Default Correlations and Large-Portfolio Credit Analysis*

Jin-Chuan Duan† and Weimin Miao‡

(First version: February 7, 2015; This Version: February 12, 2015)

Abstract

A factor model for short-term probabilities of default and other corporate exits is proposed for generating default correlations while permitting missing data. The factor model can then be used to produce portfolio credit risk profiles (default-rate and portfolio-loss distributions) by complementing an existing credit portfolio aggregation method with a novel simulation-convolution algorithm. We apply them on a global sample of 40,560 exchange-listed firms and focus on three large portfolios (the US, Eurozone-12 and ASEAN-5). Our results show the critical importance of default correlations. With default correlations, both default-rate and portfolio-loss distributions become far more right-skewed, reflecting a much higher likelihood of defaulting together. Our results also reveal that portfolio credit risk profiles evaluated at two different time points can change drastically with moving economic conditions, suggesting the importance of modeling credit risks with a dynamic system.

Keywords: Default, factor, bottom-up, missing data, convolution.

*This paper originated from several discussions with Rowan Douglas and David Simmons of Willis on how the CRI database of Risk Management Institute, National University of Singapore can be intelligently applied to help insurers manage their credit portfolios. This project benefits from RMI-CRI being a member of the Willis Research Network at the time.

†Duan is with the National University of Singapore (Risk Management Institute, Business School and Department of Economics). E-mail: bizdjc@nus.edu.sg.

‡Miao is with the National University of Singapore (Risk Management Institute). This research began when Miao was a Willis Research Fellow. E-mail: rmimw@nus.edu.sg.

1 Introduction

Default correlations are crucial information for many practical applications that involve more than one obligor. Credit derivatives such as CDS, for example, stipulate payment obligation by the protection seller in the event that the reference obligor defaults. When called upon, the protection seller may also default. Such a double default situation cannot be adequately analyzed with marginal default distributions alone. Knowledge of default correlation between the CDS reference obligor and the protector seller is critical to the analysis of CDS. A CDO with several default tranches will, for example, require one to know default correlations of all obligors underlying the collateral asset pool. A typical wholesale banking book comprises loans to many corporates. Since defaults can be clustered, one cannot adequate analyze the credit risk of a banking book without having a good handle on default correlations.

Broadly speaking, the models for default correlations fit into two categories: the bottom-up and top-down approaches. Bottom-up approaches mainly include copula models, conditionally independent defaults (CID) models, contagion models, and the model of common default events. The copula approach first appeared in Vasicek's (1987) one-factor Gaussian copula model and later in Li's (2000) Gaussian copula pricing of collateralized debt obligations (CDOs). The Gaussian copula approach was later realized to lack the needed tail dependence. Different generalizations were proposed to introduce tail dependence by employing different copulas such as Frey and McNeil (2001), Schönbucher and Schubert (2001), Hull and White (2004), and Crane and van der Hoek (2008), or using randomized recovery and factor loadings as in Anderson and Sidenius (2005). Further development into dynamic copula models for credit analysis has also emerged. The copula approach in general has to choose the copula in a rather *ad hoc* manner due to scarcity of default data. CID models assume that defaults of different obligors are driven by some common factors, observable and/or latent. This idea has been worked into both reduced-form and structural credit risk models; for example, Duffee (1999), Finger (1999), Schönbucher (2001), Driessen (2005) and Duffie, *et al* (2009). Contagion models incorporate counterparty risk; that is, default of one obligor can trigger defaults of related parties. Examples abound; for example, Davis and Lo (1999), Jarrow and Yu (2001), Giesecke and Weber (2004), and Yu (2007). Finally, default correlations can be modeled through common events as opposed to correlated default intensities/probabilities with an example of Duan (2010). In contrast, top-down approaches attempt to directly characterize portfolio credit risk without modeling default risks of individual obligors. Some examples are Schönbucher (2005), Hurd and Kuznetsov (2007), Azizpour, *et al* (2011), and Giesecke, *et al* (2011). The main drawback of top-down approaches is the assumption of homogeneous credit risks among the individual obligors of a credit portfolio. Readers may refer to the review article by Albanese, *et al* (2013) for more information on the literature.

Generally speaking, our model for default correlations falls into the CID type, but is unique in several ways. We propose a factor model on short-term probabilities of default (PDs) and probabilities of other exits (POEs) for a very large pool of obligors, and use this factor model along with factors dynamics to generate random future short-term PDs and POEs. They in turn can be combined through the standard survival-default formula to deduce long-term PDs for individual

obligors and to generate default correlations among obligors. Because survival probabilities hinge on both PDs and POEs, corporate exits, such as a merger/acquistion, need to be considered as argued in Duffie, *et al* (2007) and Duan, *et al* (2012). Default correlations in our model arise from correlated future short-term PDs and POEs due to common factors. The common factors in our model is a combination of some predetermined aggregated short-term PDs and POEs, globally or industry-wise, and some extracted factors by a principal component analysis. The predetermined common factors are meant to capture the global and industry-specific credit cycles, whereas the extracted factors are used to pick up leftover commonality in PDs and POEs. Missing data are typical due to initial public offerings, bankruptcies, mergers/acquisitions, etc., and our principal component analysis positively deals with missing data.

If the factor model were solely estimated to the historical time series of one-month PDs and POEs, the information embedded in term structures of PDs for individual obligors would not have been utilized. Our proposed method is to further calibrate parts of the factor model by matching the deduced multiperiod PDs to their observed counterparts, and the calibration only needs to be performed at the time of portfolio credit analysis.

The actual factor model estimation and calibration performed in this paper use the PDs and POEs taken from the "public-good" Credit Research Initiative (CRI) database at the Risk Management Institute (RMI) of National University of Singapore. As of January 2015, RMI-CRI produces daily updated PDs and POEs, ranging from 1 month to 5 years, on over 60,000 exchange-listed firms in 116 economies around the world, and make them freely accessible. The quantitative model underlying the RMI-CRI PDs and POEs is the forward-intensity model of Duan, *et al* (2012), in which the occurrence of defaults and other types of corporate exits are modelled by two independent, after conditioning on stochastic intensities, Poisson processes. The forward intensities for either default or other exits are functions of some macroeconomic factors and firm-specific attributes. PDs and POEs for any future period of interest can then computed with these forward intensities. For further technical details, readers are referred to the RMI-CRI Technical Report (2014).

With the default correlation model in place, a bottom-up approach to computing default-rate and portfolio-loss distributions of any subset of obligors taken from the overall pool becomes possible. In a nutshell, the bottom-up approach comprises three steps: (1) simulate future paths of 1-month PDs and POEs for the target group of obligors to some horizon of interest, and with which default correlations are generated, (2) conditional on the simulated paths, the corresponding default-rate distribution is produced using the convolution algorithm of Duan (2010) whereas the portfolio-loss distribution is generated with a novel simulation-convolution algorithm proposed in this paper, and (3) repeat the simulation many times and average the simulated default-rate and portfolio-loss distributions to obtain the desired portfolio credit risk profiles.

We apply the default correlation model and credit portfolio aggregation tools on a subsample of 40,560 exchange-listed firms in the RMI-CRI database. The subsample consists of all firms in the RMI-CRI database that satisfy the selection criterion of having at least 60 months of PDs and POEs.

Three large portfolios (the US, Eurozone-12 and ASEAN-5) are used to show the critical importance of default correlations. With default correlations, both default-rate and portfolio-loss distributions become far more right-skewed, reflecting a much higher likelihood of defaulting together. Our results also reveal that portfolio credit risk profiles evaluated at two different time points (September 2008 and December 2014) can change drastically with moving economic conditions, suggesting the importance of modeling credit risks with a dynamic system.

