Industry shocks and empirical evidences on defaults comovements.*

Md Moudud Alam†

May 26, 2010.

Abstract

It is commonly agreed that the credit defaults are correlated. However, the structure and magnitude of such dependence is not yet fully understood. This paper contributes to the current understanding about the defaults comovement in the following way. Assuming that the industries provides the basis of defaults comovement it provides empirical evidence as to how such comovements can be modeled using correlated industry shocks. Generalized linear mixed model (GLMM) with correlated random effects is used to model the defaults comovement. It is also demonstrated as to how a GLMM with complex correlation structure can be estimated through a evry simple way. Empirical evidences are drawn through analyzing quarterly individual borrower level credit history data obtained from two major Swedish banks between the period 1994 and 2000. The results show that, conditional on the borrower level accounting data and macro business cycle variables, the defaults are correlated both within and between industries but not over time (quarters). A discussion has also been presented as to how a GLMM for defaults correlation can be explained.

Key words: Credit risk, defaults contagion, GLMM, cluster correlation.

*The author is greatful to his supervisors, Kenneth Carling and Sune Karlsson, for their valuabale comments and instructions which helped a lot in improving the presentation of the paper. The author would also like to thank Xia Shen and the seminar participants at Dalarna University and Örebro University for their valuable comments.

†Dept. of Economics and Social Sciences, Dalarna University and Swedish Business School, Örebro University. Contact: Dalarna University, SE 781 88 Borlänge, Sweden; maa@du.se

1 Introduction

The comovement of defaults events, or defaults contagion, has drawn substantial attention in the recent literature on credit risk (Basel 2006). Some consequences of the failure to correctly address the defaults contagion are demonstrated in several papers mainly through simulation (see *e.g.,* Carling, Rönnegård & Rosczbach (2004)) and through some simple hypothetical examples (see *e.g.*, Das (2007) , Lucas (1995) and Zhou (2001)). Though the importance of the issue is widely accepted a common consensus on the framework to analyze the defaults contagion is yet to be achieved. The reasons are briefly reviewed in the following paragraph.

The most important reason why the debate continues is that there is not much empirical evidences to support or reject the debating hypothesis regarding the defaults comovement. Credit defaults data are not easily accecible and the defaults events are very rare. Therefore, it is difficult to estimate the defaults correlation precisely from those existing data sets which are, in most the cases, lacking sufficiently long time series observations (Zhou 2001).

Besides the problems raised from the unavailability of sufficient data and the subjective preference to particular type of credit risk model[1], the other areas of disagreement include

- Basis of comovement: Common sector factor *e.g.* industry defined by SIC (Chava & Jarrow 2004, Das, Freed, Geng & Kapadia 2006) or SNI classification (Carling et al. 2004, Rösch 2003), business cycle (Rösch 2003) *e.g.* bad/good credit regime resulted from time effect (McNeil & Wendin 2007), a combination these two (Das et al. 2006), credit rating category *e.g.* Moody's rating category: Aaa, Baa etc. (Das et al. 2006), geographic region factor (Lucas 1995) or something else.

- Structure of comovement: Whether the comovement exists only within some clusters defined by common factors (Carling et al. 2004), or both within and between common factors (Lucas 1995) or through several factors (Das et al. 2006, McNeil & Wendin 2007).

- Reasonable measure of comovement: Conventional Pearson's correlation (Carling et al. 2004) or the rank correlation such as copula correlation (Embrecht, McNeil & Straumann 2002, Das, Duffie, Kapadia & Saita 2007).

The possibility of using the industries as the basis of defaults comovements has been explored empirically in several papers *e.g.* Carling et al. (2004), McNeil & Wendin (2007) and Rösch (2003) and seems a reasonable choice. The aim of this paper is therefore to contribute to understanding of the structure and magnitude of the defaults comovements assuming the industries as the basis of such comovements. The empirical evidences are drawn through analyzing

[1] See Altman & Saunders (1998) and Carling, Jakobsson, Lindé & Roszbach (2007) for comprehensive reviews on different approaches to credit risk modeling.

individual borrower (firm) level quarterly credit history data obtained from two major Swedish banks between the period 1994-2000.

The choice of the modeling framework, in this paper, is guided by the principle of parsimony which can be explained in the following way. First, to choose a model which is capable of producing defaults correlation of a realistic magnitude. Second, use a model which does the above task by keeping the model complexity under control, in the sense that the parameter space does not grow dramatically with the sample size and parameters are easily interpretable. Third, which is capable of reproducing high/low defaults regimes and finally, the model is easily used to simulate data. The above principles are also considered as the criteria of a good model for modeling dependant defaults (Elizalde 2006).

This paper leaves the third point of disagreement by simply commenting that unless the model is completely marginally specified the simple correlation together with the specified conditional and marginal distributions is capable of producing a complete and unambiguous scenario of the dependency. However, the correlation itself may not be interpreted as a direct measure of dependency specially when the models are non-linear.

The rest of the paper is organized in the following way. Section 2 introduces the modeling framework, Section 3 provides data description and the important results, Section 4 describes some comparative implications of different candidate contagion frameworks and Section 5 concludes.

2 Statistical model

Though there is no explicit theory in the literature describing the defaults contagion, there is a fairly common agreement that the defaults of the firms cluster if the obligors are exposed to some common shocks due to their business relation (Carling et al. 2007, Das et al. 2006, Lucas 1995). Based on the above arguments, an immediate suggestion is to condition the defaults probability on macro economic indices (see *e.g.* Carling et al. (2007) and Chava & Jarrow (2004)). Though the existing literature suggest that such conditioning on macro economic and industry specific variables improve the fit of the model (Carling et al. 2007, Chava & Jarrow 2004), the recent literature on credit risk modeling also reveals that even after such conditioning a substantial residual comovement remains unexplained and may be attributed to unobservable covariates, or frailties (Das et al. 2007, Elizalde 2006).

