Maximum likelihood estimate of default correlations

Estimating asset correlations is difficult in practice since there is little available data and many parameters have to be found. Paul Demey, Jean-Frédéric Jouanin, Céline Roget and Thierry Roncalli present a tractable version of the multi-factor Merton model in which firms are sorted into homogeneous risk classes. They derive a simplified maximum likelihood approach that provides estimates in a reasonable computational time. As an application of this methodology, industrial sector correlations are estimated from S&P's data

The estimation of default correlations between obligors is a challenging issue in the management of loan portfolios. Value-at-risk and other risk indicators are sensitive to the quality of the calibration of the credit model and, in particular, to the accuracy of the estimation of default correlations. Yet there is a major problem: data is scarce. While computed estimators are known to behave well asymptotically, with small samples biases and standard errors are likely to increase dramatically. This leads to imprecision in the resulting estimators. Gordy & Heitfield (2002) suggest imposing parametric restrictions on the underlying model in order to reduce the impact of this imprecision. They compare the behaviour of the constrained estimates with Monte Carlo simulations. Here, this approach is extended to a tractable multi-factor setup in the case of homogeneous risk classes, and we compare the usual maximum likelihood estimator (MLE) – which we call the 'binomial' MLE – and a simplified estimator called the 'asymptotic' MLE, which is more tractable and behaves well with small samples.

The following section provides a reminder of some well-known features of Merton's model of default occurrences, which is also that used in the CreditMetrics model (Finger, 1999), and details some homogeneity assumptions between risk classes. We then show how the MLEs are derived and to what extent adding other constraints to the model – roughly assuming inter-risk class correlations are constant – enables us to reduce to a two-factor model for each risk class, which is clearly more tractable. We then provide a crude evaluation of the bias in the estimation using Monte Carlo simulations and compare one-factor and multi-factor models. Finally, we present two estimations of default sectorial correlations extracted from the S&P public database for 1981–2002.

Risk classes and homogeneity assumptions

Throughout this article, we assume we have succeeded in sorting all firms into a short collection of risk classes with some homogeneity properties defined below. Risk classes can be built with rating grades, geographical areas, industrial sectors or a combination of these criteria. We denote the number of risk classes as C, the number of firms belonging to the risk class c as $N_c(t)$ (the current total number of firms alive is $N(t) := \Sigma_c N_c(t)$) and the number of default occurrences within the class c in year t as $D_c(t)$. The variable of interest for our study is the 'default rate' in the risk class c called $\mu_c(t)$ and defined as:

$$\mu_c(t) := \frac{D_c(t)}{N_c(t)} \tag{1}$$

The variance of this default rate for a given risk class depends heavily on the correlation between defaults of the firms belonging to this class. The intuition behind this statement is naive: the higher the correlation is, the more likely the other firms are to default given a default event at time t. This conveys the idea that we may extract information on the correlations within (and between) risk classes from the observation of these default rates. Moreover, while the variance of default rates is an indicator of the correlation, the mean default rate calculated over time provides information on the default probabilities among each class of risk. To illustrate this, we show in figure 1 the distribution of the annual default rate in the Merton/Vasicek model with respect to the asset correlation ρ. The risk class contains 1,000 names with a common default probability equal to 20%. We check that the variance of default rates depends on the asset correlation, whereas the mean of default rates is equal to the default probability. To go a little further in the study of default correlations, we need to apply a model for describing default occurrences.

☐ **Merton's general framework.** For our study, we use the original Merton model of default events, with several factors that will be specified later. We thus consider a set of N obligors, labelled by n. The risk of each obligor is modelled through a latent variable Z_n that stands for the normalised return on the obligor's asset. As usual in Merton's model, the latent variables are described by a Gaussian vector with standardised Gaussian margins. Obligor n defaults as soon as Z_n falls below a threshold B_n, which can be mapped as the default probability (for a given maturity) of the firm n. If we call τ_n the default time of the obligor n, we have:

$$\{\tau_n \le t\} = \{Z_n \le B_n\}$$

Having such a huge number of firms to cope with in practice (up to 100 or 1,000), it is usual in Merton's framework to assume we have identified a few risk factors (labelled by $f = 1, \ldots, F$) and rewrite the latent variables (in the Gaussian assumption) as noisy linear functions of these factors[1]:

$$Z_n = \sum_{f=1}^{F} A_{f,n} X_f + \sqrt{1 - \sum_{f=1}^{F} A_{f,n}^2}\,\varepsilon_n$$

We will choose the factors in a manner consistent with the risk classes that are used to group the obligors.

