

Default correlations derived with an averaging model

Luc Hoegaerts

Model Validation & Development
Fortis Central Risk Management

April 4, 2007

Abstract

In the estimation of the credit loss distribution of a portfolio, the correlation of default between obligors has a significant impact, increasing the bank Economic Capital requirements. Since it is practically difficult to measure default correlations, they are commonly inferred from default probabilities of the obligors and correlation of the underlying assets of the obligors. Moreover, asset return data is of higher quality and availability than credit default data. In this shortpaper we discuss an averaging model that allows to create inter and intra default correlation between groups of similar clients from a large representative data set. This approach assumes that all the elements in the correlation matrix for a portfolio can be approximated by the average correlation of their peer groups in the data set. The similarity is based on a hierarchical clustering according to region, sector, rating and asset size. We describe the method from a practical perspective and discuss results in comparison with a single factor model.

1 Introduction

Credit Risk is the risk that a borrower will be unable to pay back his loan. A bank can quantify its portfolio credit risk through the measurement of the variability of the portfolio credit loss.

A loss density distribution can be associated to the portfolio, expressing the probability per aggregate loss amount over a certain horizon, eg 1 year. The basic building blocks of credit loss are Exposure-At-Default (EAD), Loss-Given-Default (LGD) and Probability of Default (PD). These main parameters at the level of single obligors determine the loss distribution.

However, considering interaction and dependency between obligors, at the level of multiple obligors, the probabilities of joint default also exert a significant influence on the shape of the loss distribution. The joint default probabilities are largely dominated by the *default correlation* between pairs of obligors. Thus default correlations constitute an essential component of credit risk.

Estimating default correlations is not so straightforward for the simple reason that data is very scarce: defaults occur not frequently, and joint defaults of pairs even less. Historical correlations based on direct observations require huge databases of long records of all kinds of firms. In practice, one resorts more often to a modelling approach. Although one obtains the estimates in an indirect manner (via asset return data), a model can forecast future correlations equally well or even better [7].

There are two major modelling approaches: averaging models and factor models. Average models assume that all the pairwise correlations for a sample of firms can be approximated by the average correlation of their peer group. Factor models assume that co-movements among asset returns are driven by one or more factors.

In this paper we deal with an averaging model that groups peers based on four characteristics: region, industry, credit rating and asset size. For illustration, we compare with a simple factor model that assumes one factor per group. Empirical results are based on a KMV dataset and turn out to be rather similar.

This shortpaper is organized as follows. Section 1 gives an introduction to the paper. Section 2 describes the parameters of credit risk and some assumptions of the correlation model. Section 3 describes the averaging model. In section 4 we describe a factor model. In section 5 we provide the clusters definition and results on a KMV dataset. Section 6 finally concludes.

2 Description of the loss model

A loss can statistically be considered as a random variable (rv), which is dependent on numerous other variables. In the Credit aggregation context, where the lender needs to account for risk of default of his obliger within the repayment term, one discerns commonly three determining factors: an amount at risk at the point of default, the degree of security and the likeliness of default.

2.1 Loss decomposition

In more formal terms we can quantify the factors as follows. Consider a portfolio of n credit risks. In this default model, the random loss L_i is decomposed into three random variables:

$$L_i = I_i (EAD)_i (LGD)_i \quad (1)$$

1. The random variable I_i is defined as the indicator variable which equals 1 if risk i leads to failure in the next period, and 0 otherwise. This is known as a Bernoulli rv and one defines:

$$I_i = \begin{cases} 1 & \text{with probability } q_i \\ 0 & \text{with probability } 1 - q_i, \end{cases} \quad (2)$$

where q_i is the *probability of default (PD)*. Remark that

$$E(I_i) = q_i \quad \text{and} \quad \text{var}(I_i) = q_i(1 - q_i). \quad (3)$$

2. The random variable $(EAD)_i$ denotes the *Exposure-At-Default* expressed in some monetary units. It is the maximal amount of loss on risk i , given that default occurs.
3. The random variable $(LGD)_i$ denotes the *Loss-Given-Default* of risk i in percentage terms. It is the percentage of the loss on policy i , given that default occurs.

