SOME ROBUST ESTIMATION TOOLS FOR MULTIVARIATE MODELS

Jan Kalina

Abstract

Standard procedures of multivariate statistics and data mining for the analysis of multivariate data are known to be vulnerable to the presence of outlying and/or highly influential observations. This paper has the aim to propose and investigate specific approaches for two situations. First, we consider clustering of categorical data. While attention has been paid to sensitivity of standard statistical and data mining methods for categorical data only recently, we aim at modifying standard distance measures between clusters of such data. This allows us to propose a hierarchical agglomerative cluster analysis for two-way contingency tables with a large number of categories, based on a regularized measure of distance between two contingency tables. Such proposal improves the robustness to the presence of measurement errors for categorical data. As a second problem, we investigate the nonlinear version of the least weighted squares regression for data with a continuous response. Our aim is to propose an efficient algorithm for the least weighted squares estimator, which is formulated in a general way applicable to both linear and nonlinear regression. Our numerical study reveals the computational aspects of the algorithm and brings arguments in favor of its credibility.

Key words: robust data mining, high-dimensional data, cluster analysis, outliers

JEL Code: C13, C14, C55

1 Introduction

Numerous procedures of multivariate statistics and data mining common for the analysis of multivariate data in economic applications are known to be sensitive to the presence of outlying and/or highly influential observations (Martinez et al., 2011). Therefore, their robust counterparts are highly desirable. Indeed, a variety of results on robust data mining or robust econometrics has been published recently (Filzmoser & Todorov, 2011; Belloni et al., 2014).

Also in the context of categorical data, standard procedures for hypothesis tests, association measures, clustering or classification trees are vulnerable to misclassified data, i.e. the methods are biased under a contaminated model, while outliers appear in discrete data quite commonly, e.g. in the form of measurement errors (Buonaccorsi, 2011). Also the simplest tools (chi-square or likelihood-ratio statistics) are heavily influenced by the presence of structural or sampling zero counts (Agresti, 2002; Neykov et al., 2014). Still, only small attention has been paid to robust methods for categorical data. In Section 2, we derive a regularized distance measure between two contingency tables with a large number of categories, which is applicable e.g. to cluster analysis.

Classical regression estimatorsfor data with a continous response suffer from the presence of outlying data (outliers) and a variety of robust statistical methods has been developed (Gentle et al., 2012). Some of them can be considered as reliable self-standing procedures suppressing the effect of data contamination. In Section 3, we propose an efficient algorithm for computation of the least weighted squares for a general context for both linear and nonlinear regression model. We discuss the choice of parameters for this algorithm.

2 Regularized cluster analysis for contingency tables

Cluster analysis represents a general information extraction methodology allowing to reveal the multivariate structure of given data and to divide multivariate data to subpopulations (Martinez et al., 2011; Kalina, 2012). It is often used as an exploratory technique for complex multivariate data and can be interpreted as an unsupervised dimensionality reduction technique. The statistical methodology however contains a gap of multivariate approaches for high-dimensional data, which are robust to the presence of noise (Kalina & Zvárová, 2013).

This section proposes a robust measure of distance between two contingency tables, which may be used as a tool within hierarchical agglomerative cluster analysis for contingency tables with a large number of categories. Cluster analysis for such tables can be performed by standard algorithms, replacing habitually used distance measures by their regularized counterparts. The result is based on a regularized measure of distance between observations from a multinomial distribution.

Tab. 1:Notation for the observed counts (example of Section 2)

	Product 1	Product 2	⋯	Product J	$\sum$
Satisfied	n_{11}	n_{12}	⋯	n_{1J}	$n_{1.}$
Unsatisfied	n_{21}	n_{22}	⋯	n_{2J}	$n_{2.}$
$\sum$	$n_{.1}$	$n_{.2}$	⋯	$n_{.J}$	n

Source: Agresti (2013)

Tab. 2: Notation for the observed counts in the k-th stratum (example of Section 2)

	Product 1	Product 2	⋯	Product J	$\sum$
Satisfied	n_{11k}	n_{12k}	⋯	n_{1Jk}	$n_{1.k}$
Unsatisfied	n_{21k}	n_{22k}	⋯	n_{2Jk}	$n_{2.k}$
$\sum$	$n_{.1k}$	$n_{.2k}$	⋯	$n_{.Jk}$	$n_{..k}$

Source: Agresti (2013)

We explain our model on an marketing example. A study of customer satisfaction has been performed with a large number of products *J*. The counts form a contingency table of size $2\,x\,J$ (Table 1). However, the study is performed on *K* different places (sub-populations, strata) and we expect the measurements to vary among different strata. The aim of the analysis is to perform a cluster analysis to find clusters of places. Thus, the observed data have the form of a set of K contingency tables $2\,x\,J$. In the k-th stratum, the observed data have the form of a contingency table, which is shown in Table 2. We do not work with Table 1, but with the total number of K tables with the form of Table 2.