☐ **Homogeneity of risk classes.** We assume the risk classes are homogeneous enough to make the following assumptions:

■ All firms within a given risk class have the same rating, that is:

$$B_n = B_c, \qquad \forall n \in c \tag{2}$$

■ Within a given class of risk, the correlation between two firms is constant, that is:

$$\rho_{m,n} = \rho_c, \qquad \forall m, n \in c \tag{3}$$

■ Given any pair of risk classes (c, d) there is a unique correlation between any pair of firms (m, n) belonging to each class, that is:

$$\rho_{m,n} = \rho_{c,d}, \qquad \forall m \in c, n \in d \tag{4}$$

[1] *Without loss of generality, the factors are taken as being independent*

The first assumption suffers from a lack of evidence, since except in the case where risk classes gather obligors sharing the same rating, it is an unrealistic hypothesis in practice. However, we believe it is a practical compromise for the unique purpose of calibrating the correlations (we would not make such an assumption for assessing the capital risk of the bank).

We can rewrite the model with our assumptions. Let us consider Σ as the following $C \times C$ matrix[2]:

$$\Sigma = \begin{pmatrix} \rho_1 & \rho_{1,2} & \cdots & \rho_{1,C} \\ \rho_{2,1} & \rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{C-1,C} \\ \rho_{C,1} & \cdots & \rho_{C,C-1} & \rho_C \end{pmatrix} \tag{5}$$

If we assume that Σ is positive and definite, each variable Z_n can be written as a linear function of a set of F factors X_f, and an idiosyncratic term ε_n:

$$Z_n = \sum_{f=1}^{F} A_{f,c} X_f + \sqrt{1-\rho_c}\,\varepsilon_n, \qquad n \in c \tag{6}$$

where obligor n belongs to the risk class c and where the $F \times C$ matrix A is a 'square root' of Σ, that is, $A^{\mathsf{T}}A = \Sigma$. Any square root corresponds to a different form of the same factor model, based on the same matrix of cross-class correlations. The factors X_f and the idiosyncratic term ε_n are independent and follow a standard Gaussian distribution. The number of factors F has to be greater or equal to C to ensure the existence of A.

Constrained/unconstrained MLEs

As in Gordy & Heitfield (2002), a maximum likelihood procedure will be carried out for estimating the default correlations. Here we derive the MLE in the general framework described above (called the 'unconstrained' model) but, in a multi-factor setup, the formula is not tractable at all. Therefore we suggest adding a new constraint on the inter-risk class correlations, so that we are able to reduce the number of factors for each risk class down to two and get a new MLE that is much easier for the numerical optimisation. Finally, we replace the last MLE (called the 'binomial' MLE) with an 'asymptotic' MLE, which is less consuming in terms of computational time.

☐ **The unconstrained model.** Using the assumptions of the last section, we can easily write the probability of default conditional on the factors $\mathbf{X}$ as:

$$P_c(\mathbf{x}) = \Phi\left(\frac{B_c - \sum_{f=1}^{F} A_{f,c} x_f}{\sqrt{1-\rho_c}}\right) \tag{7}$$

where Φ (respectively ϕ) represents the standard Gaussian cumulative distribution (respectively density) function. Conditional on the factors, the random variable[3] D_c representing the default number in the risk class c has a binomial distribution with parameters N_c – the number of firms in the risk class c at time t – and $P_c(\mathbf{x})$. Let us note:

$$\mathbf{Bin}_c(\mathbf{x}) = \binom{N_c}{D_c} P_c(\mathbf{x})^{D_c} \left(1 - P_c(\mathbf{x})\right)^{N_c - D_c} \tag{8}$$

