirical factor model (2) we obtain

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

3.2 Bartlett's method

If one considers the non-scalar structure of the covariance matrix $\mathbf{\Psi}$, a generalized least squares formula seems more appropriate:

$$\hat{\mathbf{f}}_{BA} = (\tilde{\mathbf{\Lambda}}'\tilde{\mathbf{\Psi}}^{-1}\tilde{\mathbf{\Lambda}})^{-1}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Psi}}^{-1}(\boldsymbol{\eta} - \boldsymbol{\mu}_\eta) \quad . \tag{5}$$

Note that now also the matrix $\mathbf{\Psi}$ has to be determined in advance. Again, the $\tilde{}$ indicates that it has been estimated. This formula has been proposed by Bartlett (1937). The corresponding result from the empirical factor model is

$$\hat{\boldsymbol{F}}_{BA} = (\mathbf{Y} - \mathbf{M})\,\tilde{\mathbf{\Psi}}^{-1}\,\tilde{\mathbf{\Lambda}}\;(\tilde{\mathbf{\Lambda}}'\,\tilde{\mathbf{\Psi}}^{-1}\tilde{\mathbf{\Lambda}})^{-1} \quad . \tag{6}$$

Fahrmeir et al. (1996, p. 648 and 690) remark that (5) can be regarded as a Maximum Likelihood estimator when normality for $\boldsymbol{\eta}$ is assumed. Non-normal distributed variables in $\boldsymbol{\eta}$ lead to Quasi-Maximum Likelihood estimation of loadings and scores, being still asymptotically normal distributed and consistent.

3.3 Thomson's method

3.3.1 The theoretical factor model

The method is attributed to both Thomson (1939) and Thurstone (1935). For the following see also the slightly different derivation in Fahrmeir et al. (1996). 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

$$\mathrm{MS}_j = E\left[(\hat{f}_j - f_j)^2\right] = E\left[((\boldsymbol{\eta} - \boldsymbol{\mu})'\,\mathbf{a}_j - f_j)^2\right] = \mathbf{a}_j'\,\mathbf{\Sigma}_{\eta\eta}\,\mathbf{a}_j \;-\; 2\,\mathbf{a}_j'\,cov[\boldsymbol{\eta}, f_j] \;+\; var[f_j]$$

should be minimized with respect to the vector $\mathbf{a}_j$. For the first derivative we obtain

$$\begin{aligned} \frac{\partial\,\mathrm{MS}_j}{\partial\,\mathbf{a}_j} \quad &= \quad 2\left\{\mathbf{\Sigma}_{\eta\eta}\,\mathbf{a}_j \;-\; cov[\boldsymbol{\eta}, f_j]\right\} \\ &= \quad 2\left\{(\mathbf{\Lambda}\mathbf{\Lambda}' \;+\; \mathbf{\Psi})\,\mathbf{a}_j \;-\; \boldsymbol{\lambda}_j\right\} \end{aligned}$$

where $\boldsymbol{\lambda}_j$ is an *m*-dimensional vector representing the *j*-th column of the matrix $\mathbf{\Lambda}$. Here we have used

$$cov[\boldsymbol{\eta}, f_j] \quad = \quad E\left[\begin{pmatrix} f_j\,\sum_{k=1}^{p}\lambda_{1k}f_k + u_1 \\ f_j\,\sum_{k=1}^{p}\lambda_{2k}f_k + u_2 \\ \vdots \\ f_j\,\sum_{k=1}^{p}\lambda_{mk}f_k + u_m \end{pmatrix}\right] \quad = \quad \begin{pmatrix} \lambda_{1j}var[f_j] \\ \lambda_{2j}var[f_j] \\ \vdots \\ \lambda_{mj}var[f_j] \end{pmatrix} \quad = \quad \begin{pmatrix} \lambda_{1j} \\ \lambda_{2j} \\ \vdots \\ \lambda_{mj} \end{pmatrix} = \boldsymbol{\lambda}_j \tag{7}$$

and the assumption of the orthogonal factor model that all p factors f_j (with unit variance) are uncorrelated.
Setting the vector of partial derivatives equal to zero results in

$$\mathbf{a}_j \quad = \quad (\mathbf{\Lambda\Lambda}' + \mathbf{\Psi})^{-1} \boldsymbol{\lambda}_j$$

which leads to the following estimator of scores of the j-th factor:

$$\hat{f}_j \quad = \quad \tilde{\boldsymbol{\lambda}}_j' \left(\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}' + \tilde{\mathbf{\Psi}}\right)^{-1} (\boldsymbol{\eta} - \boldsymbol{\mu}_\eta) \tag{8}$$

For all p factors jointly we arrive at

$$\hat{\mathbf{f}}_{TH} \quad = \quad \tilde{\mathbf{\Lambda}}' \left(\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}' + \tilde{\mathbf{\Psi}}\right)^{-1} (\boldsymbol{\eta} - \boldsymbol{\mu}_\eta) \tag{9}$$

3.3.2 The empirical factor model

The corresponding formula for the empirical factor model (2) is not so straightforward and to our best knowledge it has not been discussed in the literature. We therefore here present a detailed derivation.
We consider a linear estimator

$$\hat{\mathbf{F}} \quad = \quad (\mathbf{Y} - \mathbf{M})\,\mathbf{A}$$

where $\mathbf{A}$ is a $(m \times p)$ unknown matrix. We want the estimator to minimize the expression

$$\begin{aligned}
\phi &= E\left[tr\left\{(\hat{\mathbf{F}} - \mathbf{F})'(\hat{\mathbf{F}} - \mathbf{F})\right\}\right] \\
&= E\left[tr\left\{((\mathbf{Y} - \mathbf{M})\,\mathbf{A} - \mathbf{F})'((\mathbf{Y} - \mathbf{M})\,\mathbf{A} - \mathbf{F})\right\}\right] \\
&= E\left[tr\left\{\mathbf{A}'(\mathbf{Y} - \mathbf{M})'(\mathbf{Y} - \mathbf{M})\mathbf{A} \;-\; 2\,\mathbf{A}'(\mathbf{Y} - \mathbf{M})\mathbf{F} \;+\; \mathbf{F}'\mathbf{F}\right\}\right] \\
&= n\; tr\left\{\mathbf{A}'\mathbf{\Sigma}_{\eta\eta}\mathbf{A} \;-\; 2\,\mathbf{A}'\mathbf{\Sigma}_{\eta f} \;+\; \mathbf{\Sigma}_{ff}\right\}
\end{aligned}$$