Each of the counts n_{ijk} for $i\ \in \{1,2\}$, $j = 1, \dots, J$, and $k = 1, \dots, K$, represents a realization of a random variable N_{ijk}, which follows a binomial distribution. We assume the probability of success is the same in each stratum. Let π_j denote the probability of success in the *j*-th column of the table (across strata).

Cluster analysis with the task to find clusters of products (groups) among the total number of *J* products can be interpreted as a dimensionality reduction applied to the space of columns of the contingency table. It requires to measure the similarity between the contingency table corresponding to the k-th stratum and an analogous table corresponding to the k'-th stratum. Because of the large value of *J*, various such measures (e.g. phi coefficient) should be replaced by their regularized counterparts, based on regularized versions of χ^2 or G^2 test statistics.

A suitable regularization is known as a tool ensuring robustness in the context of categorical data (Hastie et al., 2008).We propose to regularize the probability p_j of the success in the *j*-th column in the form

$$p_j^* = (1-\lambda)p_j + \lambda\frac{n_{1..}}{n_{...}}, \qquad (1)$$

where p_j is the maximum likelihood estimate of success in the *j*-th column across observations and $\lambda \in (0,1)$ is a parameter. Here we use the notation

$$n_{...} = \sum_{k=1}^{K} n_{..k} \quad \text{and} \quad n_{1..} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{1jk}. \qquad (2)$$

Particularly, p_j^* can be expressed as

$$p_j^* = (1-\lambda)\frac{n_{1j.}}{n_{.j.}} + \lambda\frac{n_{1..}}{n_{...}}. \qquad (3)$$

Our main result is the optimal value of the regularization parameter, which minimizes the mean square error of (3) over all $\lambda \in (0,1)$. While the majority of regularized data mining methods relies on a cross validation in order to assess a suitable value of the parameter, we derive the optimal value of λ under the assumption that $J \rightarrow \infty$.

We propose to estimate λ by

$$min\{\lambda^*, 1\}, \qquad (4)$$

where the solution of the minimization of the mean square error has an explicit expression as

$$\lambda^* = \frac{1-\sum_{j=1}^{J}\left(\frac{n_{1j.}}{n_{.j.}}\right)^2}{(n-1)\sum_{j=1}^{J}\left(\frac{n_{1..}}{n_{...}}-\frac{n_{1j.}}{n_{.j.}}\right)^2}. \qquad (5)$$

The proof is based on the idea of asymptotically optimal value of the regularization parameter for financial time series of (Ledoit & Wolf, 2003). Nevertheless, the form of (4) ensures the value of the regularization parameter to be bounded between 0 and 1.

Now, we can express the regularized versions of the $\chi^2$and G^2 statistics as

$$\chi^{*2} = \sum_{j=1}^{J}\sum_{k=1}^{K}\left[\frac{\left(n_{1jk}-n_{.jk}p_j^*\right)^2}{n_{.jk}p_j^*} + \frac{\left(n_{2jk}-n_{.1k}(1-p_j^*)\right)^2}{n_{.jk}(1-p_j^*)}\right] \qquad (6)$$

and

$$G^{*2} = \sum_{j=1}^{J}\sum_{k=1}^{K}\left[n_{1jk}log\frac{n_{1jk}}{n_{.jk}p_j^*} + n_{2jk}log\frac{n_{2jk}}{n_{.jk}(1-p_j^*)}\right]. \qquad (7)$$

As a consequence, it is possible to define a regularized phi coefficient

$$\varphi^* = \sqrt{\frac{\chi^{*2}}{n}} \qquad (8)$$

as a measure of distance between two contingency tables. Thus, the cluster analysis can be performed by one of existing algorithms according to a selected linkage criterion. The regularized phi coefficient is not only the distance between two observations, but plays the role of a distance between two clusters. Other distance measures may be defined based on the G^{*2} statistic (Agresti, 2013).