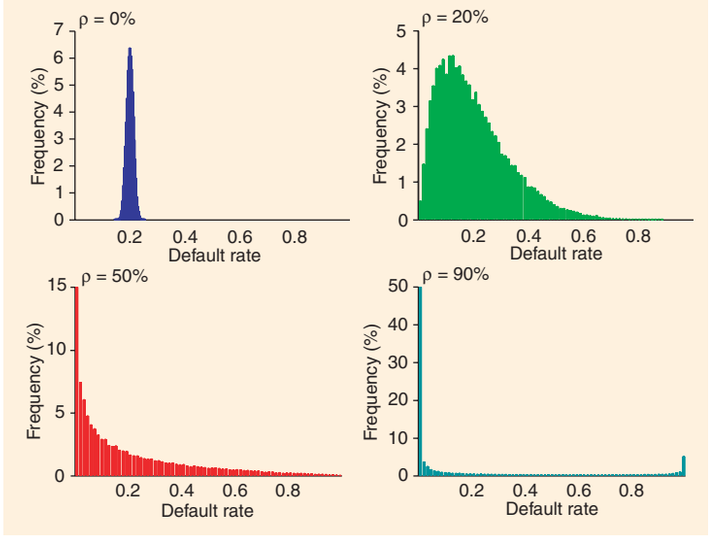
From this we can easily deduce the conditional likelihood of our observations. Summing over the distribution of each factor, we obtain the following expression for the unconditional log-likelihood:

$$\ell_t(\theta) = \log \int \cdots \int_{\mathbb{R}^F} \prod_{c=1}^{C} \mathbf{Bin}_c(\mathbf{x})\, d\Phi(\mathbf{x}) \tag{9}$$

In this expression, Φ is the standard multivariate Gaussian cumulative density function with the correlation matrix equal to the identity matrix. This log-likelihood has already been obtained by Gordy & Heitfield (2002).

This model is called 'unconstrained' because no other condition is imposed on the matrix Σ. The major difficulty we encounter is the large number of parameters to be estimated ($C(C+1)/2$). Due to the scarcity of data when dealing with default times series, estimating too many parameters is hazardous. To obtain more robust estimators, the number of parameters has to be reduced.[4] This expression becomes very intricate when the number of risk classes C increases, due to the multi-dimensional integration. A numerical solution cannot be obtained in a reasonable amount of time as soon as the number of factors becomes greater than three. For this reason, a different formulation of our initial model is explored, from which a simplified expression of the likelihood will be derived.

1. Distribution of annual default rates



☐ **The constrained model.** Let us now introduce an extra assumption on the matrix Σ. We assume that the correlation between two latent variables is unique among all risk classes, that is:

$$\rho_{c,d} = \rho, \forall c \neq d \tag{10}$$

The matrix Σ can therefore be rewritten as:

$$\Sigma = \begin{pmatrix} \rho_1 & \rho & \cdots & \rho \\ \rho & \rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & \rho_C \end{pmatrix} \tag{11}$$

We suppose $\rho \leq \min_c \rho_c$ so that Σ is positive. This is an assumption on the upper bound of the inter-risk class correlation that may not be confirmed in practice.[5] However, this assumption will prove its tractability and enable us to reduce[6] the number of estimated parameters to $C + 1$. Even if the constrained model is less general than the previous one, it is preferred for management purposes. Indeed, credit portfolio management and Raroc systems use generally parsimonious models to control their robustness.[7] For the remaining, we will only consider this 'constrained' model.

Now with the new assumption, we can rewrite our model much more easily as:

$$Z_n = \sqrt{\rho} X + \sqrt{\rho_c - \rho} X_c + \sqrt{1-\rho_c}\,\varepsilon_n, \qquad n \in c \tag{12}$$

This equation is obviously sufficient for equation (11), and this form (that is,

[2] *Σ is obviously not a correlation matrix, but its entries are the asset correlations of a sample of C obligors belonging to the different risk classes*

[3] *We drop the reference to the date t in the remainder*

[4] *As pointed by Gordy & Heitfeld (2002), the parameters are not well identified in the unconstrained model. They suggest that "identification problems can be overcome by imposing parametric restrictions"*