Mixed models, which extends the simple statistical models by allowing for unobservable random effects, are widely suggested in statistical literature for modeling clustered and correlated data (McCulloch & Searle 2001). From the perspective of the principles of parsimony, mixed models can be considered as the best practice for modeling correlated data. An extension of the linear models for non-Gaussian, exponential family, distributions is referred to as the

Generalized Linear Models (GLM) (McCullagh & Nelder 1989) while further extension of GLM with random effects terms is called Generalized Linear Mixed Models (GLMM) (McCullagh & Nelder 1989, McCulloch & Searle 2001). GLMM provides a rigorous way to model dependence in non-Gaussian response data. Examples of applying mixed models for credit risk modeling include Carling et al. (2004), McNeil & Wendin (2007) and Rösch (2003) though the second one followed a Bayesian framework while the third one did not follow the computational framework of GLMM.

Since the data set contains information on the individual firms' quarterly defaults status, a binary response mixed model, which is a special case of GLMM, is a reasonable choice for modeling the defaults probability and the defaults correlation simultaneously. The basic motivations towards this kind of model are presented in Carling et al. (2004) and McNeil & Wendin (2007). However, a common limitation of both the papers is that they use a conventional specification of the GLMM which restrict them to allow only within cluster, which is SNI industries in Carling et al. (2004) and time in McNeil & Wendin (2007), correlations of defaults. The model specification considered in this paper is, therefore, an extension of the existing practices in the sense that it allows and tests for a more complex structure of correlation with an aim to explaining the defaults contagion in a more realistic way.

2.1 Modeling correlated defaults with binomial GLMM

Let, y_{ikt} denotes the state of defaults (1 = default, 0=non-default) of loan i $(i = 1, 2, ..., n_{kt})$ in industry k $(k = 1, 2, ..., K)$ and at time $t\,(t = 1, 2, ..., T)$. Following the framework of GLMM (McCullagh & Nelder 1989) the conditional probability of default, $p_{itk} = E\left(y_{ikt}|u_{kt}\right)$ can be modelled as $y_{ikt}|u_{kt} \backsim Bin\left(1, p_{ikt}\right)$ and $g\left(p_{ikt}\right) = X_{ikt}\boldsymbol{\beta} + Z_{ikt}\mathbf{u}_t$ where, $g\left(.\right)$ is called the link function and $g\left(p_{ikt}\right) = \log\left(\frac{p_{ikt}}{1+p_{ikt}}\right)$ gives logistic mixed models (as used in McNeil and Wendin McNeil & Wendin (2007)) while $g\left(p_{ikt}\right) = \log\left(-\log\left(1-p_{ikt}\right)\right)$ gives a complementary log-log model (as used in Carling et al. (2004)) and $g\left(p_{ikt}\right) = \Phi^{-1}\left(p_{ikt}\right)$ produces a probit mixed model (which would be the case if the GLMM framework were applied in Rösch (2003)), X_{ikt} is a row vector of loan, industry and time specific covariates where each column of it corresponds to a covariate, $\boldsymbol{\beta}$ is a column vector of *fixed* effects parameters, Z_{ikt} is the ikt^{th} row of the design matrix associated with the *random* (unobservable) effects $\mathbf{u}_t = \left(u_{1t}, u_{2t}, ..., u_{Kt}\right)^T$ and $\mathbf{u}_t$'s are iid random variables having some specific distribution, $f\left(\mathbf{u}\right)$. The elements in the design matrices, $\mathbf{X}$ and $\mathbf{Z}$ are assumed to be known. Conventionally, the random effects vector, $\mathbf{u}$, is assumed to have a multivariate normal distribution with a 0 mean vector and a covariance matrix, $\mathbf{D}$, *i.e.* $f\left(\mathbf{u}\right) = \frac{|\mathbf{D}|^{-\frac{1}{2}}}{(2\pi)^{\frac{K}{2}}} \exp\left[-\frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right]$. We retain the assumption of multivariate normality of $\mathbf{u}$ and explore the possibility of both the defaults' independence between time *i.e.* $Cov\left(u_{kt}, u_{kt'}\right) = 0 \;\forall\; t \neq t'$ and dependence over time *i.e.* $Cov\left(u_{kt}, u_{kt'}\right) \neq 0 \;\forall\; t \neq t'$.

Let us denote

$$\eta_{ikt} = X_{ikt}\boldsymbol{\beta} + Z_{ikt}\mathbf{u}_t \tag{1}$$

which is called the linear predictor (McCullagh & Nelder 1989). A careful specification of (1) along with a link function determines the effect of the fixed covariates as well as the correlations on the defaults probability. Since the implications of the fixed effects specification are rather extensively studied in the existing literature on credit risk modeling and since the defaults correlation is the main interest of this study, we leave the fixed effects specification by adopting some reasonable formulation more precisely, we follow Carling et al. (2004). In the rest of this section, we summarize the existing suggestions which are available in the literature on the specification of the defaults correlation and present them in terms of the specification of the random effects in equation (1).

The specification of the random effects part in (1) determines the correlation structure and it's realization provides a measure of the random industry shocks which determines differential credit defaults concentration in different clusters (industry sectors and times). Assuming $\mathbf{u}_t = 0$ or $\mathbf{D} = \mathbf{0}$ produces independent defaults and the model reduces to simple GLM. This kind of models are often referred to as Conditionally (on $\mathbf{X}$) Independent Defaults (CID) model in the literature on credit risk modeling (Elizalde 2006).