The *Aggregate Portfolio Loss* S is the sum of the (relative) losses on the individual credit risks during the reference period:

$$S = \sum_{i=1}^n L_i = \sum_{i=1}^n I_i (EAD)_i (LGD)_i. \quad (4)$$

Each portfolio loss realization consists in fact of obligors that default or not, according to their specific parameters (borrowed amount, security, quality,

economy, randomness, etc). By simply counting such occurrences, one can determine the probability per possible aggregated loss amount, which is the portfolio *loss density distribution*.

Estimating the distribution of S is not so straightforward. The loss distribution depends on the underlying marginal distributions of LGD , PD and EAD . The specific form is ruled by the nature of the variables; whether they are deterministic or stochastic and one must unambiguously describe their (inter- and intra-) dependency structure.

2.2 Loss distribution

Some important measures are central in the assessment of the risk of a portfolio.

- *Expected Loss* (EL) is the expected level of credit losses, over the one year time horizon. Actual losses for any given year will vary from the EL, but EL is the amount that the bank should expect to lose on average. Expected Loss should be viewed as a cost of doing business rather than as a risk itself.
- The real risk arises from the volatility in loss levels. This volatility is called *Unexpected Loss* (UL). UL is defined statistically as the standard deviation of the credit loss distribution.
- The *Economic capital*, which corresponds to a quantile in the tail of the loss distribution, minus the Expected Loss.

To obtain the portfolio EL , it suffices to add the stand-alone expected losses because eg for a loss L_1, L_2 and L_3 :

$$EL(L_1 + L_2 + L_3) = EL(L_1) + EL(L_2) + EL(L_3). \quad (5)$$

However, to obtain the portfolio UL , adding the stand-alone variances is not enough information because

$$var(L_1 + L_2 + L_3) = var(L_1) + var(L_2) + var(L_3) + 2cov(L_1, L_2) + 2cov(L_1, L_3) + 2cov(L_2, L_3) \quad (6)$$

Hence, the covariances between all possible pairs of losses is required. Remark that instead of covariance, one refers more often to correlation, which is the same, apart from a normalization

$$\rho_{L_i L_j} = \frac{covar(L_i, L_j)}{\sqrt{var(L_i)}\sqrt{var(L_j)}} = \frac{E(L_i L_j) - E(L_i)E(L_j)}{\sqrt{var(L_i)}\sqrt{var(L_j)}}. \quad (7)$$

Moreover, studies on real default data show also that the magnitude and influence of these correlation terms in eq. (6) are not negligible, which is in a statistical sense logical since it entails second-order information of the loss distribution. To estimate quantiles for Ecap exactly, in fact higher-order information like multiple default correlations would be further required.

In case the *LGD* is considered to be a deterministic rv, the correlation between loss pairs equals the correlation between default pairs:

$$\rho_{L_i L_j} = \rho_{I_i I_j}. \quad (8)$$

And if *LGD* is stochastic, then it can be shown that one is function of the other. In either case, it turns out that default correlations are an additional basic element of credit risk.

2.3 Default correlations in a Merton based default model

The default of a firm is an event that depends on many factors. Attempts to describe and to find out what triggers a default is the subject of structured default models. The majority of the models are inspired by Merton [8] where default occurs if the value of the firms asset is less than its callable liabilities. The success of this type of approach is due to its analytical tractability, easy economic interpretation, and basic inputs.

The central assumption is that the level of the (log) asset value A_i of a firm i over time is assumed to fluctuate according to a normal distribution with a certain mean and volatility. This allows to compute the terms in the default correlation

$$\rho_{I_i I_j} = \frac{E(I_i I_j) - E(I_i)E(I_j)}{\sqrt{\text{var}(I_i)}\sqrt{\text{var}(I_j)}} = \frac{E(I_i I_j) - q_i q_j}{\sqrt{q_i(1 - q_i)}\sqrt{q_j(1 - q_j)}} \quad (9)$$

indirectly via a model based primarily on asset values.