[5] *Nevertheless, we believe it is true when dealing with geographical areas or industrial sectors*

[6] *For example, with 15 risk classes, the number of parameters reduces from 120 in the unconstrained model to 16 in the constrained model*

[7] *With too many parameters, it may be difficult to understand the sensitivity of results to the parameters*

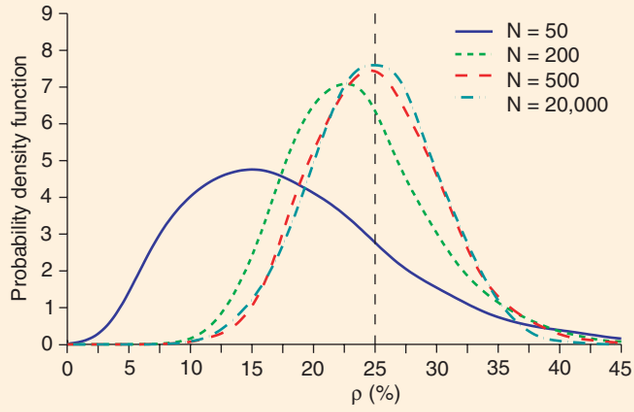
A. Statistics of the asset correlation estimates (PD = 200bp, ρ = 25%)

	Asymptotic		Binomial	
Statistics (in %)	MLE1	MLE2	MLE1	MLE2
Mean	23.7	22.5	25.2	23.6
Std. dev.	5.8	7.2	7.6	8.5
Median	23.1	21.7	24.7	23.0

B. Statistics of the asset correlation estimates (PD = 20bp, ρ = 25%)

	Asymptotic		Binomial	
Statistics (in %)	MLE1	MLE2	MLE1	MLE2
Mean	16.6	16.2	26.2	23.4
Std. dev.	8.9	9.8	11.6	12.2
Median	15.0	14.2	25.9	23.1

2. Asymptotic estimator MLE1 of asset correlation (PD = 200bp, ρ = 25%)



$F = C + 1$) is appealing because there is a natural economic interpretation for each term: Z_n can be explained by a common factor X, which affects each obligor in the same way, and by a specific factor X_c depending on the risk class c of the firm n. The part of Z_n that cannot be explained by sectorial or global factors is captured by the idiosyncratic term ε_n. In some sense, the constrained model may be interpreted as a two-factor model.

Under the Gaussian assumption for the idiosyncratic term ε_n, we can write the conditional default probability as:

$$P_c(x, x_c) = \Phi\left(\frac{B_c - \sqrt{\rho}x - \sqrt{\rho_c - \rho}x_c}{\sqrt{1-\rho_c}}\right) \tag{13}$$

Now we provide the derivation of two different MLEs to be used for the estimation.

■ **Binomial MLE.** The log-likelihood function is calculated in the same way as in the previous section. The conditional likelihood is first calculated and successively summed over the distribution of each sectorial factor and over the distribution of the common factor. The unconditional log-likelihood is[8]:

$$\ell_t(\theta) = \log \int_{\mathbb{R}} d\Phi(x) \prod_{c=1}^{C} \int_{\mathbb{R}} \mathbf{Bin}_c(x, x_c) d\Phi(x_c) \tag{14}$$

In this expression, the high-dimensional integral is replaced by a product of one-dimensional integrals, which are more tractable to compute. We can show that the expressions of likelihood (9) and (14) are mathematically equivalent assuming that $\rho \le \min_c \rho_c$.

■ **Asymptotic MLE.** Let us note as $\mu_c = D_c/N_c$ the default rate at time t in class c. When $N_c \to \infty$, and conditional on the factors $X = x$ and $X_c = x_c$, we have (according to the law of large numbers) $\mu_c \to P(x, x_c)$. Under this asymptotic assumption, we approximate the conditional default rate μ_c by its limit $P(x, x_c)$. This assumption leads to the following expression of the log-likelihood function[9]:

$$\ell_t(\theta) = \log \int_0^1 dy \prod_{c=1}^{C} \phi\left(f(y)\right) \frac{\sqrt{1-\rho_c}}{\sqrt{\rho_c - \rho}} \frac{1}{\phi\left(\Phi^{-1}(\mu_c)\right)} \tag{15}$$

with:

$$f(y) = \frac{B_c - \sqrt{1-\rho_c}\Phi^{-1}(\mu_c) - \sqrt{\rho}\Phi^{-1}(y)}{\sqrt{\rho_c - \rho}} \tag{16}$$

