www.sjm06.com

Serbian Journal of Management 12 (1) (2017) 157 - 169

Serbian Journal of Management

Review Paper

HIGH-DIMENSIONAL DATA IN ECONOMICS AND THEIR (ROBUST) ANALYSIS

Jan Kalina*

Institute of Computer Science, Czech Academy of Sciences & Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic

(Received 26 April 2016; accepted 16 September 2016)

Abstract

This work is devoted to statistical methods for the analysis of economic data with a large number of variables. The authors present a review of references documenting that such data are more and more commonly available in various theoretical and applied economic problems and their analysis can be hardly performed with standard econometric methods. The paper is focused on high-dimensional data, which have a small number of observations, and gives an overview of recently proposed methods for their analysis in the context of econometrics, particularly in the areas of dimensionality reduction, linear regression and classification analysis. Further, the performance of various methods is illustrated on a publicly available benchmark data set on credit scoring. In comparison with other authors, robust methods designed to be insensitive to the presence of outlying measurements are also used. Their strength is revealed after adding an artificial contamination by noise to the original data. In addition, the performance of various methods for a prior dimensionality reduction of the data is compared.

Keywords: Econometrics, high-dimensional data, dimensionality reduction, linear regression, classification analysis, robustness.

1. EXAMPLES OF HIGH-DIMENSIONAL DATA IN ECONOMICS

The amount of data observed in economic research and practice grows very rapidly and data are agreed to have a big potential to influence economic research as well as economic policy (Eisenstein and Lodish, 2002). Thus, economic data represent a valuable capital with an underutilized opportunity for economic decision making

* *Corresponding author:kalina@cs.cas.cz*

DOI: 10.5937/sjm12-10778

and relevance for the economic and social state of the society.

A vast number of various 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 (Belloni et al., 2013; Fan et al., 2014). A correct analysis of economic data with a large number of variables therefore becomes an emerging issue in current econometrics.

This paper is devoted to various models and methods for the analysis of high-dimensional data multivariate data. The number of variables is denoted by p and the number of measurements (observations) by n. A special attention is paid to high-dimensional data, i.e. p is assumed to be large, possibly larger than n. Often in this context, a mixture of continuous and categorical numerical variables is observed, while common statistical and data mining analysis methods are able to combine both types. Still, the analyst must be aware of the type of each variable, at least for the sake of a correct interpretation of the results.

This paper presents an overview of very recent econometric references with the aim to persuade readers about the importance of high-dimensional data for various economic tasks. As it is stressed, their analysis is far from a mere routine. Section 2 describes general challenges for the econometric analysis of high-dimensional data. Further sections 3, 4 and 5 are devoted to methods for dimensionality reduction, linear regression and classification analysis, respectively. A particular attention is paid to methods robust (insensitive) to the presence of outlying or wrongly measured observations (outliers). Section 6 compares various statistical methods on a publicly available benchmark data set with real economic data. More than that, it illustrates the performance of some recently proposed methods with promising robustness properties. Finally, Section 7 concludes the paper.

2. ANALYSIS OF HIGH-DIMENSIONAL DATA IN ECONOMICS

This section has the aim to describe challenges in the analysis of high-dimensional economic data by means of methods of multivariate statistics and data mining. The necessity of a suitable pre-processing of the data including their cleaning and exploratory data analysis is put aside. The consequent analysis of the data should be guided by the main task from the perspective of the economic problem under consideration. In any case, the analysis of high-dimensional data suffers from serious challenges, most importantly due to high dimensionality itself, but also due to outliers, dependence among variables and combination of continous and categorical variables (Fan et al., 2014). These challenges contribute to the unsuitability of standard statistical methods for high-dimensional data, which is discussed in recent economic research papers (Varian, 2014).

Specific statistical methods tailor-made for high-dimensional data have been proposed only recently, not only in statistical and econometric journals, but commonly also in journals focused on theoretical mathematics, computer science, or bioinformatics. Researchers in economics

apparently seem to be becoming aware of the trends and the arsenal of the recent methods for the analysis of high-dimensional data starts to penetrate to econometrics. However, most methods were not able to spread outside the particular community to other research fields. In addition, implementation of recent methods in commercial statistical software is running late behind the research progress.

