STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 111–132

SPARSE METHODS FOR ANALYSIS OF SPARSE MULTIVARIATE DATA FROM BIG ECONOMIC DATABASES

Daniel Kosiorowski[1], Dominik Mielczarek[2], Jerzy Rydlewski[2], Małgorzata Snarska[3]

ABSTRACT

In this paper we present a novel perspective dedicated for *sparse* high-dimensional *data* sets, i.e. data which contain many zeros among coordinates of observations. Using jointly, selected *sparse methods* recently proposed in multivariate statistics, and kernel density framework for discrete data, we outline a general perspective for bringing out useful information from big economic databases. As a framework for our considerations we take the so-called functional data analysis, which originates from Ramsay and Silverman works. In particular we use functional principal components analysis within 2D density estimation procedure proposed by Simonoff.

Key words: sparse data, sparse methods, robust methods, categorical data, big data.

1. Introduction

In recent years several authors have investigated the use of smoothing methods for sparse multinomial data. In his excellent paper Simonoff (1983) considered probabilities in a large one-dimensional sparse contingency table estimated by maximizing the likelihood modified by a roughness penalty. It was shown in his paper that if certain smoothness criteria on the underlying probability vector are fulfilled, the maximum penalty estimator is consistent in a one-dimensional table under a sparse asymptotic framework. However, a proof of sparse asymptotic consistency for multidimensional tables was not found. It was shown that the bias of kernel estimates of probabilities for cells near the

[1] Department of Statistics, Faculty of Management, Cracow University of Economics. E-mail: daniel.kosiorowski@uek.krakow.pl.

[2] Department of Differential Equations, Faculty of Applied Mathematics, AGH University of Science and Technology. E-mail: dmielcza@wms.mat.agh.edu.pl.

[3] Department of Capital Markets, Faculty of Finance, Cracow University of Economics. E-mail: snarskam@uek.krakow.pl.

boundaries of the multinomial vector often dominates the mean sum of the squared error of the estimator. However, boundary kernels contrived to correct boundary effects for kernel regression estimators can achieve the same result for these estimators. Dong and Simonoff (1994) investigated the properties of estimators based on boundary kernels and compared them to unmodified kernel estimates and maximum penalized kernel likelihood estimates. They showed that the boundary-corrected estimates usually outperform uncorrected kernel estimates and are quite competitive with penalized likelihood estimates. Shane and Simonoff (2001) considered categorical data analysis using maximum likelihood. The problem with maximum likelihood estimates is their sensitivity to outlier cells. For this reason robust alternatives to maximum likelihood estimation were proposed in Shane and Simonoff (2001). The methods include the least median of chi-squared residuals, the least median of weighted squared residuals, and methods using the least trimmed functions. They also considered equivariance and breakdown properties of the estimators. They showed that the maximum likelihood estimates break down in the presence of outlying cells, while robust estimators do not as long as the contamination point does not exceed the breakdown point. Simonoff (1998) focused on nonparametric estimation of smooth functions. He considered categorical data smoothing and constructed effective categorical likelihood smoothing estimates. He also used an appropriate likelihood function yielding cell probability estimates with many desirable properties. Such estimates can be used to construct well-behaved density estimates using local or penalized likelihood estimation. Simonoff (1998) showed advantage of the local polynomial likelihood density estimate over the penalized likelihood density estimate. Namely, it is the structure which can be manipulated to allow local variation in the amount of smoothing.

In this paper we consider the estimator of the bivariate density function proposed in Simonoff (1988) and its modifications in the context of data mining in huge economic databases which may contain outliers.

2. Estimator of two-dimensional density function

Models using categorical data usually assume that there is no relation between adjacent cells. This is not the case for continuous distributions, where many estimation procedures are based on the fact that observations falling near the approximation site do give some information about the function we are trying to estimate, whether this is a density or a regression function. This information by proximity is at the base of the modifications that have been proposed to the histogram. The classical kernel or local polynomial estimators are, in fact, clever ways to use this idea to improve upon rough estimates. This idea has been used to smooth over discrete distributions, with increased interest when few observations are available when compared with the number of cells of the underlying distribution, or when the observations tend to concentrate too much in a few cells of the support, indicating that the underlying distribution is quite peaked. Smoothing over adjacent cells does contribute to improve estimators in the similar cases. For one-dimensional distributions Simonoff (1983), Hall and Titterington (1987) smoothed the histogram with a uniform-like distribution, and Burman (1987) discretized the kernel estimator. More recently Simonoff (1995, 1996), Dong and Simonoff (1995) or Aerts et al. (1997) studied discrete versions of local polynomial estimators for higher dimensional data. Jacob and Oliveira (2011) used the local polynomial approach but with respect to a relativized L_2 - error, showing good performance for one-dimensional data. The extension of these methods to higher dimensional data introduces some difficulties.

Assume we consider objects with respect to (w.r.t.) two variables X_1 and X_2, and our aim is to estimate their joint probability density function. Our starting point is the estimator proposed in Simonoff (1995), which is based on binning the data and dedicated to sparse continuous data. Simonoff proposes to divide the range of X_1 into n_1 bins, the *i*-th bin being called I_{1i} , and to divide the range of X_2 into n_2 , the *j*-th bin being called I_{2j} .

Table 1. Illustration for binning 2D continuous data

X_1 / X_2	I_{21}	I_{22}	$\cdots$	I_{1k_2}	**total**
I_{11}	n_1	n_{k_1+1}	$\cdots$		
I_{12}	n_2	n_{k_1+2}	$\cdots$		
$\vdots$					
I_{1k_1}	n_{k_1}	n_{2k_1}	$\cdots$	$n_{k_2 \cdot k_1}$	
total					

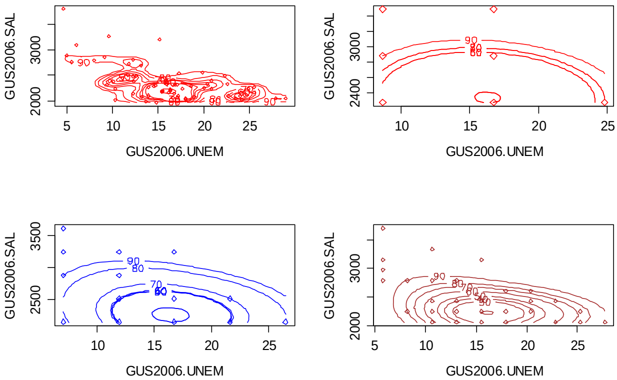

Figure 1. 2D kernel density estimates for binned data, unemployment vs. mean salary in Polish subregions in 2006

Next, let us consider $f_{2|1}(x_2 \mid x_1 \in I_{1i})$, the conditional density of x_2 given $x_1 \in I_{1i}$, $f_{1|2}(x_1 \mid x_2 \in I_{2j})$, the conditional density of x_1 given $x_2 \in I_{2j}$, the marginal densities of x_1 and x_2 to be $f_1(x_1)$ and $f_2(x_2)$. Integrating the conditional densities over the appropriate bins gives conditional probabilities:

$$P(x_1 \in I_{1i} \mid x_2 \in I_{2j}) = \int_{I_{1i}} f_{1|2}(u \mid x_2 \in I_{2j})du\,, \tag{2.1}$$

$$P(x_2 \in I_{2j} \mid x_1 \in I_{1i}) = \int_{I_{2j}} f_{2|1}(v \mid x_1 \in I_{1i})dv\,. \tag{2.2}$$

Simonoff proposes to estimate the conditional probabilities by treating each row and each column as **one-dimensional multinomial vector,** and then smooth them using **the penalized likelihood method proposed** by Simonoff (1983). The marginal probabilities were estimated using the marginal frequency estimates. He shows that when the number of rows $n_1 \to \infty$, and the number of columns $n_2 \to \infty$, then his estimator is a sparse asymptotic consistent one. For estimating the continuous density $f(x_1, x_2)$ we use an analogous technique.

