Robust statistical methods for credit risk

Metodi statistici robusti per il rischio di credito

A. Corbellini, A. Ghiretti, G. Morelli and A. Talignani

Abstract Credit risk is a relevant problem faced by banks and financial institutions. The traditional statistical models which are generally used to quantify the credit risk present several drawbacks. First, in their standard versions they are not robust and do not take into account that the data may be corrupted by several outliers. Second, when a parametric model is employed, the variable selection procedure might be severely affected by the so called masking and swamping effects. This work extends robust statistical methods to credit risk analysis, showing how the traditional approach can be greatly improved through robust methods.

Abstract *La gestione del rischio di credito è un problema particolarmente rilevante per tutte le banche e le istituzioni finanziarie. I modelli statistici tradizionalmente utilizzati per quantificare tale rischio presentano diversi svantaggi. Quando il dataset contiene alcuni outliers il fit che si ottiene attraverso i metodi di stima standard può risultare distorto e inconsistente. Inoltre, il metodo di selezione delle variabili può essere severamente influenzato dagli effetti di masking e/o swamping. L'obiettivo di questo lavoro è quello di estendere i metodi statistici robusti all'analisi del rischio di credito, mostrando come l'analisi può essere fortemente migliorata utilizzando un approccio robusto.*

A. Corbellini
Department of Economics and Management, University of Parma, Via J. F. Kennedy, 6, Parma, e-mail: aldo.corbellini@unipr.it

A. Ghiretti
Department of Statistics, Computer Science, Applications, University of Florence, Viale Morgagni, 59, Firenze e-mail: ghiretti@disia.unifi.it

G. Morelli
Department of Economics and Management, University of Parma, Via J. F. Kennedy, 6, Parma, e-mail: gianluca.morelli@unipr.it

A. Talignani
Department of Economics and Management, University of Parma, Via J. F. Kennedy, 6, Parma, e-mail: andrea.talignani@studenti.unipr.it

Key words: Credit Risk, Forward Search, Lasso, Outliers, Robustness

1 The Credit Risk framework

Credit risk is defined as the risk of default on a debt that may arise from a borrower failing to make the required payments. In order to quantify how likely a borrower will be unable to meet his debt obligation, it is customary to use the so called probability of default (PD). The PDs have been introduced in the the Basel agreements, that took into account new developments in the measurement and management of banking risks for those institutions that agreed to use the "internal ratings-based" (IRB) approach. In this approach, financial institutions and banks are allowed to use their own internal measures as primary inputs to the capital calculation. These measures, require the estimation of a set of indexes that describe the risk exposure of the institution. These risk measures are subsequently converted into risk weights and further into regulatory capital requirements by means of risk weight formulas specified by the Basel Committee. In this work we will focus on the estimate of the PD, showing how the standard statistical methods generally proposed in literature can be greatly improved when a robust approach is adopted. In literature several statistical models have been proposed to estimate the PD. Some commonly adopted examples are: logistic regression models, discriminant analysis, classification trees and so on. The main drawback of all of these procedures is that they are not robust against slight deviations from the model assumptions. In fact, real data are generally corrupted by a random number of outlying units, i.e, units that do not share the same characteristics of the majority of data. It is well known in the literature that neglect outliers might have a severe impact on the analysis, leading to biased and inconsistent estimates and misleading inference. Furthermore, when facing real data, the analyst is generally required to perform a variable selection, as many are often available, but some might not be relevant to drive the PD. As a consequence, when a parametric model such as the logistic regression is adopted, it is common practice to employ a variable selection technique. However, standard variable selection techniques such as stage-wise algorithms or the widespread LASSO may be seriously affected by few outliers, leading to miss-selected variables.
To show the great improvements that can be achieved with a robust analysis our work will consider a real data set made available by an Italian Bank, which we kindly acknowledge, but do not report for confidentiality.

2 A Forward Search approach to the LASSO

We consider a set of observations $z_i = (y_i, x_i'), i = 1, \ldots, n$, where y_i is a binary variable and x_i' is a vector of p features relative to the ith company.

We estimate the default probability of a firm by a logistic discriminant function, that is,

$$Pr(y_i = 1|x_i) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} = \frac{1}{1+e^{-x_i'\beta}} \qquad i = 1,\ldots,n.$$

$x_i' = (x_{i1},\ldots,x_{ip})$ is the vector of predictors, i.e. firm-specific characteristics and financial indexes and β is a vector of p unknown parameters.
The parameters are generally estimated by maximizing the log-likelihood function,

$$\ell(\beta) = \sum_{i=1}^{n} [y_i x_i'\beta - \log(1+e^{x_i'\beta})]. \tag{1}$$

In the data set considered one of the major concerns was to extract, from the p original features, a subset of k highly significant predictors. The LASSO (Hastie and Tibshirani, 2009) [2], performs a variable selection by means of a regularization or penalization parameter.