The PD q_i can then be analytically expressed as the probability of a standard normal random variable falling below some critical value. The expectation of the product $E(I_i I_j)$ in eq. (9) becomes equal to the joint default probability $P(I_i = 1, I_j = 1)$, because the default rv are indicator variables. And the joint default probability can be expressed as the probability of two correlated standard normal random variables both falling below given critical values:

$$P(I_i = 1, I_j = 1) = \int_{-\infty}^{\Phi^{-1}(q_i)} \int_{-\infty}^{\Phi^{-1}(q_j)} N(0, \rho_{A_i A_j}) dx dy \quad (10)$$

where $\Phi^{-1}(\cdot)$ is the inverse univariate standard normal cumulative distribution, $N(0, \rho)$ is the bivariate standard normal distribution and $\rho_{A_i A_j}$ is the correlation between the asset values.

The integral can be interpreted as the area under the joint probability distribution of asset values in which the values of assets of both firms are less than their respective default points, see figure (1).

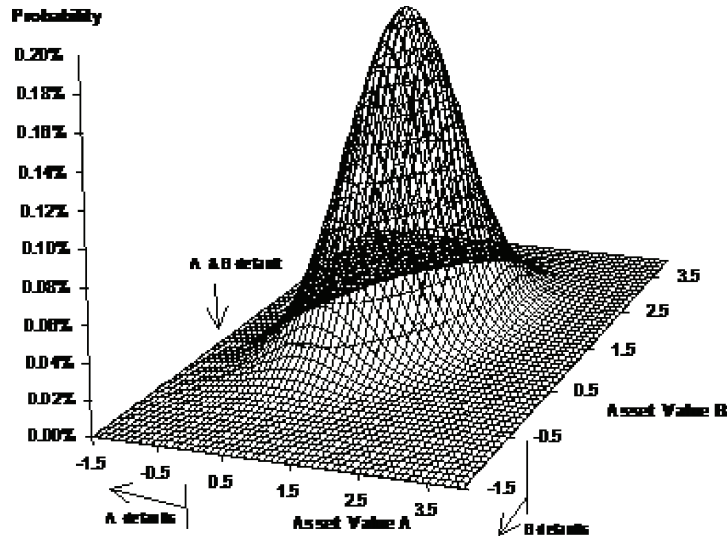


Figure 1: The joint default probability between two (correlated) assets.

The default correlation becomes in this model a function of the asset correlation and the PD's. This is convenient as data on asset values is relatively much more available and of higher quality than default data.

3 Description of the cluster averaging model

Ideally, one would compute asset correlations between each of the clients of the portfolio. In practice however, asset values over a historical period are seldom available for *each* client.

In order to reduce the granularity level, one groups the clients together in clusters. One assumes that assets of all clients within and between clusters are similarly correlated. In this approach clients are then simply each assigned to a bucket and one can use the correlation between the buckets instead.

The intra-cluster correlation between any two firms in the same cluster is calculated by taking the average of all pair-wise correlations within the cluster. The inter-cluster correlation between any two firms in different clusters is calculated by taking the average of all pair-wise correlations between the two clusters. Hence the name *averaging*.

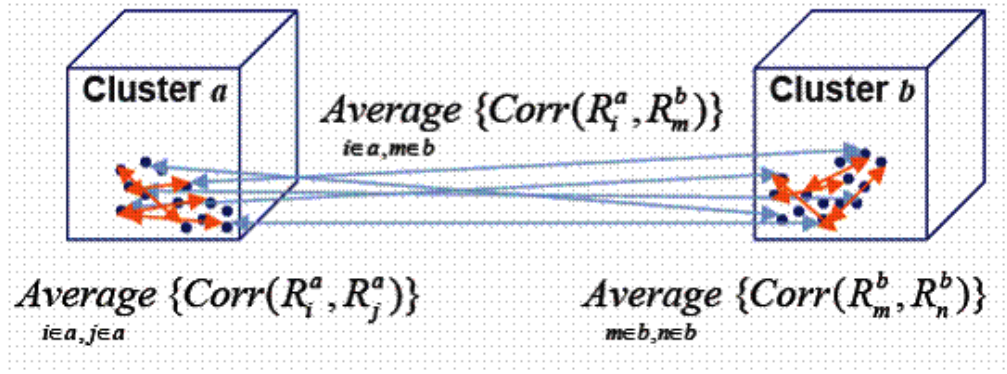


Figure 2: Schematic representation of the averaging model to derive asset correlations for clusters of firms.

