REGRESSION FOR HIGH-DIMENSIONAL DATA: FROM REGULARIZATION TO DEEP LEARNING

Jan Kalina – Petra Vidnerová

Abstract

In this paper, tools for regression modeling suitable for high-dimensional economic data are presented and discussed, including robust regularized linear regression estimators, regularization networks, and tools of deep learning. We discuss here that the analysis of high-dimensional data, so practical for specific econometric applications, requires conceptually novel tools, including robust regularized neural networks, allowing to down-weight the influence of outliers in the data, i.e. require more intricate tools compared to big data analysis. While deep learning tools do not converge for high-dimensional data, robust and regularized methods available for linear regression have not been extended to the nonlinear model yet. We model here the travel and tourism competitiveness index as a response variable explained by several tourism infrastructure characteristics to illustrate several robust tools for regression analysis. The results of the robust regularized methods are not harmed by outliers in the dataset, while it is beneficial that they (just like non-robust regularized methods) allow to order the variables according to their significance for the regression fit.

Key words: regression, neural networks, robustness, high-dimensional data, regularization

JEL Code: C45, C14, C63

Introduction

Regression modeling is well known as a fundamental task in current econometrics. However, classical estimation tools for the linear regression model are not applicable to high-dimensional data. Although there is not an agreeement about a formal definition of high-dimensional data, usually these are understood either as data with the number of variables *p* exceeding (possibly largely) the number of observations *n,* or as data with a large *p* in the order of (at least) thousands. In both situations, which appear in various field including econometrics, the analysis of the data is difficult due to the so-called curse of dimensionality (cf. Kalina (2013) for discussion). Compared to linear regression, nonlinear regression modeling with an unknown shape of the relationship of the response on the regressors requires even more intricate methods.

The aim of this paper is to overview available approaches to regression for data with a large p, including regularized linear methods (Section 1) or regularized artificial neural networks (Section 2) for nonlinear regression, underutilized in economic applications so far. We include a systematic overview of the importance of the topic of high-dimensional regression and big data analysis in econometrics (Section 3). An analysis of a tourism dataset by means of (possibly) robust regularized regression estimates is presented in Section 4, illustrating the ability to arrange variables according to their relevance (i.e. variable selection) to be their main advantageous property.

1 Robust regularized regression

In Sections 1 and 2, we consider the standard linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \qquad i = 1, \ldots, n, \tag{1}$$

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 task is to estimate the regression parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$. This section presents an overview of methods suitable for the situation with $n < p$.

The lasso estimator or the ridge regression are known tools for high-dimensional regression in (1) presented e.g. in Hastie et al. (2015). They represent penalized versions of the least squares estimator, while the lasso considers a regularization in the L_1-norm and the ridge regression in the L_2-norm. None of the two methods is however robust with respect to severely outlying measurements (outliers) in the data. The lasso (but not the ridge regression) is able to yield a sparse solution, ignoring the redundant variables completely. The computation of these regularized regression estimates is available in packages sparsereg or hdm of R software.

Recently, several robust regularized regression estimates have been proposed. They require all the regressors to be continuous. We can say that connecting robustness and regularization in the regression context is more straightforward compared to the difficulties encountered in the classification task. Nevertheless, the available robust regularized estimators have been proposed mainly for biomedical regression modeling, while high-dimensional applications in econometrics are much less frequent.

Particular examples of robust regularized regression estimators include the sparse least trimmed squares (LTS) estimator of Alfons et al. (2013), obtained by regularization of the least trimmed squares (LTS) estimator. The so-called sparse partial robust M regression (SPRM) of Hoffmann et al. (2015) is related to (robust) partial least squares (PLS) estimator for a multivariate response. A penalized quasi-likelihood M-estimator proposed by Avella-Medina and Ronchetti (2018) is also applicable to generalized linear models. Cohen-Freue et al. (2019) proposed a robust version of the elastic net estimator, combining the lasso and ridge regression estimates in an optimized way. Smucler and Yohai (2017) proposed penalized MM-estimators with oracle properties, allowing to identify correctly the set of relevant variables among the set of all variables, i.e. allowing the estimator to optimally extract the knowledge under the presence of a possibly large set of redundant variables.

Implementations of some of the available robust regularized regression estimators are available in packages robustHD, sparseLTSEigen, or sprm of R software. Extensions of the linear model (1) to the instrumental variables (IV) estimator for $n < p$ have been investigated (again) only for biomedical applications. Regularized version of the instrumental variables estimator, suitable for high-dimensional economic data, seem to remain an open problem, together with extensions to regularized versions of the generalized method of moments (GMM).

2 Regression neural networks

While neural networks are important tools for nonlinear regression modeling, only regularized neural networks are suitable for high-dimensional data. In this section, we recall the so-called regularization networks, also known as the generalized ridge estimators, and describe their connection to (much more popular) radial basis function networks.