with respect to $\mathbf{A}$ where $\mathbf{\Sigma}_{\eta f}$ is the $(m \times p)$ matrix of covariances (7) of $\boldsymbol{\eta}$ and $\mathbf{f}$ and $\mathbf{\Sigma}_{ff}$ is the $(p \times p)$ covariance matrix of the vector $\mathbf{f}$. Differentiating this expression with respect to $\mathbf{A}$ gives

$$\frac{\partial \phi}{\partial \mathbf{A}} \quad = \quad 2n\left(\mathbf{\Sigma}_{\eta\eta}\mathbf{A} \;-\; \mathbf{\Sigma}_{\eta f}\right)$$

where we have used results for matrix differentiation (Lütkepohl, 1996). Recalling the results from (7) we note that

$$\mathbf{\Sigma}_{\eta f} = \mathbf{\Lambda} \qquad .$$

Hence from setting the partial derivatives equal to zero we obtain

$$\mathbf{A} = \mathbf{\Sigma}_{\eta\eta}^{-1}\,\mathbf{\Lambda} = (\mathbf{\Lambda}' + \mathbf{\Psi})^{-1}\;\mathbf{\Lambda}$$

which leads to the so-called 'Thomson estimator' of factor scores:

$$\hat{\boldsymbol{F}}_{TH} = (\mathbf{Y} - \mathbf{M}) \left(\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Lambda}}' + \tilde{\boldsymbol{\Psi}}\right)^{-1} \tilde{\boldsymbol{\Lambda}}. \tag{10}$$

Using well-known results for inverting the sum of matrices (see Lütkepohl, 1996, p.29) and the so-called "binomial inverse theorem" (see Press, 1972, p.23) we obtain

$$\begin{aligned}
\left(\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Lambda}}' + \tilde{\boldsymbol{\Psi}}\right)^{-1} \tilde{\boldsymbol{\Lambda}} &= \left\{\tilde{\boldsymbol{\Psi}}^{-1} - \tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}}\left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\right\}\tilde{\boldsymbol{\Lambda}} \\
&= \tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}}\left\{\mathbf{I} - \left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}}\right\} \\
&= \tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}}\left\{\left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p\right) - \left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}}\right\} \\
&= \tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}}\left\{\left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p - \tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}}\right)\right\} \\
&= \tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}}\left\{\left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\left(\mathbf{I}_p\right)\right\}
\end{aligned}$$

Therefore Thomson's estimator can also be written as

$$\hat{\boldsymbol{F}}_{TH} = (\mathbf{Y} - \mathbf{M})\,\tilde{\boldsymbol{\Psi}}^{-1}\,\tilde{\boldsymbol{\Lambda}}\left(\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Psi}}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}, \tag{11}$$

In this form the estimator is usually cited, for example by Bartholomew et al. (2009, p. 574) who refer to the empirical factor model whereas Fahrmeir et al. (1996, p. 691) give the corresponding result for the theoretical factor model.

3.4 Principal component analysis

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

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

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 $\boldsymbol{\Theta}$ 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.

$$\hat{\mathbf{F}}_{PC} = \mathbf{P}. \tag{12}$$

To make the results conformable to those given above for factor analysis, the data matrix $\mathbf{Y}$ should be substituted by $\mathbf{Y} - \mathbf{M}$, i.e. deviations from the means should be considered. For more details see any textbook on multivariate analysis (Press, 1972, e.g).

4 Disclosure risk for uncorrelated variables

We now consider a certain variable uncorrelated with all other variables of a data set. (As we will see later on, in reality correlation which is almost zero will be the relevant case.) 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:

$$\mathbf{\Sigma}_{\eta\eta} \quad = \quad \left(\begin{array}{c|c} \sigma_{11} & \mathbf{0}' \\ \hline \mathbf{0} & \mathbf{\Sigma}_{22} \end{array}\right) \tag{13}$$

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.

$$\mathbf{\Lambda} \quad = \quad \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \lambda_{22} & \lambda_{23} & \dots & \lambda_{2,p-1} & \lambda_{2p} \\ 0 & \lambda_{32} & \lambda_{33} & \dots & \lambda_{3,p-1} & \lambda_{3p} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & \lambda_{m-1,2} & \lambda_{m-1,3} & \dots & \lambda_{m-1,p-1} & \lambda_{m-1,p} \\ 0 & \lambda_{m2} & \lambda_{m3} & \dots & \lambda_{m,p-1} & \lambda_{mp} \end{pmatrix} \quad = \quad \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{\Lambda}_2 \end{pmatrix} \quad .$$

Note that this implies

$$\mathbf{\Lambda}'\mathbf{\Lambda} \quad = \quad \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{\Lambda}_2'\mathbf{\Lambda}_2 \end{pmatrix} \quad \text{and} \quad (\mathbf{\Lambda}'\mathbf{\Lambda})^{-1} \quad = \quad \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \left(\mathbf{\Lambda}_2'\mathbf{\Lambda}_2\right)^{-1} \end{pmatrix} \quad .$$

It should also be noted for the following that one of the characteristic values of the covariance matrix (13) equals σ_{11}. However, this characteristic value need not to be the largest one. For example, if all variables are standardized so that the covariance matrix equals the correlation matrix, we may consider the special correlation matrix

$$\mathbf{R} \quad = \quad \left(\begin{array}{c|ccccc} 1 & 0 & 0 & \dots & 0 & 0 \\ \hline 0 & 1 & \varrho & \dots & \varrho & \varrho \\ 0 & \varrho & 1 & \dots & \varrho & \varrho \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \varrho & \varrho & \dots & 1 & \varrho \\ 0 & \varrho & \varrho & \dots & \varrho & 1 \end{array}\right)$$

with $-1/(m-1) < \varrho < 1$. Then the $m-1$ characteristic values of the lower-right block are given by

$$1+(m-1)\varrho,\ 1-\varrho,\ 1-\varrho,\dots,1-\varrho$$

whereas the m-th characteristic value equals $\theta_j = 1$ which follows from the block-diagonal structure of the matrix. However, $\theta_j = 1$ never will be largest.