Of course the next question is how to define the clusters. Grouping similar clients comes down in practice to identifying which ones have similar (driving) attributes like region, sector, rating and asset size. It is along these 4 dimensions that one tries to uniformly chunk a large dataset up in cluster parts. The number of categories to be defined in each dimension must be balanced: it must allow granularity, but sample size of each bucket should be more or less equal. Here we closely follow the choices made in [6].

4 Description of a single factor model

The clusters are defined likewise as in the averaging model, but each firm's standardized log asset return is assumed to follow a factor model. We draw from [1]. There is a factor F^a per cluster a to which all its firms' asset value are related by

$$R_i^a = \beta_a F^a + \sqrt{1 - \beta_a^2} \epsilon_i^a \quad (11)$$

where β_a is a correlation between any firm and the index of the cluster, factor F^a of cluster a is standard normally distributed, and the firm-specific variable ϵ_i^a is iid standard normally distributed. Further the cluster factor and the firm-specific term are independent of each other. The classical one factor model (as in Basle II) assumes that the factors are independent of each other. Here the multi sector factor model does not assume independent factors.

Each firm is uniquely assigned to one cluster a and a factor index F^a is constructed as an average based on unweighted log asset returns. The index-to-index correlation ρ_{ab} between any two indices is immediately calculated as $\text{corr}(F^a, F^b)$. The firm-to-index correlation β_a between any firm and its

index F^a is calculated by averaging over individual firm-to-index correlations

$$\beta_a = \frac{1}{\#A} \sum_{i \in A} \text{corr}(R_i^a, F^a) \quad (12)$$

The intra-cluster correlation between any two firms in the same cluster a is then calculated as β_a^2 . The inter-cluster correlation between any two firms in different clusters a and b is calculated as $\beta_a \rho_{ab} \beta_b$.

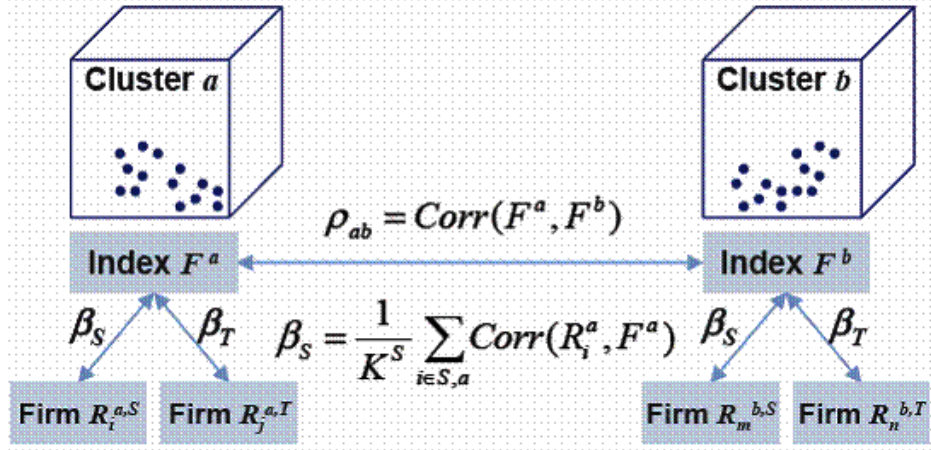


Figure 3: Schematic representation of the factor model to derive asset correlations for clusters of firms.

5 Experiments

In our example, Data from Moody's KMV Credit Monitor was used, for listed and unlisted *corporates* for a historical period of 109 months from 31/3/1997 till 30/04/2006. In total a selection of 64 843 firms was made with database information on country (71), sectors (61), EDF series (probability that firm value will fall below a pre-defined threshold within a year) and asset value series.