Some important particular econometric examples will be overviewed in Sections 3-5. Most of them have the form of computationally demanding optimization tasks. However, the properties of the recently proposed methods remain to be unclear. There is a long way from theory to econometric applications on real data and there is no agreement concerning the suitability of particular methods and recommendation in literature are dramatically different from one paper to another. Often, the assumptions of individual methods or recommended sample sizes are unclear and numerous arguments in references can be described as controversial. This chaotic situation is caused by the fact that none of the available methods performs uniformly better than the others.

A specific aspect of analysing high-dimensional data is the large sensitivity of available statistical methods to the presence of outliers. Therefore, robustification is desirable for high-dimensional data as well as for classical econometric data with a small number of variables (Kalina, 2012). Outliers can be detected prior to the data analysis (Atkinson and Riani, 2004) or robust methods can be used which do not require outliers to be detected at all (Jurečková and Picek, 2012). In addition, it is also important to investigate the numerical stability of the new methods. Even if stable algorithms for their computation exist, implementations in software may be numerically unstable and therefore unreliable (Wang and Tang, 2004).

High-dimensional data can be understood as a special case of Big Data, which appear as a phenomenon in various recent commercial as well as scientific problems. Although high-dimensional data do not fulfil a rigorous definition of Big Data (e.g. Schmarzo (2013)), they cannot be analysed by standard statistical or machine learning methods because of their specific size. The importance of Big Data in economics was discussed by Taylor et al. (2014). Their analysis is a very complicated data-driven process, because the data of various types may come from a variety of very messy (unstructured) sources, including instrumental variables, high-frequency time series, free text, granular spatial or geographical data or data from social networks. Even if the data are only numerical, the very large number of observation questions the suitability of any statistical approach standing on the assumption of random samples, because e.g. nearly every hypothesis test yields a significant result. Einav and Levin (2014) presented a broader discussion of Big Data.

3. DIMENSIONALITY REDUCTION FOR REGRESSION AND CLASSIFICATION TASKS

In order to simplify the analysis of multivariate data with $n<p$ with a possibly large value of p, dimensionality reduction is generally recommended (Lee and Verleysen, 2007) 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 (Harrell, 2001).

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 (regression, classification etc.) or to include it as an intrinsic step within the (regularized) analysis.

3.1. Prior 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.

Variable selection searches for a small set of relevant variables while recommends to ignore the remaining ones. Alternative approaches searching for linear combinations of the measurements include principal component analysis (PCA), factor analysis, methods based on information theory, correspondence analysis, or multivariate scaling. Because they suffer from the presence of outliers in the data, robust versions of prior dimensionality reduction methods have been proposed. Their review from the perspective of economic applications was given by Kalina and Rensová (2015). However, PCA as well as other methods are not suitable if the aim of the analysis is classification, i.e. if the data come from two or more different groups.

An important example of supervised prior variable selection, which takes the grouping of the data into account, is the Minimum redundancy maximum relevance (MRMR) method allowing to select a subset of variables which are relevant for the classification task penalizing for correlation among the selected variables. Its robust version, denoted as MRRMRR (Minimum regularized redundancy maximum robust relevance), was proposed by Kalina and Schlenker (2015). It uses a robust correlation coefficient as a relevance measure and a shrinkage coefficient of multiple correlation as a redundancy measure. The robustness based on implicitly assigned weights to individual observations is inspired by the least weighted squares (LWS) estimator (Víšek, 2009) and the whole method remains computationally demanding.

3.2 Regularized approaches to data analysis

Standard regression or classification procedures are numerically ill-conditioned or even infeasible for $n<p$ and a suitable regularization can be described as their modification allowing to reduce the influence of some of the variables (Carrasco et al., 2007; Pourahmadi, 2013). Thus, no prior dimensionality reduction is needed. Some approaches can be described as sparse, i.e. ignoring some variables completely, which also allows a clear interpretation of the results. Examples of particular regularized methods will be given in Section 3 and 4.

4. REGRESSION FOR A LARGE NUMBER OF REGRESSORS

This section is devoted to a review of regression estimators for data with $n<p$.