2.1 Model

We consider the regression task to model a continuous response $Y_1, \dots, Y_n$ by means of p independent variables (regressors, features) available for n observations (measurements, instances), where the values for the i-th observation ($i = 1, \dots, n$) are denoted as $X_{i1}, \dots, X_{ip}$. All regressors have to be continuous. Instead of (1), neural networks are formulated for the nonlinear regression model

$$Y_i = f(X_i) + e_i, \qquad i = 1, \dots, n, \tag{2}$$

where the shape of the function f is unknown. Neural networks seem to be still underutilized tools in econometric modeling; for example it is worth noting that (even plain neural networks) are not described in monographs on nonparametric econometrics. Recent applications of neural networks in economics include e.g. the study of Jang and Lee (2018), who considered Bayesian neural networks for predicting time series of Bitcoin processes, or the application of Livieris (2019) of recurrent neural networks to forecasting results of marketing campaigns or forecasting in the context of credit approval.

2.2 Radial basis function neural networks

Radial basis function (RBF) networks represent an important class of feedforward neural networks for $n > p$. They contain an input layer with p inputs, a single hidden layer with the total number N of RBF units (neurons), and a linear output layer. The user chooses N together with a radially symmetric function denoted here as ρ. Denoting the Gaussian density as ρ, the residuals of the RBF network can be expressed as

$$u_i = Y_i - \sum_{j=1}^{N} a_j \rho\left(\left\|X_i - c_j\right\|\right), \qquad i = 1, \dots, n, \tag{3}$$

with parameters $c_1, \dots, c_N \in \mathbb{R}^p$ and $a_1, \dots, a_N \in \mathbb{R}$, and possibly with other parameters corresponding to ρ. Parameters of RBF networks, just like those of MLPs, are typically found by means of minimizing the sum of squared residuals.

2.3 Regularization networks

Regularized versions of various types of neural networks have been available, which are suitable for data with $n < p$. Here, we describe the so-called regularization networks of Girosi et al. (1995), which remain relatively little known. Other approaches, including penalized versions of multilayer perceptrons, are popular but more complicated from the computational point of view. For example, regularized multilayer perceptrons has to be solved by means of a gradient approach to nonlinear optimization. However, it is necessary to admit that the regularization itself does not improve the robustness of the neural networks with respect to outlying measurements (Kalina and Vidnerová, 2019). In addition, there is none of the available regularized versions of neural networks implemented in R statistical software.

In the search for the regression function (2), a naïve approach

$$\min_f \sum_{i=1}^{n} (Y_i - f(X_i))^2 \tag{4}$$

is replaced by

$$\min_{f} \{\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \lambda\|f\|_K\} \tag{5}$$

where

$$\|f\|_K = \sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\beta_j K(X_i, X_j) \tag{6}$$

and K is a selected kernel. The theory of reproducible kernel Hilbert space (RKHS), as a part of functional analysis, allows to measure the distance between p-dimensional vectors by means of a selected kernel K; see Hastie et al. (2015). Commonly, the user choose K as the density of a normal distribution with expectation 0 and some fixed variance σ^2.

The solution of (5) can be derived to have the form

$$\hat{f}(x) = \sum_{i=1}^{n}\beta_i K(x, X_i), \quad x \in \mathbb{R}^p, \tag{7}$$

which of course depends on $\beta = (\beta_1, \dots, \beta_n)^T$. It is possible to find an explicit form of an estimate of β, while the task (5) is not ill-posed any more. Using now the notation

$$K = \left(K(X_i, X_j)\right)_{i,j=1}^{n}, \tag{8}$$

the task (5) can be expressed by means of the symmetric square matrix K as a penalized version of the system of normal equations

$$\min_{f} \{\sum_{i=1}^{n}\|Y - K\beta\|^2 + \lambda\beta^T K\beta\}. \tag{9}$$

By means of derivatives we find that the minimum is obtained for

$$\hat{\beta} = (K^T K + \lambda K)^{-1}K^T Y = [(K + \lambda I)K]^{-1}K^T Y = (K + \lambda I)^{-1}Y. \tag{10}$$

This estimate $\hat{\beta}$ is commonly denoted as a regularization network or generalized ridge esetimator, where the latter evokes a connection to the ridge regression (see Section 1). Particularly, the main diagonal in (10) is regularized (shifted) within the task of computing the inverse of $K + \lambda I$. Finally, the fitted value of the response for a given $x \in \mathbb{R}^p$ is obtained by means of an empirical counterpart of (7), i.e.

$$\hat{f}(x) = \sum_{j=1}^{n}\hat{\beta}_j K(x, X_j), \quad x \in \mathbb{R}^p. \tag{11}$$

Formulas (7) and (11) reveal the connection or regularization networks to RBF networks, i.e. both correspond to an RBF network with a particular choice (8) of the kernel chosen in (5).

