Università
Ca'Foscari
Venezia

Master's Degree programme – Second Cycle
(*D.M. 270/2004*)
in Economia e Finanza

Final Thesis

Ca' Foscari
Dorsoduro 3246
30123 Venezia

Asymmetric information in loan contracts: A game-theoretic and statistical approach

Supervisors
Ch. Prof. Monica Billio
Ch. Prof. Marco LiCalzi

Graduand
Francesco Benvenuti
Matriculation Number 839313

Academic Year
2015 / 2016

Abstract

In this work, we apply game-theoretic and statistical models to an open problem regarding asymmetric information in loan contracts. Under these asymmetries, the effect of higher collateral requirements on the interest rates applied by banks to borrowers is not clear. The literature has argued both for positive and a negative links, based on different hypotheses and econometric analyses. We discuss how this effect cannot be decided a-priori. In the first part we construct three game-theoretic models under different hypotheses, proving the impossibility of an univocal effect. Then, to assess what is the prevailing effect in the reality, we analyze for the first time loan big-data for millions of borrowers among various European countries, as collected by the European DataWarehouse. We examine some mathematical and practical aspects of: the Principal Component Analysis (PCA), the Principal Component Regression (PCR), the regularization theory, the LASSO and RIDGE regressions, applying them to our datasets. Finally, we combine a regression model with the Probabilistic PCA, discussing the EM algorithm in presence of sparse datasets. These datasets are characteristic of our database and others, and defining the Probabilistic PCR we propose a new technique which may show itself useful if the availability of loan data will increase over time conserving some data sparsity.

Contents

Chapter 1

Introduction

1.1 Combining game theory with econometrics

1.1.1 The problem

The present work concentrates on the influence of asymmetric information in loan contracts, examining the subject both from a theoretical and an empirical viewpoint. In particular, we consider a specific effect of the information asymmetries which affect the credit market, and our main aim is to answer the question: does collateral requirements increase or decrease interest rates of loan contracts? While this topic can be analyzed through different points of view and various subjects, we take advantage of game-theoretic models along with mathematical and statistical tools to tackle it.

In fact, as we will see in the rest of the thesis, the answer to this apparently simple question is not straightforward, since it requires a deep analysis and somewhat advanced quantitative tools to give a satisfactory conclusion. We apply, for example, utility and piecewise-defined functions, sets of borrower's variables, and probabilistic concepts to build some game-theoretic models; principal component and regularized regressions, algorithms and geometric dimension reduction techniques to analyze a huge amount of real loan data.

In doing so, we implicitly recognize the potential of quantitative reasoning applied to a purely economic question. Indeed, in the last decades, the quantitative

subjects have spread their influence in almost every field of research; this fact can be explained at least through three different considerations, which reinforce themselves reciprocally. Firstly, the significant development of these disciplines, which have been enriched of more powerful concepts over time. Secondly, the recognition of the usefulness of a formal reasoning applied to many questions arising from disparate subjects. Finally, the technology development which has made possible to take advantage of many applied mathematical methods effectively. All these aspects are characteristic of our thesis. Indeed, we rely heavily on mathematical and statistical theory, on its application to economic models and data analysis, and we use a recent database (the *ED*, see below) and the *MATLAB* software to analyze loan datasets.

To introduce the problem analyzed, we sum up the main questions which we aim to answer throughout this thesis, and which are examined in depth through the following chapters:

1. What happens to the interest rate of loan contracts when collateral of a higher value is required?

2. Is this effect influenced by information asymmetries, which are characteristic of loan contracts?

3. Is the theoretical and the empirical literature unanimous on these points?

4. How can we analyze loan big-data in an effective way? Which are the mathematical tools and the statistical models which allow us to empirically assess the interest rate-collateral relation?

5. How can we handle the problem of sparse data, which is characteristic of our and other databases, in regression models?

Asymmetric information concept is not new to finance in general, and it is a typical subject of study of game theory. Actually, it affects collateral contracts, where two players, the bank and the borrower, interact. We show that these asymmetries influence the collateral-interest rate link, since collateral can be

used as an effective tool under moral hazard and adverse selection. In fact, these situations change the conclusions reached in the analysis under symmetric information, pointing out that not any effect can be decided *a priori*, since these scenarios are equally plausible.

The economic and econometric literature has broadly discussed the topic as well, but depending on the particular assumptions made by each author, contradictory conclusions have been provided. We can expect that different contracts match some of these specific hypotheses better, hence the overall effect of collateral requirements on loan interest rates is ambiguous, as we prove theoretically. Then, this conclusion is empirically tested in the second part of the thesis, using data for millions of borrowers among various European countries. We apply to this data the Principal Component Analysis and the Principal Component Regression (PCA and PCR), some regularized regression techniques, the Cross-Validation and other statistical tools. We analyze their derivation, the theory behind these subjects and the problems related to their practical application.

We find that different effects hold for different countries and for different contracts, which is consistent with our theoretical analysis. Moreover, since other borrower's variables are considered in the analysis along with the collateral one, we provide some comments about their effect on the interest rates too, along with their reciprocal relations found through an unsupervised analysis. Finally, we notice that, while the availability of data for the current variables is likely to be improved, one characteristic feature of our database is a certain amount of sparse data. Moving from this practical consideration, we theoretically discuss different not trivial techniques and models useful under this framework (the Probabilistic PCA, the EM algorithm for sparse datasets) and we define the Probabilistic PCR moving from the basic PCR technique.

1.1.2 The link between game theory and econometrics

This work is made up of two closely interconnected parts. The first is theoretical, and we take advantage of the modelization allowed by game theory, proposing different models and formalizing the problem discussed. The second part is a statistical/econometric part, which allows us to empirically analyze a large amount of loan data in order to verify what happens in the real word. In fact, the effect of collateral requirements on loan interest rates could have been considered only from a theoretical viewpoint, or only from an empirical one, but the combination of these two aspects is very effective in this context. The rational approach of game theory, and the consequent results obtained theoretically, are enriched by an econometric analysis, where we explore what empirical data reveals. Indeed, we find that our peculiar approach to the problem presented in the first theoretical part is supported by our empirical findings. Actually, we do not propose a unique answer to the question which our thesis moves from, but our aim is to show that the effect of collateral on the loan interest rate can be either positive or negative, depending on different theoretical assumptions and on different situations which take place in the real word. We give a very short overview of the content of the following chapters.

1.2 A brief outline of the chapters

1.2.1 Chapter 2

In this chapter, we present a concise, but complete, literature review on the topic of this thesis. We move from the seminal work by Stiglitz and Weiss (1981), and then we analyze the contrasting theoretical literature which has discussed the effect of collateral on loan contracts, and in particular on the borrower's risk and on the interest rate applied. We mainly focus on the different hypotheses made by the authors, comparing them in order to understand why they reach opposite conclusions. Then, we revise the empirical literature, which has found contrasting results as well, depending on the particular data-set considered and

on the procedure chosen.

1.2.2 Chapter 3

In this chapter, we built our game-theoretic models in order to prove the *a priori* impossibility of an unambiguous effect of collateral requirements on the interest rates of loan contracts. Specifically, we show that different hypotheses, related to three different scenarios — symmetric information; moral hazard; adverse selection — lead to different conclusions, proposing a more general approach compared with the previous literature. We prove that with perfect information the presence of collateral implies a lower interest rate, but if asymmetric information is considered this effect is no longer straightforward. Our conclusions are in line with those provided by other authors, but we provide a different derivation.

1.2.3 Chapter 4

In this chapter, firstly we analyze the main mathematical and statistical aspects related to the Principal Component Analysis, for example its algebraic derivation, the choice between the covariance and the correlation matrix, the interpretation of the coefficient vectors and their associated eigenvalues, the Singular Value Decomposition. Then we focus on the Principal Component Regression method, examining its theoretical motivation and practical aspects, such as its rationale and the variable selection problem. Moreover, we briefly introduce the Cross-Validation and the *K-means* clustering in order to apply them in our empirical analysis. We conclude the chapter with the practical application of these tools to our loan big-data, examining the relation between the interest rate and collateral along with other borrower's variables.

1.2.4 Chapter 5

In this chapter, we extend the previous analysis applying other regression techniques to our data. We examine two regularized regression methods, the Ridge and the LASSO, which have an analogous in the mathematical theory of

regularization, but were independently developed by statisticians. We examine the theory behind these techniques, in particular their different shrinkage effects, justifying them both analytically and geometrically. We discuss the differences and the analogies between the PCR and these regressions, in particular their higher-bias and possible lower-variance estimators compared with the OLS as shown in literature. At the end, we apply these techniques to the loan data and we analyze the differences with the results of Chapter 4, showing that the conclusions do not change.

1.2.5 Chapter 6

In this chapter, we discuss a new possible approach which can be particularly useful for our dataset. In fact, the current RBMS datasets on the ED present a huge volume of missing data especially for the optional variables, thus we propose an approach which may be useful under the realistic hypothesis that the availability of quantitative loan data will increase over time, however conserving some data sparsity. We examine the Probabilistic PCA, an extension of the PCA which allows a stochastic modelization of variables, in order to discuss an application of the EM algorithm in the presence of missing values. Since this method is unsupervised, we extend it into a supervised one, defining the Probabilistic PCR and examining whether this extension can be applied, for example showing the necessity to orthogonalize the column vectors derived through the EM algorithm, which span the principal subspace, in order to apply the PCR model and to interpret the results in terms of the original variables.

1.3 Data description. Big-data analysis

1.3.1 Availability of loan data: The European DataWarehouse

A substantial part of this thesis has been made possible thanks to the huge collection of loan data stored in the European DataWarehouse database. The European DataWarehouse (ED) "*is the first centralised platform in Europe which*

collects, stores and distributes standardised ABS loan level data" (see the ED website). Therefore, the loan considered are an underlying for an Asset-Backed Security and the storage of loan data is motivated by the *"ABS Loan Level Initiative"*, which aims to improve the transparency of ABS markets, giving access to the market participants to this information; this initiative was conceived by the *European Central Bank*.

The specific analysis presented in this thesis has been made possible thanks to the agreement between Ca' Foscari University and the ED, which has provided the Edwin and the EDplus services. In particular, in this thesis we have analyzed millions of data collected for each single borrower associated with a residential mortgage-backed security (RMBS), for different European countries (see table 4.1 in Chapter 4). This huge amount of data requires non-traditional techniques and can be seen as an evidence of the increasing importance of the so called "big-data analysis" research field.

1.3.2 Big-data analysis: An emerging research field

One of the key characteristics of the ED data is, without doubt, their huge dimensionality. In fact, the ED offers various possibilities to aggregate them and to visualize lots of synthetic indicators, in order to make faster and easier comparisons. Nevetheless, the huge dimensionality offers a great amount of information if data is analyzed properly, as we have attempted to do in the present work.

Indeed, the importance of data analysis in recent years cannot be overstressed. This subject is important for each aspect considered, from informatics to statistics, from meteorology to natural sciences such as genetics and astronomy, from image analysis to economics.

From a more specific viewpoint, research on data analysis is one of the most promising sectors of modern mathematics. It is enough to point out a well-known consideration of Gromov (August 1998), whose remarks on the future of mathematics include to *"deal with huge amounts of loosely structured*

data", provided that *"we shall need [...] mathematical professionals able to meditate between pure and applied science"*. The subject seems very interesting, theoretically and practically too. In fact, a development of further mathematical and I.T. tools is necessary to properly take advantage of this data, which in our time is overabundant. The reduction of the so-called "Big-data" and the following extraction of their nested information can lead to a substantial improvement of our knowledge in almost every field. This explains why nowadays data analysis is one of the most prolific research fields. This development is supported by major increases in information technology. Indeed, the data is collected in bigger and bigger databases, taking advantage of increased computer memory and computing power unimaginable until a few decades ago. In this work we make these statements effective. Firstly, we take advantage of some statistical and mathematical tools to empirically assess the theoretical hypotheses considered in the game-theoretic part. In fact, we rely heavily on algebraic, mathematical analysis, computer science and probabilistic tools to analyze, reduce and interpret our data. Secondly, we propose a new approach to the principal component regression (see below) to address the problem of sparse independent variables, extending the probabilistic PCA proposed by Tipping and Bishop (1999), to a regression model.

Since the underlying variables of our data vectors are of an economic type, the final target is an econometric analysis. As a matter of fact, we analyze and discuss empirical results derived from the analysis of loan data, showing how all these tools show themselves effective in this particular context. Specifically, we discuss some useful methods which can be applied to the large datasets characteristic of the European DataWarehouse. The analytical tools proposed here are general, hence any researcher will be able to apply them in the future, to the ED datasets with a more complete loan data availability, or to other datasets.

1.3.3 Data characteristics and the problem of sparse matrices

We briefly resume the key points of our empirical analysis, in particular of our data, which are thoroughly discussed throughout the econometric part. Our analysis is a cross-sectional study, where the dependent variable is *the Interest Rate Margin* of loans associated with RMBS contracts, and the regressors are specific borrower's variables, among them the value of collateral pledged. In order to have comparable data, we preliminarily split our loan data according to the country of origin, and to the nature of the interest rate, namely fixed or floating, and in addition for the fixed rate we consider separately two shorter periods, approximately homogeneous for macroeconomic variables. For a complete description please refer to Section 4.5 of Chapter 4.

As anticipated, a major characteristic of our data is a certain sparsity among the optional variables (those which can be optionally compiled by the creditors). This problem can be considered as an intrinsic feature of many databases, thus the topic of sparse data can be seen as an interesting research topic of our age. In Chapter 6, we present some methods which follow this direction, and which may be applied to the ED database in the future.

Part I

Game Theory

Chapter 2

Literature review

2.1 Asymmetric information in financial contracts: An overview

The problem of information asymmetries is not new to the field of game theory, or to finance in general. Its importance was made apparent by the award of the *Nobel Memorial Prize in Economic Sciences* to George Akerlof, Michael Spence and Joseph Stiglitz in 2001, acknowledging their contribution in studying markets with asymmetric information; see, for example, Rosser Jr. (2003). Indeed, it is not difficult to figure how financial interactions, which involve a certain number of players and usually contrasting interests, can be frequently characterized by features such as moral hazard, private information, opportunistic actions, incentive issues and related aspects.

Many researchers have examined a broad range of topics related to these themes. With regard to the subject of the present work, asymmetric information has been shown to affect financial contracts deeply and specifically loan contracts, as mentioned in the introduction. For example, without any pretense to give a comprehensive list, asymmetric information generally leads to a conflict between bondholders and stockholders (Smith and Warner, 1979), and this situation could cause a decrease of firms' value. In fact, asymmetric information generally can be associated with a loss of social welfare, as discussed by Csóka et al. (2015) for

the case of corporate financing with moral hazard. Other authors have focused on solutions for this problem, such as regulatory devices to mitigate asymmetric information in loan contracts (Berndt and Gupta, 2009). Furthermore, since the theme of information asymmetries is recurring when two players stipulate a contract, game theory has turned out to be useful in the analysis of loan and insurance contracts. For instance, its concepts have been applied to analyze the decision to strategically default in a mortgage contract (Collins et al., 2015) or to examine the relationship between borrowers and lenders in an open economy (Claus, 2011).

On the other hand, the topic of collateral requirements has been studied under different aspects too. For example, some analyses outside the scope of our research have dealt with studying the specific effect of collateral requirement in affecting the participation of borrowers to the market (Acharya and Viswanathan, 2008) or examining how collateral can increase efficiency of the credit market (Manove et al., 2001). In this thesis, we focus specifically on the relationship between collateral requirements, borrower's risk and the interest premium in loan contracts.

In the next section we give a brief summary of the literature on this subject.

2.2 Asymmetric information in loan contracts: The role of collateral

2.2.1 Theoretical literature

A seminal work regarding the effect of imperfect information in the credit market and the role of collateral and interest rates is Stiglitz and Weiss (1981), who consider the interaction between players in a loan contract, the bank and the borrower. They argue that the decisions about the interest rate could affect both the actions (*"the incentive effect"*) and the nature (*"the adverse selection effect"*) of borrowers. For the most part, their analysis concerns credit rationing, which is a consequence of the credit market imperfection, and in some cases cannot be

modified by additional collateral requirements.

Credit rationing refers to a limitation of credit supplied by the bank, when the interest rate is no longer a factor able to rebalance the market. In order to justify this situation, firstly they prove mathematically that the expected return on a loan is a decreasing function of its risk. Therefore, a bank raising the interest rates on loans faces two opposite effects: a *"direct effect"* which is an increase of its expected return due to a higher interest rate, and an *"indirect effect"* which is due to adverse selection. If the latter effect prevails over the former, all the contracts whose interest rate r exceeds a given interest rate $\hat{r}^*$ are denied, and the basic market equilibrium is altered due to unverifiable actions of the borrower. In other terms, banks could design contracts to discourage the borrowers' actions which are in conflict with their interests, and then equilibrium credit rationing would occur. Since in their analysis, interest rates are not always a factor able to equilibrate supply and demand, we may think that collateral could do it. But under the assumption that borrowers have DARA preferences, they prove that borrowers with higher wealth, who can provide more collateral, are also prone to invest in high risk projects. Therefore, higher interests and higher collateral requirements could lead the borrower to undertake riskier investments, and this fact can decrease the advantage of banks even if collateral in itself always increases the bank return.

This model has not gone without criticism and has triggered off a huge literature on the relationship between collateral and risk in loan contracts.

An influential different view is given in Bester (1985). The author, contrary to Stiglitz and Weiss (1981), assumes that banks choose simultaneously the interest rate and the collateral amount rather than separately. In this case, he states, in equilibrium no credit rationing occurs, because now a self-selection mechanism can take place. If this assumption holds along with other conditions, like an unrestricted collateral availability, he proves that contracts with higher interest rate and lower collateral requirements are chosen by riskier borrowers. On the contrary, borrowers with lower default probability are willing to provide more

collateral. In a later paper, Bester (1987) is partially in contrast with the result of Stiglitz and Weiss (1981) once again, using similar arguments. Indeed, he suggests that lenders should design contracts to disincentive a false declaration of borrower types, and this is actually achievable through an increase in collateral requirements which implies a decrease in interest rates. However, a reduction of collateral availability caused by borrowers' wealth could lead to credit rationing and pooling of borrowers.

The same conclusion is suggested in Wang (2010), who similarly considers the possibility of different combinations of interest rate and collateral to sort borrowers, and then discusses the possible negative influence of collateral pledging on enterprise decisions of production. On the other hand, credit rationing equilibrium is recovered by Coco (1999). He argues that this conclusion holds if additional assumptions on entrepreneurs' risk preferences are made. In particular, he reconsiders the Stiglitz and Weiss' model to account for different risk attitudes of borrowers, as well as the effort required to accomplish the project. He proves that in this case, collateral might not serve as a signaling device either. This conclusion is in contrast with Bester (1985) because, when more risk-averse borrowers are associated with safer investments, the relation between their project risk and the marginal rate of substitution between interest rate and collateral may not be monotone, and this was actually one of the main assumptions made by Bester to construct its model. Nevertheless, in the last part of the analysis it is shown that the previous equilibrium implies at least a non-negative correlation between collateral and the risk of a project, and a possible positive link between interest rates and collateral.

As shown in Chan and Thakor (1987), the analysis presented in Stiglitz and Weiss (1981) depends on their assumptions made on the nature of competitive equilibrium. Indeed, they extensively discuss the role of collateral, depending both on assumptions made on the credit market and the influence of moral hazard and private information. Firstly, they examine a different hypothesis where *"all rents accrue to borrowers"* rather than to depositors, concluding that in this case

rationing no longer occurs. The intuition is that, if a perfect elastic deposit supply is presumed, now every borrower receives a loan. Secondly, they show how collateral can serve under asymmetric information. More collateral offered increases the expected borrower's loss, who as a consequence should increase his effort. On the contrary, a higher interest rate reduces the effort level: the high quality borrower contracts will have a lower interest rate and higher collateral.

Su (2010) observes that collateral is largely employed in loan contracts, hence the credit rationing as described by Stiglitz and Weiss hardly ever occurs in practice. Rationing equilibrium is under discussion in De Meza and Webb (1987) too. Indeed, differently from Stiglitz and Weiss (1981) where all borrowers' projects have an equal expected return, their model accounts for different expected returns between projects, leading to opposite results.

Other theoretical studies pick up the usefulness of collateral. In this regard, the effectiveness of collateral to reduce the information asymmetry is discussed by Broll and Gilroy (1986). They explain the dynamics of the credit market under asymmetric information, basing their model on Stiglitz and Weiss (1981) but with a greater focus on collateral requirements rather than on the interest rate. The main results are the same, but their different derivation allows a deeper examination of collateral requirements. As a matter of fact, they highlight that, contrary to the case of an efficient market when an increased demand (in this case, for credit) results in a higher equilibrium price (the interest rate), here an increase in the interest rate could lead to an adverse selection mechanism, such that low-risk borrowers might be expelled from the market, contrary to the interest of banks who would suffer major risk. This happens because, reasoning in terms of a borrower's expected profit, if the interest to pay is raised then the borrower has to find ways of increasing its expected value, thus increasing its risk to compensate. Just in this case, collateral would serve as a *"market-clearing"* device. Their analysis, focusing on the collateral variable, allows them to find a critical value $c = c^*$ for the collateral amount such that the borrower's expected profit is non-negative. An average risk σ^* is related to this value, and since they prove that $\frac{\partial \sigma^*}{\partial c} > 0$, adverse selection is

demonstrated. A positive relationship between risk and collateral is given in Chen (2006), who develops a model in order to take into consideration also the timing when collateral should be pledged.

With a slightly different approach, Chan and Kanatas (1985) investigate whether the existence of collateral is justified outside a moral hazard framework. They examine a situation of asymmetric payoff valuation between borrowers and lenders. In this case the information itself is the same, but the players of the contract have different opinions or beliefs regarding their expected payoff. This difference may exist because, denoting by x the outcome of an investment made with the borrowed capital, the expected value for the borrower (b) depends on his subjective belief, that can be represented through a cumulative distribution function $F(x)$:

$$\mathrm{E}_b(x) = \int_a^k xF(x)dx \tag{2.1}$$

for appropriate a, k, which is in general different from

$$\mathrm{E}_l(x) = \int_a^k xG(x)dx \tag{2.2}$$

where $G(x)$ is the cumulative distribution function associated with lender's belief (l). The lender is interested in this outcome because default occurs when the terminal value of the project is lower than the debt, and in this case the lender collects entirely this value. Under these assumptions, collateral is shown to be useful as a signaling device if the lender evaluates the project less than the borrower, due to the trade-off between higher collateral amount and lower interest rate taking place when asymmetric beliefs exist. Therefore, according to this model better borrowers should pledge a higher amount of collateral, in order to take advantage of a lower interest rate.

Contrarily, Flatnes and Carter (2016) propose a model in order to take into consideration the moral hazard problem. They focus on group lending contracts and demonstrate that moral hazard is heavily reduced through collateral requirements. Another specific case is examined in De Meza and Southey (1996). When the bank has more information about the risk of a project, for example in the case of a start-up, higher collateral is required from high-risk borrowers. Rajan

and Winton (1995) propose a model where collateral requirements increase as the borrower experiences financial difficulties.

Another point of discussion has been the substantial difference in the collateral value for borrowers and for banks, constituting a disincentive to request them. In fact, this is the case where collateral value is C for the borrower and $\beta \cdot C$, $\beta \in (0, 1)$ for the bank, for example because of transaction and liquidation costs. This issue is examined in Besanko and Thakor (1987), both for credit market equilibrium under monopoly and under perfect competition. Differently from Chan and Thakor (1987), they assume that collateral is not costless. They argue that, since the use of collateral is costly, a monopolist would rather extract all the borrower's surplus simply by increasing the interest rate. This result is obtained under the hypotheses of a universal risk-neutral economy, full information and collateral infinitely accessible to the borrower. Their analysis is at odds with common sense, since a riskier borrower would pay a lower interest rate compared to a safer borrower. But if asymmetric information exists, as it usually does, and if the market is competitive, the authors offer a different conclusion. The use of collateral in this case becomes a useful tool to characterize borrowers since interest rates are supposed only to cover the contractual costs and collateral is used as a disincentive for riskier borrowers from choosing the contract designed for low risk borrowers.

Another model which takes these dissipative costs into consideration is developed by Booth et al. (1991), accounting both for moral hazard and private information issues. In fact, their analysis is influenced by this hypothesis along with a more specific one, which is a distinguishing feature of their model: for any given action there is a higher margin of improvement for the bad borrower than for the good one. Among their other hypotheses there are risk-neutral borrowers, perfect competition in the credit market, a set A of two possible borrower's actions $\{\bar{a}, \underline{a}\}$ associated with different costs and which, together with the borrower type "bad" or "good", are the dependent variables for the success probability. Therefore their approach is different from Stiglitz and Weiss (1981) too, since their focus

is on the quality (the type) of borrowers rather than on their wealth. Under moral hazard, when the bank actually knows the borrower type, they conclude that collateral is generally required from bad borrowers in order to increase their loss in the case of default, consequently inducing them to choose a higher effort. The combination of an increased collateral requirement along with a higher interest rate could be more appropriate than a greater increase only of interest rates, as in the case of *"usury laws"* (Chan and Thakor, 1987) or *"accepted social norms"* which put an upper bound on the interest rates (Coco, 1999). Instead, if the bank does not know the borrower type, they achieve different results. In this last case the best action for the bank is to ask borrowers to truthfully report their type (Myerson, 1979), and then bad borrowers take advantage of the better contractual conditions offered to good borrowers. Then collateral is required both from bad and good borrowers: the positive relation between higher risk and higher collateral requirements is still not straightforward.

From a different point of view, a related discussion (Benjamin, 1978) concentrates on the additional costs for stipulating secured loan contracts. Despite these costs, it is shown that collateral can be useful in order to enforce debt contracts, as previously discussed in Barro (1976). The analysis of Bieta et al. (2008) stands apart. They challenge all the past theoretical models which investigated the role of collateral as a signaling device. The authors state that, under assumptions different from those assumed by the previous literature, collateral cannot serve to reduce the typical asymmetric information of loan contracts. In particular, they propose a model to take into consideration the continuous outcome for the risky project of borrowers. They analyze two borrowers' class, where the first class is less risky and its projects exhibit second order stochastic dominance over the second class. Under the previous assumptions, they prove that the expected payoff for the first class, in presence of collateral, is less than or equal to that for the second class, which is contrary to the bank interest due to the assumption on the risk of each class (as stated in Stiglitz and Weiss, 1981; in fact, credit rationing is recovered here). Therefore, the only case where collateral

is useful is when an equality between the two expected payoff holds, which is rare. Although this model lies on less restrictive assumptions compared with other works, it is worth noticing how it does not explain empirical results found in the plentiful previous literature and in practice, where collateral requirements are usual.

2.2.2 Empirical literature

Along with the prevalent theoretical approach, other authors have also taken advantage of an empirical analysis to assess their models, discussing the previous contrasting literature and testing different samples to verify which hypothesis was observed in practice.