To summarize this section, a tailor-made approach for the analysis of categorical data with a large number of categories and small observed samples is proposed and an asymptotic result (4) is derived. The method can be interpreted as robust to measurement errors. On the other hand, it assumes the probabilities $p_1, \ldots, p_J$ to be relatively homogeneous in order to justify a biased estimation of the probability in a particular category by means of borrowing information in (1) across all categories. The regularized association measures may be used within clustering algorithms e.g. in strategic management, credit risk management or instrumental variables estimation (Belloni et al., 2014), where a large dimensionality is commonly encountered.

3 Least weighted squares estimator for linear and nonlinear models

3.1 A general algorithm

The least weighted squares (LWS) estimator (Víšek, 2011) and its analogy for nonlinear regression, which can be denoted as the nonlinear least weighted squares (NLWS) estimator, can be described as one a few existing robust regression methods with a high breakdown point, i.e. with a high resistance (insensitivity) against outlying measurements in the data (Gentle et al., 2012). A general algorithm jointly for the LWS in a linear model and for the NLWS estimator in nonlinear regression is proposed in this section. Further, the performance of the NLWS estimator is shown on real data and we study computational aspects of the algorithm.

The LWS estimator has appealing properties like other statistical methods based on ranks of observations (Saleh et al., 2012). It has asymptotically a 100 % efficiencyof the least squares under Gaussian errors.Itsrelative efficiency was declared to be high based on numerical simulations (Víšek, 2011), compared to maximum likelihood estimatorsunder various distributional models. Extensions of the idea of implicit weights assigned to individual observations to other models (e.g. robust correlation coefficient or robust principal component analysis) turn out to yield promising results (Kalina, 2012).

The nonlinear least weighted squares (NLWS) regression estimator is based on implicit weights assigned to individual observations. The arguments for the high robustness of the LWS estimator with respect to the presence of outliers and to heteroscedasticity are valid also for the NLWS estimator.

Now, we will propose a general algorithm for an efficient computation of the LWS for both linear and nonlinear regression models. The aim is the estimation of the parameter $(\beta_1, \dots, \beta_p)^T$. We denote residuals by $(u_1, \dots, u_n)^T$. For a given estimate b of β, the residual corresponding to the i-th observation will be denoted as $u_i(b)$. The loss function for a known value of b is the value of $\sum_{i=1}^{n} w_i u_{(i)}^2(b)$.

Algorithm 1 (Least weighted squares estimator for a linear or nonlinear regression).

1. Set the value of a loss function to +∞. Select randomly p points, which uniquely determine the estimate b of regression parameters β.
2. For each observation, compute the residual and assign a weight to it based on
$$u^2_{(1)}(b) \leq u^2_{(2)}(b) \leq \ldots \leq u^2_{(n)}(b). \tag{9}$$
3. Compare the value of the loss function computed with the resulting weights with the current value of the loss function.Ifthe loss function is larger, go to step 4. Otherwise go to step 5.
4. Set the value of the loss function to the loss function and store the values of the weights. Find the estimator of $\beta_1, \ldots, \beta_p$ by weighted least using these weights. Go back to steps 2 and 3.
5. Perform steps 1 through 4 repeatedly c-times, where c is a given constant. The output (optimal) weights are those giving the global optimum of the loss function over all repetitions of steps 1 through 4.

