Antonietta Mira, Paolo Tenconi

Bayesian estimate of credit risk via MCMC with delayed rejection

2003/34

UNIVERSITÀ DELL'INSUBRIA
FACOLTÀ DI ECONOMIA

http://eco.uninsubria.it

In questi quaderni vengono pubblicati i lavori dei docenti della Facoltà di Economia dell'Università dell'Insubria. La pubblicazione di contributi di altri studiosi, che abbiano un rapporto didattico o scientifico stabile con la Facoltà, può essere proposta da un professore della Facoltà, dopo che il contributo sia stato discusso pubblicamente. Il nome del proponente è riportato in nota all'articolo. I punti di vista espressi nei quaderni della Facoltà di Economia riflettono unicamente le opinioni degli autori, e non rispecchiano necessariamente quelli della Facoltà di Economia dell'Università dell'Insubria.

These Working papers collect the work of the Faculty of Economics of the University of Insubria. The publication of work by other Authors can be proposed by a member of the Faculty, provided that the paper has been presented in public. The name of the proposer is reported in a footnote. The views expressed in the Working papers reflect the opinions of the Authors only, and not necessarily the ones of the Economics Faculty of the University of Insubria.

Bayesian estimate of credit risk via MCMC with delayed rejection

Antonietta Mira and Paolo Tenconi

Abstract. We develop a Bayesian hierarchical logistic regression model to predict the credit risk of companies classified in different sectors. Explanatory variables derived by experts from balance-sheets are included. Markov chain Monte Carlo (MCMC) methods are used to estimate the proposed model. In particular we show how the delaying rejection strategy outperforms the standard Metropolis-Hastings algorithm in terms of asymptotic efficiency of the resulting estimates. The advantages of our model over others proposed in the literature are discussed and tested via cross-validation procedures.

1. Motivation

The aim of this paper 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 % defaulted. A more detailed description of the dataset appears in Table 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 logistic regression model we propose is Bayesian, hierarchical and introduces dependency among different sectors thus addressing efficiently all the above mentioned issues.

TABLE 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%

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:

$$L(\underline{\alpha}, \underline{\beta}; y, x) = \prod_j \prod_i \theta_{i,j}^{y_{i,j}} (1 - \theta_{i,j})^{1 - y_{i,j}} \tag{2.1}$$

where

$$\theta_{i,j} = \frac{\exp(\alpha_j + \underline{x}'_{i,j} \underline{\beta})}{1 + \exp(\alpha_j + \underline{x}'_{i,j} \underline{\beta})}. \tag{2.2}$$

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

$$\pi(\underline{\alpha}, \underline{\beta}, \mu_\alpha, \sigma_\alpha | y, x) \propto L(\underline{\alpha}, \underline{\beta}; y, x) \prod_j p(\alpha_j | \mu_\alpha, \sigma_\alpha^2) \, p(\mu_\alpha) \, p(\sigma_\alpha) \, p(\underline{\beta}) \qquad (2.3)$$

A graphical representation of the proposed model appears in Figure 1.