2 Method to Construct Default Correlations

Our bottom-up approach to generating default correlations relies on multiple time series of short-term PDs and POEs. In addition to PDs, POEs are critical to deducing multiple-period default probabilities and their correlations, because survival probabilities hinge on how likely default or other corporate exits, such as a merger/acquistion, will occur. Duan, *et al* (2012) showed in their Table 1 that POEs are not negligible and they could be easily over ten times of PDs for US public firms. Our PDs and POEs are taken from the CRI database of Risk Management Institute (RMI) at National University of Singapore. Table 1 characterizes the number of firms of the RMI-CRI database over the sample period from December 1990 to December 2014 on a monthly frequency. A firm included in our sample must have PDs and POEs for at least 60 months over the sample period. The numbers under "union" indicate the counts of firms in each category that satisfy this selection criterion. The total number of firms in our sample is 40,560, which is smaller than over 60,000 firms in the RMI-CRI database. This reflects the fact that some firms are fairly new and others did not survive long enough. The table shows that the data sample used in our study contains a substantial number of firms in each industry over the entire sample period. Even in the "diversified" category, the minimum number of firms in a single month is 6.

If one faces a handful of obligors, default correlations are relatively easy to estimate as long as time series of PDs and POEs are available. With a large portfolio of several hundred or thousand obligors, one would need to impose some structure in order to handle default correlations even just over one basic period. As Table 1 shows, the number of obligors are way larger than that of time series observations, which certainly implies singular default correlation matrix. Adding to the complication is missing data. Missing data are expected simply because some firms are new and others have already ceased to exist. Missing data are an innate feature of default dataset, and handling default correlations must therefore also confront them. A natural solution for a high-dimensional data is to employ a sensible factor model, but one still needs to address missing data. In the following, we propose an intuitive factor model to handle over 40,000 exchange-listed firms in our dataset, and tackle the missing data problem in estimation. With this universal factor model is place, the bottom-up approach can be literally applied to any credit portfolio that is a subset of public firms in the world.

Our approach to default correlations comprises three steps: (1) Identify a set of predetermined credit risk factors along with factor extraction via a principal component analysis with the missing-

Table 1: Summary statistics on the number of firms monthly over the sample period

Period: December 1990 to December 2014					
Sector	**Number of firms**				
	mean	median	min	max	union
Financial	3305	3588	380	4940	6857
Basic material	1881	1917	156	2975	3854
Communications	1290	1555	176	2017	2800
Consumer (cyclical)	3418	3946	436	4662	6524
Consumer (noncyclical)	3296	3780	499	4719	6594
Diversified	258	294	6	342	490
Energy	773	708	156	1244	1660
Industrial	4239	4694	583	6079	8015
Technology	1409	1691	249	2095	2961
Utilities	410	414	147	572	771
Total	20290	22751	2797	29482	40560

The data sample comprises firms whose historical PDs and POEs are available for at least 60 months over the sample period, and they are based on the results of the RMI-CRI January 2015 calibration.

data capability; (2) Estimate the time series dynamics of the predetermined credit risk factors; (3) Further calibrate the model to the term structure of PDs at the time of application to take advantage of the information embedded in longer-term PDs.

2.1 Factor extraction and the factor model

Let n_t be the total number of firms with PDs and POEs in the RMI-CRI database at time t, and T be the last time point of the data. Let $p_{i,t}(l)$ and $q_{i,t}(l)$ denote the l-month PD and POE of firm i at month t for $i = 1, \cdots, n_t$ and $t = 1, \cdots, T$. By comparing PDs and POEs in Table 2, one notices much larger POEs vis-a-vis PDs for the whole sample and any industry, confirming our earlier assertion on the critical importance of factoring in other corporate exits. Table 2 also shows that the 1-month PDs and POEs are pretty right-skewed. We thus transform these 1-month PDs and POEs:

$$P_{i,t} = \ln\{-\ln[1 - p_{i,t}(1)]\} \quad \text{and} \quad Q_{i,t} = \ln\{-\ln[1 - q_{i,t}(1)]\}.$$

The above transformation is a natural choice because the forward intensity functions in Duan, *et al* (2012) are exponentials of some linear combination of attributes, and the 1-month PD and POE equal one minus the exponential of the product of intensity and the length of time (i.e., one month). In short, the above transformation takes PDs and POEs back to some linear combinations of attributes.

Table 2: Summary statistics of one-month PDs and POEs for different industries

	Sector	Summary statistics of 1-month PDs & POEs						
		mean	std	$q_{5\%}$	$q_{25\%}$	$q_{50\%}$	$q_{75\%}$	$q_{95\%}$
1-month PD	Financial	6.41	51.23	0.00	0.19	0.95	3.60	21.18
	Basic material	5.02	29.87	0.00	0.14	0.72	2.99	20.73
	Communications	5.62	50.80	0.00	0.87	0.52	2.35	17.29
	Consumer (cyclical)	5.03	33.10	0.00	0.15	0.74	2.87	18.65
	Consumer (noncyclical)	3.73	29.79	0.00	0.05	0.39	1.78	13.35
	Diversified	5.36	24.37	0.01	0.19	1.00	4.08	22.49
	Energy	4.85	37.99	0.00	0.06	0.43	2.22	15.78
	Industrial	4.48	27.90	0.01	0.16	0.76	2.81	17.01
	Technology	3.82	31.33	0.00	0.08	0.45	1.89	13.05
	Utilities	3.70	18.75	0.00	0.02	0.24	1.69	18.87
	All	4.85	36.13	0.00	0.11	0.65	2.65	17.60
1-month POE	Financial	43.69	30.64	12.10	23.45	37.92	56.12	92.84
	Basic material	41.56	42.93	11.43	19.77	30.77	50.99	101.19
	Communications	47.35	42.09	12.03	22.61	37.35	59.40	113.12
	Consumer (cyclical)	40.93	32.32	12.24	20.45	32.22	52.43	93.75
	Consumer (noncyclical)	45.35	38.56	12.05	21.71	36.35	58.00	105.55
	Diversified	38.35	33.74	10.97	20.39	31.50	47.60	84.97
	Energy	48.41	45.49	11.33	24.42	38.23	58.86	114.01
	Industrial	41.19	34.13	12.19	20.68	32.54	52.57	94.91
	Technology	48.72	43.30	11.96	23.24	39.63	62.09	112.06
	Utilities	34.34	21.16	10.05	19.89	30.26	43.67	72.31
	All	43.29	36.61	11.97	21.39	34.72	55.03	99.21

The data sample comprises 40,560 firms whose historical PDs and POEs are available for at least 60 months over the sample period (December 1990 to December 2014), and they are based on the results of the RMI-CRI January 2015 calibration. "std" stands for standard derivation, "q_α" represents the α-quantile. All values are in basis points and one basis point is one hundredth of one percentage.

Other transformations can also be applied; for example, marginal distributions obtained from the time series observations coupled with the inverse of standard normal distribution function can be applied to transform PDs and POEs to standard normal random variables. Such an approach will be akin to modeling default correlations through a Gaussian copula, but it would be very numerically intensive due to repeated use of this and its inverse transformations over time and for many obligors.

It is natural to expect a firm's likelihood of default being influenced by the global and industry-specific credit factors. Thus, we create such credit cycle factors by taking global and industry median transformed PDs. This approach is much like creating stock market indices. We opt not to use country median PDs because that would create too many common factors for the comprehensive global database used in this study. Because POEs are meant to reflect different aspect of corporate information, we create a pair of global common factors from both transformed PDs and POEs. Specifically, we define the PD and POE global factors as

$$F_t^{(G_P)} := \underset{1\le i\le n_t}{\text{median}}\,(P_{i,t}) \quad \text{and} \quad F_t^{(G_Q)} := \underset{1\le i\le n_t}{\text{median}}\,(Q_{i,t})\,.$$

For the industry factors, we use the Bloomberg 10-industry classification and define $K = 10$ pairs of industry factors.

$$F_{k,t}^{(I_P)} := \underset{1\le i\le n_t, i\in\text{Industry } k}{\text{median}}\,(P_{i,t}) \quad \text{and} \quad F_{k,t}^{(I_Q)} := \underset{1\le i\le n_t, i\in\text{Industry } k}{\text{median}}\,(Q_{i,t})\,, \quad k = 1,\cdots,K.$$

The global and industry factors as constructed above are certainly correlated. We thus apply the Gram-Schmidt-like process to orthonormalize these factors in pairs:

Step 1: (Orthogonalization) For $k = 1,\cdots,K$, regress the k-th industry factor pair, $F_{k,t}^{(I_P)}$ and $F_{k,t}^{(I_Q)}$, on the global factor pair and all preceding industry factor pairs.