5.1 Cluster definition

A definition of clusters was based on four dimensions: geographical region, industry sector, credit quality and asset size. Table 1 summarizes the clustering criteria:

From these corporates, a small part was set aside and considered as representative for *Publics* (very large firms with high credit quality and asset

attribute	nr	criterium
Geographical region	3	Europe; North America; Asia
Industry sector	7	Construction/Manufacturing(Hard); Manufacturing Soft; Transport Manu- facture, Communications and Utilities; Wholesale Trade and Distribution; Retail Trade and Sales; Financial Services; Services and Human Resources
Credit quality	4	0-0.85-2.15-3.97-20% Probability of De- fault (rating grade)
Asset size	4	quartiles of the region/sector/rating clus- ter asset sizes histogram

Table 1: Clustering criteria in the experiment.

size) and *Individuals* (very small firms with low credit quality and asset size). The lack of a reliable database on these extremal client groups is the main justification of the use of such approximations.

The raw data of asset values has been preprocessed for outliers. Due to debt issues (or buy-back or other corporate actions) there can be jumps in the raw asset values. Since these are not economically meaningful, the asset data series should be adjusted for such effects, otherwise resulting default correlations might be underestimated. This involved adjusting asset values for changes in debt values and then removing the top 0.25% and bottom 0.25% of the returns.

For the rating the last recorded EDF value of KMV was used. For the asset size value an average had been made over the series. Asset size bands were determined by taking the quartiles of the distribution of asset sizes in a region-sector-rating cluster.

In case that some clusters had too few firms in it, it was decided to group on the level of rating some clusters together and substitute that grouped bucket for the nearly empty buckets. In total the clustering resulted in 340 buckets for which asset values were calculated.

In order to have an asset correlation matrix that is positive semi-definite, one may apply (i) truncation of rank-one terms with negative eigenvalues in its dyadic decomposition, or (ii) adding a scaled unity matrix to the diagonal matrix in the eigenvalue decomposition.

In the computation of correlations, any null or missing values must always be treated with care as the influence can be significant. Especially the factor

model has some extra bias from zero asset values. Creation of a sector index should be only based on complete time series (ideally 107 months) of non-zero asset values, otherwise one obtains artificially lower correlations (on average relatively 1%). Effectively an index was based on less companies than are present in the cluster, and a max of 5 zeros out of 107 was allowed.

For the computation of the default correlations, the asset correlations were employed in eq. (9) with all combinations of 25 PD levels. This resulted eventually in 2200 default correlation buckets.

5.2 Results

For the purpose of comparison it is easier to report on the asset correlations, instead of the default correlations. We can make then the following observations:

- Previous studies, such as [5], have identified trends with respect to asset correlations, which are in line with common expectations. In particular, one can say that asset correlations increase with increasing asset size and that asset correlations decrease with increasing default probability. In figure 4 these relations are found back when we show asset correlations with respect to asset size and rating colour classes.
- The averaging model yields on average an asset correlation level of 5.8%, while the factor model produces slightly lower asset correlations of 5.1%. Remark that the standard deviation is 0.2% and the difference is statistically significant, but practically the results are highly similar. We summarize some results in table 2. In figure 5 we show corresponding boxplots of the asset correlations obtained by the averaging model and the factor model. In figure 5.2 we show corresponding histograms which show that both distributions are similar and slightly skewed towards higher correlations. In figure 8 the distribution of differences shows again that the results match overall rather well.
- On an absolute scale the asset correlation levels are considered as on the low side (considering around 15% on average as reference), but certainly within the realistic range of 3-20% as can be found in literature. A more extensive discussion of the levels can be found in [6]. Possible reasons could be that (i) monthly correlations bias the results downward, as correlations in general tend to lower with shorter time frames; (ii) Pearson correlation is inappropriate as a measure for linear correlation when outliers are present, as one should use then alternative robust measures like Kendall, Spearman or Biweight Midcorrelation.

