SOME ROBUST APPROACHES TO REDUCING THE COMPLEXITY OF ECONOMIC DATA

Jan Kalina

Abstract

The recent advent of complex (and potentially big) data in economics requires modern and effective tools for their analysis including tools for reducing the dimensionality (complexity) of the given data. This paper starts with recalling the importance of Big Data in economics and with characterizing the main categories of dimension reduction techniques. While there have already been numerous techniques for dimensionality reduction available, this work is interested in methods that are robust to the presence of outlying measurements (outliers) in the economic data. Particularly, methods based on implicit weighting assigned to individual observations are developed in this paper. As the main contribution, this paper proposes three novel robust methods of dimension reduction. One method is a dimension reduction within a robust regularized linear regression, namely a sparse version of the least weighted squares estimator. The other two methods are robust versions of feature extraction methods popular in econometrics: robust principal component analysis and robust factor analysis.

Key words: dimensionality reduction, Big Data, variable selection, robustness, sparsity

JEL Code: C14, C55, C29

Introduction

The amount of data observed in economic research grows very rapidly, while complex data are commonly agreed to have a big potential to influence economic research, economic policy, or everyday routine economic activities (López-Robles, 2019). Thus, economic data represent a valuable capital with an underutilized opportunity for economic decision making and relevance for the economic and social state of the society. Especially Big Data seem still undervalued by economists worldwide in terms of their significance and potential (Blazques & Domenech, 2018). A vast number of possible sources of economic data with a large number of variables include retail, finance, advertising, insurance, online trade, portfolio optimization, risk management, labor market dynamics, effect of education on earnings, customer analytics (customer analytical records), automotive industry, or stock market

dynamics (Bradlow et al., 2017). A correct analysis of economic data with a large number of available variables (features) therefore becomes an emerging issue in current econometrics.

Together with an increasing comlexity of data, econometricians begin to realize the importace of methods of dimensionality reduction (complexity reduction) for their reliable and effective analysis (Fordellone, 2019). This paper is devoted to dimensionality reduction methods for economic data and we propose here three novel robust approaches. We are especially interested in the idea of implicit weights assigned to individual observations, which comes from the idea of the least weighted squares estimator of Víšek (2011); all the three robust tools presented here exploit such implicit weighting. Particularly, Section 1 recalls the advent of complex (potentially big) data in economics and reminds the importance of their complexity reduction. Section 2 presents an overview of the main categories of complexity reduction techniques. Section 3 proposes a novel regularized version of the least weighted squares estimator in linear regression. Section 4 proposes a robust version of principal component analysis. Section 5 proposes a robustified version of factor analysis.

1 Economic Big Data

Recently, econometricians seem to experience two main directions in economic data analysis: more intensive analysis of currently available data is performed, and there are at the same time attempts for acquiring additional data, either from current data sources, or from new (non-traditional) sources. Tools for analyzing big economic data were overviewed e.g. in Choi et al. (2018), however without specifying potential sources of Big Data.

It is possible to distinguish between three following types of data sources for decision making within a given organization (company); the analysis of such managerial data often requires to combine these sources and to start the analysis by reducing their complexity, and potentially also to replace non-numerical data by numerical variables (features).

1. Internal sources of data. These may be available in the management information system (e.g. managerial reports, customer analytical reports, accounting data, or information about the work productivity and effectivity of individual employees) and may also be related to personal information (human resources), marketing evidence, quality, or risk management.
2. External sources of data.
    - Governmental data (either publicly available large administrative datasets or data presented by government or municipality offices upon request);

- Other public data (publicly available contrasts in public hospitals or macroeconomic predictions presented by statistical offices, such as prediction of inflation rate);
- Data from business associations (e.g. models of consumer behaviour for the whole population in a given country);
- Social networks or internet (data about browsing of users on websites);
- Data from mobile sensors, with a detailed description of social and economic applications in Blazques & Domenech (2018);
- Scientific literature (or popularization of science);
- Stakeholders or professional experts outside the organization.