Substituting into

$$f(x_1, x_2) = \left[f_{2|1}(x_2 \mid x_1) f_1(x_1) f_{1|2}(x_1 \mid x_2) f_2(x_2) \right]^{1/2}, \tag{2.3}$$

the kernel estimates of the conditional and marginal densities we obtain the 2D density estimate.

It is possible to generalize the estimator proposed by Simonoff for the multidimensional case. The main advantages of this estimator are relative computational simplicity in comparison to direct estimation of the multidimensional density, the effect of avoiding outlying cell propagation on the whole density estimate and its elasticity related to marginal and conditional density estimation method.

Further, we use a kernel density estimator for discrete data. Let us revise some basic notions related to this idea. Consider the estimation of a probability function defined for $X_i \in \mathrm{S} = \{0,1,...,c\text{-}1\}$.

The kernel estimator of $p(x)$

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} l(X_i, x)\ , \tag{2.4}$$

where $l(\cdot)$ is a kernel function defined by, say,

$$l(X_i, x) = \begin{cases} 1-\lambda & X_i = x \\ \lambda/(c-1) & otherwise \end{cases}, \tag{2.5}$$

and where $\lambda \in [0,(c-1)/c]$ is a “smoothing parameter” or “bandwidth”. It is easy to show

$$E\hat{p}(x) = p(x) + \lambda \left\{ \frac{1-cp(x)}{c-1} \right\}, \text{ var } \hat{p}(x) = \frac{p(x)(1-p(x))}{n} \left(1 - \lambda \frac{c}{(c-1)} \right)^2 .$$

This estimator was proposed by Aitchinson and Aitken (1976).

Theoretical results related to the Simonoff estimator (2.3) applied to binned data can be found in Simonoff (1995). Further, we use the estimator of (2.3) of the form

$$\hat{f}(x_1, x_2) = \left(\hat{f}_{2|1}(x_2 \mid x_1 \in I_{1i}) \hat{f}_1(x_1) \hat{f}_{1|2}(x_1 \mid x_2 \in I_{2j}) \hat{f}_2(x_2) \right)^{1/2} . \quad (2.6)$$

The rate of the Mean Squared Error for this estimator equals to $O(n^{-4/7})$, and is worse than the rate of the common univariate kernel estimator $O(n^{-2/3})$. This inferiority has been called the **quantitative effectiveness of smoothing**. However, it is balanced by the adaptive nature of the proposed estimator in the sense of mode determination.

It is worth noting how important is the correct choice of bins for multimodality detection of the underlying distribution. Figure 1 presents the effects of kernel density estimation of the unemployment rate and the average salary in Polish subregions in 2006 for various number of bins. Obviously, the number of bins should increase as the sample size increases. As it has been shown, it should increase with a rate $n^{2/7}$, the best rate with respect to squared error.

3. Robustness in the case of sparse contingency table

Effective analysis of high-dimensional discrete sparse data requires a special attention especially in the context of robustness of the procedure and its computational complexity. Issues related to robustness of the procedure dedicated to analysis of discrete data are not so highly developed as in the case of continuous data analysis. In the predominant part, good multivariate robust procedures are computationally very intensive. This in particular affects methods of nonparametric estimation of probability density function for high-dimensional data. As a starting point for our considerations and proposals we take pioneering works of J. Simonoff related to automatic and adaptive estimation of bivariate density function (see Simonoff, 1985, 1988, 1995), developed now by Jacob and Oliveira (see Jacob & Oliveira,2011).

Categorical data analysis is typically performed by fitting models to the observed counts in a contingency table using maximum likelihood. An inherent problem with maximum likelihood fits is their sensitivity to outlier cells, the ones whose counts are not consistent with the assumed model. Maximum likelihood estimates break down in the presence of outlying cells. It is worth noting that in categorical data analysis an outlier is a cell, i.e. a set of observations rather than a single observation, which deviates greatly from the expected count associated with the parametric model appropriate for the majority of cells.

Following Shane and Simonoff (2001), let us consider a D dimensional contingency table with d cells written as $d \times 1$ vector $n = (n_1, ..., n_d)$. Let $\mathbf{e} = (e_1, ..., e_d)$ be the vector of expected cell counts under a hypothesized model. The expected counts are $e_k = N \cdot p_k$, where N is the total sample size $\sum_{k=1}^{d} N_k$, where $\mathbf{p} = (p_1, ..., p_d)$ are theoretical cell probabilities. Assuming multinomial model for the cells we can understand robustness of the estimator in terms of goodness-of-fit statistics:

$$X^2 = \sum_{k=1}^{d} \chi_k^2(n_k, \hat{e}_k) = \sum_{k=1}^{d} \frac{(n_k - \hat{e}_k)^2}{\hat{e}_k} = \sum_{k=1}^{d} \frac{(n_k - p_k N)^2}{p_k N} \tag{3.1}$$

or equivalently the likelihood ratio goodness-of-fit statistics

$$G^2 = 2\sum_{k=1}^{d} n_k \log(n_k / \hat{e}_k) = 2\sum_{k=1}^{d} n_k \log(n_k / N \cdot p_k) \tag{3.2}$$

Let $X_{(l)}^2$ denote the $l-$ order statistics of X_k^2 . Shane and Simonoff (2001) define a robust Pearson estimate of a contingency table model as minimizing the criterion

$$\sum_{k=1}^{d} c_k X_{(k)}^2(n_k, e_k) \ , \tag{3.3}$$

where $\mathbf{c} = (c_1, ..., c_d)$ is an appropriate vector of weights.

The robust estimate according to Simonoff means a fit that is appropriate for the majority of cells and which is determined by the vectors of weights $\mathbf{c} = (c_1, ..., c_d)$. For continuous data this idea depends on the binning, the vector of weights and the measure used to assess the overall goodness of fit.

In the context of the analysis of sparse high-dimensional data for robustness of the procedure evaluation we propose to follow ideas presented in Mizera (2001). According to the ideas it is possible to define **halfspace depth** and **maximum depth based estimators** for the contingency tables. **General halfspace depth** can be defined as a measure data-analytic **admissibility of a fit with respect to the data**. **Depth** of $\mathbf{p}$ can be expressed as **the proportion of the data points** whose **omission causes** $\mathbf{p}$ to **become a nonfit**, a fit that can be uniformly dominated by another one.