In this way, the assumption of a direct or an inverse association between collateral and risk is experimentally tested in Berger and Udell (1990). The authors present a cross-section analysis to verify which hypothesis is more frequent, using a sample of over one million commercial loans from the Federal Reserve's Survey of Terms of Bank Lending. They regress the loan risk premium on collateral and other control variables at different periods, finding more frequently a positive coefficient for the former regressor. Their conclusion is that collateral is generally related with riskier loans and riskier borrowers. The authors here, contrary to Bester (1985) but in accordance with other studies (Degryse and Van Cayseele, 2000, Harhoff and Körting, 1998), have assumed a sequential decision for interests and collateral, where the latter is chosen before the former. This avoids the simultaneity problem in the estimation.

Almost contemporarily, a similar conclusion is reached by Leeth and Scott (1989). Taking a sample of 1,000 US small business loans, they find a higher default probability and a greater loan size and loan maturity, all risk indicators, for secured loans. Later, Angbazo et al. (1998) find empirical support for this conclusion, using over 4000 loan transactions registered on Loan Pricing Corporation's database, between 1987 and 1994.

A confirmation of this result is obtained in Jiménez and Saurina (2004). In

analyzing the determinants of the Default Probability (PD) of bank loans, they discuss extensively the role of collateral in this context. Their empirical results, using data from the *Credit Register of the Bank of Spain (CIR)*, suggest a higher PD for collateralized loans.

However, the conclusion is opposite to the one reached by Degryse and Van Cayseele (2000), who use data from Belgian banks, and find a negative relationship between interest rates and collateral, even if the latter is decreasing as the duration of the bank–firm relationship increases. This conclusion is supported by Capra et al. (2005). They use a sample of small and medium-sized firms in Valencia to test the role of collateral, and their analysis supports the hypothesis of a negative relationship between collateral and interest rates, which is the contract chosen by lower risk borrowers. Nevertheless, moral hazard is shown to affect the initial choice of the contract, weakening the obtained link.

Similar conclusions can be found in Janda et al. (2003), who along with an examination of the agricultural credit markets in the Czech Republic, develop a model where low-risk borrowers are subject either to collateral requirements or to rationing.

Breit and Arano (2008), consider the determinants of the price (interest rate) applied to small businesses, and hypothesize a lower risk premium if the dummy variable "Collateral" takes the value 1 (collateral required). Then they find empirically a lower interest rate when a collateral secured the loan, as in their estimation this dummy predictor is statistically significant. Similarly Agarwal and Hauswald (2010), using 2002 and 2003 US data for SME, find a negative and significant coefficient for collateral predictor regressed on the offered loan rate.

On the contrary Lehmann and Neuberger (2001), using a dummy variable for collateral too, analyze 1988 US data for 174 lines of credit, finding a positive relationship. Similarly, Pozzolo (2002) proposes both a model and an empirical analysis to justify a positive relation between interest rate and collateral. The main assumptions of his model are the different collateral value for the borrower and the lender, and the inclusion of a moral hazard problem, where borrowers

maximize their payoff choosing their effort. He argues how under these conditions an equilibrium between the borrower's profit for a given effort and the expected return of the bank (equals to the risk free investment) exists: solving the system leads both to a higher interest rate and higher collateral value for riskier contracts. This result is the same of Booth et al. (1991), who make similar hypotheses about moral hazard and costly collateral. Then, using data of bank loans to Italian non-financial firms, the author finds that secured loans are considered ex-ante riskier by banks. Brick and Palia (2007), using data from the 1993 *National Survey of Business Finances* find a higher risk premium related to collateral pledging as well. John et al. (2003) had reached the same conclusion analyzing US data from the *Securities Data Corporation*. Collateral is shown to reduce the higher credit risk also in Thailand, through an empirical analysis carried out by Menkhoff et al. (2006). Furthermore, their incidence is shown to be higher than in mature markets. Booth and Booth (2006) reach the same conclusion, analyzing data from the Securities and Exchange Commission (SEC) of loans contracts stipulated from 1987 to 1989.

A dissenting view is from Elsas et al. (2000). Examining data from five of the most important German banks, they do not find any relation between ex-ante borrower quality and collateral pledged. The same result was previously found in the empirical work by Machauer and Weber (1998) on German banks.

On the contrary Berger, Espinosa-Vega, Frame and Miller (2011) test whether collateral can be seen as a device to reduce asymmetric information, or equivalently, if a decrease in asymmetric information is related with a reduction in collateral requirements and answer positively to the question, using an original comparison of outcomes. They compared results that preceded and followed the introduction of a specific survey (the *Federal Reserve's Survey of Terms of Bank Lending*), which reduced the *ex-ante* private information: with its usage, collateral requirements lowered.

Weill and Godlewski (2006) highlight the conflicting literature on the theme, explaining the dissimilar conclusions provided by the authors as a different degree

of asymmetric information among countries, using a sample of 5843 loans from 43 countries. In particular, they analyze the link between loan risk premium and collateral: a simple OLS regression is employed concerning interest as a function of collateral, the degree of information asymmetries for each country and some other variables. Financial, accounting standards and economic development indicators are used as proxy for the degree of information asymmetries. The estimation results in a significant positive coefficient for the collateral variable, apparently not supporting its use as a device to solve adverse selection issue. However, a further analysis shows that this positive link is weakened by an increase in information asymmetries.

Differently, Berger, Frame, Ioannidou et al. (2011) suggest as a possible explanation both the different economic characteristics and the types of collateral. Then they conduct an empirical study on commercial loans of Bolivian financial institutions, and find a negative relationship between collateral amount and risk-premium, which is explained through a lower bank's loss in case of borrower's default. A similar conclusion, but with a different sample, is given by Blazy and Weill (2006), analyzing the role of collateral circumscribed to French banks and discussing how guarantees reduce loan loss when default occurs. They conclude that collateral could help to solve adverse selection problems.

A dual relationship is found in Calcagnini et al. (2009). Through an empirical analysis on Italian banks, they obtain different results depending on the nature of borrowers: using the interest rate spread as the dependent variable for a linear model, a positive coefficient for collateral predictor is estimated for firms, and a negative one for consumer households. Therefore they conclude that guarantees are used to reduce adverse selection problems for consumer households and as an incentive to mitigate moral hazard for firms.

Finally, Ono et al. (2012) examines both *ex-ante* the risk of borrowers who pledge a collateral and how guarantees influence firm *ex-post* performance using a sample from the Japanese *Surveys of the Financial Environment* (SFE). They find that guarantees are more likely to be pledged by high risk firms and that their

one year *ex-post* performance is better than firms which have not secured their loan. Specifically, they find a cost-cutting effect, which comes out in favor of the moral-hazard reduction effect of collateral, through an observed higher managerial effort in order to reduce the probability of default. A similar conclusion, but only theoretically, has been already presented in Bester (1994), who proposed an analysis where in high-risk projects there is a greater usage of collateral, pointing out that the probability of bankruptcy is reduced as collateral acts like an incentive. In agreement, collateral is implicitly assumed to be a feature required for riskier borrowers in Peydró (2013), since periods characterized by a less strict policy in granting loans are associated with a weakening in collateral requirements.

Chapter 3

Models for the relation between collateral and interest rates in loan contracts

3.1 Introduction: A reductionist approach to the problem

In this chapter, our main aim is to prove the *a priori* impossibility to decide for one unambiguous relation between collateral and interest rates of loan contracts.

We use both mathematical and intuitive arguments. Specifically, we show that different hypotheses, related to three different scenarios — symmetric information; moral hazard; adverse selection — lead to contrasting conclusions with regard to the link between collateral and interest rates. Since these general assumptions are equally plausible, the impossibility to choose which hypothesis and, hence, which conclusion is true ends our proof. Indeed, from the previous chapter, it is evident that the contradictory conclusions provided by many authors depend exclusively on their different assumptions. We highlight that these assumptions are usually rather similar, except for some details or for an additional hypothesis, but this does not prevent completely contrasting conclusions. For example, different results are obtained if borrowers' wealth is considered variable or fixed, or if the collateral usage is assumed costless or onerous.

We focus on a more general approach to tackle the question we are dealing with. The difference between our approach and the others' is the focus on some general financial and game-theoretic principles rather than on a specific aspect of the loan contracts. Moreover, our approach is mainly probabilistic, because we analyze general relations rather than a specific contract, as it befits our general discussion. Due to these features, the following analysis does not propose a unique solution. Nevertheless, this does not mean that our models are inadequate or ambiguous, because they actually achieve a unique most probable solution depending on the specific assumptions. In particular, we prove that with perfect information the presence of collateral implies a lower interest rate, but if asymmetric information comes into play the effect of a collateral requirement on the loan interest rate is no longer straightforward. Our discussion can be adapted to include more specific hypotheses, but this seems not to change our conclusion on the ambiguity of this link, since considering more precise hypotheses would change the results as seen in the literature review from Chapter 2.

3.2 A theoretical analysis under symmetric information

3.2.1 Our approach to the problem

When a bank, or in general a creditor, stipulates a loan with a borrower, it must decide the interest rate i to apply, taking both macroeconomic (the Euribor interest rate, GDP growth rate forecasts, etc.) and microeconomic variables into consideration. If we focus on the latter ones, the information available to the bank will result in a decision $D : \boldsymbol{x} \longrightarrow i$, which is a function of the borrower's variables $\boldsymbol{x}$. This function should reflect basic principles of economics and ultimately of human intrinsic nature (assuming a rational behavior). For example, a higher borrower's risk should result in a higher interest rate. Generally, the interest rate can be a function of the borrower's risk, the length of the contract, the amount borrowed, etc. Hence $D(\mathbf{x}) = i$ for an appropriate $\mathbf{x}$, which is, depending on the subject, the vector of regressors, independent variables, predictors, covariates, etc.

Sometimes, one variable x_j (assume x_j continuous) in the set $\mathbf{x}$ has a straightforward relation with the dependent variable i (called also *image, label, output,etc*):

$$\frac{\partial D}{\partial x_j} \geq 0 \quad \textit{or} \quad \frac{\partial D}{\partial x_j} \leq 0. \tag{3.1}$$

Usually, however, the effect of a variable is not *a priori* clear-cut. This is the case of collateral, for which it is unclear whether its value has a positive or a negative link with the interest rate applied. In fact, from an immediate perspective, the relationship should be negative, because borrower's risk should be lowered if collateral can be collected in case of default. But this simple interpretation does not necessarily hold, since usually the decision over the interest rate is made under asymmetric information, and the usage of collateral is largely influenced by this asymmetry.

To tackle this problem using game theory, the starting point in our analysis is to assign a role to each player, according to the well-known principal-agent model. Indeed, this model accounts for the main features of loan contracts stipulation under asymmetric information. Firstly, the payoff of the player called *principal* depends on the *actions* taken by the player called *agent*. Secondly, the principal designs the contract considering an *incentive mechanism*, in order to encourage the agent to fulfill the contractual conditions. The agent considers this incentive, along with his expected payoff and the cost of performing an action, and chooses his strategy in order to maximize his payoff.

In our analysis the bank is the principal which gives an agent (the borrower) a sum L and expects a net payoff which depends on the interest rate i. Here we assume that the bank decides the contractual conditions, and the agent receives the sum borrowed and is required to put an effort e to invest the borrowed capital. The analogy between the situation of a loan stipulation and, for example, an employment situation is that the principal gives a sum to the agent, which is a salary in the latter case and the amount borrowed in the former, and there is an agent's effort e, that can be seen as the work he performs in exchange for the salary, or the loan, increasing the payoff of the principal. Therefore, this is

a bilateral relation, established through a contract designed by the creditor, which the borrower can decide to sign, and whose final result depends critically on the agent's effort. In case of default, the borrower may not pay the borrowed capital back or, at least, we assume that the bank does not earn the contractual interest, similarly to the situation of an employee who, failing to fulfill his task, causes a loss to the principal. However, in some cases a difference arises in the bank-borrower contract, which should be taken into account and which may not be observed in the principal-employee contract, that is a common interest in the final result for the principal and for the borrower, since the latter is supposed to invest the borrowed capital in a project. In any case, the effort e is always costly for the borrower. This is a fundamental point, considering the link that it has with collateral, as discussed in Section 3.3.

In the following discussion, using a founding principle of finance and introducing the critical effort concept, we will be able to discern between risk and collateral, and to discuss what happens to the interest rates-collateral link, given a fixed risk. In fact, we can ideally split the total risk in two parts. The first part derives from the risk of any investment made using the borrowed capital, and can be priced through a higher interest rate. The second part derives from the association between a higher probability of default caused by a lower effort supplied by the borrower. Then, we discuss how asymmetric information can change the relation interest-collateral.

A presentation of the general models that some of our ideas originate from can be found in Macho-Stadler and Pérez-Castrillo (2001). We have extended them to the specific case of loan contracts, and we have adopted a more probabilistic approach. However, Section 3.2.2 and our line of reasoning regarding loan in this chapter are original.

3.2.2 A theorem based on the risk-return trade-off

This section presents a general model under symmetric information. Our main assumptions are:

1. The effort is induced by the principal in order to maximize his utility.

2. The contract is accepted by the agent if a precise condition is verified.

3. The principal can measure the agent effort e. In other words, the quantity e can be included in the contract terms. The relevance of this remark will be discussed later.

4. We consider a single period relation.

We begin defining the following concept:

Definition 1 (Critical effort)**.** *The critical effort $\tilde{e}$ is the minimum effort required for the borrower (the agent) to avoid default.*

We assume only two possible states of the world: $s_k \in S$, $k = 1, 2$, where

$$S = \begin{cases} s_1 = \text{"Default"} \\ s_2 = \text{"Solvency"} \end{cases} \tag{3.2}$$

Therefore, the bank has two possible related final payoffs: one for the state of the word s_1 and another for the complementary case. The realization of S depends on $\tilde{e}$, but since the critical effort is defined *ceteris paribus*, there are other factors which can influence the realization of a specific state of the world. These quantities are stochastic, thus they influence the conditional probability

$$\Pr(S = s_k|\tilde{e}). \tag{3.3}$$

The optimization problem for the principal becomes

$$\max \sum_{k=1}^{2} \Pr(S = s_k|\tilde{e}) \cdot U_1(\pi_k) \tag{3.4}$$

where π_k is the final payoff associated with the specific state of the world k, which depends also on the effort required to the borrower, (3.3) must be estimated by the principal in order to compute his maximum expected payoff and U_1 is an appropriate utility function associated with the lender; U_2 is the utility function associated with the borrower. This maximization problem is subject to the

restriction:

$$\sum_{k=1}^{2} \Pr(S = s_k|\tilde{e}) \cdot U_2(W) - c(e) \geq A \tag{3.5}$$

where $c : e \longrightarrow c(e)$ is the borrower's cost function. The second term of the inequality, A, is the so-called *"reservation utility"*, and intuitively represents the agent's expected utility of alternative investment opportunities, and it has to be lower than the first member. This restriction is assumed both necessary and sufficient to ensure that the contract is signed by the borrower. The argument W of the function U_2 includes both the sum borrowed L and the payoff of any external investment made with this sum, or any quantifiable additional utility obtained by the borrower using the sum L.

Under this framework we prove that the presence of collateral implies a lower interest rate.

The effect of the absence or the presence of collateral of value C on the bank's final (gross) payoff π, or equivalently except for a sign factor on the borrower's payment, can be represented through piecewise-defined functions:

$$\pi = \begin{cases} 0 & \text{if } S = s_1 \\ L(1+i) & \text{if } S = s_2 \end{cases}$$

$$\pi_c = \begin{cases} C & \text{if } S = s_1 \\ L(1+i') & \text{if } S = s_2 \end{cases}$$

or, for loans which are both secured but differ for their collateral value:

$$\pi_{c'} = \begin{cases} C' & \text{if } S = s_1 \\ L(1+i) & \text{if } S = s_2 \end{cases}$$

$$\pi_{c''} = \begin{cases} C'' & \text{if } S = s_1 \\ L(1+i') & \text{if } S = s_2 \end{cases}$$

with $C'' > C'$.

At first, let us analyze the principal's viewpoint. Clearly, if the agent does not sign the contract the game does not come into being, hence we proceed analyzing the case where the contract is signed, that is when (3.5) is satisfied.

Recall that one of the basic principle of finance is the risk-return trade-off, and besides, we assume that economic agents are risk-averse.

As far as the risk-return trade-off is concerned, even if from a theoretical point of view an injective relation between risk and return can be assumed, this hypothesis can be relaxed in order to improve the effectiveness of our conclusions. Our solution arises from a simple consideration. Theoretically, for each borrower, the preference curve f which maps σ (the risk) into r (the return) is strictly increasing. However, in practice some problems occur. In fact, σ is not observable and the domain is not continuous, because borrowers are labeled with risk-classes, for example rating-classes, which are countable, thus the domain is only *approximately* continuous. Therefore, we should consider $\hat{\sigma}$, the estimation of σ, and make the domain of f continuous, for example by using an interpolation technique. More importantly, when different borrowers are considered together, the injectivity is violated, because for each risk-class there is a point cloud where each borrower's point could have a different r-value. The consideration which solve this problem is that the mean behaves better than a single observation when looking for a general principle. This principle is applied, for example, in econometrics using regression models, or in statistical mechanics considering the behaviour of a certain number of particles, rather than their individual position, in order to formulate general laws. In other terms, we can recover an injective function by computing the barycenter $\bar{r}$ of the point cloud for each risk-class, and constructing an interpolation of these barycenters. We call g the continuous function derived through this procedure, which can be made bijective by appropriately restricting its codomain to its range $R = g(\hat{\sigma})$ (we use the conventional notation $g \triangleright R$ to indicate this restriction). Now it is acceptable to assume $\dfrac{\partial g}{\partial \hat{\sigma}} > 0$. Through this hypothesis, we prove that if collateral is required to each borrower, the interest rate must decrease.

Moving from the previous considerations, we notice that the presence of collateral, or of collateral with a higher value, decreases the expected loss of the bank if default occurs, or equivalently it increases its expected return. Thus, given a fixed expected maximum return without collateral, the same value can

be obtained applying lower interest rates if collateral is provided by borrowers. Similarly, considering a contracts whose collateral value is $C'' > C'$, the interest rate should be lower in order to maintain the principal's expected maximum return constant, that is if $C'' > C'$, then $i' < i \Leftrightarrow L(1+i') < L(1+i)$, being $L > 0$. Moreover, the borrower's risk is not changed by the presence of collateral under symmetric information, because in this context the risk is independent from collateral. But if risk with or without collateral is the same, because borrowers are the same and there are not adverse selection problems, collateral and interest rates can be considered as substitutes to price the risk borne by the bank. Therefore, when risk is fixed, if the presence of collateral increases the expected return, a contradiction of the risk-return principle arises. Thus, in order to obtain the same return, the interest rate applied must be lowered.

Given the previous general assumptions, we have proved the following

Theorem 1. *Let $C \in (0, \infty)$ be the value of collateral and i the interest rate of a loan contract. When collateral is required to each borrower, $\dfrac{\partial i}{\partial C} < 0$ holds .*

Proof. Under perfect information, collateral is not used to solve moral hazard or adverse selection problems. Therefore, from the previous discussion, the estimated risk $\hat{\sigma}$ of each borrower does not change if he provides collateral with value C. Since the restriction $g \rhd R$ is bijective, the expected total return r must be unique for a given risk. But a higher C increases r, then the interest rate i must decrease. □

It is important to underline that all these conclusions are derived *ceteris paribus*, that is fixing all the other variables that could affect the interest rate decision. We are not considering, for example, the bank strategy: the bank might choose higher interest rates *and* higher collateral requirements at the same time. But from an abstract point of view, and using only the basic financial principles, this is not possible. In other words, of course banks prefer higher collateral requirements and higher margins, but in an efficient and ideal market, with no free-lunches, perfect competition, etc, this is not feasible. A bank is rewarded only for the risk it assumes, because risk is priced, contrary to its subjective decision about the

Collateral-Interest proportion (similarly to the enterprises' choice of Equity/Debt ratio).

On the other hand, considering the situation from the agent's point of view, when the principal requires additional collateral with value C, this term must be added to (3.5) with a negative sign. But this term could change the inequality direction, such that

$$\sum_{k=1}^{2} \Pr(S = s_k|\tilde{e}) \cdot U_2(W - C) - c(e) < A. \tag{3.6}$$

In this case, the borrower does not sign the contract. But, as long as the risk of the borrower is supposed constant, this condition is not acceptable because the bank would give up on the expected payoff that would have obtained without collateral. In fact, under the previous hypotheses collateral should increase the expected payoff of banks. At the same time, under perfect competition, the agent could sign the contract with other banks for which (3.5) is verified. These banks exist as far as their remuneration is in line with the market risk premium associated with the specific risk of the borrower. The only way that a bank has to avoid this situation, verifying (3.5), is by means of a lower interest rate. Notice that interest rates are included in (3.5), because we can consider the net borrowed capital, which includes the interest to be paid, and is higher when interests are lower. Again, the bank cannot simultaneously achieve a higher i and a higher C for a given borrower.

3.3 Asymmetric information

3.3.1 The influence of moral hazard

In the previous discussion we have examined an ideal situation, where the loan contracts are not affected by asymmetric information problems. In this section we extend the previous analysis introducing these asymmetries. Firstly, we discuss how moral hazard influences our conclusions. The moral hazard problem occurs when agents' behaviour is not observable, or at least it is not verifiable by the principal. In particular, we consider the situation where the borrowers' effort e is not observable, or verifiable, after the conclusion of loan contracts. As discussed

above, this effort implies a cost $c(e)$ for the agent, then it decreases its payoff, and since e is not observable the borrower could take advantage of the situation by reducing it. Moral hazard is a widespread problem; for example in the insurance industry, insured individuals may reduce their effort to prevent the insured event (see Campbell, 2006). While in the previous section e was a fixed quantity, now the agent actively chooses the effort, because it cannot be contracted upon. Then the agent maximizes his utility function taking e into consideration, that is (using the previous notation):

$$\arg\max_{e} \sum_{k=1}^{2} \Pr(S = s_k|\tilde{e}) \cdot U_2(W) - c(e), \forall e > 0. \tag{3.7}$$

If I is the value of any investment made by the borrower with the sum L, the loan balance, the argument of the utility function U_2 contains $L \leq L + I$ if $e < \tilde{e}$, because we assume that $I = 0$ in the case of default.

In this scenario a collateral requirement can be used by the bank (the principal) to force the borrower and increase e. Considering this new condition, when $e \leq \tilde{e}$ the borrower has an incentive to increase its effort because his loss is higher. The mathematical representation of this idea is simple, because the presence of collateral just changes the maximum condition into

$$\arg\max_{e} \sum_{k=1}^{2} \Pr(S = s_k|\tilde{e}) \cdot U_2(W') - c(e) \tag{3.8}$$

where W' includes now the quantity $-C$ (we denote with k, k' all the other profits/costs associated with each case, which does not influence this discussion):

$$W' = \begin{cases} L - C + k & \text{if } S = s_1 \\ L + I + k' & \text{if } S = s_2 \end{cases}$$

To examine this condition, we take advantage of a simplifying assumption, which is not new to game theory modeling, adapting it to the specific case of loan contracts: we assume that the effort belongs to a binary set, so it can be either High or Low. Formally, $e \in \{e^H, e^L\}$. Given this new hypothesis, $c(e^H) > c(e^L)$, because $c(e)$ increases as the effort increases. The principal prefers the higher effort, e^H , and

computes the probabilities associated with each result, for a given effort: $\Pr(S = s_k|e^H)$ and $\Pr(S = s_k|e^L)$. Clearly, a lower effort is more likely to produce default (D):

$$\Pr(D|e^L) > \Pr(D|e^H). \tag{3.9}$$

This explains the principal's preference for the higher effort. If the agent maximizes his expected net utility choosing a lower unverifiable effort, the principal should design the contract in order to shift the agent's choice from e^L to e^H. However, the value I must be taken into consideration too. Therefore the inequality

$$\sum_{k=1}^{2} \Pr(S = s_k|e^H) \cdot U_2(W') - c(e^H) \geq \sum_{k=1}^{2} \Pr(S = s_k|e^L) \cdot U_2(W') - c(e^L) \tag{3.10}$$

is not always verified, and depends on the additional payoff given by I — which is more probable to exist when $e = e^H$ — compared with the higher cost associated with e^H. In all the cases where (3.10) holds, the equilibrium will be at e^H. To sum up, if both the principal and the agent would choose e^H, without incentives, there is no problem. But if the agent is prone to choose a lower effort, then the principal could design the contract to obtain a higher effort, and collateral comes into play. As a matter of fact, in our specific case the previous condition can be rewritten using the variable W' in comparison with the new variable W^* defined as:

$$W' = \begin{cases} L - C + k & \text{if } S = s_1 \\ L + I + k' & \text{if } S = s_2 \end{cases}$$

$$W^* = \begin{cases} L + k & \text{if } S = s_1 \\ L + I + k' & \text{if } S = s_2 \end{cases}$$

Clearly, the former variable is more likely to encourage an agent to choose e^H, simply because the loss for a borrower is higher in the case of default. In fact, the agent does not always choose a higher effort even when collateral is pledged, because there are other variables which influence his decision, that is why we say *"is more likely"*. Nevertheless, under appropriate conditions, which are different for each contract and for each borrower, the incentive is effective. More specifically,

in every situation where the higher cost associated with default is greater than the higher cost of choosing e^H, an agent is more likely to choose e^H. The variable quantity I can also be assumed to be a positive function of e. In order to account for this uncertainty, we say that the frequency of the borrowers which satisfy (3.10) is higher when collateral is pledged.

To formalize this intuition, we take advantage of the frequency φ, which gives the idea of a more frequent relation rather than of a relation that is always true for each borrower:

$$\varphi\left[\sum_{k=1}^{2}\Pr(S=s_k|e^H)\cdot U_2(W')-c(e^H)\geq\sum_{k=1}^{2}\Pr(S=s_k|e^L)\cdot U_2(W')-c(e^L)\right] \quad (3.11)$$

is higher than

$$\varphi\left[\sum_{k=1}^{2}\Pr(S=s_k|e^H)\cdot U_2(W^*)-c(e^H)\geq\sum_{k=1}^{2}\Pr(S=s_k|e^L)\cdot U_2(W^*)-c(e^L)\right] \quad (3.12)$$

that is, the inequality sign $\geq$ holds for more borrowers when collateral is pledged. This explains why, under moral hazard, collateral can be associated with a higher interest. As a matter of fact, higher interest rates and higher collateral requirements are compatible, though not always necessary, if they are both considered as a penalization for a riskier borrower who prefers e^L to e^H. The effect of moral hazard on collateral has been already analyzed, but with a different derivation, by Flatnes and Carter (2016) and Booth et al. (1991). The main difference in our model, as discussed in the introduction of this chapter, is its generality. Another difference is that, considering the totality of contracts rather than a single loan, we reason in terms of frequency, which fits better our more general discussion. Nevertheless, our results are consistent with these papers.