- There is again high similarity to be noted between the models when comparing histograms of intra asset correlation. From figure 5.2 we see that for both models the intra cluster asset correlations is practically centered around the same mean (9.8% with a standard deviation of 0.3%).
- The rank of the asset correlation matrix of the averaging model is fully 218, while the one of the factor model is 107. This shows that the factor model has more interdependency induced by the factors. There is considerable difference in matrix structure, which is confirmed by the relative 2-norm distance between the matrices which amounts to 16%. The correlation matrix of the factor model is already positive semi-definite, while the one of the averaging model requires a small correction.

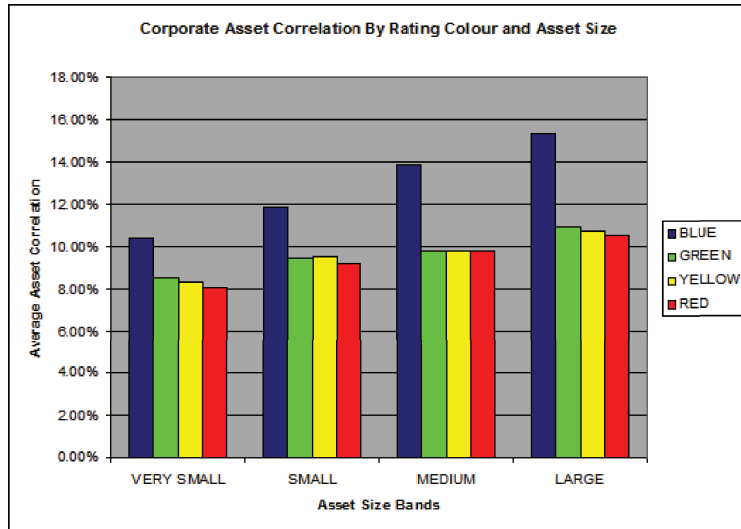


Figure 4: Asset correlations versus asset size and rating colour classes for corporates. The results confirm the intuition that asset correlations increase with increasing asset size and that asset correlations decrease with increasing default probability.

average asset correlation	averaging model	factor model
Corporates	9.8	9.7
Individuals	4.8	7.4
Publics	15.4	16.6
global intra	9.8	9.7
global inter	5.8	5.1

Table 2: Averaged asset correlation results per segment (in %).

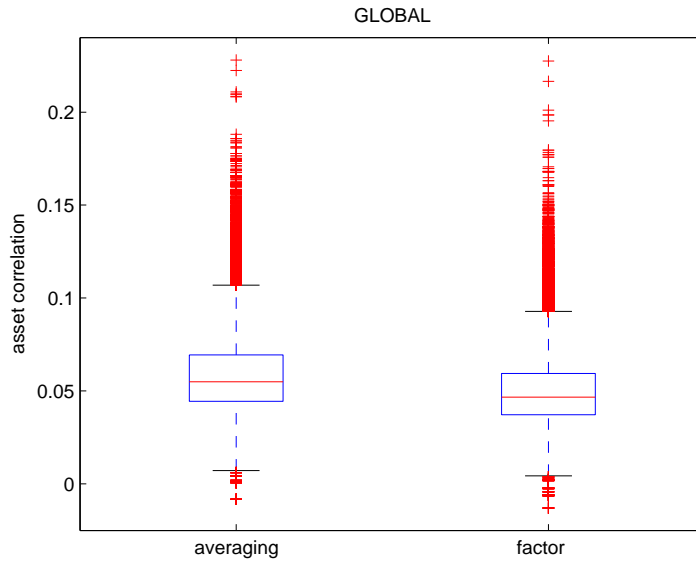


Figure 5: Boxplots of the asset correlations obtained by the averaging model and the factor model. The difference is statistically significant, but practically the levels of both models are highly similar.

