VERSITA

Journal of Official Statistics, Vol. 30, No. 1, 2014, pp. 107–122, http://dx.doi.org/10.2478/jos-2014-0006

Disclosure Risk from Factor Scores

Jörg Drechsler[1], *Gerd Ronning*[2], *and Philipp Bleninger*[3]

Remote access can be a powerful tool for providing data access for external researchers. Since the microdata never leave the secure environment of the data-providing agency, alterations of the microdata can be kept to a minimum. Nevertheless, remote access is not free from risk. Many statistical analyses that do not seem to provide disclosive information at first sight can be used by sophisticated intruders to reveal sensitive information. For this reason the list of allowed queries is usually restricted in a remote setting. However, it is not always easy to identify problematic queries. We therefore strongly support the argument that has been made by other authors: that all queries should be monitored carefully and that any microlevel information should always be withheld. As an illustrative example, we use factor score analysis, for which the output of interest – the factor loading of the variables – seems to be unproblematic. However, as we show in the article, the individual factor scores that are usually returned as part of the output can be used to reveal sensitive information. Our empirical evaluations based on a German establishment survey emphasize that this risk is far from a purely theoretical problem.

Key words: Remote data access; confidentiality; statistical disclosure control; factor analysis.

1. Introduction

The scientific community relies heavily on high quality data for the empirical validation of proposed theoretical models. However, data collection is an expensive and laborious task and thus it is prudent to use data which have already been collected by others, albeit for different reasons. Public administrations, governmental agencies and other state institutions gather valuable information on all aspects of society and there are huge benefits to be gained from broad access to these data. The crucial point is how to grant this access without violating the confidentiality guarantees given to survey respondents. Most microdata sets are collected under a pledge of confidentiality and therefore cannot be released unrestrictedly. Statistical analyses via remote access seem to offer both preservation of confidentiality and unlimited use of data. In a remote access system as we define it, the analyst uses his or her desktop computer to connect to a server on which the confidential microdata are stored. He or she can submit any query to the server, which runs the requested analysis of the microdata and returns the results to the user if the requested

[1] Institute for Employment Research, Statistical Methods, Regensburger Str. 104, Nuremberg 90478, Germany. Email: joerg.drechsler@iab.de
[2] University of Tuebingen, Mohlstraße 36, 72074 Tuebingen, Germany. Email: gerd.ronning@uni-tuebingen.de
[3] GfKSE, Nuremberg, Germany.
Acknowledgments: This research was partially supported by the "InfinitE" project funded by the German Federal Ministry of Education and Research. We thank the three referees for their valuable comments, which helped to improve the quality of the article.

output does not violate any confidentiality restrictions. The microdata never leave the secure environment of the server. However, to guarantee that the provided output does not reveal any confidential information, the list of allowed queries is generally limited in practice. The remote access solutions that have been implemented so far either define a list of queries that are not allowed (any command that is not on the list can be requested) or explicitly state which queries can be submitted.

An example of the first approach is the system implemented at the Cross-National Data Center in Luxembourg, known as LISSY (Cross-National Data Center in Luxembourg 2012a), which accepts code written for the software packages SAS, STATA or SPSS. Jobs can be submitted either per e-mail or via a job submission interface. The system does not restrict the list of allowed queries in advance. Instead, "certain syntax and comments will trigger system security alerts" (Cross-National Data Center in Luxembourg 2012b), which may terminate the job. The system will only return ASCII output, that is, no graphical output of any form will be provided. A more advanced version of the approach is also planned in the U.S. (Lucero et al. 2011).

An example of the second approach is implemented at the National Center for Health Statistics (NCHS) (Research Data Center of the National Center for Health Statistics 2012a). The NCHS system, which is called ANDRE, only accepts code written for the software packages SAS or SUDAAN. Other software packages, such as SPSS or R, can only be used on-site. Furthermore, the list of possible procedures and options is limited in advance and some procedures will automatically be adapted to avoid disclosure (Research Data Center of the National Center for Health Statistics 2012b). Finally, the website states that "[o]utput results that pose a disclosure risk will be suppressed" without any further information as to how such an output is identified. This kind of approach has also been implemented in Australia (O'Keefe and Good 2008). The Australian Bureau of Statistics provides an online tool called TableBuilder "which enables users to create tables, graphs and maps of Census data" (http://www.abs.gov.au/websitedbs/censushome.nsf/home/tablebuilder). Another online tool called DataAnalyser which additionally allows running a number of standard regression models will also be implemented soon (http://www.abs.gov.au/websitedbs/D3310114.nsf/home/About+DataAnalyser).