4.1 The least-squares solution

For the least-squares solution (4) (disregarding notation indicating estimated quantities) we obtain in this special case

$$\begin{aligned}
\boldsymbol{F}_{LS} &= (\mathbf{Y}-\mathbf{M})\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \left(\boldsymbol{\Lambda}_2'\boldsymbol{\Lambda}_2\right)^{-1} \end{pmatrix} \\
&= (\mathbf{Y}-\mathbf{M})\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2\left(\boldsymbol{\Lambda}_2'\boldsymbol{\Lambda}_2\right)^{-1} \end{pmatrix} \\
&= \left(\begin{array}{c|ccc} \begin{pmatrix} y_{11}-\mu_1 \\ y_{21}-\mu_1 \\ y_{31}-\mu_1 \\ \vdots \\ y_{n1}-\mu_1 \end{pmatrix} & & & \\ 1\cdot & \mathbf{f}_2\,, & \mathbf{f}_3\,, & \ldots\ ,\ \mathbf{f}_{p-1}, \quad \mathbf{f}_p \end{array}\right)
\end{aligned}$$

Therefore the first factor $\mathbf{f}_1$ is identical (up to an additive constant) to the data vector $\mathbf{y}_1$.

4.2 The Bartlett solution

$$\begin{aligned}
\hat{\boldsymbol{F}}_{BA} &= (\mathbf{Y}-\mathbf{M})\,\boldsymbol{\Psi}^{-1}\,\boldsymbol{\Lambda}\,(\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1} \\
&= (\mathbf{Y}-\mathbf{M})\begin{pmatrix} \psi_1^{-1} & \\ & \boldsymbol{\Psi}_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}\left(\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}'\begin{pmatrix} \psi_1^{-1} & \\ & \boldsymbol{\Psi}_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2 \end{pmatrix}\right)^{-1} \\
&= (\mathbf{Y}-\mathbf{M})\begin{pmatrix} \psi_1^{-1} & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Psi}_2^{-1}\boldsymbol{\Lambda}_2 \end{pmatrix}\left(\begin{pmatrix} \psi_1^{-1} & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Lambda}_2'\boldsymbol{\Psi}_2^{-1}\boldsymbol{\Lambda}_2 \end{pmatrix}\right)^{-1} \\
&= (\mathbf{Y}-\mathbf{M})\left(\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\Psi}_2^{-1}\boldsymbol{\Lambda}_2\left(\boldsymbol{\Lambda}_2'\boldsymbol{\Psi}_2^{-1}\boldsymbol{\Lambda}_2\right)^{-1} \end{pmatrix}\right) \\
&= \left(\begin{array}{c|ccc} \begin{pmatrix} y_{11}-\mu_1 \\ y_{21}-\mu_1 \\ y_{31}-\mu_1 \\ \vdots \\ y_{n1}-\mu_1 \end{pmatrix} & & & \\ 1\cdot & \mathbf{f}_2\,, & \mathbf{f}_3\,, & \ldots\ ,\ \mathbf{f}_{p-1}, \quad \mathbf{f}_p \end{array}\right)
\end{aligned}$$

As in the least-squares solution the first factor $\mathbf{f}_1$ is identical (up to an additive constant) to the data vector describing the variable y_1. However, the estimated factors for $j =$

2, ..., p differ from the least-squares solution.

4.3 The solution of Thomson/Thurstone

$$
\begin{aligned}
\mathbf{F}_{TH} &= (\mathbf{Y}-\mathbf{M})\left(\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}\right)^{-1}\hat{\mathbf{\Lambda}} \\
&= (\mathbf{Y}-\mathbf{M})\left(\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{\Lambda}_2\mathbf{\Lambda}_2' \end{pmatrix} + \begin{pmatrix} \psi_1 & \\ & \mathbf{\Psi}_2 \end{pmatrix}\right)^{-1}\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{\Lambda}_2 \end{pmatrix} \\
&= (\mathbf{Y}-\mathbf{M})\begin{pmatrix} (1+\psi_1)^{-1} & \mathbf{0}' \\ \mathbf{0} & \left(\mathbf{\Lambda}_2\mathbf{\Lambda}_2' + \mathbf{\Psi}_2\right)^{-1}\mathbf{\Lambda}_2 \end{pmatrix} \\
&= \left(\begin{array}{c|ccc} \frac{1}{1+\psi_1}\cdot\begin{pmatrix} y_{11}-\mu_1 \\ y_{21}-\mu_1 \\ y_{31}-\mu_1 \\ \vdots \\ y_{n1}-\mu_1 \end{pmatrix} & \mathbf{f}_2\,, & \mathbf{f}_3\,, \ \ldots\ , & \mathbf{f}_{p-1}, \quad \mathbf{f}_p \end{array}\right)
\end{aligned}
$$

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!

4.4 Factors from the principal component approach

In case of the special covariance matrix (13) one of the characteristic values, say θ_j, equals σ_{11}which, of course, is not necessarily the largest characteristic value. The corresponding characteristic vector $\mathbf{q}_1$ then must satisfy

$$
\mathbf{q}_j = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \quad .
$$

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.

5 Empirical Evidence

5.1 Data

The IAB Establishment Panel is a nationwide annual survey of enterprises in Germany representing all kinds of business conducted by the Institute for Employment Research (IAB). It includes establishments with at least one employee covered by social security. It contains a lot of 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 location, industrial sector). For further description see Fischer et al. (2008) and Kölling (2000).
Of course, establishments attending the survey do not want their sensitive information to be available. Additionally, German law restricts the release of data from public administrations (which the IAB actually is part of) to secure privacy and preserve at least de facto anonymity. Therefore the IAB Establishment Panel is maintained by a Research Data Center (FDZ) granting access, running requested analyses and controlling the output. Remote analysis and even more remote access is seen to be the gold standard for data providers. In the following we assume that the FDZ would have already implemented a remote access to the data set or at least developed a remote analysis server under regular conditions.
To make our point we try to intrude the cross-section from the year 2007. All missing values in this data set are replaced by single imputation and treated like observed values. See Drechsler (2010) 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 data set includes 12,814 completely observed enterprises.