For a contingency table with bins $\{I_{1i}\} \times \{I_{2j}\}$, $i = 1, ..., k_1$, $j = 1, ..., k_2$, we define the **depth of a fit** $\mathbf{p} = (p_1, ..., p_d)$ **as a minimal fraction of observations in the contingency table, whose replacement with other observations from the table will effect in taking the overall goodness-of-fit measure**

unacceptable value. As the overall goodness-of-fit measure we take Pearson statistics calculated for nonzero cells (we can use many other criteria functions instead, however):

$$F_{PEAR} = \sum\nolimits_{n_k \neq 0} \frac{(n_k - Np_k)^2}{Np_k} . \tag{3.4}$$

As the **robust estimator** of the model we take the **maximum depth estimator**.

In Mizera (2002) it is shown how to reformulate the general criteria (3.4) into the first order optimization. Mizera introduces the tangent depth - the depth of the fit takes a form

$$d(\mathbf{p}) = \inf_{\mathbf{u} \neq \mathbf{0}} \# \left\{ n : \mathbf{u}^T \nabla_{\mathbf{p}} F_{PEAR}(\mathbf{p}) \geq 0 \right\} . \tag{3.5}$$

where $\acute{\mathrm{N}}_p f$ denotes gradient of a function f in a point p .

Attractive breakdown point robustness of the maximum depth estimator follows from Mizera (2002).

4. Our proposals

Sparse methods could be described as methods which make interpretation of the statistical analysis easier by forcing the statistical procedure to produce sparser output that is, for example, a sparser vector of regression coefficients. As a prototype for the sparse methods one can take the ridge regression, the LASSO regression, or the ELASTIC NET. Considering regression data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\} \subset \square^{p+1}$, in ridge and LASSO regression correspondingly, as regression parameters estimates we take vectors

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 , \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq t , \tag{4.1}$$

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 , \text{ subject to } \sum_{j=1}^{p} \left| \beta_j \right| \leq t . \tag{4.2}$$

In the case of sparse PCA, taking into account the fact that an interpretation of the PCA components is conducted by examining the direction vectors known as **loadings** – we force the estimation procedure to produce sparser set of the loadings. Constraints encourages some loadings to be zero (for further details see Hastie et al. (2009)). The SCOTLASS procedure of Joliffe et al. (2003) focuses on maximum variance property of principal components by solving

$$\max \mathbf{v}^T \left(\mathbf{X}^T \mathbf{X} \right) \mathbf{v} , \text{ subject to } \sum_{j=1}^{p} \left| v_j \right| \leq t , \ \mathbf{v}^T \mathbf{v} = 1 . \tag{4.3}$$

Sparse and robust methods are relatively new and appeared during last 5 years (see Croux and Filzmoser, 2010).

Below we propose a general idea of producing a sparse and robust estimator of 2D density appealing **to functional data analysis.** The Simonoff estimator enables us to decompose 2D density estimation procedure (computationally a more complicated problem) into blocks which are estimated using 1D marginal densities and 1D conditional densities (computationally a less complicated problem).

Assuming a certain **sample of contingency tables** – for each of its cells we dispose of a certain number of marginal and conditional density estimates. We can successfully apply Functional Data Analysis (FDA) machinery to them. In particular we can use functional PCA of the estimated densities. Squared functional principal components fulfil density function postulates. We can decompose the overall density by means of them.

Let us consider functional data $x_1(t),...,x_s(t)$. Assuming we have chosen a basis $\phi_1,...,\phi_L$ (we advocate here on using basis consisted of splines), we consider representations of the data

$$x_r(t)=\sum_{j=1}^{L}c_{rj}\phi_j(t) \ , \tag{4.4}$$

where $c_{r1},...,c_{rL}$ are coefficients for r-th objects in this basis.

Coefficients $c_{r1},...,c_{rL}$ are chosen separately for every function $x_r(t)$. Assume we fixed L basis functions and then our data set consists of s functions $x_1(t),...,x_s(t)$. In the FDA we perform basic operations using $L\times s$ matrix containing object coefficients in the fixed basis (see Krzyśko et al., 2012). Introducing a quantity

$$\rho_\xi(x(t))=\int\xi(t)x(t)dt \ , \tag{4.5}$$

our aim is to find a function $\xi(t)$ which in a best way underlines a variability of the data, i.e. for which $\rho_\xi(x(t))$ takes the maximal value.

$$\textbf{FPCA GOAL}\text{: } \mu=\max_\xi\left\{\sum_{i=1}^{s}\rho_\xi^2(x(t))\right\} \text{ , under the condition } \int\xi^2(t)dt=1 . \tag{4.6}$$

It is common to use a restriction on weight function ξ , $\int\xi^2(t)dt=1$. In a similar manner as in the case of classical PCA a non-decreasing sequence of eigenvalues $\mu_1\geq\mu_2\geq\ldots\geq\mu_K$ is developed recursively: $\int\xi_j(t)\xi_l(t)dt=0$, $j=1,...,l-1$, $\int\xi_l^2(t)=1$. For further details see Ramsey et al. (2010) and Krzyśko et al. (2012).

5. Empirical examples

In order to illustrate the presented approach we used the Central Statistical Office (CSO) data concerning traceability of crimes and unemployment in Polish subregions in 2004 – 2010. We have analysed eight 5x5 contingency tables, each consisting of 66 observations. Figures 2 – 6 present kernel density estimates for marginal, conditional and joint probability distribution of the unemployment rate and traceability of crimes in Polish subregions in 2004 – 2010. Estimates were obtained using binned data presented in Table 1. Figures 7 – 18 present results of the functional PCA performed on the basis of 8 contingency tables consisting of data on traceability of crimes and the unemployment rate in Polish subregions. For simplicity of the presentation we focused only on one cell placed on the crossing of the shaded row and column in Table 2. We have performed similar analysis for the rest of the cells. It is easy to see that we can estimate the joint density of the variables using the idea of the Simonoff estimator (2.3) and using only the first or the second weight function (Fig. 9, Fig. 12, Fig. 15, Fig. 18). The output obtained in this way is much easier to interpret – the joint density function is decomposed into more evident layers. Although it is well known that the classical PCAs are not robust for outliers, several simulation studies we have performed using mixtures of various 2D discrete distributions show that our proposal seems to be robust to replacement of a small fraction of observations in the contingency table and in the spirit of Mizera (2002) ideas. It is possible, however, to directly the use robust PCA (see Croux et al., 2012) instead of classical PCA calculations during functional PCA. Our approach is computationally less intensive.

