Università della Svizzera Italiana

Faculty of Economics

Statistical Analysis for Credit Risk Modelling

Paolo Tenconi

Submitted for the degree of Ph.D. in Economics

December 2008

Advisor :	Antonietta Mira	-	University of Lugano
Examinators :	Giovanni Barone Adesi	-	University of Lugano
	Sonia Petrone	-	Bocconi University, Milan

A Rita, “Con tutto l’amore che posso...”

Ringraziamenti

Prima di tutto vorrei ringraziare la Professoressa Antonietta Mira, quale direttore di tesi, per la sua guida e il suo supporto durante tutti questi anni e a cui devo l'introduzione ai metodi Markov chain Monte Carlo.

Inoltre desidero ringraziare il Professor Giovanni Barone Adesi, direttore dell'Istituto di Finanza e docente ai corsi di dottorato, la cui opera consente, ad oggi, presso l'*Università della Svizzera Italiana*, di beneficiare di una didattica e di una ricerca di livello internazionale. Ringrazio quindi l'*Università della Svizzera Italiana* e la *Global Asset Management* per avermi consentito di godere di tale privilegio. Ringrazio la Professoressa Sonia Petrone per aver accettato di essere membro, insieme al Professor Giovanni Barone Adesi, della commissione di tesi.

Uno speciale ringraziamento alle tante persone incontrate all'Università dell'Insubria, in particolare il Professor Paolo Paruolo e i vari colleghi con cui ho potuto condividere tante idee e da cui ho potuto trarre nuovi stimoli, fra questi Christine Choirat e Fabrizio Leisen.

Da ultimo il ringraziamento più importante, ai miei genitori e in particolare a mia moglie Rita, cui questo lavoro è dedicato.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Credit risk analysis has a prominent role in the banking and financial industry, where one of the main businesses is granting credit to people and companies. The importance of credit risk evaluation was recognized since the 1988 Basel Capital Accord, which, ignoring other kind of risks, set a minimal capital requirement for banks, based on their credit exposure. Assets of banks were classified and grouped in five categories according to credit risk, then banks were required to hold capital equal to 8% of their risk-weighted assets, this is called the *standardized approach* to credit risk evaluation.

In 1996 the Basel Committee introduced a variation to the original approach, taking into account also the *market risk*, including the trading portfolio among the risky assets: shares, bonds and derivatives. The novelty the commitee introduced, was the possibility to adopt an internal method to measure the market risk, through *value at risk* methodologies. The credit risk, instead, persisted to be treated using the standardized approach.

The new Basel Capital Accord of 2004, as well as recognizing the presence of *operational risk*, introduced, concerning credit risk quantification, the possibility to create an internal rating based system (*IRB approach*), allowing bank to choose between the *foundation-IRB* by which they are required to compute only the default probability of their counterparts and the *advanced-IRB* where they need to compute also the *exposure at default*, the *loss given default* and other parameters. The *IRB approach* requires therefore an internal rating system that, after being validated, would give banks the advantage to lower the capital requirement through a good credit policy.

In the last years even companies outside the banking industry started managing credit risk by adopting frameworks similar to those required by the Basel Commitee, part of this thesis regards an experience of this kind, highlighting its worthiness. This work focuses on estimating the default probability for small and medium companies, which market data usually doesn't exist and therefore cannot be used as an indicator of perceived credit risk level, while balance sheet data is available. These enterprises are the common interlocutor for banking institutions and services providers. Dealing with large firms, banks stipulate ad hoc agreements, after a deep credit risk analysis conducted directly by analysts, while for small medium enterprises it is necessary to have tools that can automatically determine the level of credit risk. This is relevant even for the high number of cases to be processed in a short time period, however the results obtained should always be considered as a support to the analyst assessment.

1.1.1 Scope and Contributions of the Research

The main purpose of the thesis is the forecast of default event for small and medium size companies. By trying to achieve this aim, a series of issues arose, which were taken into account and analyzed during this work. The main topics we dealt with are

- Variable selection,
- Rare events,
- Variance reduction in Markov chain Monte Carlo,
- Influential observations.

The selection of variables is an issue as the number of balance sheet items is high, the *rare events* problem arose while trying to compose a flexible statistical model, able to produce sector specific parameters. This required to switch to the Bayesian paradigm, as estimates are not biased in case of rare events as pointed out in King and Zeng (2001), moreover the predictive distribution can be easily obtained. In fact, in order to deal completely with default risk, credit risk frameworks should take into account also the uncertainty of the estimated risk. The need to estimate complex bayesian models requires the use of methods to integrate numerically over high dimensional spaces, to this aim, we used *Markov chain Monte Carlo* techniques. These methods come with convergence issues and results embed an error that should be reduced as much as possible. So, by using the *delayed rejection* strategy developed in Tierney and Mira (1999), an improvement over the well known *Metropolis-Hastings* algorithm, which is usually the first choice, we tried to overcome these problems. A second way to reduce the variance in Markov chain Monte Carlo we investigated, is based on the extension to statistical models of the *zero variance principle* introduced by Assaraf and Caffarel (1999) in the physics literature. The last issue we faced is the role played by influential observations on parameter estimates. This is a significant problem with default data, response variables pose no problems as are naturally bounded, while explanatory data may present aberrant observations and ruin results. Therefore we addressed this kind of problem by working on the *weighted likelihood score equations* methodology as proposed in Markatou et al. (1997).

1.2 Thesis Organisation

Chapter 2: Credit Risk in the Energy Market

This chapter is aimed at creating a statistical model for predicting the default events for a company providing energy services, exposed to credit toward a substantial number of small and medium enterprises. For these companies balance sheets are available. Much of the work in this part of the thesis is primarily aimed at data, because the anticipation of default should be linked to variables not readily available. We derive a large number of balance sheet ratios and select the best subset in term of prediction performance. The identified model, once found the consensus of financial analysts, was subsequently implemented. After being applied to new customers, once new data became available, its out of sample performance was later investigated. The results confirmed forecast effectiveness and temporal stability, we conclude that the model can be used as an effective internal rating system.

Chapter 3: Hierarchical Bayesian Modelling of Credit Risk

In this chapter we start using the Bayesian paradigm, analyzing default data provided by a leading italian bank, regarding small-medium companies asking for a loan. In this data set the default event is much rare, therefore maximum likelihood estimates are biased, while Bayesian estimates are not, as shown in King and Zeng (2001). Financial analysts are convinced about the difference of baseline default risk among sectors, so we chose to model this heterogenity treating hierarchically a subset of parameters. This choice was reasonable, allowing flexible sectorial estimates, while avoiding unstable results due to a further increase of default rareness in each sector. Parameter estimates were obtained through the *delayed rejection* algorithm developed in Tierney and Mira (1999), whose performace was compared to the *Metropolis-Hastings* algorithm (Hastings (1970)) in term of autocorrelation of generated chains. This chapter has been published with some minor changes in Mira and Tenconi (2004).

Chapter 4: Zero Variance Markov Chain Monte Carlo

The Markov chain Monte Carlo efficiency issue, partially discussed in the previous chapter, is the central theme here. The goal is the variance reduction through the *zero variance principle*, introduced in the physics literature in Assaraf and Caffarel (1999). In this chapter we focused on the adaptation of this principle to the statistical estimation problem. Starting from simple cases where it was possible to obtain analytical solutions with the variance reduced to zero, we moved gradually to more complex problems. In the end we tested the zero variance principle on the estimate of the hierarchical model proposed in the previous chapter, where we obtained a $78,95\%$ average variance reduction. Some considerations regarding statistical models and their properties, allowed the developement of solutions that make the methodology more agile, avoiding the calibration process.

Chapter 5: Weighted Likelihood Equations for Resistant Credit Risk Modelling

Often internally handled balance sheet data, present aberrant observations. These sometimes are purely recording errors, more commonly observations are correct but still extreme. It is necessary to handle this matter, otherwise estimates may be incorrect as stated in the robustness literature. With such data the risk is increased by the practice, in default prediction, to select all defaulted companies while subsetting the non defaulted ones in a random way, this is done to reduce the huge amount of data to be treated. In this chapter we adopt the approach of *weighted likelihood equations* (Markatou et al. (1997)), which represents a way to deal with extreme data in addiction to the min-max (Huber (1981)) and infinitesimal approach, see Hampel et al. (1986). The literature on weighted likelihood equations has not yet reached maturity, it started dealing with discrete data models and later was extended to some continuous models, see Basu and Lindsay (1994), Agostinelli (1997), Markatou et al. (1998). The idea we propose in this chapter, is to extend the weighted likelihood equation methodology to the class of generalized linear models in unitary fashion. We derive a modified version of *iterative reweighted least squares* algorithm, while regarding the logistic regression, widely used in this thesis, we derive analytically the *Newton-Raphson* equations. Results, though at an early stage, appear to be encouraging, the weighting scheme adopted underweights only leverage points that are not coherent with the theoretical model, while coherent leverage points are not underweighted.

References

C. Agostinelli. A one-step robust estimator based on the weighted likelihood methodology. Technical report, 16, Dipartimento di Scienze Statistiche, Universita' di Padova, 1997.

R. Assaraf and M. Caffarel. Zero-Variance principle for Monte Carlo algorithms. *Physical Review letters*, 83, 23:4682–4685, 1999.

A. Basu and B. Lindsay. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of Institute of Statistical Mathematics*, 46:683–705, 1994.

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics: The Approach based on Influence Functions*. Wiley, 1986.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

P. J. Huber. *Robust Statistics*. Wiley, 1981.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9:2:137–163, 2001.

M. Markatou, A. Basu, and B. Lindsay. Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference*, 57:215–232, 1997.

M. Markatou, A. Basu, and B. Lindsay. Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93:740–750, 1998.

A. Mira and P. Tenconi. Bayesian estimate via credit risk via mcmc with delayed rejection. *Stochastic Analysis, Random Fields and Applications IV in Progress in Probability*, 26: 277–291, 2004.

L. Tierney and A. Mira. Some adaptive Monte Carlo methods for bayesian inference. *Statistics in Medicine*, 18:2507–2515, 1999.

Chapter 2

Credit Risk in the Energy Market

2.1 Introduction

The credit risk management has been growing in the past 20 years, mainly in the financial and banking industry. This is natural, given their core activity, aimed at connecting economic agents having a surplus of money with agents requiring some kind of lending.

In the past years, also institutions not belonging to the financial area, have been starting adopting credit risk frameworks similar to the ones developed in the banking industry. As an example among them, there are firms operating in the energy market, that became interested in statistical methods as a tool to process an extensive amount of units requiring energy services. The liberalization process in this market, started in Europe in the late Nineties, played surely an important role.

Regarding the energy market, another important reason why credit risk modelling is critical, is related to the nature of the provided services. In fact, if a default occurs, the content of the contract between the provider and the user is lost. This is not the case, for example, for companies operating in the automitive leasing area.

In this chapter, a model built for an energy market institution (*EI* hereafter), providing energy services to small and medium firms, is outlined. The purpose of the analysis is twofold:

- Investigate the usefullness of balance sheet ratios analysis, identifying relevant causal relationships with default;
- Predict the default event, providing an automatic rating system, intended as a primary tool for credit analysts.

The model has been conceived after some brainstormings with internal financial analysts.

2.2 The Data

The typical agreement between energy providers and their customers, is to settle any credit after a given period of energy supply, whose length is usually about three or four months. As pointed out previously, this is the reason why energy suppliers front a credit risk problem, especially because the supplied good cannot be rescued as it is destroyed during its use.

The data that will be used in the following analysis, is composed by 1564 companies to whom the EI provided, in the past, energy services. For each of them the whole balance sheet was available[1], together with their behaviour regarding payments, expressed in term of time lasted since the agreed settlement date. This set of data reduced to 1067 after removing missing and illogical values.

2.2.1 Response Variable

Regarding the credit status of these firms, the rule adopted by *EI* financial analysts is to assign the default status to companies which have a payment delay higher than a given treshold, usually of 60 days. In Table 2.1 is reported the frequency of default occurrencies.

[1] Balance sheet data was obtained from Cerved, http://www.cerved.com

	Abs. Freq.	Rel. Freq.
Default	150	9.59 %
Non-Default	1414	90.41 %

Table 2.1: Default status

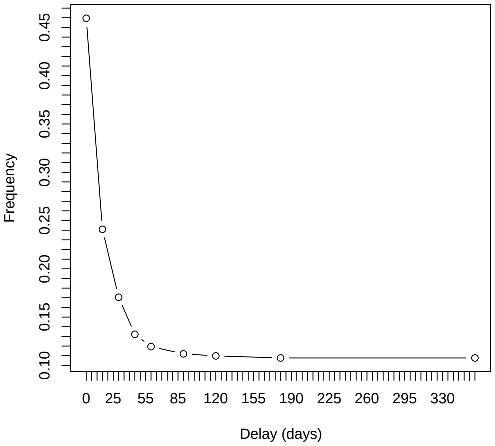

Figure 2.1: Payment delay

The choice of these threshold is not so arbitrary, as analysts' experience shows that after a given delaying period, firms almost surely will never pay their debt. In Figure 2.1 it is shown clearly this concept, in fact after 60 days there is almost no reduction of the number of firms paying their debt to the EI.

2.2.2 Explanatory variables

The analysis is developed by using some extra balance sheet data, such as the *seniority* of the firm, expressed as years since its foundation, the geographical location and a set of balance sheet ratios. The idea is that the default can be predicted (see Altman (1968)) and that forecasts can be made for companies going to become *EI* customers. We build a large set of balance sheet ratios, aimed at capturing the whole aspects of company management.

Many of them have a similar meaning, so care is needed to avoid the selection of collinear variables. We divide the generated ratios into five categories (income, equity, development, liquidity and working capital, financial structure), although some of them could be placed in more than one category. In our analysis, the average time lag between default and the issue of balance sheet, is of one and half year, therefore we expect that ratios with low variability over time are more relevant to predict the default status.

2.2.2.1 Income

The capability of the company to generate an income is a necessary, yet not sufficient, condition to persist in the non default status. What we want to capture, first of all, is the ability of the firm to sell its services, this is obtained by analyzing the *turnover* ratio. Then we are interested on the return generated by the company in term of operational and non operational activity. To this aim we consider the following indicators:

$$\text{ROE} = \frac{\text{Free Tax Profit -Taxes}}{\text{Equity}}$$

$$\text{ROA} = \frac{\text{Free Tax Profit -Financial Result-Extraordinary Result}}{\text{Total Assets}}$$

$$\text{ROI} = \frac{\text{EBIT}}{\text{Net Capital Employed}}$$

$$\text{Turnover} = \frac{\text{Net Sales}}{\text{Total Assets}}$$

$$\text{I2} = \frac{\text{Profit Before Extraordinary Item}}{\text{Total Assets}}$$

$$\text{I5} = \frac{\text{EBIT}}{\text{Interest and Similar Income} \setminus \text{Interest and Similar Expenses}}$$

$$\text{I7} = \frac{\text{EBIT}}{\text{Equity +Financial Long Term Debts}}$$

2.2.2.2 Equity

In this category we put ratios that measure the amount of firm activities covered by internal resources. These ratios are inversely related to some financial structure ratios. The *total assets* variable is inserted with the aim to capture any link between the dimension of firms and their default probability. We expect that companies with a low presence of debts have a low default probability in the medium term, so these variables should be relevant with a medium/long time lag as in our setting.

$\text{Equity Ratio 1} = \text{ER}_1 = \dfrac{\text{Equity}}{\text{Long Term Assets}}$

$\text{Equity Ratio 2} = \text{ER}_2 = \dfrac{\text{Equity}}{\text{Tangible Assets}}$

$\text{Leverage} = \dfrac{\text{Total Assets}}{\text{Equity}}$

$\text{Capitalization Ratio} = \text{CR} = \dfrac{\text{Equity}}{\text{Total Financial Debts}}$

$\text{I3} = \dfrac{\text{Equity}}{\text{Total Liabilities and Debts}}$

I6=Total Assets

2.2.2.3 Development

In the attempt to capture the dynamics of the management we compute the variation, over two years, of the firm's net income, as a synthesis of its whole performance. We build the following indicator

Development=Net Income t[0]-Net Income t[-2]

where $t[0]$ is the income available from the last balance sheet, while $t[-2]$ is the income recorded two years before the last balance sheet available.

2.2.2.4 Liquidity and Working Capital

The analysis of short term firm's balance sheet items, taken alone, is relevant if the time lag between the analysis and the default status is very short. However, we have a time lag of one and half year on average, so we expect these variables to be relevant only if analyzed jointly with other long term structural variables (equity and financial structure). This is why we also considered the following ratios:

$\text{Instant Liquidity} = \dfrac{\text{Cash and Cash Equivalents}}{\text{Short Term Total Debts}}$

$\text{Current Ratio} = \text{CR} = \frac{\text{Current Assets}}{\text{Short Term Debts}}$

$\text{Working Capital Coverage} = \frac{\text{Short Term Financial Debts}}{\text{Working Capital}}$

$\text{Current Assets Intensity} = \frac{\text{Current Assets}}{\text{Net Sales}}$

$\text{Quick Ratio 1} = \text{QR}_1 = \frac{\text{Current Assets-Inventory}}{\text{Equity+Long Term Debt}}$

$\text{Quick Ratio 2} = \text{QR}_2 = \frac{\text{Current Assets-Inventory}}{\text{Total Assets}}$

$\text{Cash Asset Ratio} = \text{CAR} = \frac{\text{Liquid Assets}}{\text{Short Term Debts}}$

$\text{I1} = \frac{\text{Net Working Capital}}{\text{Total Assets}}$

$\text{Net Working Capital 1} = \text{NWC}_1 = \frac{\text{Current Assets}}{\text{Equity+Long Term Debts}}$

$\text{Net Working Capital 2} = \text{NWC}_2 = \frac{\text{Current Assets-Short Term Debts}}{\text{Total Assets}}$

$\text{Net Working Capital 3} = \text{NWC}_3 = \frac{\text{Current Assets}}{\text{Total Assets}}$

$\text{I4} = \frac{\text{Net Working Capital}}{\text{Total Operating Revenue } \backslash \text{Total Assets}}$

$\text{Working Capital} = \text{WC} = \frac{\text{Current Assets-Liquid Assets}}{\text{Short Term Debts-Short Term Financial Debts}}$

2.2.2.5 Financial Structure

The following variables measure the impact of debts, some of them are inversely related to the equity ratios, while the remaining capture the impact of short term debts.

$\text{Short Term Debts Intensity} = \frac{\text{Short Term Debts}}{\text{Net Sales}}$

$\text{Short Term Debts Impact} = \frac{\text{Short Term Financial Debts}}{\text{Total Financial Debts}}$

$\text{Long Term Assets Coverage Ratio} = \frac{\text{Equity+Long Term Financial Debts}}{\text{Long Term Assets}}$

$\text{I8} = \frac{\text{Total Assets}}{\text{Equity}}$

2.3 The Model

The main purpose behind the statistical modelling of credit risk data, is the selection of financial ratios and extra balance sheet data, able to predict the default event. The fitting capability of the statistical model is not the unique criterion taken into account, as a first requirement there is the interpretability of results, discarding the so called *black box* models, as it is not possibile to interpret the role played by the selected variables and their obedience to the economic and financial rules. The second criterion choosen is to build a pasimonious model, but covering, as widely as possibile, the balance sheet of the firm, that is to say the following areas:

- Financial structure;
- Income capability;
- Equity equilibrium;
- Liquidity.

The last criterion is to discard variables that lead to results conflicting with the economic rationality. Given that, the starting point is the unconditional default probability (Table 2.1).

A very naive method to assess default, is to associate to each firm the historical frequency of default as the future probability to manifest a default status. This approach is reasonable if there is no data describing the status of each debtor, which happens usually when the credit exposure is very low and there is no convenience to collect explanatory data. In our case this corresponds to the *null model*, i.e. a model with intercept only.

The statistical model should be able, for healthy firms, to associate a default prediction lower than the historical average, while increasing the default prediction for the non healthy ones. This corresponds to a reduction of deviance of the fitted model, relative to the null one, as will be described later.