Step 2: (Normalization) Normalize all factors to be of mean 0 and standard deviation 1.

In the sequel, we refer to the industry factors as those after the above orthogonormalization, and continue to use the same notations to represent them.

The global and industry factors, even though being the main drivers, are unlikely to fully capture the co-movement of credit risks across firms. We thus proceed to extract some unknown common factors to supplement information on co-movements. We first remove the effects of the global and industry factors from the transformed 1-month PDs and POEs, and then apply the principal component analysis (PCA) to produce other common factors. The specific procedure is as follows.

Step 1: For each firm i, regress $P_{i,t}$ on $F_t^{(G_P)}$ and $F_{k,t}^{(I_P)}$, $k = 1,\cdots,K$. Denote the regression residuals by a T-dimensional vector $Z_i^{(P)}$. Although regression residuals are stored in a full-length T-dimensional vector, some entries may be missing because a firm's 1-month PDs may not run the full sample length.

Step 2: For each firm i, regress $Q_{i,t}$ on $F_t^{(G_Q)}$ and $F_{k,t}^{(I_Q)}$, $k = 1, \cdots, K$. Denote the regression residuals by a T-dimensional vector $Z_i^{(Q)}$. Again, this residual vector may have missing entries.

Step 3: Stack all regression residuals together to form a matrix Z of size $T \times 2n$, i.e., $Z := (Z_1^P, \cdots, Z_n^P, Z_1^Q, \cdots, Z_n^Q)$. Extract the first r principal components of Z as the additional common factors (orthogonal to the global and industry factors), and denote them by $F_j^{(O)}$, $j = 1, \cdots, r$. (See Appendix A for technical details on the PCA method that handles missing data. In our implementation, r is set to five.)

To summarize, the factor model used for default correlations will be comprised of 27 factors: one PD global factor, one POE global factor, ten PD industry factors, ten POE industry factors, and five other latent common factors.

Each individual firm is expected to respond to common factors in a different manner, and the channels of influence can be identified by regressing each firm's transformed 1-month PDs and POEs. It is naturally to think that the PDs are only affected by PD-related factors, and likewise for POEs. Therefore, for firm i, the following pair of factor regressions is conducted and each regression contains 16 regressors ($K = 10$ and $r = 5$):

$$P_{i,t} = \beta_i^{(P)} + \beta_i^{(G_P)} F_t^{(G_P)} + \sum_{k=1}^{K} \beta_{k,i}^{(I_P)} F_t^{(I_P)} + \sum_{j=1}^{r} \beta_{j,i}^{(O_P)} F_j^{(O)} + \varepsilon_{i,t}^{(P)}, \quad t = 1, \cdots, T, \tag{1}$$

$$Q_{i,t} = \beta_i^{(Q)} + \beta_i^{(G_Q)} F_t^{(G_Q)} + \sum_{k=1}^{K} \beta_{k,i}^{(I_Q)} F_t^{(I_Q)} + \sum_{j=1}^{r} \beta_{j,i}^{(O_Q)} F_j^{(O)} + \varepsilon_{i,t}^{(Q)}, \quad t = 1, \cdots, T. \tag{2}$$

It is quite plausible that a firm is simultaneously influenced by several industries due to its multiple business lines. However, most firms are likely to be influenced by only a few industry factors. In our implementation, the above regressions are conducted stepwise with a cut-off p-value of 10%. When a factor is insignificant, we will fix the specific regression coefficient at zero to avoid the instability caused by sampling errors.

2.2 Factors dynamics

Short-term PDs and POEs are expected to have time dynamics. With the factor model, one could argue that the time dynamics comes through the common factors. Indeed, the predetermined common factors are autocorrelated. Due to the pairwise orthogonalization described earlier, it makes sense to also model them pairwise by a first-order two-dimensional vector autoregression (VAR) to caputure the time dynamics of the predetermined common factors. For the latent factors extracted by PCA, they are grouped together as a five-dimensional system ($r = 5$). Specifically, let

Table 3: VAR(1) Estimates for the global and financial industry factors

Factor pair	VAR(1) parameter estimates						
	A_{11}	A_{12}	A_{21}	A_{22}	Γ_{11}	Γ_{22}	Γ_{12}
Global	0.990	-0.016	0.010	0.974	0.034	0.054	0.010
	(0.013)	(0.012)	(0.016)	(0.015)	(0.003)	(0.005)	(0.003)
Financial	0.925	-0.016	0.090	0.883	0.163	0.172	0.009
	(0.030)	(0.028)	(0.031)	(0.029)	(0.016)	(0.017)	(0.011)

Each factor pair comprises the median transformed PD and POE for the category. The sample runs monthly from December 1990 to December 2014. Values inside the brackets are standard errors of parameter estimates.

$\mathbf{F}_t$ represent any of the vectors:

$$\begin{bmatrix} F_t^{(G_P)} \\ F_t^{(G_Q)} \end{bmatrix}, \qquad \begin{bmatrix} F_{k,t}^{(G_P)} \\ F_{k,t}^{(G_Q)} \end{bmatrix} (1 \le k \le K) \qquad \text{and} \quad \begin{bmatrix} F_{1,t}^{(O)} \\ \vdots \\ F_{r,t}^{(O)} \end{bmatrix}.$$

Note that all factors are of mean zero by construction. We can represent the VAR(1) model for any group of factors as

$$\mathbf{F}_t = \mathbf{A}\mathbf{F}_{t-1} + \mathbf{E}_t, \tag{3}$$

where $\mathbf{A}$ is a time-invariant square matrix and $\mathbf{E}_t$ is a vector of error terms.

Table 3 provides the estimation results for the global factor pair of median PD and POEs. Also presented are the results for the financial industry factor pair. Autocorrelation is clearly evident in Table 3, but cross-autocorrelation is less clear. The results for the other nine industries and five latent factors are not shown here to conserve space, but the results are qualitatively similar.

2.3 Calibration to the term structures of PDs

Up until this point, the factor model estimation has not utilized anything beyond historical time series of one-month PDs and POEs. When term structures of PDs are available, and which is the case with the RMI-CRI database, one can devise a way to further calibrate the factor model to term structures of PDs to take advantage of additional information.

The estimated factor model in equations (1) and (2) can be combined with the factors dynamics in equation (3) to simulate future paths of 1-month PDs and POEs for any group of obligors over any horizon of interest. With one set of simulated paths in place, one can compute, by the standard survial/default formula, individual default probabilities, conditional on the paths, for those obligors over various horizons. These conditional individual default probabilities can then be

averaged over many simulations to arrive at PDs for different horizons and obligors, which should in principle match up with their observed term structure of PDs. In reality, model mis-spefication and estimation errors will prevent two sets to match exactly unless the PD term structures are deduced from the estimated factor model. Mismatch in fact presents an opportunity for the user to further calibrate parts of the factor model.