Table 2. A contingency table – traceability of crimes in Polish sub-regions in 2010

2010	X_{11}= 50.4	X_{12}= 58.4	X_{13}= 66.4	X_{14}= 74.4	X_{15}= 82.4	TOTAL
X_{21}=5.6	3	2	1	2	0	8
X_{22}=9.8	1	4	5	1	2	13
X_{23}=14.0	0	1	3	14	8	26
X_{24}=18.2	0	0	1	9	3	13
X_{25}=22.4	0	0	1	5	0	6
TOTAL	4	7	11	31	13	66

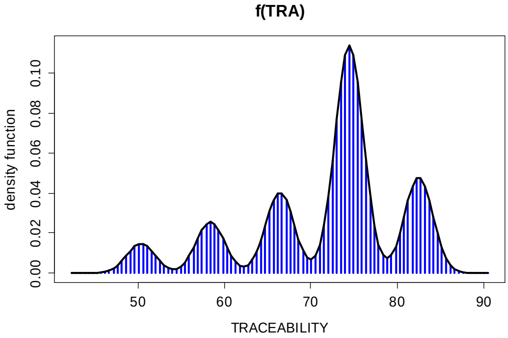

Figure 2. Kernel estimate of marginal density – traceability of crimes in Polish sub-regions in 2010

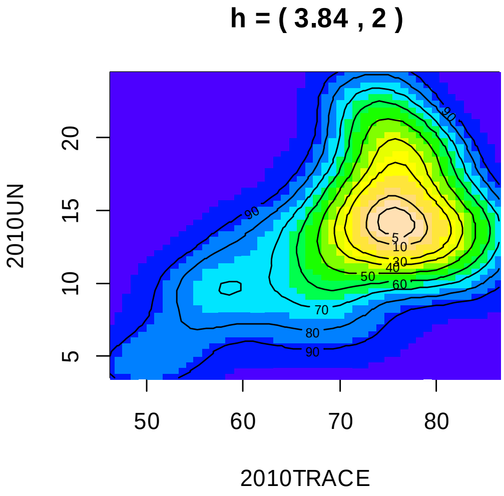

Figure 3. 2D kernel density estimate of unemployment rate vs. traceability of crimes in Polish sub-regions in 2010

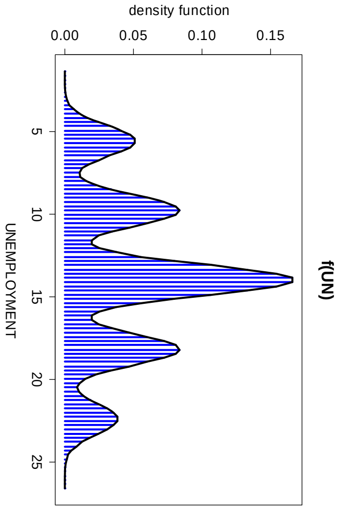

Figure 4. Kernel estimate of marginal density – unemployment rate in Polish sub-regions in 2010

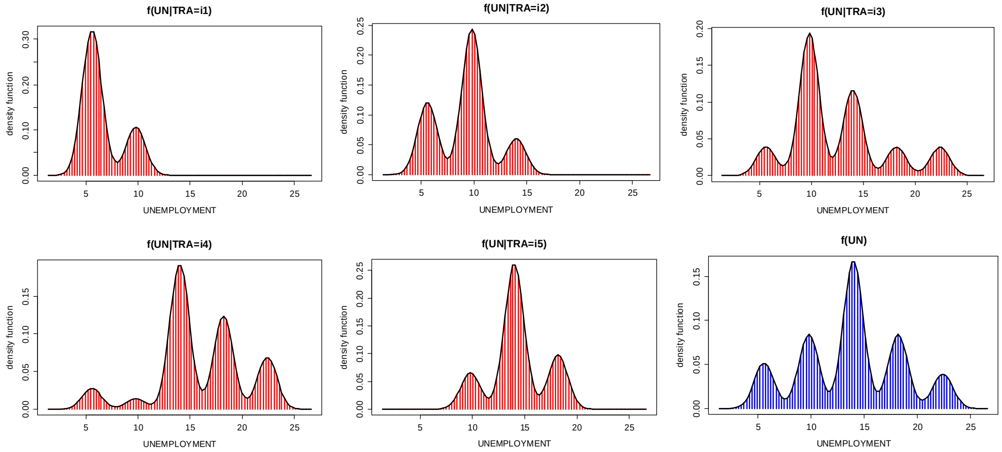

Figure 5. Conditional density estimate of unemployment rate under the condition that traceability of crimes takes value i1,…, i5. Last graph represents the unconditional density estimate of unemployment

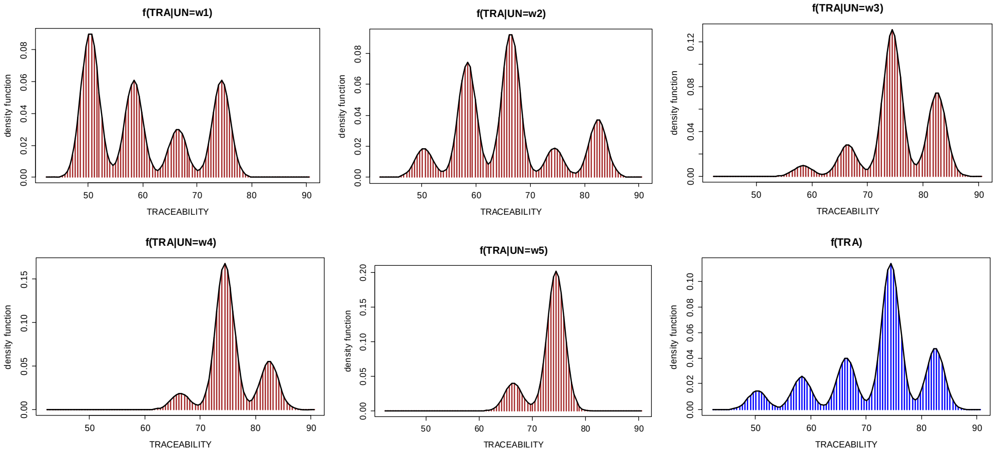

Figure 6. Conditional density estimate of traceability of crimes under the condition that unemployment rate takes value w1,…,w5. Last graph represents the unconditional density estimate of traceability

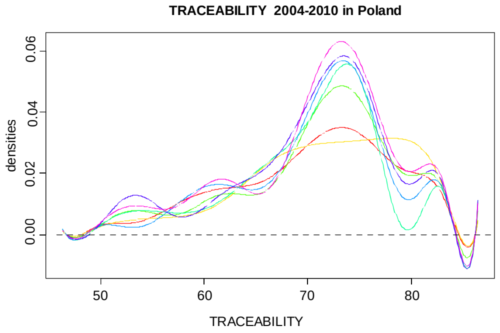

Figure 7. Density estimates for traceability of crimes in Polish subregions in 2004–2010

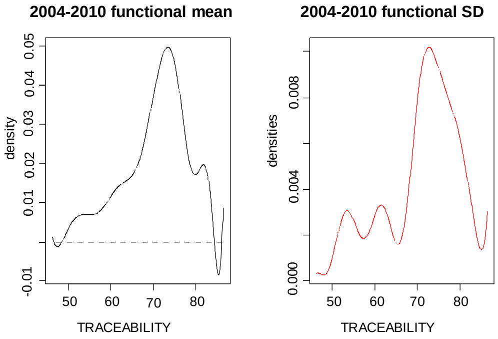

Figure 8. Functional mean (left) and functional SD (right) for density estimates for traceability of crimes in Polish subregions in 2004–2010