The lasso estimator (least absolute shrinkage and selection operator) represents a general regularized estimation procedure suitable not only for the linear regression model, but also for nonlinear models such as logistic regression. In the standard linear regression model

$$Y_i=\beta_1 X_{1i}+\cdots+\beta_p X_{pi},\ i=1,\ldots,n, \qquad (1)$$

the lasso estimator is obtained by solving the optimization procedure

$$\min_b \sum_{i=1}^{n} u_i^2 \ \text{ subject to } \ \sum_{j=1}^{p} |b_j| \leq t \qquad (2)$$

for a specific $t>0$, where the vector of parameters $\beta=(\beta_1,\ldots,\beta_p)^T$ is estimated by $b=(b_1,\ldots,b_p)^T$. The regularization (2) in L_1-*norm* allows the method to perform variable selection simultaneously with statistical modeling (Bühlmann and van de Geer, 2011) in such a way that many variables obtain a zero estimate of the regression parameter. Thus, the lasso eliminates completely the influence of the least relevant variables and reduces the parameters of the remaining ones by shrinking towards zero. Applications of the lasso to economic modeling include the work by Belloni et al. (2013), who modelled economic growth in different countries, or the regularized (shrinkage) generalized method of moments (GMM) estimator for estimating parameters in a spatially autoregressive process (Ahrens and Bhattacharjee, 2015).

The Dantzig selector of Candes and Tao (2007) can be described as a regularized version of the regression median, penalized for violating the normal equations. Equivalently, the Dantzig selector can be defined as the estimator, which has the smallest value of the largest residual (under some additional constraints). It possesses oracle properties, i.e. the user does not have to specify the number of selected variables and the estimator is able to adapt to unknown conditions and identify a correct subset model with the optimal rate of convergence. However, because it is based on the regression median, it is highly sensitivite to violations of the assumption of normally distributed data. The Dantzig selector was used in practice e.g. to analyse data on income mobility by Liu (2014).

Other approaches suitable for regression modeling with $n<p$ include e.g. the partial least squares. Available tailor-made approaches for specific econometric models include the lasso two-stage least squares (2SLS) estimator of Zhu (2015), who proposed to compute the lasso estimator within both stages of the 2SLS estimator. Nevertheless, robust regression estimators for a large p remain to be an open problem.

5. CLASSIFICATION ANALYSIS

Multivariate data with $n<p$ will be considered which are observed in two or more different and known groups. The task of classification analysis aims at constructing a classification rule allowing to assign a new observation to one of the groups. Credit scoring (Liu et al., 2008; Lessmann et al., 2015) represents an important example of a classification task and logistic regression represents the most common tool for learning its classification rule (Greene, 2012). However, its parameters are estimated

in an unstable way for $n<p$, they have an unnecessarily large variability and therefore a lasso-type logistic regression is recommended (Belloni et al., 2015).

Standard linear discriminant analysis (LDA) is computationally infeasible for n<p, but there is a good experience with its regularized versions (Pourahmadi, 2013) based on a regularized covariance matrix in the form

where S denotes the pooled covariance

$$S^* = \lambda S + (1-\lambda)I, \quad \lambda > 0, \qquad (3)$$

matrix (across groups) and I is a unit matrix. An explicit expression for the asymptotically optimal value of λ for $p \to \infty$ was derived by Ledoit and Wolf (2003), originally for the context of financial time series. However, LDA is less used in econometric applications because of its more specific assumptions compared to logistic regression (Hastie et al., 2009). In particular, LDA is more sensitive to the presence of outliers.

Available robust versions of LDA include the MWCD-LDA (Kalina, 2012) based on the MWCD estimator of the mean and covariance matrix of multivariate data. The MWCD estimator assigns weights to individual observations with the aim to suppress outliers and was inspired by robust regression of Višek (2008) or Višek (2009). The MWCD itself, being feasible only for $n>p$, was proposed by Roelant et al. (2009). It estimates the expectation by a weighted mean with such permutation of given weights, which minimizes the determinant of the weighted covariance matrix across all permutation. At the same time, the MWCD yields an estimate the covariance matrix, which is namely equal to the weighted covariance matrix with the same permutation of the weights.

A regularized MWCD estimator for $n<p$ was proposed by Kalina et al. (2015). Starting with an initial robust estimator of the covariance matrix, which is a non-singular matrix, the determinant in each step of the iterative algorithm is regularized in the form of (3) to ensure the determinant to be non-zero. The classification rule MWCD-LDA is constructed as a version of LDA replacing standard means and covariance matrix by their (regularized) MWCD counterparts.