However, 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 data set 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 sufficient to obtain exact information on sensitive attributes in the data set. 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 counterstrategies. Restricted remote access following the second approach described above is an attempt to circumvent this dilemma by only

allowing computations that are considered safe under all circumstances. However, as a consequence, the set of allowed queries will be very limited and many users will find this set too restricted 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. However, 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 to a minimum. This is an ambitious goal and it is not clear whether this goal can ever be achieved.

While the risks from releasing microlevel information of the original data are obvious, it is less obvious that microlevel information is a byproduct of several data analysis tools and that this byproduct might pose a risk although the final output of interest might not be problematic. Regression procedures provide microlevel output such as fitted values or residuals, and model-fitting checks, such as Q-Q plots or Cook's distance, provide information on the individual level at least for the outliers (arguably the most interesting individuals for an intruder). Although at first sight it seems impossible to learn anything about the reported microdata values from these diagnostic plots, Sparks et al. (2008) illustrate the risks that might result if these analytics tools are provided in a remote access system without further restrictions. For this reason, the remote access system that is planned for the U.S. will, for example, provide Q-Q plots that are based on synthetic data. Sparks et al. (2008) also suggest a number of additional protective measures that can be taken to avoid these kinds of disclosures and argue that no information on the individual level should be released in general. To our knowledge, all agencies that have implemented a remote access environment so far have followed this advice.

In this article we provide another example of why monitoring the output of any analysis and suppressing all microlevel information is generally a good strategy. Factor analysis is very popular in the social sciences since it can be applied in a wide range of explorative and confirmatory tasks and it would be a severe drawback of remote access if this kind of analysis was not possible. On the other hand, as we will illustrate in this article, there is a risk of disclosure if unrestricted factor analysis is allowed. However, this risk can easily be avoided if the individual factor scores are not revealed to the analyst. Since researchers will usually only be interested in the factor loadings for the different variables included in the model, we do not see any disadvantages in not providing the individual factor scores. If information on the individual factors is considered necessary, graphical displays of the winsorised data could be provided akin to the disclosure prevention measures described in Sparks et al. (2008).

The remainder of the article is organized as follows: Following a brief description of factor analysis methods, 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 set of variables could be identified in the data set that is uncorrelated with the variable to be disclosed, henceforth called the variable of interest. 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 data set 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 (IAB) in Germany. The article concludes with some final remarks.

2. Some Basic Facts on Factor Analysis

Factor analysis and the closely related method of principal components are widely used in all fields of social science, in particular in psychology and sociology where "latent" variables, such as ability and satisfaction, are modelled frequently. More recently, the method has also been employed in modern time series analysis when factor-augmented vector autoregression models (FAVAR) are considered (see, for example, Stock and Watson 2002). The aim of the approach is to reduce the empirical information from a large set of continuous variables to a small set of (latent) factors. In the following we describe the basic concept briefly. A detailed description can be found in any standard textbook on the topic (see, for example, Press 2005).

Consider a set of m random variables $\eta = (\eta_1, \eta_2 \ldots, \eta_m)'$ with

$$E[\eta] = \mu_\eta, cov[\eta] = \Sigma_{\eta\eta}$$

for which n observations are available leading to the $(n \times m)$ data matrix $Y = (y_1, y_2, \ldots, y_m)$. 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

$$\eta - \mu_\eta = \Lambda \mathbf{f} + \mathbf{u}, \tag{1}$$

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

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

Since the factors are assumed to be orthogonal with $cov[\mathbf{f}] = I_p$, where I_p is the identity matrix of dimension $(p \times p)$, as well as independent of $\mathbf{u}$, we obtain what is called the "fundamental equation"

$$\Sigma_{\eta\eta} = \Lambda\Lambda' + \Psi.$$

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

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

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

$$M = \iota_n \otimes \mu_\eta'.$$

Here ι_n is an n-vector of ones and $\otimes$ denotes the Kronecker product, that is, the corresponding mean from the vector μ_η is subtracted from each observation in Y in (2).