3.3.2 The influence of adverse selection

In addition to moral hazard, in this section we examine a different problem related with asymmetric information, called adverse selection. Consider the situation where each borrower has more information than the principal regarding his own quality. To formalize this situation, we make use of a binary set, and

we suppose that the quality q of a borrower can be either *good* (G) or *bad* (B). A good borrower is an agent who is less likely to default. On the contrary, a bad borrower is an agent associated with a higher probability of default. We assume that the quality is not influenced by the effort e discussed in the previous section. For example, a higher probability of default may be caused by a riskier borrower (who invests in riskier projects).

In this case different agents should get different contractual conditions, and in particular better agents should obtain better conditions in terms of the interest rate. But this is not easily achievable if information is not perfect, since the bank does not know which are the true good borrowers. On the other hand, bad borrowers try to take advantage of better contractual conditions, obfuscating their true category. At the same time, an adverse-selection effect takes place if banks raise the interest rate, because this leads to exclude lower risk borrowers and to retain those who are riskier. In this case banks could design the contract in order to discern and select between good and bad borrowers, requiring a signal. Obviously, the contract should be designed so that good borrowers benefit from signalling their true type.

This concept is very intuitive: if an agent proves his good quality to the principal, then he can take advantage of its features, otherwise he achieves a lower utility. In the specific case of loan contracts, we can examine if a good agent has the possibility to obtain better contractual conditions, that is a lower interest rate. We prove that this question has an affirmative answer. The main point here is that borrowers are not interested in revealing their true type (=quality) if they can obtain better conditions by cheating on their quality, and they are prone to reveal this information if they can take advantage of it.

Above in this chapter, we have discussed how the borrower's payoff is changed when he provides collateral, representing with the variables W' and W^* the two different situations of absence and presence of collateral.

In this case, if an agent regards himself as a good borrower ($q = G$), in order to compute his expected payoff he uses a (strictly) lower weight $\Pr(S = s_k|q = G)$ for

the sum associated with the event D than a bad borrower:

$$\Pr(D|q = B) > \Pr(D|q = G). \tag{3.13}$$

If adding collateral, or equivalently increasing its value, allows the borrower to prove his good quality, and to obtain better contractual conditions, he will use it as a signal, benefiting from a lower interest rate. On the other hand, providing collateral of higher value may not be the optimal choice for a bad borrower if the expected cost of this signal is higher than the lower interest paid. On the other hand, from the principal's point of view, collateral is used in order to distinguish better borrowers and to apply them a lower interest rate, therefore, *ceteris paribus*, when collateral is provided, interest should be lower. Thus, collateral can be seen as a tool that mitigates the adverse selection problem. Using the previous notation, when a lower interest rate does not compensate the borrower's higher expected cost of providing collateral, and assuming a common cost function $c(e)$ for all borrowers, (3.13) implies

$$\sum_{k=1}^{n} \Pr(S = s_k|q = B) \cdot U_2(W') - c(e) < \sum_{k=1}^{n} \Pr(S = s_k|q = G) \cdot U_2(W') - c(e). \tag{3.14}$$

As in the general case, the presence of collateral is associated with a lower interest rate, but in this case the presence of collateral allows the bank to discriminate between good and bad borrowers. This result is consistent with Bester (1985), Bester (1987), Chan and Thakor (1987), but derived through a more general reasoning. Moreover, while the main results are the same, our discussion can be adapted to include any more specific hypothesis. For example, taking into consideration also the hypothesis of the difference in the collateral value for borrowers and for banks, Besanko and Thakor (1987) suggest the same conclusion. Indeed, this similar result is justified by the construction of our model, because even if $\beta \cdot C$, $\beta \in (0, 1)$, is the collateral value C for the bank, we have discussed how this tool has a direct effect only on the borrower's decision. In other terms, the value remains C for the borrower, hence the incentive or the signal effects remain effective.

When moral hazard and adverse selection are both present, the effect is not clear

if we consider the totality of contracts. As a matter of fact, we have discussed how collateral can be used as a tool in both cases, but it is associated with, respectively, higher and lower interest rate. Both e and the borrowers' quality q variables influence the probability of default (D). In fact we can write

$$P(S = D) = f(e, q, \cdot) \tag{3.15}$$

where $\cdot$ stands for all the other variables affecting this probability. Moreover,

$$P(S = D|q = B, e = e^L) \leq P(S = D|q = B) \tag{3.16}$$

because we have assumed that a lower effort increases the default probability; $P(a|b)$ is the conditional probability. These quantities are unknown to the lender, since q and e are unobservable under asymmetric information. Therefore, the decision on the interest rate cannot be easily associated with collateral requirements. In truth, the bank could achieve both a signal and an incentive raising at a proper level the latter quantity, but at least a negative effect arises at the same time, that is a loss of surplus. This loss is due to the lost contracts associated with those borrowers who consider the cost of additional collateral too high. For example, these borrowers could belong to the group (e^L, B), because without the second assumption on their quality, some of them would have signed the contract under only moral hazard, evaluating e^H profitable as discussed in Section 3.2.2. This loss of lender's profit is intuitive, since the bank should lose some payoff in a situation where it is penalized by lack of information. The results of this discussion are consistent with Booth et al. (1991), who similarly cannot derive a straightforward relation in the presence of both moral hazard and adverse selection.

In conclusion: higher interest rates can be associated with more collateral under moral hazard, because collateral requirements can be used as an incentive to increase the effort. Lower interest rates are compatible with more collateral under symmetric information and adverse selection. We have proved our thesis on the *a priori* impossibility to establish a unique monotone relation under general assumptions, because different hypotheses may reasonably apply.

In the next part of this work, we demonstrate empirically that these conclusions are supported by an analysis of the effect of collateral on the interest rate among some different European countries, obtaining contrasting results dependent on the country considered.

Part II

An econometric approach

Chapter 4

Principal Component Analysis for supervised learning

> [The Econometric Society] main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems and that are penetrated by constructive and rigorous thinking similar to that which has come to dominate the natural sciences [...]. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics.
>
> Ragnar Frisch

4.1 A bridge between supervised and unsupervised learning

In analyzing any type of data, researchers can encounter two categories, usually called labeled and unlabeled data. Suppose that a collection $\mathfrak{C}$ of data belongs to a set $\boldsymbol{\Theta}$ of a given dimension n and the relation $f : \boldsymbol{\Theta} \longmapsto \boldsymbol{\Phi}$ holds, then we call $\mathfrak{C}$ *unlabeled*, while the collection $\tilde{\mathfrak{C}} \in (\boldsymbol{\Theta}, \boldsymbol{\Phi})$ is called a *labeled* data set, whose dimension is $(n \times m)$ with m being the dimension of $\boldsymbol{\Phi}$. In other words, the case of two data collections may be considered, $x_i \in \boldsymbol{\Theta}$, $i = 1, 2, ..., n$, and $y_j \in \boldsymbol{\Phi}$, $j = 1, 2, ..., n$, and the map $\psi : x_i \longmapsto y_j$. Then, if the researcher can (or decides to) examine only the collection $\{x_i\}$, the independent variables are unlabeled because the dependent variables are not considered. On the contrary, if the pairs $\{x_i, y_j\}$ are available, some *labels* are *attached* to $\boldsymbol{x}$: this is why this is known as *labeled data* case, where the dependent variables can be employed to analyze the relation ψ. We highlight that even in the case of unlabeled data a relation can exist, and it can be either known or unknown, but the focus (or the availability) is on the dependent variable. It is also worth underlying that this distinction is not discriminating, as a greater data availability in the case of labeled collections might suggest, because unlabeled data can be used for various purposes, for example in the classification domain.

As a result of both this conceptual and practical distinction, an ensuing difference arises between supervised and unsupervised learning. We broadly call *supervised learning* the methods whose primary goals are to understand the underlying relation ψ between the dependent variable (usually a single variable y) and the independent variables (usually a vector of variables, $\boldsymbol{x}$). Usually in practice more than one realization for each variable is observed, thus we replace the dependent variable with a column vector $\boldsymbol{y}$ and the independent variables with a matrix $\boldsymbol{X}$, where each row represents a realization and each column a variable.

The most well-known supervised learning method is the regression analysis. On the other hand, the cluster analysis can be addressed as an unsupervised learning method, because it aims to analyze the latent structure of a dataset without considering the divergence between its input and output variables. Another

well-known unsupervised method is the Principal Component Analysis (also called PCA). This very useful, and commonly employed, technique aims to reduce a set of variables from a space Θ to a subspace spanned by the principal component vectors, that are orthogonal vectors constructed so that they preserve as much as possible the variation of the original dataset. Principal component analysis is not really a new technique. In its early form it was proposed by Hotelling (1933) and Pearson (1901), but its widespread use followed both the increased computing power and its development as a general statistical method (see among others, Rao, 1964). Moreover, it was modified in a great number of ways: for example in Chapter 6 we examine the Probabilistic PCA, a modification which can be particularly useful to analyze large sparse datasets. A comprehensive treatise of principal component analysis can be found in Jolliffe (2002).

Principal component analysis is commonly used as an unsupervised learning method, since it involves a linear transformation of the independent variables, but in truth it has been used in a supervised environment too, for example to improve the interpretability or to reduce the number of regressors. In our following application we examine in depth this use, achieving two levels of analysis at the same time. On one hand, we are showing its differences compared to the standard regression model, discussing its effects and problems related, for example, in selecting of a subset for the principal components. At the same time we are going to adapt and to take advantage of these econometric tools to address the problem about the analysis of loan borrowers' variables presented under the general framework of the game theory. Therefore, the successive analysis provides some empirical evidence about the relation between collateral and borrowers' risk as modeled in the first part, and a further analysis about other borrowers' variables.

4.2 Principal component analysis background

4.2.1 Introduction to the Principal Component Analysis and algebraic derivation of principal components

Let $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times v}$, $n, v \in \mathbb{N}$, the matrix of data whose j-th column, $j = 1, 2, .., v$, is the column vector of observations for the j-th variable and whose i-th row, $i = 1, 2, .., n$, refers to the row vector of variables for the i-th observation, denoted by x_i. Denote with $\mathbf{\Sigma}$ the square matrix of Variance-Covariance for the random variables x. As stated above, this method looks for a linear transformation of the original variables that sequentially maximizes the variance of the transformation, imposing the orthogonality of the new vectors found, that are called *principal components*. Let α be a vector of v constants, $\alpha_{1,j}$, and consider the linear combination $\sum_{j=1}^{v} \alpha_{1,j} x_j$ and its variance

$$\mathrm{Var}\, \boldsymbol{\alpha'}_1 \boldsymbol{x} = \boldsymbol{\alpha'}_1 \mathbf{\Sigma} \boldsymbol{\alpha}_1. \tag{4.1}$$

Hence the optimization problem is:

$$\max \boldsymbol{\alpha'}_1 \mathbf{\Sigma} \boldsymbol{\alpha}_1 \tag{4.2}$$

under a constraint arbitrarily chosen, $\boldsymbol{\alpha'}_1 \boldsymbol{\alpha}_1 = 1$. This procedure is iterative, in the sense that it consists of successive maximizations of the quantity $\boldsymbol{\alpha'}_j \mathbf{\Sigma} \boldsymbol{\alpha}_j$ imposing the linear independence condition $Cov[\boldsymbol{\alpha'}_{j-1} \boldsymbol{x}, \boldsymbol{\alpha'}_j \boldsymbol{x}]$, $j = 2, .., v$. It is easy to verify that the principal components are related with the algebraic concept of eigenvalue and eigenvector: this is a direct consequence of their definition. The derivation of the PCA can be found, among others, in Jolliffe (2002), Smith (2002), Kramer (2013) and James et al. (2013)) and implies some some mathematical hypotheses that here we attempt to explain. Let us consider the definition of eigenvalues and eigenvectors (Roman, 2005) :

Definition 2 (Eigenvalues and Eigenvectors)**.** *Let V be a linear space over a field $\mathbb{F}$, and let $\tau \in \ell(V)$. A scalar $\lambda \in \mathbb{F}$ is an eigenvalue for τ if there exists a non-zero vector $v \in V$:*

$$\tau v = \lambda v \tag{4.3}$$

In this case, v is called an eigenvector for τ.

Here $\ell(V)$ represents the set of all linear operators ψ on V, $\psi : V \longrightarrow V$. Consider the derivation of the first two principal components: as the procedure is iterative, it is immediate to extend the result to the j-th principal component, $j = 3, 4, ..., v$. Using the standard Lagrange multipliers method we write the function to be maximized as:

$$f(\boldsymbol{\alpha}_1, \lambda) = \boldsymbol{\alpha'}_1 \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha'}_1 \boldsymbol{\alpha}_1 - 1) \tag{4.4}$$

Thus, imposing the first order condition:

$$\frac{\partial f}{\partial \boldsymbol{\alpha}_1} = \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - \lambda \boldsymbol{\alpha}_1 = 0 \tag{4.5}$$

and rewriting the equation in the form corresponding to Definition 2,

$$\boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \lambda \boldsymbol{\alpha}_1 \tag{4.6}$$

it can be stated that λ is an eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_1$ is its eigenvector, but only if the conditions of Definition 2 hold: having the same form is not a sufficient condition. In fact, the concept of eigenvalue can be extended to matrices, and this is not surprising since the multiplication map $\boldsymbol{\Gamma} v$, where $\boldsymbol{\Gamma}$ is a matrix, is a linear transformation. We can wonder if $\lambda \in \mathbb{F}$ such that $\boldsymbol{\Gamma} v = \lambda v$, $v \in V$ exists. Indeed, if this number exists for a non-zero vector v, then it is defined as an eigenvalue for the matrix $\boldsymbol{\Gamma}$. Another useful concept is given by the following definition:

Definition 3 (Spectrum of a matrix)**.** *The spectrum $Spec(\Gamma)$ of Γ is the set of all eigenvalues of the matrix Γ.*

It can be proved that the eigenvalue λ is not unique in $Spec(\Sigma)$ when $v > 1$. On the contrary, from the derivation of the remaining $v - 1$ principal components, a number of $v - 1$ related eigenvalues λ_j can be computed, with a very important property: they are in (strictly) decreasing order, $\lambda_{j-1} > \lambda_j$. This can be derived from the previous optimization problem:

$$\max \boldsymbol{\alpha'}_1 \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 \Longrightarrow \tag{4.7}$$

$$\boldsymbol{\alpha'}_1 \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \boldsymbol{\alpha'}_1 \lambda_1 \boldsymbol{\alpha}_1 = \lambda_1 \boldsymbol{\alpha'}_1 \boldsymbol{\alpha}_1 = \lambda_1 \boldsymbol{\alpha'}_1 \boldsymbol{\alpha}_1 = \lambda_1 \cdot 1 \tag{4.8}$$

given the orthonormality for the vectors $\boldsymbol{\alpha'}_1$ and $\boldsymbol{\alpha}_1$. Thus, the quantity that maximizes the variance is λ_1, which is associated with the first PC (principal component).

To derive the second principal component we write:

$$\begin{cases} \max \boldsymbol{\alpha'}_2 \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha'}_2 \boldsymbol{\alpha}_2 = 1 \\ Cov[\boldsymbol{\alpha'}_1 \boldsymbol{x}, \boldsymbol{\alpha'}_2 \boldsymbol{x}] = 0 \end{cases} \tag{4.9}$$

or equivalently,

$$\begin{cases} \max \boldsymbol{\alpha'}_2 \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha'}_2 \boldsymbol{\alpha}_2 = 1 \\ \boldsymbol{\alpha'}_2 \boldsymbol{\alpha}_1 = 0 \end{cases} \tag{4.10}$$

Then, using again the Lagrange multiplier method,

$$f(\boldsymbol{\alpha}_2, \lambda, \eta) = \boldsymbol{\alpha'}_2 \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha'}_2 \boldsymbol{\alpha}_2 - 1) - \eta(\boldsymbol{\alpha'}_2 \boldsymbol{\alpha}_1) \tag{4.11}$$

$$\frac{\partial f}{\partial \boldsymbol{\alpha}_2} = \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda \boldsymbol{\alpha}_2 - \eta \boldsymbol{\alpha}_1 = 0 \tag{4.12}$$

and multiplying by $\boldsymbol{\alpha'}_1$ the second member gives

$$\boldsymbol{\alpha'}_1 \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda \boldsymbol{\alpha'}_1 \boldsymbol{\alpha}_2 - \eta \boldsymbol{\alpha'}_1 \boldsymbol{\alpha}_1 = 0 \Longrightarrow \tag{4.13}$$

$$\eta \cdot 1 = 0 \tag{4.14}$$

and

$$\boldsymbol{\Sigma} \boldsymbol{\alpha}_2 = \lambda \boldsymbol{\alpha}_2 \tag{4.15}$$

Again, the last equation shows that λ and $\boldsymbol{\alpha}_2$ are, respectively, the eigenvalue and the eigenvector of the matrix $\boldsymbol{\Sigma}$. As stated above, the same conclusions are reached proceeding in this way for the remaining variables. According to the usual notation, we call *Principal Components* (PCs) the linear combinations $\boldsymbol{\alpha'}_j \cdot \boldsymbol{x}$ for $j = 1, 2, .., v$ and *loadings* or simply coefficients the objects $\boldsymbol{\alpha'}_j$.

While principal components are usually derived analytically, an alternative geometrical interpretation exists. It can be shown (see for example Jolliffe, 2002, pp. 18-19) that principal components are interpretable geometrically as

the principal axes of the v-dimensional ellipsoids $E = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$. Indeed, E is by definition the generalized form for an ellipsoid, and it can be proved that for a given matrix $\boldsymbol{\Sigma}^{-1}$ the principal axes of the ellipsoid are its eigenvectors, which correspond to the PCs in this case.

4.2.2 The Singular Value Decomposition

Another very relevant algebraic result for the principal component analysis is the singular value decomposition (SVD). A formal discussion of the SVD can be found in Candilera and Bertapelle (2011); in this section we do not denote the vectors and the matrices in bold, since the context makes the notation evident. Let $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be a linear application, $K = \ker \phi$, $X = K^{\perp} \subset \mathbb{R}^n$ and $Y = im\phi$, with the usual notations for the kernel, the image and the orthogonal complement. Now define:

$$\phi|_X := X \longrightarrow Y \tag{4.16}$$

and

$$\psi(x) \cdot x' := \phi(x) \cdot \phi(x'), \forall x, x' \in X \tag{4.17}$$

which respectively represent an isomorphism, since $\mathbb{R}^n = X \oplus \ker \phi$, and a symmetric endomorphism $\psi : X \longrightarrow X$. If $u_1, u_2, ..., u_v$ is an orthonormal basis for X, then

$$\psi(x) = \sum_{i=1}^{v} \phi(x) \cdot \phi(u_i) u_i. \tag{4.18}$$

The spectral theorem guarantees the existence of an orthonormal basis for X which is made up of the eigenvectors of ψ. This theorem reminds one of the main features of PCA, the orthonormality of the principal component vectors justifying the connections of this mathematical technique with the PCA. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_v > 0$ be the ordered set which corresponds to the collection of the eigenvalues λ_i^2 associated with the i-th vector of the basis. The values λ_i, $i = 1, ..., v$, are called the singular values for the matrix $A \in M_{m \times n}(\mathbb{R})$. If the rank of the matrix A is r, and we define two matrices P and Q such that $P'P = I_r$, $Q'Q = I_r$ and S is the diagonal

matrix whose non-zero elements are λ_i, $i = 1, ..., r$ the equality

$$A = PSQ' \tag{4.19}$$

is called the Singular Value Decomposition. Let Σ be the Variance-Covariance matrix for the data collected in A, then

$$\Sigma \propto A'A = (PSQ')'PSQ' = QS^2Q' \tag{4.20}$$

so the PCs are given by

$$AQ = PS \tag{4.21}$$

The SVD result is largely adopted in the numerical algorithms for the PCA analysis, allowing a fast calculation of its values.

4.2.3 Sample properties of principal components and the choice between covariance and correlation matrix

In the previous sections we have discussed the population properties, because we have implicitly assumed the matrix $\mathbf{\Sigma^{-1}}$ to be known and $\boldsymbol{x}$, the vector of variables, to be a row vector with v columns, but the sample properties are needed to apply the principal component analysis to real data. In any case, these properties share lots of similarities with the population ones. Let $\boldsymbol{x_1}$,$\boldsymbol{x_2}$,..., $\boldsymbol{x_n}$ be the vectors for the n observations; the already defined matrix $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times v}$ is the matrix whose i-th row is $\boldsymbol{x_i}$. In this case (see, for example Jolliffe, 2002) we can define $\tilde{z} = \boldsymbol{\alpha'}_j \boldsymbol{x}_i$ as *the score* for the i-th observation on the j-th PC. Then the *sample* variance to maximize, in order to derive the first PC, is

$$\frac{1}{n-1} \sum_{i=1}^{n} (\tilde{z}_{i1} - \bar{z}_1)^2 \tag{4.22}$$

under the same constraint as before. The derivation that follows is conceptually the same as the population case.

A specific practical precaution which is usually applied in the empirical analyses is the standardization of variables. This is a very important feature to take advantage of. For example, it allows us to consider the basic form of the singular

value decomposition, but more than anything it prevents a wrong analysis due to a possibly incomparable data size. As a matter of fact, the standardization

$$\tilde{x}_j = \frac{x_j - \bar{x}_j}{\sigma_{jj}}, j = 1, 2, ..., v \tag{4.23}$$

provides the possibility to compute the PCs using the *Correlation matrix* instead of the Variance-Covariance matrix. In Von Storch and Zwiers (1999), Section 13.1.10, it is proved that the principal components (also called Empirical Orthogonal Functions) are invariant under an orthogonal transformation:

$$\boldsymbol{Z} = \boldsymbol{\Gamma X} : \boldsymbol{X} = \boldsymbol{\Gamma}^{-1}\boldsymbol{Z} \tag{4.24}$$

$$\boldsymbol{\Sigma}_{ZZ} = \boldsymbol{\Gamma \Sigma}_{XX}\boldsymbol{\Gamma}' \tag{4.25}$$

$$\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}^{-1} \tag{4.26}$$

but not under every transformation. In general, the principal components do not correspond to those that are found using the Variance-Covariance matrix $\boldsymbol{\Sigma}$ after an appropriate reverse transformation. Nevertheless, frequently different units of measurements are used in practice to measure different variables and some variables may be of a different order of magnitude than others. In all these cases, using the correlation matrix is more appropriate. The standardization of variables can provide a standardized method to follow before applying PCA, preventing some principal components from being overrated, and in general avoiding all the errors due to incompatible measures among the variables. Moreover, as it was independently demonstrated by Hotelling (1933) and Meredith and Millsap (1985), the use of correlation matrix can be seen as the appropriate method to maximize the quantity:

$$\sum_{j=1}^{v}\sum_{k=1}^{q} r_{jk}^2 \tag{4.27}$$

where $1 \leq q \leq v$ is the number of the elements of y resulting from the orthonormal (linear) transformation $\boldsymbol{y} = \boldsymbol{B'x}$, $\boldsymbol{B'} \in \mathbb{R}^{v \times q}$, and r_{jk}^2 representing the squared correlation between the j-th variable and the k-th principal component. This particular criterion can be thought as a way to characterize an "optimal" subspace

of q-dimensions, since $1 \leq q \leq v$, and it is maximized when the first q eigenvalues of the *Correlation* matrix are computed.

Another interesting property can be pointed out considering again the transformation(s) $\boldsymbol{y_i} = \boldsymbol{B'x_i}$, where now the index $i > 1$ highlights the presence of two or more random vectors $\boldsymbol{x}$. One way to choose the matrix $\boldsymbol{B}$ is to solve the problem

$$\max \sum_{h=1}^{n} \sum_{k=1}^{n} (\boldsymbol{y}_h - \boldsymbol{y}_k)'(\boldsymbol{y}_h - \boldsymbol{y}_k). \tag{4.28}$$

Associating the index $i > 1$ with the number of observations n, and q with dimension of the subspace where they are projected, it can be proved that $\boldsymbol{B}$ must be equal to *the first* q columns of the loading coefficient matrix. The last two theorems presented above suggest the idea of retaining a fewer number of principal components. In fact, if the first few PCs explain an (arbitrary) substantial part of the data variance, these theorems strengthen the idea to retain only them. This is one of the main advantages of the PCA method, together with a possible improved interpretability with respect to the original variables.

These properties suggest a link between the principal component analysis and regression models. Indeed, we can wonder if the principal component vectors play some role in the regression analysis: using the standard notation for the regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, the next theorem paves the way to discuss this topic in greater details.

Theorem 2. *Let* $\boldsymbol{\Gamma}$ *be a* $v \times v$ *orthogonal matrix and* $\boldsymbol{\varphi} = \boldsymbol{\Gamma^{-1}\beta}$*. If* $\hat{\boldsymbol{\varphi}}$ *is the* OLS *estimator of* $\boldsymbol{\varphi}$*, the elements of* $\hat{\boldsymbol{\varphi}}$ *have, successively, the smallest possible variance if* $\boldsymbol{\Gamma} = (\alpha)_{ij}$*. Then the elements of* $\boldsymbol{Z} = \boldsymbol{X\Gamma}$ *are the sample PCs for* $\boldsymbol{x}$*.*

Proof. If $\hat{\boldsymbol{\varphi}}$ is an OLS estimator, then $\boldsymbol{\sigma}_{\hat{\boldsymbol{\varphi}}} \propto (\boldsymbol{Z'Z})^{-1}$, hence $\boldsymbol{\sigma}_{\hat{\boldsymbol{\varphi}}} \propto \boldsymbol{B'(X'X)^{-1}B}$. Considering the trace $tr(\boldsymbol{B_j'(X'X)^{-1}B_j})$, which we want to minimize, it can be shown that $\boldsymbol{B_j}$ must consist of the last j columns associated with the last j eigenvectors of $\boldsymbol{(X'X)^{-1}}$, which are the first j eigenvectors of $\boldsymbol{(X'X)}$ since they are reciprocal. □

These considerations suggest we should verify whether principal component

analysis may be usefully employed in the linear regression. Indeed, even if the previous theorem does not suggest any link between PCA and the dependent variable (please note that the y variable is not involved in the reasoning, but only in the OLS estimate) it suggests a possible usage of the PCA as a dimensional reduction device, along with conceivable improved interpretational and estimation properties, which now we are going to discuss.

4.3 Principal Component Regression

The main idea of Principal Component Regression (PCR) is to take advantage of the results given by the PCA, a traditional unsupervised learning method, into the standard regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, which is a supervised method.

Assume the usual hypotheses for the linear model and its error terms; see for example Wooldridge (2015), Chapter 2. As we do in the empirical analysis, here we assume that each variable is measured about its mean and that the independent variable matrix is standardized, which is convenient in the case of a different unit of measurement among variables as discussed above.

