Institute for Employment Research

The Research Institute of the Federal Employment Agency

Disclosure Risk from Factor Scores in a Remote Access Environment

26 October 2011, Tarragona (Spain)

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality

Philipp Bleninger
Jörg Drechsler
Gerd Ronning

Motivation

- Remote data access
  - gives access from a desktop machine to a host server
  - with anonymised data on-screen but analyses run on original data
  - outputs (automatically) checked for violations
- Disclosure risk
  - compromises confidentiality even via indirect access to the actual data
  - through inferential statistical disclosure
- To date focus on linear regression, e.g. Reznek (2003), Reznek/Riggs (2004, 2005), Gomatam et al. (2005), Bleninger et al. (2011), Vogel (2011)
- At present application of factor analysis, e.g. Ronning et al. (2010), Ronning/Bleninger (2011)

Factor Analysis

- Task: Dimension reduction and feature extraction

→ Reduction of m observed variables Y to $p < m$ unobserved latent common factors F

- Factor model: $Y - M = F\Lambda' + U$
- Fundamental equation: $\Sigma_Y = \Lambda\Lambda' + \Psi$
- $(n \times m)$ mean matrix $M = \iota_n \otimes \mu'_Y$

$(n \times m)$ data matrix

$$Y = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1m} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nm} \end{pmatrix}$$

$(m \times p)$ factor loading matrix

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

$(n \times p)$ common factor matrix

$$F = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1p} \\ \vdots & \vdots & & \vdots \\ f_{n1} & f_{n2} & \dots & f_{np} \end{pmatrix}$$

$(m \times m)$ uniqueness matrix

$$cov[U] = \Psi = \begin{pmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_m \end{pmatrix}$$

Factor Analysis

- Factor loadings Λ
  - Evaluation of the optimal number p^* of factors using scree plots or tests of sphericity
  - Estimation of Λ according to the fundamental equation
  - Rotation of Λ optimizing interpretability with regards to contents, using different criteria preserving orthogonality (e.g. varimax) or not (e.g. oblimin)
- Factor scores
  - Estimation of F from Principal Component Analysis
  - Estimation of F from the factor model using the least squares approach, Bartlett's method or Thomson's/Thurstone's method

Setting up Disclosing Factor Analysis

- Use the variable Y_1 of interest and any additional variables $Y_2, \ldots, Y_m$ which are uncorrelated to the former resulting in the covariance matrix

$$\Sigma_{\boldsymbol{Y}} = \begin{pmatrix} \sigma_{11} & 0 & 0 & \dots & 0 \\ 0 & \sigma_{22} & \sigma_{23} & \dots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \sigma_{m2} & \sigma_{m3} & \dots & \sigma_{mp} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0' \\ 0 & \Sigma_2 \end{pmatrix}$$

- Extract an arbitrary number of factors. The variable of interest has to get its own factor which is highly connected only to it and is not connected to the other variables.

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_{22} & \lambda_{23} & \dots & \lambda_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \lambda_{m2} & \lambda_{m3} & \dots & \lambda_{mp} \end{pmatrix} = \begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}$$

Setting up Disclosing Factor Analysis

- Orthogonality of the factors guarantees the unique relationship between the target variable and its single factor.
- The factor's scores $f_{11}, ..., f_{n1}$ serve as disclosing approximations of the real values of the variable of interest:

$$y_{i1} - \mu_1 = \sum_{k=1}^{p} \lambda_{1k} f_{ik} + u_{i1} = f_{i1} \quad , i = 1, \ldots, n$$

→ with the assumptions of F_1 being the single factor of Y_1 and following of $\lambda_{11} = 1$, $\lambda_{1k} = 0$ $(k = 2, ..., p)$ and $u_{i1} = 0$

- Estimations $\hat{f}_{i1}$ of the factor scores are not only unbiased and consistent estimations of f_{i1} but also accurate estimations of the real values y_{i1}.