In a finite sample, D_c may be equal to zero, which is not possible asymptotically. This problem arises because the number of observations N_c may not be enough in practice. One way to reconcile data with our model is to consider that the real value of μ_c is not zero, but is inferior to a threshold $\mu_{\min}^c = 1/N_c$. Let us consider $\mathcal{U}$, the set of risk classes for which the default rate at time t is strictly positive: $\mathcal{U} = \{c \in \{1, \ldots, C\} \mid \mu_c > 0\}$. We obtain the following generalised formula for the log-likelihood:

$$\ell_t(\theta) = \log \int_0^1 \prod_{c \in \mathcal{U}_t} \phi\left(f(y)\right) \frac{\sqrt{1-\rho_c}}{\sqrt{\rho_c - \rho}} \frac{1}{\phi\left(\Phi^{-1}(\mu_c)\right)} \prod_{c \notin \mathcal{U}_t} \left[1 - \Phi\left(f(y)\right)\right] dy$$

In this section, two expressions of log-likelihood functions have been developed. The first one – binomial likelihood – is the function obtained by Gordy & Heitfield (2002). However, by imposing the restriction described above on the cross-risk class correlation matrix, we were able to convert the multi-dimensional integral into a product of one-dimensional integrals. The binomial likelihood obtained is more convenient, especially in a multi-factor setup. The second form of likelihood function – asymptotic likelihood – assumes that the number of firms in each class is large enough to allow us to approximate the random variable representing default rate by its limit. By maximising each likelihood – using numerical methods – we deduce the corresponding maximum likelihood estimators, respectively called 'binomial' and 'asymptotic' estimators.

Estimating the bias with Monte Carlo simulations

All the calculated MLEs are asymptotically unbiased. However, in small sample conditions, biases usually appear. The aim of this section is to give a rough evaluation of these biases using Monte Carlo simulations of the data.[10] We will stress the differences between the one-factor and two-factor (or the constrained multi-factor) frameworks.

---

[8] To derive this equation, one can compute the conditional (given $X = x$) likelihood, which is equal to:

$$\mathbb{P}\left(D_c = d_c \forall_c \mid X = x\right) = \prod_c \int_{\mathbb{R}} \mathbf{Bin}_c(x, x_c) d\Phi(x_c)$$

and sum over the distribution of X

[9] One way to derive this likelihood is to calculate the probability of the default rate conditionally to the factor X:

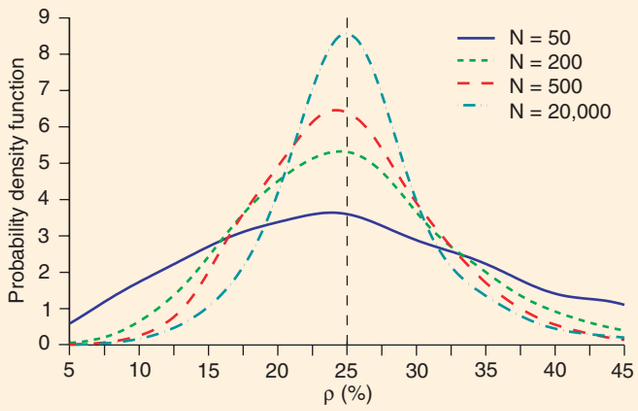
$$\begin{aligned} \mathbf{G}_c(m_c) &= \mathbb{P}\left(\mu_c \le m_c \mid X = x\right) \\ &= \mathbb{P}\left(\Phi\left(\frac{B_c - \sqrt{\rho}x - \sqrt{\rho_c - \rho}x_c}{\sqrt{1-\rho_c}}\right) \le m_c\right) \\ &= 1 - \Phi\left(\frac{B_c - \sqrt{1-\rho_c}\Phi^{-1}(m_c) - \sqrt{\rho}\Phi^{-1}(x)}{\sqrt{\rho_c - \rho}}\right) \end{aligned}$$

and the cumulative joint distribution function of default rates:

$$\begin{aligned} \mathbf{G}(m_1, \ldots, m_C) &= \mathbb{P}\left(\mu_1 \le m_1, \ldots, \mu_C \le m_C\right) \\ &= \int_{\mathbb{R}} d\Phi(x) \prod_{c=1}^{C} \mathbb{P}\left(\mu_c(t) \le m_c \mid X = x\right) \end{aligned}$$