5.2 Estimation of factor loadings

Since in empirical analysis often the data are transformed by the logarithmic function in order to reduce skewness and kurtosis, instead of turnover we will use

$$y^*_{1,i} = \log(turnover_i + 1)$$

in the factor analysis. Note that the transformation leads to a variable which is approximately normal distributed, leading to Maximum Likelihood estimation of the corresponding loadings and scores. However, it should be noted, too, that the intruder finally wants to know turnover itself so that we have to identify the estimated turnovers by using the exponential function. As will become clear from our example, even these additional transformations will not offer security against disclosure of the original turnovers. To

obtain the actual values we re-compute the estimated turnovers by the exponential function.
The other auxiliary variables are chosen strategically according to the modelling described above, without regards to content. They are all more or less uncorrelated with the turnover and its logarithm. Therefore the empirical correlation matrix **R** (see Table 1) approximates the assumption of zero correlation in section 4 very well. Altogether we are using the following eight variables which we will refer to by their abbreviation.

1. turnover (turn.) from sales after taxes, resp. its logarithm (lgturn.)
2. investments in EDP, information and communication technology equipment (inv.)
3. total number of civil servant aspirants (asp.)
4. total number of vacant positions for unskilled, low-, semi-, and skilled workers (vac.w.1)
5. number of vacancies notified to employment office for unskilled, low-, semi-, and skilled workers (vac.w.2)
6. number of vacancies notified to employment office for qualified employees (vac.em.)
7. employees with wage subsidies (sub.)
8. employees older than 50 with wage subsidies (sub.50)

As we can see from the first column of Table 1 the logturnover is more or less independent from the other variables. The latter have any interrelations we actually do not care about.
In a real data setting zero correlation will hardly occur and hence exact results as

Table 1: Empirical correlation matrix **R** of the logarithmic turnover and the auxiliary variables

	lgturn.	inv.	asp.	vac.w.1	vac.w.2	vac.em.	sub.	sub.50
lgturn.	1.0000	0.0587	0.0082	0.0536	0.0374	0.1193	0.0984	0.0513
inv.	0.0587	1.0000	-0.0075	0.0057	0.0083	0.0440	0.0020	0.0111
asp.	0.0082	-0.0075	1.0000	-0.0003	-0.0004	-0.0011	0.0015	0.0045
vac.w.1	0.0536	0.0057	-0.0003	1.0000	0.9249	0.0925	0.0285	0.0199
vac.w.2	0.0374	0.0083	-0.0004	0.9249	1.0000	0.0905	0.0222	0.0160
vac.em.	0.1193	0.0440	-0.0011	0.0925	0.0905	1.0000	0.0641	0.0853
sub.	0.0984	0.0020	0.0015	0.0285	0.0222	0.0641	1.0000	0.7901
sub.50	0.0513	0.0111	0.0045	0.0199	0.0160	0.0853	0.7901	1.0000

given in section 4 are unelikely. In fact, there is an interrelation between the number

of vacancies notified to the employment office for qualified employees and the turnover ($\varrho_{1,5} = 0.10$) resp. its logarithm ($\varrho_{1,5} = 0.12$), but it is low, especially in comparison to the other correlations in the matrix (e.g. $\varrho_{6,7}$ and $\varrho_{3,4}$), and in the end it will turn out to be almost negligible.
All data manipulations and analyses are done using the procedure `factanal` of the statistical software package `R 2.9.0` (R Development Core Team, 2008). It offers Maximum Likelihood estimation of the loadings using starting values for the variances of the specific factors, i.e. the uniquenesses ψ_j, according to Jöreskog (Lawley and Maxwell (1971), p. 31) for the quasi-Newton-method of maximization.
Usually the number p of factors is evaluated using Bartlett's test of sphericity (Tobias and Carlson, 1969) or using screeplots (Fahrmeir et al., 1996). This would lead to extraction of only two factors. However, here we choose $p = 4$ arbitrarily considering only its purpose of disclosure.
As in the initial matrix of estimated loadings both wage subsidies load too high on the

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

third factor disturbing its relation with the turnover and diminishing its desired usability for disclosure, we rotate the factors. Rotation is done according to the Varimax-criterion (Kaiser, 1958). We choose this criterion for two reasons: First, high loadings of each variable should only result for a few factors and the rest should be near zero. Second, it rotates the factors orthogonally, so we maintain the structure of the loadings. Third, it is the most common criterion and therefore it is unsuspicious. In our case the resulting loading matrix $\tilde{\mathbf{\Lambda}}$ of rotated factors is just as we want it to be, i.e. it supports disclosure (see Table 2).
The loadings $\tilde{\lambda}_{j,k}, \quad j = 1,\ldots,8\,; k = 1,\ldots,4$ given in Table 2 come very close to the loadings matrix required for a disclosure. Obviously we are able to extract one single factor (the third factor), almost perfectly loading ($\tilde{\lambda}_{1,3} = 0.9867$) solely on the sensitive turnover and therefore almost perfectly correlated only with the target variable. The scores of the third factor will almost equal the true logarithmic values of the variable.
Looking at the uniqueness (variance of specific factors) $cov[\mathbf{u}] = \tilde{\Psi}_j$ in Table 3 we rec-

ognize that the logturnover is very well explained by its factor and that there is almost no variance $\tilde{\psi}_1$ left for its specific factor u_1 (Note that the statistical software package R turns all near-zero values in $\tilde{\mathbf{\Psi}}$ to a default value of 0.005 to avoid problems with optimization. Anyway, in our case we get the desired, very small $\tilde{\psi}_1$.). This factor model obviously serves our purpose of disclosure very well.