As the default event is dichotomous, we model the *response,* y_i, as a Bernoulli random variable

$$y_i \sim Be\left(\theta_i\right) \;\;,\;\; i = 1, \ldots, n \tag{2.1}$$

where the default probability, $\theta_{i,}$ is modelled as

$$\theta_i = \frac{\exp\left(\mathbf{x}'_i\beta\right)}{1+\exp\left(\mathbf{x}_i'\beta\right)}, \tag{2.2}$$

while $\mathbf{x}_i$ is the data of the firm and β is the set of parameters to be estimated. What we get is a *logistic regression* model. We chose this kind of model as results are interpretable, moreover the estimation is quick, this is relevant as we are investigating a large amount of explanatory variables. The loglikelihood for the model at hand is

$$\ell\left(\beta\right) = \sum_{i=1}^{n}\left\{y_i \ln\left(\frac{\exp\left(\mathbf{x}'_i\beta\right)}{1+\exp\left(\mathbf{x}_i'\beta\right)}\right) + \left(1-y_i\right)\ln\left(1-\frac{\exp\left(\mathbf{x}'_i\beta\right)}{1+\exp\left(\mathbf{x}_i'\beta\right)}\right)\right\}. \tag{2.3}$$

The selection of variables is guided by the reduction of the *Akaike Information Criterion*, which, for a model with k explanatory variables, is given by $\text{AIC} = 2k - 2\ln\left(\ell\left(\beta\right)\right).$ As an approximate test we use the reduction of *deviance* ((Azzalini, 2004)), we can rewrite the loglikelihood in the *exponential form*, as the Bernoulli belongs to the exponential family

$$\ell\left(\beta\right) = \sum_{i=1}^{n}\left\{\frac{\omega_i}{\psi}\left[y_i\eta_i - b\left(\eta_i\right)\right] + c_i\left(y_i,\psi\right)\right\}$$

for the case at hand we have $\psi = 1$, $\omega_i = 1$, $\eta_i = \ln\frac{\theta_i}{1-\theta_i} = \mathbf{x}_i'\beta$, $b\left(\eta_i\right) = \ln\left(1+e^{\eta_i}\right)$ and $c_i\left(y_i,\psi\right) = 0$. Given that, $E\left[Y\right] = \mu = b'\left(\eta\right),$ the deviance can be expressed as

$$D\left(y,\hat{\mu}\right) = -2\sum_{i=1}^{n}\left\{\omega_i\left[\left(y_i\hat{\eta}_i - b\left(\hat{\eta}_i\right)\right) - \left(y_i\tilde{\eta}_i - b\left(\tilde{\eta}_i\right)\right)\right]\right\}$$

where $\tilde{\eta}_i$ is obtained by estimating the *full model* (the one having as many parameters as the number of observations), while $\hat{\eta}_i$ is the estimated model. We call $\frac{D}{\psi}$ the *normalized deviance* and for two nested models $M_2 \subset M_1$, we have that, approximately, $\frac{D\left(y,\hat{\mu}_{p_2}\right)-D\left(y,\hat{\mu}_{p_1}\right)}{\psi} \xrightarrow{d} \chi^2_{p_1-p_2}$, where p_1 and p_2 are the number of parameters.

During the model selection, care is devoted to avoid too much correlated explanatory variables, moreover the purpose is to build a model able to capture as much as possible all the relevant aspects pertaining to the enterprise management, as expressed in the balance sheet and related to the default event. In Table 2.2 it is summarized the identified model, obtained through a stepwise procedure using the Akaike Information Criterion.

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercept	-2.4693	0.1387	-17.80	0.0000
I3	-0.9183	0.4105	-2.24	0.0253
I7	-0.1587	0.0835	-1.90	0.0574
ROA	-0.2672	0.1141	-2.34	0.0192
Turnover	-0.4270	0.1411	-3.03	0.0025
Current Ratio	-0.4719	0.2991	-1.58	0.1146
Seniority	-0.1788	0.1136	-1.57	0.1155

Table 2.2: Identified model

At a first stage, analysts decided to keep included in the model the *seniority* variable, despite the fact it is weakly significant from a statistical point of view. This decision was made, while wating for additional data confirming this choice (later analysis rejected this decision, see Table 2.8). We decided to mantain *current ratio* given its importance regarding the liquidity aspect of the firm. Testing the selected model against the *null* model, we have that for the null model the deviance is 708.6 and 659.6 for the model estimated. The p-value of the test statistic, approximatively distributed as χ^2_{7-1}, is nearly zero, therefore we have a significant reduction of deviance. In the following table it is shown the correlation matrix of selected regressors

	I3	I7	ROA	Turnover	Current Ratio	Seniority
I3	1.00	−0.03	0.01	−0.16	0.45	0.02
I7	−0.03	1.00	−0.39	−0.00	−0.03	−0.02
ROA	0.01	−0.39	1.00	0.14	−0.00	0.05
Turnover	−0.16	−0.00	0.14	1.00	−0.05	−0.08
Current Ratio	0.45	−0.03	−0.00	−0.05	1.00	0.13
Seniority	0.02	−0.02	0.05	−0.08	0.13	1.00

Table 2.3: Correlation matrix

Table 2.3 shows that there are no collinearity problems. The selected explanatory variables cover widely the various firm activities, we can divide them into five categories, to get a better understanding of results. The identified categories are:

1. **Equity equilibrium**

   - I3: Represents company capitalization.

2. **Income equilibrium**

   - ROA: Measures company's profitability on the equity and the financial debts.
   - I7: Measures how effectively the operating result (EBIT) of a company yields profit to the equity and the financial debts.

3. **Liquidity and working capital**

   - Current ratio: measures the short term liquidity.

4. **Commercial equilibrium**

   - Turnover: represents the firm ability to transform its invested capital into income.

5. **Extra balance sheet data**

   - Seniority: years lasted since company's foundation.

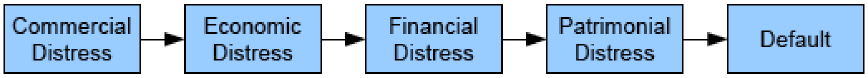

Figure 2.2: Distress chain

We have obtained reasonable results under an economical point of view: companies having a good commercial efficiency, able to generate profits, with a good short term liquidity and a sufficient degree of equity have less chance to get defaulted. The analysis conducted is static, but we argue there are sequential steps, leading to the default status (Figure 2.2): at a first stage, there is a *commercial* distress, here the company isn't able to sell its services, this causes an *economic* damage; after that, given it is unable to produce internal resources, there is an increase of *debts*, the last stage is reached when the *equity* is compromised.

2.4 Model Evaluation

In graph 2.3 the histogram of default probabilities generated by the model is shown, the dashed line represents the historical average default frequency. The dispersion over the historical average is a first insight about the capability of the model to separate companies in subgroups with distinct risk.

2.4.1 Discriminative Power

At a first stage, it is recommended to evaluate the discriminating power of the model, also if it will be used only as a ranking instrument. In graph 2.4 two curves, representing correctly classified defaults and non-defaults is shown, by varying the default probability treshold beyond which customers are refused.

To get a better explanation of the discriminative power of the model, we use the *ROC curve*. Defining

- *sensibility* the proportion of predicted positives with respect to all the true positives

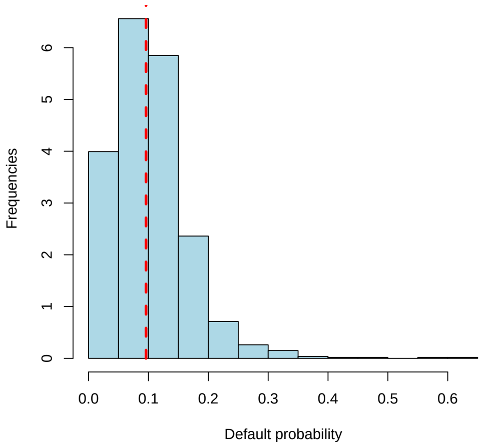

Figure 2.3: Generated default

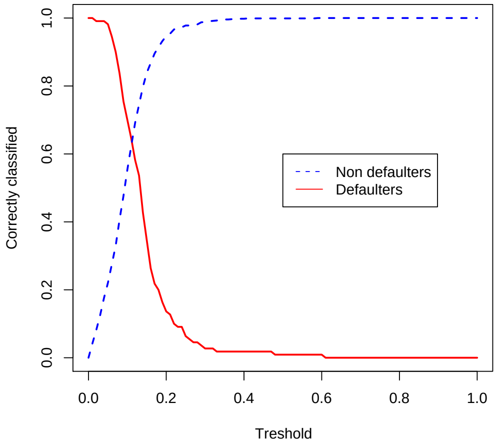

Figure 2.4: Treshold curve

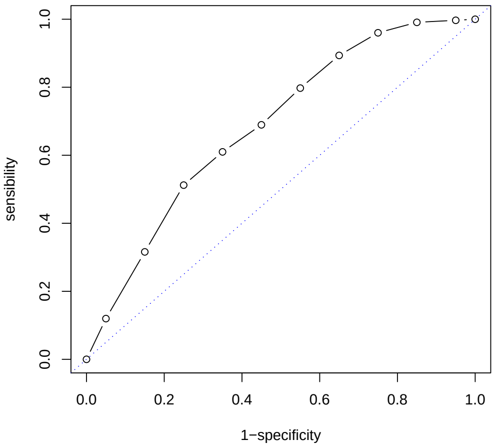

Figure 2.5: ROC curve

	Predicted Default	Predicted Regular
Actual Default	2.73%	7.63%
Actual Regular	11.96%	77.68%

Table 2.4: Confusion Matrix

- *specificity* the proportion of predicted negatives with respect to all the true negatives,

the ROC curve (Figure 2.5) is given by the points on a plane whose coordinates are $(1 - specificity, sensibility)$, obtained for the set of all possibile values of the treshold. A random classification of companies would produce a bisector line, so the area between this line and the curve generated by the model gives a visual insight of the predictive power of the model.

As an example, if we choose a 16% treshold we obtain two *confusion matrices*, the first one (Table 2.4) shows that we have a 19.59% probability to predict a wrong status. It is also worth investigating the conditional error, reported in Table 2.5, where it is shown that if we condition the analysis to the creditworth firms, we have a 13.34% chance to get wrong predictions, while this probability goes up to 73.64% for bad companies.

Using the treshold in the example, we may notice that, with respect to the *naive forecast*,

	Predicted Default	Predicted Regular
Actual Default	26.36%	73.64%
Actual Regular	13.34%	86.66%

Table 2.5: Conditional confusion matrix

	Predicted Default	Predicted Regular
Actual Default	0	-(1-r)
Actual Regular	-i	i

Table 2.6: Loss function

having nearly the same probability to correctly detect a creditworth company, the model probability to detect defaults is more than triple.

2.4.2 Choice of the Treshold

In case of a very large number of customers, where only a subset of them can be accepted, there is a true need to find a good treshold. This is not the case when the commercial policy leads to accept most of customers, however it is usually required a warranty whose amount is related to the degree of estimated default risk. Here we elaborate an economical criterion to drive the choice of the treshold. Depending on the commercial and risk policy it is possibile to build a *loss function*, which can be minimized by choosing the optimal classification treshold.

Such a *loss function* can be constructed in a way that the economic damage in case of default is incorporated, together with the failure income in case of default prediction for healthy companies. As an example if we define

- r : recovery rate;
- i : profit rate;

we can build Table 2.6, describing the *loss function* whose content is given by

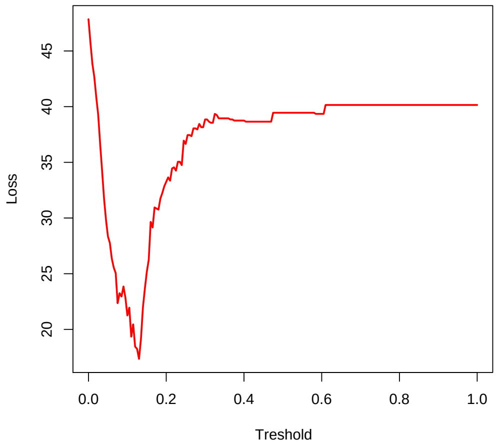

Figure 2.6: Loss curve

- i : net profit on a regular transaction;
- $-(1-r)$: the loss of capital invested, minus the expected recovery rate in case of a non detected default;
- $-i$: the failure income given that a healthy company is rejected.

In Figure 2.6 , it is exposed the loss curve for $i = 5\%,\ r = 20\%$.

The optimal treshold is the one minimizing the loss function. In this case the minimum in achieved in 0.13. This approach can be extended including other factors, such as the size of the transaction.

2.4.3 Internal Rating

Beside the use of the statistical model as a discrimination tool, to accept or refuse companies, there is a second way to use it. This second kind of use is aimed at creating internal rating classes, so companies are usually accepted, but, given their rating, a warranty is required, related to the estimated degree of risk.

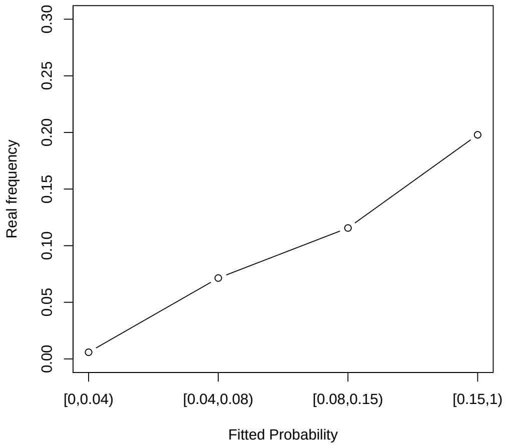

Figure 2.7: Rating classes, predicted and observed defaults

For this purpose we create four rating classes, say *low/medium/high/very high* risk and we put in each firms depending on the predicted default probability, as generated by the model. In Figure 2.7 these rating classes are shown together with their associated *realized default frequencies*. In the first class there is a realized default frequency of 0.5%, in the second one we have a 7.14% frequency, in the third class it jumps to 11.56%, while in the last class it is about 19.80%. So we can conclude that, at least relative to the investigated sample, the model is a reliable tool to create an internal rating system. We will see in the following, the predictive behaviour of the model as discriminative and rating engine.

2.5 Out of Sample Prediction

The following data has been collected after the estimation of the statistical model previously exposed. To the acquired new companies was associated a default probability, as obtained from the statistical model, later, it was registred their behaviour in terms of delay of payment (i.e. the default status).

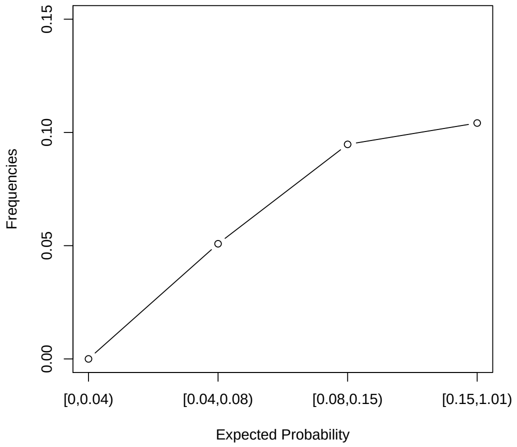

Figure 2.8: Rating classes, forecasted and observed defaults

2.5.1 Internal Rating

In Figure 2.8 it is shown the performance of the model as a rating tool. On the abscissa there are four risk classes, while on the ordinate axis is represented the actual frequency of default within each class. At a first sight, it is clear that the higher the estimated riskness, the higher is the actual default frequency. In particular

- *very low risk*: here we put all firms with estimated default probability in the range $[0; 4\%)$, for this class no defaults were registered;
- *medium-low risk*: here are stored companies with a $[4\%; 8\%)$ estimated default probability, so against an expected 6% default probability a realized 5% default probability has been registered;
- *medium-high risk*: given an expected 11.5% default probability, in the $[8\%; 15\%)$ class, a 10% in term of default frequency was observed;
- *high risk*: in this class were put all firms with a default rate equal or higher than 15%, this is the only class with an expected default probability not close to the observed

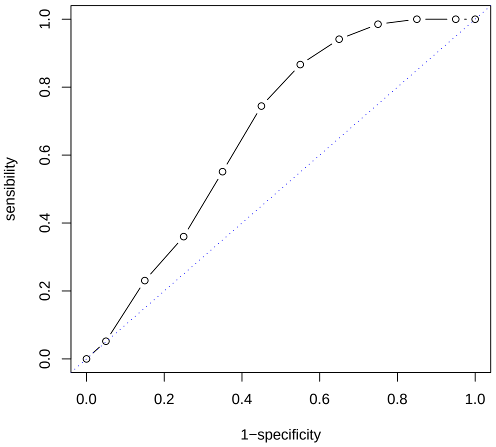

Figure 2.9: ROC curve

	Predicted Default	Predicted Regular
Actual Default	17.65%	82.35%
Actual Regular	14.72%	85.28%

Table 2.7: Confusion matrix: conditional distributions

default frequency. However it should be noticed that a few companies were put into this rating class.

2.5.2 Discriminative Power

As in the previous section, to illustrate the discrimitative power of the model, we construct the *ROC curve*. Graph 2.9 confirms the out of sample predictive power of the statistical model, with respect to a random selection of default status. Choosing again a 16% treshold we obtain the conditional distribution confusion matrix represented in Table 2.7.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.6687	0.3044	−2.20	0.0280
I3	−0.5789	0.2384	−2.43	0.0152
I7	−0.0511	0.0231	−2.21	0.0269
ROA	−3.0401	1.0225	−2.97	0.0029
Turnover	−0.5499	0.1805	−3.05	0.0023
Current Ratio	−0.3774	0.2118	−1.78	0.0748
Seniority	−0.0063	0.0067	−0.93	0.3530

Table 2.8: Updated model

2.6 Model Updating

After having acquired new companies, the set of data was enlarged and the statistical model was estimated again (Figure 2.8), in order to get additional support of its validity. We expect variables to remain significant from a statistical point of view, this is the case and, in addition, the decision to keep in the model *current ratio* has been rewarded, in fact now there is a almost fully significant power of the t-test statistic. On the other hand the behavioural variable, *seniority*, has definitively lost its statistical meaning, so it has to be removed for any subsequent default modelling.

References

A. Alberici. *L'Analisi di Bilancio per Fidi Bancari.* Franco Angeli, 2008.

E. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23, 4:589–609, 1968.

A. Azzalini. *Inferenza Statistica, Una Presentazione Basata sul Concetto di Verosimiglianza.* Springer-Verlag Italia, 2004.

L. A. Bernstein. *Financial Statement Analysis.* IRWIN, 1993.

S. Castelli and S. Gatti. *Il Corporate Lending.* Bancaria Editrice, 2003.

G. De Laurentis, F. Saita, and A. Sironi, editors. *Rating Interni e Controllo del Rischio di Credito.* Bancaria Editrice, 2004.

L. Fahrmeir and G. Tutz. *Multivariate statistical modelling based on generalized linear models.* Springer, 2001.

B. Fornari. *Gli Indici Aziendali.* Franco Angeli, 2007.

F. Metelli. *Basilea2 Che cosa cambia.* Il Sole 24 ORE, 2005.

A. Montrone. *Il Sistema delle Analisi di Bilancio per la Valutazione dell'Impresa.* Franco Angeli, 2005.

E. Paravani. *Analisi Finanziaria.* McGraw-Hill, 2002.

A. Resti, editor. *Misurare e Gestire il Rischio di Credito nelle Banche.* Alpha Test, 2001.

M. R. Tyran. *Gli Indici Aziendali.* Il Sole 24 ORE, 2004.

AA. VV. *Professione Credit Manager.* IPSOA, 2003.

Chapter 3

Hierarchical Bayesian Modelling of Credit Risk

3.1 Introduction

The aim of this chapter is to estimate the default probability (DP) of companies that apply to banks for loan. The explanatory variables available to us are performance indicators derived from the balance sheet of each company and the knowledge of the macro-sector to which the company belongs. For privacy reasons we do not report how the 4 performance indicators are obtained and the 7 sectors identified. The data set (Banca Intesa, BCI) consists of 7513 companies of which 1.615 %. A more detailed description of the dataset appears in Table 3.1 where the unbalanced design is apparent.

The main issues related to DP prediction are: the events of interest are rare (thus bias and consistency problems arise); the different sectors might present similar behaviors relative to risk of defaulting; expert analysts have, typically, strong prior opinions on DP. The model we propose is a hierarchical Bayesian logistic regression, and introduces dependency among different sectors thus addressing efficiently all the above mentioned issues.

Table 3.1: Summary of the dataset.