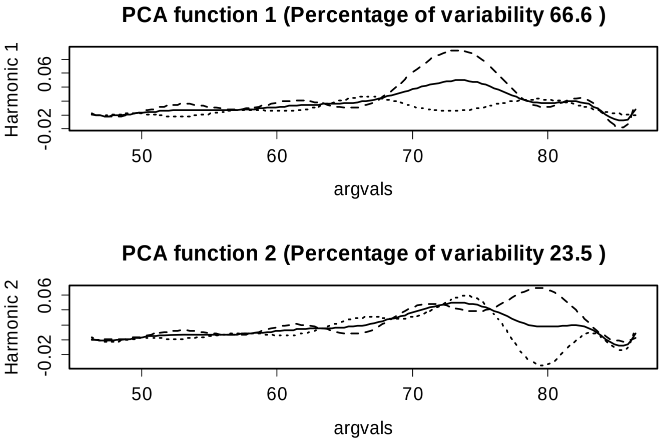

Figure 9. First and second weight functions (analogues of the eigenvectors) for density estimates for traceability of crimes in Polish subregions in 2004 – 2010

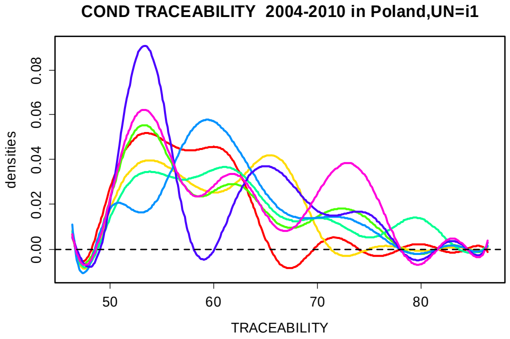

Figure 10. Density estimates for conditional traceability of crimes in Polish subregions in 2004–2010, condition unemployment rate = i1

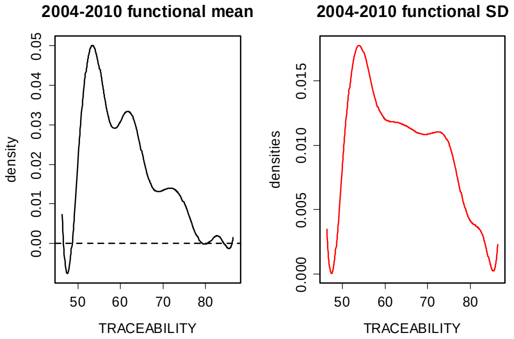

Figure 11. Functional mean (left) and functional SD (right) for conditional density estimates for traceability of crimes in Polish subregions

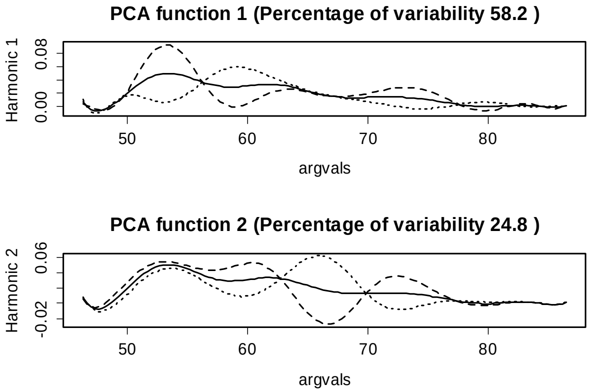

Figure 12. First and second weight functions (analogues of the eigenvectors) for conditional density estimates for traceability of crimes in Polish subregions in 2004–2010

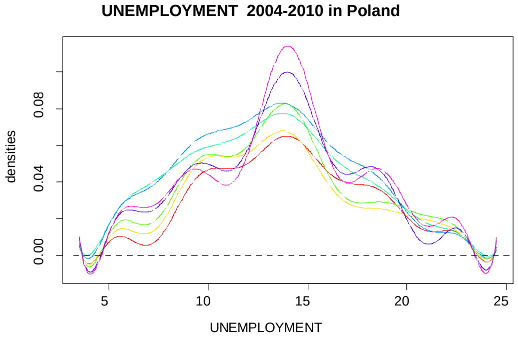

Figure 13. Density estimates for unemployment rate in Polish subregions in 2004–2010

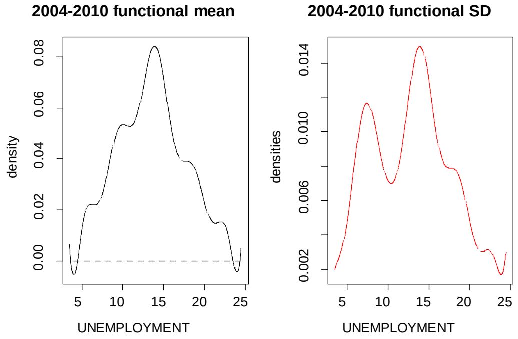

Figure 14. Functional mean (left) and functional SD (right) for density estimates for unemployment in Polish subregions in 2004–2010

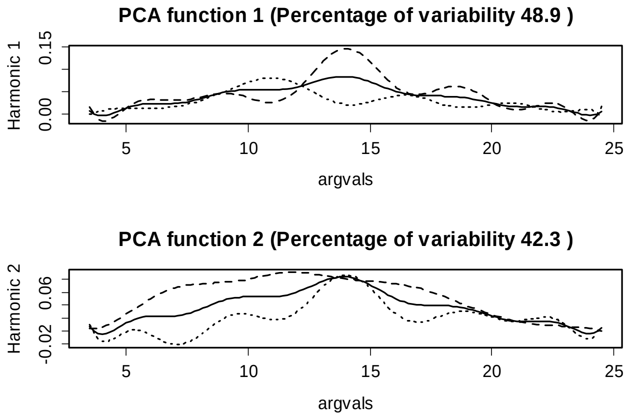

Figure 15. First and second weight functions (analogues of the eigenvectors) for density estimates for unemployment rate in Polish subregions in 2004–2010

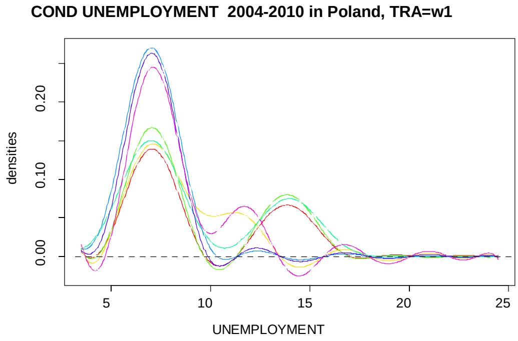

Figure 16. Density estimates for conditional traceability of crimes in Polish subregions in 2004–2010, condition unemployment rate = i1