We will call (2) the "empirical factor model", whereas (1) will be called the "theoretical factor model".

If the estimated matrix Λ has a block-diagonal structure, particular factors can be related to a subset of the vector η, which helps to interpret these factors. However, it is well known that this estimated matrix is not unique: Take any $(m \times m)$ orthogonal matrix W and it will by definition satisfy $WW' = I_m$. Keeping this in mind, we can rewrite (1) as

$$\eta - \mu_\eta = (\Lambda W)(W'\mathbf{f}) + \mathbf{u},$$

where $\Lambda^* = \Lambda W$ would represent the factor-loading matrix and $\mathbf{f}^* = W'\mathbf{f}$ the vector of factors. The multiplication of the factor-loading matrix by any orthogonal matrix is called rotation of this matrix. Usually, the matrix W is chosen such that for each factor the loading on a subset of variables is as large as possible and the loading on the remaining variables is as small as possible, so that a "simple structure" is obtained which facilitates the interpretation of factors. One way to achieve this is to find the orthogonal matrix that maximizes the variance of the squared factor loadings. This is the well-known varimax criterion (see, for example, Press 2005 Ch. 10.6 for details).

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 Λ is known or rather has been estimated in an earlier step indicated by the symbol $\sim$ placed above the relevant quantities. Hence, the resulting estimates of $\mathbf{f}$ depend on the method by which the factor-loading matrix was determined. In all cases $\tilde{\Lambda}$ may represent either the original or the rotated factor-loadings. We will only present the results for the empirical model (2) as 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 empirical factor model (2) can be seen as a regression model with unknown matrix F which can be estimated by least squares. The resulting estimator is

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

Note that the transpose of $\hat{F}_{LS}$ is just the standard OLS estimate from linear regression. Horst (1965) seems to have been one of the first to use this approach (McDonald and Burr 1967, p. 386).

3.2. Bartlett's Method

Considering the nonscalar structure of the covariance matrix Ψ, a generalized least squares formula seems more appropriate:

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

Note that in this case the matrix Ψ also has to be determined in advance. This method has been proposed by Bartlett (1937). Fahrmeir et al. (1996, pp. 648, 690) remark that (4) can be regarded as a maximum likelihood estimator when normality for η is assumed. Non-normally distributed variables in η lead to quasi-maximum likelihood estimation of loadings and scores, still being asymptotically normally distributed and consistent.

3.3. Thomson's Method

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

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

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:

$$\hat{F}_{TH} = (Y - M)\left(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}\right)^{-1}\hat{\Lambda}. \tag{5}$$

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

$$\Sigma_{\eta\eta} = Q\Theta Q',$$

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

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

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

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

For more details see any textbook on multivariate analysis, such as, Press (2005).

4. Disclosure Risk from Factor Analysis

In this section we will illustrate scenarios in which the factor scores disclose sensitive information. We show analytically that a severe risk of disclosure exists if at least one variable can be identified in the data set that is (almost) uncorrelated with the variable of interest. As we show later in the empirical example (Subsection 5.3), potential variables can be selected by inspecting the correlation matrix.

For concreteness, let us assume that η_1 is the variable of interest so that the covariance matrix has the following block diagonal structure:

$$\Sigma_{\eta\eta} = \begin{pmatrix} \sigma_{11} & 0' \\ 0 & \Sigma_{22} \end{pmatrix} \tag{7}$$

where Σ_{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 one factor "loading" only on the first variable and the remaining $p-1$ factors having zero loading weight on this variable. Note that this implies

$$(\Lambda'\Lambda)^{-1} = \begin{pmatrix} 1 & 0' \\ 0 & (\Lambda_2'\Lambda_2)^{-1} \end{pmatrix} \tag{8}$$

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

$$F_{LS} = F_{BA} = \left(1 \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right).$$

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 $j = 2, \ldots, p$ 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):

$$F_{TH} = \left(\frac{1}{1+\psi_1} \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right).$$

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

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

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

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.

