A robust multinomial logit model for evaluating judges' performances

Un modello multinomiale robusto per valutare la performance dei giudici.

Ida Camminatiello and Antonio Lucadamo

Abstract Principal component multinomial regression is a method for modelling the relationship between a set of high-dimensional regressors and a categorical response variable with more than two categories. This method uses as covariates of the multinomial model a reduced number of principal components of the regressors. Because the principal components are based on the eigenvectors of the empirical covariance matrix, they are very sensitive to anomalous observations. Several methods for robust principal component analysis have been proposed in literature. In this study we consider ROBPCA method. The new robust approach will be applied for assessing judges' performances.

Abstract *La regressione multinomiale sulle componenti principali è un metodo per modellare la relazione tra un set di regressori ad alta dimensionalità e una variabile di risposta nominale con più di due modalità. Questo metodo usa come covariate del modello multinomiale un numero ridotto di componenti principali estratte dai regressori. Poiché le componenti principali si basano sugli autovettori della matrice di covarianza empirica, sono molto sensibili alle osservazioni anomale. Diversi metodi robusti per l'analisi in componenti principali sono stati proposti in letteratura. In questo studio consideriamo il metodo ROBPCA. Il nuovo approccio robusto sarà applicato per valutare la performance dei giudici.*

Key words: principal component analysis, multinomial logit model, outliers, judges' performances.

[1] Ida Camminatiello, University of Campania; email: ida.camminatiello@unicampania.it
Antonio Lucadamo, University of Sannio; email: antonio.lucadamo@unisannio.it

1. Introduction

The court computerization of the last decades allows us to create available databases with complete information about the judicial flows. Here, we aim to focus on the causes of different judges' performances in the court of Naples.

The dataset shows strongly correlated regressors, so the most proper statistic methodology to analyse this kind of data could be Principal component multinomial regression (Camminatiello, Lucadamo, 2010; Lucadamo, Leone, 2015).

A previous research on Florence Court (Camminatiello, Lombardo, Durand, 2017) highlighted the presence of outliers among judges.

The aim of the paper is to study the dependence relationship between the judges' performances and some indicators of the judges' workload taking into account multicollinearity and outlier problems which make the estimation of the multinomial model parameters inaccurate because of the need to invert nearsingular and ill-conditioned information matrices.

A robust method for logistic regression (Rousseeuw, Christmann, 2003) and robust logistic ridge regression (Ariffin, Midi, 2014) have been proposed in literature, we propose a robust approach for the principal component multinomial regression (PCMR).

We proceed in the following way. In the second section we describe the PCMR. In the third section we list the most important robust methods for estimating the variance/covariance matrix and propose a robust approach to PCMR. In the fourth section we apply our robust approach for evaluating the judges' efficiency and calculate the correct classification rate for comparing three different models.

2. From multinomial logit regression to robust methods for principal component multinomial regression.

Multinomial logit model (MNL) is the simplest model in discrete choice analysis when more than two alternatives are in a choice set. The model becomes unstable when there is multicollinearity among predictors (Ryan, 1997). To improve the estimation of the MNL parameters, Camminatiello and Lucadamo (2010) proposed the PCMR.

PCMR uses as covariates of the multinomial model a reduced number of principal components (PCs) of the regressors. Because these components are based on the eigenvectors of the empirical covariance matrix, they are very sensitive to anomalous observations. Several methods for robustifying principal component analysis (PCA) have been proposed in literature.

If the number of observations is sufficiently large with respect to the number of variables, the classical covariance matrix can be replaced by minimum covariance determinant (MCD) estimator, minimum volume ellipsoid (MVE) estimator (Rousseeuw and Leroy 1987), S-estimators (Davies 1987), reweighted MCD

(Rousseeuw and van Zomeren, 1990), FAST-MCD (Rousseeuw and Van Driessen, 1999).

For high-dimensional data, a ROBust method for PCA, called ROBPCA, has recently been developed (Hubert, Rousseeuw and Vanden Branden, 2012). ROBPCA starts by reducing the data space to the affine subspace spanned by *n* observations. A convenient way to perform it is by a singular value decomposition of the mean-centred data matrix. In the second stage $h < n$ "least outlying" data points are found by Stahel-Donoho affine invariant outlyingness. In the third stage, the algorithm robustly estimates the location and scatter matrix of the data points projected into a subspace of small to moderate dimension by using the FAST-MCD estimator. ROBPCA ends by yielding the robust principal components. Like classical PCA, the ROBPCA method is location and orthogonal equivariant.

2.1. Robust Principal Component Multinomial Regression - RobPCMR

Several authors applied ROBPCA to formulate other robust techniques (Hubert, Vanden Branden, 2003; Hubert, Verboven, 2003; Rousseeuw, Christmann, 2003). We investigate using ROBPCA before PCMR to deal with multicollinearity and outlier problems in MNL. We proceed in the following way.