The machine learning methodology contains a battery of classification tools applicable also to highly correlated variables. An overview of machine learning methods for credit scoring (Liu et al., 2008) shows a good experience with support vector machines (SVM) and neural networks. Nevertheless, SVM has been criticized for high-dimensional data by other authors (Ratner, 2012). Particularly, disadvantages of SVM include the impossibility to find optimal values of the parameters, tendency to overfitting, relying on a too large number of support vectors, and unavailability of a clear interpretation of the results. Multilayer perceptrons are commonly claimed to be universal classifiers suitable for high-dimensional data from the theoretical point of view. However, their implementation is often computationally infeasible for n<p, e.g. in R statistical software with default setting of the parameters. A number of econometric papers appraises machine learning methods, but this good experience is based mainly on results of numerical simulations.

6. EXAMPLE

It is worth mentioning that various methodological econometric papers use only

simulated data (Florens and Simoni, 2012). Still, analysing real data is usually much more complicated. The authors of this paper performed an intensive research for publicly available economic data for $n<p$, but unfortunately, were not able to find such data. Therefore, a popular publicly available economic data set with $n>p$ will be now analyzed. While this data set has been repeatedly analyzed by other authors, the novelty of this work is a comparison of very recent methods including the MWCD-LDA of Kalina (2012) and MRRMRR (Kalina and Schlenker, 2015), which have not been sufficiently compared with available classification methods on real data.

The Australian Credit Approval data set from the UCI Library (Lichman, 2013) will be now analyzed. This benchmark data set for credit scoring algorithms (Lessmann et al., 2015) contains n=690 cases and 6 continuous and 8 categorical independent variables. The variables contain personal information about credit cards and their proprietors, therefore the precise meaning of the variables remains confidential. The computations are performed in *R* software package. Each variable is treated as continuous. Among the 8 categorical variables, there are 7 binary ones and one with three categories, while the third category appears only rarely. The downloaded data set does not contain missing values. Although there were missing values in the original data set, the downloaded version from Lichman (2013) has all missing values replaced by the mean of the particular variable.

The aim of the analysis is to solve the classification task into two groups, i.e. to predict if an individual client is credible or not. Several classification methods are used and the most common machine learning methods on:

(1) Raw data,

(2) The set of 3 (or 4) principal components,

(3) The set of 3 (or 4) most relevant variables selected by the MRRMRR variable selection algorithm. Here, the robust correlation coefficient with linear weights is used as the relevance measure and regularized coefficient of determination as the redundancy measure (Kalina and Schlenker, 2015).

The classification methods include logistic regression, LDA, MWCD-LDA (with linearly decreasing weights; Kalina, 2012), a neural network (a multilayer perceptron with 1 hidden layer) and SVM. In addition, a lasso-logistic regression (lasso-LR, i.e. lasso estimator in the logistic regression model) is used as an example of approaches of Section 3.2. Default values of parameters for the methods which are implemented in R software are used. It is justifiable that each of these methods is applied to a mixture of continuous and categorical data.

The results of the classification performance on the original data are listed in Table 1. There, the classification accuracy is evaluated, which is defined as the ratio of correctly classified observations to the total number of observations. If PCA is used as a prior dimensionality reduction method, a considerable loss of the classification performance is remarkable. The fourth principal component is actually quite important for the classification task and improves the classification performance by the largest amount compared to other components. The first two components seem to contribute to the separation between the groups very little. MRRMRR is superior to

Table 1. Comparison of classification accuracy of different methods with different choices of dimensionality reduction. PCA uses 3 or 4 main principal components and MRRMRR 3 or 4 most relevant variables.

Classification method	Dimensionality reduction				
		PCA		MRRMRR	
	None	3	4	3	4
Logistic regression	0.88	0.67	0.76	0.86	0.87
Lasso-LR	0.90	0.67	0.76	0.86	0.87
LDA	0.86	0.64	0.73	0.86	0.86
MWCD-LDA	0.88	0.66	0.74	0.86	0.86
Neural network	0.85	0.65	0.74	0.83	0.84
SVM	0.90	0.66	0.76	0.85	0.86

PCA and leads to reliable results already with 3 selected variables, with a negligible loss compared to the analysis over all variables jointly.