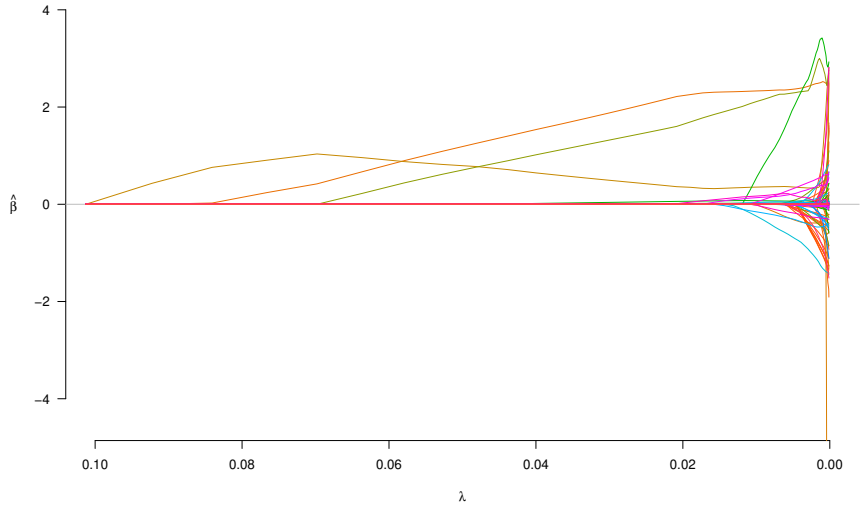

Fig. 1 Plot of the trajectories of the estimated coefficients, as a function of λ, obtained with the group LASSO with no outliers

When the LASSO is applied to the logistic model, the resulting penalized version of the log-likelihood function to be maximized becomes

$$\ell_{\lambda}^{L}(\beta) = \sum_{i=1}^{n}[y_i x_i'\beta - \log(1+e^{x_i'\beta})] + \lambda \sum_{j=1}^{p}|\beta_j|.$$

By selecting a value of λ sufficiently large, the L_1 penalty used in the LASSO, forces some of the coefficient estimates to be exactly equal to zero. The selection of λ is generally performed by cross validation and in the trivial case $\lambda = 0$ the log-likelihood (1) is maximized.
We will denote the optimal value of λ obtained by cross validation with λ_{optim}.
In our case, we have p predictors, h of which are qualitative (with $h \ll p$), and it is common to code those variables with dummies. However, it might happen that, a group of dummies represents the same variable, therefore it is compulsory to assume that the p predictors belong to G distinct groups. Notice that each quantitative variable form a distinct group. The parameter ρ_g is such that we either select or neglect the entire group. This is the standard approach suggested by Yuan e Lin (2007) [4] via the group LASSO, expanded by Meier, Van De Geer and Bühlmann (2008), [3] where groups of features can be included together into or out of a model. The group LASSO estimator is obtained by maximizing the following penalized log-likelihood function

$$\ell_{\lambda}^{GL}(\beta) = \sum_{i=1}^{n}[y_i x_i'\beta - \log(1+e^{x_i'\beta})] + \lambda \sum_{g=1}^{G}\sqrt{\rho}_g\|\beta_g\|_2$$

where β_g is the vector of parameters associated with the predictors in group g.
In our work we propose a novel use of the Forward Search , see Atkinson and Riani (2000) [1], coupled with a LASSO regularization technique. This allows to detect multiple outliers and perform a selection of the most significant explanatory variables in a robust way.

In order to detect outliers and departures from the fitted regression model, the Forward Search uses the group LASSO to fit the model to subsets of m observations. The initial subset of m_0 observations is chosen robustly as follows:

1 fit the group-LASSO and select λ_{optim} by performing a cross validation

2 obtain the residuals for all the observations given the k features obtained at the previous step

3 sort all the residuals

4 repeat steps from 1-3 ten-thousand times and store the results in a data matrix

5 the subset of units that minimize the median of the residuals is selected.