As a final remark, we try to shed some light on the question of what influence the $m-1$ "remaining" variables in the factor model have on the accuracy of the results. Most importantly, whenever at least one variable highly correlated with the variable of interest is included in the factor model, there will be no eigenvector loading on the variable of interest alone and no disclosure will be possible. Clearly, if the correlation with the variable of interest is exactly zero for all variables included in the set of variables in m, the theoretical results above imply exact reproduction of the vector $\mathbf{y}_1$ no matter how many additional variables are included in the set of variables in m. In this case, one variable would be sufficient and adding variables that are (even slightly) correlated with the variable of interest will decrease the level of accuracy. In practice, the correlation is never exactly zero, as illustrated in Table 1 from our empirical example in Section 5. However, it would still make sense in terms of prediction accuracy to only pick the variable with the lowest correlation with the variable of interest. Nevertheless, it might generally be advantageous from the perspective of an intruder to include some additional variables in the model to avoid submitting queries that look overly suspicious. In this case it would be the best strategy to pick a predefined set of variables, say eight to ten, consisting of those variables with the lowest empirical correlation with the variable of interest. This is the strategy we follow in our empirical evaluations in the next section.

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 a range of background information (e.g., regional information, industrial sector). For further information see Fischer et al. (2009) 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 is only granted to external researchers at the IAB's research data center (RDC). The RDC, which was established in 2004, provides researchers with access to microdata for noncommercial empirical research in the fields of social security and employment. Most of the surveys conducted at the IAB and samples from the administrative data of the Federal Employment Agency are available for on-site analysis (see Heining 2009 or http://fdz.iab.de for further details).

Researchers can also submit queries to the RDC that are run on the original data by the staff of the RDC (remote execution). In this case the results are reported back to the researchers only after the output has been carefully checked for confidentiality violations (if the researcher analyzes the data onsite, only the results that are intended to be used outside of the rooms of the RDC will be checked). Finally, some surveys are also available as scientific use files (unlike public use files, scientific use files are only available to the scientific community). Currently, all confidentiality checks are performed manually, so the attack described in this article would be detected. Nevertheless, as remote access is seen to be the future for data providers, we use the data set to illustrate that unrestricted factor analysis in a remote access setting would be problematic in terms of disclosure risk.

For our empirical evaluations we use the cross-section from the year 2007 of the survey. All missing values in this data set are replaced by single imputation and treated as 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, that is, the revenue. Thus, we exclude all establishments that do not report turnover, such as nonindustrial organizations, regional and local authorities and administrations, financial institutions and insurance companies. The remaining data set includes 12,814 fully 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

$$lgturn_i = \log(turnover_i + 1), \tag{9}$$

where $turnover_i$ is the turnover in euros for establishment i. The 1 is added to ensure that all values are strictly positive before the log transformation, because some establishments report a turnover of zero. 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 disclosure attack outlined above, we need to identify variables that are (almost) uncorrelated with this variable. It should not be difficult for an intruder to obtain this information because correlation matrices are not usually considered

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 IT (inv.)	0.0587
Total number of civil servant aspirants (asp.)	0.0082
Total number of vacant positions for workers (vac.w.1)	0.0536
Number of vacancies for workers reported to employment agency (vac.w.2)	0.0374
Number of vacancies for qualified employees reported to employment agency (vac.em.)	0.1193
Employees with wage subsidies (sub.)	0.0984
Employees over 50 with wage subsidies (sub.50)	0.0513

to provide a high risk of disclosure. Table 1 lists the eight variables that we use 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 estimated almost exactly with this scenario.

Usually, factor analysis starts by inspecting the eigenvalues of the covariance matrix or correlation matrix to determine the number of factors p to be used in the model. Only the largest eigenvalues are selected with the understanding that the variability of Y is sufficiently explained by this subset. Based on the correlation matrix both the Kaiser criterion (Kaiser 1958) and the scree test (Fahrmeir et al. 1996) would suggest selecting $p = 4$ for our set of variables. However, inspecting the eigenvalues is not helpful in our setting since we need to make sure that the factor that loads on the variable of interest alone is also included in the model. As noted earlier, it can be shown that the relevant eigenvalue need not be one of the largest eigenvalues (see Ronning and Bleninger 2011 for more details). Therefore, the intruder should choose a large $p \leq m$ and examine all estimated factors. Alternatively, he or she could simply try alternative values of p. We found the ideal number of factors by evaluating the full range of possible factors. The loading matrix for $p = 4$ (after rotation based on the varimax criterion) is presented in Table 2 and it is obvious that in this case the third factor loads primarily on turnover and thus this factor model is ideally suited for a disclosure attack.