Tables 2 and 3 show that under usual circumstances this factor analysis would be

Table 3: Matrix $\tilde{\mathbf{\Psi}}$ of uniquenesses

	lgturn.	inv.	asp.	vac.w.1	vac.w.2	vac.em.	sub.	sub.50
lgturn.	0.0050							
inv.		0.9633						
asp.			0.9996					
vac.w.1				0.0208				
vac.w.2					0.1257			
vac.em.						0.9327		
sub.							0.3652	
sub.50								0.0061

rejected instantly by a serious researcher looking for relevant results and not for disclosure. There are three variables, namely the investments in EDP, information and communication technology equipment, the total number of civil servant aspirants and finally the number of vacancies notified to employment office for qualified employees, not sufficiently loading on any common factor, but being characterized by very high uniquenesses pointing at major specific factors.

5.3 Estimation of factor scores

In the next step, we estimate the matrix $\hat{\mathbf{F}}$ of scores of the factors with the rotated loadings from Table 2. We use Bartlett's solutions (5) as well as Thomson's solution (9) as outlined above. Having estimated the score values, we solve the empirical model for the estimated logturnover $\hat{\mathbf{y}}_\mathbf{1}^*$ using its mean μ_1^* under the assumption that the mean of a (transformed) variable is available in remote access. Of course an intruder is not interested in logarithms of turnovers. As already mentioned we re-compute the estimated turnovers to obtain estimations $\hat{y}_{1,i}$ of the actual values of every establishment $i = 1, \ldots, n$ in the data set by

$$\hat{y}_{1,i} = exp\left\{\hat{y}_{1,i}^*\right\} - 1 \qquad .$$

Obviously this is a crude method leading to biased results (especially as the expectation is biased), but as we will see its actual effect on the disclosure is very small: Despite the log-transformation and exp-retransformation we are able to disclose the true values

almost perfectly.
For $\hat{\mathbf{y}}_\mathbf{1}$ being only an approximation of the true $\mathbf{y}_\mathbf{1}$ we need a criterion for the reliability of the disclosure. To asses the actual degree of disclosure we use the difference between real and approximated turnover relative to the real turnover

$$\delta_i = \frac{\hat{y}_{1,i} - y_{1,i}}{y_{1,i}} \quad , i = 1, \ldots, n\,.$$

For illustration we will show scatter plots of these differences including a Locally

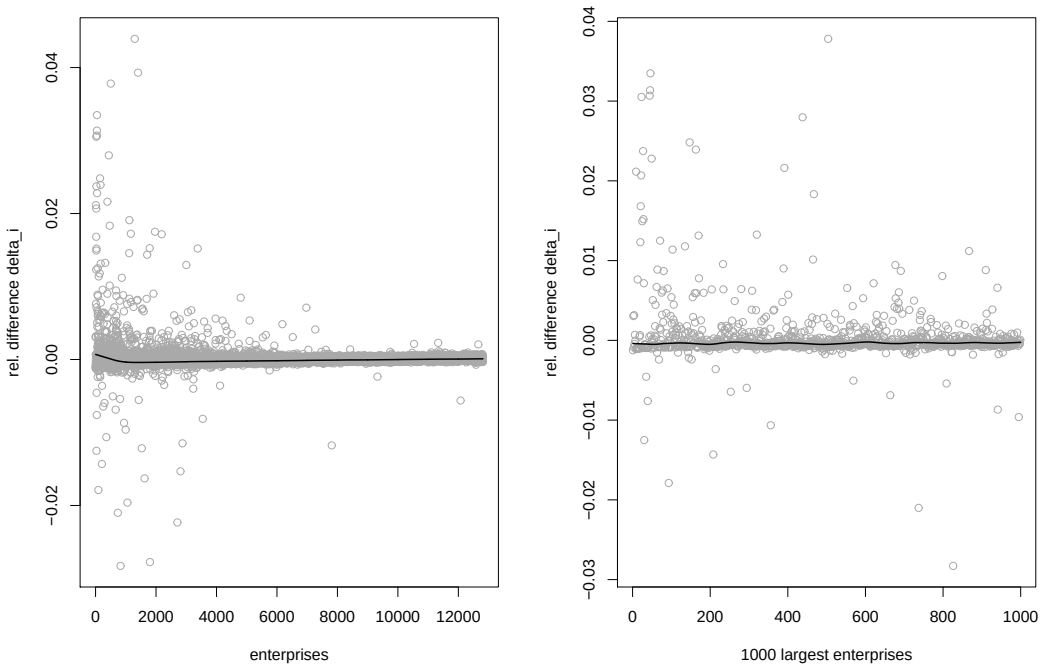

Figure 1: Relative differences δ_i for Bartlett's factor scores (with LOWESS)

Weighted Scatter plot Smoother (Cleveland, 1979, LOWESS). It illustrates the differences over all establishments very well without further assumptions. The LOWESS is based on a local polynomial fitted into the 10% nearest neighbours weighted tricubically. The scatter plots are interpreted in the manner that the disclosure is good if the differences δ_i are dispersed narrowly around zero. The smoothers are interpreted in the manner that in the case of disclosure the LOWESS is a straight horizontal line on zero. In any other cases it shows the bias in the estimated values regarding disclosure.

First we present the disclosure using Bartlett's method to estimate factor scores. The solution in section 4.2 shows that the factor scores approximate the data vector very well. Of course, here the factor is not exactly equal to the variable as there are empirical