Let $\mathbf{X} = \left[\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_p\right]$ be a set of *p* quantitative regressors and $\mathbf{y}$ a categorical response variable with more than two categories observed on *n* statistical units.

At first step, robust principal component multinomial regression (RobPCMR) creates the robust PCs of the regressors $\mathbf{Z} = \left[\mathbf{z}_1, \dots, \mathbf{z}_i, \dots, \mathbf{z}_p\right]$ which are linear combinations of the original variables $\mathbf{Z} = \mathbf{XV}$. At second step the multinomial model is carried out on the set of robust PCs. At third step, the number of robust PCs, $a < p$, to be retained in the model, is selected according to different tools (Camminatiello and Lucadamo, 2010).

At fourth step, the multinomial model is carried out on the subset of robust PCs chosen. The probability, for the individual *i*, to choose the alternative *c* can be expressed in terms of *a* robust PCs as:

$$\pi_i^{(a)}(c) = \frac{\exp\left\{\sum_{j=1}^{p}\sum_{k=1}^{a} z_{ik} v_{kj} \beta_{jc}^{(a)}\right\}}{\left\{\sum_{b=1}^{s}\exp\sum_{j=1}^{p}\sum_{k=1}^{a} z_{ik} v_{kj} \beta_{jb}^{(a)}\right\}} = \frac{\exp\left\{\sum_{k=1}^{a} z_{ik} \gamma_{kc}^{(a)}\right\}}{\left\{\sum_{b=1}^{s}\exp\sum_{k=1}^{a} z_{ik} \gamma_{kb}^{(a)}\right\}} \qquad (1)$$

where $\gamma_{kb}^{(a)} = \sum_{j=1}^{p} v_{kj} \beta_{jb}^{(a)}$ are the robust coefficients to be estimated on the subset of *a* robust PCs and $\beta_{jb}^{(a)}$ are the robust PCMR parameters obtained after the extraction of the *a* components.

Finally, the robust MNL parameters can be expressed in function of original variables (**X** matrix)

$$\mathbf{Z}^{(a)}\boldsymbol{\gamma}^{(a)} = \mathbf{X}\mathbf{V}^{(a)}\boldsymbol{\gamma}^{(a)} = \mathbf{X}\boldsymbol{\beta}^{(a)} \quad (2)$$

where $\boldsymbol{\beta}^{(a)} = \mathbf{V}^{(a)}\boldsymbol{\gamma}^{(a)}$ is the matrix of robust parameters expressed in function of original variables; $\boldsymbol{Z}^{(a)}$ is the matrix of robust PCs; $\boldsymbol{\gamma}^{(a)}$ is the matrix of robust parameters on a robust PCs for the s alternatives; $\mathbf{V}^{(a)}$ is the matrix of robust eigenvectors.

To measure the performance of a method, several criteria can be utilised (Camminatiello, Lombardo, Durand, 2017; Camminatiello, Lucadamo, 2010). Here, we focus on rate of well classified which we expect higher compared to PCMR and MNL.

3. A robust model to predict judges' performances.

Our study concerns the causes of different judges' performances in the court of Naples. The performance evaluation is based, among others, on the time that each judge employs to solve the disputes. According to many available publications about similar problems (Camminatiello, Lombardo, Durand, 2017), we aim to study if number of: pendings, hearings, dossiers, incoming and defined proceedings can influence the judges' performances. The response variable is on categorical scale, with four modalities from 1 (Low) to 4 (High), measured for 136 judges.

To evaluate how judges' performances can be influenced by explicative variables, we divide our sample in two sub-samples. The first one, composed by the 70% of the observations, is the sample used to estimate the model parameters (estimation sample). The second one (validation sample) is considered to test the goodness of the obtained estimates. In both the cases we calculate the rate of well classified judges and we compare the results obtained by applying the MNL, the PCMR and the RobPCMR.

In table 1 the results obtained by the three methods on the estimation sample are shown. From the second to the fifth column we have the percentage of well classified, calculated for four different models: the MNL estimated on all the regressors and on the two significant ones (via stepwise regression); the PCMR run on the first PC (which accounts for 93.4% of the variance, has the eigenvalue higher than one and furthermore is the only significant regressor); the RobPCMR carried out on the first robust PC (which accounts for 89.8 % of the variance, has the eigenvalue higher than one and is the only significant regressor).

It is easy to observe that for the estimation sample the classical MNL performs better than other models: it could be due to an over-fitting problem.

Table 1: Percentage of correct classified calculated for four different models on the estimation sample

	MNL (all)	***MNL (sign)***	***PCMR (1)***	***RobPCMR (1)***
% correct classified	70.1%	63.9%	59.8%	58.8%

To verify the goodness of the techniques it is then necessary to consider the results on the validation sample (table 2).

In this case it is evident that for all the methods, but for the RobPCMR, the percentages are lower than before.