The likelihood is then obtained by differentiation of $\mathbf{G}$

3. Binomial estimator MLE1 of asset correlation (PD = 200bp, ρ = 25%)



C. Statistics of the MLE2 estimates (mixture of PD, ρ = 25%)

	Asymptotic		Binomial	
Statistics	**ρ (in %)**	**B**	**ρ (in %)**	**B**
Mean	16.9	−2.47	22.3	−2.53
Std. dev.	8.1	0.12	10.0	0.17
Median	15.5	−2.48	21.3	−2.53

D. Statistics of the MLE1 asset correlation estimates (two-factor model)

	Asymptotic			Binomial		
Statistics (in %)	ρ_1	ρ_2	ρ	ρ_1	ρ_2	ρ
Mean	19.9	12.9	6.5	19.9	10.7	7.5
Std. dev.	4.8	3.1	3.1	6.4	4.3	3.7
Median	19.5	12.6	6.3	19.4	10.3	7.2

To obtain our Monte Carlo estimates, a synthetic default database that is supposed to represent historical data is first generated. For the sake of simplicity, we will consider here a single probability of default (PD) for all the firms in our sample – recall we want a crude estimation of the bias. We consider a sample of maturity $T = 20$ years, which is about the same as in public databases. For each time t, the total number of firms N_t is supposed to be constant and equal to N.

Two kinds of estimators are studied: first, considering that the value of threshold $B = \Phi^{-1}(PD)$ is known, the only parameter to estimate is the asset correlation ρ. This full information estimator is called MLE1. Another possibility is to assume that the value of threshold B is unknown, and thus to estimate it jointly with ρ. The resulting limited information estimator is called MLE2. These estimators are calculated using both the binomial model and the asymptotic one.

□ **The bias within the one-factor model.** Here, the single-factor model is used, which is less demanding in terms of running time. The study is then extended to a two-factor model, to try to check the validity of our analysis closer to real-world conditions.

We assume that the asset correlation ρ is equal to 25%. Tables A and B present some elementary statistics on the distribution of MLE1 and MLE2, extracted from a synthetic sample of $N = 200$ firms. Gordy & Heitfield (2002) point out that the 'binomial' MLE2 – with limited information – presents a downward bias, and that the phenomenon is even more acute when the default probability decreases. The same conclusions are reached here. The 'binomial' MLE1 behaves approximately in the same way. Figures 2 and 3 show the distribution of MLE1 estimators, extracted from the binomial and asymptotic log-likelihood, when N varies. The 'asymptotic' estimator presents an important downward bias when the number of firms N is small, but it is decreasing with N. The behaviour of the 'binomial' estimator is better even in small sample conditions. However, we notice that the variance of the distribution is very important for both estimators. In our example, we observe that the dispersion decreases when the number of firms increases. We obtain similar results when the number of years T increases (see Gordy & Heitfeld, 2002, for the binomial estimator). However, it is important to notice that the 'asymptotic' estimator converges to the true value when both N and T tend to infinity, whereas the 'binomial' estimator converges to the true value only when T grows to infinity.

The previous simulations were based on the assumption that all the firms have the same default probability. Another case is now tested, where the firms in the synthetic sample belong to different rating groups, and thus have different default probabilities. For each date t, a sample of 200 assets values is simulated, corresponding to 100 firms with annual default probability of 20 basis points and 100 firms with annual default probability of 100bp. We want to test the robustness of our estimates when the rating groups firms belong to are unknown. The MLEs are thus calculated considering that all the firms have the same rating, hence the same probability of default and the same threshold B, which is unknown. The results are shown in table C. Compared with the binomial estimator, the 'asymptotic' estimator presents a larger downward bias but the variance in its distribution is lower.