Disclosure Risk from Factor Scores

- Estimation of factor scores by Bartlett's method

$$
\begin{aligned}
\hat{F} &= (Y-M)\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda)^{-1} \\
&= (Y-M)\begin{pmatrix} \psi_1^{-1} & \\ & \Psi_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}\left(\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}'\begin{pmatrix} \psi_1^{-1} & \\ & \Psi_2^{-1} \end{pmatrix}\begin{pmatrix} 1 & 0' \\ 0 & \Lambda_2 \end{pmatrix}\right) \\
&= (Y-M)\begin{pmatrix} \psi_1^{-1} & 0' \\ 0 & \Psi_2^{-1}\Lambda_2 \end{pmatrix}\left(\begin{pmatrix} \psi_1^{-1} & 0' \\ 0 & \Lambda_2'\Psi_2^{-1}\Lambda_2 \end{pmatrix}\right)^{-1} \\
&= (Y-M)\left(\begin{pmatrix} 1 & 0' \\ 0 & \Psi_2^{-1}\Lambda_2\left(\Lambda_2'\Psi_2^{-1}\Lambda_2\right)^{-1} \end{pmatrix}\right) \\
&= \left(\begin{array}{c|ccc} \hat{f}_1 = \begin{pmatrix} y_{11}-\mu_1 \\ \vdots \\ y_{n1}-\mu_1 \end{pmatrix} & \hat{f}_2, & \dots, & \hat{f}_p \end{array}\right)
\end{aligned}
$$

Data

- IAB Establishment Panel
  - Nationwide survey of establishments
  - Establishments with at least one employee covered by social security
  - Sampled from the German Social Security Data, stratified by the number of employees, the region and the branches
  - Cross section: panel wave from 2007
  - → 12,814 completely observed establishments
- Sensitive information of interest Y_1
  - Turnover from sales (in Euro, excluding taxes)
  - → transformed by the logarithm

$$y_{i1} = lgturn_i = \log(turnover_i + 1)$$

Setting up the Data Intrusion

$Y =$

- Logarithmic turnover (lgturn.) from sales
- Investments in EDP, information and communication technology (inv.)
- Number of civil servant aspirants (asp.)
- Number of vacant positions for workers (vac.w.1)
- Number of vacancies notified to employment office for workers (vac.w.2)
- Number of vacancies notified to employment office for qualified employees (vac.em.)
- Number of employees with wage subsidies (sub.)
- Number of employees older than 50 with wage subsidies (sub.50)

	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.		1.0000	-0.0075	0.0057	0.0083	0.0440	0.0020	0.0111
asp.			1.0000	-0.0003	-0.0004	-0.0011	0.0015	0.0045
vac.w.1				1.0000	0.9249	0.0925	0.0285	0.0199
vac.w.2					1.0000	0.0905	0.0222	0.0160
vac.em.						1.0000	0.0641	0.0853
sub.							1.0000	0.7901
sub.50								1.0000

Table: Empirical correlation matrix R

Estimating Disclosing Loadings

- The target variable lgturn is put in the first column of Y.
- An adequate number of factors is extracted to obtain a single factor for the target variable.
- Λ is estimated with Maximum Likelihood by Newton maximization using starting values of Jöreskog.
- Λ is rotated orthogonally with the Varimax-criterion

	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

Table: Rotated Matrix $\hat{\Lambda}$ of estimated loadings

Estimating Disclosing Scores

- Estimation of the disclosing factor scores $\hat{f}_{i1}$ from the factor model using Bartlett's method.
- Re-standardisation of the scores with μ_1 being estimated from the data and u_{i1} being supposed to be 0 after checking ψ_1.
- Re-computation of the estimated turnovers $\hat{y}_{i1}$

$$turn\hat{o}ver_i = exp\left\{lg\hat{t}urn_i\right\} - 1\ .$$

- Assessment of the disclosure risk

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

Disclosure Risk from Factor Scores

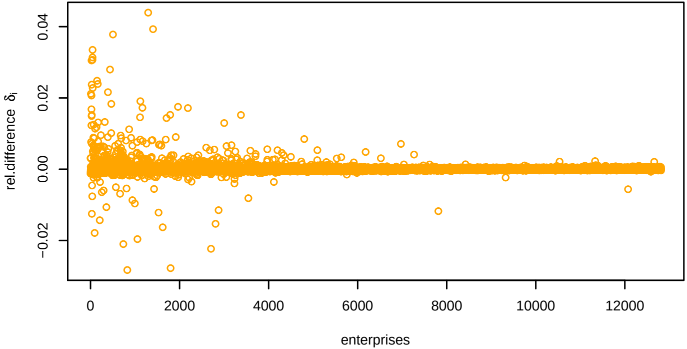

Figure: Scatter plot of relative differences δ_i for Bartlett's factor scores

Conclusion

- Disclosure from factor scores works very well
  - though log-transformation and exp-retransformation of the target variable may lead to biased expectations.
  - ignoring the regular extraction of the number of common factors according to eigenvalues.
  - despite the non-recurrence from factor space back to variable space, especially if variables were transformed before analysis.
- Still there are many standard analyses to contain the risk of disclosure, e.g. Cluster Analysis.

Thank you for your attention

Philipp Bleninger
Jörg Drechsler
Gerd Ronning
Contact: philipp.bleninger@iab.de