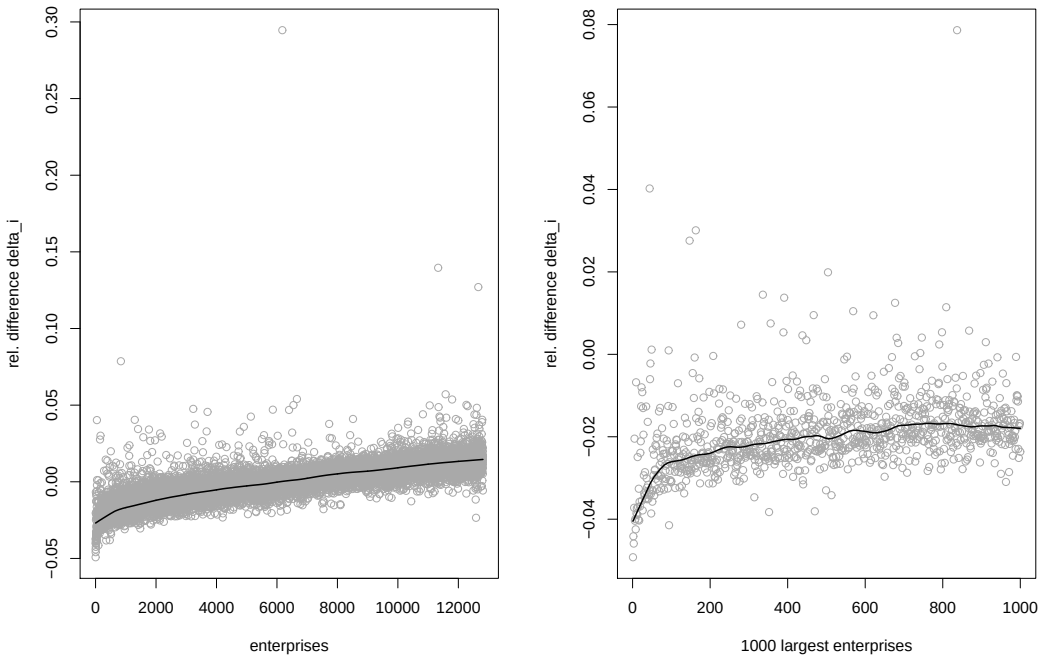

Figure 2: Relative differences δ_i for Thomson's factor scores (with LOWESS)

correlations greater zero disturbing the disclosure. However, both the average relative difference $\bar{\delta} = 2.781e-06$ and the variance of the differences $s_\delta^2 = 2.913e-06$ are almost zero resulting in a nearly perfect disclosure of virtually all the establishments' turnover. The disclosure is illustrated in Figure 1. The left panel shows the results for the whole Establishment Panel and the right panel shows the result specifically for the 1000 largest establishments as they may be of special interest to an intruder. Please note that the enterprises are ordered from the largest to the smallest, i.e. in the figures the size of establishments decreases from the left to the right. In the left panel the leftmost enterpises already contain the 1000 largest establishments from the right panel, but a a more detailed view can be insightful. As one can see the relative differences between the true turnovers and the turnovers approximated from the factor scores are very low without relevant dispersion.The LOWESS is almost a straight line on zero though a small window (of only 10 %) was chosen which usually leads to wiggliness of the smoother.

We also estimate the factor scores using Thomson's method. Again the scores have to be re-scaled and re-computed. The vector δ of relative differences again mirrors the degree of disclosure of the true turnover of the establishments. Figure 2 shows the results the same way as above. As we can see the factors scores estimated with Thomson's method also disclose the true values with an average of $\bar{\delta} = 0.0001$ ($s_\delta^2 = 0.0002$). As

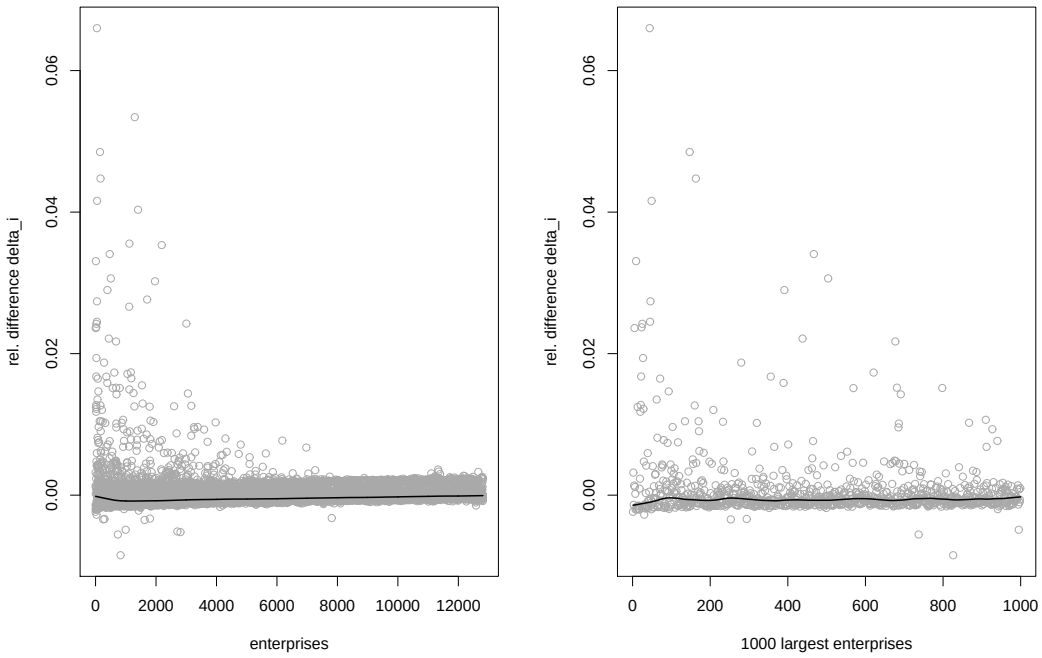

Figure 3: Relative differences δ_i for Thomson's ψ-corrected factor scores (with LOWESS)

we stated in section 4.3 Thomson's solution leads to estimates biased by the variance of the turnover not explained by its factor, namely by the multiplicator $\frac{1}{1+\psi_1}$. Indeed, according to Figure 2 we underestimate the true turnovers, as the relative differences deviate systematically from zero towards negative values. For the largest enterprises in the data which we consider to be of main interest to an intruder the true values are even more underestimated. We can see the systematic bias especially from the LOWESS in Figure 2.

To obtain disclosure we have to correct the estimations by the multiplicator $1+\tilde{\psi}_1$. Table 3 reports a uniqueness $\tilde{\psi}_1 = 0.005$ of the logturnover. Indeed, according to Figure 3, the estimations $\hat{y}_{1,i}$ from the corrected Thomson-scores are generally much closer to the real turnovers. In particular, we do not underestimate the turnovers anymore and consequently the relative differences are not only diminished but also there is no systematic deviation below zero (see especially the right panel of Figure 3).