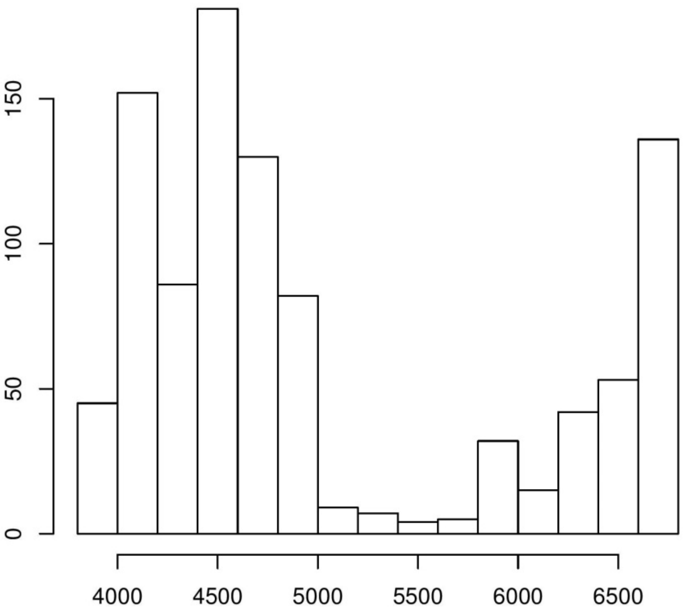
The weighted estimator in Step 4 is a classical weighted least estimator, either in linear or nonlinear case. A suitable choice of c has not been investigated even for the linear regression. Two illustrative examples on real data will follow. In Section 3.2, a suitable choice of c is investigated for the LWS estimator. In Section 3.3, computational aspects of the NLWS using Algorithm 1 are studied.

3.2 Example: Investment data

Our aim is to investigate the optimal number of iterations in Step 5 of Algorithm 1. We work with $n = 22$ values of real gross domestic product (GDP) and real gross private domestic investment (INVEST) in the United States in the years from 1980 to 2001. Both variables are expressed in 10^9 of USD. The data are copied from the website www.stls.frb.org/fred, whilethey originally come from the U.S.Department of Commerce. We use a linear regression model, where the response INVEST is explained by the regressor GDP.

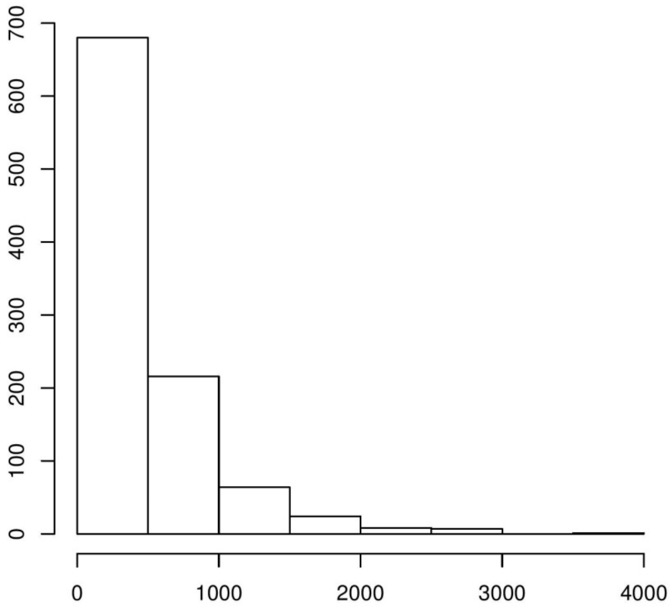
Table 3 presents estimates of the least squares (LS) and LWS for the linear regression model, together with values of $\sum_{i=1}^{n} w_i u^2_{(i)}(b)$,which is the loss function of the LWS estimator. The data do not contain any severe outliers.

Fig. 1: A study of computational aspects of the NLWS estimator. The loss function evaluated in particular 1000 iterations of Algorithm 1 in the example of Section 3.2.



Source: own computation

Fig. 2: A study of computational aspects of the LWS estimator on the example of Section 3.2. The number of iterations needed to reach the optimal loss out of the total number of 1000 performed iterations.



Source: own computation

Tab. 3: Results of the example of Section 3.2. Estimates of parameters in a linear regression model for the Investment data, which are computed with the least squares and the LWS estimator with linearly decreasing weights.

	Least squares	LWS estimator
Intercept	-582.0	-465.4
Slope	0.239	0.221
$\sum_{i=1}^{n} w_i u_{(i)}^2(b).$	4219.9	3910.3

Source: own computation

Tab. 4:Results of the example of Section 3.3. Estimates of parameters in a nonlinear regression model by the least squares (LS) and NLWS with linearly decreasing weights.

	Raw data	Contaminated data
Least squares – Intercept	190.8	193.9
Least sqaures – Slope	0.060	0.067
NLWS –Intercept	191.4	191.6
NLWS– Slope	0.061	0.061

Source: own computation

Let us now search for a suitable value of c for obtaining the the minimal value of the loss function, which equals in our example 3910.3. Figure 1 shows thenumber of iterations needed to reach the optimal loss out of the total number of 1000 independently performed iterations. Thus, the choice $c = 1\,000$ seems to be very safe in order to find the minimum of the loss function. This will be further verified in a more sophisticated study.