We only allow the parameters of individual error terms of the factor model to be calibrated to match with their respective term structures of PDs. Choosing to focus on error terms is natural because (1) each firm has one pair of error terms (i.e., PD and POE) and one term structure of PDs, and (2) any change to the common factors dynamics would have global implications that inevitably complicate the calibration task. The pair of error terms is assumed to have a VAR(1) structure: for firm i,

$$\begin{bmatrix} \varepsilon_{i,t}^{(P)} \\ \varepsilon_{i,t}^{(Q)} \end{bmatrix} = \begin{bmatrix} \mu_i^{(P)} \\ \mu_i^{(Q)} \end{bmatrix} + \begin{bmatrix} \rho_i^{(P)} & 0 \\ 0 & \rho_i^{(Q)} \end{bmatrix} \begin{bmatrix} \varepsilon_{i,t-1}^{(P)} \\ \varepsilon_{i,t-1}^{(Q)} \end{bmatrix} + \begin{bmatrix} e_{i,t}^{(P)} \\ e_{i,t}^{(Q)} \end{bmatrix}, \qquad t \geq T+1, T+2, \cdots, \tag{4}$$

where $\left(e_{i,t}^{(P)}, e_{i,t}^{(Q)}\right)' \sim N(\mathbf{0}, \Sigma_i)$. Let θ_i represent all the parameters in equation (4), i.e., $\theta_i := \left(\mu_i^{(P)}, \mu_i^{(Q)}, \rho_i^{(P)}, \rho_i^{(Q)}, \Sigma_i\right)$. Thus, there are seven parameters to be calibrated at time T, the calibration (or application) time.

Denote by $\hat{p}_{i,T}(\theta_i; k)$ and $\hat{q}_{i,T}(\theta_i; k)$ the PD and POE for firm i over $(T, T+k]$ generated by the calibration model at parameter θ_i. With the previously estimated factors dynamics and factor loadings in place, the k-period PD at time T can be computed as the expected value of some function of future one-period PDs and POEs; that is,

$$\hat{p}_{i,T}(\theta_i; k) = p_{i,T}(1) + E_T\left\{\sum_{s=1}^{k-1} \hat{p}_{i,T+s}(\theta_i; 1) \prod_{\tau=0}^{s-1} \left[1 - \hat{p}_{i,T+\tau}(\theta_i; 1) - \hat{q}_{i,T+\tau}(\theta_i; 1)\right]\right\}, \tag{5}$$

where $E_T(\cdot)$ is expectation taken at time T for those random future one-month PDs and POEs.

Calibration is to search for the seven unknown parameters so as to match the model PDs with their observed counterparts over a set of selected default prediction horizons defined by C; that is, solve the following minimization problem firm by firm:

$$\min_{\theta_i} \sum_{k \in C} \left(\frac{\hat{p}_{i,T}(\theta_i; k)}{p_{i,T}(k)} - 1\right)^2. \tag{6}$$

In our implementation, the expectation in equation (5) is approximated by an average from a 1000-path Monte Carlo simulation. The chosen prediction horizons for calibration are $2, 3, 4, \cdots, 24$ months.

The summary statistics on the term structure of PDs used in calibration is given in Table 4. To conserve space, we only present those for a subset of selected default prediction horizons. It is clear from this table that PDs for different prediction horizons are all highly right-skewed.

Table 4: Summary statistics of PDs for selective horizons

Horizon	Summary statistics of PDs						
	mean	std	$q_{5\%}$	$q_{25\%}$	$q_{50\%}$	$q_{75\%}$	$q_{95\%}$
1-month	4.85	36.13	0.00	0.11	0.65	2.65	17.60
3-month	14.44	81.73	0.01	0.46	2.35	8.97	54.53
6-month	29.05	125.96	0.05	1.34	6.00	20.87	113.09
12-month	59.00	181.65	0.30	4.71	17.09	51.27	231.99
24-month	116.79	246.07	2.23	17.66	49.18	121.74	437.55

The data sample comprises 40,560 firms whose historical PDs are available for at least 60 months over the sample period (December 1990 to December 2014), and they are based on the results of the RMI-CRI January 2015 calibration. "std" stands for standard derivation, "q_α" represents the α-quantile. All values are in basis point and one basis point is one hundredth of one percentage.

3 Large-Portfolio Credit Analysis

Our credit portfolio aggregation methods rely on two facts: (1) default correlations arise from correlated future short-term PDs, and (2) conditional on future short-term PD paths, default events are actually independent. These two features enable us to come up with suitable algorithms for generating default-rate and portfolio-loss distributions (see Appendix B). In a nutshell, we combine the estimated factor model and factors dynamics in equations (1), (2) and (3) with the calibrated residual dynamics in equation (4) to simulate future paths of 1-month PDs and POEs for any target group of obligors to any future time point of interest. As a consequence, multiperiod default probabilities deduced from these random paths exhibit default correlations through co-movements induced by the common factors. Individual obligors' multiperiod default probabilities, exposures at default and recovery rates, conditional on the random paths, can then be aggregated to the portfolio level to compute quantities of interest. Averaging the conditional outcomes over many simulations then yields the final results. This bottom-up approach is naturally applicable to small or very large credit portfolios.

The algorithms presented in Appendix B provide the credit portfolio aggregation tools that handle default correlations, but they can also be used when there are no default correlations. The only difference is that one should directly apply marginal PDs of different obligors in those algorithms. In other words, simulation according to the factor model to generate default correlations is no longer needed. For computing portfolio-loss distribution, however, sampling defaulting obligors as decribed in Appendix B.2 still needs to performed, but it can be done separately to any degree of accuracy without being tied to the simulation of the factor model.

Three large credit portfolios are used to demonstrate our bottom-up approach to portfolio credit analysis. These portfolios are, respectively, all exchange-listed firms in the US, Eurozone-12 (Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands,

Portugal and Spain) and ASEAN-5 (Indonesia, Malaysia, the Philippines, Singapore and Thailand) satisfying the selection criterion of having at least 60 months of PDs and POEs. As explained earlier, our PDs and POEs are taken from the NUS-RMI database based on its January 2015 calibration.

The factor model in equations (1) and (2) has fairly good performance for these three credit portfolios. The average R^2 on the US firms up to December 2014 is 79.2% (57.3%) for PDs (POEs). In the case of Eurozone-12, the average R^2 is 80.1% for PDs and 83.7% for POEs. For firms in ASEAN-5, the average R^2 is 80.3% for PDs and 78.2% for POEs, respectively. Calibration to the term structure of PDs also delivers good performance. The average relative deviation of the calibrated versus observed PD across all horizons and firms is quite small, and they are 1.70%, 0.81% and 1.41% for the US, Eurozone-12 and ASEAN-5 in September 2008. In December 2014, the average relative deviations are 1.72%, 1.19% and 1.38% for these three groups.

Recovery rate and exposure are essential to the determination of loss distributions. In our analysis, all firms are assumed to share the same recovery rate distribution for simplicity. Our assumed recovery rate distribution is normal with mean 0.4 and standard deviation 0.2, and truncated to the meaningful support of $[0, 1]$. We also assume homogeneous obligors in terms of the exposure at default. It should be noted that the simulation-convolution algorithm for portfolio-loss distributions presented in Appendix B.2 can easily handle heterogenous exposures and/or recovery rate distributions. Our algorithms for default-rate and portfolio-loss distributions require of generating random paths of PDs and POEs, and we use 1,000 random paths in the implementation.

Figure 1 displays the comparison in September 2008 of the 12-month default-rate and portfolio-loss distributions with and without default correlations, and the three sets of two plots are for the three large credit portfolios considered in this analysis. Evidently, default correlations can significantly impact portfolio credit risk profiles (i.e., default-rate and portfolio-loss distributions) at the height of the 2008-09 global financial crisis. It is particularly so for the US, which is known to have suffered a severe blow during that financial crisis. When default correlations are suitably handled, both default-rate and portfolio-loss distributions are, as expected, right-skewed. Without default correlations, one should actually expect to see, due to central limit theorem, normally distributed default-rate and portfolio-loss distributions when the horizon is lengthened and/or the number of firms increases. With default correlations through common factors, however, converging to normality will never happen, because the random mixture effect caused by common factors will not dissipate.