3. The algorithm

We use a MCMC algorithm (Tierney, 1994) to simulate observations from (2.3), 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
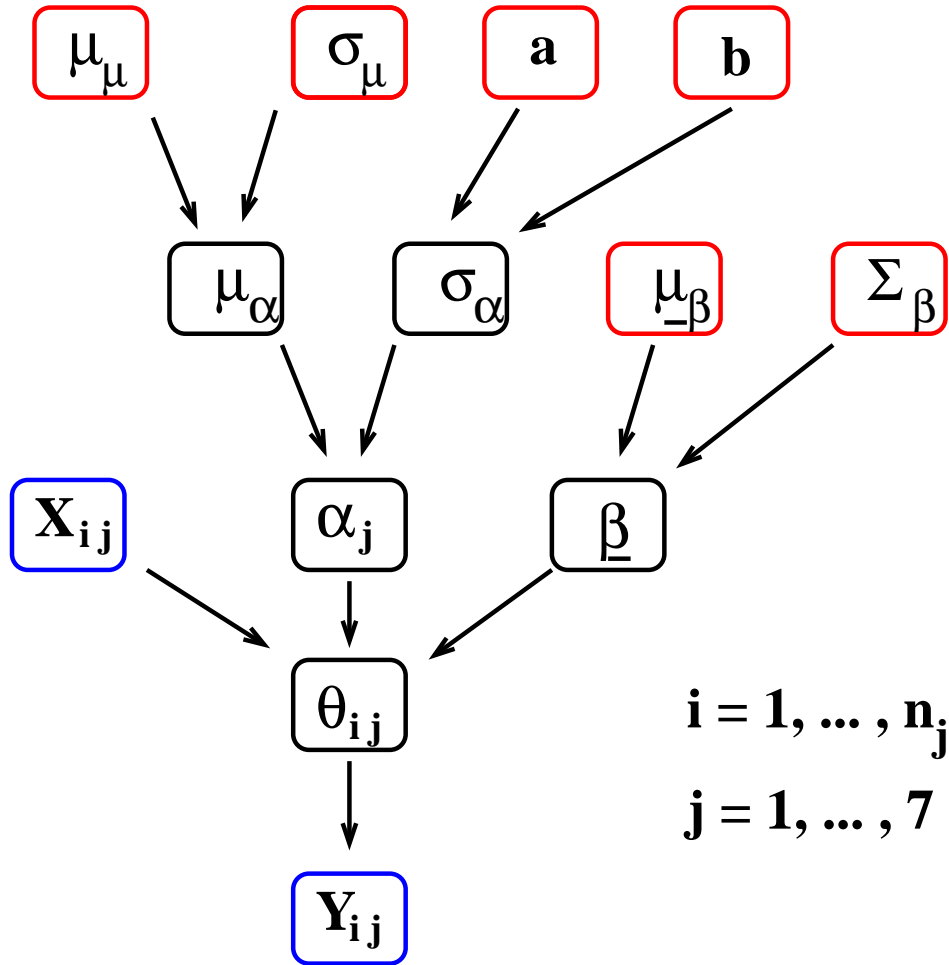
FIGURE 1. Graphical representation of the model.

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.

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 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 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
b_1	0.15
b_2	0.4
b_3	0.3
b_4	0.15

The simulation results are presented in Table 3 where the mean along the sample path is reported for both the MH and the DR chain. The numbers in Table 3 and 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 then 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. In 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), by computing the 0.25 and the 0.975 quantiles of the simulated values.

For comparison purposes, in Table 3 we also report the MLE (maximum likelihood estimates) of the logistic regression parameters, $\underline{\alpha}$ and $\underline{\beta}$, obtained

TABLE 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_α^2	2.97 (0.293)	0.695 ; 7.28	2.21 (0.123)	0.65 ; 5.18		

using a standard Newton-Raphson procedure. When computing the MLE we use (2.1) 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 shows, this dummy variable is justified also by the Bayesian analysis, since the estimated value of 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 2 two such distributions (for company 30 in sector 6 and company 20 in sector 2) are plotted: notice the long
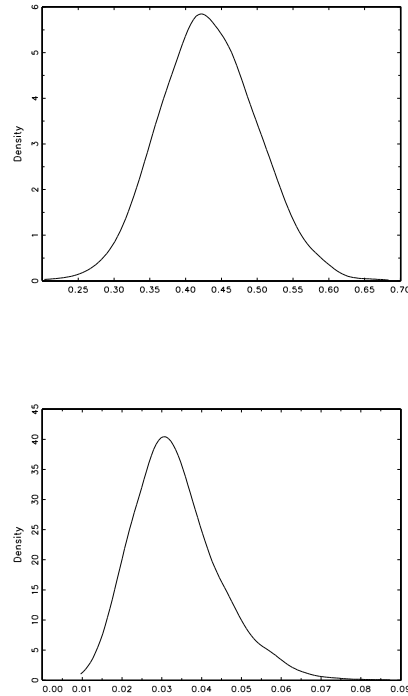
FIGURE 2. Posterior density estimate of DP: company 30 in sector 6 (top); company 20 in sector 2