Table 2: Percentage of correct classified calculated for four different models on the validation sample

	MNL (all)	***MNL (sign)***	***PCMR (1)***	***RobPCMR (1)***
% correct classified	53.8%	56.4%	53.8%	59.0%

In fact, looking at the table 2, we can notice that the MNL, considering all the variables, leads to the same classification rate obtained by the PCMR with only one component, while for the MNL, taking into account only the two significant variables, the percentage of well classified increases. It is surprising that RobPCMR result shows a rate of correct classification higher than before (59.0% against 58.8%). This may indicate the ability of the method in parameter estimation.

It is also interesting to notice what happens when we consider more components in the analysis, both for PCMR and for RobPCMR.

For this reason we show in table 3 and 4 the results obtained when we consider a different number of components as explicative variables.

Table 3: Percentage of correct classified, at varying the number of components, for the estimation sample

Number of components	***PCMR***	***RobPCMR***
1	59.8%	58.8%
2	60.8%	60.8%
3	60.8%	63.9%
4	61.9%	63.9%
5	70.1%	62.9%

For the PCMR, the results on the estimation sample show that, when the number of components increases, the classification improves. Considering all the components the result is equal to that obtained using all the variables in the classical MNL (Camminatiello and Lucadamo, 2010).

For RobPCMR instead, there is an improving in the correct classification rate at beginning, but when we consider all the components, the result is lower than one obtained with 3 and 4 components.

If we consider the validation sample the results confirm both for the PCMR and for the RobPCMR that the selection of the significant components (in this case explaining the most part of the variability too) is an useful solution to obtain a good rate of correct classification.

Table 4: Percentage of correct classified, at varying the number of components, for the validation sample

Number of components	*PCMR*	*RobPCMR*
1	53.8%	59.0%
2	53.8%	59.0%
3	51.3%	56.4%
4	48.7%	58.9%
5	51.3%	56.4%

Obviously for RobPCMR, as already done in previous studies for PCMR, a complete simulation study is necessary to generalize the results.

4. Conclusion and perspective

In this paper we carried out a robust model for evaluating the judges performances in presence of outliers and strongly correlated covariates. .To solve these problems, we proposed to use as covariates of the multinomial model a reduced number of robust PCs of the predictor variables.
The application showed that the proposed approach is a valid alternative on real data. However, an extensive simulation study is needed in order to verify that it is resistant towards many types of contamination, to compare the results with other robust approaches for PCA proposed in literature and to select optimal dimension of the model. The procedure should lead to lower variance estimates of model parameters comparing to PCMR. The variance can be estimated by bootstrap resampling.
Finally, an extension to MNL of the approaches proposed in literature, for dealing with multicollinear and outlier problems in the logit model (Ariffin, Midi, 2014) could be interesting as well as an extension to ordinal logit regression of the approach here proposed.

5. References

1 Ariffin, S.B., Midi H.: Robust Logistic Ridge Regression Estimator in the Presence of High Leverage Multicollinear Observations. In: 16th Int. Conf. Math. Comput. Methods Sci. Eng. pp 179-184 (2014)

2 Camminatiello, I., Lucadamo, A.: Estimating multinomial logit model with multicollinear data. Asian Journal of Mathematics and Statistics 3 (2), 93-101 (2010)

3 Camminatiello, I., Lombardo, R., Durand, J.F.: Robust partial least squares regression for the evaluation of justice court delay. Qual Quant 51 (2), 813-27 (2017). https://doi.org/10.1007/s11135-016-0441-z

4 Davies, L.: Asymptotic Behavior of S-Estimators of Multivariate Location and Dispersion Matrices. The Annals of Statistics 15, 1269-1292 (1987).

5 Hubert, M., Rousseeuw, P.J., Vanden Branden, K.: ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics 47 (1), 64-79 (2012) doi: 10.1198/004017004000000563
6 Hubert, M., Vanden Branden, K.: Robust Methods for Partial Least Squares Regression. *J Chemometr* 17, 537-549 (2003)
7 Hubert, M., Verboven, S.: A Robust PCR Method for High-Dimensional Regressors. *J Chemometr* 17, 438-452 (2003)
8 Lucadamo, A, Leone, A.: Principal component multinomial regression and spectrometry to predict soil texture. *J Chemometr*., **29** (9), 514-520 (2015).
9 Rousseeuw, P. J., Christmann, A.: Robustness Against Separation and Outliers in Logistic Regression. Computational Statistics and Data Analysis 43, 315-332.
10 Rousseeuw, P.J. Leroy, A.M.: Robust regression and Outlier Detection. Wiley, New York (1987).
11 Rousseeuw, P.J., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212-223 (1999)
12 Rousseeuw, P. J., Van Zomeren, B. C.: Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association 85, 633-651 (1990).
13 Ryan, T.P.: Modern Regression Methods. Wiley, New York (1997)