Figures 2, 3 and 4 report the comparison between September 2008 and December 2014 for the US, Eurozone-12 and ASEAN-5, respectively. These figures reveal as expected that credit risk profiles (default-rate or portfolio-loss distributions) change with economic conditions. The three economies in December 2014 were in a far better shape as compared to September 2008. Not surprisingly, their credit risk profiles had improved significantly.

Figure 1: Comparison of default-rate and portfolio-loss distributions with and without correlations for three credit portfolios in September 2008

The default-rate and portfolio-loss distributions, with and without default correlations, as of September 2008 for a 12-month period are provided for all exchange-listed firms in the US, Eurozone-12 and ASEAN-5, respectively. Estimation of the factor model and factors dynamics is conducted with data up to September 2008. Calibration to the term structure of PDs is performed at month end of September 2008.

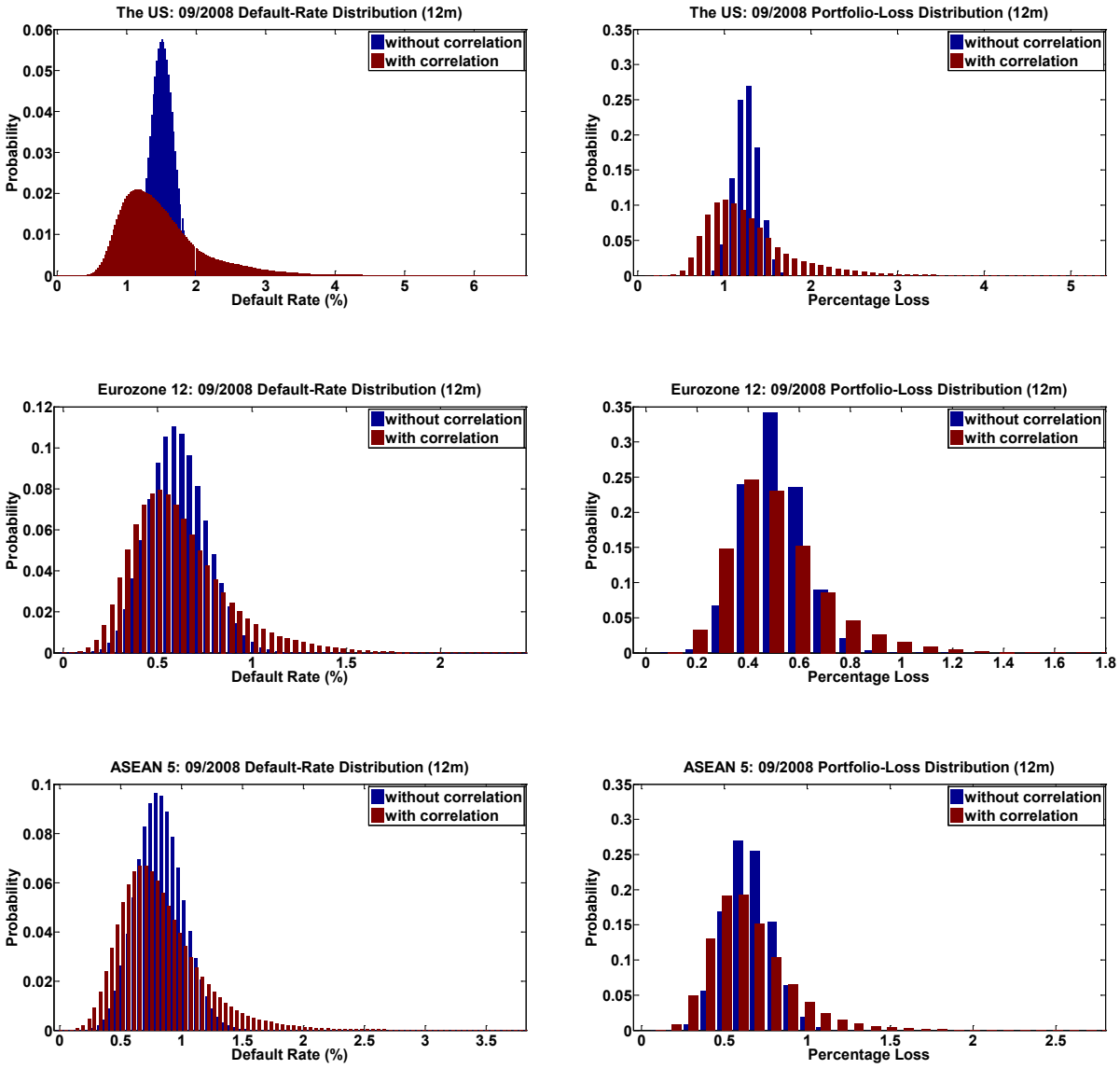


Figure 2: Comparison of default-rate and portfolio-loss distributions (with correlation) for the US portfolio between September 2008 and December 2014

The default-rate and portfolio-loss distributions with default correlations for all exchange-listed firms in the US as of September 2009 and December 2014. The distributions for three future periods (3 months, 12 months and 24 months) are presented. Estimation of the factor model and factors dynamics is conducted with data up to September 2008 and December 2014, respectively. Calibration to the term structure of PDs is performed at month end of September 2008 and December 2014, respectively.