3. Own research performed by the organization, e.g. marketing research or customer satisfaction surveys; intensive attention has been paid to methods for appropriate formulating of questions for economic surveys (Taylor et al., 2014).

2 Dimensionality reduction in econometrics

In order to simplify the analysis of complex multivariate data, dimensionality reduction is generally recommended in various economic applications (e.g. in Wilson (2018)) in spite of losing some relevant information. Parsimonious economic models, i.e. simple models with a small set of relevant variables, may enable a good comprehensibility of the result from the economic point of view. They may even improve the results compared to those obtained with full data. On the other hand, if the set of variables is reduced to a too small number of relevant ones, the results may be severaly biased.

Dimensionality reduction should be tailored for the particular economic task/problem as well as the statistical task of the analysis (i.e. regression, instrumental variables estimation, classification, clustering). For all the situations, there are two basic ways of reducing the dimensionality, namely to perform it as a prior step before the analysis (classification, regression, etc.) or to include it as an intrinsic step within the analysis, which may possibly exploit regularized methods. Let us now explain the difference between a prior dimensionality reduction and an intrinsic dimensionality reduction.

Prior dimensionality reduction represents a preliminary or assistive step prior to the particular analysis task, i.e. is performed prior to the regression modeling or for example learning a classification rule.

[I] Variable selection searches for a small set of relevant variables while recommends to ignore the remaining ones.

[II] Feature extraction methods search for linear combinations of the measurements. Popular methods include principal component analysis (PCA), factor analysis, methods based on information theory, correspondence analysis, or multivariate scaling (see e.g. the monograph by Greene (2017)).

Both variable selection and feature extraction methods may suffer from the presence of outliers in the data, and some their robust versions have been already available. Still, it should be stressed that principal component analysis as the most popular method is not suitable if the data come from two or more different groups (i.e. for a mixture of populations), while supervised dimension reduction methods are preferable.

Intrinsic dimensionality reduction may be performed within regression or classification tasks by means of regularization (Fan et al., 2020). Such methods yield typically sparse solutions, exploiting information only from some variables while ignoring completely the remaining observations. Although regularization may bring local robustness to small changes of the data, we can say in general that regularized methods may suffer from outliers.

3 Implicitly weighted lasso estimator

Assuming now the standard linear regression model, we propose here a novel esimator denoted as LWS-L1; the abbreviation reveals the L1-regularization and thus also an intrinsic dimensionality reduction, which is performed in the same spirit as for the lasso estimator. We assume the standard linear regression model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where $Y_1, \ldots, Y_n$ are values of a continuous response variable and $e_1, \ldots, e_n$ are random errors (disturbances) with a common value of $var\, e_i = \sigma^2$ with $\sigma > 0$. The lasso estimator represents an L_1-regularized estimator of the vector $\beta = (\beta_1, \ldots, \beta_p)^T$. It is suitable for correlated regressors, however it remains vulnerable to the presence of outliers in the data. Therefore, some robust alternatives have been already proposed, including the LTS-lasso of Alfons et al. (2013), which is based on the least trimmed squares (LTS) estimator. As the solution is typically sparse, especially if p is large and the regressors are correlated, the estimator performs an intrinsic variable selection, as explained in Section 2.

We exploit here the idea of the least weighted squares (LWS) regression estimator of Víšek (2011) for (1), which is able to combine high robustness with high efficiency. It downweights less reliable data points by a set of continuous weights; if the outliers in (1) get zero weights, then the estimator may attain a high breakdown point. The magnitudes of non-negative weights $w_1, \dots, w_n$ must be chosen by the user, while the weights are assigned to particular observations after a permutation, which is determined automatically only during the computation based on the residuals. Other approaches to rank methods in regression are also known to yield robust results (Saleh et al., 2012).