	Dimension	% Default
Sector 1	63	0%
Sector 2	638	1.41%
Sector 3	1342	1.49%
Sector 4	1163	1.63%
Sector 5	1526	1.51%
Sector 6	315	9.52%
Sector 7	2466	0.93%

3.1.1 Rare Events

The logistic regression we are going to propose is based on the assumption that an observed dichotomous variable Y is modeled as

$$Y_i \sim \text{Bernoulli}\left(Y_i|\pi_i\right)$$

with

$$\pi_i = \frac{1}{1+\exp\left(-\mathbf{x}_i^{'}\beta\right)},$$

while β is an unknown set of parameters. An alternative way to look at this model is to imagine the existence of an unobserved latent variable Y^* distributed as a logistic distribution

$$Y^* \sim \text{Logistic}\left(Y^*|\mu_i = \mathbf{x}_i^{'}\beta\right)$$

where the observed Y variable is related to the unobserved one Y^* through the fact that $Y_i = 0$ if $Y^* \leq 0$, while $Y_i = 1$ if $Y^* > 0$. The model is the same as

$$\begin{aligned} p\left(Y_i = 1|\beta\right) &= \pi_i = p\left(Y^* > 0|\beta\right) \\ &= \int_0^\infty \text{Logistic}\left(Y^*|\mu_i\right) dY_i^* = \frac{1}{1+\exp\left(-\mathbf{x}_i^{'}\beta\right)}. \end{aligned}$$

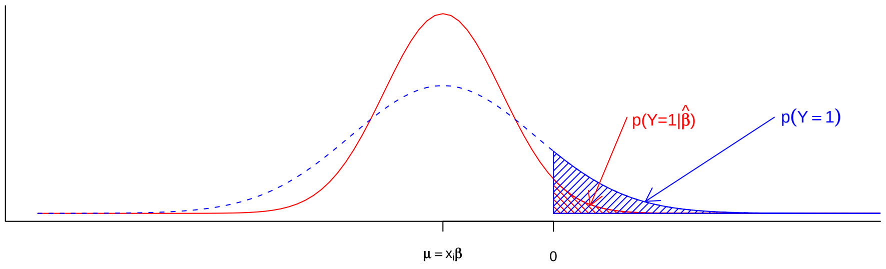

Figure 3.1: Logistic latent variable

The use of maximum likelihood probability estimates

$$p\left(Y_i = 1|\hat{\beta}\right) = \frac{1}{1 + \exp\left(-\mathbf{x}_i'\hat{\beta}\right)}, \tag{3.1}$$

may create biased, lower estimated, results for rare events. The reason is that we are ignoring parameter uncertainty, as pointed out in King and Zeng (2001). To avoid this we should instead integrate over the estimated parameter distribution

$$P\left(Y_i = 1\right) = \int P\left(Y_i|\beta\right) p\left(\beta\right) d\beta.$$

The effect caused by using maximum likelihood paramter $\hat{\beta}$, instead of averaging over the estimated parameter distribution can be observed in Figure 3.1.1.

To deal with this kind of matter we decided to adopt the bayesian paradigm, because parameters are treated as random variables, therefore it is possibile to integrate over their distribution once a sample is obtained through Markov chain Monte Carlo simulation methods. Moreover any random effects structure for parameters can be introduced and inference easly conducted. This is relevant given the different baseline risks present in each sector.

3.2 The Model

We use a logistic regression, that is we model the logit of the default probability, as a linear function of the explanatory variables. In the sequel we use the following notation, indicating vectors with underlined letters:

- n_j: number of companies belonging to sector j, $j = 1, \cdots, 7$;
- $y_{i,j}$: binary observation on company i ($i = 1, \cdots, n_j$), belonging to sector j. The value one indicates a default event;
- $\underline{x}_{i,j}$: 4×1 vector of explanatory variables (performance indicators) for company i belonging to sector j;
- $\underline{\alpha}$: 7×1 vector of intercepts, one for each sector;
- $\underline{\beta}$: 4×1 vector of slopes, one for each performance indicator.

The parameters of interest are $\underline{\alpha}$ and $\underline{\beta}$. We will, informally, indicate by y and x all the observations on the dependent and explanatory variables respectively.

Adopting a logistic regression model gives rise to the following likelihood:

$$L(\underline{\alpha}, \underline{\beta}; y, x) = \prod_j \prod_i \theta_{i,j}^{y_{i,j}} (1 - \theta_{i,j})^{1-y_{i,j}} \tag{3.2}$$

where

$$\theta_{i,j} = \frac{\exp(\alpha_j + \underline{x}'_{i,j}\ \underline{\beta})}{1 + \exp(\alpha_j + \underline{x}'_{i,j}\ \underline{\beta})}. \tag{3.3}$$

Following the Bayesian paradigm, prior distributions are assigned to the parameters of interest, in particular we take the prior on $\underline{\beta}$, $p(\underline{\beta})$, to be a four dimensional normal centered at zero ($\underline{\mu}_\beta = \underline{0}$) and with the identity matrix times 64 as the covariance matrix (Σ_β).

The intercepts, α_j, are assumed to have normal prior distributions, $p(\alpha_j|\mu_\alpha, \sigma_\alpha^2)$, independent only given the parameters μ_α and σ_α^2. The mean μ_α, is unknown with normal hyper prior, $p(\mu_\alpha)$, centered at zero and with variance equal to 64. The prior on the variance is a Gamma(a, b) distribution with mean equal to 5 and variance equal to 9.

The values of the known hyper parameters have been fixed so that the corresponding priors are fairly vague. Prior information on DP, elicited by expert analysts (not available to us), can be incorporated when assigning the values of these hyper parameters. Typically expert analysts express opinions on the DP, $\theta_{i,j}$, (rather than $\underline{\alpha}$ and $\underline{\beta}$) by assigning them a mean value and a level of confidence or a variance. Given these measures of location and spread a beta distribution is assumed on these probabilities and the values of $\underline{\alpha}$ and $\underline{\beta}$ matching the assigned prior distributions can be inferred using the inverse logit transformation.

The model implemented has been estimated both using informative and non-informative priors centered in zero with a very high variance (results reported). The evidence gained using fictitious informative priors suggests that, in our setting, the estimates are robust relative to the choice of the prior parameters due to rather large amount of data that causes the prevalence of the likelihood over prior influence in the posterior.

The distribution of interest, the posterior of the slopes, intercepts and hyper parameters, is proportional to

$$\pi(\underline{\alpha}, \underline{\beta}, \mu_\alpha, \sigma_\alpha|y, x) \propto L(\underline{\alpha}, \underline{\beta}; y, x) \prod_j p(\alpha_j|\mu_\alpha, \sigma_\alpha^2)\, p(\mu_\alpha)\, p(\sigma_\alpha)\, p(\underline{\beta}) \tag{3.4}$$

A graphical representation of the proposed model appears in Figure 3.2

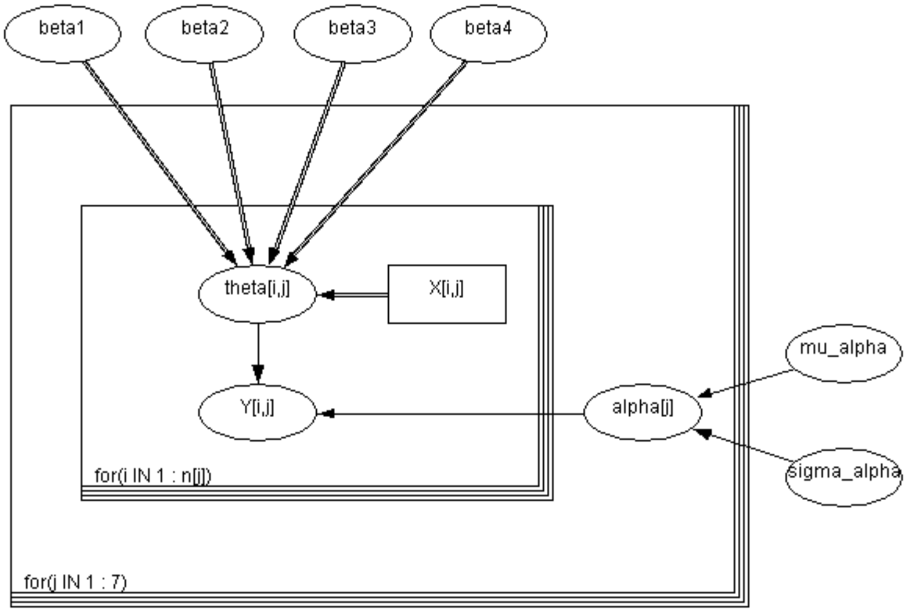

Figure 3.2: Graphical representation of the model

3.3 The Algorithm

We use a MCMC algorithm (Tierney (1994)) to simulate observations from (3.4), the 13-dimentional posterior distribution of interest. To improve the performance of the standard Metropolis-Hastings algorithm (MH) we adopt the delaying rejection (DR) strategy (Tierney and Mira (1999), Green and Mira (2001)) with a single delaying step. This means that, upon rejection of a proposed candidate move, instead of advancing the simulation time and retaining the same position (as in a standard MH sampler), a second stage candidate is proposed and accepted with a probability computed to preserve detailed balance relative to the target distribution (Tierney and Mira (1999)). If this second stage proposal is accepted the chain moves there, otherwise the same position is retained. In either case, only at this point, time is advanced. The advantage of the DR strategy is that the resulting algorithm dominates the standard MH since it produces estimates with a smaller asymptotic variance, in other words the DR dominates the corresponding single stage MH sampler in the Peskun ordering (Peskun (1973)) as proved by Tierney and Mira (1999). Also, the proposal distribution, which is typically hard to tune in regular MH samplers, can be improved upon rejection that is, the second stage proposal can be different from the first stage one and we are allowed to "learn" from previously rejected candidates (without loosing the Markovian property). This allows to locally tune the proposal with a partially (within sweep) adaptive strategy. Different forms of adaptation can be adopted. As suggested in Green and Mira (2001) the first stage proposal should permit "bold" moves (having high variance, for example), and should be simple to obtain and to sample from. The design of higher stage proposals can require more computational time (using for example more accurate approximations of the target at the current position of the chain) and should propose more "timid" moves. Along these lines, a possible strategy to update the proposal, expecially in a varying dimentional setting, is to use the "zeroth order method" suggested by Brooks et al. (2003) to design the first stage proposal, the "first order method" (more computationally intensive) at the second stage and so on.

We tried different updating schemes: single variable updating and block updating of all the variables of interest at once. The former strategy shows a much better performance than the latter for both the MH and the DR due to the fact that the range of variability of $\underline{\alpha}$ and $\underline{\beta}$ is quite different. We will thus only report the simulation results of the random scan single site updating scheme.

3.4 Simulation Results

The results reported were obtained by running a simulation of length 1024 ($= 2^{10}$) after a burn-in of 150 steps. Both the DR and the MH were started in the same position, namely all the variables are initialized at zero. Convergence to the core of the distribution happens quite fast thus the choice of the relatively short burn-in and length of the simulation. The proposal distributions are all normals centered at the current position of the chain thus leading to a random walk Metropolis-Hastings algorithm. As suggested in Green and Mira

(2001) the first stage proposal is over dispersed and σ_1 (the spread of the first stage proposal), for the various parameters, has been set, after having run 5 pilot simulations, equal to the values reported in Table 3.2. The second stage proposal has a $\sigma_2 = \sigma_1/2$. The comparison in terms of efficiency of the resulting estimates is made with a MH that uses the same Normal proposals but with spread equal to $(\sigma_1 + \sigma_2)/2$.

Table 3.2: Values of σ_1 used for the first stage proposal in the DR.

α_1	1.2
$\alpha_2, \cdots, \alpha_7, \mu_\alpha$	0.4
σ_α	3
β_1	0.15
β_2	0.4
β_3	0.3
β_4	0.15

The simulation results are presented in Table 3.3 where the mean along the sample path is reported for both the MH and the DR chain. The numbers in Table 3.3 and 3.5 have been obtained by averaging 5 independent runs of DR and MH to reduce the simulation bias. We report in parenthesis the standard deviations obtained over these 5 runs: the DR estimates appear to be more stable than the MH ones. The drawback of DR is that, in this particular application, it takes a time almost twice as long to run, compared to the MH. At this regard we point out that the code is written in GAUSS, an interpreted language, thus comparisons between DR and MH, that take simulation time into account, are not very meaningful.

Credible (confidence) intervals at 95 % level are also derived from the MCMC simulation (Table 3.3), by computing the 0.25 and the 0.975 quantiles of the simulated values.

For comparison purposes, in Table 3.3 we also report the MLE (maximum likelihood estimates) of the logistic regression parameters, $\underline{\alpha}$ and $\underline{\beta}$, obtained using a standard Newton-Raphson procedure. When computing the MLE we use 3.2 as the likelihood with a dummy variable for the intercept of sector 6 since the data show a much higher percentage of defaults here (in the sequel we will refer to this model as the "classical" model). As Table 3.3 shows, this dummy variable is justified also by the Bayesian analysis, since the estimated value of

Table 3.3: Estimates and credible (confidence) intervals of the parameters of interest for the Bayesian (MH and DR) and the classical model (MLE).

	MH Est. (sd)	MH Cred. Int.	DR Est. (sd)	DR Cred. Int.	MLE	ML Conf. Int.
α_1	-7.06 (0.008)	-10.10 ; -4.83	-6.76 (0.003)	-9.09; -4.98	-5.25	-5.66 ; -4.84
α_2	-5.47 (0.085)	-6.26 ; -4.82	-5.49 (0.115)	-6.20 ; -4.84	-5.25	-5.66 ; -4.84
α_3	-5.21 (0.020)	-5.75 ; -4.72	-5.21 (0.014)	-5.72 ; -4.74	-5.25	-5.66 ; -4.84
α_4	-4.99 (0.005)	-5.59 ; -4.45	-5.01 (0.002)	-5.58 ; -4.51	-5.25	-5.66 ; -4.84
α_5	-5.34 (0.237)	-5.93 ; -4.83	-5.36 (0.108)	-5.93 ; -4.86	-5.25	-5.66 ; -4.84
α_6	-4.03 (0.074)	-4.71 ; -3.46	-4.06 (0.067)	-4.67 ; -3.54	-3.54	-4.41;-2.66
α_7	-6.48 (0.024)	-7.15 ; -5.87	-6.50 (0.055)	-7.09 ; -5.97	-5.25	-5.66 ; -4.84
β_1	-0.10 (0.035)	-0.20 ; 0.01	-0.10 (0.075)	-0.18 ; 0.0	-0.083	-0.16; -0.002
β_2	-1.50 (0.050)	-2.35 ; -0.84	-1.54 (0.066)	-2.29 ; -0.85	-1.08	-1.65; -0.51
β_3	-1.38 (0.053)	-1.73 ; -1.06	-1.37 (0.071)	-1.66 ; -1.09	-1.13	-1.47 ; -0.79
β_4	0.06 (0.042)	-0.026 ; 0.14	0.07 (0.064)	-0.01 ; 0.13	0.08	-0.001 ; 0.16
μ_α	-5.49 (0.054)	-6.72 ; -4.34	-5.47 (0.097)	-6.43 ; -4.55		
σ^2_α	2.97 (0.293)	0.695 ; 7.28	2.21 (0.123)	0.65 ; 5.18		

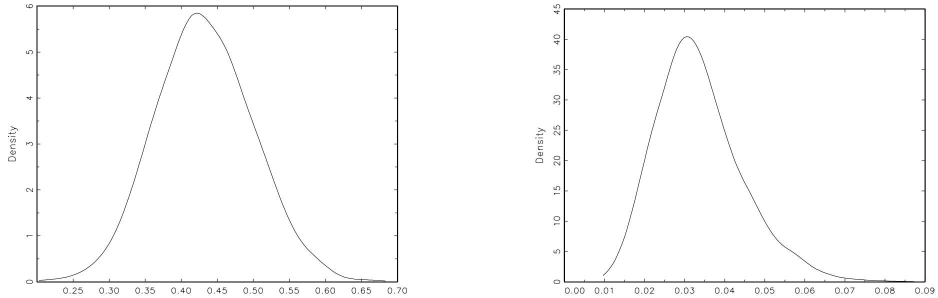

Figure 3.3: Posterior density estimate of DP: company 30 in sector 6 (top); company 20 in sector 2

the parameters in this sector are significantly different from the others. This dummy causes the MLE and the confidence interval for the intercept of sector 6 to be different from the others.

We preferred a generalized linear regression parametric model (versus, for example, a neural network) since the signs of the estimated $\underline{\beta}$ parameters are amenable for a financial interpretation: Variable 1 measures the overall economic performance of the firm and, as the estimate suggests, there is a negative relationship with the default probability; Variable 2 is related to the ability of the firm to pick-up external funds, the interpretation of this coefficient sign can be ambiguous; Variable 3 is related to the ability of the firm to generate cash flow to finance its short term activities, the negative sign of the parameter is expected; Variable 4 measures the inefficiency in administrating commercial activities, the obvious correlation with default probability is highlighted by the estimated parameter.

For each company we also derive the estimated posterior distribution of the DP by using a normal kernel density estimator on the values of $\theta_{i,j}$ computed at each point in time during the simulation. In Figure 3.3 two such distributions (for company 30 in sector 6 and company 20 in sector 2) are plotted: notice the long right tail behavior in the bottom picture which is quite common for companies with low risk.

Various estimates of the DP can be computed. Table 3.4 summaries the results obtained for

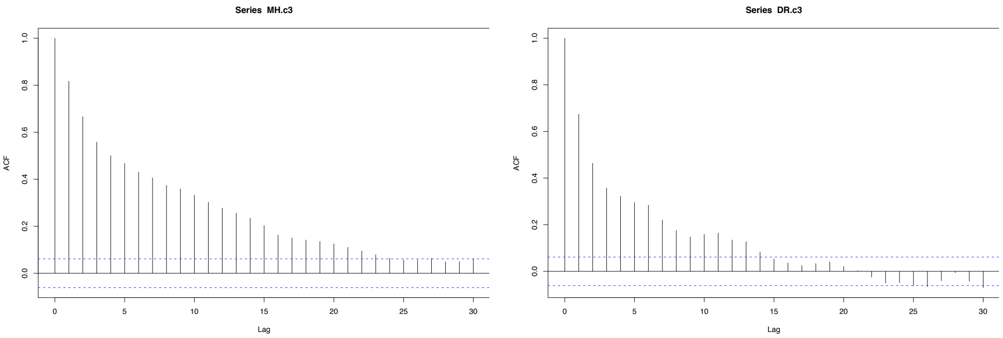

Figure 3.4: Autocorrelation functions for α_3: MH (left) and DR

the two companies mentioned above. In the first column we report the value obtained using formula 3.3 and substituting for α_j and $\underline{\beta}$ the estimates obtained with the DR algorithm by averaging over the whole simulation. In the second column we average the 1024 values of $\theta_{i,j}$ simulated at each step of the DR algorithm by substituting for α_j and $\underline{\beta}$ in 3.3the values of these parameters at that step in the simulation (these are the same values of $\theta_{i,j}$ used to get the kernel density estimator). In the last column the estimates of the DP obtained by ML are reported. As we can clearly see the MLE highly underestimates the probabilities of interests while the Bayesian estimates, in particular the ones reported in the second column, obtained by integrating over the posterior distribution of $\theta_{i,j}$, do not suffer from this drawback.

Table 3.4: Estimates of DP for company 30 in sector 6 and company 20 in sector 2.

	plug in posterior mean of $\underline{\alpha}$ and $\underline{\beta}$	posterior mean of $\theta_{i,j}$	MLE
$\hat{\theta}_{30,6}$	0.431	0.434	0.37169
$\hat{\theta}_{20,2}$	0.032	0.034	0.02576

All the estimates so far reported have been obtained from the DR simulation, unless otherwise specified. Similar values would be obtained from the MH sampler since both the algorithms produce Markov chains with the proper stationary distribution and both have converged according to the performed diagnostics. As pointed out before, the difference between the MH and the DR is in the asymptotic variance of the resulting estimators.

To compare the performance of the two samplers, in Figure 3.4 we present the graphs of the autocorrelation function (ACF) for one of the parameters of interest, α_3. The picture shows that the ACF for the DR is below the one obtained using the MH. This fact, true for all the parameters, is a signal of better mixing of the DR chain which explores the state space in a more efficient way.

For comparison purposes we also estimate the integrated auto correlation time, $\tau = \sum_{k=-\infty}^{\infty} \rho_k$, where $\rho_k = \text{cov}_P\{\phi(X_0), \phi(X_k)\}/\sigma^2$, ϕ is the function of interest (we have taken $\phi(x) = x$), and σ^2 is the finite variance of ϕ under the posterior π. To estimate τ we used Sokal's adaptive truncated periodogram estimator Sokal. The results are presented in Tables 3.5 and 3.6 and show that, for all the parameters of interest, the DR outperforms the MH.

Table 3.5: Estimates of τ for α with MH and DR.

	α_1	α_2	α_3	α_4	α_5	α_6	α_7
MH	26.9	50.1	43.2	50.3	54.6	60.6	60.2
DR	17.0	18.4	28.1	28.4	30.1	32.3	35.1

Table 3.6: Estimates of τ for β and the hyper-parameters with MH and DR.