6 Conclusions

There is an increasing demand from researchers for micro data from varying sources underlying restrictions preserving confidentiality and privacy. We address this problem in a remote analysis and remote access setting. We restrict to the risk of disclosure from factor analysis. We choose factor analysis for being a widely used technique, incorporated in every standard statistical package. In addition, it is usually based on micro data instead of covariance. On first sight you would not expect it to bear the risk of disclosure for micro data as it is usually used for dimension reduction and feature extraction, i.e. for the compression of detailed information.

Disclosure risk from factor scores arises if a single variable is almost uncorrelated with all other variables in the data set. An intruder may choose the target and the set of other variables strategically as we illustrate quite drastically in our empirical example. Though we do not have optimal circumstances with uncorrelated variables and though we perform profound transformations and re-transformations of the target variable, the risk of disclosure remains serious. Most important, our paper shows that disclosure risk is not limited to a certain variant of factor analysis but it exists in all alternative approaches for estimating factor scores. Although the principal components method is the most common method of factor extraction and score estimation, we do not give empirical evidence. However, we provide theoretical considerations about the disclosure risk which the principal components method bears. In the end this method is as risky as the others.

The usual definition of confidentiality aims at individual privacy and the disclosure of single values. Through factor analysis whole data vectors are revealed and you cannot assign the values to individuals immediately. Anyhow we want this approach to be seen as the first step towards violation of individual privacy. In our view you cannot ignore the risk of an intruder revealing whole data vectors.

Of course, once discovered, this kind of attacks can be easily prevented by suppressing estimation methods, detailed parameter and score estimations and model diagnostics (e.g. the uniquenesses). Even the output of the factor scores could be suppressed and instead the researcher is only allowed to use them for further analyses (e.g. Principal Component Regression). However, we want to emphasize that this abuse of factor analysis is only a showcase. Data providers granting access to sensitive data need to be sensitized to statistical disclosure. There are many ways to obtain sensitive information using standard analyses and not all of them are that obvious.

References

Bartholomew, D., I. Deary, and M. Lawn (2009). The origin of factor scores: Spearman, thomson and bartlett. *British Journal of Mathematical and Statistical Psychology* (62), 569 – 582.

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

Buch, C., S. Eickmeier, and E. Prieto (2010). Macroeconomic factors and micro-level bank risk. *Deutsche Bundesbank Discussion Paper Series 1: Economic Studies* (20).

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*, 829–836.

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

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

Gomatam, S., A. Karr, J. Reiter, and A. Sanil (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.

Kendall, M. (1957). *A Course in Multivariate Analysis*. Charles Griffin and Company Ltd.: London.

Kölling, A. (2000). The iab-establishment panel. *Journal of Applied Social Science Studies* (120), 291–300.

Lawley, D. and A. Maxwell (1971). *Factor Analysis as a Statistical Method* (Second ed. ed.). Butterworths.

Lütkepohl, H. (1996). *Handbook of Matrices*. Wiley: New York.

McCallum, B. (1970). Artificial orthogonalization in regression analysis. *The Review of Economics and Statistics* (52), 110– 113.

McDonald, R. and E. Burr (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.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Stock, J. and M. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

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.

Tobias, S. and J. Carlson (1969). Bartlett's test of sphericity and chance findings in factor analysis. *Psychometrika 4*(3), 375 – 377.

A Equicorrelation

We assume that pairwise correlation of random variables $X_1, \ldots, X_r$ is equal to ϱ with

$$\frac{1}{r-1} > \varrho > 1.$$

The correlation matrix $\mathbf{R}$ can then be written as

$$\mathbf{R} \quad = \quad (1-\varrho)\,\mathbf{I}_r \;+\; \varrho\,\boldsymbol{\iota}_r\boldsymbol{\iota}_r' \qquad .$$

where $\mathbf{I}_r$ is the $(r \times r)$ identity matrix and $\boldsymbol{\iota}_r$ is an r dimensional vector of ones. Characteristic values λ_j of this matrix are given by

$$\begin{aligned} \lambda_1 &= 1-(r-1)\varrho \\ \lambda_j &= 1-\varrho\,, \;\; j = 2, \ldots, r\,, \end{aligned} \tag{14}$$

and the corresponding characteristic vectors $\mathbf{x}_j$ satisfy the following conditions:

$$\begin{aligned} \mathbf{x}_1 &= \frac{1}{\sqrt{r}}\,\boldsymbol{\iota}_r \\ \boldsymbol{\iota}_r'\,\mathbf{x}_j &= 0\,, \;\; j = 2, \ldots, r\,, \end{aligned} \tag{15}$$

Proof of (14): Using theorem 8.4.3 or 8.4.4 of Graybill (1969) we can write the characteristic equation as

$$det(\mathbf{R} - \lambda\,\mathbf{I}_r) \quad = \quad (1 + (r-1)\varrho - \lambda)\,(1-\varrho-\lambda)^{(r-1)} \quad = \quad 0$$

from which the above results follow.

Proof of (15): The characteristic vector corresponding to the first characteristic value λ_1 must satisfy

$$(\mathbf{R} - (1+(r-1)\varrho)\mathbf{I})\;\mathbf{x}_1 \quad = \quad \mathbf{0}$$

or

$$((1-\varrho)\,\mathbf{I}_r \;+\; \varrho\,\boldsymbol{\iota}_r\boldsymbol{\iota}_r' \;-\; (1+(r-1)\varrho)\mathbf{I})\;\mathbf{x}_1 \quad = \quad \mathbf{0}$$

leading to

$$\mathbf{x}_1 \quad = \quad \frac{\boldsymbol{\iota}_r'\,\mathbf{x}_1}{r}\,\boldsymbol{\iota}_r \qquad .$$

Since the characteristic vectors are normalized, $\mathbf{x}_1$ must satisfy

$$1 \quad = \quad \mathbf{x}_1'\mathbf{x}_1 \quad = \quad \frac{1}{r}\,(\,\boldsymbol{\iota}_r'\,\mathbf{x}_1\,)^2 \qquad .$$

Taking into account that all elements of $\mathbf{x}_1$ must be equal, we arrive at

$$\mathbf{x}_1 \quad = \quad \frac{1}{\sqrt{r}}\,\boldsymbol{\iota}_r \qquad .$$

Using the same approach again for $\lambda_2, \ldots, \lambda_r$ the corresponding characteristic vectors $\mathbf{x}_j$ must satisfy

$$((1-\varrho)\,\mathbf{I}_r \;+\; \varrho\,\boldsymbol{\iota}_r\boldsymbol{\iota}_r' \;-\; (1\;-\;\varrho)\mathbf{I})\;\mathbf{x}_j \quad = \quad \mathbf{0}$$

leading to

$$\varrho\,\boldsymbol{\iota}_r\boldsymbol{\iota}_r'\,\mathbf{x}_j \quad = \quad \mathbf{0}$$

from which

$$\boldsymbol{\iota}_r'\,\mathbf{x}_j \quad = \quad 0$$

follows.