Independent and identical $\mathbf{u}_t$ *i.e.* $\mathbf{u}_t = \mathbf{u} = (u_1, u_2, ..., u_K)^T$ and $\mathbf{D} = \sigma_u^2 \mathbf{I}_K$ (say) incorporate only a symmetric intra-sector (industry) dependencies (SID) where the realization of u_k denotes the k:th industry specific random effect (shock) which is identical over time. The SID model with a probit link would produce a similar model to Rösch (2003). The SID model is the simplest kind of GLMM. A drawback of the latter is that u_k's suffer from a lack of economic interpretation *e.g.* if u_k's are the random industry shocks then it is not reasonable that they are identical over time.

Independent but non-identical $\mathbf{u}_t$ and $\mathbf{D} = diag\left\{\sigma_k^2\right\}_{k=1,2..K}$ produces non-symmetric intra-sector correlation (NIC) and the realizations of $\mathbf{u}_t$ gives the vector of industry specific shocks at time t. The last model is exactly what was used in Carling et al. (2004). Since the defaults concentration is likely to vary over time, the NIC model makes sense for credit risk modeling. However, it is hard to believe that the industry shocks are independent between industries.

The companies in different industries are related through their business relations. Hence, a shock coming to any industry is likely to transmit to companies of the other industries through the business relations of the companies. Therefore, we propose an extension of the NIC model with an unstructured covariance matrix for the random effects *i.e.* we propose the $\mathbf{D}$ matrix to be a general, symmetric and positive-definite matrix. We call this last model as the general industry shocks (GIS) model.

A time dependency among $\mathbf{u}_t$'s might also be possible (see *e.g.* McNeil & Wendin (2007)). In such situation a vector autoregressive (Reinsel 1997) specification of $\mathbf{u}_t$ would be more reasonable

than a simple auto regressive specification as it is presented in McNeil & Wendin (2007). We denote a GIS model with a first order vector auto regressive $\mathbf{u}_t$'s as the time-dependent industry shocks (TIS) model and express the dependency structure of $\mathbf{u}_t$ as

$$\mathbf{u}_t = \mathbf{B}\mathbf{u}_{t-1} + \boldsymbol{\varepsilon}_t \tag{2}$$

where, $\mathbf{B}$ is a $K \times K$ square matrix with absolute eigen values less than 1 and $\boldsymbol{\varepsilon}_t$ is a vector of white noise. The data set permits us to estimate all these 5 different specification of the random effects and the results are presented in section 3.

2.2 Choice of the link function

The choice of link function, $g(.)$, is another point of disagreement among binary mixed models for credit risk modeling. It is mentioned early that there are three widely used alternative choices for the link function for binomial mixed models. Carling et al. (2004) used a complementary log-log link, McNeil & Wendin (2007) used a logit link which is also called the canonical link (McCullagh & Nelder 1989) for binomial models and is the most frequently used link for credit risk modeling (Altman & Saunders 1998) and the model used in Rösch (2003) is equivalent to the binomial GLMM with a probit link.

A theoretical comparison between different links is given in McCullagh & Nelder (1989). From the measures of the goodness of fit, logit and probit links are often indistinguishable (McCullagh & Nelder 1989). Both the logit and the probit links are symmetric around 0 and provides almost equal prediction around $p = 0.5$. The logit link, however, has flatter tails than the probit link. Complementary log-log link is not symmetric around 0 and infinitely slower in prediction than logit link near $p = 1$ though it is almost similar to logit link when p is small.

Besides the mathematical features of the link function, the interpretation of the parameters are also to be considered as a criteria for a good model. This is one of the reasons why we do not prefer machine learning methods such as neural network (Altman & Saunders 1998). The complementary log-log model was chosen in Carling et al. (2004) since in that case the model parameters has a proportional hazard model explanation in the latent variable scale. However, it is well known that the model parameters estimated through a logistic model can also be interpreted as the parameters of a discrete-time survival model under non-informative censoring (Chava & Jarrow 2004). The parameters of the logistic model also have the proportional odds explanation and the computation of the model parameters become relatively easy since the logit link is the canonical link which make many analytical derivations simple (Alam 2008). Therefore, we propose a logit link for analyzing the data.

3 Data analysis and results

We re-analyze the data set of Carling et al. (2004)[2] to explore the possibility of more complex structure of dependency among credit defaults. The data set contains information on credit history of 100,926 small and medium size *aktiebolag*, the approximate Swedish equivalent of US corporations and UK limited companies, from two major Swedish banks between the 2^{nd} quarter of 1994 and the 2^{nd} quarter of 2000. It should be noted that the above number of companies are not randomly sampled from the portfolios of the banks rather they represent the total number of companies in the portfolios of the two banks during the above mentioned period. Besides the bank data on credit status, the data set contains information on accounting data of the borrower companies *e.g* sales, assets, liabilities etc., different remarks of the credit bureau and two macro variables being the slope of the yield curve and the output gap. The richness of the data set enables us to estimate all 5 models discussed in section 2.1.

Industries are defined by merging the SNI industries (see www.scb.se) at the first two digits level, in a way that closely resembles the industry definition of Carling et al. (2004). Since, it was not possible to construct exactly the same industry definition presented in Carling et al. (2004) by merging SNI industries, the industry definition used in this paper differs slightly from that of Carling et al. (2004). Furthermore, we consider only 6 industries in contrast to 7 industries used in Carling et al. (2004). Such a reduction of the number of industries was necessary to insure the stability of the computation which would otherwise be hampered since with 7 industries the proportion of defaults is very close to 0 in some quarters (see Venables & Ripley (2002) for further discussion on this computational issue).