	β_1	β_2	β_3	β_4	μ_α	σ_α^2
MH	10.0	64.5	23.4	5.6	15.9	20.2
DR	7.2	38.1	20.9	4.2	14.6	15.6

To compare the predictive performance of the Bayesian versus the classical logistic regression model a cross-validation analysis has been performed. In Figures 3.6, 3.7,3.8,3.9,3.10,3.11 and 3.12 we represent, for each sector, the predicted default and not default detected by the Bayesian and classical model estimated via MLE (there is no graph for not defaulted companies for sector 1 since no defaults were observed). To estimate the two models we used 70% of the total observations while the remaining sample was used to validate the model. The two samples (training and validation) are randomly selected but balanced in that they have the same proportion of defaults for each sector as in the original sample. On the x-axis the observation number is indicated, on the y-axis the default probability. For the graphs on the right hand side we would like these probabilities to be as high as possible and, comparing the classical (solid line) with the Bayesian model (dashed line) we detect that the proposed model outperforms the classical one for every sector except the last one (sector 7) which is the sector with observed smallest default frequency (excluding sector 1, which is a residual sector). As for the graphs on the left hand side, there are companies that, according to both models, would not receive any credit line despite the fact that they showed no default, that is, both models misclassify these companies and, the Bayesian model is more inclined toward this.

To have an overall feeling of the comparative performance of the two models we computed, on the test sample, the root mean squared error of classification:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\theta}_i)^2}$$

where y_i is either zero or one and $\hat{\theta}_i$ is the estimated default probability (for simplicity we slightly change the notation here). This performance indicator has been computed on the test sample for both defaulted and not defaulted companies (thus having $n = 30\% \times 7513 = 2254$) and also for the subset of defaulted companies alone as well as for the subset of not defaulted ones. The results are reported in Table 3.7 and show the overall better performance of the Bayesian model.

Table 3.7: Estimated root mean squared error

	MLE	Bayesian
all	0.1282	0.1273
not defaulted	0.0280	0.0272
defaulted	0.9646	0.9591

Finally, in Figure 3.5, we show how the percentage of correct classification for defaulted (right picture) and not defaulted (left picture) companies varies as the threshold defined to classify them ranges between zero and one. Again the proposed Bayesian model outperforms the classical one for practically all values of the threshold.

3.5 Conclusions

The proposed model presents various advantages. First the fact that the output of the Bayesian approach is the estimate of the posterior distribution of the DP of each company. Having a distribution instead of a punctual value, we obtain a more complete and informative picture of the quantity of interest, that's to say the parameter uncertainty is also and easly taken into account during default prediction.

The second advantage is that our procedure does not suffer from bias problems which are typical for rare events (King and Zeng (2001)). Also, the hierarchical model allows parametric flexibility among sectors to estimate DP, while allowing sectors with tiny data to receive strength from the data available from other sectors, we get therefore more reliable results.

To compare the predictive performance of the Bayesian versus the classical model we performed a cross-validation analysis. By computing the root mean squared error of classification and the percentage of correct classification for a varying threshold, we show how the Bayesian model overall outperforms the classical one.

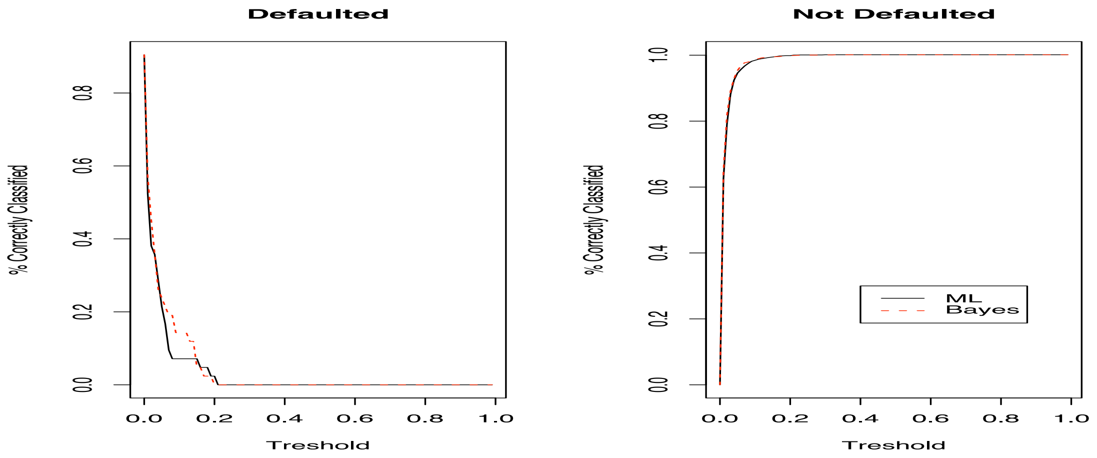

Figure 3.5: Percentage of correct classification for defaulted (left) and not defaulted companied as the classification threshold varies.

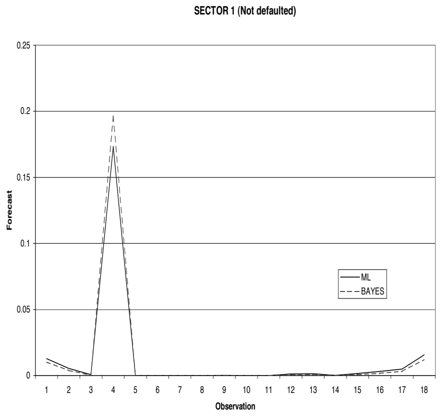

Figure 3.6: Sector 1.

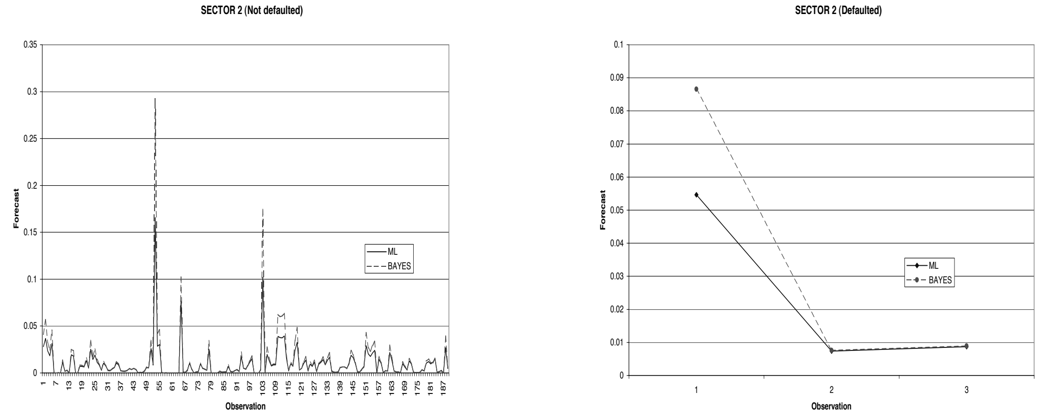

Figure 3.7: Sector 2.

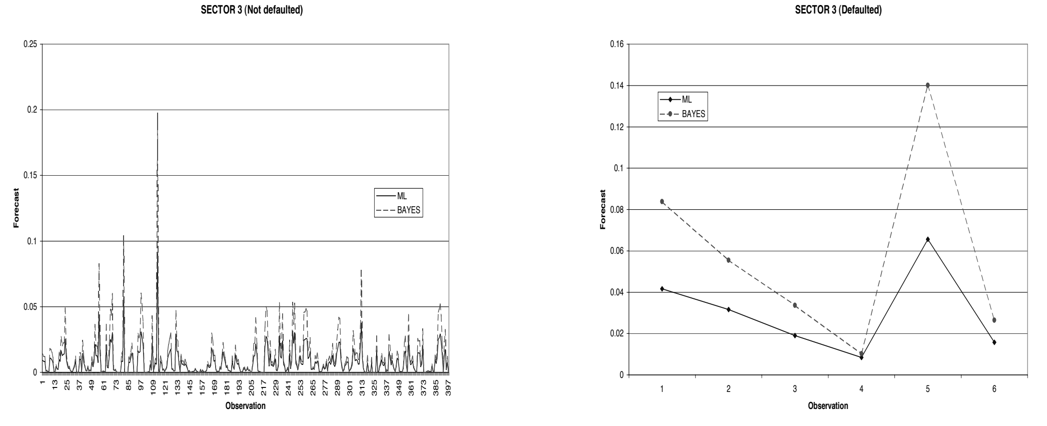

Figure 3.8: Sector 3.

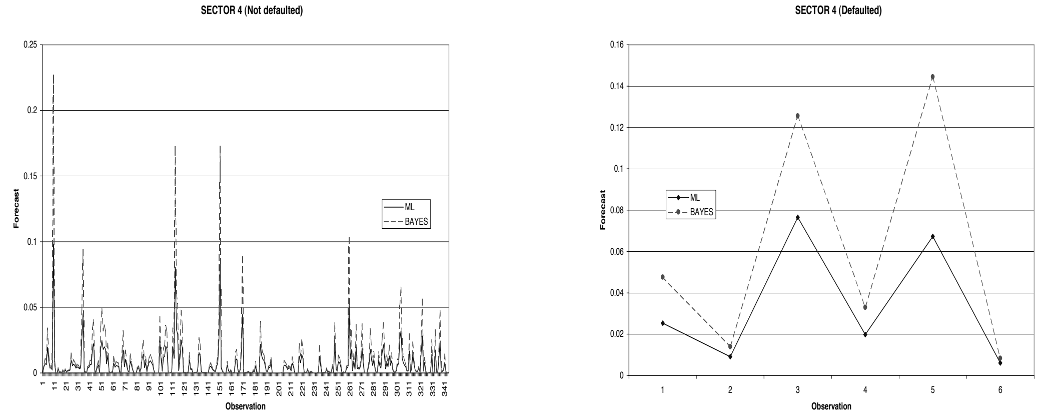

Figure 3.9: Sector 4.

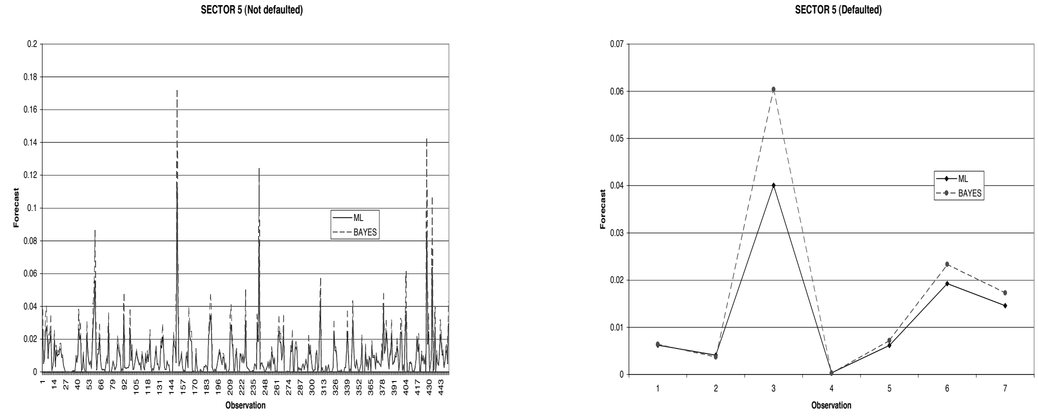

Figure 3.10: Sector 5.

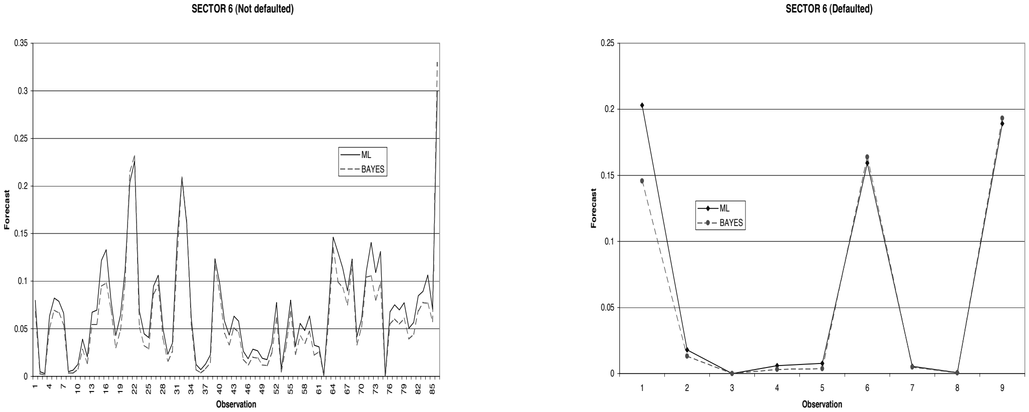

Figure 3.11: Sector 6.

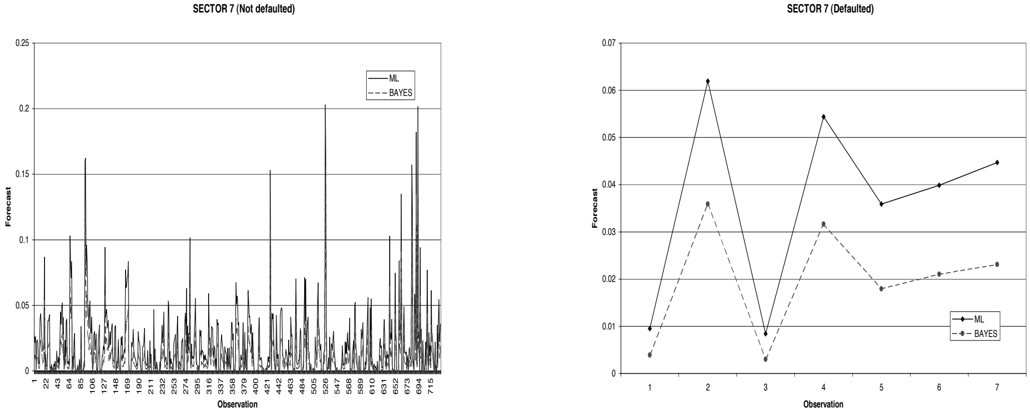

Figure 3.12: Sector 7.

References

S. P. Brooks, P. Giudici, and G.O. Roberts. Efficient construction of reversible jump MCMC proposal distributions. *Journal of the Royal Statistical Society, Series B*, 65:3–55, 2003.

P. J. Green and A. Mira. Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika*, 88:1035–1053, 2001.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9:2:137–163, 2001.

P. H. Peskun. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

A. D. Sokal. Monte Carlo methods in statistical mechanics: foundations and new algorithms. Cours de Troisième Cycle de la Physique en Suisse Romande, Lausanne 1989.

L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22: 1701–1762, 1994.

L. Tierney and A. Mira. Some adaptive Monte Carlo methods for bayesian inference. *Statistics in Medicine*, 18:2507–2515, 1999.

Chapter 4

Zero variance Markov chain Monte Carlo

We propose a general purpose variance reduction technique for Markov Chain Monte Carlo (MCMC) estimators based on the zero-variance principle introduced in the physics literature by Assaraf and Caffarel (1999, 2003). The potential of the new idea is illustrated with some toy examples and a real application to Bayesian inference for credit risk estimation.

4.1 Main idea

We are interested in estimating the expected value of a function f with respect to a, possibly unnormalized, probability distribution π:

$$\mu_f = \frac{\int f(x)\pi(x)dx}{\int \pi(x)dx}. \tag{4.1}$$

Markov chain Monte Carlo methods (Metropolis et al. (1953), Hastings (1970), Tierney (1994)), estimate integrals using a large but finite set of sample points, $x^i, i = 1, \cdots, N$, collected along the sample path of an ergodic Markov chain, P, having π (normalized) as its unique stationary and limiting distribution:

$$\hat{\mu}_f = \frac{1}{N}\sum_{i=1}^{N} f(x^i). \tag{4.2}$$

We have that

$$\mu_f = \hat{\mu}_f + \Delta\mu_f,$$

where $\Delta\mu_f$ is the statistical error associated with the fact that the length of the simulated Markov chain path, N, is finite. For large enough N, standard statistical arguments lead to the following expression of the error:

$$\Delta\mu_f = K_f \frac{\sigma_f}{\sqrt{N}}$$

where the constant K_f is proportional to the amount of correlation along the sampled chain and σ_f is the standard deviation of f under π (assumed to be finite).

Recent literature (Peskun (1973), Liu (1996), Tierney (1998), Tierney and Mira (1999), Mira and Geyer (2000), Green and Mira (2001)), aimed at reducing the statistical MCMC error, $\Delta\mu_f$, by reducing the correlation along the Markov chain, that is, by reducing K_f. See Mira (2001) for a review.

In this chapter we suggest instead to reduce the error, decreasing its second component, σ_f, by replacing f with a different function, $\tilde{f}$, obtained by properly re-normalizing f. The function $\tilde{f}$ is constructed so that its expectation, under π, equals μ_f, but its variance with respect to π is smaller. This is a standard variance reduction technique used in Monte Carlo simulation, see Gilks et al. (1996). The novelty of the zero-variance principle is that, to define $\tilde{f}$, an operator, H, and a trial function, ϕ, are introduced. We require that H is Hermitian (symmetric for finite state spaces, and real in all practical applications) and

$$\int H(x,y)\sqrt{\pi(y)}dy = 0. \tag{4.3}$$

The trial function $\phi(x)$ is a rather arbitrary function which is only required to be integrable.

We define the renormalized function to be

$$\tilde{f}(x) = f(x) + \frac{\int H(x,y)\phi(y)dy}{\sqrt{\pi(x)}} = f(x) + \Delta f(x). \tag{4.4}$$

As a consequence of (4.1) and (4.3) we have that

$$\mu_f = \mu_{\tilde{f}}, \tag{4.5}$$

that is, both functions f and $\tilde{f}$ can be used to estimate the desired quantity via Monte Carlo or MCMC simulation. However, the statistical error of the resulting estimator can be very different. The optimal choice for (H, ϕ), i.e. the one that leads to zero-variance, can be obtained by imposing that $\tilde{f}$ is constant and equal to its average, that is, by requiring

$$\sigma_{\tilde{f}} = 0,$$

which is equivalent to require that

$$\tilde{f} = \mu_f.$$

The latter, together with (4.4), leads to the fundamental equation:

$$\int H(x,y)\phi(y)dy = -\sqrt{\pi(x)}[f(x) - \mu_f]. \tag{4.6}$$

In most practical applications equation (4.6) cannot be solved exactly, still, we propose to find an approximate solution in the following way. First choose H verifying (4.3) (in Section 4.2 we will suggest two general recipes to construct H). Second, parametrize ϕ and optimally choose the parameters by minimizing $\sigma_{\tilde{f}}$ over a finite set of points generated according to the Markov chain P. Finally, a much longer MCMC simulation is performed using $\hat{\mu}_{\tilde{f}}$ instead of $\hat{\mu}_f$ as the estimator. Note that the proposed approach can be used to obtain variance reduction also in Monte Carlo simulation if we can get i.i.d. draws from the target distribution π. This is what is done in Assaraf and Caffarel (1999, 2003).

4.2 Choice of H

In this section the rationale to choose the operator H, both for discrete and continuos settings, is illustrated.

4.2.1 Discrete case

Denote with $P(x,y)$ a transition matrix reversible with respect to π (we identify a Markov chain with the corresponding transition matrix of kernel):

$$\pi(x)P(x,y) = \pi(y)P(y,x), \qquad \forall x, y.$$

The following choice of H

$$H(x,y) = \sqrt{\frac{\pi(x)}{\pi(y)}}[P(x,y) - \delta(x-y)]$$

satisfies the requirements, where $\delta(x-y)$ is the Dirac delta function: $\delta(x-y) = 1$ if $x = y$ and zero otherwise. With this choice of H, letting $\tilde{\phi} = \frac{\phi}{\sqrt{\pi}}$, equation (4.4) becomes:

$$\tilde{f}(x) = f(x) - \int P(x,y)[\tilde{\phi}(x) - \tilde{\phi}(y)]dy. \tag{4.7}$$

The main difficulty with (4.7) is the evaluation of the integral.