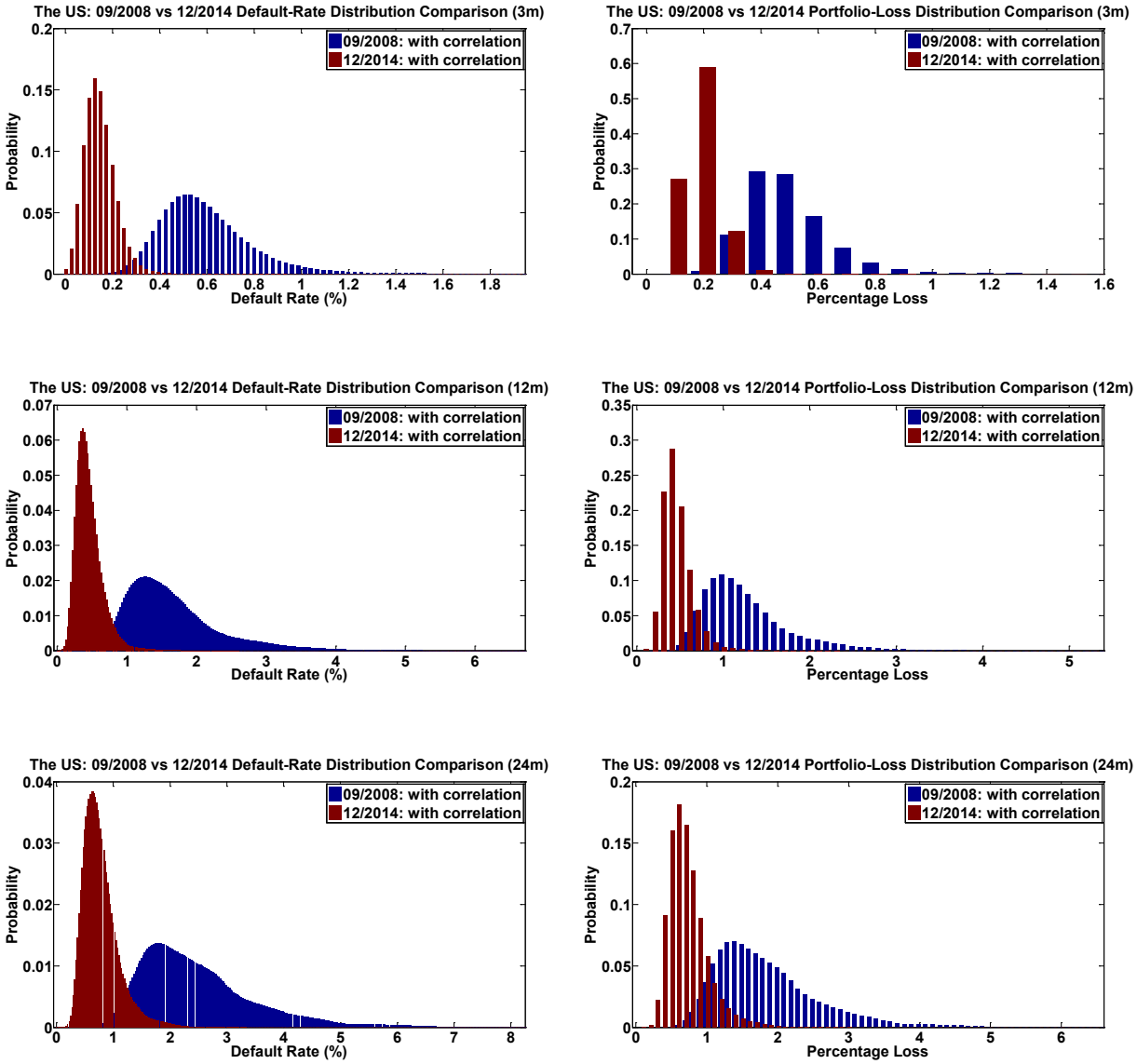


Figure 3: Comparison of default-rate and portfolio-loss distributions (with correlation) for the Eurozone-12 portfolio between September 2008 and December 2014

The default-rate and portfolio-loss distributions with default correlations for all exchange-listed firms in Eurozone-12 as of September 2009 and December 2014. The distributions for three future periods (3 months, 12 months and 24 months) are presented. Estimation of the factor model and factors dynamics is conducted with data up to September 2008 and December 2014, respectively. Calibration to the term structure of PDs is performed at month end of September 2008 and December 2014, respectively.

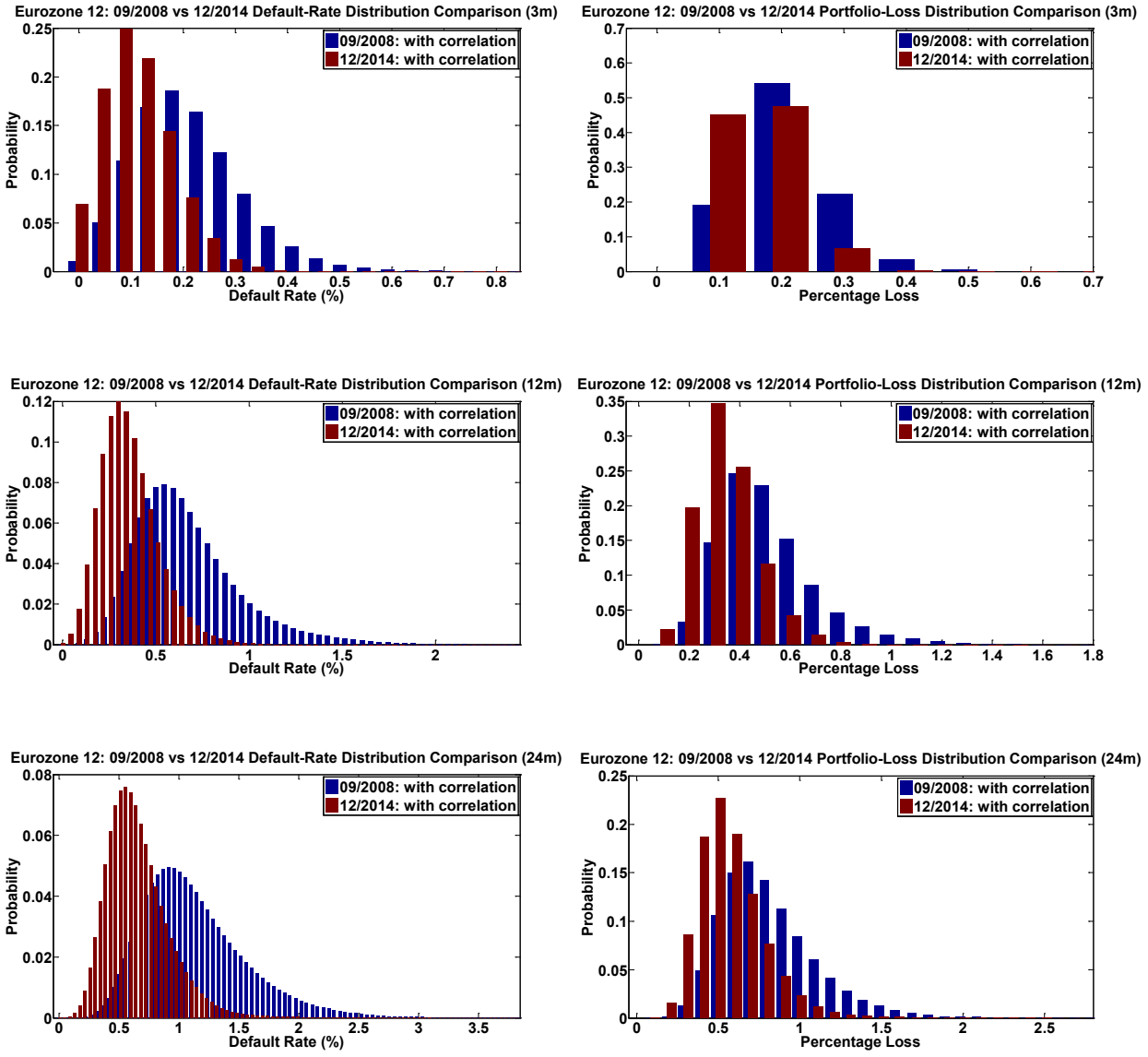
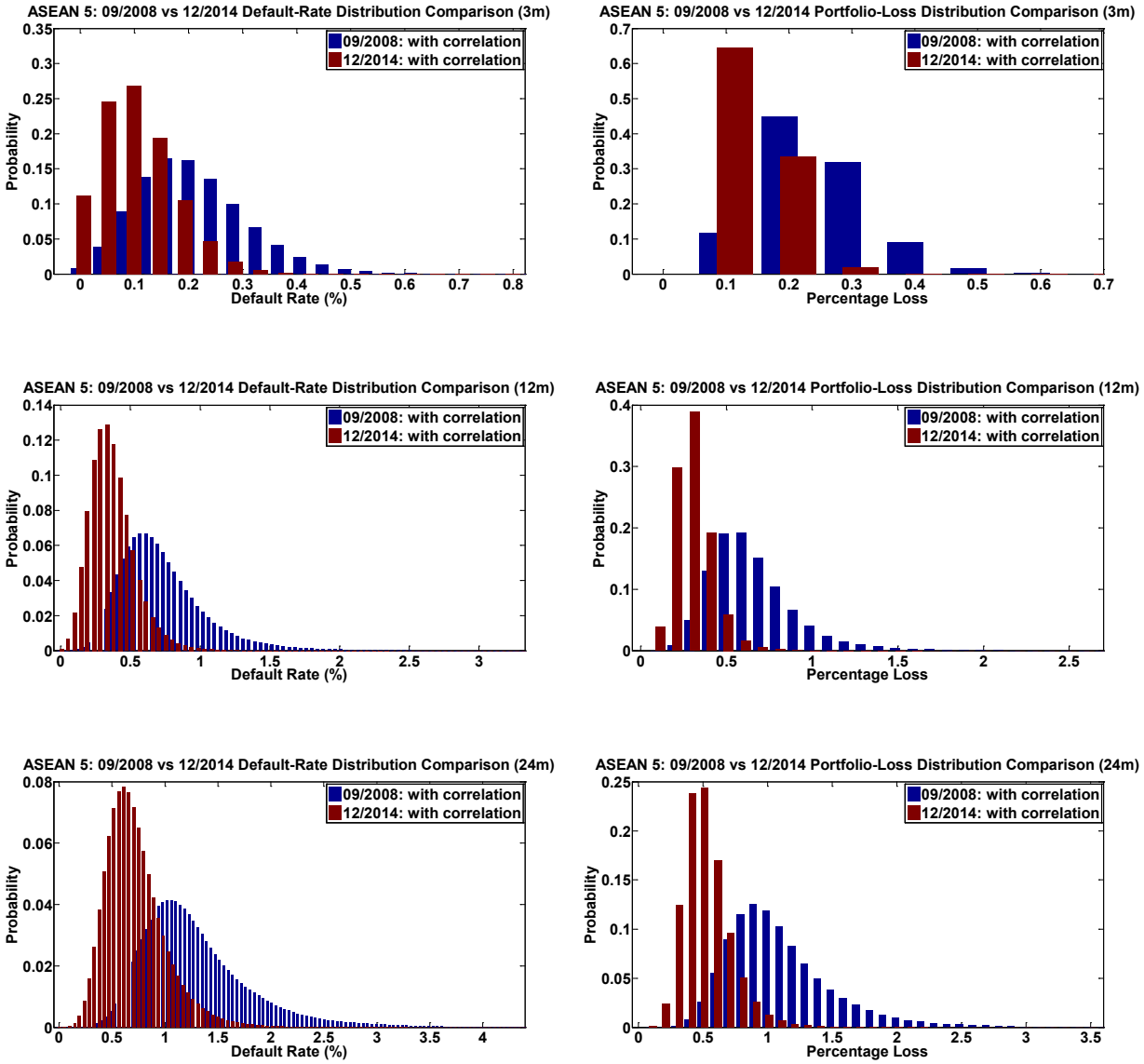