To give the forma definition, let us consider a fixed estimate $b = (b_1, \dots, b_p)^T \in \mathbb{R}^p$ of the vector parameter β. It will be useful to denote the residual corresponding to the i-th observation as

$$u_i(b) = Y_i - b_1 X_{i1} - \cdots - b_p X_{ip}, \quad i = 1, \dots, n. \tag{2}$$

Ordering squared residuals as

$$u^2_{(1)}(b) \leq u^2_{(2)}(b) \leq \cdots \leq u^2_{(n)}(b), \tag{3}$$

we recall the formal definition of the LWS estimator of β in the form

$$\arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^n w_i u^2_{(i)}(b). \tag{4}$$

Using the same notation, we now define the novel LWS-L1 estimator by means of

$$\arg \min_{b \in \mathbb{R}^p} \left[\sum_{i=1}^n w_i u^2_{(i)}(b) + \lambda \sum_{j=1}^p |b_j| \right]. \tag{5}$$

A suitable value of the regularization parameter $\lambda > 0$ may be found using cross-validation. The computation may exploit an adapted version of the FAST-LTS algorithm, which is a standard tool for computing (i.e. approximating) the LTS estimator. The LWS-L1 estimator represents a robust regularized estimator of β, which is resistant against multicollinearity in the model (cf. Kalina et al., 2019).

4 Robust regularized principal component analysis

While principal component analysis represents a popular feature extraction method, it suffers from the presence of outliers. Its regularized versions have been available, and we propose here a novel robust regularized version based on idea of the LWS estimator, which was

recalled in Section 3. Let us start by recalling the (implicitly weighted) LWS-correlation coefficient. This is defined for two n-dimensional vectors $X = (X_1, \dots, X_n)^T$ and $Y = (Y_1, \dots, Y_n)^T$ as

$$r_W(x, y; w) = \frac{\sum_{i=1}^n w_i(X_i - \bar{X}_W)(Y_i - \bar{Y}_W)}{\sqrt{\sum_{i=1}^n [w_i(X_i - \bar{X}_W)^2] \sum_{i=1}^n [w_i(Y_i - \bar{Y}_W)^2]}} \tag{6}$$

with weights $w = (w_1, \dots, w_n)^T$ obtained by the LWS regression estimator, where

$$\bar{X}_W = \sum_{i=1}^n w_i X_i \quad \text{and} \quad \bar{Y}_W = \sum_{i=1}^n w_i Y_i. \tag{7}$$

Algorithm 1. (Robust regularized correlation matrix)

Let us consider p-dimensional data vectors $X_1, \dots, X_n$.

1. Compute the LWS-correlation matrix

$$R_{LWS} = \left(R_{ij}^{LWS}\right)_{i,j=1}^p, \tag{8}$$

where R_{ij}^{LWS} is equal to the LWS-correlation coefficient between $(X_{1j}, \dots, X_{nj})^T$ and $(X_{1i}, \dots, X_{ni})^T$.

2. Compute

$$S_{LWS} = \left(S_{ij}^{LWS}\right)_{i,j=1}^p, \tag{9}$$

as proposed in Kalina and Tichavský (2020), using R_{LWS}.

3. Compute the value of the regularization parameter $\delta^* \in [0,1]$ as

$$\delta^* = \frac{2\sum_{i=2}^p \sum_{j=1}^{i-1} \widehat{var}(S_{ij}^{LWS})}{2\sum_{i=2}^p \sum_{j=1}^{i-1} \left(S_{ij}^{LWS}\right)^2} \tag{10}$$

i.e. in the same way as in Schäfer & Strimmer (2005).

4. Compute $S^* = (1 - \delta^*)S_{LWS} + \delta^* I$, where I denotes a unit matrix.

5. Compute

$$R^* = (R_{ij}^*)_{i,j=1}^p, \tag{11}$$

where

$$R_{ij}^* = \frac{S_{ij}^*}{\sqrt{S_{ii}^* S_{jj}^*}}, \quad i, j = 1, \dots . p. \tag{12}$$

Algorithm 2. (Robust regularized principal component analysis)

Let us consider p-dimensional data vectors $X_1, \dots, X_n$.