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 4 summaries the results obtained for the two companies mentioned above. In the first column we report the value obtained using formula (2.2) 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 (2.2) the 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 ML 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.
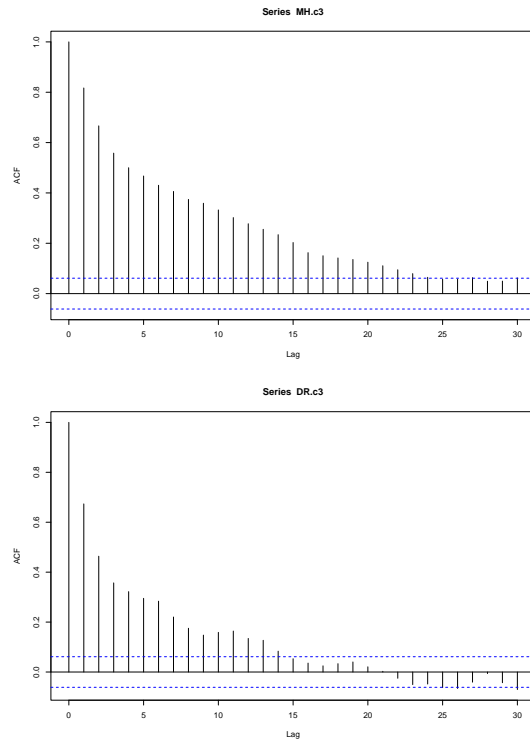
FIGURE 3. Autocorrelation functions for α_3: MH (left) and DR

TABLE 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 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, 1989). The results are presented in Tables 5 and 6 and show that, for all the parameters of interest, the DR outperforms the MH.

TABLE 5. Estimates of τ for $\underline{\alpha}$ 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 6. Estimates of τ for $\underline{\beta}$ 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 5, 6,7,8,9,10 and 11 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 $n = 30\%7513 = 2254$) 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 7 and show the overall better performance of the Bayesian model.

TABLE 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 4, 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.

5. Conclusions and extensions

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 (as obtained by classical MLE approaches) allows the construction of credible intervals and the possibility to estimate quantiles to derive performance indicators such as the analog of the Value at Risk for the default probability. We thus obtain a more complete and informative picture of the quantity of interest. Furthermore, the linear parametric model adopted allows for a coherent economic interpretation of the estimated parameters.

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 to estimate DP of sectors where no default event was observed by taking strength from the data available from other sectors that present a similar behavior relative to credit risk.

Finally, since the joint posterior distribution of the DP associated to all companies in all sectors is available one can derive the joint posterior estimates of the risk associated to a specific sector or to a specific portfolio of loans. The aggregation involved in computing the risk of a given portfolio should take the covariance structure derived by the posterior distribution into account.

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.

We plan to investigate a model where different slopes are allowed for different sectors and indicators of the economic cycle will be included among the explanatory variables. A further extension will incorporate time into the analysis: the performance indicators derived from the balance sheets can be updated every 3 months. Finally, partition models or mixture models can be used to partition the companies in different sectors.

References

[1] P. J. Green and A. Mira, *Delayed Rejection in Reversible Jump Metropolis-Hastings*, Biometrika, **88** (2001) 1035–1053.

[2] G. King and L. Zeng, *Logistic Regression in Rare Events Data*, Political Analysis, **9:2** (2001), 137–163.

[3] P. H. Peskun, *Optimum Monte Carlo sampling using Markov chains*, Biometrika, **60** (1973), 607–612.

[4] 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).

[5] L. Tierney, *Markov Chains for Exploring Posterior Distributions*, Annals of Statistics, **22** (1994), 1701–1762.