3.1 Estimation of the GLMM

The maximum likelihood estimation of the GLMMs encounter some difficulties since the random effects in a GLMM are not observable. A reasonable solution of the above problem is to integrate them out of the likelihood function. The integration is a challenging task specially for binomial GLMMs since the joint likelihood function has such a complicated functional form that the analytical solution of the integration is impossible. Under the conventional specification of the GLMM, computation of the integral is carried out via some numerical integration methods *e.g.* Gauss-Hermite quadrant, Marcov-chain Monte-Carlo method or Laplace approximation, or the integration is by-passed through the Generalized Estimating Equations (GEE) method (McCulloch & Searle 2001). All the above methods have some advantages and disadvantages but their implementations in available computer packages, *e.g.* SAS and R, does not allow to have an unstructured covariance matrix, $\mathbf{D}$, for the random effects (Littell, Milliken, Stroup &

[2] A part of the data are also analyzed in Carling et al. (2007). The above paper also provides an extensive discussion on the justification of the fixed effects part of the model.

Wolfinger 1996, Venables & Ripley 2002).

Alam & Carling (2008) and Alam (2008) provide feasible ways to handle those complex situations for large data sets. It is shown in Alam & Carling (2008) that, in such complex situations and with large data sets, the estimation of the model parameters can be carried out through fixed effects specification while Alam (2008) provided a two-step pseudo likelihood (2-PL) method, based on a Laplace approximation, to carry out the parameter estimation. Alam (2008) also provided an analytical expression to check the applicability of the fixed effects approach suggested in Alam & Carling (2008). This paper has applied the fixed effects approach proposed in Alam & Carling (2008) to estimate the model parameters. A Bayesian analysis is also carried out to check the credibility of the model parameter estimates of the fixed effects approach.

3.2 The results and model comparison

The objective of this paper, as stated early, is to study the mechanism of the credit risk transmission. Therefore, the results on the fixed covariate effects are of minor interest and are not presented in this paper. However, some of those results are available in Alam (2008) and Carling et al. (2004). It is worth noting that all the candidate models contained the same fixed effects specification as presented in Carling et al. (2004) except for the fixed effects specification of the model where the macro variables are not included. Since it is already known that the CID models does not perform well because of its inability to handle dependent response we do not make any comparison with that model rather we offer a comparison between other three models namely, the SID, NIC and GIS models.

The SID model has only one covariance parameter and its estimate is 0.25. The estimate of the covariance parameter, $\mathbf{D}$, obtained from the NIC and GIS models are given in Table 1. The second column of Table 1 presents the estimated variances of the random effects (industry shocks) in the NIC model. It is woth noting that the covariances between the random industry shocks are 0 for the NIC model. Columns 3-9 in Table 1 present the estimated variances and the covariances of the random industry shocks in the GIS model *e.g.*, the value 0.23 at the row 1 and column 3 in Table 1 presents the variance of the industry schocks for industry 1 (Public & subsdized sector) while 0.12 at the row 2 and column 3 in Table 1 presents the covariance of the industry schocks for industry 1 (Public & subsidized sector) and industry 2 (Wholesale &

retail).

Table 1 Estimate of covariance parameters from NIC and GIS model

Industry	NIC model	GIS Model					
1. Public & subsidized sector	0.36	0.23					
2. Wholesale & retail	0.26	0.12	0.12				
3. Transport & commnunication	0.25	0.14	0.11	0.12			
4. Manufacturing	0.11	0.14	0.09	0.09	0.11		
5. Construction, forest and others	0.23	0.14	0.11	0.11	0.08	0.12	
6. Estate and finance	0.42	0.18	0.12	0.11	0.12	0.12	0.23

Table 1 reveals that the NIC model provides bigger estimate the variance parameters (diagonal elements of $\mathbf{D}$) compared with the GIS model. It is reasonable since the effects of omitted covariances are likely to increase the estimate of variances. Some other consequences of omitting off-diagonal elements in $\mathbf{D}$ is discussed in the following section. Since the above estimates are obtained from a very large data set and the estimated off-diagonal elements of $\mathbf{D}$ in GIS are not very close to 0, we can conclude that the conventional specification of random effects may not be applicable for modeling credit risk.

We also estimated the $\mathbf{D}$ matrix using the 2-PL (Alam 2008) technique. The mean absolute difference between the estimates of non-redundant parameters of $\mathbf{D}$ obtained from these two method was found to be 0.06. The above fact indicates that the estimates of $\mathbf{D}$ matrix obtained from the fixed effect approach and the two-step pseudo likelihood approach are close to each other.

To check the credibility of the estimates, a Bayesian analysis is also carried out. In the Bayesian analysis, the GIS model is estimated using R and Bugs[3]. Very weekly informative priors, as suggested in Gelman & Hill (2007), are used so that the estimates come close to those would be obtained through the likelihood method. The initial exercises with Bayesian analysis revealed that the Bugs implementation were extremely slow with this huge data set. Therefore, in stead of using the original data, a sample of the data is used for the Bayesian analysis. The sample is selected in the following way. We choose all the defaults companies and a random sample of equal size of non-default companies from the original data set. A consequence of such a sampling scheme with the logistic model is that it provides a biased estimate of the intercept term but the odds ratio parameter estimates are not affected by the sampling (McCullagh & Nelder 1989). As is expected, the estimates produced by Bayesian analysis are not very different from those obtained through the likelihood method. Therefore, we do not report all the results in this paper but a 95% MCMC high-density interval of $\mathbf{D}$ obtained from the Bayesian analysis is reported in the appendix.