Considering the matrix of scores $\boldsymbol{Z}$ whose elements are associated with each observation, that is, using the previous notations, $\tilde{z} = \boldsymbol{\alpha'}_j \boldsymbol{x}_i$, $j = 1, 2, ..., v$ or $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A}$. If we write the linear model using $\boldsymbol{Z}$ as the matrix of dependent variables, the following relations hold:

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\varphi} + \varepsilon \tag{4.29}$$

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{I}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}'\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{\varphi} \tag{4.30}$$

where

$$\boldsymbol{\varphi} := \boldsymbol{A}'\boldsymbol{\beta} \tag{4.31}$$

and since the matrix of coefficients **A** is orthogonal by derivation.

The OLS estimator $\hat{\boldsymbol{\beta}}$ can be derived, after having computed the OLS estimator $\hat{\boldsymbol{\varphi}}$, through the inverse relation $\hat{\boldsymbol{\beta}} = \boldsymbol{A}\hat{\boldsymbol{\varphi}}$. Usually only a subset of vectors of the score matrix is retained, such that

$$\boldsymbol{y} = \boldsymbol{Z}_{\boldsymbol{S}}\boldsymbol{\varphi}_{\boldsymbol{S}} + \tilde{\varepsilon} \tag{4.32}$$

for $Z_S \subset Z$.

The main advantages of this approach are at least four.

First of all, the possibility of taking advantage of the interpretational properties of the PCs, which are computed separately and before the regression. However, in reconstructing the estimated values for β, the PCR offers an estimate for the original independent variables.

Secondly, it may be interesting to study our data using an approach different from the previous empirical studies, and which usually has been found helpful in dealing with large dataset.

Thirdly, and really worth noticing, combining PCA analysis with the linear regression is a particular aspect of a more general theory developed on the class of biased estimators, as will be clearer further on in this and the following chapters. This is extremely important considering the resulting possibility to compare the estimates given by different methods, both to assess if there is an unambiguous relationship independent of the method chosen, and to choose the more convenient model in terms of a common defined statistical performance. In fact, the possible reduction of the Mean Squared Error (MSE) — as discussed in the section on the Cross-Validation technique — can be used to choose the most suitable regression model to apply to our dataset.

Furthermore, the PCA analysis can be extended to examine the general problem of missing data, which is a feature of our loan data, and allows us to propose a regression model based on the results of the Probabilistic PCA (see Chapter 6).

We cannot omit another advantage of this technique, that is related to the multicollinearity problem. This problem is quite frequent in practice, especially when a large number of regressors is considered, and occurs when a specific variable can be written as a linear combination of another variable included in the estimation too. More formally, and without loss of generality, this is the case where $x_1 = \alpha + \beta x_2$ for some appropriate constant vectors α and β. This has been studied in econometrics for a long time, since in this case the traditional OLS estimators for the linear regression model are subjected to a different range of

problems. Choosing as the independent variables a subset of the scores derived from PCA can considerably reduce the variance of these estimators. The reason, which will be analytically evident in the next section, is that if the first $k \leq v$ columns of the score matrix are chosen, then the variance of the estimator for $\boldsymbol{\beta}$, $\mathrm{Var}\,\hat{\boldsymbol{\beta}}$, can be reduced.

In the next section, we are going to discuss in depth some of its properties, its interpretation, some problems related with both variable selection and score selection, and generally every feature necessary in order to make this technique effective in our successive empirical analysis.

4.4 Variable selection and interpretation in PCA and PCR

4.4.1 Interpretation and selection of the optimal principal component subset

Considering the principal component analysis as an unsupervised method, the interpretation of the principal components and the choice of the optimal subset are two of the most relevant aspects in practice. A discussion and some empirical applications of these subjects are presented, among others, in Rao (1964), Blackith and Reyment (1971), Diamantaras and Kung (1996) and Jolliffe (2002).

The interpretation of the principal components, $\boldsymbol{\alpha'}_j\boldsymbol{x}$, concerns its elements and first and foremost their relative magnitude and sign. To understand better their meaning, we can represent a j-th principal component as a column vector of v rows:

$$\begin{pmatrix} z_{j1} \\ z_{j2} \\ .. \\ z_{jv} \end{pmatrix}$$

Firstly, it is important to compare each element of this vector on a relative rather than an absolute scale, not depending on the measurement unit. Hence, the first step consists of standardizing each element, according to its relative magnitude.

A second relevant step is to consider the sign of their absolute relative values. In fact, to analyze the coefficients, it can be useful to represent them as a vector whose elements are the $+$ or $-$ sign, if their absolute relative value is greater than an arbitrary chosen level, and with another symbol (let $*$) if they are not significant (in this context without any reference to the equivalent statistical term), operating the following transformation:

$$\begin{pmatrix} z_{j1} \\ z_{j2} \\ .. \\ z_{jv} \end{pmatrix} \Longrightarrow \begin{pmatrix} +/-/* \\ +/-/* \\ .. \\ +/-/* \end{pmatrix}$$

The rounding of optimal estimates implied in this procedure, as shown in Bibby (1980), has more advantages than negative aspects. Indeed, when done carefully, this rounding reduces the original variance explained by a PC only of a negligible percentage: in this paper the average loss of the explained variance, after a rounding of one decimal, was $\frac{\lambda_1}{300}$ for the first principal component in $\mathbb{R}^4$. Therefore, the loss in accuracy is overcompensated by the gain in interpretation.

In any case commenting only one sign, considered separately, would be wrong: what the contrast of these signs suggest is of importance. Obviously, if only the contrast counts, then the signs of each PC are arbitrary. Therefore, the examination of this contrast, made for each PCs one at a time, could suggest the phenomenon responsible for the original data variation (percent) associated with the considered principal component. The interpretation should be done with regard to the underlying nature and characteristics of the real case examined. For example, in our successive empirical study on loan data we analyze whether a clear interpretation of the first two PCs, obtained from the available independent variables, is possible, that is if an economic interpretation is feasible.

Since the derivation of PCs is essentially algebraic, this cannot be guaranteed. However, this is not the only added value of this technique. In fact, another feature of PCA is the dimensionality reduction of the original space allowed by selecting an optimal subspace among the v principal components. Specifically, it

can be shown (see below in this section) that choosing the first $k \leq v$ principal components retains, by construction, a subspace with the maximum possible amount of variance of the original space.

This can be formalized using the previous notation: the choice of a subset $\boldsymbol{Z_S} \subset \boldsymbol{Z}$ is made such that $\boldsymbol{Z_S} = \boldsymbol{Z_1} \cup \boldsymbol{Z_2} \cup ... \cup \boldsymbol{Z_k}$. Moreover, we have already discussed the inverse relation between the arrangement of the column vectors belonging to the score matrix and the percentage of the original variance they explain.

It is interesting to make some theoretical consideration on this result, which is one of the most important concepts of PCA and one of the most discussed topics of PCR (see 4.4.2). From (4.2.1) we know that the equation

$$\boldsymbol{\Sigma\alpha}_j = \lambda\boldsymbol{\alpha}_j, j = 1, 2, ..., v \tag{4.33}$$

holds, as well as the orthogonal constraint $\boldsymbol{\alpha'}_j\boldsymbol{\alpha}_j = \|\alpha_j\|^2 = 1$. Now let us define the following:

Definition 4 (Hermitian matrix)**.** *A matrix $\boldsymbol{A} = \boldsymbol{A}^\dagger$, where $\boldsymbol{A}^\dagger$ is the conjugate transpose of $\boldsymbol{A}$, is an Hermitian matrix.*

Suppose $\boldsymbol{A}$ is the Variance-Covariance matrix: indeed, a symmetric matrix $\boldsymbol{A}$ which takes values in the field $\mathbb{F} = \mathbb{R} \subseteq \mathbb{C}$ is a special case of an Hermitian matrix.

Definition 5 (Rayleigh-Ritz ratio)**.** *The ratio $R(\boldsymbol{A}, \boldsymbol{x}) := \frac{\boldsymbol{x'Ax}}{\boldsymbol{x'x}}$ is called Rayleigh-Ritz ratio.*

The Rayleigh-Ritz theorem, discussed among others in Horn and Johnson (2012), provides a link between the non-decreasing eigenvalues $\lambda_1 \leq \lambda_2 \leq .... \leq \lambda_j$ of the Hermitian matrix and the maximization problem with orthonormal constraint, as takes place in PCA. The reverse order of the indexes for λ here are only due to opposite conventions: using the standard notation for PCA, we write $\boldsymbol{\Sigma\alpha}_j = \lambda\boldsymbol{\alpha}_j$ and derive

$$R(\boldsymbol{\Sigma}, \boldsymbol{\alpha}_j) = \frac{\boldsymbol{\alpha'}_j\lambda\boldsymbol{\alpha}_j}{\boldsymbol{\alpha'}_j\boldsymbol{\alpha}_j} = \lambda_j. \tag{4.34}$$

The variance explained by the j-th PCs, $\operatorname{Var} PC_j$, can be computed through its associated eigenvalue. Indeed, from the algebraic construction presented above, it

follows immediately that:

$$\operatorname{Var} PC_j = \frac{\lambda_j}{\sum_{j=1}^{v} \lambda_j}. \tag{4.35}$$

Thus retaining only the first $k \leq v$ PCs lead to keeping an amount of the original variance of $\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{v} \lambda_j} < \frac{\sum_{j=1}^{v} \lambda_j}{\sum_{j=1}^{v} \lambda_j} = 1$.

The main point here is: if k is small, maybe equal to 2 or 3, and at the same time the variation explained by these first k PCs is high, and the cumulative variation is also high, the reduction allowed by the PCA is effective since it retains a large amount of the original variation. The rule to choose the optimal level to be retained is not unequivocal. Some authors have discussed various procedures to choose the optimal level, as Mandel (1972), Eastment and Krzanowski (1982) and Sugiyama and Tong (1976). In these studies a discussion on the optimal dimensional reduction of the original space is presented, under specific hypotheses and simulation studies. On this point a wide discussion has been made in Al-Kandari and Jolliffe (2001), where different variable selection criteria are compared. However, we are not going to analyze these and other related studies, since our focus is on the variable selection in the PCR and in practice usually the decision is made by rule of thumb: a level between 0.80 and 0.90 is considered satisfactory. Moreover, retaining a high percentage of the variance may not be the only purpose of the researcher. In fact, as discussed among others by Ferré (1995), the goal is not always to optimally fit the original dataset. In the case of PCR, where a dependent variable is present and then the predictive aspect is predominant, a deeper and probably more essential discussion is imperative.

4.4.2 Choosing the optimal principal component subset in PCR. Solutions and variations of PCR

In the principal component regression the score vectors are substituted to the original independent variables as predictors. In doing so, it is common to retain as regressors only the row vectors associated with the first principal components derived from the previous PCA, and this is in fact one of the most straightforward usage of PCA in the regression analysis, which can improve both the interpretation

and the estimates given by the *OLS* estimator. For example, in his examination on the variable selection in linear regression, Hocking (1976) explains the principal component regression as a technique which deletes the last PCs. This is the approach supported by Mansfield et al. (1977), Mosteller and Tukey (1977) and Gunst and Mason (1980) but criticized by Jolliffe (1982).

Among all the possible methods, a very simple one consists of fixing a cut-off level h^* for the respective eigenvalue, such that the PC is included in the analysis if its eigenvalue $h_j > h^*$. The cut-off level must be chosen considering whether a previous standardization has been made: if this holds, a level between 0.01 and 0.1 may be appropriate.

In truth, even though the first principal components retain the higher percentage of the variance among the independent variables, there is not any certainty that their correlation with the dependent variable will be maximum. A possible conflicting choice can arise when a principal component is associated with a relative low eigenvalue but has a high predictive importance for the y variable. For example, in the analysis of Smith and Campbell (1980) among the most significant regressors in PCR for the response variable there are some PCs associated with a low eigenvalue. Their analysis considered chemical data, but in Hill et al. (1977) the same happens for economic variables, such as the Gross National Product and the Unemployment rate. Their results suggest that all of the six PCs (they used six original variables) should be included, even if the last explain a small amount of the variance. Empirical opposite conclusions are given in Berk (1984) with six different datasets. His result, even though it points out that this problem does not always take place, cannot be considered as the only possible case.

Ideally, the optimal subset for the Principal Component Regression has the form $\boldsymbol{Z_S} \subset \boldsymbol{Z}$, where $\boldsymbol{Z_S} = \bigcup_{j=1}^{k} \boldsymbol{Z_i}$ is only a special case. It is evident that there is practically not an unequivocal solution to this problem, and for this reason various ideas have been proposed in literature.

On the one hand, as for the unsupervised PCA analysis, the selection can be

made through various rules of thumb. As a matter of fact, every choice made for an optimal level of some determined parameter can be useful in practice. A simple but efficient rule could consist of including both the first PCs and the other potential PCs which present a low-variance but a significant link with the dependent variable. One of the weaknesses of these techniques is the inevitable presence of an arbitrary choice.

On the other hand, the usage of the PCA combined with a linear regression model suggests the possibility of applying the same test of significance and the statistics which are usually employed for the standard linear regression. One possible approach involves a decision of an optimal level for the variance inflation factor (VIF), which is usually employed to detect the multicollinearity when lots of variables are present. A less arbitrary choice is to use the T-test applied to the principal components, as discussed in Foucart (2000). This test measures the independent contribution of each score vector to the PCR. Even if the T-test, as shown by Mason and Gunst (1985), is not perfectly symmetric for low and high-variance components, it can constitute a valid solution even if too many variables could be retained (Jolliffe, 2002).

Another common coefficient used in the linear regression, the $\bar{R}^2$, is taken into consideration by Lott (1973) for the problem we are discussing. He suggests choosing the subset S which satisfies the condition

$$\max_S \bar{R}^2_S = \max_S (1 - (1 - R^2_S)\frac{n-1}{n-v-1}). \tag{4.36}$$

With a completely different approach, other authors have suggested some variations of the classical PCR itself, since this selection problem seems to be an intrinsic characteristic of the principal component regression.

Literature on this topic is varied. Hawkins (1973) and Webster et al. (1974) independently proposed including the dependent variable in the principal component analysis, a method called latent root regression. A variation of the latent root regression which employs the Cholesky factorization is presented in Hawkins and Eplett (1982). Oman (1991) discuss a shrinkage of the least square estimators on the first principal components. Another possible approach is the

partial least squares regression. In our empirical study we focus on the previous statistical methods, rather than on the variations of the PCR. Since the number of principal components is four, sometimes these methods, in accordance with the estimated MSE, supply the conclusion of retaining all the variables, and in other cases to retain their own subset.

4.4.3 Overfitting. The Cross-Validation technique and its applications to PCR

In our empirical analysis we largely use the so-called Cross-Validation (CV) technique, a tool which compared with the previous solutions proposed for the optimal subset selection could lead to a less-arbitrary choice, or at least which presents some advantages. The Cross-Validation is largely employed in practice, despite its conceptual simplicity, even to test the performance of increasingly complex models (James et al., 2013), and is adopted in this work for the validation of other regression techniques too. It is worth pointing out that this technique can be advantageously employed thanks to the sufficient contemporary computing power, which has remarkably increased compared with the past.

The CV belongs to the resampling methods, whose name derives from the repeated extraction of a sample set that these methods perform. The theoretical basis for the use of the CV method in the principal component regression can be found in Hill et al. (1977). In their discussion of an optimal subset selection for the PCR, the authors suggest considering the mean squared error (MSE), which for a general predictor of y is defined as

$$\rho = MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{4.37}$$

This definition of MSE corresponds to those used for the "strong criterion" proposed by the authors, which is to prefer the subset a to the subset b if

$$MSE^a < MSE^b \tag{4.38}$$

This quantity must be estimated as in the case of the MSE for an estimator, since in this case we are interested in $\hat{\rho}$ for the population and not only for the

sample which the values $\hat{y}_i$ were computed on: the CV technique allows this. These considerations find support in literature: Krzanowski (1987), Mertens et al. (1995) and Diana and Tommasi (2002) discuss the application of the CV to the selection of the PCs, because it measures the quality of the fit for different choices of regressors in the PCR.

To briefly discuss the Cross-Validation technique, we consider some preliminary concepts. For a dataset of n observations of the dependent and independent variables:

$$O = \{(y_1, \boldsymbol{x_1}), (y_2, \boldsymbol{x_2}), ..., (y_n, \boldsymbol{x_n})\}. \tag{4.39}$$

the training set $T \subseteq O$ of the data is defined as the observations which are available to *train*, that is to fit, the model. On the contrary, the test set $K \subseteq O : K \cap T = \emptyset$ contains those observations which have been excluded by the training set, or however those which are unavailable to fit the data. In fact, this last subset is very useful to compare the predictive values, computed through the training set, with a different dataset. For example, loosely speaking, the MSE computed on the training set has less importance than the test MSE, $\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2$, $i \in K$, since we are interested in a model which prevents large errors when the new elements of K are added and the predictions $\hat{y}_i$ have been made using the trained model. This concept is related to the statistical problem of overfitting. As the word says, when too many parameters are included in a model, it $over$-fits (and then does not simply fit) the training data set because it adapts very precisely (at limit perfectly) to the available observations. But the problem is that a very bad fit occurs when a new observation is included, since its flexibility is very poor. In other terms, a compromise between the flexibility of a model and an optimal fit for the training data should be reached; otherwise the predictive power of the model would be insufficient. The *Bias-Variance trade-off*, a well-known concept in statistics, must be considered to fit a model.

Formalizing, the expected MSE for the test set can be written as

$$EMSE(test) = E(y_i - \hat{y}_i)^2 = \operatorname{Var}\hat{y}_i + Bias^2(\hat{y}_i) + \operatorname{Var}\varepsilon \tag{4.40}$$

$i \in K$, ε the error term. For a fixed $\operatorname{Var}\varepsilon$, the previous discussion suggests finding the best balance between the $\operatorname{Var}\hat{y}_i$ and the $Bias(\hat{y}_i)$, such that $EMSE(test)$ is minimized. These considerations hold for the MSE of an estimator as well: in the PCR regression made on a subset of scores $\boldsymbol{Z}$, contrary to the ordinary linear regression model, the $Bias(\tilde{\boldsymbol{\beta}}) \neq 0$. Using the previous notation,

$$\boldsymbol{y} = \boldsymbol{Z\varphi} + \boldsymbol{\varepsilon} \Longrightarrow \hat{\boldsymbol{\varphi}} = (\boldsymbol{ZZ'})^{-1}\boldsymbol{Z'y} = \boldsymbol{\Lambda^{-2}Z'y} \tag{4.41}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix whose elements are $\lambda_{kk}^{1/2}$, the k-th largest eigenvalues of the matrix $\boldsymbol{X'X}$. Then,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\hat{\boldsymbol{\varphi}} = \boldsymbol{A\Lambda^{-2}A'X'y} = \sum_{i=1}^{v} \lambda_{ii}^{-1}\boldsymbol{a_k a_k' X'y}. \tag{4.42}$$

Since for the OLS estimator $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, if in the PCR the set of regressors is restricted to a subset of scores $1, 2, ..., s-1$, with a related estimator $\tilde{\boldsymbol{\beta}}$ the $E(\tilde{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$ of a quantity $\sum_{i=s}^{v} \boldsymbol{a_k a_k'\beta}$. On the other hand, it can be shown using the SVD that

$$\operatorname{Var}\tilde{\boldsymbol{\beta}} = \sigma^2 \sum_{i=1}^{s-1} \lambda_{ii}^{-1}\boldsymbol{a_k a_k'}. \tag{4.43}$$

Since among the retained scores there are usually those associated with larger eigenvalues, the reciprocal factors λ_{ii}^{-1} associated with the excluded variables are smaller. Therefore (4.43) is smaller, thus the variance is lower than that of the traditional linear regression estimators. To compute the overall effect of these two opposite effects, the Cross-Validation algorithm randomly split the available data set in two subsets. The training set is used to train the model; the validation set is used to evaluate the statistical performance. When the reduction of the variance overcompensates the greater bias the MSE is lower, as we illustrate in our empirical analysis. A substantial problem could arise if, choosing the splitting rule, the training set is excessively restricted, as fewer observations always imply a worse statistical performance. That is the reason why in practice usually a k-fold Cross Validation is employed, which means that the observations are randomly split into k subsets (groups or folds) of comparable cardinality, of which the last $k-1$ are used to train the model and the first is the validation set; this procedure

is repeated k times (algorithm 1). When $n = k$ the procedure is called LOOCV; it retains the largest training set possible, but it is computationally expensive and its estimates have a lower bias but a higher variance compared with the 10-fold CV. In our empirical analysis, as it is frequent in practice, we use a 10-fold CV.

Algorithm 1 CV algorithm

procedure K-FOLD CROSS VALIDATION
    Split the data in k subsets of comparable size
    for i=1:k **do**
        Train the model on {1,2,...,i-1, i+1,...,k} subsets
        Compute the MSE_i on the i-th validation subset
    end for
    $MSE_{cv} = \frac{1}{k} \sum_{j=1}^{k} MSE_j$
end procedure

4.4.4 Selecting variables in PCR using an iterative F-test

Another aspect of the principal component regression, which is mostly unheeded in practice, is the variable selection among the original predictors. In fact, we have discussed the selection for the principal components but not for the independent variables which the principal component analysis is applied to. This subject is characteristic when dealing with the standard linear regression, but most studies tend to retain all the original variables for the principal components included in the regression. As a matter of fact (depending on its definition) the regression applied to the PCA can be seen as a method which however retains all the variables and which concerns only the score vectors.

Nevertheless, some authors have explored this topic to examine whether a variable selection in the PCR can improve results. In the empirical part we adopt an F-test, as proposed by Mansfield et al. (1977), to show the significance of our predictors.

Considering a principal component regression model where only a subset

$1, 2, ..., s$ of scores is retained, then

$$\hat{\boldsymbol{Y}} = \bar{\boldsymbol{Y}} + \sum_{j=1}^{s} \boldsymbol{u_j}\hat{\boldsymbol{\varphi}}_{\boldsymbol{j}} \tag{4.44}$$

with $\boldsymbol{u} = \boldsymbol{z'A}, \boldsymbol{z} \in \boldsymbol{Z}$. The Sum of Squared Residuals (RSS) becomes:

$$RSS = \sum_{i=1}^{n} (\boldsymbol{Y_i} - \hat{\boldsymbol{Y}}_{\boldsymbol{i}})^2. \tag{4.45}$$

The test statistic used to compare two different models is analogous to the F-statistics for a linear regression model. Firstly, the increase in the residual sum of squares u_1 is computed deleting from the original variable set the j-th variable $\boldsymbol{x_j}$ one at a time. The minimum of the set $\{u_j\}$, $j = 1, ..., v$ is:

$$u_i^k = \min u_j. \tag{4.46}$$

The i-th variable is deleted if, considering the F distribution with $(1, l)$ degrees of freedom

$$\frac{\min u_i^k}{MSE} \leq F(m, l). \tag{4.47}$$

The procedure is iterative: if it exists, in the second step $k + 1$ the minimum of the increased RSS value is computed, and the test to decide if the variable is statistically irrelevant becomes

$$\frac{(\min u_i^{k+1} - \min u_i^k)}{MSE} \leq F(m, l). \tag{4.48}$$

However, since in our analysis the number of available complete variables is not large and the economic interpretation has to be considered too, we apply this algorithm only to give an example of the significance of our variables, stopping at the first step. The choice is made one at a time for each variable with thousands of observations, hence for $F(m, l)$ it can be assumed $m = 1$ and $l \longrightarrow \infty$.

4.5 Empirical results on loan data

In this section we apply the principal component regression to loan data, along with the variable selection techniques, the statistical tests and the other aspects which have been discussed so far in this chapter. Our aim is to verify what these

results suggest, from an economic point of view, and specifically what is the effect of collateral and of other borrower variables on the residential mortgage-backed security (RMBS) interest rate. The key points of our analysis are:

1. It is a cross-sectional study on different large datasets collected on the European DataWarehouse (ED).

2. A preliminary split of this data is done considering the nature of the interest rate, namely fixed or floating. Indeed, the fixed interest rate margin, being established in a specific date, is not precisely comparable taking an extended period of time. Thus, we consider separately two shorter periods approximately homogeneous for the macroeconomic variables, in particular for the Euribor interest rate: from 2004 to 2006 (before the crisis, higher margins) and from 2011 to 2016 (post-crisis, lower margins).

3. The dependent variable, collected in the column vector y, is *the Interest Rate Margin* of the loan contracts associated with a RMBS for each borrower.

4. The explanatory variables, collected in the matrix $\boldsymbol{X}$, are *the Primary Income* of the borrower, *the Loan Term*, *the Original Balance* of the loan and the *Loan to Value ratio*. The variable *Debt to Income* was excluded due to issues of multicollinearity.

5. Our analysis is performed separately on the main European Union countries for which the data are available: Belgium, UK, Spain, France, Italy, Germany, Ireland and Netherlands. This allows us to analyze the possible differences among these countries.

Therefore, for each country, an original matrix $\boldsymbol{A}$ of n rows (the number of observations, ranging between 133153 to 1264432) and $v = 7$ columns is available. In this analysis, contrary to the case examined in the last chapter, all rows with at least one Nan (not available) data are deleted. Our goal is to examine the empirical results on the relation between $\boldsymbol{X}$= [*Primary Income, Original Balance, Loan Term, Loan to Value*] and $\boldsymbol{y}$= [*Interest Rate Margin*] as given by the PCR.

The dimensionality of the matrices analyzed is resumed in table 4.1. While the number of observations n is variable, the number of columns is always 7: two of them contain the loan origination date and the interest rate type, four contain the independent variables and one the response variable.

Table 4.1: The dimensionality of data analyzed

Country	$n =$ Number of rows (observations)
Belgium	$133,153$
UK	$370,146$
Spain	$1,264,432$
France	$994,601$
Germany	$489,402$
Ireland	$242,331$
Netherlands	$775,920$
Italy	$476,241$

It is important to highlight that the primary goal of this analysis is not to propose a *predictive model* for the interest rate. Indeed, other borrower's variables foreseen in the ED taxonomy are currently unavailable and moreover the interest rate can depend on further variables, such as macro-economics variables, the economic cycle expectations, the risk-adversity of the lender and, as usually happens in social sciences analyses, the imponderable individual decision made for each client. Our goal, however, is *to examine the overall effect* of the variables collected for each borrower, as they are available in the ED, which are very specific and relevant for the interest rate decision. As a matter of fact, both the economic intuition and the statistical significance in our results guarantee their influence on the interest rate decision. From a statistical viewpoint, these statements are equivalent to state that the significance of our predictors are of more interest than the total variance of the dependent variable explained by the model. The MSE here is used as a number which can be advantageously employed to compare the accuracy of the different methods applied.

On the other hand, the high significance level of our regressors (the score variables), combined with a lower variance for the estimators induced by the PCR when a subset of the PCs is chosen, and above all the high number of observations keep up the accuracy of our estimates. The unsupervised principal component analysis made before the regression is of interest too.