4.2.2 Continuous case

If $x \in \Re^d$ we can consider the operator:

$$H = -\frac{1}{2}\sum_{i=1}^{d}\frac{\partial^2}{\partial x_i^2} + V(x) \tag{4.8}$$

where $V(x)$ is constructed to fulfill equation (4.3):

$$V(x) = \frac{1}{2\sqrt{\pi(x)}} \sum_{i=1}^{d} \frac{\partial^2 \sqrt{\pi(x)}}{\partial x_i^2}. \tag{4.9}$$

In this setting we have that

$$\tilde{f}(x) = f(x) + \frac{H\phi(x)}{\sqrt{\pi(x)}}. \tag{4.10}$$

This is the function we use in the examples considered in the sequel. To obtain the first and second order derivatives we used the R function “hessian” from the library “numDeriv” which evaluates an approximate Hessian of a scalar function using finite differences. Note that, for calculating $\tilde{f}$ with the operator (4.8), the normalizing constant of $\pi(x)$ is not needed.

4.3 Choice of ϕ

The optimal choice of ϕ is the *exact solution* of the fundamental equation (4.6). In real applications, typically, only *approximate solutions*, obtained by numerically minimizing $\sigma_{\tilde{f}}$, are available. In other words, we select a functional form for ϕ, typically a polynomial, parametrized by some coefficients, and optimize those coefficientsby minimizing the fluctuations of the resulting $\tilde{f}$, obtained by Monte Carlo or Markov chain Monte Carlo simulations. The particular form of ϕ is very dependent on the problem at hand, that is on π, and on f. However an important point to notice is that, if we parametrize ϕ in terms of $c = \int \phi(x)dx$ and then minimize $\sigma_{\tilde{f}}$ with respect to c, the optimal choice of c is

$$c = -\frac{[E_\pi(f(x)\Delta f(x))]^2}{E_\pi(\Delta f(x))^2}$$

and, for this value of the parameter, from (4.4) we obtain

$$\sigma_{\tilde{f}}^2 = \sigma_f^2 - \frac{[E_\pi(f(x)\Delta f(x))]^2}{E_\pi(\Delta f(x))^2}. \tag{4.11}$$

Since the correction factor in (4.11), that leads from σ_f^2 to $\sigma_{\tilde{f}}^2$, is always negative, regardless of the choice of ϕ, a variance reduction in the MCMC estimator is obtained by replacing f with $\tilde{f}$ in (4.2).

4.4 Examples of variance reduction in Monte Carlo case

In this section we present a few toy examples to demonstrate the power of the proposed zero-variance technique. In particular we consider as target:

1. Univariate and bivariate Gaussian distributions,
2. Univariate and bivariate Student-T distributions.

The functions of interest, f, are:

- $f(x) = x$ and $f(x) = x^2$ in the univariate case,
- $f(x_1, x_2) = x_1$, $f(x_1, x_2) = x_1^2$ and $f(x_1, x_2) = x_1 x_2$ in the bivariate case.

These are the typical quantities of interest in a Bayesian setting, where π is the posterior distribution and one is interested in evaluating the posterior mean, variance and covariance of the parameters. In the tables we report $\hat{\mu}_f$ as defined in (4.2) and $\hat{\sigma}_f^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(f\left(x^i\right) - \hat{\mu}_f\right)^2$. The estimates $\hat{\mu}_{\tilde{f}}$ and $\hat{\sigma}_{\tilde{f}}^2$ are similarly defined and also reported.

In the results presented we sample $N = 150$ iid values from π, unless otherwise stated.

4.4.1 Univariate Gaussian distribution

Consider as target a normal distribution, $N(\mu, \sigma^2)$, with non-normalized density $\pi(x) = \exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})$. In this case the theoretical functions ϕ that solve (4.6) are respectively for

Table 4.1: $N(\mu = 1, \sigma^2 = 2)$, $f_1(x) = x$, $f_2(x) = x^2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$
$\hat{\mu}_f$	0.912	1	2.824	3
$\hat{\sigma}_f^2$	2.013	2.28e-22	9.377	3.53e-21

Table 4.2: Univariate Student-T with $g = 5$, $f_1(x) = x$, $f_2(x) = x^2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$
$\hat{\mu}_f$	-0.271	1.65e-12	1.834	1.666
$\hat{\sigma}_f^2$	1.778	5.19e-22	20.536	1.32e-23

Table 4.3: Bivariate Normal, $(\mu_1, \mu_2) = (2, 1)$, $(\sigma_1, \sigma_2) = (2, 1)$, $\rho = 0.6$, $f_1 = x_1$, $f_2 = x_1^2$, $f_3 = x_1 x_2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$	f_3	$\tilde{f}_3$
$\hat{\mu}_f$	1.703	2	6.518	8	2.582	3.2
$\hat{\sigma}_f^2$	3.654	3.48e-20	7.199	4.63e-18	13.053	5.76e-20

Table 4.4: Bivariate Student-T, $g = 7$, $f_1 = x_1$, $f_2 = x_1^2$, $f_3 = x_1 x_2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$	f_3	$\tilde{f}_3$
$\hat{\mu}_f$	-0.220	1.31e-11	1.398	1.4	0.231	-1.06e-12
$\hat{\sigma}_f^2$	1.363	8.85e-21	8.843	4.07e-19	1.724	2.03e-22

$f_1(x) = x$ and $f_2(x) = x^2$:

$$\begin{aligned} &\phi_1(x) = (-2\sigma^2 x)\sqrt{\pi\,(x)}; \\ &\text{and} \\ &\phi_2(x) = (-\sigma^2 x^2 - 2\mu\sigma^2 x)\sqrt{\pi\,(x)}. \end{aligned}$$

In Table 4.1 we show Monte Carlo simulation results for $f_1(x) = x$, $f_2(x) = x^2$ and the associated $\tilde{f}_1(x)$ and $\tilde{f}_2(x)$ for a $N(\mu = 1, \sigma^2 = 2)$ target. Despite the small sample ($N = 150$), a great reduction in variability of the estimator (estimated through the sample variance), is achieved and the final variance is nearly zero.

4.4.2 Univariate Student-T distribution

In this section we proceed as in the previous one but taking the univariate Student-T distribution with $g > 2$ degrees of freedom, as the target. In this case the non-normalized density is $\pi(x) = \left(1 + \frac{x^2}{g}\right)^{-\frac{g+1}{2}}$ and the theoretical functions ϕ that solve (4.6) are, respectively, for $f_1(x) = x$ and $f_2(x) = x^2$:

$$\begin{aligned} \phi_1\,(x) &= \left(\frac{2}{3}\frac{1}{1-g}x^3 + 2\frac{g}{1-g}x\right)\sqrt{\pi\,(x)} \\ \text{and} & \\ \phi_2\,(x) &= \left(\frac{1}{2}\frac{1}{2-g}x^4 + \frac{g}{2-g}x^2\right)\sqrt{\pi\,(x)}. \end{aligned}$$

Also in this case the Monte Carlo simulation results, displayed in Table 4.2, for a Student-T distribution with $g = 5$ degrees of freedom, show an estimated variance close to zero.

4.4.3 Bivariate Gaussian

We consider here a two dimensional vector, $\underline{x} = (x_1, x_2)$, having a bivariate normal distribution with mean vector $\underline{\mu} = (\mu_1, \mu_2)$, standard deviations $\underline{\sigma} = (\sigma_1, \sigma_2)$, and correlation coefficient ρ. The theoretical ϕ functions for $f_1(\underline{\mathrm{x}}) = x_1$, $f_2(\underline{\mathrm{x}}) = x_1^2$ and $f_3(\underline{\mathrm{x}}) = x_1 x_2$, are, respectively:

$$\begin{aligned}
\phi_1(\underline{\mathrm{x}}) &= \left(-2\sigma_1^2 x_1 - 2\rho\sigma_1\sigma_2 x_2\right)\sqrt{\pi(\underline{x})};\\
\phi_2(\underline{\mathrm{x}}) &= \Bigg\{ \left[-\rho^2\frac{\sigma_1^4}{\sigma_1^2+\sigma_2^2} - \sigma_1^2\left(1-\rho^2\right)\right] x_1^2 + \left[-\rho^2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right] x_2^2 + \left[-\rho\frac{\sigma_1^3\sigma_2}{\sigma_1^2+\sigma_2^2}\right] x_1 x_2\\
&\quad + \left[-2\mu_1\sigma_1^2 + 2\rho\frac{\sigma_1^3\sigma_2}{\sigma_1^2+\sigma_2^2}\mu_2 - 2\rho^2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\mu_1\right] x_1\\
&\quad + \left[-2\sigma_1\sigma_2\mu_1\rho - 2\rho\frac{\sigma_1\sigma_2^3}{\sigma_1^2+\sigma_2^2}\mu_1 + 2\rho^2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\mu_2\right] x_2 \Bigg\}\sqrt{\pi(\underline{x})};\\
\phi_3(\underline{\mathrm{x}}) &= \Bigg\{ \left[-\rho\frac{\sigma_1^3\sigma_2}{\sigma_1^2+\sigma_2^2}\right] x_1^2 + \left[-\rho\frac{\sigma_1\sigma_2^3}{\sigma_1^2+\sigma_2^2}\right] x_2^2 + \left[-2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right] x_1 x_2\\
&\quad + \left[-2\frac{\sigma_1^4}{\sigma_1^2+\sigma_2^2}\mu_2 - 2\rho\frac{\sigma_1\sigma_2^3}{\sigma_1^2+\sigma_2^2}\mu_1\right] x_1 + \left[-2\frac{\sigma_2^4}{\sigma_1^2+\sigma_2^2}\mu_1 - 2\rho\frac{\sigma_1^3\sigma_2}{\sigma_1^2+\sigma_2^2}\mu_2\right] x_2 \Bigg\}\sqrt{\pi(\underline{x})}.
\end{aligned}$$

We consider first a standard bivariate Gaussian target and then move on to the case where $\underline{\mu} = (2, 1), \underline{\sigma} = (2, 1)$ and $\rho = 0.6$. In Table 4.3 we report the Monte Carlo simulation results obtained, for the latter case, and we have again a near zero-variance.

4.4.4 Bivariate Student-T

We conclude the Monte Carlo simulation with theoretical knowledge of the exact ϕ's functions, with the simulation of a bivariate Student-T distribution. The theoretical ϕ functions for $f_1(\underline{\mathrm{x}}) = x_1$, $f_2(\underline{\mathrm{x}}) = x_1^2$ and $f_3(\underline{\mathrm{x}}) = x_1 x_2$, are, respectively:

$$\begin{aligned}
\phi_1(\underline{\mathrm{x}}) &= \left(\frac{2}{2-3g} x_1^3 + \frac{2}{2-3g} x_1 x_2^2 + \frac{6g}{2-3g} x_1 \right) \sqrt{\pi(\underline{\mathrm{x}})};\\
\phi_2(\underline{\mathrm{x}}) &= \left\{ \left[\frac{1}{4} \frac{3-2g}{(2-g)(1-g)} \right] x_1^4 + \left[-\frac{1}{4} \frac{1}{(2-g)(1-g)} \right] x_2^4 + \left[\frac{1}{2} \frac{1}{2-g} \right] x_1^2 x_2^2 \right.\\
&\quad \left. + \left[\frac{1}{2} \frac{g(3-2g)}{(2-g)(1-g)} \right] x_1^2 + \left[-\frac{1}{2} \frac{g}{(2-g)(1-g)} \right] x_2^2 \right\} \sqrt{\pi(\underline{\mathrm{x}})};\\
\phi_3(\underline{\mathrm{x}}) &= \left(\frac{1}{2} \frac{1}{1-g} x_1^3 x_2 + \frac{1}{2} \frac{1}{1-g} x_1 x_2^3 + \frac{g}{1-g} x_1 x_2 \right) \sqrt{\pi(\underline{\mathrm{x}})}.
\end{aligned}$$

The simulation results, for a Student-T with 7 degrees of freedom, are reported in Table 4.4 and confirm the reduction of variance toward zero.

4.4.5 A first discussion of the gained insight

As shown in the previous subsection, in the Monte Carlo framework this method works well when the theoretical ϕ is available. However, in most practical applications, two problems may arise:

1. The impossibility to get iid samples from the target distribution;
2. The unavailability of the theoretical ϕ.

To overcome the first problem one could use MCMC simulation techniques, however it would be questionable if the zero-variance machinery introduced, works properly also in a MCMC setting. The answer is affirmative, indeed it is straightforward to show that, when the exact ϕ is available, i.e. ϕ satisfying equation (4.6), $Cov(\tilde{f}(x^0), \tilde{f}(x^k))$ is also zero for all k. Our simulations confirm this fact, moreover, also when the ϕ function is not the exact one, the variance is reduced dramatically.

The second problem is more delicate but, as pointed out in Section 4.3, any choice of ϕ reduces the variance. A good choice of ϕ remains an open question that we want to address here. In the previous examples we observed that the theoretical ϕ's take the form $P(x)\sqrt{\pi\left(x\right)}$ where $P(x)$ is a polynomial. Furthermore, we noticed the influence of the following two factors on the degree of the polynomial $P(x)$:

a) The degree of the function $f(x)$;

b) The structure of the target.

Regarding the first of the two factors, a simple suggestion would be to control it by imposing $P(x)$ to have the same degree of $f(x)$. This is what we have observed for the optimal ϕ in the Gaussian case (both univariate and bivariate), when $f(x) = x$ or x^2. This can be justified, for other targets, by resorting to normal approximation arguments. The second factor varies strongly among problems faced, so it is difficult to give a general suggestion, however our experimental results are robust to misspecification of the ϕ function. As an example we tried to impose a first order $P(x)$ for an univarite Student-T target distribution, whose theoretical ϕ requires a third degree polynomial, when one is interested in $E_\pi(x)$, i.e. $f(x) = x$. We obtained a promising 93% variance reduction: a performance only a little worse with respect to the exact ϕ. This robustness, to misspecification of the degree of the $P(x)$ polinomial in the ϕ function, is mantained also for MCMC samples.

4.5 Examples of Variance reduction in MCMC case

We revisit the examples studied in Section 4.4, by running a Markov chain using the exact theoretical ϕ and we obtain results similar to the Monte Carlo case. In Tables 4.5, 4.6, 4.7, 4.8 we report the simulation results. These are obtained by simulating 1000 points with a random walk Metropolis Hastings with an optimally scaled Normal proposal distribution (see Roberts and Rosenthal, 2001) and then discarding the first 850 points so that the number of MCMC actual points compares to the number of MC draws used in Section 4.4 (i.e. $N = 150$). For the central limit theorem of ergodic averages

$$N^{1/2}\left(\hat{\mu}_f^{(N)} - \mu_f\right) \rightarrow N\left(0, \nu_f^2\right).$$

The following result gives an expressions for υ_f^2 for Markov chains

$$\nu_f^2 = \sigma_f^2 + 2\sum_{i=2}^{\infty} Cov\left(f\left(x^1\right), f\left(x^i\right)\right),$$

the ratio

$$eff_{\hat{\mu}_f} = \frac{\sigma_f^2}{\nu_f^2}$$

is a measure of efficiency of the Markov chain for estimating μ_f, see Roberts (1996b). A similar expression for $\nu_{\tilde{f}}^2$ also holds

$$\nu_{\tilde{f}}^2 = \sigma_{\tilde{f}}^2 + 2\sum_{i=2}^{\infty} Cov\left(\tilde{f}\left(x^1\right), \tilde{f}\left(x^i\right)\right).$$

In the following tables $\hat{\sigma}_f^2$ and $\hat{\sigma}_{\tilde{f}}^2$ are variances computed within each chain and we refer to them simply as the first of the two components contributing to the efficiency of the Markov chain for estimating μ_f and $\mu_{\tilde{f}}$. In subsection 4.5.2 however we compute $\hat{\nu}_f^2$ and $\hat{\nu}_{\tilde{f}}^2$ estimating variances of $\hat{\mu}_f^N$ and $\hat{\mu}_{\tilde{f}}^N$ obtained over different chains.

4.5.1 Gaussian-Gaussian model

Consider the following Bayesian model for s iid observations y_i:

$$l(y_i|\theta) \sim N(\theta, \sigma_y^2) \qquad i = 1, \cdots, s;$$

where σ_y^2 is the known variance and θ is the parameter of interest. We assume a conjugate Normal prior:

$$h(\theta) \sim N(\mu_\theta, \tau_\theta^2)$$

where μ_θ and τ_θ^2 are known hyperparameters. It is well known that posterior distribution of the parameter of interest is

$$\pi(\theta|y_1, \cdots, y_s) = N(\mu_\pi, \sigma_\pi^2)$$

Table 4.5: $N(\mu = 1, \sigma^2 = 2)$, $f_1(x) = x$, $f_2(x) = x^2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$
$\hat{\mu}_f$	0.080	1	3,193	3
$\hat{\sigma}_f^2$	2.563	4.8e-20	13.209	1.31e-19

Table 4.6: Univariate Student-T with $g = 5$, $f_1(x) = x$, $f_2(x) = x^2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$
$\hat{\mu}_f$	0.095	2.08e-12	1.55	1.666
$\hat{\sigma}_f^2$	1.551	1.08e-22	4.077	6.51e-24

Table 4.7: Bivariate Normal, $(\mu_1, \mu_2) = (2, 1)$, $(\sigma_1, \sigma_2) = (4, 1)$, $\rho = 0.6$, $f_1 = x_1$, $f_2 = x_1^2$, $f_3 = x_1 x_2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$	f_3	$\tilde{f}_3$
$\hat{\mu}_f$	1,683	2,549	5.366	8	2.136	3.2
$\hat{\sigma}_f^2$	2	2,01e-16	33.937	1.193e-14	7.14	7.11e-17

Table 4.8: Bivariate Student-T, $g = 7$, $f_1 = x_1$, $f_2 = x_1^2$, $f_3 = x_1 x_2$.

	f_1	$\tilde{f}_1$	f_2	$\tilde{f}_2$	f_3	$\tilde{f}_3$
$\hat{\mu}_f$	-0.09	7.29e-10	1.049	1.4	-0.038	-4,31e-12
$\hat{\sigma}_f^2$	1.04	1.02e-17	5.44	1.92e-17	1.254	1.95e-21

where

$$\mu_\pi = \frac{\mu_\theta \sigma_y^2 + s\tau_\theta^2 \overline{y}}{\sigma_y^2 + s\tau_\theta^2}$$

and

$$\sigma_\pi^2 = \frac{\sigma_y^2 \tau_\theta^2}{\sigma_y^2 + s\tau_\theta^2},$$

here $\overline{y}$ is the sample mean. In this setting we considered $f(\theta) = \theta$ and:

$$\phi(\theta) = \phi_1(\theta)$$

where $\phi_1(\theta)$ is the function defined in section 4.4.1. As a concrete example we used $\sigma_y = 3, \mu_\theta = 0, \tau_\theta = 3$ and generated the actual sample of size $s = 10$, from a Gaussian distribution with mean equal to one and standard deviation equal to 3. The target posterior distribution has $\mu_\pi = 1.7487$ and $\sigma_\pi^2 = 0.904$. The summary of f and $\tilde{f}$, computed on $N = 500$ MCMC sampled values (after a burn-in of 100) , are presented in Table 4.9. Again, the advantage in terms of variance reduction, using $\tilde{f}$ in place of f, is clear.

Table 4.9: Bayesian Gaussian-Gaussian model, $f(\theta) = \theta$, $N = 500$.

	f	$\tilde{f}$
$\hat{\mu}_f$	1.7736	1.7399
$\hat{\sigma}_f^2$	0.8838	0.0362 (96% reduction)

4.5.2 Poisson-Gamma model

As a second model we consider the well known *Poisson-Gamma* model where:

$$l(y_i|\theta) \sim Po(\theta), \quad i = 1, \cdots, s;$$
$$h(\theta) \sim Ga(\alpha = 4, \beta = 4).$$

We generate $s = 30$ values from a $Po(\theta = 4)$ distribution, we then

1. run a first MCMC simulation of length 1000 with a burn-in of 100;
2. minimize the variance of $\tilde{f}$, obtained using ϕ_1 in 4.4.1, on this first simulation and save the numerically optimized parameters;
3. run 100 parallel MCMC chains, each of length 10000 (after a burn-in of 150 steps);
4. compute, on each chain, μ_f, $\mu_{\tilde{f}}$ and the resulting between chain variances, $\hat{\nu}_f^2$ and $\hat{\nu}_{\tilde{f}}^2$.

We are interested in the first moment of the posterior distribution, in this case we have the exact solution:

$$\frac{\beta + \sum_{i=1}^{s} y_i}{\alpha + s} = 4.058824.$$

The inspection of parallel chains[1], for example at 500 iterations, shows that $\bar{\hat{\mu}}_f = 4.060625$, $\bar{\hat{\mu}}_{\tilde{f}} = 4.058843$ and $\hat{\nu}_f^2 = var\left(\hat{\mu}_f\right) = 0.0150$ while $\hat{\nu}_{\tilde{f}}^2 = var\left(\hat{\mu}_{\tilde{f}}\right) = 0.001777$. A variance reduction of 87% is achieved.