We investigate the number of iterations (i.e. repetitions of Steps 1-4) needed to reach precisely the value of the optimal loss. We repeated 1000 times the computation of 1000 iterations of Algorithm 1 with a random selection of p points in Step 1. Each time, we evaluated the value of$\sum_{i=1}^{n} w_i u_{(i)}^2(b)$, which is shown in Figure 2. The mean value is 458 iterations. In the average, it requires to compute 458 iterations to obtain the optimal value of the loss function. The number of iterations turns out to be less than 1000 in 87.9 % of cases. Indeed, the choice $c = 1000$ is safe in a large percentage of cases, but only the choice $c = 10\,000$ would ensure reaching the optimum in more than 99 % of cases and remains to be our recommended choice.

3.3 Example: Puromycin data

The aim of this example is to reveal the robustness of the NLWS estimator compared to the nonlinear least squares. We use a Puromycin data set, available in the package *datasets* of R software. The reaction velocity *Y* is explained as a response of the substrate concentration *X* in the nonlinear regression model

$$Y_i = \frac{\beta_1 X}{\beta_2 + X} + e_i, \quad i = 1, \dots, n, \tag{10}$$

where the aim is to estimate regression parameters $\beta_1$and β_2. To reveal the strength of the method, we also consider a contaminated data set, obtained by modifying the value of the observation in the Puromycin data set. Particularly, the concentration of the first observation was modified from 0.02 to 0.05 to become the only outlier in the data set.

The results of the least squares and NLWS estimators are shown in Table 4. The advantage of the NLWS is revealed on the contaminated data set, where it yields reliable values, while the least squares estimator is heavily influenced by the contamination. The constant $c = 10\,000$ in Algorithm 1 seems to be very sufficient also in nonlinear regression model.

To summarize the contribution of Section 3, it proposes a general algorithm for an implicitly weighted regression estimator (denoted as NLWS), encompassing both the linear and nonlinear framework. The performance of the estimator and mainly the computational aspects of the new algorithms are revealed in numerical illustrations. Based on the result of the computations, we recommend to choose $c = 10\,000$ for moderate samples sizes for Algorithm 1. Besides, the NLWS estimator turns out to perform reliably on a data set contaminated by outlying measurements. Such robustness is highly desirable in the analysis of real data and brings an argument in favour of the NLWS estimator.

Acknowledgment

The research was supported by the Czech Science Foundation projects No. 13-01930S and No. 13-17187S.

References

1. Agresti, A. (2013). *Categorical data analysis.* 3rd edn. New York: Wiley.
2. Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29-50.

3. Buonaccorsi, J.P. (2010). *Measurement error: models, methods, and applications.*Boca Raton: Chapman & Hall/CRC.
4. Filzmoser, P., & Todorov, V. (2011). Review of multivariate statistical methods in high dimension. *Analytica Chinica Acta*, *705*, 2-14.
5. Gentle, J.E., Härdle, W.K., & Mori, Y. (2012). *Handbook of Computational Statistics.* 2nd edn. Berlin: Springer.
6. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd edn. New York: Springer.
7. Kalina, J. (2012). Implicitly weighted methods in robust image analysis. *Journal of Mathematical Imaging and Vision*, *44*, 449-462.
8. Kalina, J., & Zvárová, J. (2013). Decision support systems in the process of improving patient safety. *E-health technologies and improving patient safety: Exploring organizational factors.* Hershey: IGI Global, 71-83.
9. Kalina, J. (2014). Classification methods for high-dimensional genetic data. *Biocybernetics and Biomedical Engineering*, *34*, 10-18.
10. 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.
11. Martinez, W.L., Martin, R.D., & Yohai, V.J. (2011). *Exploratory data analysis with MATLAB*. 2nd edn. London: Chapman & Hall/CRC.
12. Neykov, N.M., Filzmoser, P., & Neytchev, P.N. (2014). Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. *Statistical Papers*, *55*(1), 187-207.
13. 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*, 311-328.
14. Víšek, J. Á. (2011). Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*,*47*, 179-206.

Contact

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