Comparing different classification methods, SVM and lasso-logistic regression yield the best results for the full data set. It is worth noting that the lasso-version of the logistic regression gives better results compared to the standard version of the logistic regression in spite of the fact that $n>p$. In this situation, a routine analyst would probably not think of using a lasso-type method. However, this superiority is lost for a reduced data set. Particularly, the lasso-version of the logistic regression does not bring any benefit compared to the standard version. Both version of the logistic regression turn out to be the best methods for the reduced data sets.

Further, an artificially generated noise is added to the data and again various classification methods are used to assign each of the observations to one of the two groups. Particularly, contamination by noise with Gaussian normal distribution is used. The computations for such contaminated data were repeated 100 times and the averaged results of the classification accuracy are given in Table 2. In an analogous way, a noise from Cauchy distribution with probability density function

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}, \tag{4}$$

was used and the results are given in Table 3.

A slight contamination by normal outliers can be recognized on the results over the whole data set. The classification accuracy obtained with PCA with 3 principal components is surprisingly improved compared to results over raw data. However, the results remain on a similar level if 4 principal components are used. To explain this, the first two principal components are strongly influenced by noise, then the third one happens to be more relevant for the classification by a mere chance and the fourth and the other principal components

Table 2. Comparison of classification accuracy of different methods with different choices of dimensionality reduction for data with continuous regressors contaminated by normal distribution $N(0,\sigma^2)$ with $\sigma=5$.

Classification method	Dimensionality reduction				
		PCA		MRRMRR	
	None	3	4	3	4
Logistic regression	0.86	0.69	0.68	0.83	0.84
Lasso-LR	0.88	0.69	0.68	0.83	0.84
LDA	0.86	0.64	0.65	0.84	0.84
MWCD-LDA	0.87	0.66	0.66	0.84	0.85
Neural network	0.84	0.67	0.66	0.81	0.83
SVM	0.88	0.68	0.69	0.83	0.85

are less relevant compared to those computed from the raw data. On the other hand, MRRMRR turns out to be able to resist the presence of normally distributed noise in the data. The choice of the best method depends on the particular dimensionality reduction method and even on the number of components or variables selected. MRRMRR still allows, like for the non-contaminated data, to classify the two groups almost as reliably as if the full data set is used.

Table 3. Comparison of classification accuracy of different methods with different choices of dimensionality reduction for data with continuous regressors contaminated by Cauchy distribution

Classification method	Dimensionality reduction				
		PCA		MRRMRR	
	None	3	4	3	4
Logistic regression	0.86	0.67	0.67	0.82	0.83
Lasso-LR	0.87	0.67	0.67	0.82	0.83
LDA	0.86	0.64	0.63	0.82	0.83
MWCD-LDA	0.87	0.66	0.70	0.84	0.85
Neural network	0.85	0.65	0.66	0.82	0.80
SVM	0.88	0.65	0.76	0.84	0.85

If the Cauchy-distributed noise contaminates the data slightly, the performance of PCA drops even more. SVM turns out to remain the most reliable classification method. MWCD-LDA seems to be as reliable as SVM in some situations or only slightly weaker, although it is not designed for non-normal contamination. Other classification methods lose their ability to classify the two groups more considerably. There seems no advantage of using the lasso-logistic regression compared to the standard logistic regression. Because the data after contamination by a Cauchy noise contain severe outliers, the results clearly document that the lasso-type regularization cannot ensure robustness with respect to outliers.

7. CONCLUSIONS

Data with a large number of variables bring a big potential and represent an important force for the development of economics (Schmarzo, 2013; Baesens, 2014). The analysis of such data sets may play an irreplacable role in the commercial sphere as well as in the contribution to the development of economic theory (Taylor et al., 2014). Important tasks of the analysis of high-dimensional data with a large number of variables and a small number of observations include linear regression, classification analysis, clustering or time series analysis.

This paper is focused on the analysis of econometric data with a large number of variables by means of linear regression and classification analysis. The analysis has to face serious challenges and requires experience as well as suitable methods of multivariate statistics and machine learning, far from just a routine or mechanical analysis by available software. Surprisingly, current econometrics textbooks are not prepared for the analysis of high-dimensional data. To a large extent, this is determined by the fact that reliable (and comprehensible) methods have been proposed only recently. Therefore, this paper presents a review of recently proposed methods applicable to economic data.