Having taken this into consideration, our examination is made on the available micro-data as best we can in order to consider the individual characteristic of a huge number of contracts, in contrast with the previous empirical analyses made on this subject; we do not exclude that in the future, as data availability may improve, these results may be reconsidered and extended in order to include new predictors (see also Chapter 6). Due to this consideration, in order to strengthen the results found by these estimates, we flank an unsupervised learning method, known as *K-means clustering*, to the PCR analysis. This method is by construction independent of the other possible missed predictors. The main idea behind this clustering method is to partition the dataset into K disjoint sets, which contain similar data. The similarity is usually evaluated as a specific distance between the observations and a centroid. Intuitively, the algorithm based on the work by Arthur and Vassilvitskii (2007) starts from a randomly chosen observation considered as the first of k centroids. Then it assigns each observation to its closest centroid, chosen again randomly with an assigned probability and depending on a specified distance. Each distance is computed as $d_i = d(x_i, c_i)$, where x_i are the points which belong to the set with centroid c_i. The process is then repeated n times, and finally converges to the lowest possible sum of distances d_i. This leads to a classification of points based on their distances from their closest centroid.

In our empirical analysis, we show in an example that the clustering is in accordance with our estimates. In particular we partition the observations in two clusters, considering the *Loan to Value* and the *Interest Rate Margin* variables: we find that the positive coefficient estimated in the PCR for the *Loan to Value* variable corresponds to the cluster whose centroid is associated with higher interest rates and higher loan to value. As discussed in the fifth chapter, the results

of the present section are coherent with those computed with other regression techniques.

Now we comment the main results given by our analysis, which is conducted separately on the available data for the fixed and for the floating interest rates. In this section we present an overview of these results, along with some representative tables. The complete analysis for each country can be found in Appendix A.

At the first step, a correlation analysis on the matrix $\boldsymbol{X}$ is done for each country: $\hat{\Sigma}_J$, where J is the country's name, represents the matrix of correlations for the independent variables. There is a certain similarity among these matrices. In nearly all the cases for the floating interest rates, the correlation between the *Primary Income* and the *Loan Term* is negative, as is also often that between the *Primary Income* and the *Loan to Value*, while the other correlations are positive. These estimates suggest that in general the borrowers who require a longer loan and who pledge less collateral have a lower primary income. This fact is completely coherent with common sense, because a borrower who has a low income is more likely to require a loan with a high duration and he probably cannot pledge much collateral. For the fixed interest rates these conclusions do not hold, since the correlation matrix has always positive entries (except for the Netherlands and Germany at one of its entries). This fact points out the difference between the fixed and the floating interest rate contracts: a fixed interest rate seems to be required from borrowers with higher *Primary Income*, *Loan to Value* and *Loan Term* together. There are not too high correlations, and then no evidence of multicollinearity: notice that the *Loan to Value* variable depends on the collateral pledged (the denominator) and then is not predictable knowing the *Original Balance* (the numerator). However, to assess the multicollinearity among the independent variables we employed the Belsley collinearity diagnostics. Some of these tests are presented in the Appendix (see for example table A.1.1). The condition indices are all low, and this strongly supports the absence of multicollinearity (Belsley et al., 2005). The principal component analysis allowed us to detect and delete in advance the main outliers (see figure 2), even if the results did not change due to the high number of observations considered.

$$\hat{\Sigma}_{Belgium} = \begin{bmatrix} 1.0000 & & & \\ 0.21860 & 1.0000 & & \\ -0.10556 & 0.29683 & 1.0000 & \\ -0.10189 & 0.14899 & 0.21622 & 1.0000 \end{bmatrix}$$

Table 4.2: The correlation matrix for Belgian data. The negative correlation between the *Primary Income* and the variables *Loan Term/Loan to Value* is a recurring pattern in our analysis.

Singular values	Condition indices
1^{st}	1.0000
2^{nd}	1.2070
3^{rd}	1.4272
4^{th}	1.7504

Table 4.3: Belsley collinearity diagnostics on the UK data. There is no evidence of multicollinearity.

The first principal components, by construction, explain most of the data variation. Since the variables considered are four, the last PCs do not explain a very low variation, as would happen with many more components, but generally the first two components explain more than 65% of the original variation, and the first three around 85%.

Principal component number	Variance explained
Component 1	40.207
Component 2	25.044
Component 3	19.134
Component 4	15.616

Table 4.4: Principal Component Analysis on Spanish data: the first three components explain around 85% of the original variation.

Regarding the interpretation of the first two principal components there is a clear recurring pattern. For each country, the first principal component explains most of the variations, measuring the overall level of the loan variables. This is a

common result in PCA: it simply shows that the greatest source of variation is the magnitude of the variables. As the entries of the correlation matrices are mainly positive, this could have been expected. On the contrary, the second principal component, both for the floating and for the 2004-2006 fixed interest with only two exceptions (UK and 2004-2006 fixed Belgium), contrasts two groups. The distinction is made between the groups [*Primary Income*, *Original Balance*] and [*Loan Term*, *Loan to Value*], which geometrically can be represented as in figure 6 (see Appendix A). A possible economic interpretation has been anticipated during the correlation analysis. A higher primary income leads both to the possibility to pledge more collateral and to exploit the loan for a a shorter time. The last consideration is supported by the possibility of a more rapid repayment which a wealthier borrower actually can do. The other variable, the *Original Balance*, is positively related to the primary income, which means that the loans of a greater amount are associated with a higher primary income. This is straightforward. Finally, the contrast between the original balance and the Loan to Value ratio suggests that the increase of collateral value is more than proportional than the increase in the original balance. This supports the logical idea of a positive link between the amount granted and the collateral required. These conclusions are again quite different for the fixed interest rates. Among four countries analyzed, two (Belgium and Netherlands) exhibit the same contrast between the first variable (*Primary Income*) and the other three nations. This means that when the original balance is higher, the fixed interest rate contract generally does not require more collateral, and at the same time that a higher primary income is in contrast with the original borrowed sum. With more countries, it would have been interesting to examine if this was a specific feature of the post-crisis fixed rate contracts.

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	$+$	$+$
Original Balance	$+$	$+$
Loan Term	$+$	$-$
Loan to Value	$+$	$-$

Table 4.5: The coefficient signs of the first two principal components computed on Irish data. The second PC contrasts the *Primary Income* and the *Original Balance* with the other two variables.

The theoretical considerations about the optimal PCs subset choice examined through this chapter are implemented in our analysis. The number of PCs retained ranges between 3 and 4 (in the last case, all the components are retained). The selection is based on various decision rules, but since the number of components is not very high, they usually are not very discordant. As we pointed out at the beginning of this section, this analysis deleting the last PC retains most of the original variation ($\approx 85\%$), but for a less-arbitrary decision three criteria were considered: the Rule of Thumb/RSS, the T-statistic and the Cross-Validation MSE. As is displayed in figure 4, the cross-validation method allows us to consider the subset with the lowest estimated MSE. The initial rapid decrease, when too few PCs are retained, stops when more than two PCs are chosen, coherently with the theoretical discussion about the overfitting (Section 4.4.3). The T-statistic is usually highly significant for at least three of the four components. In the tables of Appendix A, the PCs reported are those with a highly significant p-value (usually under 1%). Finally, the rule of thumb based on the RSS comparisons, and on the total variation explained, sometimes suggests a strong relation between the last PC and the dependent variable, cannot support its removal from the analysis. The MSE varies among the estimates for each country, ranging from 0.28 to 3.62, with a lot of values around 1. The other question, the selection of the independent variables, is applied to the first few analyses mainly to support the decision to retain all four original variables. As a matter of fact, the F-statistic test always rejects the null hypothesis, implying that all the predictors are significant, as can

be expected since there are not many independent variables.

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3
T-statistic ($\alpha = 0.1$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3

Table 4.6: The results of the different decision rules discussed in the theoretical part, applied to German data. In this case the CV technique suggests to retain the first three PCs, minimizing the Mean Squared Error.

The final estimates $\hat{\beta}$, reconstructed for the original variable terms starting from the results obtained through the principal component regression on the dependent variable *Interest Rate Margin*, are presented for each country and both for floating and fixed interest rates in the tables of Appendix A. The comment is made on the sign rather than on the magnitude and is always to be intended as a result of a general pattern (considering the totality of the contracts) rather than applicable to each singular contract. In the following chapter we analyze whether the effect given by a particular coefficient is weak due to its magnitude, employing shrinkage estimators.

Starting from the floating loan analysis, the primary income influences positively the variable y in each country except for the Netherlands, where its sign is negative. Economically, this coefficient suggests that the level of the interest rate applied in a (RMBS) loan is higher for borrowers with higher income. Among the possible explanation there is the greater propensity to pay for the wealthier borrowers, given the decreasing marginal utility theory. On the contrary, the original balance is negatively related to the dependent variable, except for France, Ireland and the Netherlands. An intuitive explanation is that probably a higher amount borrowed allows the debtor to obtain better contractual conditions. The effect of the third variable, the loan term, is more frequently positive. The related coefficient is greater than 0 in Belgium, Spain, Germany, Ireland and the Netherlands. A negative coefficient is estimated for the UK, France and Italy.

Examining the overall effect for each country, since a geographical pattern is not clearly evident, the analysis suggests that in some countries the interest rate can be higher as a remuneration for the longer loan term, in other countries that a longer loan agreement implies more favorable conditions.

Finally, the question posed in the first part of this work is answered by the coefficient estimated for the Loan to Value variable (see Section 4.6). Given the amount borrowed, from the definition of the Loan to Value, a positive coefficient implies a negative link between collateral and interest rates. A higher interest rate is associated with a higher Loan to Value, and then with lower collateral requirements, in Belgium, the UK, Spain, Germany and Ireland. In the other three nations, Italy, the Netherlands and France the latter relationship is positive. This empirically supports our thesis on the *a priori* impossibility to decide for one theoretical hypothesis. In fact, for the first group of countries the prevalent effect is the signalling related to the case of asymmetric information between the borrower and the lender. In the second minority group the moral hazard seems to be prevalent: as we proved in our model, *"higher interest rates and higher collateral requirements are compatible if they are both considered as a penalization for a riskier borrower"*.

The economic analysis for fixed interest rates is similar, and the empirical conclusions support the hypothesis of a negative link between collateral and interest rates. Specifically, both for 2004-2006 and for 2011-2016 periods, the coefficient estimated for the Loan to Value is positive for 6 of the 7 cases analyzed. Moreover, the coefficient for the variable Original Balance is negative in every country except for France, which is an exception also for the floating interest rate for this coefficient.

To support the estimated coefficient for the Loan to Value variable, as discussed above, a cluster analysis is done for example on the UK and German data, see figures 1 and 5 (Appendix A).

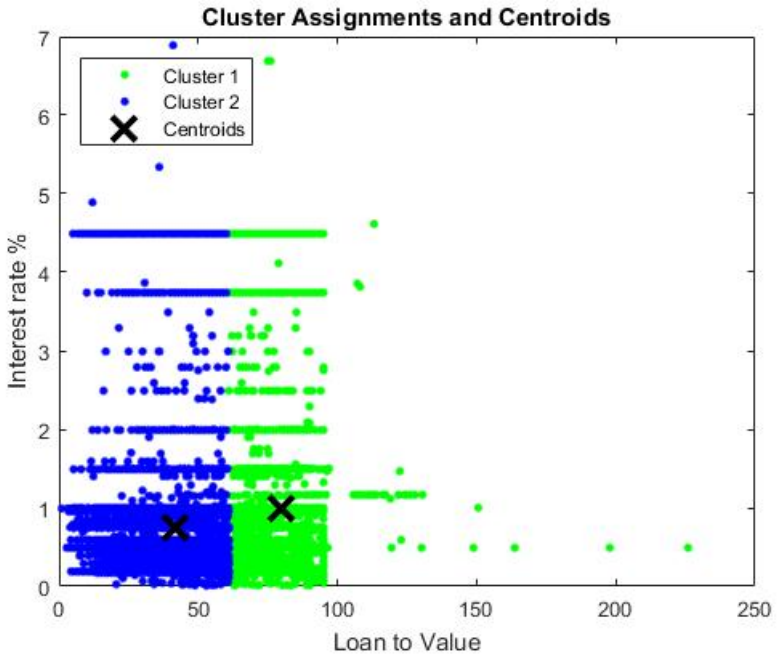

Figure 1: The positive coefficient for the variable *Loan to Value*, found through the Principal Component Regression applied to UK data, is confirmed by a cluster analysis: the centroid for the contracts with a higher *Loan to Value* corresponds to a higher interest rate margin.

This graphic suggests at least two interesting considerations. Firstly, as we have affirmed above, the effect estimated is the prevailing one, and it is not valid for every contract. As a matter of fact in these graphics it can be seen that each contract is associated to its Loan to Value and its Interest Rate, and there are contracts with a high loan to value and low interest rate, as well as the opposite.

There is not a clear cluster. This confirms again the existing different game theory effects which are specific for each contract.

On the other hand, considering the two centroids the second has both a higher Loan to Value and a higher Interest Rate. This link is not so strong, but it is the prevailing overall effect, coherently with the positive value for this coefficient estimated by the PCR. In other words, the main point here is that considering the total of the contracts, those with a *higher* Loan to Value tend to be clustered into the group whose centroid has a *higher* interest rate, since this *minimizes* the possible distance for these contracts.

These results imply another important consideration. An empirical analysis which derives its conclusions from a small number of observations, or from an analysis on a single country, can be misleading and biased since there is not a link which is nearly always valid. In fact, the discordant coefficients we have estimated indicate the prevalent effect observed on millions of different contracts of different countries. But, more importantly, it indicates that for different contracts, and for different countries, different hypotheses could be valid. In the next chapter we discuss other possible biased methods, which are based on the regularization, and analyze whether their estimates are in line with the conclusions reached in this section.

4.6 A summary table

We highlight the most important results of the empirical analysis in this summary table. The other results, and the estimates for each country, are presented in the immediately following appendix.

Country	Effect of collateral on floating interest rates
Belgium, the UK, Spain, Germany, Ireland	$-$
Italy, the Netherlands, France	$+$
	Effect of collateral on fixed interest rates
Belgium	$+$
The UK, Spain, Germany, Ireland	$-$
Italy, the Netherlands, France	$-$
	Other relevant statistics
Number of PCs usually retained	PC1, PC2, PC3
Variance explained by the first three PCs	$\approx 85\%$ (on average)

Appendix A

Empirical results. Principal component regression

A.1 Analysis for floating interest rates

A.1.1 Belgium

$$\hat{\Sigma}_{Belgium} = \begin{bmatrix} 1.0000 & & & \\ 0.21860 & 1.0000 & & \\ -0.10556 & 0.29683 & 1.0000 & \\ -0.10189 & 0.14899 & 0.21622 & 1.0000 \end{bmatrix}$$

Table A.1: Belsley collinearity diagnostics

Singular values	Condition indices
1^{st}	1.0000
2^{nd}	1.1085
3^{rd}	1.3453
4^{th}	1.5868

Table A.2: Variation explained in PCA. Belgium

Principal component number	Variance explained
Component 1	36.185
Component 2	29.450
Component 3	19.993
Component 4	14.372

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	$*$	$+$
Original Balance	$+$	$+$
Loan Term	$+$	$-$
Loan to Value	$+$	$-$

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/RSS	PC1, PC2, PC4
T-statistic ($\alpha = 0.05$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

F-test		
$Min(\frac{\min u_i^k}{MSE}) = 4.1575$	$F-stat = 3.84$	Variables to exclude: None

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.0199	$+$
Original Balance	-0.0115	$-$
Loan Term	0.1257	$+$
Loan To Value	0.10723	$+$
MSE	1.9465	

A.1.2 UK

$$\hat{\Sigma}_{UK} = \begin{bmatrix} 1.0000 & & & \\ 0.15728 & 1.0000 & & \\ -0.019896 & 0.13989 & 1.0000 & \\ 0.0056817 & 0.24755 & 0.46365 & 1.0000 \end{bmatrix}$$

Table A.3: Belsley collinearity diagnostics

Singular values	Condition indices
1^{st}	1.0000
2^{nd}	1.2070
3^{rd}	1.4272
4^{th}	1.7504

Table A.4: Variation explained in PCA. UK

Principal component number	Variance explained
Component 1	39.940
Component 2	27.414
Component 3	19.610
Component 4	13.036

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	−
Original Balance	+	+
Loan Term	+	−
Loan to Value	+	−

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3
T-statistic ($\alpha = 0.04$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

PCR final results		
Variable	Coefficient	Effect
Intercept	1.3315	
Primary Income	0.035627	+
Original Balance	−0.42502	−
Loan Term	−0.043499	−
Loan To Value	0.19186	+
MSE	2.2989	

A.1.3 Spain

$$\hat{\Sigma}_{Spain} = \begin{bmatrix} 1.0000 & & & \\ 0.021481 & 1.0000 & & \\ -0.0084544 & 0.35704 & 1.0000 & \\ -0.0037192 & 0.23818 & 0.31844 & 1.0000 \end{bmatrix}$$

Table A.5: Variation explained in PCA. Spain

Principal component number	Variance explained
Component 1	40.207
Component 2	25.044
Component 3	19.134
Component 4	15.616

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	+
Original Balance	+	*
Loan Term	+	*
Loan to Value	+	*

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3, PC4
T-statistic ($\alpha = 0.04$)	PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.0036458	+
Original Balance	−0.078313	−
Loan Term	0.050054	+
Loan To Value	0.026419	+
MSE	0.2882	

A.1.4 France

$$\hat{\Sigma}_{France} = \begin{bmatrix} 1.0000 & & & \\ 0.15917 & 1.0000 & & \\ -0.19793 & 0.21913 & 1.0000 & \\ -0.069169 & 0.16672 & 0.26524 & 1.0000 \end{bmatrix}$$

Table A.6: Variation explained in PCA. France

Principal component number	Variance explained
Component 1	36.338
Component 2	28.935
Component 3	19.443
Component 4	15.283

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	*	+
Original Balance	+	+
Loan Term	+	*
Loan to Value	+	*

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3
T-statistic ($\alpha = 0.04$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.31300	+
Original Balance	0.029993	+
Loan Term	−0.13342	−
Loan To Value	−0.020734	−
MSE	0.7192	

A.1.5 Germany

$$\hat{\Sigma}_{Germany} = \begin{bmatrix} 1.0000 & & & \\ 0.15681 & 1.0000 & & \\ -0.020135 & 0.13352 & 1.0000 & \\ 0.0055376 & 0.24718 & 0.46287 & 1.0000 \end{bmatrix}$$

Table A.7: Variation explained in PCA. Germany

Principal component number	Variance explained
Component 1	32.710
Component 2	26.907
Component 3	24.750
Component 4	15.633

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	*	+
Original Balance	*	+
Loan Term	+	−
Loan to Value	+	*

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3
T-statistic ($\alpha = 0.04$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.021645	+
Original Balance	−0.34960	−
Loan Term	0.31400	+
Loan To Value	0.28106	+
MSE	3.0071	

A.1.6 Ireland

$$\hat{\Sigma}_{Ireland} = \begin{bmatrix} 1.0000 & & & \\ 0.22666 & 1.0000 & & \\ -0.055005 & 0.19398 & 1.0000 & \\ 0.032021 & 0.26106 & 0.43792 & 1.0000 \end{bmatrix}$$

Table A.8: Variation explained in PCA. Ireland

Principal component number	Variance explained
Component 1	40.600
Component 2	28.405
Component 3	17.183
Component 4	13.812

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	+
Original Balance	+	+
Loan Term	+	−
Loan to Value	+	−

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2
T-statistic ($\alpha = 0.04$)	PC1, PC2 PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.016572	+
Original Balance	0.033523	+
Loan Term	0.022233	+
Loan To Value	0.026827	+
MSE	1.8919	

A.1.7 Netherlands

$$\hat{\Sigma}_{Netherlands} = \begin{bmatrix} 1.0000 & & & \\ 0.33716 & 1.0000 & & \\ -0.034690 & -0.025170 & 1.0000 & \\ 0.084148 & 0.035977 & 0.069199 & 1.0000 \end{bmatrix}$$

Table A.9: Variation explained in PCA. Netherlands

Principal component number	Variance explained
Component 1	33.982
Component 2	26.680
Component 3	22.870
Component 4	16.467

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	*
Original Balance	+	−
Loan Term	*	+
Loan to Value	*	+

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3, PC4
T-statistic ($\alpha = 0.04$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

PCR final results		
Variable	Coefficient	Effect
Primary Income	−0.20686	−
Original Balance	0.17934	+
Loan Term	0.046619	+
Loan To Value	−0.18101	−
MSE	1.2649	

A.1.8 Italy

$$\hat{\Sigma}_{Italy} = \begin{bmatrix} 1.0000 & & & \\ 0.1434 & 1.0000 & & \\ -0.0186 & 0.2230 & 1.0000 & \\ -0.0006 & 0.2184 & 0.3803 & 1.0000 \end{bmatrix}$$

Table A.10: Variation explained in PCA. Italy

Principal component number	Variance explained
Component 1	39.048
Component 2	26.6360
Component 3	18.8450
Component 4	15.4740

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	+
Original Balance	+	+
Loan Term	+	−
Loan to Value	+	−

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3
T-statistic	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.0153	+
Original Balance	−0.2197	−
Loan Term	−0.0652	−
Loan To Value	−0.0531	−
MSE	2.4944	

A.2 Analysis for fixed interest rates: 2004-2006

A.2.1 Belgium

$$\hat{\Sigma}_{Belgium} = \begin{bmatrix} 1.0000 & & & \\ 0.065811 & 1.0000 & & \\ 0.0024815 & 0.34966 & 1.0000 & \\ 0.021600 & 0.25591 & 0.32396 & 1.0000 \end{bmatrix}$$

Table A.11: Belsley collinearity diagnostics

Singular values	Condition indices
1^{st}	1.0000
2^{nd}	1.2736
3^{rd}	1.4795
4^{th}	1.6085

Table A.12: Variation explained in PCA. Belgium

Principal component number	Variance explained
Component 1	40.653
Component 2	25.064
Component 3	18.571
Component 4	15.713

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	−
Original Balance	+	+
Loan Term	+	−
Loan to Value	+	−

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3
T-statistic ($\alpha = 0.04$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3

F-test		
$Min(\frac{\min u_i^k}{MSE}) = 4.37$	$F-stat = 3.84$	Variables to exclude: None

PCR final results		
Variable	Coefficient	Effect
Primary Income	-0.011049	$-$
Original Balance	-0.31753	$-$
Loan Term	-0.25217	$-$
Loan To Value	-0.089406	$-$
MSE	1.9704	

A.2.2 Netherlands

$$\hat{\Sigma}_{Netherlands} = \begin{bmatrix} 1.0000 & & & \\ 0.075851 & 1.0000 & & \\ 0.011198 & -0.050778 & 1.0000 & \\ 0.0022840 & 0.097338 & 0.036386 & 1.0000 \end{bmatrix}$$

Table A.13: Variation explained in PCA. Netherlands

Principal component number	Variance explained
Component 1	28.138
Component 2	25.729
Component 3	24.829
Component 4	21.303

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	$+$	$-$
Original Balance	$+$	$-$
Loan Term	*	$+$
Loan to Value	$+$	$+$

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC4
T-statistic ($\alpha = 0.04$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

PCR final results		
Variable	Coefficient	Effect
Primary Income	-0.0022467	$-$
Original Balance	-0.026443	$-$
Loan Term	0.043101	$+$
Loan To Value	0.079255	$+$
MSE	0.7705	

A.2.3 Italy

$$\hat{\Sigma}_{Italy} = \begin{bmatrix} 1.0000 & & & \\ 0.13736 & 1.0000 & & \\ 0.0021499 & 0.38335 & 1.0000 & \\ 0.0075177 & 0.32641 & 0.43651 & 1.0000 \end{bmatrix}$$

Table A.14: Variation explained in PCA. Italy

Principal component number	Variance explained
Component 1	44.363
Component 2	25.649
Component 3	16.240
Component 4	13.748

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	+
Original Balance	+	+
Loan Term	+	−
Loan to Value	+	−

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC4
T-statistic ($\alpha = 0.04$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.0045507	+
Original Balance	−0.091774	−
Loan Term	0.28609	+
Loan To Value	0.00094786	+
MSE	3.6221	

A.3 Analysis for fixed interest rates: 2011-2016

A.3.1 Belgium

$$\hat{\Sigma}_{Belgium} = \begin{bmatrix} 1.0000 & & & \\ 0.47777 & 1.0000 & & \\ 0.13899 & 0.47612 & 1.0000 & \\ 0.22159 & 0.29693 & 0.43491 & 1.0000 \end{bmatrix}$$

Table A.15: Variation explained in PCA. Belgium

Principal component number	Variance explained
Component 1	50.978
Component 2	23.496
Component 3	16.292
Component 4	9.2340

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	+
Original Balance	+	−
Loan Term	+	−
Loan to Value	+	−

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC4
T-statistic ($\alpha = 0.1$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.017098	$+$
Original Balance	-0.23177	$-$
Loan Term	0.19269	$+$
Loan To Value	0.28076	$+$
MSE	0.7578	

A.3.2 France

$$\hat{\Sigma}_{France} = \begin{bmatrix} 1.0000 & & & \\ 0.060393 & 1.0000 & & \\ -0.056128 & 0.39090 & 1.0000 & \\ 0.0046009 & 0.21268 & 0.23404 & 1.0000 \end{bmatrix}$$

Table A.16: Variation explained in PCA. France

Principal component number	Variance explained
Component 1	39.175
Component 2	25.416
Component 3	20.619
Component 4	14.790

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	*	+
Original Balance	+	+
Loan Term	+	−
Loan to Value	+	*

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3 ,PC4
T-statistic ($\alpha = 0.1$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3, PC4

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.0063327	+
Original Balance	0.048759	+
Loan Term	0.13448	+
Loan To Value	0.030659	+
MSE	0.8843	

A.3.3 Germany

$$\hat{\Sigma}_{Germany} = \begin{bmatrix} 1.0000 & & & \\ 0.13260 & 1.0000 & & \\ -0.0037796 & 0.24019 & 1.0000 & \\ 0.015499 & 0.18177 & 0.27145 & 1.0000 \end{bmatrix}$$

Table A.17: Variation explained in PCA. Germany

Principal component number	Variance explained
Component 1	36.940
Component 2	25.894
Component 3	19.613
Component 4	17.553

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	+
Original Balance	+	+
Loan Term	+	−
Loan to Value	+	−

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3
T-statistic ($\alpha = 0.1$)	PC1, PC2, PC3, PC4
Cross-Validation MSE	PC1, PC2, PC3

PCR final results		
Variable	Coefficient	Effect
Primary Income	0.012740	+
Original Balance	−0.12717	−
Loan Term	−0.070708	−
Loan To Value	0.017765	+
MSE	0.4668	