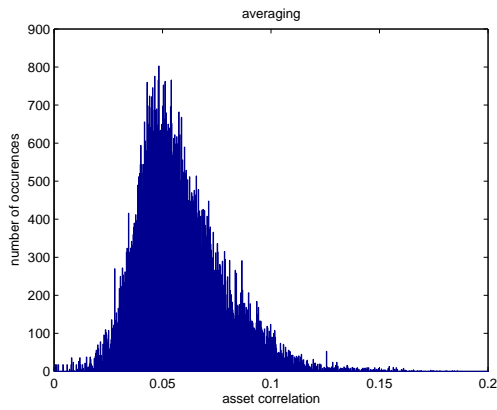


Figure 6: Histogram of overall asset correlations of the averaging model. Slightly skewed towards higher correlations.

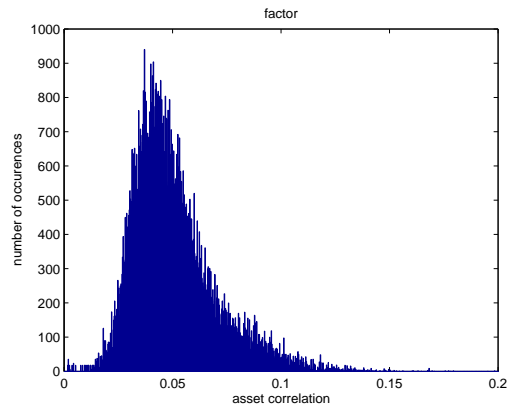


Figure 7: Histogram of overall asset correlations of the factor model. Slightly skewed towards higher correlations.

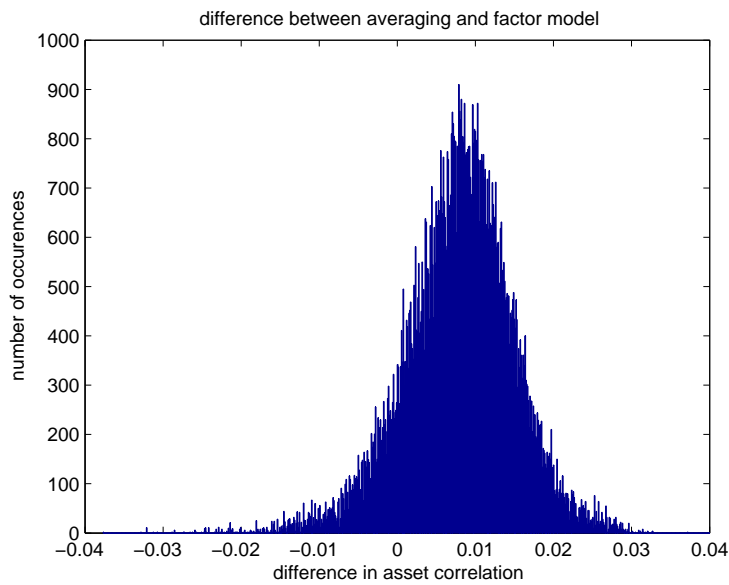


Figure 8: Histogram of the differences between asset correlations of the averaging and factor model. On average there is a 0.8% difference.

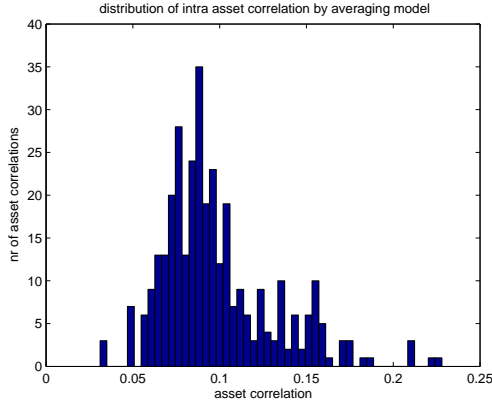


Figure 9: Histogram of intra asset correlations for the averaging model.

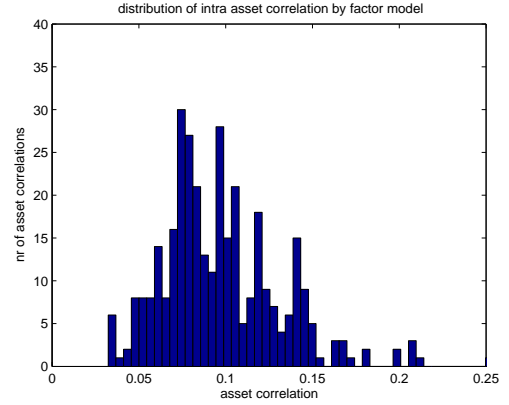


Figure 10: Histogram of intra asset correlation for the factor model.