1. Using Algorithm 1, compute R^*.
2. Compute the eigenvalues of R^* and denote them as $\lambda_1^*, \dots, \lambda_p^*$. The corresponding eigenvectors will be denoted as $z_1^*, \dots, z_p^*$.
3. Find r as the minimal integer fulfilling

$$\sum_{j=1}^{r} \lambda_j^* \geq 0.9 \tag{13}$$

   and $r \leq p$.
4. Each of the observations X_i is replaced by the set of the first r principal components $z_1^{*T} X_i, \dots, z_r^{*T} X_i$ for $i = 1, \dots, n$.
5. The observations are replaced by the set of the transformed observations

$$(z_1^{*T} X_1, \dots, z_r^{*T} X_1)^T, \dots, (z_1^{*T} X_n, \dots, z_r^{*T} X_n)^T. \tag{14}$$

The method reduces the dimension of the observations from p to a smaller value r. The robustness of such approach is ensured by the robustness of each step of the computation.

5 Robust factor analysis

Factor analysis, commonly used in economic data analysis, is based on the assumption that the observed data can be explained by means of a small number of latent variables (factors). The model of factor analysis can be expressed for the i-th observation as

$$\begin{aligned} X_{i1} - \mu_1 &= \gamma_{11} f_{i1} + \cdots + \gamma_{1t} f_{it} + e_{i1,} \\ &\vdots \\ X_{ip} - \mu_p &= \gamma_{p1} f_{i1} + \cdots + \gamma_{pt} f_{it} + e_{ip,} \end{aligned} \tag{15}$$

where $i = 1, \dots, n$. A particular observation $X_i = (X_{i1}, \dots, X_{ip})^T$ is explained by means of latent factors $f_{i1}, \dots, f_{it}$, parameters $\mu_1, \dots, \mu_p$ and $\gamma_{11}, \dots, \gamma_{pt}$, and noise $e_{i1}, \dots, e_{ip}$. The model (15) can be expressed in the matrix notation as

$$X_i - \mu = \Gamma f_i + e_i, \quad i = 1, \dots, n. \tag{16}$$

In contrary to the principal component analysis, it is not assumed that the latent variables explain the whole variability of the observed data. The part (component) of the variability of an individual variable, which is explained by the latent factors, is denoted as communality.

There are various approaches to estimating the parameters in (15). Let us denote by S the empirical covariance matrix obtained from all observations $X_1, \dots, X_n$. It is assumed that $S = \Gamma\Gamma^T + var\ e$ and that the matrix $var\ e$ is diagonal. Thus, we estimate the off-diagonal elements of the matrix $T = S - var\ e$ directly using off-diagonal elements of S. Diagonal elements of T can be estimated by an iterative procedure. This allows us to obtain an estimate of the whole matrix T. Further, such matrix Γ is searched for, which fulfils $\Gamma\Gamma^T = T$. However, if U is a given (any) orthogonal matrix, also the matrix $\Gamma^* = \Gamma U$ fulfils $\Gamma^*\Gamma^{*T} = T$, which complicates such search. In other words, the latent variables are not determined uniquely. Various approaches for estimating Γ have been proposed:

- Principal component analysis,
- Method of principal factors,
- Iterated method of principal factors,
- Maximum likelihood method,
- Minimization of residuals.

We propose a novel robust approach to factor analysis based on estimating Γ by the robust principal component analysis of Section 4. In such approach, the empirical covariance matrix S is replaced by S_{LWS} of Kalina and Tichavský (2020). The robust principal component analysis method for estimating parameters of factor analysis can be explained by means of the spectral decomposition of the matrix T (obtained from S_{LWS}) in the form $T = Q\Lambda Q^T$. Let us denote by Q_t the matrix containing the first t columns of Q and by $\Lambda_t^{1/2}$ the diagonal matrix, the diagonal elements of which are equal to square roots of the first t diagonal elements of Λ. The matrix Γ is then in the novel robust approach determined as $\Gamma = Q_t\Lambda_t^{1/2}$. Alternatively, the matrix S can be replaced by the robust regularized correlation matrix defined in Algorithm 1.