The performance of the CID, SID and NIC models are compared, in terms of their ability to produce efficient estimates of the value-at-risk and the defaults distribution, through simulation

[3]Practical impetementation was done with OpenBugs 3.0.0, a free and opensource software which is downlodable from http://mathstat.helsinki.fi/openbugs/ .

in Carling et al. (2004) and it is found that the NIC model outperforms the CID and the SIC models. Somewhat similar conclusion is also drawn by McNeil & Wendin (2007). Here, we offer a comparison between NIC and GIS in terms of goodness of fit measured by Akaike's Information Criterion (AIC) which is calculated by using the following formula.

$$AIC = -2l_{PL} + 2p \tag{3}$$

where, l_{PL} is the pseudo likelihood[4] evaluated at the pseudo maximum likelihood estimate of the parameters and p is the total number of parameters including the fixed effects and the covariance parameters. The calculated AIC for the GIS model is 11665896 which is only 0.19% lower than the AIC of the NIC model (11688545). The smaller AIC indicated better fit of the GIS model. However, from the closeness, in relative sense, of the above AIC values one might suspect that the two models might not be substantially different from each other in terms of the goodness of fit. It should be noted that the AIC of the GIS model was calculated using the 2-PL approach, not the fixed effects approach.

We also estimated model (2) using the fixed effects approach (Alam 2008) i.e. **B** parameters are estimated using the fixed effects estimates of industry and time interaction parameters as though they are the true realization of the random effects. A least square estimate of **B** is found insignificant using a likelihood ratio test at the 5% level (see Reinsel (1997) pp. 92 for detailed test procedure). However, it should be noted that there are only 25 realizations of the random effects vector while its covariance matrix contains 21 $(6 \times (6+1)/2)$ non-redandant parameters. Therefore, the power of the above likelihood ratio test may be questionable.

4 Model implied effects on defaults comovements

In the previous section we have shown the estimates of the random effects variances and covariances obtained from different models. From the estimate of the covariance parameters of the random effects one can produce a correlation measure between different companies at the latent variable scale which is often referred as implied assets correlations (McNeil & Wendin 2007) but such a measure should be interpreted with due caution (Embrecht et al. 2002). In this section we offer a discussion on the magnitude of defaults dependencies, in terms of defaults correlations, implied by the models. We also demonstrate, with a simple example, how the GIS model transmits the contagion effect from one industry to another.

For linear mixed models, correlation between two responses does not depend on the covariates, only the random effects determine the correlation. But, it is not necessarily the case for non-linear models *e.g.* for binomial mixed models. The magnitude of the implied defaults correlation can not be explained in the same way as it is done for the linear mixed models. On

[4]The pseudo likelihood is obtained using Laplace approximation. See detailed derivation in Alam (Alam 2008).

the contrary, it requires rather complicated computation. For the GIS model given in (1), the correlation among defaults can be calculated in the following way.

$$cor\left(y_{ikt}, y_{jk't}\right) = \frac{cov\left(y_{ikt}, y_{jk't}\right)}{\sqrt{var\left(y_{ikt}\right) var\left(y_{ik't}\right)}} \tag{4}$$

The marginal distributions of y_{ikt}'s are binary. Therefore, their marginal means and marginal variances are $E\left(y_{ikt}\right)$ and $E\left(y_{ikt}\right)\left(1 - E\left(y_{ikt}\right)\right)$ respectively. Now, the marginal means of the responses of the GIS model are given by

$$E\left(y_{ikt}|X_{ikt}\right) = \int_{-\infty}^{\infty} \frac{\exp\left[X_{ikt}\boldsymbol{\beta}+u_{kt}\right]}{1+\exp\left[X_{ikt}\boldsymbol{\beta}+u_{kt}\right]} \frac{1}{\sqrt{2\pi}d_{kk}} \exp\left[-\frac{1}{2d_{kk}^2}u_{kt}^2\right] du_{kt} \tag{5}$$

Equation (5) has no closed form solution but it can be evaluated numerically using Monte-Carlo technique. Alternatively, $E\left(y_{ikt}\right)$ can be approximated as

$$E\left(y_{ikt}|X_{ikt}\right) \approx \frac{\exp\left[X_{ikt}\boldsymbol{\beta}^*\right]}{1+\exp\left[X_{ikt}\boldsymbol{\beta}^*\right]} \tag{6}$$

where, $\boldsymbol{\beta}^* = \frac{1}{\sqrt{1+\frac{256}{75\pi}d_{kk}}}\boldsymbol{\beta}$ (see McCulloch & Searle (2001), pp. 107). Once we have $E\left(y_{ikt}|X_{ikt}\right)$ in hand, $var\left(y_{ikt}|X_{ikt}\right)$ can be computed easily. Since all the expecations caclulated in this section are conditional on the observed covariate, X, we omit writing $y_{ikt}|X_{ikt}$ all the time; instead we denote $E\left(y_{ikt}|X_{ikt}\right)$ as $E\left(y_{ikt}\right)$ since the covraites are allways there as the contioned term. Equation (5) also holds for the NIC model. When, $k = k'$ the covariance between two responses can be calculated as

$$cov\left(y_{ikt}, y_{jkt}\right) = E_{u_{kt}}\left(E\left(y_{ikt}|u_{kt}\right)E\left(y_{jkt}|u_{kt}\right)\right) - E\left(y_{ikt}\right)E\left(y_{jkt}\right)$$