A.3.4 Netherlands

$$\hat{\Sigma}_{Netherlands} = \begin{bmatrix} 1.0000 & & & \\ 0.13260 & 1.0000 & & \\ -0.051453 & 0.035089 & 1.0000 & \\ -0.069043 & 0.086023 & 0.049338 & 1.0000 \end{bmatrix}$$

Table A.18: Variation explained in PCA. Netherlands

Principal component number	Variance explained
Component 1	32.984
Component 2	27.154
Component 3	23.817
Component 4	16.045

Coefficient Matrix		
Variable	PC1	PC2
Primary Income	+	−
Original Balance	+	+
Loan Term	*	+
Loan to Value	*	+

Choosing the optimal subset	
Decision rule	PCs retained
Rule of Thumb/ RSS	PC1, PC2, PC3
T-statistic ($\alpha = 0.1$)	PC1, PC2, PC3
Cross-Validation MSE	PC1, PC2, PC3

PCR final results		
Variable	Coefficient	Effect
Primary Income	−0.10927	−
Original Balance	−0.023524	−
Loan Term	−0.0090878	−
Loan To Value	0.21362	+
MSE	0.8470	

A.4 List of figures

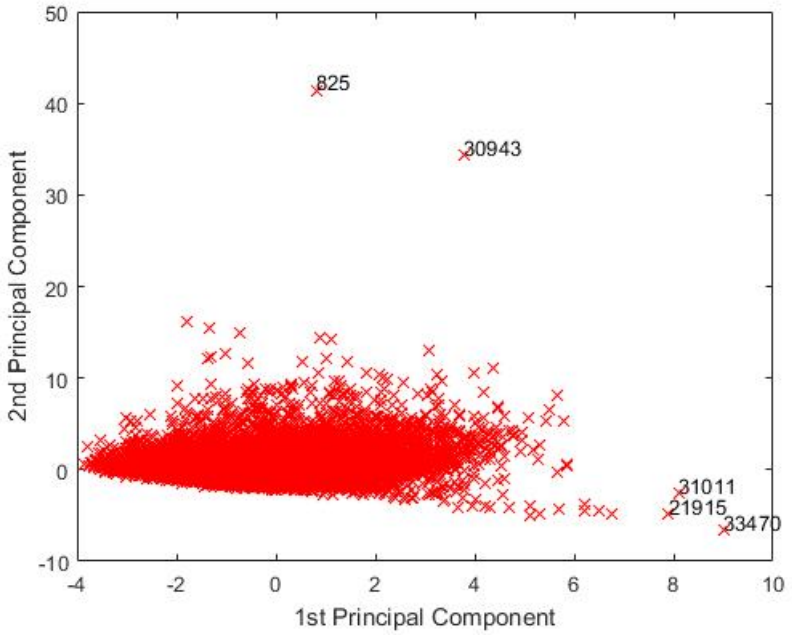

Figure 2: Detecting outliers with PCA. (Belgium)

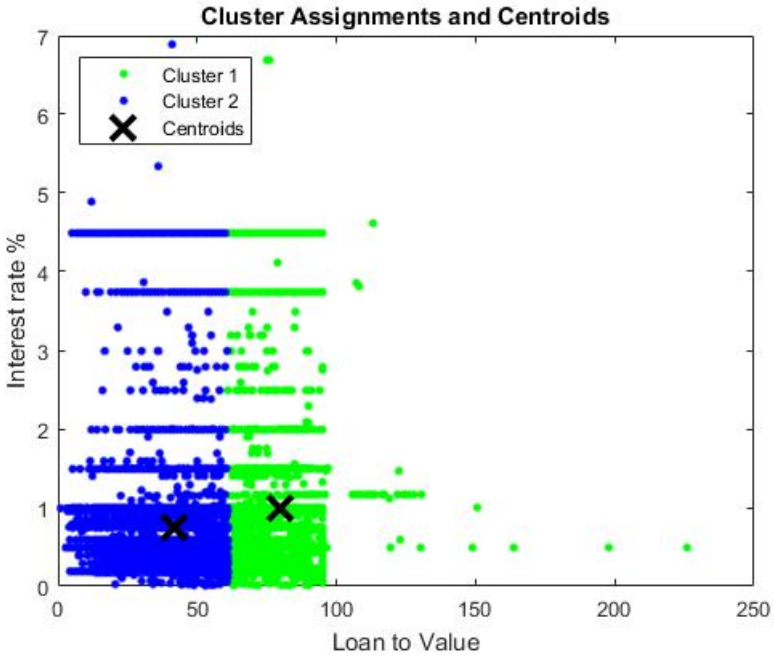

Figure 3: Cluster analysis for the data of the UK

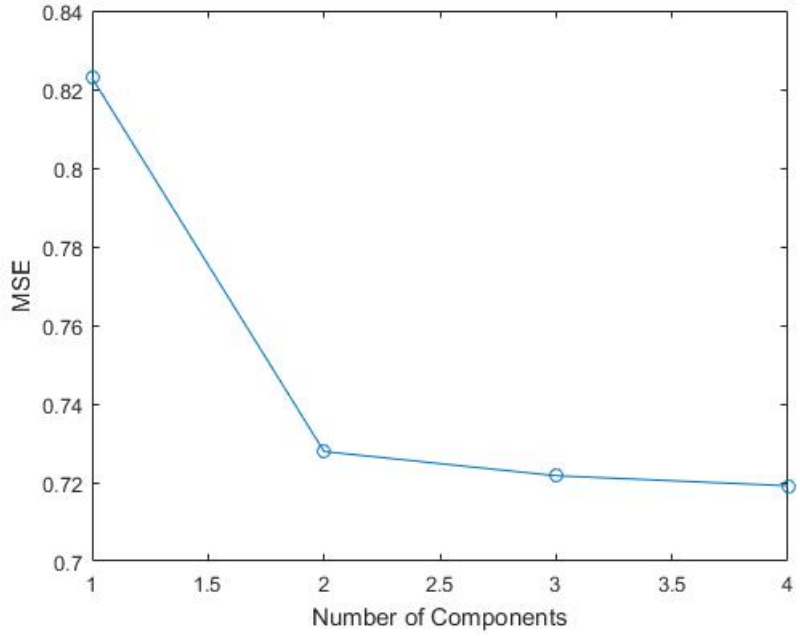

Figure 4: 10-fold Cross Validation MSE for the PCR. France.

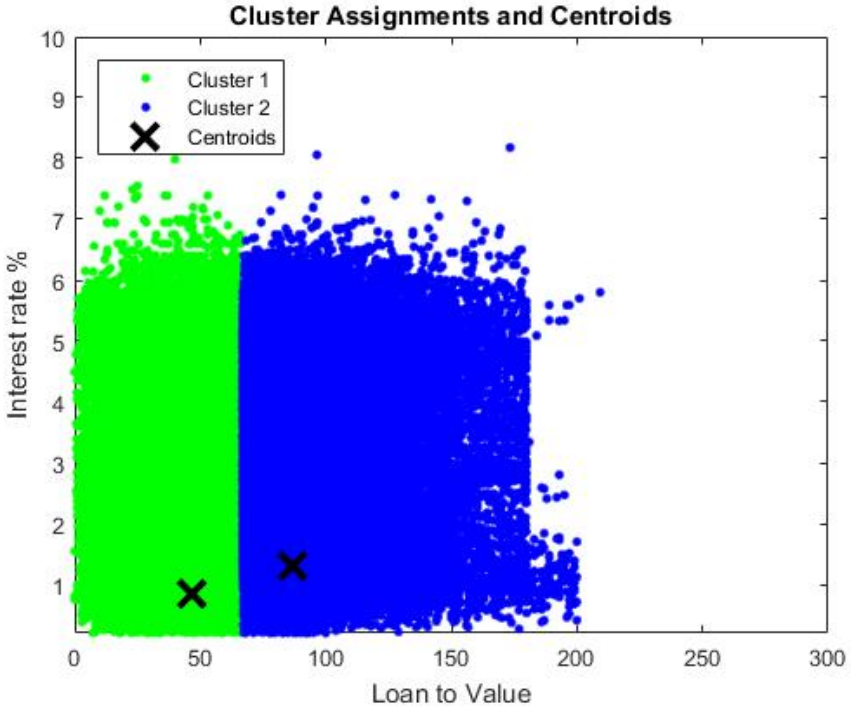

Figure 5: Cluster analysis for German data.

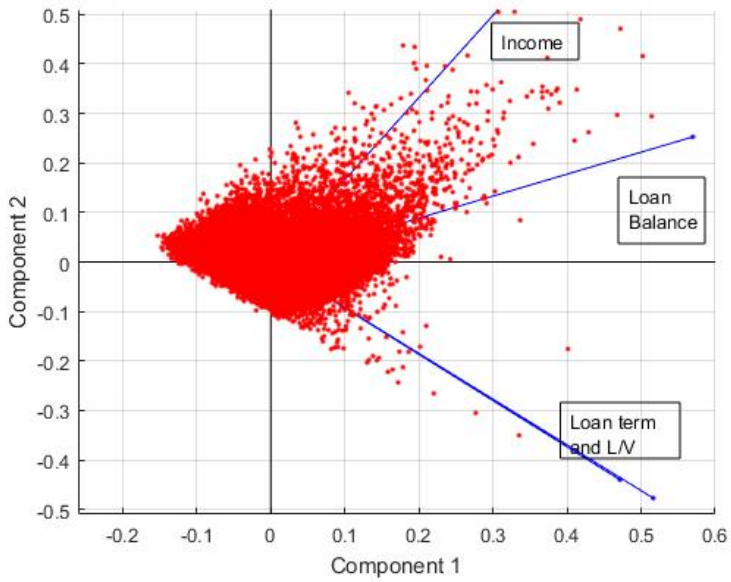

Figure 6: Geometrical interpretation of the first two PC: the principal axes for the *Primary income* and the *Original balance* are both positive in the first PC and opposite in the second PC.

Chapter 5

Shrinkage methods for supervised learning

5.1 Introduction to the LASSO and ridge regressions

5.1.1 Mathematical background: The regularization of ill-posed problems and the L^p-norm

In this section, we briefly discuss the mathematical background of regularization. As a matter of fact this technique, which has been applied to regression models, is strictly connected to other important fields of mathematics, where a more general theory has been developed. To have a clearer and, ultimately, more insightful vision of the regularization methods, it is worth briefly examining this theory; a systematic treatise on this subject can be found in Nair (2009). To explain the meaning of the term *ill-posed* problem, let consider the model

$$Tx = y \tag{5.1}$$

where T is an operator between two normed linear spaces, $T : X \longrightarrow Y$. In this general framework, we want to find a solution $\tilde{x}$ for a given $y \in Y$. Suppose we have found it, and consider a perturbation of the system, such that the new value y^* is close to y. Is the new solution $\tilde{x}^*$ close to $\tilde{x}$? If the reply is negative, the problem is called *ill-posed*; it is otherwise *well-posed*. The need for a close solution

research can be seen under two different perspectives for our purposes. First, the overfitting problem connected with the bias-variance trade-off (see Section 4.4.3) is a statistical counterpart for this question. Furthermore, a solution for $Tx = y$ can be found if and only if a corresponding y exists in the image subset. Otherwise an element $\tilde{x}_0$ should be found, so that $T\tilde{x}_0$ is close to y. Clearly, under this framework the function F

$$F : x \longrightarrow \|Tx - y\| \tag{5.2}$$

should be minimized:

$$\tilde{x}_0 : inf\{\|Tx - y\|, x \in X\}. \tag{5.3}$$

This is equivalent to the least squares solution except for a change of terminology. The regularized regressions discussed in Section 5.1.2 can be used, for example, when the condition $\boldsymbol{X'X} \sim \boldsymbol{I}$ is violated, as in the case of multicollinearity. More generally, if the problem is ill-posed a solution does not always exist. In this case, a general set of tools which find a stable approximation for an ill-posed problem can be adopted. They are called regularization techniques.

Definition 6 (Regularization family)**.** *Consider the operator family* $\{R_\alpha\}_{\alpha \geq 0}$ *such that* $R_\alpha y \longrightarrow T^+ y$ *for* $\alpha \longrightarrow 0$*. This family is a regularization family.*

The matrix denoted as T^+ is the Moore–Penrose inverse of T. It can be proved that $T^+ : R(T) + R^\perp(T) \longrightarrow X$, and therefore the importance of regularization is given by the following theorem:

Theorem 3. *Let* $T : X \longrightarrow Y$ *be a linear operator. The following statements are equivalent:*

1. $Tx = y$ *has a least squares solution*
2. $y \in R(T) + R^\perp(T)$

where $R^\perp(T)$ *is the standard notation for the orthogonal complement of the subspace* $R(T) := \{y : y \in R(T)\}$.

Proof. The proof is based on a third condition, which is proved to be equivalent to the second. Then, it suffices to prove that the first and the third conditions are

equivalent. Let $y \in Y$, and define the orthogonal projection $P : Y \longrightarrow Y$ onto $\bar{R}(T)$, the closure of R(T). The third condition is: $Tx = Py$. The second condition and the third are then equivalent, since $P(y) \in R(T) \Longleftrightarrow y \in R(T) + R^{\perp}(T)$. Moreover it can be proved that inf{$\|v - y\|, v \in \bar{R}(T)$}= $\|v - y\|$. Therefore, a least squares solution is $\|Tx_0 - y\| = \|Py - y\|$. From this equation we derive that x_0 is a least squares solution if and only if $Tx_0 = Py$. □

Among all the possible regularization methods, here the Tikhonov regularization is of interest.

Definition 7 (Tikhonov functional)**.** *The functional* $x \longrightarrow \|Tx - y\|^2 + \alpha\|x\|^2$, $x \in X_0$ *is called the Tikhonov functional.*

The solution which minimizes this functional is then the Tikhonov regularization solution. The difference between this minimization problem and those without any additional parameter is that in this case a solution always exists. The concept of Tikhonov regularization corresponds to the ridge regression and a modification of this regularization method is related to the LASSO.

The minimization problem we have generally considered until now was defined on a general norm. However in practice some particular norms are chosen, according to their specific properties related to the regularized problem (see the next section). While the Euclidean norm is the most known, its generalization considers this norm only as a particular case (Kolmogorov and Fomin, 1957):

Definition 8 (L^p-norm)**.** *An* L^p*-norm, for a vector* $x \in R^n$*, is the norm* $\|x\|_p = (\sum_{k=1}^{n} |x_k|^p)^{\frac{1}{p}}$.

If $p = 2$ the norm is called Euclidean: $\|x\|_2 := \sqrt{(x_1^2 + ... + x_n^2)}$. If $p = 1$ the correspondent ℓ^1-norm is also called *Manhattan distance*. A geometrical interpretation of different norms is given in the next section to justify the different results of the ridge and the LASSO regressions.

5.1.2 The regularization applied to regression models

The regularization applied to the regression framework was suggested for the first time by Hoerl (1962), in order to control the least squares estimation $\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$ for the usual linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Nevertheless, the later work by Hoerl and Kennard (1970) is considered the landmark for this type of technique. In their paper the original definition of the so-called *ridge estimator*, whose name comes from the similarities with the quadratic response functions, is given through a modification of the OLS estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$. The ridge estimator is defined as:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}} = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'Y}, k \geq 0 \tag{5.4}$$

The regression model whose estimator is $\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}$ is called ridge regression. They prove that the following relationship holds:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}} = (\boldsymbol{I} + k(\boldsymbol{X'X})^{-1})^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}. \tag{5.5}$$

The ridge estimator can be derived, through simple calculations, from the minimization problem

$$\min\left[\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{v}\beta_j x_{ij})^2 + k\sum_{j=1}^{v}\beta_j^2\right] = \min\left[RSS + k\sum_{j=1}^{v}\beta_j^2\right]. \tag{5.6}$$

It is clear that this formulation is the same as the Tikhonov regularization for ill-posed problems; an empirical analysis on the stability of the ridge trace ($\beta_j^k, \forall k$) is provided in Beaton et al. (1976). Compared with the least squares, the ridge formulation implicates a *shrinkage effect*. Clearly for $k = 0$ the least squares estimate is computed. The shrinkage terminology refers to the consequence that the absolute value of the estimators $\{\beta_j\}$, $j = 1, ..., v$, are skewed toward zero, an effect caused by the *shrinkage penalty* $k\sum_{j=1}^{v}\beta_j^2$. This effect can explained from the formulation of the minimization problem, since a lower value is reached if the coefficients are smaller (this effect is known as penalization). The only estimator which is not penalized is β_0, because it represents the intercept.

The parameter k is arbitrary: however, the optimal (in the MSE sense) shrinkage parameter is usually computed with the CV approach (algorithm 1).

Clearly, for $k \longrightarrow \infty$, the shrinkage effect is maximum. In practice a value $\tilde{k} < \infty$ is adopted: this means that all the original variables are retained ($\beta_j \neq 0$, $\forall j$). Then, a resulting question could be if a modification of the original ridge regression would improve the predictability or the interpretation, setting some coefficients exactly to 0. This is the reason behind the LASSO regression formulation, proposed for the first time by Tibshirani (1996), which can be formulated with a similar expression to the ridge changing the constraint:

$$\min \left[\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{v} \beta_j x_{ij})^2 + k \sum_{j=1}^{v} |\beta_j| \right] = \min \left[RSS + k \sum_{j=1}^{v} |\beta_j| \right] \tag{5.7}$$

since this formulation is equivalent to the original definition for the LASSO, which consists in finding the values $(\hat{\alpha}, \hat{\beta})$ which solve the following minimization problem:

$$arg\min \{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{v} \beta_j x_{ij})^2 \} s.t. \sum_{j=1}^{v} |\beta_j| \leq t. \tag{5.8}$$

Recalling Definition 8, the value p of the L^p-norm equals 1 in the LASSO and 2 for the ridge regression. This fact leads to a possible a geometrical explanation. The ridge regression differs from the LASSO since the constraint is a circle (a disk) for the former and to a square for the latter.

$$\ell^1 \text{ norm} = \|x\|_1 := (|x_1| + ... + |x_n|)$$

$$\ell^2 \text{ norm} = \|x\|_2 := \sqrt{(x_1^2 + ... + x_n^2)}$$

Considering this distinction, a common explanation is that at the minimum, the ellipses which represent the constant RSS points, intersect a vertex of the square, then a coordinate axis, in the LASSO case: see figure 7 for the $\mathbb{R}^2$ case (James et al., 2013). The same intuition can be extended to $\mathbb{R}^n$.

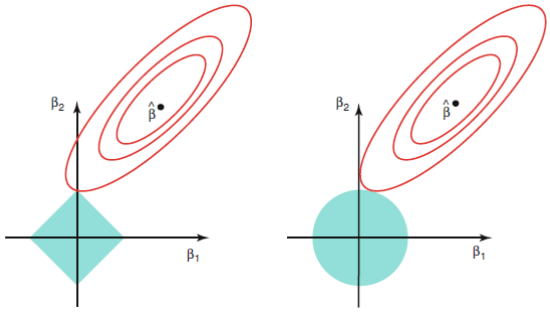

Figure 7: Geometrical interpretation of LASSO (square) and ridge (circle) constraints. (Source: Chapter 3 of James et al., 2013).

The difference can also be interpreted analytically, through convex optimization theory (Elad, 2010). Consider a maximization problem in ℓ^1:

$$\min \|\boldsymbol{x}\|_1 : \boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}. \tag{5.9}$$

As a consequence all the solutions give an equal penalty:

$$\pi_{min} := \|\boldsymbol{x_{opt}}\|_1 < \infty \tag{5.10}$$

where x is a vector. From the triangle inequality

$$\|\boldsymbol{x}^{\boldsymbol{1}}_{\boldsymbol{opt}} + \boldsymbol{x}^{\boldsymbol{2}}_{\boldsymbol{opt}}\|_1 \leq \|\boldsymbol{x}^{\boldsymbol{1}}_{\boldsymbol{opt}}\|_1 + \|\boldsymbol{x}^{\boldsymbol{2}}_{\boldsymbol{opt}}\|_1 < 2\pi_{min} \tag{5.11}$$

it is seen that all the solutions are close. Let $\|\boldsymbol{x_{opt}}\|_1$ be a solution with $k > n$ non-zeros, where n is the dimension of the column space for the matrix $\boldsymbol{A}$. From a well-known result of linear algebra, then the k elements are linearly dependent in $\mathbb{R}^n$, namely a non-trivial vector $\boldsymbol{v}$ exists, such that

$$\boldsymbol{A}\boldsymbol{v} = \boldsymbol{0} \tag{5.12}$$

For an appropriate small ϵ the relation

$$\boldsymbol{x} = \boldsymbol{x_{opt}} + \epsilon\boldsymbol{v} \tag{5.13}$$

still satisfies $\boldsymbol{A}\boldsymbol{x} = 0$, then it is still a solution (even if not the optimal one) by definition:

$$\forall|\epsilon| \leq \min_i \frac{|x^i_{opt}|}{|h^i|}, \|\boldsymbol{x}\|_1 = \|\boldsymbol{x_{opt}} + \epsilon\boldsymbol{v}\|_1 \geq \|\boldsymbol{x_{opt}}\|_1 \tag{5.14}$$

The condition on $|\epsilon|$ is needed to ensure that any element of $\boldsymbol{x}$ does not change its sign in that neighbourhood. The key point here is that the previous inequality is in fact an equality

$$\forall|\epsilon| \leq \min_i \frac{|x^i_{opt}|}{|v^i|}, \|\boldsymbol{x}\|_1 = \|\boldsymbol{x_{opt}} + \epsilon \boldsymbol{v}\|_1 = \|\boldsymbol{x_{opt}}\|_1 \tag{5.15}$$

since it must be valid both for positive and negative ϵ. Then, the $\|\boldsymbol{x_{sol}}\|_1$ solution is not changed by an addiction or subtraction of $\boldsymbol{v}$.

This result paves the way to demonstrate the possible sparsity deriving from an ℓ_1 minimization problem, since it is possible to consider a new solution $\boldsymbol{x}^{\boldsymbol{k-1}}_{\boldsymbol{opt}}$ where one entry is 0; the superscript $(k-1)$ stands for the number of non-empty elements of the vector. It is sufficient to set $\epsilon = -\frac{x^i_{opt}}{v^i}$ for an appropriate index i.

This procedure can be applied if $k > n$, or as long as the linear dependence holds. This proves the link between ℓ_1 norm and sparse solutions.

Interestingly, the regularization theory applied to regression models can also be justified through a Bayesian approach, as discussed in Lindley and Smith (1972) and Tibshirani (1996). The main idea is that, in a regression analysis, prior distributions $p(\boldsymbol{\beta})$ can be chosen for a specific model parameter and the cases of the ridge and LASSO correspond to an appropriate choice for these distributions.

5.2 Differences and analogies with the PCR. Some properties of biased estimators

The three regression methods discussed till now, the ridge, the LASSO and the PCR, share some common properties; in fact, they are strictly linked as discussed in literature. Indeed, the estimator of the ridge regression, $\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}$, is biased as well as $\hat{\boldsymbol{\beta}}_{\boldsymbol{PCR}}$, as is shown in Hoerl and Kennard (1970) and Vinod (1978):

$$Bias(\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}) = E(\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}} - \boldsymbol{\beta}) = -k\,(\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{\beta} \tag{5.16}$$

The LASSO estimator is biased too: it is not possible to derive an analytical form, but simulation studies show that $Bias(\hat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}})_{k^1} > Bias(\hat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}})_{k^2} \neq 0$ if $k^1 > k^2$ where k is the shrinkage parameter.

As for the PCR, the rationale behind defining a biased estimator, which is clearly a negative property, is the bias-variance trade-off. Both $\hat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}$ could lead to a lower MSE compared with $\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$, since their variance could be lower. Again, simulations are needed to compute the variance for the LASSO, and the results show that $Var(\hat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}})_{k^1} < Var(\hat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}})_{k^2}$ if $k^1 > k^2$. Similar conclusions are reached for ridge regression. However, the variance for the ridge estimator can also be derived analytically, as well as its MSE (Hoerl and Kennard, 1970): if

$$\boldsymbol{Z} := [\boldsymbol{I} + k(\boldsymbol{X'X})^{-1})^{-1}] \tag{5.17}$$

then

$$Var(\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}) = \boldsymbol{Z}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}(Var\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{Z'} = \sigma^2\boldsymbol{Z}(\boldsymbol{X'X})^{-1}\boldsymbol{Z'} \tag{5.18}$$

which is decreasing as k increases, due to the definition of the matrix $\boldsymbol{Z}$ where k is raised to the power of minus one. The total effect on the MSE accounts both for a higher bias and a lower variance:

$$MSE(\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}) = Tr(Var\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}) + \boldsymbol{\beta'}(\boldsymbol{Z} - \boldsymbol{I})'(\boldsymbol{Z} - \boldsymbol{I})\boldsymbol{\beta} \tag{5.19}$$

implies

$$MSE \propto Var(\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}) + Bias^2(\hat{\boldsymbol{\beta}}_{\boldsymbol{ridge}}) \tag{5.20}$$

Therefore, the performance of the three biased estimators examined up until now can be compared in terms of MSE, and in terms of interpretability too.

The similarities between the ridge and the principal component regression suggest a more general analysis. Indeed, as we have already discussed, the LASSO tends to reduce the given set of independent variables contrary to the ridge and the PCR. Hocking et al. (1976) define a general class of biased estimators

$$\tilde{\boldsymbol{\gamma}} = \boldsymbol{B}\hat{\boldsymbol{\gamma}} \tag{5.21}$$

where $\boldsymbol{B}$ is a diagonal matrix whose non-zero elements are $b_i = \sum_{j=i}^{v} a_j$, which should be set depending on the specific regression model. Indeed, they show that different choices for a_j are related to different biased estimators, including those of PCR and ridge regression. Trenkler and Trenkler (1984) extend the analysis of Hsuan

(1981) on the analogies between the PCR and the ridge regression in the case of multicollinearity, showing that in some cases the Euclidean distance between the respective two estimators can be bounded. A simulation study comparing some biased estimators is given by Hoerl et al. (1986), and the PCR seems to perform worse than the ridge regression; however, this result could depend on the procedure carried out. The same conclusion, in terms of a lower MSE, is achieved in Al-Hassan and Al-Kassab (2009) using Monte Carlo simulations.

5.3 Empirical results compared with the PCR

In this section we briefly discuss the results presented in Appendix B. In particular we examine the effects of regularization methods compared with the previous estimates; for an economic explanation of these results, the variable description, the method followed and any other remarks we refer to the previous chapter, Section 4.5.