Figure 4: Comparison of default-rate and portfolio-loss distributions (with correlation) for the ASEAN-5 portfolio between September 2008 and December 2014

The default-rate and portfolio-loss distributions with default correlations for all exchange-listed firms in ASEAN-5 as of September 2009 and December 2014. The distributions for three future periods (3 months, 12 months and 24 months) are presented. Estimation of the factor model and factors dynamics is conducted with data up to September 2008 and December 2014, respectively. Calibration to the term structure of PDs is performed at month end of September 2008 and December 2014, respectively.



4 Conclusion

We propose a practical method for generating default correlations among obligors of a large credit portfolio. This method hinges on creating a sensible factor model for short-term PDs and POEs. The common factors are a combination of some intuitive predetermined factors by aggregating PDs and POEs and those latent ones determined by the principal component analysis. This factor model is meant to effectively handle high dimensionality inherent in default correlations for a large number of obligors, and serves to generate default correlations through correlated short-term PDs and POEs. Since missing data are an expected feature of default dataset, our factor extraction has been conducted with a suitable treatment of missing data. Portfolio credit analysis requires tools to aggregate individual defaults/losses to the portfolio level, and for this we utilize the convolution algorithm of Duan (2010) to compute default-rate distribution. In addition, we come up with a novel simulation-convolution algorithm for computing portfolio-loss distribution that factors in heterogenous individual credit exposures and recovery rates at default.

We apply the default correlation model and credit portfolio aggregation tools on a sample of 40,560 exchange-listed firms taken from the RMI-CRI database hosted by the National University of Singapore. Three large portfolios (the US, Eurozone-12 and ASEAN-5) are used to show the critical importance of default correlations. The results also reveal that portfolio credit risk profiles can change drastically with moving economic conditions, suggesting the importance of modeling credit risks with a dynamic system.

Future research can go in several directions; for example, refining the factor model and factors dynamics. For practical usage, however, the most urgent challenge is to come up with sensible ways to handle newly listed firms that have no data or a very short data history for estimating their factor loadings. Likewise, private firms present a similar challenge. Some private firms are significant players in credit markets. Being able to include them in portfolio credit analysis can be essential to many applications.

Appendices

A: Factor extraction with missing data

Principal component analysis (PCA) fits a high-dimensional data matrix by a low-dimensional representation, and for which the axes are defined as principal components. A complete data matrix Z of size $T \times n$ contains n obligors and T time series observations. Let $F_1, \cdots, F_r$ (T-dimensional vectors) denote its first r principal components. Let I_r denote the identity matrix of size r. Matrix $F := (F_1, \cdots, F_r)$ of size $T \times r$ (associated with some matrix B of size $r \times n$) is the solution to the following least squares problem:

$$\min_{F,B} \sum_{i,t} \left(Z_{i,t} - (FB)_{i,t}\right)^2 \quad \text{s.t. } BB^T = I_r. \tag{7}$$

The above minimization is related to another matrix factorization, known as the singular value decomposition (SVD). It is rather standard these days to compute the principal components using the (partial) SVD because it admits many efficient algorithms.

In real-life applications, Z is an incomplete data matrix with potential many entries missing, and the aforementioned method is not applicable. However, one can modify the least squares problem to

$$\min_{F,B} \sum_{(i,t)\in\Omega} \left(Z_{i,t} - (FB)_{i,t}\right)^2 \quad \text{s.t. } BB^T = I_r, \tag{8}$$

where Ω denotes the set of indices for non-missing entries. Stock and Watson (2002) proposed an EM algorithm-based iterative method to solve this minimization problem. Since the objective function of the least squares problem (8) is proportional to the log-likelihood under the normality assumption, a simple EM algorithm can be constructed for likelihood maximization (or least squares minimization), and it is summarized as follows:

Step 0: Start with $Z^{(0)}$ where $Z^{(0)}_{i,t} := Z_{i,t}$ for $(i,t) \in \Omega$ and $Z^{(0)}_{i,t} := 0$ for $(i,t) \notin \Omega$. (Note: Set missing values to 0 because the entries of Z are supposed to have their means equal to 0.) Set $k := 0$.

Step 1: Compute the optimal solution $\left(\widehat{F}^{(k)}, \widehat{B}^{(k)}\right)$ (via the partial SVD) to the minimization problem:

$$\min_{F,B} \sum_{i,t} \left(Z^{(k)}_{i,t} - (FB)_{i,t}\right)^2 \quad \text{s.t. } BB^T = I_r, \tag{9}$$

Step 2: Set $Z^{(k+1)}$ where $Z^{(k+1)}_{i,t} := Z_{i,t}$ for $(i,t) \in \Omega$ and $Z^{(k+1)}_{i,t} := \left(\widehat{F}^{(k)}\widehat{B}^{(k)}\right)_{i,t}$ for $(i,t) \notin \Omega$.

Step 3: If converged, stop; otherwise, set $k := k+1$ and go to **Step 1**.

The above algorithm is also known as the singular value projection (SVP) algorithm in the field of matrix completion; see Jain, *et al* (2010).

B: Algorithms for default-rate and portfolio-loss distributions

Consider a credit portfolio of n obligors. The algorithms are for computing the distributions for this portfolio's default rates and loss rates over a specified time period. For obligor i ($i = 1, \cdots, n$), let p_i, r_i and z_i denote its PD over the time period, recovery rate at default, and exposure amount, respectively. We assume random recovery rates with known distributions, and they may be different across different obligors. If recovery rates are constants, the algorithm for portfolio-loss distribution can be significantly simplified. The method for the default-rate distribution is taken from Duan (2010). The algorithm for the portfolio-loss distribution is new, which takes advantage of the default-rate distribution.

B.1: Default-rate distribution

Denote the default-rate distribution of this portfolio by $Q(k/n)$, $k = 0, \cdots, n$. Obviously, $Q(k/n) = P_n(k)$, where $P_i(k)$ represents the probability of k defaults in this portfolio after first i out of n obligors have been considered. Instead of tacking the default numbers that have negligible probabilities, we truncate $P_i(k)$ at some default number, k^*, beyond which the probability is smaller than a tiny threshold, say, $\tau = 10^{-6}$. The following convolution method follows that of Duan (2010), but is stated in an algorithm style. The convolution algorithm is as follows:

Step 0: Let $P_0(0) := 1$. Set $k^* := 0$ and $i := 1$.