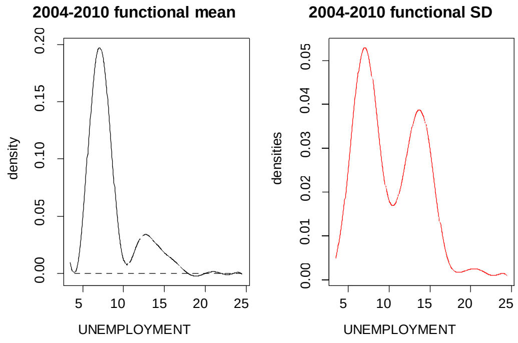

Figure 17. Functional mean (left) and functional SD (right) for conditional density estimates for traceability of crimes in Polish subregions

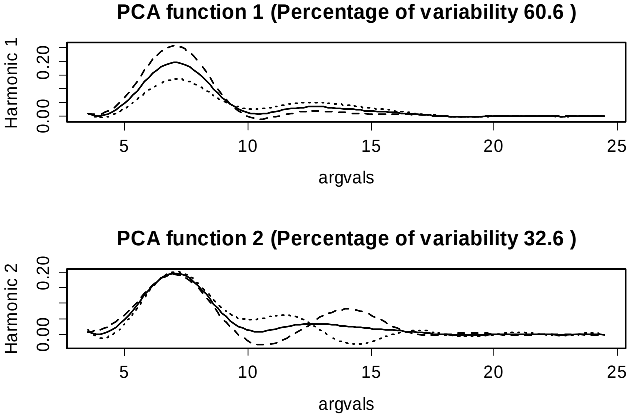

Figure 18. First and second weight functions (analogues of the eigenvectors) for density estimates for unemployment rate in Polish subregions in 2004–2010

6. The random matrix theory for detecting dependency between variables in a huge contingency table

Consider now that a contingency table, i.e. a data frame of p_1 input factors and p_2 output factors is observed continuously at n consecutive time moments. Let Y_{ia} be the value of the i-th ($i = 1,\ldots, p_1$) random variable at the a-th time moment ($a = 1,\ldots,n$); together, they make up a rectangular $p_1 \times n$ matrix $\mathbf{Y}$. Analogously, let X_{jb} be the value of the j-th ($j = 1,\ldots, p_2$) random variable at the b-th time moment ($b = 1,\ldots,n$); together, they make up a rectangular $p_2 \times n$ matrix $\mathbf{X}$. **In general** p_1, p_2, n **can be very large**. Further, we will assume that $p_1, p_2, n \to \infty$ but $p_1 / n = c_1$ and $p_2 / n = c_2$ are fixed. Under null hypothesis, each Y_{ia} and X_{jb} is supposed to be drawn from a Gaussian probability distribution, and that they have mean values zero. Specifically, the aim is to test the hypothesis:

H_0 **: x and y are independent; against** H_1 **: x and y are not independent,**

where $\mathbf{x} = (x_1,\ldots,x_{p_1})^T$ and $\mathbf{y} = (y_1,\ldots,y_{p_2})^T$. Without loss of generality, suppose that $p_1 \le p_2$.

It is well known that the canonical correlation analysis (CCA) deals with the correlation structure between two random vectors. Draw n independent and identically distributed (i.i.d.) observations from these two random vectors $\mathbf{x}$ and

y respectively, and group them into $p_1 \times n$ random matrix $\mathbf{X} = (x_1, \ldots, x_n) = (X_{ij})_{p_1 \times n}$ and $p_2 \times n$ random matrix $\mathbf{Y} = (y_1, \ldots, y_n) = (Y_{ij})_{p_2 \times n}$, respectively. The CCA seeks the linear combinations $\mathbf{a}^{\mathrm{T}}$ **x** and $\mathbf{b}^{\mathrm{T}}$ **y** that are most highly correlated, that is to maximize

$$\gamma = Corr(a^T x, b^T y) = \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} a}\sqrt{b^T \Sigma_{YY} b}} \tag{6.1}$$

where Σ_{XX} and Σ_{YY} are the population covariance matrices for x and y respectively, and Σ_{XY} is the population covariance matrix between x and y .

After finding the maximal correlation r_1 and associated vectors a_1 and b_1, CCA continues to seek a second linear combination $a_2^T x$ and $b_2^T y$ that has the maximal correlation among all linear combinations uncorrelated with $a_1^T x$ and $b_1^T y$. This procedure can be iterated and successive canonical correlation coefficients $\gamma_1, \ldots, \gamma_{p_1}$ can be found. It turns out that the population canonical correlation coefficients $\gamma_1, \ldots, \gamma_{p_1}$ can be recast as the roots of the determinant equation

$$\det(\Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY'} - \gamma^2 \sigma_{XX}) = 0 \tag{6.2}$$

This equation can be replaced by:

$$\det(G_{XY} D_{YY}^{-1} G_{XY'} - r^2 D_{XX}) = 0 \tag{6.3}$$

$$D_{XX} = \frac{1}{n} XX^T \qquad D_{YY} = \frac{1}{n} YY^T \qquad G_{XY} = \frac{1}{n} XY^T$$

We also think of D_{XX} , D_{YY} and G_{XY} as sample covariance matrices. However, due to dimensionality curse these are not consistent estimators of population covariance matrices, when the dimensions p_1 and p_2 are both comparable to the sample size n. As a consequence, it is conceivable that the classical likelihood ratio statistics do not work well in the high dimensional case.

Moreover, $r_1^2, r_2^2, \ldots, r_{p_1}^2$ are the eigenvalues of the matrix

$$S_{XX} = D_{XX}^{-1} G_{XY} D_{YY}^{-1} G_{XY^T} \tag{6.4}$$

Evidently, D_{XX}^{-1} and D_{YY}^{-1} do not exist when $p_1 > n$ and $p_2 > n$. For this reason we also consider the eigenvalues of the **regularized matrix**

$$T_{XY} = D_{tX}^{-1} G_{XY} D_{XY}^{-1} G_{XY^T} , \tag{6.5}$$

where $D_{tX}^{-1} = (\frac{1}{n}XX' + tI_{P_1})^{-1}$, t is a positive constant number and I_{p_1} is a $p_1 \times p_1$ identity matrix.

In addition to proposing statistics for testing we will also establish the limit of the ESD of regularized sample canonical correlation coefficients and central limit theorems (CLT) of linear functionals of the classical and regularized sample canonical correlation coefficients $r_1, r_2, \ldots, r_{p_1}$, respectively. To derive the CLT for linear spectral statistics of classical and regularized sample canonical correlation coefficients, the strategy is to first establish the CLT under the Gaussian case, the entries of X are Gaussian distributed. In the Gaussian case, the CLT for linear spectral statistics of the matrix S_{XY} can be linked to that of an F -matrix, which was investigated in Bai and Silverstein (1995).