Firstly, the signs found for the coefficients with the PCR are not changed by the LASSO or the ridge regression. Mainly, the dissimilarities caused by the regularization in our analysis are the differences of the coefficient magnitudes, in any case usually negligible. Moreover the MSE estimated is not significantly changed by the application of these regression methods. This effect can be partially explained by the low number of predictors, which instead are abundant in other analyses where the LASSO and ridge regression are usually employed. In other terms, we find that in our analysis the PCR did not perform noticeably worse than these regularized methods (see figure 8). Hence, generally the results of the present section support those found in the previous analysis. In a few cases the LASSO estimates differentiate for a sparser solution compared with the ridge and the PCR. For example, in the case of UK (see B.1.2, Appendix B), the *Primary Income* and the *Loan Term* have a zero coefficient. This is a consequence of the low absolute value estimated for these coefficients in the ridge regression (and in the PCR), combined with the minimization condition posed on the MSE. Then,

in this case the weak effect of these two variables is totally nullified by the ℓ^1 regularization. On the other hand, the two coefficients estimated for the *Original Balance* and the *Loan to Value* maintain their sign, in spite of a change in their magnitude. This suggests that the overall positive relationship between the loan to value and the interest rate margin, already discussed in the previous chapter for the PCR and confirmed by the cluster analysis, is still valid for the UK.

Again, these results based on a huge number of contracts confirm the absence of an univocal effect among the main European countries.

Appendix B

Empirical results: ridge and LASSO

B.1 Analysis for floating interest rates

B.1.1 Belgium

ridge regression. Belgium			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	0.019925	+	None
Original Balance	−0.011468	−	None
Loan Term	0.12567	+	None
Loan To Value	0.10720	+	None
MSE	1.9468		None

LASSO. Belgium			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.019670	+	None
Original Balance	−0.011174	−	None
Loan Term	0.12544	+	None
Loan To Value	0.10703	+	None
MSE	1.9469		None

B.1.2 UK

ridge regression. UK			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	0.035627	+	None
Original Balance	−0.42502	−	None
Loan Term	−0.043499	−	None
Loan To Value	0.19186	+	None
MSE	2.4414		None

LASSO. UK			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.00000	No Effect	Sparse solution
Original Balance	−0.33450	−	None
Loan Term	0.00000	No Effect	Sparse solution
Loan To Value	0.084944	+	None
MSE	2.3133		None

B.1.3 Spain

ridge regression. Spain			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	0.0036458	+	None
Original Balance	−0.078313	−	None
Loan Term	0.050054	+	None
Loan To Value	0.026419	+	None
MSE	0.28839		None

LASSO. Spain			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.00000	No effect	Sparse solution
Original Balance	−0.071431	−	None
Loan Term	0.044837	+	None
Loan To Value	0.022421	+	None
MSE	0.2880		None

B.1.4 France

ridge regression. France			
Variable	Coefficient	Effect	Differences with PCR
Intercept	1.1067		
Primary Income	0.31300	+	None
Original Balance	0.029993	+	None
Loan Term	−0.13342	−	None
Loan To Value	−0.020734	−	None
MSE	0.72032		None

LASSO. France			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.31286	+	None
Original Balance	0.029653	+	Same as ridge
Loan Term	−0.13318	−	None
Loan To Value	−0.020485	−	None
MSE	0.7202		None

B.1.5 Germany

ridge regression. Germany			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	0.021645	+	None
Original Balance	−0.34960	−	None
Loan Term	0.31400	+	None
Loan To Value	0.28106	+	None
MSE	3.0074		None

LASSO. Germany			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.020765	No Effect	Sparse solution
Original Balance	−0.34867	−	None
Loan Term	0.31351	No Effect	Sparse solution
Loan To Value	0.28024	+	None
MSE	3.0074		None

B.1.6 Ireland

ridge regression. Ireland			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	0.014565	+	None
Original Balance	0.032035	+	None
Loan Term	0.0082966	+	None
Loan To Value	0.041476	+	None
MSE	1.8922		None

LASSO. Ireland			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.010218	+	None
Original Balance	0.029541	+	None
Loan Term	0.0048329	+	Shrinkage effect compared with PCR
Loan To Value	0.038981	+	None
MSE	1.8921		None

B.1.7 Netherlands

ridge regression. Netherlands			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	-0.20686	$-$	None
Original Balance	0.17934	$+$	None
Loan Term	0.046619	$+$	None
Loan To Value	-0.18101	$-$	None
MSE	1.2683		None

LASSO. Netherlands			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	-0.20204	$-$	None
Original Balance	0.17404	$+$	None
Loan Term	0.042956	$+$	None
Loan To Value	-0.17752	$-$	None
MSE	1.2682		None

B.1.8 Italy

ridge regression. Italy			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	0.0134	$+$	None
Original Balance	-0.2181	$-$	None
Loan Term	-0.0931	$-$	None
Loan To Value	-0.0262	$-$	None
MSE	2.4938		None

LASSO. Italy			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.0129	+	None
Original Balance	−0.2177	−	None
Loan Term	−0.09296	+	None
Loan To Value	−0.0260	−	None
MSE	2.4912		None

B.2 Analysis for fixed interest rates: 2004-2006

B.2.1 Belgium

ridge regression. Belgium			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	-0.015696	–	None
Original Balance	-0.29455	–	None
Loan Term	-0.28577	–	None
Loan To Value	-0.075664	–	None
MSE	2.1868		Higher
LASSO. Belgium			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.00000	No effect	Sparse solution
Original Balance	-0.21632	–	None
Loan Term	-0.21391	–	None
Loan To Value	0.00000	No effect	Sparse Solution
MSE	1.9991		Lower than ridge; Higher than PCR

B.2.2 Netherlands

ridge regression. Netherlands			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	-0.0022467	$-$	None
Original Balance	-0.026443	$-$	None
Loan Term	0.043101	$+$	None
Loan To Value	0.079255	$+$	None
MSE	0.77223		None

LASSO. Netherlands			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.00000	No effect	Sparse solution
Original Balance	-0.013738	$-$	None
Loan Term	0.032022	$+$	None
Loan To Value	0.066244	$+$	None
MSE	0.7711		None

B.2.3 Italy

ridge regression. Italy			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	0.0045507	$+$	None
Original Balance	-0.091774	$-$	None
Loan Term	0.28609	$+$	None
Loan To Value	0.00094786	$+$	None
MSE	3.6529		None

LASSO. Italy			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.00000	No effect	Sparse solution
Original Balance	-0.033094	$-$	None
Loan Term	0.22840	$+$	None
Loan To Value	0.00000	No effect	Sparse solution
MSE	3.6278		None

B.3 Analysis for fixed interest rates: 2011-2016

B.3.1 Belgium

ridge regression. Belgium			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	0.017098	$+$	None
Original Balance	-0.23177	$-$	None
Loan Term	0.28076	$+$	None
Loan To Value	0.0095291	$+$	None
MSE	0.75798		None

LASSO. Belgium			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.017060	$+$	None
Original Balance	-0.23171	$-$	None
Loan Term	0.28072	$+$	None
Loan To Value	0.0095181	$+$	
MSE	0.7580		None

B.3.2 France

ridge regression. France			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	0.0063327	+	None
Original Balance	0.048759	+	None
Loan Term	0.13448	+	None
Loan To Value	0.030659	+	None
MSE	0.88429		None

LASSO. France			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.0054513	+	None
Original Balance	0.048307	+	None
Loan Term	0.13388	+	None
Loan To Value	0.030023	+	None
MSE	0.8843		None

B.3.3 Germany

ridge regression. Germany			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	6.7606×10^{-8}	+	None
Original Balance	-0.13070	$-$	None
Loan Term	-0.064863	$-$	None
Loan To Value	0.014495	+	None
MSE	0.46687		None

LASSO. Germany			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	0.014554	+	None
Original Balance	−0.13028	−	None
Loan Term	−0.064447	−	None
Loan To Value	0.013945	+	None
MSE	0.4668		None

B.3.4 Netherlands

ridge regression. Netherlands			
Variable	Coefficient	Effect	Differences with PCR
Primary Income	−0.10851	−	None
Original Balance	−0.024294	−	None
Loan Term	−0.0089454	−	None
Loan To Value	0.21393	+	None
MSE	0.75798		None

LASSO. Netherlands			
Variable	Coefficient	Effect	Differences with PCR and ridge results
Primary Income	−0.10820	−	None
Original Balance	−0.023938	−	None
Loan Term	−0.0084896	−	None
Loan To Value	0.21347	+	
MSE	0.8474		None

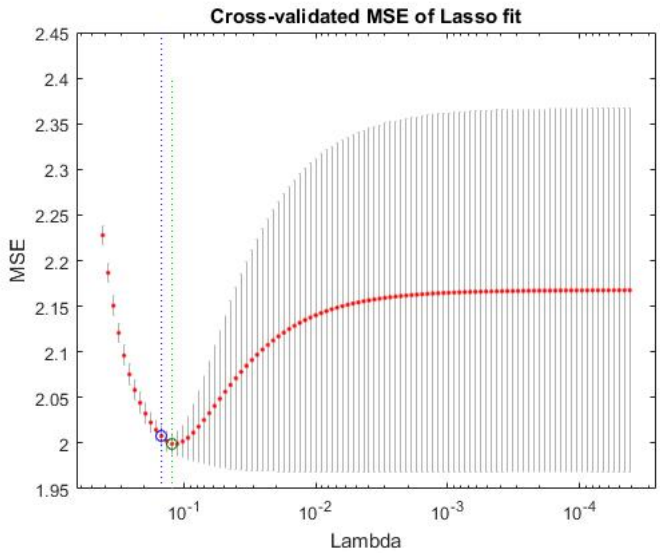

Figure 8: An illustration of the Cross-Validation method. The minimum value is highlighted with a green marker.

Chapter 6

Probabilistic Principal Component Regression

6.1 Introduction: How to deal with data sparsity

In this chapter we discuss a new possible approach which extends the previous analysis to the case of data vectors — which in our case are borrower's variables — characterized by a minor lack of entries. This problem is technically called data-sparsity.

In fact, the following discussion is strictly related both to the theoretical and the empirical topics discussed in the previous two chapters, and may show itself useful under the realistic hypothesis that the availability of quantitative RMBS loan data, collected on the European DataWarehouse, will increase over time, however conserving some data sparsity. Specifically, we notice two intrinsic characteristics of this database and others:

1. Some important variables may be added over time to those already accessible, or however the availability of data for the current variables is likely to be improved. Two reasons to believe this scenario to be true are both the increasing effort of the ED to improve data compilation made by banks, and the relative recent creation of this database.

2. However, an intrinsic problem of missing data must be acknowledged. Indeed,

some banks might continue to leave some values empty, some errors might occur during the data compilation or for other various reasons which can happen in the practice of data storing, the presence of empty entries should not be unexpected.

These considerations are supported by the features of the current RBMS datasets on ED, which present a huge volume of missing data especially for the optional variables. In fact, the present amount of missing data is currently too high to include in the previous analysis other variables, as for example the *"Guarantor Income"* or the *"Additional Collateral Value"*. For this reason, the most complete empirical analysis which we were able to propose, for each country and with the current dataset, has been done with the predictors employed in the previous two chapters. Here we present a specific solution to handle with missing data. In particular, we discuss mainly the theoretical approach to this problem, since the current data-sparsity is too high to allow an analysis similar to Chapters 4 and 5. It will be possible to take advantage empirically of this approach, applying it to each individual country, if the percentage of missing data decreases in the future.

6.2 The Probabilistic PCA model

In Chapter 4, the derivation of the principal component analysis was essentially based on algebraic calculations applied to some statistical concepts, such as the variance-covariance matrix or the linear independence. However, to introduce the Probabilistic Principal Component Analysis (PPCA) as an extension of the classical PCA it is helpful to discuss its corresponding geometrical derivation (as for example in Wang, 2011). The geometrical interpretation of the PCA is closer to the original Pearson's idea and highlights its feature of dimensionality reduction technique.

The first passage of the PCA derivation in Chapter 4 was the standardization and the centering of the independent variables. Regarding variables standardization, here we suppose that the data are either comparable or are analyzed after their standardization. As far as centering is concerned, it is allowed

in this context provided that this transformation does not change the distance among the original data. But this condition is actually verified, since centering is a translation, and in $\mathbb{R}^n$ any translation is an isometry: $T_{\bar{x}}$ is the translation operator

$$T_{\bar{x}} f(\boldsymbol{x}) = f(\boldsymbol{x} - \bar{x}). \tag{6.1}$$

Then, the second step is to derive the first principal component, maximizing under an arbitrary constraint the variance of the original dataset. Intuitively, this is equivalent to find a vector called *the first principal direction* such that the projection of the data onto its direction retains most of the original variation. Clearly, this direction is a 1-dimensional subspace of $\mathbb{R}^v$, v being the dimension of the row-space of the data matrix, using the same notation of the previous chapters. By definition, the first principal component is this vector, since the projections of each observation onto this 1-dimensional subspace form the score vector. Then, to find the other principal components it is sufficient to find successively the other principal directions, with the additional constraint of orthogonality among them. Analytically, let define the quantity δ as the Frobenius norm of the matrix $\boldsymbol{X}$:

$$\delta(\boldsymbol{X}) = \sum_{i=1}^{n} \|\boldsymbol{x_i^2}\| = \|\boldsymbol{X}\|_F. \tag{6.2}$$

Then, we need to find the direction which maximizes $\delta(\boldsymbol{X})$, which is the line S_1 associated with the first principal direction (or axis) $\boldsymbol{w_1}$. Since the data are centered, this direction passes through the origin. Ideally $\boldsymbol{w_1} \in \mathbb{S}^{v-1}$, which is a sphere, because it represents all possible directions in the space. Therefore

$$T_{\boldsymbol{w_1}} : \mathbb{R}^v \longrightarrow S_1 \tag{6.3}$$

$$\boldsymbol{w_1} = \underset{\boldsymbol{a} \in \mathbb{S}^{v-1}}{\arg\max}\, \delta(T_{\boldsymbol{a}}(\boldsymbol{X})). \tag{6.4}$$

The loading vector related to the first principal component becomes:

$$\boldsymbol{\alpha}_{n,1} = T_{\boldsymbol{w_1}}(\boldsymbol{X}) = [T_{\boldsymbol{w_1}}(\boldsymbol{x_1}), \cdots, T_{\boldsymbol{w_1}}(\boldsymbol{x_n})]' \tag{6.5}$$

or

$$\boldsymbol{\alpha}_{n,1} = [y_{1,1}, ..., y_{1,n}]'. \tag{6.6}$$

The remaining variables $\boldsymbol{w_j}$, $j = 2, 3, ..., v$, are the solutions of

$$\boldsymbol{w_j} = \underset{\boldsymbol{a} \in S_{v-1}^{\perp} \cap \mathbb{S}^{v-1}}{\arg\max} \ \delta(T_{\boldsymbol{a}}(\tilde{\boldsymbol{X}}_{v-1})). \tag{6.7}$$

where

$$\tilde{\boldsymbol{X}}_{v-1} = \left\{ \boldsymbol{x}_i - \sum_{j=1}^{v-1} \boldsymbol{w}_j y_{j,i}, 1 \leq i \leq n \right\} \tag{6.8}$$

because the maximization problem now takes into consideration the orthogonality constraint, and the procedure is iterative. As discussed above, this procedure implicates a loss of information, because it reduces the original data space. Specifically, the principal component analysis minimizes the error e (=the loss) deriving from the reconstruction of the original values through the orthogonal linear projections found with the procedure just discussed. Hence e is defined as:

$$e = \sum_{i=1}^{n} \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n\| \tag{6.9}$$

where

$$\hat{\boldsymbol{x}}_n = \boldsymbol{W}\boldsymbol{t}_n \tag{6.10}$$

$$\boldsymbol{t}_n = \boldsymbol{W'}\boldsymbol{x_n} \tag{6.11}$$

$$\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_v) \tag{6.12}$$

with the notations of Chapter 4.

However, one of the main features of the present discussion is the absence of a stochastic model associated to the independent variables, which is surely a simplification. Starting from this consideration, Tipping and Bishop (1999) propose an extension of the classical PCA to include a probabilistic model, whence the name Probabilistic Principal Component Analysis. In the next section we discuss how this model can be useful in the presence of missing data, while in this section we present their original derivation of the PPCA, and its main properties. The starting point is to consider a subject related to the PCA, the factor analysis. In this last model, an original v-dimensional vector of variables $\boldsymbol{x}$ is expressed as a linear function of a q-dimensional vector $\boldsymbol{t}$:

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{t} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{6.13}$$

where $\boldsymbol{\mu}$ is a vector of constants and $\boldsymbol{W}$ is a $v \times q$ matrix. When $q < v$ the dimensional reduction effect is evident. The error term $\boldsymbol{\epsilon}$, is usually assumed to be $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$. If at the same time it is assumed $\boldsymbol{t} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, then $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{WW'} + \boldsymbol{\Psi})$. The parameters $\boldsymbol{W}$ and $\boldsymbol{\Psi}$ have to be estimated via maximum likelihood, in particular the space spanned by the column vectors of $\boldsymbol{W}$ typically do not correspond to the principal component subspace. Moreover, considering the usual hypothesis assumed in literature for the residual variances (*isotropic error model*),

$$\forall i, \psi_i \in \boldsymbol{\Psi} : \psi_i = \sigma^2 \tag{6.14}$$

Tipping and Bishop (1999) criticize the usual related hypothesis of a known variance σ^2. For this reason, starting from the works by Lawley (1953) and Anderson and Rubin (1956) on the connection between the PCA and the factor analysis, they extend these results considering the case where σ^2 is estimated from observed data and proving that the maximum likelihood estimators $\boldsymbol{W}_{ML}$ and σ^2_{ML} are linked to the PCA. If

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \tag{6.15}$$

then

$$\boldsymbol{x}|\boldsymbol{t} \sim \mathcal{N}(\boldsymbol{Wt} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \tag{6.16}$$

and

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I} + \boldsymbol{WW'}). \tag{6.17}$$

The related log-likelihood, derived through standard calculations, is

$$\mathcal{L} = -\frac{N}{2}\left\{v \ln(2\pi) + \ln|\boldsymbol{C}| + tr(\boldsymbol{C^{-1}S})\right\} \tag{6.18}$$

where N is the number of observations, and

$$\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x_n} - \boldsymbol{\mu})(\boldsymbol{x_n} - \boldsymbol{\mu})' \tag{6.19}$$

$$\boldsymbol{C} = \sigma^2 \boldsymbol{I} + \boldsymbol{WW'}. \tag{6.20}$$

Then, the authors prove that (6.18) is minimized setting

$$\boldsymbol{W}_{ML} = \boldsymbol{U}_q(\boldsymbol{\Lambda}_q - \sigma^2 \boldsymbol{I})^{0.5}\boldsymbol{R} \tag{6.21}$$

where $\boldsymbol{U}_q$ is the matrix whose columns are the principal eigenvectors of $\boldsymbol{S}$, $\boldsymbol{\Lambda}_q$ is the diagonal matrix of the related eigenvalues λ_j, $1 \leq j \leq q$, and $\boldsymbol{R}$ is a $[q \times q]$ orthogonal rotation matrix. The key result is that $\boldsymbol{W}_{ML}$ is equal to the matrix of the principal component coefficient, or at least converges when it is estimated through the EM algorithm - necessary when some data are missing, hence the model specified in (6.13) maps the latent space into the principal subspace. The ML estimator for σ^2 has an evident interpretation:

$$\sigma^2_{ML} = \frac{1}{v-q} \sum_{j=q+1}^{v} \lambda_j \tag{6.22}$$

is directly proportional to the variance lost applying the PCA. Given the model just discussed, in the following section we examine how the Expectation-Maximization (EM) algorithm can be usefully applied to the PPCA in the presence of missing values.

6.3 PPCA for sparse matrices: The EM algorithm with missing data

As it was briefly discussed in Section 6.1, the problem of missing data, or equivalently of sparse data matrices, is very relevant not only to our analysis but definitely to a lot of other empirical analyses. Candès (2014) defines the sparsity as:

Definition 9 (Sparsity)**.** *We shall say that a signal $x \in \mathbb{C}^n$ is sparse, when most of the entries of x vanish. Formally, we shall say that a signal is s-sparse if it has at most s nonzero entries. One can think of an s-sparse signal as having only s degrees of freedom.*

It is clear from this definition that the sparsity is associated with incomplete information (called also energy, signal, etc). To handle with this problem, the EM algorithm has proven useful. Since the EM algorithm can be applied in order to estimate $\boldsymbol{W}_{ML}$ and σ^2_{ML} in the PPCA, the link between the probabilistic principal component analysis and sparsity becomes clear. The EM algorithm (Dempster

et al., 1977) consists of two steps, namely the Expectation and the Maximization steps, and computes the maximum likelihood (ML) estimator of some parameters — in this case of $\boldsymbol{W}_{ML}$ and σ^2_{ML} — when some variables are unobserved — in this case the latent variables $\boldsymbol{t}_n$. In particular, in order to estimate the ML parameters starting from an initial guess, these two steps find iteratively:

1. (*E-step*) the *expected value* of the log-likelihood function for the conditional distribution of the hidden variables, given the observed data and the current parameter estimates,
2. (*M-step*) the *ML parameters*, given the distribution assumed in the E-step.

The following estimator derivation is given again by Tipping and Bishop (1999), in Section 3 and Appendix B. Starting from the log-likelihood for the complete data:

$$\mathcal{L}_C = \sum_{n=1}^{N} ln\{p(\boldsymbol{x}_n, \boldsymbol{t}_n)\} \tag{6.23}$$

where, considering the distributions for $\boldsymbol{x}_n$ and $\boldsymbol{t}_n$ defined in the previous section,

$$p(\boldsymbol{x}_n, \boldsymbol{t}_n) = (2\pi\sigma^2)^{-v/2} \exp\left(-\frac{\|\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{t}_n - \boldsymbol{\mu}\|}{2\sigma^2}\right) (2\pi)^{-q/2} \exp\left(-\frac{\|\boldsymbol{t}_n\|}{2}\right) \tag{6.24}$$

and where the complete data include both the observations $\boldsymbol{x}_n$ and the latent variables $\boldsymbol{t}_n$, the EM algorithm computes:

1. (*E-step*) The expected value of $\mathcal{L}_C$ for the conditional distribution $p(\boldsymbol{t}_n|\boldsymbol{x}_n, \boldsymbol{W}, \sigma^2)$.
2. (*M-step*) The maximum value of the result obtained in the *E-step* with respect to $\boldsymbol{W}$ and σ^2.

The final results are:

$$\boldsymbol{W}_{ML} = \left\{\sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})\langle \boldsymbol{t}_n \rangle'\right\} \left(\sum_{n=1}^{N} \langle \boldsymbol{t}_n \boldsymbol{t'}_n \rangle\right)^{-1} \tag{6.25}$$

$$\sigma^2_{ML} = \frac{1}{Nv} \sum_{n=1}^{N} \{\|\boldsymbol{x}_n - \boldsymbol{\mu}\|^2 - 2\langle \boldsymbol{t}_n \rangle' \boldsymbol{W'}_{ML}(\boldsymbol{x}_n - \boldsymbol{\mu}) + tr(\langle \boldsymbol{t}_n \boldsymbol{t'}_n \rangle \boldsymbol{W'}_{ML} \boldsymbol{W}_{ML})\} \tag{6.26}$$

where the terms

$$\langle \boldsymbol{t}_n \rangle = \boldsymbol{M}^{-1} \boldsymbol{W'}(\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{6.27}$$

$$\langle \boldsymbol{t}_n \boldsymbol{t'}_n \rangle = \sigma^2 \boldsymbol{M}^{-1} + \langle \boldsymbol{t}_n \rangle \langle \boldsymbol{t}_n \rangle' \tag{6.28}$$

$$\boldsymbol{M} = \boldsymbol{W'W} + \sigma^2 \boldsymbol{I} \tag{6.29}$$

are incorporated in the *E-step* and are included in the iterative procedure. In fact, this algorithm is applied iteratively until convergence, determined through a tolerance level, is obtained. In presence of missing data, as discussed extensively in Ilin and Raiko (2010), this algorithm can be modified as follows. Firstly, (6.13) must be rewritten element-wise as:

$$\boldsymbol{x}_{ij} = \boldsymbol{w'}_i \boldsymbol{t}_j + \mu_i + \epsilon_{ij}, \forall i, j \in O \tag{6.30}$$

where O stands for "Observed Data". Then, contrary to the case of fully-observed data, the EM algorithm is applied not to the matrix $\boldsymbol{T}$ but to each vector $\boldsymbol{t}_j$, and $\boldsymbol{\mu}$ must be included in the computation at each step. These differences are a consequence of the inclusion of the missing observations together with the latent variables to construct the original complete values.

It is important to point out how these considerations hold only if data are *missing at random*, which means that their missingness occurs at random and does not depend on the unobserved data or on any latent variable characteristic. The missingness which takes place in ED may present these characteristics, because it seems not dependent on any specific value or feature of borrowers' variables.

6.4 Probabilistic Principal Component Regression: A new possible approach

In this section we present a new possible approach which combines the linear regression technique with the PPCA for missing data. In particular, we take advantage of the traditional PCR derivation, where the score vectors are the dependent variables, but this time using the scores computed through the PPCA in the presence of missing data. This extension must be done carefully, because all the conditions of the original principal component regression must be respected to derive a similar model. For example, we show the necessity to orthogonalize the

column vectors of $\boldsymbol{W}_{ML}$ derived through the EM algorithm in order to apply the PCR model and to interpret the results in terms of the original variables.

Therefore, after having briefly recalled the PCR derivation, we verify step by step the similarities and the differences which arises when the PPCA scores are used as independent variables for the regression. Our purpose is to verify that the traditional PCR model does not prevent the use of the score vectors derived through the model discussed in this chapter, or in other terms that this approach is theoretically correct. The main advantage of this method, which motivates its derivation, is the possibility to include in the PCR the variables which presents missing data too.

Consider the traditional linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6.31}$$

along with the assumptions mentioned in Section 4.3. If we assume that the values of each variable have the same distribution, we can still center and standardize the data using the mean and the variance, computed for each variable, using only the observed values. Indeed, with the previous assumption, we do not make any mistake in probability. Let us denote with the underscore ν the variables which refer to the PPCA; then $\boldsymbol{Z}_\nu$ is the score matrix derived through the PPCA in the presence of missing data. In the classic PCR, we wrote the linear model using $\boldsymbol{Z}$ as the matrix of dependent variables:

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\varphi} + \boldsymbol{\varepsilon} \tag{6.32}$$

The critical point is: does the relation derived in the classic PCR

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{I}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}'\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{\varphi} \tag{6.33}$$

with

$$\boldsymbol{\varphi} := \boldsymbol{A}'\boldsymbol{\beta} \tag{6.34}$$

still hold? At first sight, the answer is negative because $\boldsymbol{A}$ is assumed to be orthogonal. In fact, Tipping and Bishop (1999) discuss how at convergence the

columns of $\boldsymbol{W}_{ML}$ spans the principal subspace, but they may not be orthonormal because the relation

$$\boldsymbol{W'}_{ML}\boldsymbol{W}_{ML} = \boldsymbol{R'}(\boldsymbol{\Lambda}_q - \sigma^2\boldsymbol{I})\boldsymbol{R} \tag{6.35}$$

derived directly from (6.21) and where $\boldsymbol{R}$ is an arbitrary rotation matrix, is not diagonal when $\boldsymbol{R} \neq \boldsymbol{I}$.