The second term in the right hand side of the above equation can be calculated using (5) while the first term can be calculated as

$$E_{u_{kt}}\left(E\left(y_{ikt}|u_{kt}\right)E\left(y_{jkt}|u_{kt}\right)\right) = \int_{-\infty}^{\infty} \frac{\exp\left[X_{ikt}\boldsymbol{\beta}+u_{kt}\right]}{1+\exp\left[X_{ikt}\boldsymbol{\beta}+u_{kt}\right]} \frac{\exp\left[X_{jkt}\boldsymbol{\beta}+u_{kt}\right]}{1+\exp\left[X_{jkt}\boldsymbol{\beta}+u_{kt}\right]} \frac{\exp\left[-\frac{u_{kt}^2}{2d_{kk}^2}\right]}{\sqrt{2\pi}d_{kk}} du_{kt} \tag{7}$$

Again, we need Monte-Carlo integration to evaluate the above expression. The above calculation is also the same for the NIC model. Now, if $X_{ikt} = X_{jkt}$, i.e. when two loans have the same observable characteristics, then the above expectation can be given as

$$E\left(E\left(y_{ikt}|u_k\right)E\left(y_{jkt}|u_k\right)\right) = E\left(\left(E\left(y_{ikt}|u_k\right)\right)^2\right)$$

$$\Rightarrow cov\left(y_{ikt}, y_{jkt}\right) = E\left(\left(E\left(y_{ikt}|u_k\right)\right)^2\right) - \left(E\left(y_{ikt}\right)\right)^2$$

Still, we can not avoid numerical integration. But, one thing that we see for sure is that any pair of loans in the same sector with the identical covariates, or the identical linear predictor values, have the same correlation. When $k \neq k'$ the basic form of equation (7) does not change

however, the integration is to be taken over the bivariate normal distribution of $(u_{kt}, u_{k't})$. It is worth noting that for NIC model the correlation is 0 for $k \neq k'$.

In order to calculate the above correlations we consider a portfolio of two loans with equal background characteristics and we chose the fixed effects part of the linear predictor to vary between -3 and 3 since the above interval covers over 90% probability under logistic model. Correlations are calculated between the firms having the same value of $X\boldsymbol{\beta}$. The estimate of $\mathbf{D}$ is taken from the respective model and the integrations are evaluated using Monte-Carlo technique. Equation (6) is used to check the credibility of the Monte-Carlo integrations. The results are summarized in Figure 1.

From Table 1 and Figure 1 we can conclude that high variance of the industry random effects imply high defaults correlation but the exact magnitude of the defaults correlation depends heavily on the realization of fixed effects. Now, consider the fixed effects part as the measure of the credit quality of a firm. Then, the above results is consistent with the results of Zhou (2001) where it is concluded that the peak of default correlation depends on the credit quality of the underlying firms. It is worth noting that the analysis in Zhou (2001) is based on a completely different modeling approach and the data set used there is concerning corporate defaults. Although the GIS model implies a correlation of 0.08 between any two companies in industry 1 (Public and subsidized sector), Figure 1 shows that the marginal correlation can at best be 0.05. Figure 1 also depicts that ignoring between industry correlation would produce higher estimate of intra industry correlation between two firms.

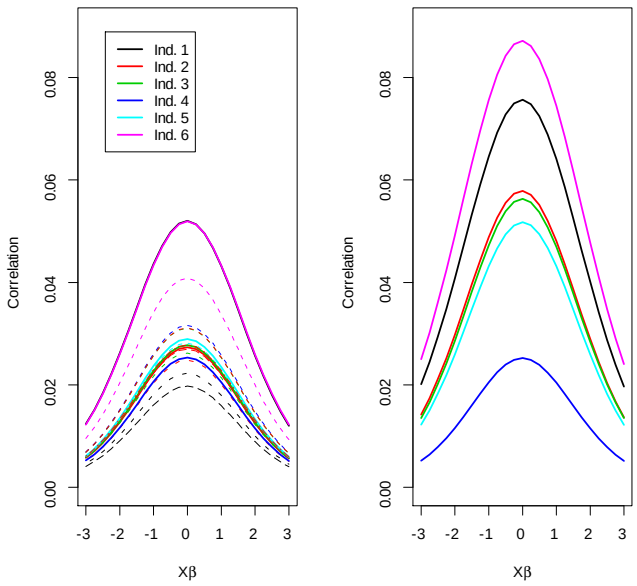

Figure 1 Defaults correlations implied by GIS model (left) and NIC model (right). Solid lines present within industry correlation and broken lines present between industry correlation.

Though, Table 1 and Figure 1 show very small estimate of defaults correlations between companies within and between industries their effect on bringing about a defaults concentration may not be negligible. For example, let us consider industry 4 (Manufacturing) which has the lowest within industry correlation. Now assume, the other 5 industries are exposed to high negative shocks; to be specific let us assume shocks coming to those 5 industries are the realizations of the 95th percentile from the respective marginal distributions. Given the shocks to the other 5 industries, $\mathbf{u}_2 = (u_1 = 1.14, u_2 = 0.96, u_3 = 0.97, u_5 = 0.98, u_6 = 0.14)$, the distribution of random shocks coming to Manufacturing industry, $\mathbf{u}_1 = u_4$, have a normal distribution with mean 0.795 and variance 0.015. Under the above situation, the marginal probability (MP) of defaults, joint probability of defaults (JP) of two similar firms in the same industry and defaults correlation (Cor) between two similar firms in industry 4 are shown in Figure 2. A comparison of those measures with other model specifications namely the NIC and the CID are also presented in Figure 2.