3 Deep learning

The potential of big economic data seems to have been acknowledged, at least in theoretical papers, and deep learning represents the methodology suitable for their analysis. The computer science community has attempted to find deep learning applications in econometrics and to draw attention of econometricians. However, little attention has been paid to the important question in which economic applications big data emerge and which tasks are the most useful ones, leading to using particular big data analysis tools. Big data, as repeatedly defined in the literature, require (among others) a large number *n* of measurements. It is clear that big data can be hardly obtained in economics by means of standard ways and throughout all branches of economic applications, e.g. in management in commercial companies. Specific non-traditional sources of big economic (and social) data were described by Blazques and Domenech (2018) as the internet, social networks, and urban or mobile sensors. In these specific domains, the analysis of big data is of course useful. On the whole, however, we have to say that it is typically not so easy to obtain big data in economics compared to some other disciplines (including natural sciences or medicine).

Deep learning is a broad concept encompassing a variety of particular machine learning tools for the analysis of big data, popular in current computer science. Tools for analyzing big economic data were overviewed e.g. by Varian (2014), with a focus on dimensionality reduction techniques, however without specifying potential sources of big economic data. Convolutional neural networks represent one important class of deep learning tools, suitable especially for data in the form of images; they exploit specific heuristics including dimensionality reduction in a specific form, which is unsuitable for economic regression data. Recurrent neural networks are another class within the deep learning methodology, suitable for non-stationary economic time series, which are long and which are at the same time modeled by a large number of regressors. Deep networks obtained as direct deep analogies of available neural networks (e.g. multilayer perceptrons) are conceptually analogous to standard (shallow) approaches and bring difficulties only from the computational point of view. They can be robustified by the same approaches as the standard (shallow) networks, e.g. using the approach of Kalina and Vidnerová (2019), only with a need for more efficient algorithms.

We can say that the analysis of big (numerical) data does not bring conceptually new challenges, as it requires only computationally more efficient algorithms, while the analysis of high-dimensional data requires new approaches and presents a real challenge for econometricians. Clearly, high-dimensional data cannot be analyzed by deep learning tools, which require a large n and can hardly converge to the solution for data with $n < p$.

It is already evident that we have to strictly distinguish between big data and high-dimensional data. Concerning the importance of high-dimensional regression in economics, the task with high-dimensional data seems useful again only in specific tasks, e.g. in the analysis of panel data with a large number of panels, or portfolio optimization. In other areas, such as macroeconomic modeling, high-dimensional data are much less to be encountered. Also in management science, other tools (e.g. of game theory, operations research or classification analysis), especially for a small p, seem to be much more common. From the methodological point of view, dimensionality reduction (i.e. prior to performing the regression modeling) is not needed. Not even a (possibly highly) robust dimensionality reduction (e.g. that of Kalina and Schlenker (2015)) is needed, if regularized approaches ensuring a sparse solution are used.

4 Example

Our aim is to compare the regression methods described in this paper, namely robust regularized regression estimators and regularization neural networks. However, it is difficult to find a publicly available high-dimensional economic dataset, as we were actually not able to find in publicly available repositories, and it remains difficult to simulate realistic data with $n < p$. Thus, we take resort to a regression dataset with $n > p$. The Travel and Tourism Competitiveness Index (TTCI) dataset with $p = 12$ and $n = 141$, which was previously analyzed by (robust) linear regression (especially regression quantiles) in Kalina et al. (2019), is analyzed here. The tourist service infrastructure measured across 141 countries is modeled as a response of 12 characteristics of TTCI, while all variables come from the year 2015. All the 12 regressors are continuous random variables. Web perform all the computations in R software.

We present the results of the least squares, lasso, and LTS-lasso, where the last considers the default value of trimming. Table 1 presents values of MSE and trimmed MSE (TMSE) computed by means of various regression methods. TMSE is formally defined as

$$TMSE(\alpha) = \frac{1}{h}\sum_{i=1}^{h} r_{(i)}^2, \qquad (12)$$

where h is integer part of αn, $\alpha \in [0.5,1)$ is a fixed constant (ensuring $n/2 \leq h \leq n$), and squared prediction errors are arranged as $r_{(1)}^2 \leq \cdots \leq r_{(n)}^2$. We use $\alpha = 3/4$ here.

Further, we performed the variable selection by means of *t*-tests for the least squares estimator; the backward selection yields 5 significant variables here. These are reported in Table 2. None of the two regularized estimators is able to induce sparsity here, as the dataset is too small, and the final model uses all the 12 regressors. As the lasso and LTS-lasso perform ordering of the variables with respect to their importance for the regression task (i.e. for explaining the response), Table 2 presents the variables arranged according to this relevance. We do not present deep learning results here, because it was revealed as highly unstable for this dataset, particularly because *n* is insufficient for using deep learning at all.