Table 2. Rotated Matrix $\tilde{\Lambda}$ 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

5.3. Estimation of Factor Scores

In the next step, we estimate the matrix of factor scores $\hat{F}$ based on the rotated loadings from Table 2. For purpose of brevity, we limit our evaluation to Bartlett's (4) and Thomson's (5) solution. We note that the least squares solution and principal component analysis will provide similar results. Once we have 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

$$\hat{turn}_i = exp\left\{\widehat{lgturn}_i\right\} - 1.$$

We note that the transformation will lead to a small bias in the estimated turnover since in general $E(\log(y_i)) \neq \log(E(y_i))$. 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

$$\delta_i = \frac{\hat{turn}_i - turnover_i}{turnover_i}, \quad i = 1, \ldots, n.$$

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 sorted in ascending order based on the number of employees.

Looking at the scatter plots, we find that using Bartlett's method the estimated turnover is very close to the true turnover for almost all establishments. The relative difference δ is less than 0.5% for 99.3% of the establishments.

For Thomson's method, we notice that the relative differences are generally larger than for Bartlett's method (note that the scale of δ differs between the scatter plot for Thomson's method (middle panel) and the scatter plots for Barlett's method (left panel) and Thomson's method after correction (right panel)). More than 40% of the estimated

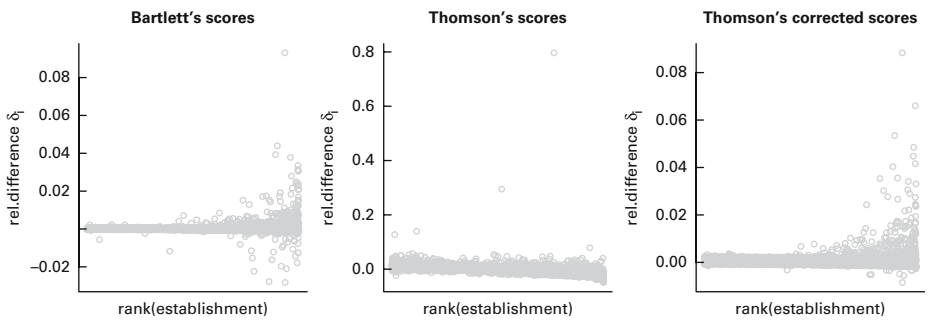

Fig. 1. Relative differences δ_i between the true turnover and the turnover estimated from the factor scores obtained through Bartlett's method and Thomson's method with/without correction. Establishments are sorted in ascending order based on the number of employees.

turnovers differ by more than 1% from the true value and the difference can be up to 80%. We also find a trend in the relative differences. The turnover derived from the factor scores overestimates the true turnover for the smallest establishments. This effect is decreases continuously 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 transformed turnover according to (9) and $\hat{y}_1$ is its estimate based on Thomson's method without correction, the difference between the two quantities is given by

$$\hat{y}_1 - y_1 = \frac{\psi_1}{1+\psi_1} \begin{pmatrix} \mu_1 - y_{11} \\ \vdots \\ \mu_1 - y_{1n} \end{pmatrix}, \tag{10}$$

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 negative trend for the relative difference when going from the smallest establishments to the largest. 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 99.0% of the establishments, having a relative difference of less than $\pm 0.5\%$. In fact, the estimated turnover never differs by more than $\pm 8.9\%$ from the true turnover. Thus the risk of disclosure is comparable to the risk when Bartlett's method is applied.

6. Conclusions

There is an increasing demand among 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 article we have 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 individual microlevel values 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 the method used to estimate the factor scores. Of course, zero correlation is unrealistic in practice but our empirical example illustrates that a very close approximation of the microlevel 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 by applying the procedure outlined in this article, the intruder will only obtain a full vector of estimated microlevel values. Even if these estimates are very close to the true values, this will not necessarily lead to disclosure if the intruder is not able to link this information to individual units in other databases.