[6] L. Tierney and A. Mira, *Some adaptive Monte Carlo methods for Bayesian inference*, Statistics in Medicine, **18** (1999), 2507–15.
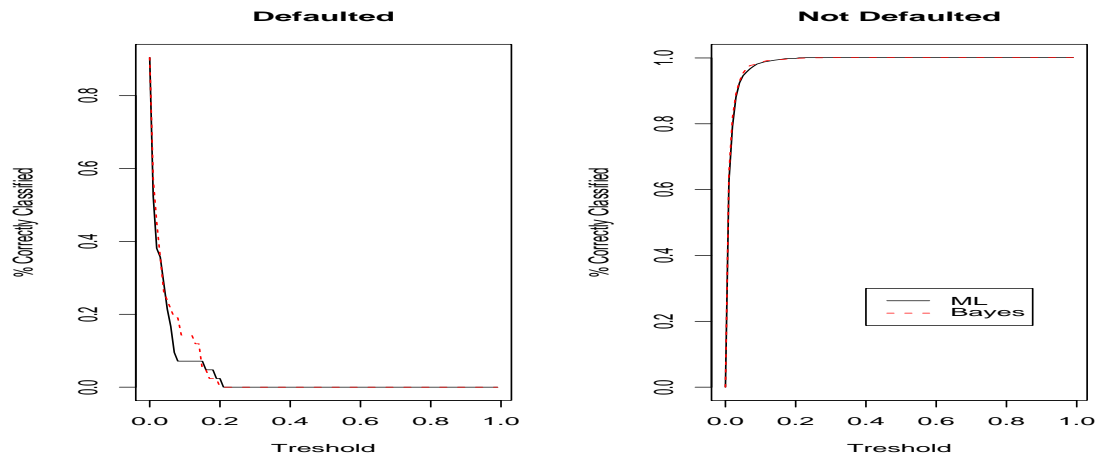
FIGURE 4. Percentage of correct classification for defaulted (left) and not defaulted companied as the classification threshold varies.
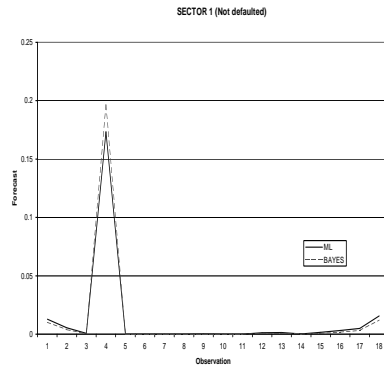

FIGURE 5. Sector 1.

Dipartimento di Economia, Università dell'Insubria, Via Ravasi 2, 21100 Varese, Italia
E-mail address: amira@uninsubria.it