□ **The bias within the two-factor model.** In practice, the number F of factors depends on the public database used and the number of risk classes, but it is typically around 15 when sectors are considered. Therefore, we need to check the behaviour of our estimates when the number of factors increases. Since the running time would be almost infinite for a two-factor model with 15 risk classes, we limit ourselves to checking the behaviour of the estimates extracted from a two-factor model with two risk classes. Again, what we are interested in are rough ideas of the biases in the estimations.

We simulate assets values from the following risk class correlations matrix:

$$\Sigma = \begin{pmatrix} \rho_1 & \rho \\ \rho & \rho_2 \end{pmatrix} = \begin{pmatrix} 20\% & 7\% \\ 7\% & 10\% \end{pmatrix} \qquad (17)$$

The size of our sample is still assumed to be $T = 20$ years. The number of firms is $N = 200$ for the two risk classes and all time t. We set $PD = 200$bp for all firms. We assume that PDs are known and calculate maximum likelihood estimates for parameters ρ, ρ_1 and ρ_2 only. Table D presents some statistics of the estimators.

The bias of the estimator seems to often be lower than the one obtained in the one-factor model. This is not immediately intuitive since the bias is a complicated function of all parameters (PDs, true correlations, etc). However, as suggested by one referee, this may be explained by observing that, in the two-factor case, we have more information for estimating each interclass correlation due to the presence of the (correlated) other class.

Estimation using S&P data

Rating agencies such as Moody's and Standard & Poor's annually provide databases of default rates sorted by category of obligors. Firms are grouped either by rating, sector of activity or geographical location. Analysing these historical series of data enables us to estimate the parameters of dependence of default events across each class of risk. In our study, we used S&P data giving default times series of firms grouped by sector of activity. Hence the expression 'class of risk' in our article always refers to this sectorial distribution (with no distinction between high-yield and investment-grade companies). To our knowledge, this is the first study to provide

[10] *We use 5,000 trials for all Monte Carlo experiments in this section*

E. Asymptotic and binomial MLE2 estimates of the asset correlations extracted from the S&P database

	$\bar{N}_c$	$\bar{\mu}_c$	Two-factor		Single-factor	
			Asymptotic	Binomial	Asymptotic	Binomial
Aerospace/automobile	301	2.08%	13.3%	13.9%	13.7%	11.6%
Consumer/service sector	355	2.37%	12.2%	10.6%	12.2%	8.9%
Energy/natural ressources	177	2.10%	23.2%	25.5%	16.2%	14.5%
Financial institutions	424	0.57%	17.0%	16.4%	12.0%	9.5%
Forest/building products	282	1.90%	18.1%	18.8%	28.6%	31.5%
Health	135	1.27%	12.9%	10.6%	13.1%	13.2%
High technology	131	1.66%	15.0%	16.4%	12.9%	10.6%
Insurance	166	0.61%	26.3%	34.3%	13.6%	17.8%
Leisure time/media	232	3.01%	13.8%	9.4%	17.2%	12.0%
Real estate	133	1.01%	43.2%	52.4%	48.7%	53.0%
Telecoms	100	1.91%	22.9%	29.1%	27.0%	34.0%
Transportation	146	2.02%	17.7%	11.1%	12.8%	10.4%
Utilities	206	0.43%	14.4%	18.7%	10.4%	17.5%
Inter-sector			7.2%	9.4%		

maximum likelihood estimates of default correlations within and across obligors grouped by industrial sector.[11]

The S&P database of defaults spans 22 years between 1981 and 2002. For each year, the database reports the number of firms by sector of activity and by rating grade, and the observed number of defaults. Firms are divided into $S = 13$ sectors of activity. In this study, we do not take into account the rating grade of each obligor, but we assume that sectors are homogeneous enough to have a unique asset correlation for all firms in the same sector.[12] The risk classes correspond then to these 13 sectors of activity. We report asset correlation estimates in table E (we use the MLE2 estimator, because we assume the thresholds are unknown).