Still, most legislation requires that no individual information be released to the public, no matter whether a direct link is possible or not. Furthermore, it is often easy to attribute some of the obtained values to specific units, such as the largest turnover in the data set, for example.

Finally, we wish to stress that it is not the aim of this article to call for more restrictive data access. Factor analysis is a useful and widely used method that should be available to researchers in a remote access system. We only wish to raise awareness of the fact that this kind of attack is possible if no countermeasures are taken. Once identified, these attacks can be prevented easily by not reporting individual factor scores, since applied analysts are not usually interested in these scores. Following Brandt et al. (2010), who provided general guidelines for output checking when data are disseminated, the factor loadings of the different variables can be considered "safe" outputs that can be released without restrictions. The individual factor scores, on the other hand, should be classified as "unsafe", and extra measures are necessary if these scores are to be provided. Simply checking the correlation between the factor scores and the variables in the data set, for example, could be a useful tool for avoiding disclosure. The factor scores can be suppressed if the bivariate correlation with any variable in the data set is higher than an agency-defined threshold, say 0.995. Alternatively, preventive measures, such as providing only graphical displays of the winsorised factor scores or other measures akin to the measures suggested by Sparks et al. (2008), could be implemented. Finally, as suggested by one of the referees, output perturbation could also be applied. As the name indicates, this approach guarantees confidentiality by only perturbing the output of the queries; the underlying microdata are not altered. This approach has been discussed for other query types such as survival analysis (see, for example O'Keefe et al. 2012) and the original setup for ε-differential privacy (Dwork 2006) was also developed around this idea. Identifying the best perturbation approach when providing individual factor scores would be an area for future research. The aim of this article was more generally 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.

Appendix. Derivations of the Factor Scores if One Variable is Uncorrelated With the Other Variables in the Model

The Least Squares Solution

$$F_{LS} = (Y - M)\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}\begin{pmatrix} 1 & 0' \\ 0 & (\Lambda_2'\Lambda_2)^{-1} \end{pmatrix} = (Y - M)\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2(\Lambda_2'\Lambda_2)^{-1} \end{pmatrix}$$

$$= \left(1 \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right)$$

Bartlett's Method

$$\hat{F}_{BA} = (Y - M)\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda)^{-1}$$

$$= (Y - M)\begin{pmatrix} \psi_1^{-1} & \\ & \Psi_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}\left(\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}'\begin{pmatrix} \psi_1^{-1} & \\ & \Psi_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}\right)^{-1}$$

$$= (Y - M)\begin{pmatrix} \psi_1^{-1} & 0' \\ 0 & \Psi_2^{-1}\Lambda_2 \end{pmatrix}\left(\begin{pmatrix} \psi_1^{-1} & 0' \\ 0 & \Lambda_2'\Psi_2^{-1}\Lambda_2 \end{pmatrix}\right)^{-1}$$

$$= (Y - M)\begin{pmatrix} 1 & 0' \\ 0 & \Psi_2^{-1}\Lambda_2\left(\Lambda_2'\Psi_2^{-1}\Lambda_2\right)^{-1} \end{pmatrix}$$

$$= \left(1 \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right)$$

The Solution of Thomson/Thurstone

$$F_{TH} = (Y - M)\left(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi}\right)^{-1}\hat{\Lambda}$$

$$= (Y - M)\left(\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2\Lambda_2' \end{pmatrix} + \begin{pmatrix} \psi_1 & \\ & \Psi_2 \end{pmatrix}\right)^{-1}\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}$$

$$= (Y - M)\begin{pmatrix} (1 + \psi_1)^{-1} & 0' \\ 0 & \left(\Lambda_2\Lambda_2' + \Psi_2\right)^{-1}\Lambda_2 \end{pmatrix}$$

$$= \left(\frac{1}{1 + \psi_1} \cdot \begin{pmatrix} y_{11} - \mu_1 \\ \vdots \\ y_{n1} - \mu_1 \end{pmatrix} \middle| \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_{p-1}, \mathbf{f}_p \right)$$

7. References

Bartlett, M. (1937). The Statistical Conception of Mental Factors. British Journal of Psychology, 28, 97–104. DOI: http://www.dx.doi.org/10.1111/j.2044-8295.1937.tb00863.x

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.