But this problem is easily solved by orthogonalization of $\boldsymbol{W}_{ML}$. This is actually possible through, for example, the application of the singular value decomposition to find an orthonormal basis for the matrix. We recall that linear independence is not sufficient for the orthogonality condition, even if it is necessary.

Therefore, the orthonormality for the coefficient vectors of the matrix $\boldsymbol{A}_\nu$ is recovered. This allows to replicate the passages of Chapter 4 and in particular the following equations are valid for the Probabilistic Principal Component Regression too:

$$\boldsymbol{X\beta} = \boldsymbol{XI\beta} = \boldsymbol{XA}_\nu\boldsymbol{A'}_\nu\boldsymbol{\beta} = \boldsymbol{Z}_\nu\boldsymbol{\varphi}_\nu \tag{6.36}$$

$$\boldsymbol{\varphi_\nu} := \boldsymbol{A'}_\nu\boldsymbol{\beta} \tag{6.37}$$

$$\boldsymbol{Z}_\nu = \boldsymbol{XA}_\nu \tag{6.38}$$

Due to the orthonormality of $\boldsymbol{A}_\nu$, $\hat{\boldsymbol{\beta}}_\nu$ can be still computed, using the OLS estimator $\hat{\boldsymbol{\varphi}}_\nu$, as $\hat{\boldsymbol{\beta}}_\nu = \boldsymbol{A}_\nu\hat{\boldsymbol{\varphi}}_\nu$. Clearly, being the PPCA score different from those computed through the classic PCA, the final $\hat{\boldsymbol{\beta}}$ will be different.

Finally, in the practice of the classic PCR a subset of vectors of the score matrix is usually retained, which lead us to write the following definition for the *Probabilistic Principal Component Regression (PPCR)*:

Definition 10 (Probabilistic Principal Component Regression)**.** *Using the notation discussed above, we define the Probabilistic Principal Component Regression as:*

$$\boldsymbol{y} = \boldsymbol{Z}_\nu^S\boldsymbol{\varphi}_\nu^S + \tilde{\boldsymbol{\varepsilon}} \tag{6.39}$$

for $\boldsymbol{Z}_\nu^S \subset \boldsymbol{Z}_\nu$.

This definition is a modification of the traditional regression model $\boldsymbol{y} = \boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ in the following sense. Firstly, $\tilde{\boldsymbol{\varepsilon}}$ denotes the traditional error term for this particular

model (as in the case of the PCR, Chapter 4). But the relevant difference concerns the independent variables, which in the PPCR model are the score vectors derived through the PPCA.

This is the most important feature of the Probabilistic Principal Component Regression, because formulating this model using as regressors the score vectors computed through the PPCA — instead of the original independent variables, or of the scores provided by the traditional PCA — provides a solution to the practical problem of making a supervised analysis when some data are missing, as in the case of our dataset. Essentially, Definition 10 extends the unsupervised PPCA analysis to a regression model, preserving the interpretational properties of the classic PCR, but allowing the matrix of independent variables to be sparse. As we discussed previously in this chapter, this extension is not straightforward and combines different statistical, algebraic and numerical tools. Indeed, the PPCA technique, the EM algorithm, an orthogonalization of the matrix $\boldsymbol{W}_{ML}$ and the results of a linear regression, along with the possibility to compute their outputs, are required in order to apply the PPCR.

As we highlight in the final chapter, we leave the practical implementation of the PPCR to future research, not necessary on loan datasets; however, if the application to the European DataWarehouse datasets becomes possible due to a reduction of data sparsity, a more comprehensive empirical analysis on sparse loan variables may be supported by the PPCR.

Chapter 7

Conclusions

In this thesis, we have analyzed the open problem regarding the effect of collateral on the loan interest rates. In particular, we have studied this topic both theoretically and empirically through a quantitative approach, in order to give a new contribution to the existing literature.

In the first part of this work, we have proved the theoretical impossibility to support an unidirectional effect, constructing different game-theoretic models.

In particular, under symmetric information, we have shown that some fundamental financial assumptions lead to a bijection between risk and return when the totality of loan contracts — rather than an individual loan — are considered. This has allowed us to prove Theorem 1, which affirms that the interest rate should be decreased by a higher collateral value.

However, asymmetric information is characteristic of loan contracts, hence we have adapted the principal-agent model to the specific case of loan contracts, in order to consider, for example, the default probability or the value of an investment made with the borrowed capital. We have mathematically proved how collateral is an effective tool which mitigates adverse selection and moral hazard problems. These asymmetries change the previous result, because the effect of collateral now depends on which hypothesis is assumed.

In the second part, we have provided a rather extensive discussion of the mathematical and statistical theory behind the Principal Component Analysis,

the Principal Component Regression, the LASSO and the Ridge regularized regressions. We have applied these methods to loan big-data collected by the ED. The purpose of this second part was to prove that the theoretical results obtained in the first part of this work were supported by real loan contracts, as we have indeed found out.

The main advantages of our approach to the empirical analysis are at least four, as discussed in Chapter 4.

First of all, the possibility of combining an unsupervised analysis with a supervised one. In fact, we have included other borrower's variables along with the value of the collateral pledged, and their reciprocal effects have been analyzed with the PCA method. Secondly, it was convenient to study our data using some approaches — different from the simpler ones used in the previous empirical studies — which fit better to the huge volume of data analyzed in this thesis. Indeed, these models provide a computational advantage compared with the more traditional regression methods. For example, the OLS estimator requires a huge amount of memory to be computed, which is a problem when applied to big-data. Thirdly, the PCR, the LASSO and the Ridge lie within a more general theory which considers the class of biased estimators. This has allowed us to discuss their common features and to compare the estimates given by these different methods, both to assess if there was an unambiguous relationship, independent of the method chosen, and to choose the possible more convenient model in terms of a common defined statistical performance if necessary. Finally, the PCA analysis can be considered the starting point to examine the general problem of missing data, which is a feature of our loan data. In fact, this consideration led us to propose a new regression model based on the Probabilistic PCA (see Chapter 6), that may be applied to our loan datasets in the future, provided that the data sparsity decreases.

The huge collection of loan data available on the ED datasets has played an essential role. As far as we know this is the first study which examines this data through the methods discussed above, and definitely it is one of the few

works present in literature which uses millions of data to assess empirically the collateral-interest rates link, and which compares different countries using common variables of loan contracts.

The results derived in the empirical part supported our thesis on the impossibility to decide for a specific hypothesis among those presented in the game-theoretic part. Indeed, the different coefficients of the collateral variable associated to different countries suggest that the prevalent effect can be either moral hazard or adverse selection: they both take place in the real word. Moreover, some statistical values and a cluster analysis have also pointed out that the prevailing effect obtained from the data analysis reflects a general trend, but that this trend is not strong, which is another confirmation of its ambiguity.

In the final chapter, we have examined the Probabilistic Principal Component Analysis approach in the presence of missing data, given the large amount of sparse datasets which characterize the ED database. We have discussed how the EM algorithm, which can be applied to derive the score vectors through the PPCA, is an useful approach to handle with missing data, as highlighted in literature. Finally, we have checked whether the PPCA can be applied to a regression model, showing that this is feasible after some algebraic expedients, defining the Probabilistic PCR. In the context of this thesis, this approach may reveal itself useful if the availability of data will increase over time, however conserving some data sparsity. The more comprehensive empirical analysis resulting from this approach, if data will allow it reducing their sparsity, is left to future research.

The other future research themes related to this work are varied and numerous, given the various topics discussed throughout this thesis. Among them, an application of other models from game theory to analyze the link between collateral and interest rates from other point of views may be interesting, as well as an examination of a specific aspect of collateral in order to include further hypotheses — even if the theoretical literature has already discussed broadly this last point. As far as our empirical analysis is concerned, an extension could be done applying other unsupervised techniques to the loan data. Another possibility may be to

include some qualitative variables, providing additional insights to our analysis. Other future research topics related to this thesis concern more the theoretical part as, for example, the extension of the statistical methods presented or a further examination of their properties (however, these topics have been already discussed widely in literature) or the problem of sparsity in the supervised learning context.

Bibliography

Acharya, V. V. and Viswanathan, S. (2008), 'Moral hazard, collateral and liquidity'. Working Paper, New York University.

Agarwal, S. and Hauswald, R. (2010), 'Distance and private information in lending', *Review of Financial Studies* **23** (7), pp. 2757–2788.

Al-Hassan, Y. M. and Al-Kassab, M. M. (2009), 'A Monte Carlo Comparison between Ridge and Principal Components Regression Methods', *Applied Mathematical Sciences* **3** (42), pp. 2085–2098.

Al-Kandari, N. M. and Jolliffe, I. T. (2001), 'Variable selection and interpretation of covariance principal components', *Communications in Statistics-Simulation and Computation* **30** (2), pp. 339–354.

Anderson, T. W. and Rubin, H. (1956), 'Statistical inference in factor analysis', *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability* **5**, pp. 111–150.

Angbazo, L. A., Mei, J. and Saunders, A. (1998), 'Credit spreads in the market for highly leveraged transaction loans', *Journal of Banking and Finance* **22** (10), pp. 1249–1282.

Arthur, D. and Vassilvitskii, S. (2007), 'k-means++: The advantages of careful seeding', *in* 'Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms', Society for Industrial and Applied Mathematics, pp. 1027–1035.

Barro, R. J. (1976), 'The loan market, collateral, and rates of interest', *Journal of Money, Credit and Banking* **8** (4), pp. 439–456.

Beaton, A. E., Rubin, D. B. and Barone, J. L. (1976), 'The acceptability of regression solutions: Another look at computational accuracy', *Journal of the American Statistical Association* **71** (353), pp. 158–168.

Belsley, D. A., Kuh, E. and Welsch, R. E. (2005), *Regression diagnostics: Identifying influential data and sources of collinearity*, John Wiley & Sons.

Benjamin, D. K. (1978), 'The use of collateral to enforce debt contracts', *Economic Inquiry* **16** (3), pp. 333–359.

Berger, A. N., Espinosa-Vega, M. A., Frame, W. S. and Miller, N. H. (2011), 'Why do borrowers pledge collateral? New empirical evidence on the role of asymmetric information', *Journal of Financial Intermediation* **20** (1), pp. 55–70.

Berger, A. N., Frame, W. S., Ioannidou, V. et al. (2011), 'Reexamining the empirical relation between loan risk and collateral: The roles of collateral characteristics and types'. Federal Reserve Bank of Atlanta Working Paper Series.

Berger, A. N. and Udell, G. F. (1990), 'Collateral, loan quality and bank risk', *Journal of Monetary Economics* **25** (1), pp. 21–42.

Berk, K. N. (1984), 'Validating regression procedures with new data', *Technometrics* **26** (4), pp. 331–338.

Berndt, A. and Gupta, A. (2009), 'Moral hazard and adverse selection in the originate-to-distribute model of bank credit', *Journal of Monetary Economics* **56** (5), pp. 725–743.

Besanko, D. and Thakor, A. V. (1987), 'Collateral and rationing: Sorting equilibria in monopolistic and competitive credit markets', *International Economic Review* **28** (3), pp. 671–689.

Bester, H. (1985), 'Screening vs. rationing in credit markets with imperfect information', *The American Economic Review* **75** (4), pp. 850–855.

Bester, H. (1987), 'The role of collateral in credit markets with imperfect information', *European Economic Review* **31** (4), pp. 887–899.

Bester, H. (1994), 'The role of collateral in a model of debt renegotiation', *Journal of Money, Credit and Banking* **26** (1), pp. 72–86.

Bibby, J. (1980), 'Some effects of rounding optimal estimates', *Sankhyā: The Indian Journal of Statistics. Series B* **42** (3/4), pp. 165–178.

Bieta, V., Broll, U. and Siebe, W. (2008), 'The banking firm: The role of signaling with collaterals'. Dresden Discussion Paper in Economics.

Blackith, R. E. and Reyment, R. A. (1971), *Multivariate morphometrics*, Academic Press.

Blazy, R. and Weill, L. (2006), 'Le rôle des garanties dans les prêts des banques françaises', *Revue d'Économie Politique* **116** (4), pp. 501–522.

Booth, A., Thakor, A. V. and Udell, G. F. (1991), 'Secured lending and default risk: Equilibrium analysis, policy implications and empirical results', *The Economic Journal* **101** (406), pp. 458–472.

Booth, J. R. and Booth, L. C. (2006), 'Loan collateral decisions and corporate borrowing costs', *Journal of Money, Credit and Banking* **38** (1), pp. 67–90.

Breit, E. and Arano, K. (2008), 'Determinants of Loan Interest rates: Evidence from the Survey of Small business finances (SSBF)'. SSRN Electronic Journal.

Brick, I. E. and Palia, D. (2007), 'Evidence of jointness in the terms of relationship lending', *Journal of Financial Intermediation* **16** (3), pp. 452–476.

Broll, U. and Gilroy, M. B. (1986), 'Collateral in banking policy and adverse selection', *The Manchester School* **54** (4), pp. 357–366.

Calcagnini, G., Farabullini, F., Giombini, G. et al. (2009), 'Loans, interest rates and guarantees: Is there a link?'. Working Paper, Università degli Studi di Urbino.

Campbell, D. E. (2006), *Incentives: Motivation and the economics of information*, Cambridge University Press, 2nd Edition.

Candès, E. J. (2014), 'Mathematics of sparsity (and a few other things)'. Proceedings of the International Congress of Mathematicians, Seoul, South Korea.

Candilera, M. and Bertapelle, A. (2011), *Algebra Lineare e Primi Elementi di Geometria*, McGrowHill.

Capra, C. M., Fernandez, M., Ramirez-Comeig, I. et al. (2005), 'Moral hazard and collateral as screening device: Empirical and experimental evidence'. Working Paper, Emory University.

Chan, Y.-S. and Kanatas, G. (1985), 'Asymmetric valuations and the role of collateral in loan agreements', *Journal of Money, Credit and Banking* **17** (1), pp. 84–95.

Chan, Y.-S. and Thakor, A. V. (1987), 'Collateral and competitive equilibria with moral hazard and private information', *The Journal of Finance* **42** (2), pp. 345–363.

Chen, Y. (2006), 'Collateral, loan guarantees, and the lenders' incentives to resolve financial distress', *The Quarterly Review of Economics and Finance* **46** (1), pp. 1–15.

Claus, I. (2011), 'The effects of asymmetric information between borrowers and lenders in an open economy', *Journal of International Money and Finance* **30** (5), pp. 796–816.

Coco, G. (1999), 'Collateral, heterogeneity in risk attitude and the credit market equilibrium', *European Economic Review* **43** (3), pp. 559–574.

Collins, A. J., Harrison, D. M. and Seiler, M. J. (2015), 'Mortgage Modification and the decision to strategically default: A game theoretic approach', *Journal of Real Estate Research* **37** (3), pp. 439–470.

Csóka, P., Havran, D. and Szűcs, N. (2015), 'Corporate financing under moral hazard and the default risk of buyers', *Central European Journal of Operations Research* **23** (4), pp. 763–778.

De Meza, D. and Southey, C. (1996), 'The borrower's curse: Optimism, finance and entrepreneurship', *The Economic Journal* **106** (435), pp. 375–386.

De Meza, D. and Webb, D. C. (1987), 'Too much investment: A problem of asymmetric information', *The Quarterly Journal of Economics* **102** (2), pp. 281–292.

Degryse, H. and Van Cayseele, P. (2000), 'Relationship lending within a bank-based system: Evidence from European small business data', *Journal of Financial Intermediation* **9** (1), pp. 90–109.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society. Series B* **39** (1), pp. 1–38.

Diamantaras, K. I. and Kung, S. Y. (1996), *Principal Component Neural Networks: Theory and applications*, John Wiley & Sons.

Diana, G. and Tommasi, C. (2002), 'Cross-validation methods in principal component analysis: A comparison', *Statistical Methods and Applications* **11** (1), pp. 71–82.

Eastment, H. and Krzanowski, W. (1982), 'Cross-validatory choice of the number of components from a principal component analysis', *Technometrics* **24** (1), pp. 73–77.

Elad, M. (2010), *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, 1st Edition.

Elsas, R., Krahnen, J. P. et al. (2000), 'Collateral, default risk, and relationship lending: An empirical study on financial contracting'. Working Paper, Center for Financial Studies.

Ferré, L. (1995), 'Selection of components in principal component analysis: A comparison of methods', *Computational Statistics and Data Analysis* **19** (6), pp. 669–682.

Flatnes, J. E. and Carter, M. R. (2016), 'A little skin in the microfinance game: Reducing moral hazard in joint liability group lending through a mandatory collateral requirement'. Working Paper, Agricultural and Applied Economics Association.

Foucart, T. (2000), 'A decision rule for discarding principal components in regression', *Journal of Statistical Planning and Inference* **89** (1), pp. 187–195.

Gromov, M. (August 1998), 'Report of the Senior Assessment Panel of the International Assessment of the U.S. Mathematical Sciences, adapted as: Possible trends in mathematics in the coming decades', *Notices of the American Mathematical Society* **45** (7), pp. 846–847.

Gunst, R. F. and Mason, R. L. (1980), *Regression analysis and its application: A data-oriented approach*, CRC Press.

Harhoff, D. and Körting, T. (1998), 'Lending relationships in Germany–Empirical evidence from survey data', *Journal of Banking and Finance* **22** (10), pp. 1317–1353.

Hawkins, D. M. (1973), 'On the investigation of alternative regressions by principal component analysis', *Applied Statistics* **22** (3), pp. 275–286.

Hawkins, D. M. and Eplett, W. (1982), 'The Cholesky factorization of the inverse correlation or covariance matrix in multiple regression', *Technometrics* **24** (3), pp. 191–198.

Hill, R., Fomby, T. B. and Johnson, S. (1977), 'Component selection norms for principal components regression', *Communications in Statistics-Theory and Methods* **6** (4), pp. 309–334.

Hocking, R. R. (1976), 'A biometrics invited paper. The analysis and selection of variables in linear regression', *Biometrics* **32** (1), pp. 1–49.

Hocking, R. R., Speed, F. and Lynn, M. (1976), 'A class of biased estimators in linear regression', *Technometrics* **18** (4), pp. 425–437.

Hoerl, A. E. (1962), 'Application of ridge analysis to regression problems', *Chemical Engineering Progress* **58** (3), pp. 54–59.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12** (1), pp. 55–67.

Hoerl, R. W., Schuenemeyer, J. H. and Hoer1, A. E. (1986), 'A simulation of biased estimation and subset selection regression techniques', *Technometrics* **28** (4), pp. 369–380.

Horn, R. A. and Johnson, C. R. (2012), *Matrix Analysis*, Cambridge university press, 2nd Edition.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology* **24** (6), pp. 417–441 and 498–520.

Hsuan, F. C. (1981), 'Ridge regression from principal component point of view', *Communications in Statistics-Theory and Methods* **10** (19), pp. 1981–1995.

Ilin, A. and Raiko, T. (2010), 'Practical approaches to principal component analysis in the presence of missing values', *Journal of Machine Learning Research* **11**, pp. 1957–2000.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An introduction to statistical learning*, Springer.

Janda, K. et al. (2003), 'Credit guarantees in a credit market with adverse selection', *Prague Economic Papers* **12** (4), pp. 331–349.

Jiménez, G. and Saurina, J. (2004), 'Collateral, type of lender and relationship banking as determinants of credit risk', *Journal of Banking and Finance* **28** (9), pp. 2191–2212.

John, K., Lynch, A. W. and Puri, M. (2003), 'Credit ratings, collateral, and loan characteristics: Implications for yield', *The Journal of Business* **76** (3), pp. 371–409.

Jolliffe, I. (2002), *Principal component analysis*, Springer, 2nd Edition.

Jolliffe, I. T. (1982), 'A note on the use of principal components in regression', *Applied Statistics* **31** (3), pp. 300–303.

Kolmogorov, A. N. and Fomin, S. V. (1957), *Elements of the theory of functions and functional analysis. Vol. 1, Metric and normed spaces*, Graylock Press.

Kramer, O. (2013), *Dimensionality reduction with unsupervised nearest neighbors*, Springer.

Krzanowski, W. (1987), 'Cross-validation in principal component analysis', *Biometrics* **43** (3), pp. 575–584.

Lawley, D. (1953), 'A modified method of estimation in factor analysis and some large sample results', *Uppsala Symposium on Psychological Factor Analysis. Nordisk Psykologi's Monograph Series* **3**, pp. 35–42.

Leeth, J. D. and Scott, J. A. (1989), 'The incidence of secured debt: Evidence from the small business community', *Journal of Financial and Quantitative Analysis* **24** (3), pp. 379–394.

Lehmann, E. and Neuberger, D. (2001), 'Do lending relationships matter? Evidence from bank survey data in Germany', *Journal of Economic Behavior and Organization* **45** (4), pp. 339–359.

Lindley, D. V. and Smith, A. F. (1972), 'Bayes estimates for the linear model', *Journal of the Royal Statistical Society. Series B* **34** (1), pp. 1–41.

Lott, W. F. (1973), 'The optimal set of principal component restrictions on a least-squares regression', *Communications in Statistics-Theory and Methods* **2** (5), pp. 449–464.

Machauer, A. and Weber, M. (1998), 'Bank behavior based on internal credit ratings of borrowers', *Journal of Banking and Finance* **22** (10), pp. 1355–1383.

Macho-Stadler, I. and Pérez-Castrillo, J. D. (2001), *An introduction to the economics of information: Incentives and contracts*, Oxford University Press, 2nd Edition.

Mandel, J. (1972), 'Principal components, analysis of variance and data structure', *Statistica Neerlandica* **26** (3), pp. 119–129.

Manove, M., Padilla, A. J. and Pagano, M. (2001), 'Collateral versus project screening: A model of lazy banks', *The RAND Journal of Economics* **32** (4), pp. 726–744.

Mansfield, E. R., Webster, J. T. and Gunst, R. F. (1977), 'An analytic variable selection technique for principal component regression', *Applied Statistics* **26** (1), pp. 34–40.

Mason, R. L. and Gunst, R. F. (1985), 'Selecting principal components in regression', *Statistics and Probability Letters* **3** (6), pp. 299–301.

Menkhoff, L., Neuberger, D. and Suwanaporn, C. (2006), 'Collateral-based lending in emerging markets: Evidence from Thailand', *Journal of Banking and Finance* **30** (1), pp. 1–21.

Meredith, W. and Millsap, R. E. (1985), 'On component analyses', *Psychometrika* **50** (4), pp. 495–507.

Mertens, B., Fearn, T. and Thompson, M. (1995), 'The efficient cross-validation of principal components applied to principal component regression', *Statistics and Computing* **5** (3), pp. 227–235.

Mosteller, F. and Tukey, J. W. (1977), *Data analysis and regression: A second course in statistics*, Prentice Hall.

Myerson, R. B. (1979), 'Incentive compatibility and the bargaining problem', *Econometrica: Journal of the Econometric Society* **47** (1), pp. 61–74.

Nair, M. T. (2009), *Linear operator equations: Approximation and regularization*, World Scientific.

Oman, S. D. (1991), 'Random calibration with many measurements: An application of stein estimation', *Technometrics* **33** (2), pp. 187–195.

Ono, A., Sakai, K. and Uesugi, I. (2012), 'The effects of collateral on firm performance', *Journal of The Japanese and International Economies* **26** (1), pp. 84–109.

Pearson, K. (1901), 'On lines and planes of closest fit to systems of points in space.', *Philosophical Magazine* **2** (11), pp. 559–572.

Peydró, J.-L. (2013), 'Credit cycles and systemic risk', *Els Opuscles del CREI* (35).

Pozzolo, A. F. (2002), 'Secured lending and borrowers' riskiness'. Working Paper, Banca d'Italia, Research Department.

Rajan, R. and Winton, A. (1995), 'Covenants and collateral as incentives to monitor', *The Journal of Finance* **50** (4), pp. 1113–1146.

Rao, C. (1964), 'The use and interpretation of principal component analysis in applied research', *Sankhya A* **26** (4), pp. 329–358.

Roman, S. (2005), *Advanced linear algebra*, Springer, 3rd Edition.

Rosser Jr., J. B. (2003), 'A Nobel prize for asymmetric information: The economic contributions of George Akerlof, Michael Spence and Joseph Stiglitz', *Review of Political Economy* **15** (1), pp. 3–21.

Smith, C. W. and Warner, J. B. (1979), 'On financial contracting: An analysis of bond covenants', *Journal of Financial Economics* **7** (2), pp. 117–161.

Smith, G. and Campbell, F. (1980), 'A critique of some ridge regression methods', *Journal of the American Statistical Association* **75** (369), pp. 74–81.

Smith, L. I. (2002), 'A tutorial on principal components analysis'. Working Paper, Cornell University.

Stiglitz, J. E. and Weiss, A. (1981), 'Credit rationing in markets with imperfect information', *The American Economic Review* **71** (3), pp. 393–410.

Su, X. (2010), 'A Re-examination of Credit rationing in the Stiglitz and Weiss Model'. Technical report, Norwegian School of Economics.

Sugiyama, T. and Tong, H. (1976), 'On a statistic useful in dimensionality reduction in multivariable linear stochastic system', *Communications in Statistics-Theory and Methods* **5** (8), pp. 711–721.

Tibshirani, R. (1996), 'Regression Shrinkage and Selection via the Lasso', *Journal of the Royal Statistical Society. Series B* **58** (1), pp. 267–288.

Tipping, M. E. and Bishop, C. M. (1999), 'Probabilistic Principal Component Analysis', *Journal of the Royal Statistical Society. Series B* **61** (3), pp. 611–622.

Trenkler, D. and Trenkler, G. (1984), 'On the Euclidean distance between biased estimators', *Communications in Statistics-Theory and Methods* **13** (3), pp. 273–284.

Vinod, H. D. (1978), 'A survey of ridge regression and related techniques for improvements over ordinary least squares', *The Review of Economics and Statistics* **60** (1), pp. 121–131.

Von Storch, H. and Zwiers, F. W. (1999), *Statistical analysis in climate research*, Cambridge university press.

Wang, J. (2011), *Geometric structure of high-dimensional data and dimensionality reduction*, Springer.

Wang, Y.-L. (2010), 'Does collateral cause inefficient resource allocation?', *Journal of Economics and Business* **62** (3), pp. 220–233.

Webster, J. T., Gunst, R. F. and Mason, R. L. (1974), 'Latent root regression analysis', *Technometrics* **16** (4), pp. 513–522.

Weill, L. and Godlewski, C. J. (2006), 'Does collateral help mitigate adverse selection? A cross-country analysis'. MPRA Working Paper.

Wooldridge, J. M. (2015), *Introductory econometrics: A modern approach*, Nelson Education.