After the initial subset has been obtained, the search procedure starts. The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest residuals. For each m $(m_0 \leq m \leq n-1)$, by a graphical monitoring of the maximum standardized residual among the units included in the set,

$s_{i,max}$, and the minimum standardized residual for the units excluded from the fit, $s_{i,min}$, it is possible to detect a potential outlier.
It is worth to be noted that, as outlined at step 2, at each iteration of the Forward Search, the residuals are calculated for all units, notwithstanding the fact that these residuals are stemming from different sets of variables. Moreover, since in the Forward Search at each step the residuals are ordered, and several units or groups of units might join and leave the subset, some problems concerning the identification of the model may arise. To overcome this drawback we impose that in every subset generated by the search at each step, there will be a percentage of defaulted firms equal to the original percentage found in the whole data. The defaulted firms included in the subset are selected among those with smallest residual.

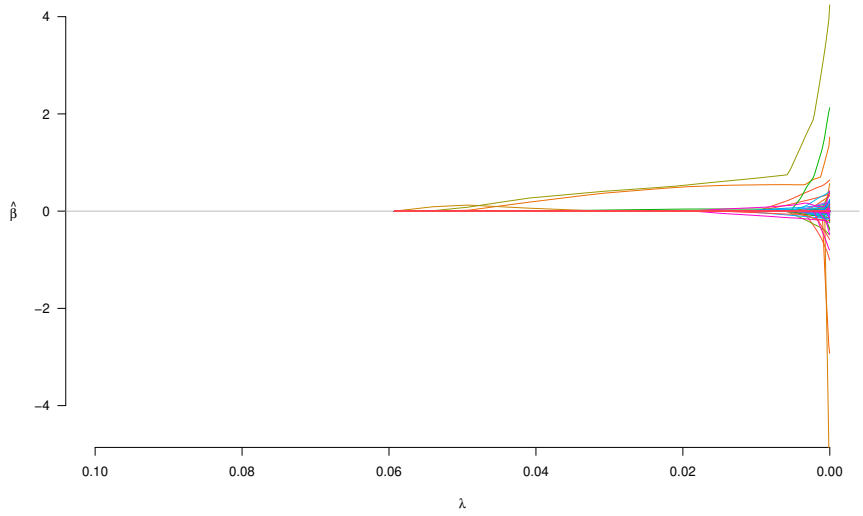

Fig. 2 Trajectory plot of the estimated coefficients, as a function of λ, obtained with the group LASSO when a 15% contamination is introduced

3 Preliminary results

As we have previously mentioned standard statistical techniques can be seriously affected by the presence of outliers in the data.
In order to show the impact that a small contamination might provoke on the group LASSO we introduce a 15% contamination into the Bank dataset and subsequently we perform the group-LASSO on the contaminated data.

The contamination is performed by adding a percentage of 15% additive outliers to the response variable Y. Figure 1 and Figure 2 show the trajectories of the estimated parameters over different values of lambda, before and after the contamination.
The different trajectories in the two plots show clearly that when some contamination is introduced the spurious units affect consistently the variable selection technique.
Figure 3 highlights the effect that the spurious units introduced have on the cross validation. First, the λ_{optim} selected in the contaminated scenario results lower than the one selected with the clean data. Second, the cross validation error in the contaminated case results sensibly larger for all the values of λ. Last, but not least, the number of groups selected with the introduction of the additive outliers drops remarkably to a number of 14.

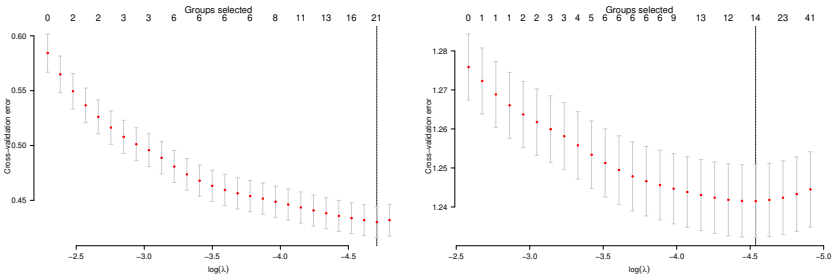

Fig. 3 Plots of λ estimates. Left panel, uncontaminated data, right panel, 15% of contamination

The significant drop in the number of selected groups as well as the increase in the cross validation error, suggest that there were swamping and masking effects induced by the introduction of several outliers.
The aim of our work is to perform a robust calibration of the value of λ whose choice is not affected by the spurious units and, at the same time, identify the most influential units by means of the forward plots.

References

1. Atkinson, A., Riani, M.: Robust Diagnostic Regression Analysis. Springer Verlag (2000).
2. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer (2009).
3. Meier, Lukas and Van De Geer, Sara and Bühlmann, Peter: The group lasso for logistic regression. Journal of the Royal Statistical Society, Series B **70**(1), 53-71 (2008)
4. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B, **68**(1), 49-67 (2007)