Brandt, M., Franconi, L., Guerke, C., Hundepool, A., Lucarelli, M., Mol, J., Ritchie, F., Seri, G., and Welpton, R. (2010). Guidelines for the Checking of Output Based on Microdata Research. Final report of ESSnet sub-group on output SDC.

Cross-National Data Center in Luxembourg (2012a). Available at: http://www.lisdatacenter.org (accessed January 17, 2014).

Cross-National Data Center in Luxembourg (2012b). Available at: http://www.lisdatacenter.org/data-access/lissy/best-practices/ (accessed January 17, 2014).

Drechsler, J. (2011). Multiple Imputation in Practice – a Case Study Using a Complex German Establishment Survey. Advances in Statistical Analysis, 95, 1–26. DOI: http://www.dx.doi.org/10.1007/s10182-010-0136-z

Dwork, C. (2006). Differential Privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming (ICALP), 1–12.

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

Fischer, G., Janik, F., Müller, D., and Schmucker, A. (2009). The IAB Establishment Panel – Things Users Should Know. Schmollers Jahrbuch – Journal of Applied Social Science Studies, 129, 133–148. DOI: http://www.dx.doi.org/10.3790/schm.129.1.133

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. DOI: http://www.dx.doi.org/10.1214/088342305000000043

Heining, J. (2009). The Research Data Centre of the German Federal Employment Agency: Data Supply and Demand Between 2004 and 2009. RatSWD working paper, 129.

Horst, P. (1965). Factor Analysis of Data Matrices. New York: Holt, Rinehart & Winston.

Kaiser, H. (1958). The Varimax Criterion for Analytic Rotation in Factor Analysis. Psychometrika, 23, 3, 187–200. DOI: http://www.dx.doi.org/10.1007/BF02289233

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

Lucero, J., Freiman, M., Singh, L., You, J., DePersio, M., and Zayatz, L. (2011). The Microdata Analysis System at the U.S. Census Bureau. SORT, Special Issue: Privacy in Statistical Databases, 77–98.

McDonald, R. and Burr, E. (1967). A Comparison of Four Methods for Constructing Factor Scores. Psychometrika, 32, 381–401. DOI: http://www.dx.doi.org/10.1007/BF02289653

O'Keefe, C., Sparks, R., McAullay, D., and Loong, B. (2012). Confidentialising Survival Analysis Output in a Remote Data Access System. Journal of Privacy and Confidentiality 4. Available at: http://repository.cmu.edu/jpc/vol4/iss1/6 (accessed January 17, 2014).

O'Keefe, C.M. and Good, N.M. (2008). A Remote Analysis Server – What Does Regression Output Look Like? In Privacy in Statistical Databases, J. Domingo-Ferrer and Y. Saygin (eds), vol 5262 of Lecture Notes in Computer Science. New York: Springer, 270–283.

Press, S. (2005). Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference, (2nd edn). New York: Dover Publications.

Research Data Center of the National Center for Health Statistics (2012a). Available at: http://www.cdc.gov/rdc/B2AccessMod/ACs230.htm (accessed January 17, 2014).

Research Data Center of the National Center for Health Statistics (2012b). Available at: http://www.cdc.gov/rdc/Data/B2/SASSUDAANRestrictions.pdf (accessed January 17, 2014).

Ronning, G. and Bleninger, P. (2011). Disclosure Risk From Factor Scores. Technical Report, IAW Discussion Papers 73. Available at: http://www.iaw.edu/w/IAWPDF.php?id=886&name=iaw_dp_73.pdf (accessed January 17, 2014).

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C., Duncan, J., Keighley, T., and McAullay, D. (2008). Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-preserving Analytics. Comput Methods Programs Biomed, 91, 208–222. DOI: http://www.dx.doi.org/10.1016/j.cmpb.2008.04.001

Stock, J. and Watson, M. (2002). Forecasting Using Principal Components From a Large Number of Predictors. Journal of the American Statistical Association, 97, 1167–1179. DOI: http://www.dx.doi.org/10.1198/016214502388618960

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

Thurstone, L. (1935). The Vectors of Mind. Chicago: University of Chicago Press.

Received May 2012
Revised April 2013
Accepted September 2013