Conclusions

Data with a large number of variables bring a big potential to economic research and at the same time represent an important force for the development of the economies worldwide. Datasets with a large number of variables may appear in various tasks including linear

regression, classification analysis, clustering or time series analysis. The analysis of such datasets, which is hardly possible without an appropriate and reliable dimensionality reduction, may play an irreplacable role in the commercial sphere as well as in the contribution to the development of economic theory.

This paper is focused on robust dimensionality reduction methods suitable for econometric data, which may be contaminated by outliers. The LWS-L1 estimator represents a novel intrinsic variable selection approach based on regularization, where the latter starts to be acknowledged as a useful tool in the analysis of economic data with a large number of variables. In addition to reducing the dimensionality, it has a potential to divide variables to clusters and to reduce or remove correlation among variables.

The novel robust versions of principal component analysis and factor analysis search for linear combinations (latent factors) behind the data. Their robustification uses the idea of implicit weighting, which comes again from the LWS estimator. We plan to perform numerical experiments with the novel methds over simulated as well as real economic data to investigate the performance of the newly proposed methods, i.e. to reveal their robustness and at the same time their performance over non-contaminated datasets (without outliers).

Acknowledgment

The work was supported by the project 21-05325S (“Modern nonparametric methods in econometrics”) of the Czech Science Foundation.

References

Alfons, A., Croux, C., Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics*, 7(1), pp. 226-248.

Blazques, D., Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting & Social Change*, 130, pp. 99-113.

Bradlow, E.T., Gangwar, M., Kopalle, P., Voleti, S. (2017). The role of Big Data and predictive analytics in retailing. *Journal of Retailing*, 93, pp. 79-95.

Choi, T.M., Wallace, S.W., Wang, Y. (2018). Big Data analytics in operations management. *Production and Operations Management*, 27, 1868-1883.

Fan, J., Ke, Y., Wang, K. (2020). Factor-adjusted regularized model selection. *Journal of Econometrics*, 216, pp. 71-85.

Fordellone, M. (2019). *Statistical analysis of complex data. Dimensionality reduction and classification methods.* LAP LAMBERT Academic Publishing, Mauritius.

Greene, W.H. (2017). *Econometric analysis*. 8th edn. Pearson, London.

Kalina, J., Tichavský, J. (2020). On robust estimation of error variance in (highly) robust regression. *Measurement Science Review*, 20, pp. 6-14.

Kalina, J., Vašaničová, P., Litavcová, E. (2019). Regression quantiles under heteroscedasticity and multicollinearity: Analysis of travel and tourism competitiveness. *Ekonomický časopis*, 67, pp. 69-85.

López-Robles, J.R., Rodríguez-Salvador, M., Gamboa-Rosales, N.K., Ramirez-Rosales, S., Cobo, M.J. (2019). The last five years of Big Data Research in economics, econometrics and finance: Identification and conceptual analysis. *Procedia Computer Science*, 162, pp. 729-736.

Saleh, A.K.M.E., Picek, J., Kalina, J. (2012). R-estimation of the parameters of a multiple regression model with measurement errors. *Metrika*, 75, pp. 311-328.

Schäfer, J., Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4, Article 32.

Taylor, L., Schroeder, R., Meyer, E. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big Data & Society*, 1(2), pp. 1-10.

Víšek, J.Á. (2011). Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*, 47, pp. 179-206.

Wilson, P.W. (2018). Dimension reduction in nonparametric models of production. *European Journal of Operational Research*, 267, pp. 349-367.

Contact

Jan Kalina

The Czech Academy of Sciences, Institute of Information Theory and Automation,

Pod Vodárenskou věží 4, 182 00 Prague 8, Czech Republic

& The Czech Academy of Sciences, Institute of Computer Science,

Pod Vodárenskou věží 2, 182 00 Prague 8, Czech Republic

kalina@cs.cas.cz