IAW-Diskussionspapiere

Die IAW-Diskussionspapiere erscheinen seit September 2001. Die vollständige Liste der IAW-Diskussionspapiere von 2001 bis 2008 (Nr. 1-44) finden Sie auf der IAW-Internetseite www.iaw.edu/publikationene/iaw-diskussionspapiere.

IAW-Diskussionspapiere seit Juli 2008:

Nr. 45 (Oktober 2008)
Effects of Dismissal Protection Legislation on Individual Employment Stability in Germany
Bernhard Boockmann / Daniel Gutknecht / Susanne Steffes

Nr. 46 (November 2008)
Trade's Impact on the Labor Share: Evidence from German and Italian Regions
Claudia M. Buch / Paola Monti / Farid Toubal

Nr. 47 (März 2009)
Network and Border Effects: Where Do Foreign Multinationals Locate in Germany?
Julia Spies

Nr. 48 (März 2009)
Stochastische Überlagerung mit Hilfe der Mischungsverteilung (Stand: 18. März 2009 – Version 49)
Gerd Ronning

Nr. 49 (April 2009)
Außenwirtschaftliche Verbindungen der deutschen Bundesländer zur Republik Österreich
Anselm Mattes /Julia Spies

Nr. 50 (Juli 2009)
New Firms – Different Jobs? An Inquiry into the Quality of Employment in Start-ups and Incumbents (Stand: 28. Juli 2009 – Version 1.3)
Andreas Koch / Jochen Späth

Nr. 51 (Juli 2009)
Poverty and Wealth Reporting of the German Government: Approach, Lessons and Critique
Christian Arndt / Jürgen Volkert

Nr. 52 (August/September 2009)
Barriers to Internationalization: Firm-Level Evidence from Germany
Christian Arndt / Claudia M. Buch / Anselm Mattes

Nr. 53 (September 2009)
IV-Schätzung eines linearen Panelmodells mit stochastisch überlagerten Betriebs- und Unternehmensdaten
Elena Biewen / Gerd Ronning / Martin Rosemann

Nr. 54 (November 2009)
Financial Constraints and the Margins of FDI
Claudia M. Buch / Iris Kesternich / Alexander Lipponer / Monika Schnitzer

Nr. 55 (November 2009)
Offshoring and the Onshore Composition of Tasks and Skills
Sascha O. Becker / Karolina Ekholm / Marc-Andreas Muendler

Nr. 56 (November 2009)
Intensifying the Use of Benefit Sanctions – An Effective Tool to Shorten Welfare Receipt and Speed up Transitions to Employment?
Bernhard Boockmann / Stephan L. Thomsen / Thomas Walter

Nr. 57 (November 2009)
The Responses of Taxable Income Induced by Tax Cuts – Empirical Evidence from the German Taxpayer Panel
Peter Gottfried / Daniela Witczak

Nr. 58 (November 2009)
Reformoption Duale Einkommensteuer – Aufkommens- und Verteilungseffekte
Peter Gottfried / Daniela Witczak

IAW-Diskussionspapiere

Nr. 59 (Februar 2010)
The Impact of Horizontal and Vertical FDI on Labor Demand for Different Skill Groups
Anselm Mattes

Nr. 60 (Februar 2010)
International M & A: Evidence on Effects of Foreign Takeovers
Anselm Mattes

Nr. 61
The Impact of Regional Supply and Demand Conditions on Job Creation and Destruction (Februar 2010)
Raimund Krumm / Harald Strotmann

Nr. 62 (April 2010)
The Effects of Foreign Ownership Change on the Performance of German Multinational Firms
Christian Arndt / Anselm Mattes

Nr. 63 (April 2010)
The Export Magnification Effect of Offshoring
Jörn Kleinert / Nico Zorell

Nr. 64 (März 2010)
Kundenbetreuung aus einer Hand im SGB II? – Integration versus Spezialisierung von Fallmanagement, Vermittlung und materiellen Leistungen
Harald Strotmann / Martin Rosemann / Sabine Dann / Christine Hamacher

Nr. 65 (Mai 2010)
The Combined Employment Effects of Minimum Wages and Labor Market Regulation – A Meta-analysis
Bernhard Boockmann

Nr. 66
Remote Access – Eine Welt ohne Mikrodaten ?? (Stand: 20.06.2010, Version 18) (Juni 2010)
Gerd Ronning / Philipp Bleninger / Jörg Drechsler / Christopher Gürke

Nr. 67
Opening Clauses in Collective Bargaining Agreements: More Flexibility to Save Jobs? (Oktober 2010)
Tobias Brändle / Wolf Dieter Heinbach

Nr. 68
Interest Rate Policy and Supply-side Adjustment Dynamics (Dezember 2010)
Daniel Kienzler / Kai Schmid

Nr. 69
Should Welfare Administration be Centralized or Decentralized? Evidence from a Policy Experiment (Dezember 2010)
Bernhard Boockmann / Stephan L. Thomsen / Thomas Walter / Christian Göbel / Martin Huber

Nr. 70
Banks in Space: Does Distance Really Affect Cross-Border-Banking? (Februar 2011)
Katja Neugebauer

Nr. 71
An Almost Ideal Wage Database Harmonizing the ILO Database October Inquiry (Februar 2011)
Daniela Harsch / Jörn Kleinert

Nr. 72
Disclosure Risk from Interactions and Saturated Models in Remote Access (Juni 2011)
Gerd Ronning

Nr. 73
Disclosure Risk from Factor Scores (Juli 2011)
Gerd Ronning / Philipp Bleninger