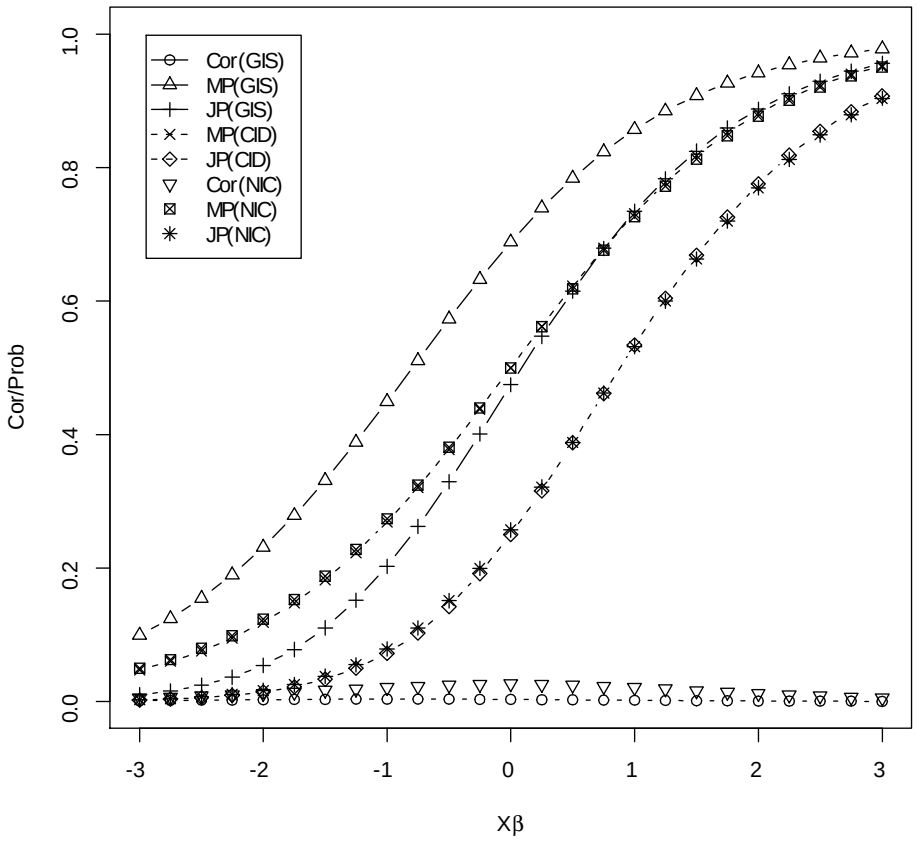

Figure 2 Effects of industry-shocks transmission in terms of marginal pobability of defaults (MP), joint probability of defaults (JP) and defaults correlation (Cor) in industry 4 implied by the GIS, NIC and CID models.

The CID and the NIC models do not take between industry correlation into account. Therefore, given the knowledge that the other industries are hit by tremendous shocks they still provide marginal defaults probabilities as though there was no such information. While the GIS model shows that under such condition the marginal probability of defaults in a particular industry gets much higher in some cases twice as high compared to that implied by NIC model while CID does the worst. However, in the situations when there is no external shocks are evident there would not be any visible different between these three models.

5 Concluding discussion

This paper provides empirical evidence on the different issues of disagreement regarding the defaults comovement. The results reveal that the GIS model is the best choice for modeling defaults probability. However, the NIC model may do a reasonable job. The issue that the NIC model misses is that the defaults events may be correlated between industries. Thus the NIC model fails to adjust for any contagion effect due to a shock in some specific industry.

In additiona, the empirical results do not provide any evidence in favour of time dependencies in the industry shocks and defaults correlation. This conlcusion is contradictory to many others' remarks (see *e.g.* Das (2007), Lucas (1995), McNeil & Wendin (2007), Rösch (2003) and Zhou (2001)). But, when people argue in favour of time varying defaults correlation they offer, generally, the following arguments. First, since the quality of the loan is dynamic over time and defaults correlation varies over the quality of the loans, the defaults correlation should be dynamic (Zhou 2001). However, if the above statement is true then the fixed effects part in the GIS model can take care of such variation. Second, since after every remarkable failure in the financial sector, *e.g.* when Enron fails, certain business regulations are changed hence the defaults probability and the defaults correlation reduces in the following time periods (Lucas 1995).

A third issue that has been evident from the discussion in section 4 is that the Basel II suggestion of considering an overall assets correlation of 0.2 may be far higher than the real correlation, especially for loans to small and medium size enterprises. A similar conclusion was also drawn in McNeil & Wendin (2007) and Rösch (2003). The estimated magnitude of the within industry assets correlation, which is found to be between 3.2-6.5% in GIS model, matches fairly well with the findings of Rösch (2003) however, it is much lower than those reported in McNeil & Wendin (2007) (6-8%) and Das et al. (2006) (0-19% as per their best model though it varies over time). In section 4 it is also demonstrated that such an assets correlation does not make sense when the assets correlation is translated into the defaults correlation (see also (Embrecht et al. 2002)).

From the estimate of the $\mathbf{D}$ matrix one can see that the random industry shocks are highly

correlated between industries. The correlations of industry shocks, for some industry pairs, are as high as 0.9. The above feature leaves an impression that a model allowing for an overall defaults correlation between all the companies, as it is suggested in Basel II, might also work well in modeling defaults probability. But, such a model will not be capable of incorporating an experience of shocks to some industries in estimating the defaults correlation in other industries except for saying that there the same correlation in all the industries.