Figure 4.1 depicts the results obtained considering the variance among means computed on different chains, while in Figure 4.2 the convergence for one of these chains is shown.

4.5.3 Logistic regression

We now consider a logistic regression model, commonly used in statistical applications. We simulate dependent binary data as follows:

$$l(y_i|\theta) \sim Be(\theta_i) \qquad , \qquad i = 1, ..., 100;$$

$$\theta_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \ , \ x_i \sim N(0, 1);$$

[1] With the symbol $\bar{\hat{\mu}}_f$ we intend the empirical average of the mean of f computed on different chains, see Roberts (1996b), similarly for $\bar{\hat{\mu}}_{\tilde{f}}$. With the symbol $\hat{\nu}_f^2$ we mean the variance of $\hat{\mu}_f$ computed over different chains, similarly for $\hat{\nu}_{\tilde{f}}^2$.

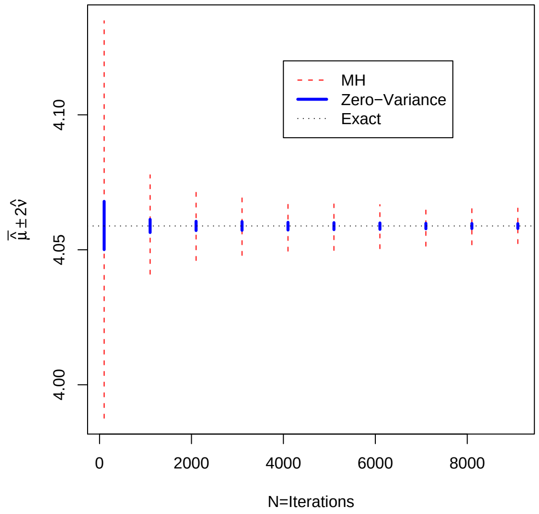

Figure 4.1: Poisson-Gamma: parallel chains

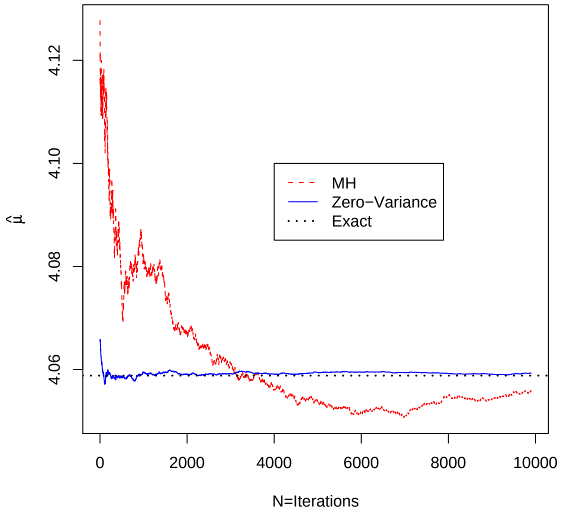

Figure 4.2: Poisson-Gamma: single chain

setting $\beta_0 = 0.5$, $\beta_1 = 1.5$. Then we estimate a Bayesian logistic regression, using an uninformative prior on each parameter.

The posterior distribution does not have a closed form, however we resort to its normal approximation and therefore choose a ϕ with the same structure of the optimal ϕ for the normal case. In Tables 4.10 and 4.11, we report the resulting variance reductions, obtained using 300 MCMC sampled values, after a burn-in of 1000.

Table 4.10: Logit model, $f(\beta_0, \beta_1) = \beta_0$, $N = 300$.

	f	$\tilde{f}$
$\hat{\mu}_f$	0.5676	0.5629
$\hat{\sigma}_f^2$	0.0923	0.0018 (80% reduction)

Table 4.11: Logit model, $f(\beta_0, \beta_1) = \beta_1$, $N = 300$.

	f	$\tilde{f}$
$\hat{\mu}_f$	2.0089	1.9758
$\hat{\sigma}_f^2$	0.1839	0.0122 (93% reduction)

4.6 A simplified credit risk model

We now consider a real application, estimating the parameters of a logistic regression for creditworthiness. We analyze a sample of 124 firms that gave rise to problematic credit and a sample of 200 healthy firms (so that $s = 324$). The models proposed is the following

$$\pi\left(\underline{\beta}|y, x\right) \propto \prod_{i=1}^{s} \theta_i^{y_i} \left(1 - \theta_i\right)^{1-y_i} p\left(\underline{\beta}\right), \tag{4.12}$$

$$\ell\left(y_i|\theta_i\right) \sim Be(\theta_i) \ , \quad \theta_i = \frac{\exp\left(\underline{x}_i^T \underline{\beta}\right)}{1 + \exp\left(\underline{x}_i^T \underline{\beta}\right)} \ , \qquad i = 1, \cdots, s;$$

where $\underline{\mathrm{x}}_i$ is a vector of four balance sheet indicators, including the intercept. We use a non informative improper prior distribution on $\underline{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$. This real data set has already been analyzed in Mira and Tenconi (2004), where a random effects model in the intercept was assumed.

We run an initial Markov chain using a canonical Metropolis Hastings of length 300 (after a burn in of 700) and over this initial sample we estimate the optimal parameters of the ϕ function for each j dimension and for $f_j(\underline{\beta}) = \beta_j$, $j = 1, \cdots, 5$

$$\phi^j\left(\underline{\beta}\right) = \left(\gamma_1^j\beta_1 + \gamma_2^j\beta_2 + \gamma_3^j\beta_3 + \gamma_4^j\beta_4 + \gamma_5^j\beta_5\right)\sqrt{\pi\left(\underline{\beta}|y,x\right)}\ , \qquad j = 1, \ldots, 5.$$

The optimization gave the estimates reported in Table 4.12.

Table 4.12: Credit risk model, estimated ϕ parameters

j	$\hat{\gamma}_1^j$	$\hat{\gamma}_2^j$	$\hat{\gamma}_3^j$	$\hat{\gamma}_4^j$	$\hat{\gamma}_5^j$
1	-0.09457704	-0.01333198	-0.05751499	-0.04640937	0.01208364
2	-0.01507528	-0.15816491	0.05934955	0.01612018	0.05508849
3	-0.05629736	0.06052546	-0.19269449	0.01473065	-0.03554821
4	-0.046095866	0.019266392	0.014117965	-0.101136218	0.003513808
5	0.0105972810	0.0597459884	-0.0345264631	0.0001133164	-0.0624642257

The estimated mean and variance for each model parameter are reported in Table 4.13.

Table 4.13: Credit risk model, initial $N = 300$ sample estimation

j	$\hat{\mu}_{f_j}$	$\hat{\mu}_{\tilde{f}_j}$	$\hat{\sigma}^2_{f_j}$	$\hat{\sigma}^2_{\tilde{f}_j}$	% variance reduction
1	-1.4761	-1.4339	0.0507	0.0015	97.04
2	-1.0337	-1.0138	0.0664	0.0018	97.28
3	-0.2858	-0.2830	0.0825	0.0043	94.78
4	-0.9687	-0.9746	0.0630	0.0007	98.88
5	0.8279	0.7756	0.0317	0.0012	96.21

Table 4.14: Credit Risk Model, $N = 6\,000$

j	$\hat{\mu}_{f_j}$	$\hat{\mu}_{\tilde{f}_j}$	$\hat{\sigma}^2_{f_j}$	$\hat{\sigma}^2_{\tilde{f}_j}$	% variance reduction
1	-1.4045	-1.4431	0.0435	0.0032	92.64
2	-0.9831	-1.0122	0.0795	0.0028	96.47
3	-0.2810	-0.3078	0.1081	0.0097	91.02
4	-0.9466	-0.9716	0.0523	0.0007	98.66
5	0.7737	0.7762	0.0323	0.0019	94.11

After performing a longer MCMC simulation of length 6000 (with a burn in of 1000 points), we obtain the results reported in Table 4.14, while after 600 000 MCMC iterations (with a burn in of 100 000) we achieve the results reported in Table 4.15. So with 50 000 iterations only, the zero-variance estimator (second column of Table 4.14) is close to the 500 000 standard MCMC estimator (first column of Table 4.15). This means that, to have results similar to the variance reduction technique we have introduced, one should run a 100 times longer Markov chain.

Table 4.15: Credit Risk Model, N=500 000

j	$\hat{\mu}_{f_j}$	$\hat{\sigma}^2_{f_j}$
1	-1.4354	0.0450
2	-1,0138	0.0820
3	-0.2941	0.0950
4	-0.9709	0.0510
5	0.7778	0.0310

4.7 Some tricks to speed up the simulation

When the operator defined in (4.8) is used, the function $\tilde{f}$ takes the form

$$\tilde{f}(x) = f(x) + \frac{(H\phi)(x)}{\sqrt{\pi(x)}}.$$

$H\phi(x)$ has to be computed on each point in the sample path, therefore when unavailable analytically, we must compute numerically the second order derivative that appears in the H operator. This is a time-consuming operation, however, by using the tricks we illustrate in the sequel, we are able to speed up the necessary computations.

We have suggested to use functions having the form $\phi(x) = P(x)\sqrt{\pi(x)}$ where $P(x)$ is a polynomial. As the following theorem shows, this choice reduces the calculation of $H\phi(x)$ to a first order derivative.

Theorem 1. *Assume*

$$\phi(x) = P(x)\sqrt{\pi(x)}$$

where $P(x)$ is a polynomial. Then

$$(H\phi)(x) = -\frac{1}{2}\sum_{i=1}^{d}\left[\sqrt{\pi(x)}\frac{\partial^2}{\partial x_i^2}P(x) + 2\left(\frac{\partial}{\partial x_i}P(x)\right)\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right]. \tag{4.13}$$

Before giving a proof of the proposition, some comments are required. Indeed, in (4.13) a second order derivative still appears but it is applied to a polynomial function and can thus be computed analytically. This theorem therefore reduces the computation to the first order derivative of the square root of the target. Also recall that the target has been already evaluated over all possible values x during the MCMC simulation: these values can thus be stored and re-used in the evaluation of $\tilde{f}$.

Proof. We must take the derivative of $\phi(x)$ twice with respect to a generic coordinate i:

$$\frac{\partial^2}{\partial x_i^2}\phi(x) = P(x)\frac{\partial^2}{\partial x_i^2}\sqrt{\pi(x)} + \sqrt{\pi(x)}\frac{\partial^2}{\partial x_i^2}P(x) + 2\left(\frac{\partial}{\partial x_i}P(x)\right)\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right).$$

Then

$$\begin{aligned}(H\phi)(x) &= \left(-\frac{1}{2}\sum_{i=1}^{d}\frac{\partial^2}{\partial x_i^2}\phi(x)\right) + \phi(x)V(x) \\ &= -\frac{1}{2}\sum_{i=1}^{d}\left[P(x)\frac{\partial^2}{\partial x_i^2}\sqrt{\pi(x)} + \sqrt{\pi(x)}\frac{\partial^2}{\partial x_i^2}P(x) + 2\left(\frac{\partial}{\partial x_i}P(x)\right)\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right] \\ &\qquad + \frac{1}{2}\sum_{i=1}^{d}P(x)\frac{\partial^2}{\partial x_i^2}\sqrt{\pi(x)} \\ &= -\frac{1}{2}\sum_{i=1}^{d}\left[\sqrt{\pi(x)}\frac{\partial^2}{\partial x_i^2}P(x) + 2\left(\frac{\partial}{\partial x_i}P(x)\right)\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right]\end{aligned}$$

□

Corollary 2. *Suppose that $\phi(x) = P(x)\sqrt{\pi(x)}$ and $P(x) = P(x_1, \ldots, x_d)$ is a first degree polynomial in $\mathbb{R}^d$, i.e.*

$$P(x) = \sum_{i=1}^{d} a_i x_i, \qquad a_i \in \mathbb{R}.$$

Then

$$(H\phi)(x) = -\sum_{i=1}^{d}\left[a_i\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right].$$

Proof. It follows from Theorem 1 by noting that

$$\frac{\partial}{\partial x_i}P(x) = a_i \quad , \quad \frac{\partial^2}{\partial x_i^2}P(x) = 0$$

□

Remark 3. With the previous theorem $\tilde{f}$ becomes

$$\tilde{f}(x) = f(x) - \frac{1}{\sqrt{\pi(x)}}\sum_{i=1}^{d}\left[a_i\left(\frac{\partial}{\partial x_i}\sqrt{\pi(x)}\right)\right].$$

A "logarithmic" version of the previous formula is also available:

$$\tilde{f}(x) = f(x) - \frac{1}{2}\sum_{i=1}^{d}\left[a_i\left(\frac{\partial}{\partial x_i}\ln\pi(x)\right)\right].$$

Corollary 4. *In the Credit Risk model of section 4.6*

$$(H\phi^j)(\underline{\beta}) = -\sum_{i=1}^{5}\left[\gamma_i^j\left(\frac{\partial}{\partial\beta_i}\sqrt{\pi\left(\underline{\beta}|y,x\right)}\right)\right].$$

Proof. It follows from Corollary 2. □

We conclude this section with an intuition useful to avoid the numerical optimization, necessary to find a ϕ close to the optimal one. In the credit risk model we noticed a great similarity of the matrix $\Gamma = \{\gamma_i^j\}_{i,j=1,\ldots,5}$, reported in table 4.12, to the matrix $-2\hat{\Sigma}$, where $\hat{\Sigma}$ is the estimated (from the MCMC output) covariance matrix of the target distribution.

This intuition is confirmed by the theoretical ϕ we obtained for the normal case when $f(\underline{x}) = x_i$, described in section 4.4.1 and 4.4.3, as the coefficients of the polynomial are the elements of the i-th row of the target covariance matrix, $\hat{\Sigma}_\pi$, multiplied by -2. In the next subsection we will use the above mentioned tricks to reduce the variance in a complex credit risk model that builds on the simplified one introduced in Section 4.6.

4.7.1 An extended credit risk model

It is commonly accepted that the amount of credit risk is different among sectors. In Mira and Tenconi (2004), a hierarchical logistic regression model was proposed with the purpose to capture the sector specific baseline risks and to obtain a best fit of the data. This model is reproposed here to investigate the zero-variance principle on a highly parametrized model. The data contains 7513 firms allocated among $j = 7$ sectors, firm specific balance sheet indicators, x_{ij}, and default events, y_{ij}, $i = 1, \ldots, n_j$. The model presents a hierarchical structure in the intercepts α_j, allowing for greater variation among sectors, overcoming at the same time overfitting issues:

$$\pi\left(\underline{\alpha},\underline{\beta},\mu_a,\sigma_\alpha^2|y,x\right)\propto\prod_{j=1}^{7}\prod_{i=1}^{n_j}\theta_{ij}^{y_{ij}}\left(1-\theta_{ij}\right)^{1-y_{ij}}\prod_{j=1}^{7}p\left(\alpha_j|\mu_\alpha,\sigma_\alpha^2\right)p\left(\mu_\alpha\right)p\left(\sigma_\alpha^2\right)p\left(\underline{\beta}\right)$$

$$\theta_{ij}=\frac{\exp\left(\alpha_j+\underline{x}_{ij}^T\underline{\beta}\right)}{1+\exp\left(\alpha_j+\underline{x}_{ij}^T\underline{\beta}\right)}$$

with

$$\begin{aligned}\underline{\beta}&\sim MN\left(0,\sigma^2=64I_4\right)\ ,\\ \alpha_j|\mu_\alpha,\sigma_\alpha^2&\sim N\left(\mu_\alpha,\sigma_\alpha^2\right)\ ,\\ \mu_\alpha&\sim N\left(0,\sigma^2=64\right)\ ,\\ \sigma_\alpha^2&\sim Ga(\alpha=\frac{9}{5},r=\frac{25}{9}).\end{aligned}$$

We focus on the functionals $f_k\left(\underline{\eta}\right)=\eta_k$ where $\underline{\eta}$ is the vector of all parameters, $\underline{\eta}=(\underline{\alpha},\underline{\beta},\mu_\alpha,\sigma_\alpha)$. The ϕ functions are choosen as in Section 4.6 and the following steps are taken:

1. A Markov chain of lenght 50 000 is run, discarding the first 10 000 steps as burn-in, to obtain a sample from $\pi\left(\underline{\eta}|y,x\right)$;

2. The target variance-covariance matrix of $\underline{\eta}$, Σ_π, is estimated along the chain simulated at step 1. This estimate, $\hat{\Sigma}$, is used to parametrize the ϕ functions to compute $\tilde{f}$ with the "fast version" of our algorithm, i.e.

$$\tilde{f}_k\left(\underline{\eta}\right)=f_k\left(\underline{\eta}\right)-2\hat{\Sigma}\times\nabla\ln\left(\pi\left(\underline{\eta}|y,x\right)\right);$$

3. We evaluate $\tilde{f}_k\left(\underline{\eta}\right)$ on a second MCMC sample of length 3 000.

The results, in terms of variance reduction, for all parameters of interest, are presented in Table 4.16 which shows an average variance reduction of 78,95%. If we exclude the hyper parameters, η_{12} and η_{13}, which are of little interest for credit risk estimation, the variance reduction goes up to 85,49%.

Table 4.16: Variance reduction for complex credit risk model

k	η_k	$\hat{\mu}_{f_k}$	$\hat{\mu}_{\tilde{f}_k}$	$\hat{\sigma}^2_{f_k}$	$\hat{\sigma}^2_{\tilde{f}_k}$	% variance reduction
1	$\eta_1 = \alpha_1$	-6.5122	-6.4548	1.8261	0.7731	57.67
2	$\eta_2 = \alpha_2$	-5.3699	-6.5122	0.1546	0.0166	89.24
3	$\eta_3 = \alpha_3$	-5.1055	-5.1296	0.0884	0.0113	87.21
4	$\eta_4 = \alpha_4$	-4.8881	-4.9179	0.0876	0.0086	90.16
5	$\eta_5 = \alpha_5$	-5.2247	-5.2446	0.0869	0.0112	87.14
6	$\eta_6 = \alpha_6$	-3.9072	-3.9560	0.1057	0.0170	83.91
7	$\eta_7 = \alpha_7$	-6.3274	-6.3539	0.1097	0.0131	88.06
8	$\eta_8 = \beta_1$	-0.0942	-0.0901	0.0032	0.0005	83.83
9	$\eta_9 = \beta_2$	-1.2452	-1.2649	0.0999	0.0078	92.23
10	$\eta_{10} = \beta_3$	-1.4105	-1.4295	0.0415	0.0049	88.26
11	$\eta_{11} = \beta_4$	0.0870	0.0868	0.0027	0.0002	92.73
12	$\eta_{12} = \mu_\alpha$	-5.2806	-5.3548	0.3840	0.1114	70.98
13	$\eta_{13} = \sigma_\alpha$	1.3738	1.4248	0.1883	0.1601	15.00

4.8 Rao-Blackwellization

Rao-Blackwellization (Casella and Robert (1996)), can be seen as a special case of the variance reduction technique proposed in this chapter. The Rao-Blackwellization idea is to replace $f(x^i)$ in $\hat{\mu}$ by a conditional expectation, $E_\pi[f(x^i)|h(x^i)]$, for some function h or to condition on the previous value of the chain thus using $E[f(x^i)|x^{i-1} = x]$ instead. Changing an expectation with a conditional expectation naturally reduces the variance of the resulting MCMC estimator. The functions $E_\pi[f(x^i)|h(x^i)]$ and $E[f(x^i)|x^{i-1} = x]$ can be considered as special instances of $\tilde{f}$ which do not minimize $\sigma_{\tilde{f}}$ but certainly reduce it. This suggests general guidelines that can be adopted to construct ϕ based on which we obtain $\tilde{f}$. In real applications, typically $E_\pi[f(x^i)|h(x^i)]$ or $E[f(x^i)|x^{i-1} = x]$ are not available in closed form, still, the researcher may have some intuition on the parametric form of such functions (or estimate them via pilot runs of the Markov chain). This intuition might aid the design of ϕ.

4.9 Conclusions

We have presented the advantages, in a statistical setting, of a general purpose variance reduction technique which has been originally suggested in the physics literature (Assaraf and Caffarel (1999)). Not only the zero-variance physics principle has been adapted to the statistical framework, but it has also been extended from Monte Carlo to Markov chain Monte Carlo simulation. The extent by which the variance of Monte Carlo and MCMC estimators can be reduced, is illustrated via some toy examples and a complex credit risk Bayesian model, fitted to a real dataset. The overall performance of the proposed technique is quite astonishing: in simple cases zero variance is indeed achieved, while in more complicated models, when the exact solution to the fundamental equation cannot be obtained analytically, a variance reduction between 80% and 95% is obtained. Moreover, useful tricks are proposed to dramatically speed up the application of the method to statistical modelling. Connections with the Rao-Blackwellization principle known in the MCMC literature are explored and exploited to better apply the zero-variance technique in a Bayesian setting.