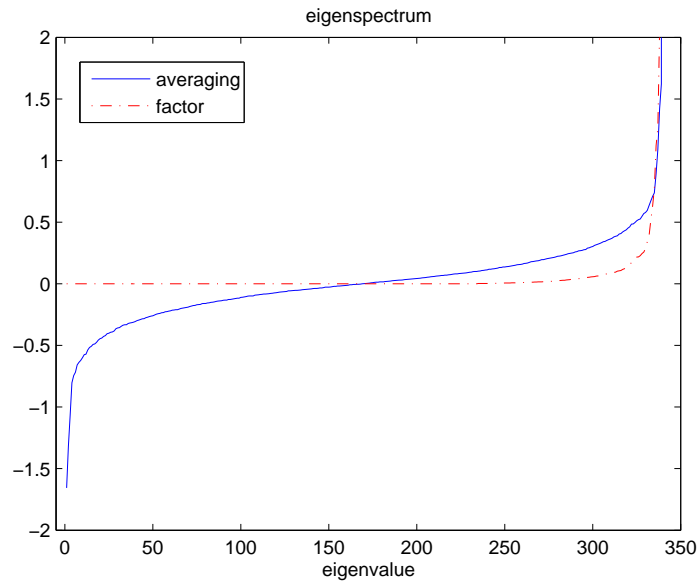


Figure 11: The eigenspectrum in ascending order (ranging from -2 to 20) for the averaging model (full line) and the factor model (dash-dotted line). The rank of the asset correlation matrix of the averaging model is fully 218, while the one of the factor model is 107. This shows that the factor model allows for more interdependency induced by the factors. There is considerable difference in structure, which is confirmed by the relative 2-norm distance between the matrices which amounts to 16%.

6 Conclusions

In this paper we described an approach to derive default correlations, important parameters that allow credit risk estimation and portfolio optimization. We discussed an averaging model that groups peers based on four characteristics: region, industry, credit rating and asset size. For illustration, we compared with a simple factor model that assumes one factor per group.

The factor model makes additional assumptions (parametric approach) and the dependency through a factor may result in slightly biased outcomes. On one hand, the averaging model uses minimal assumptions (semi-parametric approach) and this results in a more robust, data-driven estimate. On the other hand, the factor model is less computationally intensive.

Overall, we can conclude that asset correlations are mainly in the range between 3% and 10%. Further inter and intra asset correlations are highly similar for both models.

Acknowledgements

Any views expressed within this document represent those of the author and not necessarily those from Fortis. The author wishes to thank Steven Vanduffel, Andrew Chernih and Ivan Goethals for helpful discussions.

References

- [1] K. Duellmann, M. Scheicher & C. Schmieder (2006). Asset correlations and credit portfolio risk - An empirical analysis.
- [2] Frey, A. McNeil & M.A. Nyfeler (2001). Modelling Dependent Defaults: Asset Correlations Are Not Enough! Working Paper, Department of Mathematics, ETHZ, Zurich.
- [3] A. Pitts (2004). Correlated Defaults: let's go back to the data. Risk Magazine, June.
- [4] A. de Servigny and O. Renault, Default correlation : empirical evidence
- [5] J.A. Lopez (2002), The Empirical Relationship between Average Asset Correlation, Firm Probability of Default and Asset Size.
- [6] Chernih, A., Vanduffel, S. & Henrard, L. (2006). Asset Correlations: Shifting Tides?
- [7] B. Zeng & J. Zhang (2001). An Empirical Assessment of Asset Correlation Models. Moody's KMV Research Paper.
- [8] Merton, R. (1974): On the Pricing of Corporate Debt: The Risk Structure of Interest Rates, Journal of Finance, vol. 29, pp. 449-470.