Istituto di Finanza, Università della Svizzera Italiana, Via Buffi 1, 6900 Lugano, Svizzera
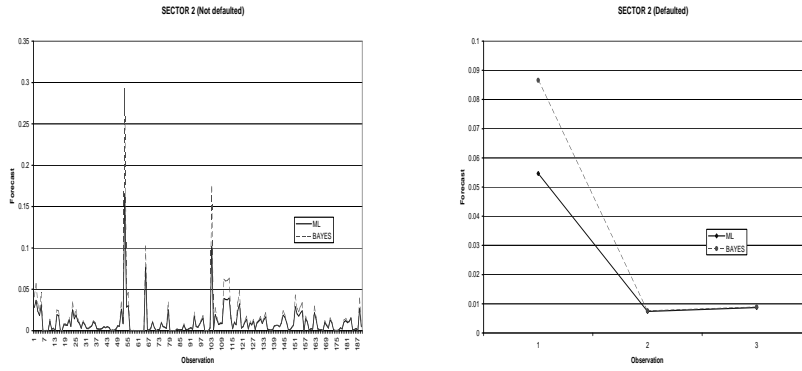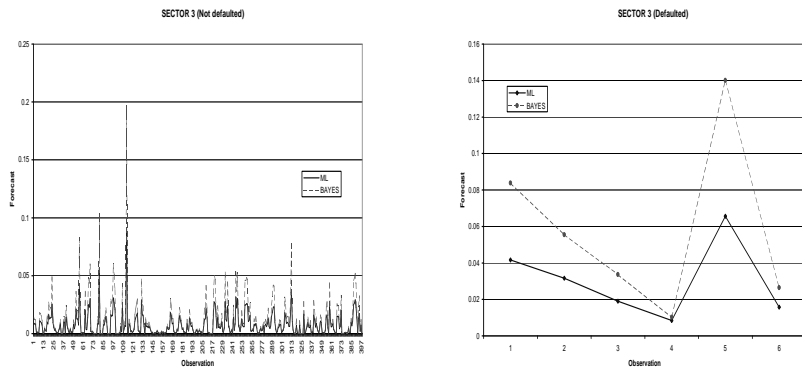*E-mail address*: paolo.tenconi@lu.unisi.ch
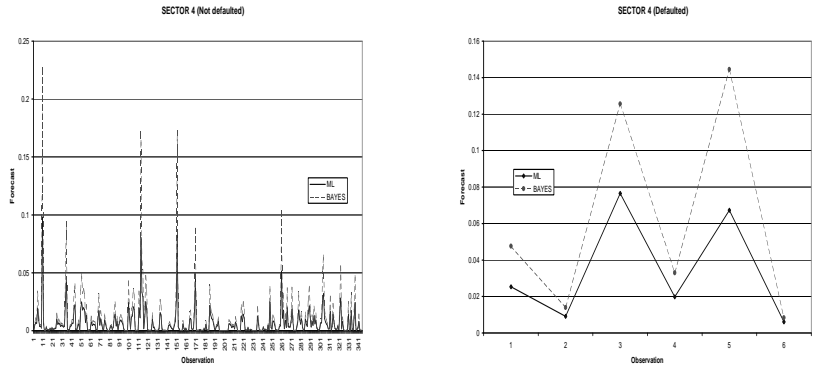
FIGURE 6. Sector 2.
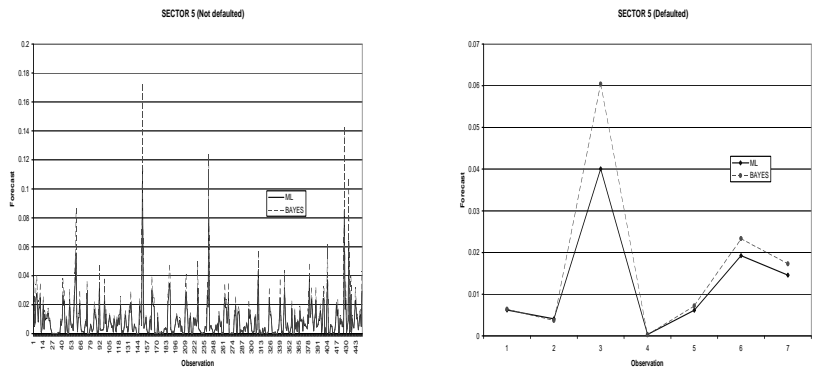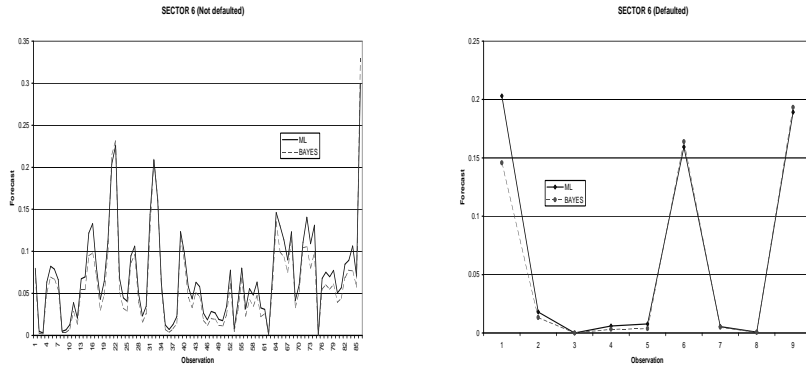

FIGURE 7. Sector 3.

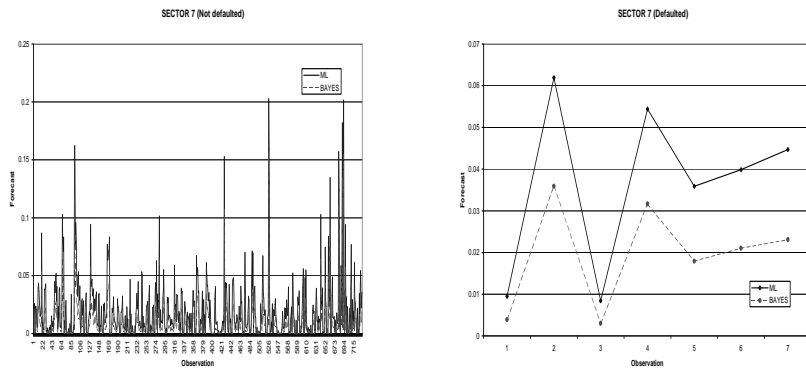FIGURE 8. Sector 4.

FIGURE 9. Sector 5.

FIGURE 10. Sector 6.

FIGURE 11. Sector 7.