We do acknowledge that any other modeling approach, *e.g.* copula approach, neural network or support vector machines, may be capable of predicting defaults correlation with the same precision as the GLMM models does but an advantage of the GLMM approach is that the model parameters have very natural explanations, e.g. the random effects can be explained as the random industry stocks and the fixed effects part as the quality of the loan. The GLMM approach also provides us with an instrument to adjust the defaults probability of the loans in some certain industries while some other industries are speculated to have experienced a sudden shocks.

References

Alam, M. (2008), An efficient estimation of the generalized linear mixed models with correlated random effects, *in* P. Brito, ed., 'Proceedings of COMPSTAT'2008', Vol. II: Contributed Papers, Physica-Verlag, Heidelberg, pp. 853–861.

Alam, M. & Carling, K. (2008), 'Computationally feasible estimation of the covariance structure in generalized linear mixed models (GLMM)', *Journal of Statistical Computation and Simulation* **78**(12), 1227–1237.

Altman, E. I. & Saunders, A. (1998), 'Credit risk measurement: development over the last 20 years', *Journal of Banking and Finance* **21**, 1721–1742.

Basel (2006), Studies on credit risk concentration: an overview of the issues and a synopsis of the results from the research task force project, Working Paper 15, Basel Committee on Banking Supervision, Bank for International Settlements, Basel.

Carling, K., Jakobsson, T., Lindé, J. & Roszbach, K. (2007), 'Corporate credit risk modeling and the macro economy', *Journal of Banking and Finance* **31**, 845–868.

Carling, K., Rönnegård, L. & Rosczbach, K. (2004), Is firm interdependence within industries important for portfolio credit risk?, Working Paper 168, Sverige Riskbank.

Chava, S. & Jarrow, R. A. (2004), 'Bankuptcy prediction with industry effects', *Review of Finance* **8**, 537–569.

Das, S. R. (2007), 'Basel ii: Correlation related issues', *Journal of Financial Services Research* **32**, 17–38.

Das, S. R., Duffie, D., Kapadia, N. & Saita, L. (2007), 'Common failings: how corporate defaults are correlated', *Journal of Finance* **62**, 93–117.

Das, S. R., Freed, L., Geng, G. & Kapadia, N. (2006), 'Correlated default risk', *The Journal of Fixed Income* (Fall 2006), 7–32.

Elizalde, A. (2006), Credit risk models i: defaults correlation in intensity models, CEMFI Working Paper 0605, CEMFI, Madrid.

Embrecht, P., McNeil, A. J. & Straumann, D. (2002), Correlation and dependence in risk management: properties and pitfalls, *in* M. A. H. Dempster, ed., 'Risk Management: Value at Risk and Beyond', Cambridge University Press, New York.

Gelman, A. & Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, New York.

Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. (1996), *The SAS system for mixed models*, Cary, North Carolina.

Lucas, D. J. (1995), 'Defaults correlation and credit analysis', *The Journal of Fixed Income* (March 1995), 76–87.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall, London.

McCulloch, C. E. & Searle, S. R. (2001), *Generalized Linear and Mixed Models*, Wiley, New York.

McNeil, A. J. & Wendin, J. P. (2007), 'Bayesian inference for generalized linear mixed models of portfolio credit risk', *Journal of Empirical Finance* **14**(2007), 131–149.

Reinsel, G. C. (1997), *Elements of Multivariate Time Series Analysis*, Springer, New York.

Rösch, D. (2003), 'Correlation and business cycles of credit risk: evidence from bankruptcies in germany', *Financial Market and Portfolio Management* **17**(3), 309–331.

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, Springer, New York.

Zhou, C. (2001), 'An analysis of defaults correlation and multiple defaults', *The Review of Financial Studies* **14**(2), 555–576.

A Appendix

A.1 Results from the Bayesian analysis.

The Bayesian hierarchical model specification for GIS model:

$y_{ikt}|\mathbf{u}_t \backsim Bin\left(1, p_{ikt}\right); i = 1, 2, ..., n_{kt}, k = 1, 2, ..K(K = 6), t = 1, 2, ..., T\,(T = 25)$

$logit(p_{ikt}) = X_{ikt}\beta + u_{kt}$

$\mathbf{u}_t = (u_{1t}, u_{2t}, ..., u_{Kt})^T \backsim N_K\left(\mathbf{0}, \mathbf{D}\right); \mathbf{u}_t \perp \mathbf{u}_{t'} \forall t \neq t'$

$\beta_j \backsim iid\ N\left(0, 100\right); j = 1, 2, ...,ncol(X)\,.$

$\mathbf{D} \backsim W^{-1}\left(\mathbf{I}_6, \right)$ *i.e.* **D** follows an inverted Wishart distribution with 7 degrees of freedom and the scale matrix being an identity matrix.

Estimated 95% MCMC high-density interval for **D** for the above model is given in table A.1.

Table A.1 An element-wise 95% high density interval of **D** from the Bayesian analysis (calculated with 5000 MCMC samples)

2.5th Percentile						97.5th Percentile					
0.21	0.13	0.14	0.11	0.11	0.17	0.73	0.52	0.54	0.49	0.48	0.74
0.13	0.15	0.14	0.10	0.11	0.08	0.52	0.53	0.50	0.43	0.44	0.52
0.14	0.14	0.14	0.10	0.11	0.10	0.54	0.50	0.51	0.42	0.43	0.55
0.11	0.10	0.10	0.06	0.08	0.08	0.49	0.43	0.42	0.42	0.38	0.51
0.11	0.11	0.11	0.08	0.07	0.07	0.48	0.44	0.43	0.38	0.42	0.50
0.17	0.08	0.10	0.08	0.07	0.23	0.74	0.52	0.55	0.51	0.50	0.98