Dimensionality reduction is needed for the analysis of high-dimensional data. However, variable selection represent only one (and the standard) possibility. It allows to simplify consequent computations and sometimes also to divide variables to clusters and reduce or remove correlation among variables (Leskovec et al., 2014). On the other hand, important regularized regression or classification methods are overviewed, which reduce the dimensionality within their computation. Although regularization has been denoted as a trick for econometricians (Varian, 2014), it can be described as a useful tool which deserves a wider attention in the analysis of economic data sets. In addition, attention to methods which are robust to outliers is paid. These are however available only for dimensionality reduction and classification analysis but not linear regression.

In addition to the review, the performance of important available methods is illustrated on a public benchmark data set with real economic data. In the examples, the authors use their own robust methods from previous papers (MWCD-LDA, MRRMRR).

In this paper, various standard as well as recently proposed methods are used to analyse a data set on credit scoring. The analysis of the original data shows that the MRRMRR variable selection allows to reduce the dimensionality of the data in a

much more effective way compared to the standard PCA. This is caused by the fact that PCA does not take the grouping of the data into account and remains puzzled. The recently proposed MWCD-LDA seems to be superior to the standard LDA, which is an argument in its favour. MWCD-LDA is able to overcome standard machine learning algorithms on this data set, but reaches their performance after a prior dimensionality reduction. It also turns out that a lasso-regularized regression model, namely the lasso-logistic regression, performs very reliably, which justifies its common usage in various sorts of classification tasks.

Finally, the results of the analysis of the credit scoring data after their artifical contamination by noise will be discussed. MWCD-LDA appears to be reasonably robust and attains a comparable classification accuracy with SVM. These two methods are recommendable for both normal and Cauchy noise. Other methods turn out to be more depending on the distribution of the data and less resistant to data contamination. Besides, MRRMRR is able to clearly outperform the PCA, which is disoriented on the data set containing two groups of observations. The contribution of the examples can be summarized by saying that arguments in favor of the methods MWCD-LDA and MRRMRR are found, which were both recently proposed as promising methods for high-dimensional data.

Acknowledgements

This work is supported by the Czech Science Foundation grants 13-01930S and 17-07384S.

ВИШЕ-ДИМЕНЗИОНИ ПОДАЦИ У ЕКОНОМИЈИ И ЊИХОВА РОБУСТНА АНАЛИЗА

Jan Kalina

Извод

Овај рад је посвећен статистичким методама анализе економских података уз велики број промењивих. Аутори представљају преглед референци, чиме документују да су такви подаци све чешће доступни у различитим проблемима теоријске и примењене економије и да се њихова анализа тешко може учинити применом стандардних економетрисјких метода. Рад је фокусиран на вишедимензионе податке, који имају мали број посматрања, и дају преглед недавно предложених метода за њихову анализу у контексту економетрије, у првом реду у области смањења димензионалности, линеране регресије и класификационој анализи. Даље, перформансе различитих метода су представљене на јавно доступним подацима за бенчмаркинг, у виду сета података о рангирању кредита. У циљу поређења са другим ауторима, коришћени су такође робустни методи, који немају осетљивост на утицај екстрема у мерењима. Њихова снага је оцењена тек додавањем вештачке контаминације путем сигнала шума уведеног у полазне податке. Као додатак, поређене су перформансе различитих метода за претходну редукцију димензионалности.

Кључне речи: Економетрија, више-димензини подаци, смањење димензионалности, линеарна регресија, класификациона анализа, робустност

References

Ahrens, A., & Bhattacharjee, A. (2015). Two-step lasso estimation of the spatial weights matrix. Econometrics, 3, 128-155.

Atkinson A., & Riani M. (2004). Exploring multivariate data with the forward search. New York, NY, USA: Springer.

Baesens, B. (2014). Analytics in a big data world. New York, NY, USA: Wiley.

Belloni, A., Chernozhukov, V., & Hansen, C.B. (2013). Inference for high-dimensional sparse econometric models. In Acemoglu, D., Arellano, M., & Dekel, E. (Eds.), Advances in Economics and Econometrics, 10th World Congress, Vol. 3. Cambridge, UK: Cambridge University Press.