Tab. 1: Prediction error measures evaluated for three regression estimators over the TTCI dataset

	MSE	TMSE
Least squares	0.449	0.160
Lasso	0.461	0.162
LTS-lasso	0.453	0.161

Source: own computation

Tab. 2: Significant variables in the TTCI dataset arranged according to their relevance in the regression fit, starting from the most significant to the least significant

Least squares	6, 5, 9, 1, 10
Lasso	6, 5, 9, 1, 3, 10, 4, 7, 2, 11, 12, 8
LTS-lasso	5, 6, 3, 10, 9, 7, 8, 11, 1, 12, 4, 2

Source: own computation

Conclusions

This paper recalls regression estimators suitable for high-dimensional econometric modeling, namely linear methods (robust regularized regression estimators) as well as nonlinear ones (regularized neural networks). Robust regularized regression estimator for high-dimensional data allow to select a subset of the relevant measurements and to distinguish between relevant and redundant regressors in the model. Still, their assumption of the model to be linear is a serious limitation, although the approximation of the real relationship of the response on the regressors by a linear model is often (but not always) perceived as a reasonable approximation in economic applications.

Regularization networks are more flexible; they require to find a suitable architecture, which can be performed by computational approaches (cross validation). In the analysis of a nonlinear trend, regularization networks are more suitable than (standard or robust) regularized regression estimators, as they are able to perform nonlinear modeling. We formulate open questions in the area of regularization networks and deep learning, namely the need to improve their robustness to outliers. However, regularization networks do not contain as many additional characteristics as lasso or LTS-lasso; see Kalina et al. (2019) for a more detailed analysis and statistical interpretation of the tourism dataset. Here, robust regularized regression estimates are used to illustrate the ability to arrange variables according to their relevance (i.e. variable selection) to be one of their main advantageous properties. The numerical results of the analysis of the TTCI dataset do not show many differences between robust and non-robust methods, which indicates the dataset not to contain severe outliers.

The regularized networks depend on all the available measurements and therefore do not possess a global robustness. Our work may also be perceived as preparation for our future study of robust regularized neural networks, which seem to be still missing in the literature. Particularly, we would like to propose inter-quantile estimates by means of regularized neural networks, or to use regularized neural networks with the loss function of the (highly robust) least weighted squares estimator.

Acknowledgment

The work was supported by the projects 19-05704S and 18-23827S of the Czech Science Foundation. Thanks to Eva Litavcová and Petra Vašaničová for providing the TTCI dataset.

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), 226-248

Avella-Medina, M. & Ronchetti, M. (2018). Robust and consistent variable selection in high-dimensional generalized linear models. Biometrika, 105(1), 31-44

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

Cohen-Freue, G., Kepplinger, D., Salibián-Barrera, M., & Smucler, E. (2019). Robust elastic net estimators for variable selection and identification of proteomic biomarkers. Annals of Applied Statistics, 13(4), 2065-2090

Girosi, F., Jones, M., & Poggio, T. (1995). Regularization theory and neural networks architectures. Neural Computation, 7(2), 219-269

Hastie T., Tibshirani R., & Wainwright M. (2015). Statistical learning with sparsity. The lasso and generalizations. CRC Press, Boca Raton

Hoffmann, I., Serneels, S., Filzmoser, P. & Croux, C. (2015). Sparse partial robust M regression. Chemometrics and Intelligent Laboratory Systems, 149, 50-59

Jang, H. & Lee, J. (2018). An empirical study on modeling and prediction of bitcoin prices with Bayesian neural networks based on blockchain information. IEEE Access, 6, 5427-5437

Kalina, J. (2013). Highly robust methods in data mining. Serbian Journal of Management, 8(1), 9-24

Kalina J., Schlenker A. (2015). A robust supervised variable selection for noisy high-dimensional data. BioMed Research International, 2015, 320385

Kalina, J., Vašaničová, P., & Litavcová, E. (2019). Regression quantiles under heteroscedasticity and multicollinearity: Analysis of travel and tourism competitiveness. Ekonomický časopis, 67(1), 69-85

Kalina, J. & Vidnerová, P. (2019). Robust training of radial basis function neural networks. Lecture Notes in Artificial Intelligence, 11508, 113-124

Livieris, I.E. (2019). Forecasting economy-related data utilizing weight-constrained recurrent neural networks. Algorithms, 12, 85

Smucler, E. & Yohai, V.J. (2017). Robust and sparse estimators for linear regression models. Computational Statistics and Data Analysis, 111, 116-130

Varian, H.R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28

Contact

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
kalina@cs.cas.cz

Petra Vidnerová
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
petra@cs.cas.cz