4.10 Appendix

In this appendix we give some explanation how the ϕ function can be identified for simple cases.

Univariate Case

The aim is to find a general form for the solution to the *fundamental equation* in (4.4). For the univariate case, the continuous version of the the fundamental equation is the following

$$H(x)\phi(x) = -\pi(x)^{1/2}[f(x) - \mu_f] \tag{4.14}$$

with $\mu_f = \frac{\int_{-\infty}^{+\infty} f(x)\pi(x)\mathrm{d}x}{\int_{-\infty}^{+\infty} \pi(x)\mathrm{d}x}$. The solution of the previous equation can be found in the form

$$\phi = \pi(x)^{1/2} \cdot P(x)$$

where $P(x)$ is an integrable function, but not necessarily a polinomial. Substituting in (4.14) the previous expression for ϕ and using the operator $H(x)$, the fundamental equation we obtain is

$$\frac{\mathrm{d}^2 P(x)}{\mathrm{d}x^2} + \pi(x)^{-1}\frac{\mathrm{d}\pi(x)}{\mathrm{d}x}\frac{\mathrm{d}P(x)}{\mathrm{d}x} = 2[f(x) - \mu_f]. \tag{4.15}$$

By setting

- $\pi(x)^{-1}\frac{\mathrm{d}\pi(x)}{\mathrm{d}x} = A(x)$
- $g(x) = 2[f(x) - \mu_f]$

then (4.15) becomes

$$\frac{\mathrm{d}^2 P(x)}{\mathrm{d}x^2} + A(x)\frac{\mathrm{d}P(x)}{\mathrm{d}x} = g(x) \tag{4.16}$$

which is a linear differential equation, not homogeneous with variable coefficients. To solve it, we perform a variable transformation, by setting $y = \frac{\mathrm{d}P(x)}{\mathrm{d}x}$, so that (4.16) becomes

$$\frac{\mathrm{d}y}{\mathrm{d}x} + A(x)y = g(x). \tag{4.17}$$

The integral of y solution is the P solution we are looking for. The general solution for (4.17), available from literature, is

$$y_{\text{gen}} = \underbrace{c_1 e^{-\int A(x)\mathrm{d}x}}_{\text{homogeneous sol.}} + \underbrace{e^{-\int A(x)\mathrm{d}x} \int g(x) e^{+\int A(x)\mathrm{d}x}}_{\text{particular solution}}, \tag{4.18}$$

but

$$\int A(x)\mathrm{d}x = \int \frac{1}{\pi(x)} \frac{\mathrm{d}\pi(x)}{\mathrm{d}x} \mathrm{d}x = \ln(\pi(x)),$$

then (4.18) becomes

$$y_{\text{gen}} = \underbrace{c_1 \frac{1}{\pi(x)}}_{\text{homogeneous sol.}} + \underbrace{\frac{1}{\pi(x)} \int g(x)\pi(x)\mathrm{d}x}_{\text{particular sol.}}. \tag{4.19}$$

Substituting the definition of $g(x)$ and integrating (4.19) we obtain the solution for $P(x)$:

$$P(x)_{\text{gen}} = \underbrace{\int c_1 \frac{1}{\pi(x)}}_{\text{homogeneous sol.}} + \underbrace{\int \frac{2}{\pi(x)} \int [f(x') - \mu_f]\pi(x')\mathrm{d}x'\mathrm{d}x}_{\text{particular sol.}}. \tag{4.20}$$

We are interested to a solution $\phi(x) = \pi(x)^{1/2}P(x)$ which can be normalized (that's to say $\int_{-\infty}^{+\infty} |\phi(x)|^2 \mathrm{d}x < \infty$) but not necessarily the most general. It is sufficient for ϕ to verify (4.14), or that $P(x)$ satisfies (4.16). It is sufficient, therefore, to consider a particular solution $P(x)$, or the second addendum of (4.20). Then the solution for (4.14) is given by the following

$$\phi(x) = \pi(x)^{1/2} \int \frac{2}{\pi(x)} \int [f(x') - \mu_f]\pi(x')\mathrm{d}x'\mathrm{d}x \tag{4.21}$$

with

$$\frac{\int_{-\infty}^{+\infty} f(x)\pi(x)\mathrm{d}x}{\int_{-\infty}^{+\infty} \pi(x)\mathrm{d}x}.$$

Using distributions $\pi(x)$ with known parameters and simple $f(x)$ functions, such as polinomials, it is possibile to find an exact solution, such as the ones found in the previous sections for the normal and t-student distribution.

As an example for (4.21), suppose to have

- $\pi(x) = \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = N(x)$
- $f(x) = x$
- $\mu_f = \mu$

then, by substituting them into (4.21) we obtain

$$\begin{aligned}
\phi\left(x\right) &= N(x)^{1/2} \int 2N(x)^{-1} \int [x' - \mu] \underbrace{\exp\left[-\frac{(x'-\mu)^2}{2\sigma^2}\right]}_{N(x')} \mathrm{d}x'\mathrm{d}x \\
&= N(x)^{1/2} \int 2N(x)^{-1} \int (-\sigma^2)\frac{\mathrm{d}}{\mathrm{d}x'} \exp\left[-\frac{(x'-\mu)^2}{2\sigma^2}\right] \mathrm{d}x'\mathrm{d}x \\
&= N(x)^{1/2} 2(-\sigma^2) \int N(x)^{-1}N(x)\mathrm{d}x \\
&= -2\sigma^2 N(x)^{-1/2} \int \mathrm{d}x \\
&= -2\sigma^2 x N(x)^{1/2}.
\end{aligned}$$

Multivariate Case

For the d-dimensional case the equation (4.15) becomes a partial differential equation (PDE)

$$\sum_{i}^{d} \frac{\partial^2 P(\underline{x})}{\partial x_i^2} + \pi(\underline{x})^{-1} \sum_{i}^{d} \frac{\partial \pi(\underline{x})}{\partial x_i} \frac{\partial P(\underline{x})}{\partial x_i} = 2[f(\underline{x}) - \mu_f]. \tag{4.22}$$

Considering the particular case with π gaussian, we can disinguish two cases, depending on the independence of variables. If they are independent, the functional form of the multivariate normal and the one of the homogeneous differential operator associated to (4.22) allows this equation to be separable. Given that, the homogeneous PDE associated to (4.22) transforms into a system of ordinary linear differential equations, of the same form, for each variable. Once available the homogeneous solutions, the particular solutions for each variable can be obtained as previously explained, for any form assuming the function $g(\underline{x}) \equiv 2[f(\underline{x}) - \mu_f]$. On the contrary, for dependent variables, as for the bivariate normal case in section 4.4.3, the PDE it is not easly separable and the solution requires a specific treatment for each case.

References

R. Assaraf and M. Caffarel. Zero-variance zero-bias principle for observables in quantum monte carlo: Application to forces. *The Journal of Chemical Physics*, 119, 20:10536–10552, 2003.

R. Assaraf and M. Caffarel. Zero-Variance principle for Monte Carlo algorithms. *Physical Review letters*, 83, 23:4682–4685, 1999.

R. Casella and C. P. Robert. Rao-Blackwellization of sampling schemes. *Biometrika*, 83, 1: 81–94, 1996.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman & Hall, 1996.

P. J. Green and A. Mira. Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika*, 88:1035–1053, 2001.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

J. S. Liu. Peskun theorem and a modified discrete-state Gibbs sampler. *Biometrika*, 83: 681–682, 1996.

N. Metropolis, A. E. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

A. Mira. Ordering and improving the performance of Monte Carlo Markov chains. *Statistical Science*, 16:340–350, 2001.

A. Mira and C. J. Geyer. On reversible Markov chains. *Fields Inst. Communic.: Monte Carlo Methods*, 26:93–108, 2000.

A. Mira and P. Tenconi. Bayesian estimate via credit risk via mcmc with delayed rejection. *Stochastic Analysis, Random Fields and Applications IV in Progress in Probability*, 26: 277–291, 2004.

P. H. Peskun. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

G. O. Roberts. Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996b.

L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22: 1701–1762, 1994.

L. Tierney. A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, 8:1–9, 1998.

L. Tierney and A. Mira. Some adaptive Monte Carlo methods for bayesian inference. *Statistics in Medicine*, 18:2507–2515, 1999.

Chapter 5

Weighted Likelihood Equations for Resistant Credit Risk Modelling

5.1 Introduction

The presence of some extreme observations may ruin statistical estimates, the body of statistical robustness has investigated this matter widely. For example to overcome this problem, Huber (1964) proposed the *minimax approach*, while Hampel (Hampel (1968, 1974)) the one based on *influence functions.* In this chapter we consider the *weighted likelihood equations* approach, described in Markatou et al. (1997). The idea is to downweight surprising observations, these are defined in terms of *Pearson residuals*, we propose a way to produce these residuals for *generalized linear models*, in order to extend to this class of models the methodology. As a special case we deal with logistic regression, as it is widely used for default risk prediction, our aim is to downweight aberrant observations on the design space only if they produce predictions that contrast with the whole model, this property is verified through artificially generated data. The method is finally applied to real data to produce resistant credit risk estimates.

5.2 The Methodology

5.2.1 Weighted Likelihood Equation Estimators

Given an observed set of data $\{X_1, X_2, ..., X_n\}$, assumed to be drawn from a sample distribution $m_\beta(x)$, we define $u(x,\beta) = \nabla_\beta \ln(m_\beta(x))$ the maximum likelihood score function, being ∇_β the gradient with respect to β . Then the *maximum likelihood estimator* for β is the solution for $\sum_{i=1}^n u(x_i, \beta) = 0$. In the *weighted likelihood equation* approach, instead, an estimator for β is obtained as a solution to the following equation

$$\sum_{i=1}^{n} w\left(X_i, M_\beta, \hat{F}\right) \nabla_\beta \ln\left(m_\beta\left(X_i\right)\right) = 0 \tag{5.1}$$

where $w(X_i, M_\beta, \hat{F})$ is a *weight function* taking values between 0 and 1 and describing the inconsistency of the c.d.f. of the chosen theoretical model M_β with respect to the empirical distribution function $\hat{F}$ of observed data. The concept of *outlier* used to describe the discrepancy between M_β and $\hat{F}$ is the one of *surprising observation* (Lindsay (1994)), that is a value that occurs in a small probability region under the model m_β. In Markatou et al. (1997), as the observations are countable under the model, to define an outlier the concept of *Pearson residual* is used

$$\delta\left(x\right) = \frac{d\left(x\right)}{m_\beta\left(x\right)} - 1 \tag{5.2}$$

where $d(x)$ is the number of observations in a given cell, while $m_\beta(x)$ is the number obtained under the model. For a continuous model the previous definition is not appropriate, as $d(x)$ and $m_\beta(x)$ are not comparable. Basu and Lindsay (1994) addressed this defining

$$\delta\left(x\right) = \frac{f^*\left(x\right)}{m^*_\beta\left(x\right)} - 1, \tag{5.3}$$

where

$$f^*\left(x\right) = \int k\left(x, t, h\right) d\hat{F}\left(t\right), \tag{5.4}$$

is the smoothed empirical distribution function, obtained using k as a smoothing kernel with h as a bandwidth, while

$$m_{\beta}^{*}(x) = \int k(x,t,h)\, dM_{\beta}(t) \tag{5.5}$$

is the smoothed model density, obtained applying the same kernel used to obtain $f^{*}(x)$. By smoothing both the data and the model, using the same kernel, if the model is correctly specified, guarantees for a fixed x, that $\delta(x)$ converges in probability to zero (see Agostinelli (1997)). The method of weighted likelihood attempts to downweight large Pearson residuals. This is achieved through the following weighting formula

$$w\left(x, M_{\beta}, \hat{F}\right) = \frac{A(\delta(x)) + 1}{\delta(x) + 1}, \tag{5.6}$$

where A is a function defined in $[-1, +\infty)$, having the properties that $A(0) = 0$, $A^{'}(0) = 1$, $A^{'}(\delta) > 0$ and twice differentiable. In particular Lindsay (1994) chooses A to be a *residual adjustment* function (RAF) wich guarantees a link with minimum distance methods. It has been shown in Lindsay (1994) that the behaviour of the RAF in the tails guides the robustness properties of the corresponding estimators. For example if $A(\delta) \sim \sqrt{\delta}$ for $\delta \to +\infty$ gives a 50% breakdown point. It's also possibile to truncate weights and introduce an extra parameter, k, controlling for the degree of robustness

$$w\left(x, M_{\beta}, \hat{F}\right) = \left\{\min\left\{\frac{\max\left[A(\delta) + 1; 0\right]}{\delta(x) + 1}; 1\right\}\right\}^{k} \quad , \quad k \geq 0 \tag{5.7}$$

the estimates are still efficient, for a correct model, as $A(\delta)$ is still increasing. An alternative way to build weights, adopted in Markatou et al. (1998), is

$$w\left(x, M_{\beta}, \hat{F}\right) = \left\{1 - \frac{\delta(x)^{2}}{(\delta(x) + 2)^{2}}\right\}^{k} \quad , \quad k \geq 0 \tag{5.8}$$

The presented methodology has been introduced for a logistic regression with observations grouped in cells in Markatou et al. (1997) and then extended to normally distributed obervations in Markatou et al. (1998) and Agostinelli (1997). In the following subsection we will introduce a convenient way to define Pearson residuals in order to extend the framework to the class of generalized linear models.

5.2.2 Residuals for Generalized Linear Models

In the context of generalized linear models (GLMs) we assume m_β to belong to the exponential class of distributions. Let $\{y_1, \dots, y_n\}$ be response observations which are related to their corresponding vectors of covariates $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. The probability density for y_i is assumed to be

$$f\left(y_i, \theta_i, \psi\right) = \exp\left\{\frac{\zeta_i\left[y_i\theta_i - \kappa\left(\theta_i\right)\right]}{\psi} + c_i\left(y_i, \psi\right)\right\} \tag{5.9}$$

where c_i and κ are known functions, ζ_i is a known constant, while the following relations hold $\mu_i = E(y_i) = \kappa^{'}(\theta_i)$, $V(\mu_i) = \kappa^{''}(\theta_i)$ and $V(y_i) = \psi V(\mu_i)$. In the class of generalized linear models $g(\mu_i) = \mathbf{x}_i^{'}\beta$, where g is called the *link function.* The dispersion parameter ψ is usually common across observed data and for our porposes it is assumed constant.

For GLMs the residuals commonly used are the Pearson residuals

$$r_{p,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V\left(\hat{\mu}_i\right)}}$$

and the deviance residuals, where assuming $t(y, \mu) = y\theta - \kappa(\theta)$, are defined as

$$r_{d,i} = 2\{t(y_i, y_i) - t(y_i, \hat{\mu}_i)\}^{1/2}\text{sign}(y_i - \hat{\mu}_i).$$

Both residuals as argued in Pierce and Schafer (1986) tend to normality for $\psi/\mu \to 0$ with rate $O(\psi^{1/2})$ for Pearson residuals and rate $O(\psi)$ for deviance residuals and are exactly normal for the normal linear model. Deviance residuals are exactly normal for inverse gaussian responses too. In other cases, when ψ/μ is large, they are not guaranteed to converge to the normal distribution, moreover, deviance residuals do not have zero means and constant variances.

Dunn and Smith (Dunn and Smyth (1996)) build *randomized quantile residuals* having the property to be normally distributed, for exponential family distribution responses. Indicating with $F(y, \mu, \psi)$ the cumulative distribution function for the model, for continuous responses quantile residuals are defined as

$$r_{q,i} = \Phi^{-1} \left\{ F \left(y_i, \hat{\mu}_i, \hat{\psi} \right) \right\} \tag{5.10}$$

where Φ is the standard normal cumulative distribution function. If a consistent estimator for β and ϕ is used, then $r_{q,i}$ converges to a standard normal. If responses are discretely distributed, then quantile residuals are defined as

$$r_{q,i} = \Phi^{-1}(u_i) \quad , \quad u_i \sim U(a_i, b_i) \tag{5.11}$$

where $a_i = lim_{y \to y_i^-} F(y_i, \hat{\mu}_i, \hat{\psi})$ and $b_i = F(y_i, \hat{\mu}_i, \hat{\psi})$. In Dunn and Smyth (1996) these residuals are used as a diagnostic tool and are applied to some generalized linear models, in order to check graphically any inconsistency with data.

By using this kind of residuals it is possibile to mutuate the continuous version of Pearson residuals as defined in (5.3), in this case we set $m^*_\beta(r_{q,i}) \sim N(0, 1)$, while computing f^* on the empirical cumulative distribution function of generated quantile residuals. Using this technique it is possibile to apply the weighted likelihood equations methodology to other generalized linear models.

5.2.3 Weighting Scheme

Assuming a dependency model with a set of stochastic regressors

$$m(y_i, x_i) = m_\beta(y_i | x_i) \, m(x_i)$$

we may want to extend the weighting scheme to take into account aberrant observation on the design space too. Given the errors z_i as defined in (5.11), we assume that

$$m\left(y_i, x_i\right) = m\left(z_i\right) m\left(x_i\right),$$

we define joint weights as

$$\delta\left(z_i, x_i\right) = \frac{f^*\left(z_i, x_i\right)}{m^*\left(z_i, x_i; \theta\right)} - 1$$

being θ the parametric space. The above expression downweights high leverage points, whether they are related to big prediction errors or not. So, seeking Agostinelli (1997), we define

$$w\left(\delta_z, \delta_x\right) = w\left(\delta_z\left(z_i, \theta_z\right)\right) w_x\left(\delta_z\left(z_i, \theta_z\right), \delta_x\left(x_i, \theta_x\right)\right),$$

while

$$w_x\left(\delta_z, \delta_x\right) = w\left(\delta_x\left(x_i, \theta_x\right)\right) + \left(1 - w\left(\delta_x\left(x_i, \theta_x\right)\right)\right) \psi\left(\delta_z\left(z_i, \theta_z\right)\right);$$

ψ is a function such that $\lim_{\delta_z \to 0} \psi\left(\delta\left(z\right)\right) = 1$, while $\lim_{\delta_z \to \infty} \psi\left(\delta\left(z\right)\right) = 0$. We set $\psi\left(\delta_z\right) = w\left(\delta_z\right)$, in Agostinelli (1997) it was suggested a logistic type function. This way of weighting observations allow to underweight leverage points whose residuals are not coherent with the assumed model, while high leverage points associated to low residuals are not downweighted, this (Agostinelli (1997)) avoids a decrease of estimator efficiency.