Belloni, A., Chernozhukov, V., & Wei, Y. (2015). Honest confidence regions for a regression parameter in logistic regression with a large number of controls. Available: http://arxiv.org/abs/1304.3969 (February 20, 2016).

Buhlmann, P., & van de Geer, S. (2011). Statistics for high-dimensional data. Berlin, Germany: Springer.

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. Annals of Statistics, 35, 2313-2351.

Carrasco, M., Florens, J.-P., & Renault, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. Pp. 5633-5751 in Handbook of Econometrics, Volume 6, Part B.

Einav, L., & Levin, J.D. (2013). The data revolution and economic analysis. NBER working paper No. 19035.

Eisenstein, E.M., & Lodish, L.M. (2002). Marketing decision support and intelligent systems: Precisely worthwhile or vaguely worthless? Pp. 436-454 in Weitz B.A., Wensley R. (Eds.), Handbook of marketing. London, UK: SAGE.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National Science Review, 1, 293-314.

Florens, J.-P., & Simoni, A. (2012). Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized prior. Journal of Econometrics, Vol. 170, 458-475.

Greene, W.H. (2012). Econometric Analysis. 7th edn. Harlow, UK: Pearson Education Limited.

Harrell, F.E. (2001). Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York, NY, USA: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Data mining, inference, and prediction. New York, NY, USA: Springer.

Jurečková, J., Picek, J. (2012). Methodology in robust and nonparametric statistics. Boca Raton, FL, USA: CRC Press.

Kalina, J. (2012). On multivariate methods in robust econometrics. Prague Economic Papers, 21, 69-82.

Kalina, J., & Rensova, D. (2015). How to reduce dimensionality of data: Robustness point of view. Serbian Journal of Management, 10, 131-140.

Kalina, J., & Schlenker, A. (2015). A robust and regularized supervised variable selection. BioMed Research International, Article 320385.

Kalina, J., Schlenker, A., & Kutílek, P. (2015). Highly robust analysis of keystroke dynamics measurements. Pp. 133-138 in Proceedings SAMI 2015, 13th International Symposium on Applied Machine Learning Intelligence and Informatics. Budapest, Hungary: IEEE.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance, 10, 603-621.

Lee, J.A., & Verleysen, M. (2007). Nonlinear dimensionality reduction. New York, NY, USA: Springer.

Leskovec, J., Rajaraman, A., & Ullman, J. (2014). Mining of massive datasets, 2nd edn. Cambridge, UK: Cambridge University Press.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247, 124-136.

Lichman, M. (2013). UCI Machine Learning Repository. Available: http://archive.ics.uci. edu/ml (February 20, 2016). Irvine, CA, USA: University of California.

Liu, B., Yuan, B., & Liu, W. (2008). Classification and dimension reduction in bank credit scoring system. Lecture Notes in Computer Science, 5263, 531-538.

Liu, D. (2014). Essays in theoretical and applied econometrics. Montreal, Canada: Concordia University.

Pourahmadi, M. (2013). High-dimensional covariance estimation. Hoboken, NJ, USA: Wiley.

Ratner, B. (2012). Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data, 2nd edn. Boca Raton, FL, USA: CRC Press.

Roelant, E., Van Aelst, S., & Willems, G. (2009). The minimum weighted covariance determinant estimator. Metrika, 70, 177-204.

Schmarzo, B. (2013). Big data: Understanding how data powers big business. New York, NY, USA: Wiley.

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, 1-10.

Varian, H.R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28, 3-28.

Víšek, J.Á. (2008). The implicit weighting of GMM estimator. Bulletin of the Czech Econometric Society, 15, 3-29.

Víšek, J.Á. (2009). The least weighted squares I. The asymptotic linearity of normal equations. Bulletin of the Czech Econometric Society, 15, 31-58.

Wang, X., & Tang, X. (2004). Experimental study on multiple LDA classifier combination for high dimensional data classification. Lecture Notes in Computer Science, Vol. 3077, 344-353.

Zhu, Y. (2015). Sparse linear models and l1-regularized 2SLS with high-dimensional endogenous regressors and instruments. Available: http://arxiv.org/pdf/1309.4193 (February 20, 2016).