We make the following assumptions:

1. $p_1 = p_1(n)$ and $p_2 = p_2(n)$ with $p_1 \to c_1$ and $p_2 \to c_2$, $c_1, c_2 \in (0,1)$ as $n \to \infty$
2. $X = (X_{ij})_{i,j=1}^{p_1,n}$ and $Y = (Y_{ij})_{i,j=1}^{p_2,n}$ satisfy $X = \Sigma_{XX}^{1/2} W$ and $Y = \Sigma_{YY}^{1/2} V$, where $W = (w_1, \ldots, w_n) = (W_{ij})_{i,j=1}^{p_1,n}$ consists of i.i.d. real random variables $\{W_{ij}\}$ with $EW_{11} = 0$ and $E|W_{11}|^2 = 1$; $V = (v_1, \ldots, v_n) = (V_{ij})_{i,j=1}^{p_2,n}$ consists of i.i.d. real random variables $\{V_{ij}\}$ with $EV_{11} = 0$ and $E|V_{11}|^2 = 1$; $\Sigma_{XX}^{1/2}$, $\Sigma_{YY}^{1/2}$ are Hermitian square roots of positive definite matrices Σ_{XX} and Σ_{YY} .
3. $F^{\Sigma_{XX}} \to^{D} H$ a proper cumulative distribution function.

By the definition of the matrix S_{XY}, the classical canonical correlation coefficients between x and y are the same as those between w and v when w, v are i.i.d.

We now introduce some results from random matrix theory and free probability theory as presented by Voiculescu (1991).

Definition 6.1: Denote the ESD of any $n \times n$ matrix A with real eigenvalues $\mu_1 \le \mu_2 \le \ldots \le \mu_n$

$$F^{A}(x) = \frac{1}{n}\#\{i : \mu_i \le x\}, \tag{6.6}$$

where $\#\{\ldots\}$ denotes the cardinality of the set $\{\ldots\}$.

Theorem 6.2: When the two random vectors x and y are independent and each of them consists of i.i.d Gaussian random variables, under Assumptions 1 and 2, the empirical measure of the classical sample canonical correlation coefficients $r_1, r_2, \ldots, r_{p_1}$ converges in probability to a fixed distribution whose density is given by

$$\rho(x) = \frac{\sqrt{(x-L_1)(x+L_1)(L_2-x)(L_2+x)}}{\pi c_1 x(1-x)(1+x)} \quad , \qquad (6.7)$$

$x \in [L_1, L_2]$, and atom size of $\max(0,(1-c_2)/c_1)$ at zero and size $\max(0,1-(1-c_2)/c_1)$ at unity, where $L_1 = |\sqrt{c_2 - c_2 c_1} - \sqrt{c_1 - c_1 c_2}\,|$ and $L_2 = |\sqrt{c_2 - c_2 c_1} + \sqrt{c_1 - c_1 c_2}\,|$.

Here, the empirical measure of $r_1, r_2, \ldots, r_{p_1}$ is defined as in the ESD with μ_i replaced by r_i.

Let us now introduce the test statistics. Under Assumption 1 and Assumption 3, if $Y = \sigma_1 W$ and $X = \sigma_2 W$ with $p_1 = p_2$ and both Σ_1 and Σ_2 being invertible, then $S_{XY} = 1$, which implies that the limit of $F^{S_{XY}}(x)$ is a degenerate distribution. Thus, we consider the following statistics

$$S_n = \int x dF^{S_{XY}}(x) = \frac{1}{p_1}\sum_{i=1}^{p_1} r_i^2. \qquad (6.8)$$

In the classical CCA, the maximum likelihood ratio test statistics with fixed dimensions is

$$MLR_n = \sum_{i=1}^{p_1} \log(1-r_i^2). \qquad (6.9)$$

Note that the density $\rho(x)$ has atom size of $\max(0,1-(1-c_2)/c_1)$ at unity. Thus, the normalized statistics MLR_n is not well defined when $c_1 + c_2 > 1$ (because $\int \log(1-x^2)dx$ is not meaningful). In addition, even when $c_1 + c_2 \le 1$, the right end point of $\rho(x)$, L_2, can be equal to one so that some sample correlation coefficients r_i are close to one. For example, $L_2 = 1$ when $c_1 = c_2 = 1$. This in turn causes a big value of the corresponding $\log(1-r_i^2)$. **Therefore, MLR_n is not stable.**

Here we would like to point out that the idea of testing independence between two random vectors x and y by the CCA is based on the fact that the lack of correlation between x and y is equivalent to independence between them when the random vector of size (p1 +p2) consisting of the components of x and y is a Gaussian random vector.

In addition, it can be proved that

$$\mathrm{Tr}(G_{XY}^{H_1} - G_{XY}^{H_0}) = O_p(n) \qquad (6.10)$$

ALGORITHM FOR THE PROCEDURE – "DOUBLE SPARSITY ALGORITHM"

STEP 1. Preparation of the dataset

Now we will extend our consideration to the case of n consecutive observations. First, let us divide all variables into two subsets, i.e. focus on p_1 input factors X_a $(a = 1,\dots,p_1)$ and p_2 output factors Y_α $(\alpha = 1,\dots,p_2)$ with the total number of observations being n. All series of observations are standardized to have zero mean and unit variance. The data can be completely different or can be the same variables but observed at different times. First, one has to remove potential correlations inside each subset, otherwise it may interfere with the out-of-sample signal. To remove the correlations inside each sample we form two correlation matrices which contain information about in-the-sample correlations:

$$\mathbf{D}_{\mathbf{XX}} = \frac{1}{n} XX^T, \qquad \mathbf{D}_{\mathbf{YY}} = \frac{1}{n} YY^T$$

STEP 2. Diagonalization

The matrices are then diagonalized, provided $n > p_1, p_2$, and the empirical spectrum is compared to the theoretical Bai, Silverstein (1995) result

$$\rho(x) = \frac{1}{2\pi x} \mathbf{Re}\sqrt{(x - L_1)(L_2 - x)} \; L_1 = (1 \pm \sqrt{c_1})^2 \qquad L_1 = (1 \pm \sqrt{c_1})^2$$

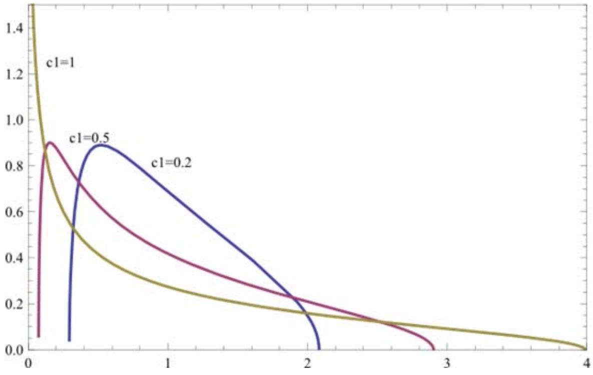

Figure 19. The spectrum of the single sparse matrices D_{XX} and D_{YY} when null hypothesis holds (i.e., there are no internal temporal correlations. The eigenvalues of ESD, which lie much below the lower edge of the spectrum, represent the redundant factors inconsistent with the null hypothesis)

STEP 3. Reconstruction

One can then construct a set of uncorrelated unit variance input variables $\hat{X}$ and output variables $\hat{Y}$

$$\hat{X}_{w_i} = \frac{1}{\sqrt{nw_i}} W^T X_i \qquad \hat{Y}_{v_j} = \frac{1}{\sqrt{nv_j}} V^T Y_j$$

where V,U, λ_a, λ_α are the corresponding eigenvectors and eigenvalues of D_{XX}, D_{YY}.