5.2.4 Robustness Properties

Given the *wle* estimator at hand

$$\int w\left(x, M_\beta, F\right) \nabla_\beta \left(\ln M_\beta\left(x\right)\right) dF\left(x\right) = 0 \tag{5.12}$$

we can study its robustness properties through the *influence function* and its *breakdown properties.* In the previous equation if $F = M_\beta$, then the estimator is Fisher consistent as the true value β is contained among the solutions of the equation (5.1). Suppose to have the following contaminated distribution

$$F_{\varepsilon}\left(x\right)=\left(1-\varepsilon\right)F\left(x\right)+\varepsilon\Delta_{y}\left(x\right)\quad,\quad\varepsilon\in\left(0,1\right)$$

where Δ_y is a distribution having all its mass concentrated in y. The *influence function* for an estimator $T\left(F\right)$ is given by

$$\mathrm{IF}\left(x,T,F\right)=\lim_{\varepsilon\to0^{+}}\left\{\frac{\left[T\left(\left(1-\varepsilon\right)F+\varepsilon\Delta_{x}\right)-T\left(F\right)\right]}{\varepsilon}\right\}. \tag{5.13}$$

For the *wle* estimator the influence function (Markatou et al. (1997)), is given by

$$\frac{\partial}{\partial\varepsilon}\beta_{\varepsilon}|_{\varepsilon=0}=A\left(F\right)B\left(y,F\right) \tag{5.14}$$

where

$$A\left(F\right)\ =\ \left\{\int w^{'}\left(\delta\left(t\right)\right)u\left(t,\beta_{0}\right)u^{T}\left(t,\beta_{0}\right)\left(\delta\left(t\right)+1\right)dF\left(t\right)+\int w\left(\delta\left(t\right)\right)\left(-\nabla u\left(t,\beta_{0}\right)dF\left(t\right)\right)\right\}^{-1}$$

$$B\left(y,F\right)=w\left(\delta\left(y\right)\right)u\left(y,\beta_{0}\right)+w^{'}\left(\delta\left(y\right)\right)u\left(y,\beta_{0}\right)\left(\delta\left(y\right)+1\right)-\int w^{'}\left(\delta\left(t\right)\right)f\left(t\right)\frac{u\left(t,\beta_{0}\right)}{m_{\beta_{0}}}dF\left(t\right).$$

The influence function if $M=M_{\beta_0}$ is the same influence function of a maximum likelihood estimator

$$\frac{\partial}{\partial\varepsilon}\beta_{\varepsilon}|_{\varepsilon=0}=\left\{\int-\nabla_{\beta}u\left(t,\beta_{0}\right)dM_{\beta}\left(t\right)\right\}^{-1}u\left(y,\beta_{0}\right) \tag{5.15}$$

it is unbounded, however (Lindsay (1994)) it can be misleading stopping a first order level analysis. However the influence function is useful to derive the asymptotics properties of the *wle* estimator. We have in fact that the asymptotic variance of $\sqrt{n}T\left(F\right)$ is given by

$$\Sigma_{\beta}=A\left(F\right)E\left\{B\left(Y,F\right)B\left(Y,F\right)^{T}\right\}A^{T}\left(F\right), \tag{5.16}$$

this can be estimated with

$$\hat{\Sigma}_{\beta} = A\left(\hat{F}\right)\left\{\frac{1}{n}\sum\left[B\left(X_i\hat{F}\right)B\left(X_i,\hat{F}\right)^T\right]\right\}A^T\left(\hat{F}\right). \tag{5.17}$$

5.2.5 Iterative Reweighted Least Squares Algorithm

To solve (5.1) an iterative algorithm is required, one can use a modified version of the well known *iterative reweighted least squares*, used to get an estimate for generalized linear models. In our setting we deal with a weighted version of the score equations

$$\frac{\partial\tilde{\ell}}{\partial\beta_j} = w_i\frac{\partial\ell_i}{\partial\beta_j} = w_i\frac{\partial\ell_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}$$

where

$$\begin{aligned}
\frac{\partial\ell_i}{\partial\theta_i} &= \frac{y_i-\kappa'\left(\theta_i\right)}{\psi/\zeta_i} = \frac{y_i-\mu_i}{\psi/\zeta_i}\\
\frac{\partial\mu_i}{\partial\theta_i} &= \kappa''\left(\theta_i\right) = \frac{\omega_i\mathrm{Var}\left\{Y_i\right\}}{\psi}\\
\frac{\partial\eta_i}{\partial\beta_j} &= x_{ij}
\end{aligned}$$

so we have

$$\frac{\partial\tilde{\ell}}{\partial\beta_j} = w_i\frac{y_i-\mu_i}{\psi/\zeta_i}\frac{\psi/\zeta_i}{\mathrm{Var}\left\{Y_i\right\}}\frac{\partial\mu_i}{\partial\eta_i}x_{ij}$$

and the weighted likelihood equations are

$$\sum_{i=1}^{n} w_i\frac{\left(y_i-\mu_i\right)x_{ij}}{\mathrm{Var}\ \left\{Y_i\right\}}\frac{\partial\mu_i}{\partial\eta_i} = 0\ \ ,\ \ j=1,\ldots,p.$$

These equations can be solved using a Newton-Raphson like method, by which

$$\beta^{(s+1)} = \beta^{(s)} + \left[I\left(\beta^{(s)}\right)\right]^{-1}u^{(s)} \tag{5.18}$$

where $u^{(s)}$ are the likelihood equations at iteration s and $I\left(\beta^{(s)}\right)$ is the negative of the expected Fisher information matrix

$$\begin{aligned} -E\left\{\frac{\partial^2 \ell_i}{\partial\beta_j\partial\beta_k}\right\} &= E\left\{\left(w_i\frac{(y_i-\mu_i)\,x_{ij}}{\mathrm{Var}\left\{Y_i\right\}}\frac{\partial\mu_i}{\partial\eta_i}\right)\left(w_i\frac{(y_i-\mu_i)\,x_{ij}}{\mathrm{Var}\left\{Y_i\right\}}\frac{\partial\mu_i}{\partial\eta_i}\right)\right\} \\ &= w_i^2\frac{x_{ij}x_{ik}}{\mathrm{Var}\left\{Y_i\right\}}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2, \end{aligned}$$

this in matrix notation can be expressed as

$$I(\beta) = X^T\tilde{\Lambda}X \tag{5.19}$$

where

$$\tilde{\Lambda} = \begin{bmatrix} \tilde{\lambda}_{11} & 0 & \cdots & 0 \\ 0 & \tilde{\lambda}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\lambda}_n \end{bmatrix}$$

with elements

$$\tilde{\lambda}_i = w_i^2\frac{1}{\mathrm{Var}\left\{Y_i\right\}}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$$

Regarding (5.18) it is the same to write

$$I\left(\beta^{(s)}\right)\beta^{(s+1)} = I\left(\beta^{(s)}\right)\beta^{(s)} + u^{(s)} \tag{5.20}$$

where for the right member

$$I\left(\beta^{(s)}\right)\beta^{(s)} + u^{(s)} = X^T\tilde{\Lambda}^{(s)}z^{(s)}$$

the right member of this equation is

$$\sum_j \left[\sum_i w_i^2 \frac{x_{ih} x_{ij}}{\text{Var}\left\{Y_i\right\}} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right] \beta_j^{(s)} + \sum_i w_i \frac{\left(y_i - \mu_i^{(s)} \right) x_{ih}}{\text{Var}\left\{Y_i\right\}} \left(\frac{\partial \mu_i}{\partial \eta_i} \right),$$

this is a $z^{(s)}$ vector whose $i-th$ element is

$$z_i^{(s)} = \eta_i^{(s)} + \left(y_i - \mu_i^{(s)} \right) / \left(\frac{\partial \mu_i}{\partial \eta_i} w_i \right).$$

Given that, we can write (5.20)

$$\left(X^T \tilde{\Lambda}^{(s)} X \right) \beta^{(s+1)} = X^T \tilde{\Lambda}^{(s)} z^{(s)},$$

then we solve for $\beta^{(s+1)}$ and we get

$$\beta^{(s+1)} = \left(X^T \tilde{\Lambda}^{(s)} X \right)^{-1} X^T \tilde{\Lambda}^{(s)} z^{(s)}.$$

The recursive evaluation of this formula is known as *iterated reweighted least squares* and converges to the solution of equation (5.1). At each iteration w_i weights are computed and treated as constant values.

5.2.6 Newton-Raphson for Logistic Regression

The Newton-Raphson algorithm is a common tool used to maximize functions. It requires the computation of the gradient vector and hessian matrix, if possibile they should be derived analitically, otherwise numerical derivatives are used. As we are dealing with default events and their statistical modelling, we derive the Newton-Rapshon steps for the weighted logistic regression, giving the expressions for the gradient and hessian matrix. The weights w_i are obtained as previously .

The logistic regression log-likelihood is given by

$$\ell\left(\beta\right)=\sum_{i=1}^{n}\left\{w_i\left[y_i\ln\left(\frac{\exp\left(\mathbf{x}'{}_i\beta\right)}{1+\exp\left(\mathbf{x}_i'\beta\right)}\right)+\left(1-y_i\right)\ln\left(1-\frac{\exp\left(\mathbf{x}'{}_i\beta\right)}{1+\exp\left(\mathbf{x}_i'\beta\right)}\right)\right]\right\}, \tag{5.21}$$

indicating with

$$\theta\left(\mathbf{x}_i'\beta\right)=\frac{\exp\left(\mathbf{x}'{}_i\beta\right)}{1+\exp\left(\mathbf{x}_i'\beta\right)}$$

we rewrite (5.21) as

$$\ell\left(\beta\right)=\sum_{i=1}^{n}\left\{w_i\left[y_i\ln\theta\left(\mathbf{x}_i'\beta\right)+\left(1-y_i\right)\ln\theta\left(-\mathbf{x}_i'\beta\right)\right]\right\}.$$

The Newton-Raphson algorithm is based on recursions

$$\beta_t=\beta_{t-1}+\left[\frac{\partial^2\ell}{\partial\beta\partial\beta'}\right]^{-1}\times\frac{\partial\ell}{\partial\beta}. \tag{5.22}$$

For the case at hand we have that the gradient vector is

$$\begin{aligned}
\frac{\partial\ell\left(\beta\right)}{\beta} &= \sum_{i=1}^{n}\left\{w_i\left[y_i\mathbf{x}_i-\frac{\exp\left(\mathbf{x}_i'\beta\right)\mathbf{x}_i}{1+\exp\left(\mathbf{x}_i'\beta\right)}\right]\right\}\\
&= \sum_{i=1}^{n}\left\{w_i\left[y_i\mathbf{x}_i-\theta\left(\mathbf{x}_i'\beta\right)\mathbf{x}_i\right]\right\}\\
&= \sum_{i=1}^{n}\left\{w_i\left[\mathbf{x}_i\left(y_i-\theta\left(\mathbf{x}_i'\beta\right)\right)\right]\right\}\\
&= \sum_{i=1}^{n}\left\{w_i\mathbf{x}_i\left(y_i-\hat{y}_i\right)\right\}\\
&= \mathbf{X}'\mathrm{diag}\left(\mathbf{w}\right)\left(\mathbf{y}-\hat{\mathbf{y}}\right)
\end{aligned} \tag{5.23}$$

while the hessian matrix is

$$\begin{aligned}\frac{\partial^2 \ell\left(\beta\right)}{\partial\beta\partial\beta'} &= \frac{\partial}{\partial\beta}\left[\sum_{i=1}^{n}\left\{w_i\mathbf{x}_i\left(y_i-\hat{y}_i\right)\right\}\right]\\ &= -\sum_{i=1}^{n}\left\{w_i\mathbf{x}_i\frac{\partial}{\partial\beta}\theta\left(\mathbf{x}_i^{'}\beta\right)\right\};\end{aligned}$$

given that

$$\begin{aligned}\frac{\partial}{\partial\beta}\theta\left(\mathbf{x}_i^{'}\beta\right) &= \frac{\partial}{\partial\beta}\left[\frac{\exp\left(\mathbf{x}_i^{'}\beta\right)}{1+\exp\left(\mathbf{x}_i^{'}\beta\right)}\right]\\ &= \frac{\mathbf{x}_i\exp\left(\mathbf{x}_i^{'}\beta\right)}{\left[1+\exp\left(\mathbf{x}_i^{'}\beta\right)\right]^2}\\ &= \frac{\exp\left(\mathbf{x}_i^{'}\beta\right)}{1+\exp\left(\mathbf{x}_i^{'}\beta\right)}\frac{1}{1+\exp\left(\mathbf{x}_i^{'}\beta\right)}\mathbf{x}_i^{'}\\ &= \frac{\exp\left(\mathbf{x}_i^{'}\beta\right)}{1+\exp\left(\mathbf{x}_i^{'}\beta\right)}\left(1-\frac{\exp\left(\mathbf{x}_i^{'}\beta\right)}{1+\exp\left(\mathbf{x}_i^{'}\beta\right)}\right)\mathbf{x}_i^{'}\\ & \hat{y}_i\left(1-\hat{y}_i\right)\mathbf{x}_i^{'}\end{aligned}$$

we have that the following expression for the hessian

$$\begin{aligned}\frac{\partial^2 \ell\left(\beta\right)}{\partial\beta\partial\beta'} &= -\sum_{i=1}^{n}\left\{w_i\mathbf{x}_i\hat{y}_i\left(1-\hat{y}_i\right)\mathbf{x}_i^{'}\right\} \qquad (5.24)\\ &= \mathbf{X}^{'}\left[\operatorname{diag}\left(\hat{\mathbf{y}}\left(1-\hat{\mathbf{y}}\right)^{'}\right)\operatorname{diag}\left(I_n\mathbf{w}\right)\right]\mathbf{X}.\end{aligned}$$

Therefore, the weighted logistic regression Newton-Raphson step is given by

$$\beta_t = \beta_{t-1} + \left\{\mathbf{X}^{'}\left[\operatorname{diag}\left(\hat{\mathbf{y}}\left(1-\hat{\mathbf{y}}\right)^{'}\right)\operatorname{diag}\left(I_n\mathbf{w}\right)\right]\mathbf{X}\right\}^{-1}\mathbf{X}^{'}\operatorname{diag}\left(\mathbf{w}\right)\left(\mathbf{y}-\hat{\mathbf{y}}\right). \qquad (5.25)$$

	$\alpha = 0$	$\alpha = 0,1$	$\alpha = 0,2$	$\alpha = 0,3$	$\alpha = 0,4$	$\alpha = 0,5$
Max. Lik.	$1,96$	$3,13$	$4,47$	$5,99$	$7,35$	$8,72$
wle	$1,96$	$1,92$	$1,95$	$2,23$	$\{1,97;8,94;14,72\}$	$\{2,42;6,67;13,52\}$
wle-qr	$1,96$	$1,98$	$1,96$	$3,39$	$\{2,27;9,53;11,90\}$	$\{3,46;5,68;13,63\}$

Table 5.1: Poisson roots

5.3 Applications

We test the method proposed in the previous sections, while also making a comparison with other suggested robust procedures. We will refer to our method with *wle-qr,* i.e. *weighted likelihood equations* through *quantile residuals*, while with *wle* we will refer to the method proposed in Markatou et al. (1997). Logit is a short name we use to indicate a standard logistic regression.

5.3.1 Poisson Data

As a first example in Table 5.1, we compare the results obtained in Markatou et al. (1997), by simulating contaminated poisson data. We generate $N = 100$ data from a $Po(\lambda = 2)$ and subsequently contaminate it as following: $(1 - \alpha)Po(\lambda = 2) + \alpha Po(\lambda = 15)$ with $\alpha \in \{0,1; 0,2; 0,3; 0,4; 0,5\}$. We choose

$$A\left(\delta\right) = 2\left\{\sqrt{\delta + 1} - 1\right\},$$

define weights as in (5.7) using $k = 1$ and, regarding the bandwidth, we set $h = 0,5$. Results we obtain are similar to those obtained by Markatou et al. (1997), we experience the arising of multiple roots as the degree of contamination increases. The presence of multiple roots requires therefore care during estimation, Markatou et al. (1998) propose a method called *bootstrap root search* to deal with this kind of matter.

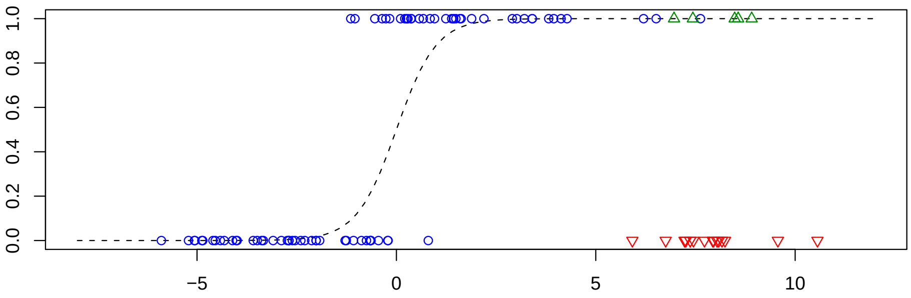

Figure 5.1: Contrasting and non contrasting leverages

5.3.2 Logistic Regression

Similarly as in Agostinelli (1997), we generate a set of $n = 100$ data from a distribution $X \sim N(\mu = 0, \sigma = 3)$ and generate Bernoulli distributed data $y_i \sim Be(\theta_i)$, with $\theta_i = \exp(\alpha + x_i\beta) / (1 + \exp(\alpha + x_i\beta))$, so that to have $\beta = 2$. Then we contaminate a ε proportion of such data with $X^{(c)} \sim N(\mu = 8, \sigma = 1)$, by replacing the last $\varepsilon \times n$ observations, for this data we produce a ν proportion of dependent data not coherent with the model built (i.e. $y_i = 0$), while the remaining part is coherent $(y_i = 1)$. The generated data and the true relation can be observed in Figure 5.1, where circles are uncontaminated data, while triangles contaminate the sample, however only reversed triangles are not coherent with the true model, while remaining triangles are. The objective is to verify the behaviour of the estimator and weights generated for each observation, to check that coherent observations are not downweighted. We choose $\nu = 0,8$ and run a logistic regression on not contaminated part of data, while on contaminated data we run both the logistic regression and the weighted version. Regarding the weighting scheme we choose a bandwidth equal to $0,5$ and we set $k = 1$ both on design space and response space. The weight function is the the one defined in (5.8), results are reported in Table 5.2. The slope of logistic regression , second term in parentheses in the first column, is strongly influenced by leverage points, while the *wle-qr* estimator, second column, is resistant. Regarding the downweighted observations,

ε	Logit	wle-qr
0%	(0.8589; 2.1751)	(0.8712; 2.187)
5%	(0.1934; 0.7007)	(0.9708; 2.2339)
10%	($-$0.2560; 0.2356)	(0.8847; 2.1922)
15%	($-$0.2756; 0.2091)	(1.090; 2.360)
20%	($-$0.4527; 0.1369)	(0.9826; 2.3081)
30%	($-$0.68879; 0.03718)	(0.8639; 2.0356)

Figure 5.2: Contaminated logistic regression

we have for contaminated values whose observations contrast with the model, the following summary measures (Table 5.2)

Min	1st Qu.	Median	Mean	3rd Qu.	Max
6.90e-47	4.13e-34	2.23e-30	4.21e-19	6.17e-26	5.83e-18

Table 5.2: Contrasting leverage weights

while for observations coherent with the model we have (Table 5.3)

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.849	0.854	0.871	0.883	0.899	0.939

Table 5.3: Coherent leverage weights

therefore, the methodology proposed is able to discard contrasting leverages only.

5.3.3 Credit risk model

To test on a larger dimentional space the weighted logistic regression, we compose a subset of data described in section 3.1, using all defaulted events, while creating a subset of 200 undefaulted companies. In Table 5.4 it is shown that in presence of extreme observations, using a standard logistic regression we may obtain unstable results (as for the β_1 parameter, where even the sign of the relation changed). We compare our results with the *Conditionally*

Unbiased Bounded Influence Estimator (Kunsch et al. (1989)) and a *Mallows type* resistant estimator (Carroll and Pederson (1993)).

We use as bandwidth $h = 1$, while we set $k = 1$ for both response and explanatory data which we assume to be drawn from a multivariate normal distribution. In Table 5.4 are reported parameters and in parenthesis the t-test statistic.

	Logit	wle-qr	Mallows	Cubif
β_0	-1.3775 (-6.24)	-1.3500 (-6.403)	-1.6990 (5.743)	-1.6948 (-6.106)
β_1	0.1470 (2.54)	-0.6583 (-2.583)	-0.5313 (-1.545)	-0.5539 (-1.886)
β_2	-0.8773 (-2.60)	-0.3606 (-1.179)	-1.0416 (-1.938)	-0.9783 (-2.048)
β_3	-1.0659 (-5.30)	-0.9182 (-4.343)	-0.9221 (-4.010)	-0.9477 (-4.141)
β_4	0.6434 (3.53)	1.051 (4.481)	1.2390 (4.377)	1.2963 (4.596)

Table 5.4: Default risk, logistic regression

The results obtained by our weighted logistic regression are similar to the ones obtained using other robust procedures. There is a disagreement on β_2, we believe the reason is due to the distribution of the covariate which is much far from normal, this will be taken into account in future research, it has to be noticed however that the coefficient is not statistically significant.

References

C. Agostinelli. A one-step robust estimator based on the weighted likelihood methodology. Technical report, 16, Dipartimento di Scienze Statistiche, Universita' di Padova, 1997.

A. Basu and B. Lindsay. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of Institute of Statistical Mathematics*, 46:683–705, 1994.

R. J. Carroll and S. Pederson. On robustness in the logistic regression model. *Journal of the Royal Statistical Society (B)*, 55:693–706, 1993.

P. K. Dunn and G. K. Smyth. Randomized quantile residuals. *Journal of Compututational and Graphical Statististics*, 5:236–244, 1996.

F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393, 1974.

F. R. Hampel. *Contributions to the Theory of Robust Estimation.* PhD thesis, University of California, Berkeley, 1968.

P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.

L. Kunsch, L. Stefanski, and R. Carroll. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, 84:460–466, 1989.

B. Lindsay. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, 22:1081–1114, 1994.

M. Markatou, A. Basu, and B. Lindsay. Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference*, 57:215–232, 1997.

M. Markatou, A. Basu, and B. Lindsay. Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93:740–750, 1998.

D. A. Pierce and D. W. Schafer. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81:977–986, 1986.