We observe that the 'binomial' and 'asymptotic' estimators are relatively close in value. Yet a noticeable difference appears for some sectors, such as insurance, real estate, telecom or transportation. It seems that the greater differences occur when the number of firms in a particular sector is low (in table E, $\bar{N}_c$ represents the average number of firms by sector whereas $\bar{\mu}_c$ represents the average default rates by sector). This remark suggests that these differences may be caused by a convergence problem due to scarcity of data. If we compare these results with estimates based on the single-factor model, we observe that asset correlation estimates based on the two-factor model are slightly greater on average than the ones based on the single-factor model.

■ **Remark 1.** If we pool all the sector activities to define only one risk class, the 'binomial' and 'asymptotic' estimates of correlation are respectively 8.3% and 10.1%. By using rating grades as risk classes, estimates of correlation are of the same order (which confirms the results of Gordy & Heitfeld, 2002). Compared with results with sector activities as risk classes, default correlations are smaller. We think we may underestimate default correlations when we define risk classes as rating grades, because rating grades are not homogeneous enough to have a unique default correlation.

Conclusion

This article extends the framework of Gordy & Heitfield (2002), which calculates the MLE of asset correlations. Using a simplified version of the Merton factor model for modelling default events and under realistic assumptions, our assumptions enable us to estimate default correlations using the maximum likelihood approach with reasonable computational times. Two kinds of estimators are formulated. The first, the 'binomial' estimator, is based on the joint distribution of the number of defaults of different risk classes. The second one, the 'asymptotic' estimator, is based on the joint limit distribution of the default rates of different risk classes. This estimator imposes more restrictive assumptions, but it allows a significant simplification in the log-likelihood calculations.

Monte Carlo simulations are performed to assess the bias of the resulting estimators. With small samples, which is the practice with databases of default rates, the estimators may produce some bias. Unfortunately, we cannot ignore the fact that default data is very sparse and thus the distribution of the estimators may be significantly dispersed. As a preliminary study, we estimate default correlations extracted from the S&P database. We obtain relatively closed results for asymptotic and binomial estimates. Moreover, using a one-factor model may underestimate default correlations. Similar results are obtained when we use rating grades as risk classes. We conclude that default correlation estimators are highly dependent on the definition of risk classes, and they may be underestimated. However, we should be cautious since Monte Carlo simulations suggest that default correlation estimation is a very hard task. ■

Paul Demey, Jean-Frédéric Jouanin, Céline Roget and Thierry Roncalli all work in the groupe de recherche opérationelle within the risk management group at Crédit Agricole. **They would like to thank Antoine Frachot, former head of the groupe de recherche opérationelle, for helpful comments and advice, and Gaël Riboulet, who worked with them on an early draft. They are also grateful to Pierre-Henri Floquet, Christian Redon and Frédéric Zana from the Crédit Agricole-Calyon finance division for stimulating discussions. They also thank Olivier Renault from S&P for providing some data, and two anonymous referees for their suggestions and stimulating discussions.** **Email: paul.demey@creditlyonnais.fr, jean-frederic.jouanin@creditlyonnais.fr, celine.roget@creditlyonnais.fr, thierry.roncalli@creditlyonnais.fr**

[11] *A pioneer study is that of Frye (2000), which estimates default correlation in a one-factor model. Gordy & Heitfield (2002) provides maximum likelihood estimates by rating classes (see also Düllmann & Scheule, 2003). The closest study is the one performed by Servigny & Renault (2003). They provide estimates of default correlations within and across sectors of activity using the relationship between joint default probabilities and default correlations*

[12] *We assume that the PDs are the same within a sector, which is not the case in practice. It is a practical compromise to estimate default correlations*

REFERENCES

Düllmann K and H Scheule, 2003
Determinants of the asset correlations of German corporations and implications for regulatory capital
Working paper

Finger C, 1999
Conditional approaches for CreditMetrics portfolio distributions
CreditMetrics Monitor, April

Frye J, 2000
Depressing recoveries
Risk November, pages 108–111

Gordy M and E Heitfield, 2002
Estimating default correlations from short panels of credit rating performance data
Federal Reserve Board

Servigny A and O Renault, 2003
Correlation evidence
Risk July, pages 90–94