Finally, we can reproduce the asymmetric $p_1 \times p_2$ cross-correlation matrix G between the $\hat{Y}$ and $\hat{X}$:

$$G = \hat{X}\hat{Y}^T .$$

Under the null hypothesis of independence between X and Y, the ESD should follow the distribution with density (see, Snarska 2012)

$$\rho_{\mathbf{G}}(x) = \max(1-c_1, 1-c_2)\delta(x) + \max(c_1+c_2-1, 0)\delta(x-1) + \frac{\operatorname{Re}\sqrt{(x^2-s_-)(x_+-s^2)}}{\pi x(1-x^2)},$$

where $x_\pm = c_1 + c_2 - 2c_1c_2 \pm 2\sqrt{c_1c_2(1-c_1)(1-c_2)}$ are the two positive roots of the quadratic expression under the square root. It is easy to see the fact that in the limit $n \to \infty$ at fixed p_1, p_2 all singular values collapse to zero as they should since there are no true correlations between X and Y; the allowed band in the limit $c_1, c_2 \to 0$ becomes: $x \in \left[|\sqrt{c_1} - \sqrt{c_2}|, \sqrt{c_1} + \sqrt{c_2} \right]$. When $c_1 \to c_2$, the support becomes $x \in [0, 2\sqrt{c_1(1-c_1)}]$ (plus a δ function at $x = 1$ when $c_1 + c_2 > 1$), while when $c_1 = 1$, the whole band collapses to a δ function at $x = \sqrt{1-n}$. For $c_1 + c_2 \to 1^-$ there is an initial singularity of $\rho(x)$ $x = 1$ diverging as $(1-x)^{-1/2}$. Ultimately, $c_1 \to 0$ at fixed c_2, one finds that the whole band collapses again to a δ function at $x = \sqrt{c_2}$.

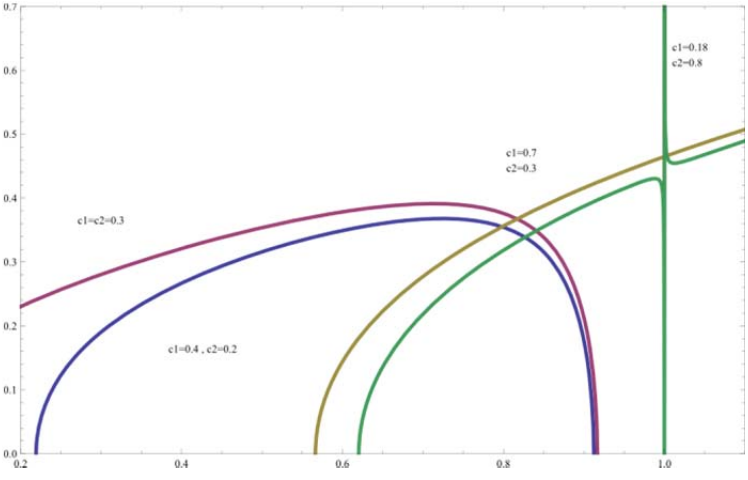

Figure 20. Theoretical distribution of singular values for G_{XY} under validity of null hypothesis. The eigenvalues of ESD, which lie much below the lower edge of the spectrum, represent the redundant factors inconsistent with the null hypothesis

7. Conclusions

A common application of the statistical procedures has changed business and the economy. Statistics have changed the ways we reason in a public debate, form our opinions, manage banking systems, perform interventions in a certain market, allocate energy stored in the capital between competing investments.

The innovative nature of the outlined approach to big economic databases analysis is manifested in formation of a complete methodology for a robust analysis of sparse high-dimensional discrete data in the economy. Our approach is still being developed and we hope to obtain interesting results in the near future. We are convinced that our proposal could find several applications in the on-line economy and exploration of the official statistics databases.

Acknowledgements

Daniel Kosiorowski thanks for financial support from Polish National Science Center grant UMO-2011/03/B/HS4/01138.

REFERENCES

CROUX, C., FILZMOSER, P., FRITZ, H., (2012). Robust Sparse Principal Component Analysis, Technometrics

DONG, J., SIMONOFF, J. S., (1994). The Construction and Properties of Boundary Kernels for Smoothing Sparse Multinomials. Journal of Computational and Graphical Statistics. Vol. 3, No. 1, 57–66.

HASTIE, T., TIBSHIRIANI, R., FRIEDMAN, J., (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer.

JACOB, P., OLIVEIRA, P. E., (2011). Local smoothing with given marginals, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2011.561436

JOLIFFE, I. T., TRENDAFILOV, N. T., UDDIN, M., (2003). A modified principal component technique based on the lasso, Journal of Computational and Graphical Statistics 12: 531–547.

KRZYŚKO, M., GÓRECKI, T., DERĘGOWSKI, K., (2012). Jądrowa i Funkcjonalna Analiza Składowych Głównych – spotkanie PTS o. w Poznaniu (referat dostępny na stronach PTS o. w Poznaniu http://www.stat.gov.pl/pts/)

MIZERA, I., (2002). On Depth and Depth Points: a Calculus. *The Annals of Statistics* (30), 1681–1736.

RAMSAY, J. O., HOOKER, G., GRAVES, S., (2010). *Functional Data Analysis with R and Matlab*, Springer, New York.

SHANE, K. V., SIMONOFF, J. S., (2001). A robust approach to categorical data analysis, Journal of Computational and Graphical Statistics, Vol. 10, No. 1, 135–157.

SILVERSTEIN, J., BAI, Z., (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. Journal of Multivariate Analysis 54, (2), 175–192.

SIMONOFF, J. S., (1985). An improved goodness-of-fit statistic for sparse multinomials, Journal of the American Statistical Association, Vol. 80, No. 391, 671–677.

SIMONOFF, J. S., (1988). Detecting outlying cells in two-way contingency tables via backward-stepping, Technometrics, Vol. 30, No. 3, 339–345.

SIMONOFF, J. S., (1995). A simple, automatic and adaptive bivariate density estimator based on conditional densities, Statistics and Computing, Vol. 5, 245–252.

SIMONOFF, J. S., (1983). A penalty function approach to smoothing large sparse contingency tables. The Annals of Statistics. Vol. 11, No. 1, 208–218.

SIMONOFF, J. S., (1998). Three sides of smoothing: categorical data smoothing, nonparametric regression, and density estimation. International Statistical Review, Vol. 66, No. 2, 137–156.

SNARSKA, M., (2012). A random matrix approach to dynamic factors in macroeconomic data, Acta Phys. Pol A, 121 (2B), 110–120**.**

VOICULESCU, D. V., (1991). Limit laws for random matrices and free products, Invent. Math. 104, 201.