Step 1: Forward the convolution to the next obligor by letting

$$P_i(k) := \begin{cases} P_{i-1}(0)(1-p_i) & \text{for } k = 0 \\ P_{i-1}(k)(1-p_i) + P_{i-1}(k-1)p_i & \text{for } k = 1, \cdots, k^* \\ P_{i-1}(k-1)p_i & \text{for } k = k^* + 1 \end{cases}$$

Step 2: Reset the truncated point $k^* := \max_{1 \le k \le k^*+1} \{P_i(k) \ge \tau\}$. If $i = n$, reset $P_n(.)$ vector by dividing the vector with its own sum to ensure no loss of probability, let

$$Q(k/n) = \begin{cases} P_n(k) & \text{for } 0 \le k \le k^* \\ 0 & \text{for } k^* < k \le n \end{cases}$$

and stop. Otherwise, reset $i := i + 1$, and go to **Step 1**.

B.2: Portfolio-loss distribution

With the default-rate distribution in place, we still need to deal with the significant complication associated with heterogeneity across obligors in terms of their individual exposures and recovery rates. We assume recovery rates are independent across obligors at the time of their defaults. Convolution can again be applied to figure out the loss distribution if the identity of the defaulters are known. Consider, for example, three defaults with the probability of such an occurrence taken from the default-rate distribution. But the three defaulters can be any three out of the total n obligors, which means many possible combinations of three obligors. Exhausting all combinations, knowing the likelihood of each combination, and computing the convoluted loss distribution for each combination can be a daunting task. Here, we propose a novel simulation-convolution method. For each number of defaults with its probability from the default-rate distribution, randomly sample a combination using the appropriate probability, and then compute the convoluted loss distribution for this combination of obligors. Repeat this simulation-convolution many times to obtain the portfolio-loss distribution for each given number of defaults.

Let $0 = c_0 < c_1 < \cdots < c_{m-1} < c_m = 1$ be an partition of the percentage loss domain $[0, 1]$ into m equally sized subintervals, and define $I_j := (c_{j-1}, c_j]$, $j = 1, \cdots, m$. Denote by $Q_n(I_j)$,

$j = 1, \cdots, m$, the discretized loss distribution of n obligors that is associate with the partition. For each obligor, we convert its individual loss distribution into one that is measured as a fraction of the total portfolio exposure and based on the partition; that is,

$$q_i(I_j) = \text{Prob}\left(\frac{z_i(1-r_i)}{\sum_{i=1}^n z_i} \in I_j \right), \quad i = 1, \cdots n, \ j = 1, \cdots, m.$$

Similar to the algorithm for the default-rate distribution, we set $\tau = 10^{-6}$ to truncate the loss distribution function. The algorithm is as follows:

Step 0: Obtain the distribution for the number of defaults, i.e., $P_n(k)$, $k = 0, \cdots, k^*$, as in the convolution algorithm for the default-rate distribution.

Step 1: For each $k = 1, \cdots, k^*$, simulate a set of indices of the k defaulting obligors, say $i_k^{(1)}, \cdots, i_k^{(k)}$, one at a time using an appropriate probability for each of the remaining obligors. Specifically, let A_l be the set of remaining obligors after sampling out the first l obligors. The probability for sampling $i_k^{(l+1)}$ is $\frac{p_i \prod_{\{j \in A_l, j \neq i\}} (1-p_j)}{\sum_{i \in A_l} p_i \prod_{\{j \in A_l, j \neq i\}} (1-p_j)}$ for $i \in A_l$.

Step 2: Use convolution to calculate the loss distribution conditional on this simulated obligors. The randomized loss distribution after completing l out of k obligors is denoted by $\tilde{Q}_k^{(l)}(I_j)$, $j = 1, \cdots, m$. Apply a step-specific threshold, $\tau_k := \frac{\tau}{P_n(k)}$, to truncate the loss distribution. (Note: τ_k is used to make sure that each random copy of $\tilde{Q}_k^{(k)}(\cdot)$ after multiplying by $P_n(k)$ later is effectively truncated by τ. This treatment is used to speed up the algorithm without losing accuracy.)

Step 2.0: Let $\tilde{Q}_k^{(0)}(I_0) := 1$. Set $s_1 := 0$, $s_2 := 0$ and $l := 1$.

Step 2.1: Set $t_1 := \min_{1 \le j \le m} \{q_{i_k^{(l)}}(I_j) \ge \tau_k\}$ and $t_2 := \max_{1 \le j \le m} \{q_{i_k^{(l)}}(I_j) \ge \tau_k\}$. Set $j_1 := s_1 + t_1$ and $j_2 := s_2 + t_2$. For $j_1 \le j \le j_2$, let

$$\tilde{Q}_k^{(l)}(I_j) := \tilde{Q}_k^{(l-1)}\big(I_{\max\{s_1, j-t_2\}}\big) q_{i_k^{(l)}}\big(I_{\min\{t_2, j-s_1\}}\big) + \cdots + \tilde{Q}_k^{(l-1)}\big(I_{\min\{s_2, j-t_1\}}\big) q_{i_k^{(l)}}\big(I_{\max\{t_1, j-s_2\}}\big).$$

Step 2.2: If $l = k$, let

$$\tilde{Q}_k^{(k)}(I_j) = \begin{cases} \tilde{Q}_k^{(k)}(I_j) & \text{for } j_1 \le j \le j_2 \\ 0 & \text{for } 1 \le j < j_1 \text{ and } j_2 < j \le m \end{cases}$$

reset $\tilde{Q}_k^{(k)}(.)$ vector by dividing the vector with its own sum to ensure no loss of probability, and stop. Otherwise, set $s_1 := \min_{j_1 \le j \le j_2} \{\tilde{Q}_k^{(l)}(I_j) \ge \tau_k\}$, $s_2 := \max_{j_1 \le j \le j_2} \{\tilde{Q}_k^{(l)}(I_j) \ge \tau_k\}$ and $l := l + 1$, and go to **Step 2.1**.

Step 3: Compute the randomized version of the portfolio-loss distribution by $\tilde{Q}_n(I_j) := \sum_{k=1}^{k^*} P_n(k) \tilde{Q}_k^{(k)}(I_j)$ for $j = 1, \cdots, m$.

Repeat the above steps to obtain many random copies of $\tilde{Q}_n(\cdot)$, and then average to obtain a Monte Carlo estimate of $Q_n(\cdot)$.

The percentage loss partition needs to be fine enough to achieve good approximation accuracy, and we set $m = 10,000$ in our implementation. In terms of the number of random copies of $\tilde{Q}_n(\cdot)$, we sync it with the number of future one-month PD simulations. When, say, 1,000 future PD paths are simulated, we will just generate one random copy corresponding to each PD simulation to obtain 1,000 random copies.

References

[1] Albanese, C., Li, D., Lobachevskiy, E., and Meissner, G. (2013), "A comparative analysis of correlation approaches in finance," *Journal of Derivatives* 21(2), 42-66.

[2] Andersen, L., and Sidenius, J. (2005), "Extensions to the Gaussian copula: Random recovery and random factor loadings," *Journal of Credit Risk* 1(1), 29-70.

[3] Azizpour, S., Giesecke, K. and Kim, B. (2011), "Premia for correlated default risk," *Journal of Economic Dynamics and Control* 35(8), 1340-1357.

[4] Crane, G. and van der Hoek, J. (2008), "Using distortions of copulas to price synthetic CDOs," *Insurance: Mathematics and Economics* 42(3), 903-908.

[5] Driessen, J. (2005), "Is default event risk priced in corporate bonds?" *Review of Financial Studies* 18(1), 165-195.

[6] Duan, J.-C. (2010), "Clustered defaults," Working paper, Risk Management Institute, National University of Singapore.

[7] Duan, J.-C., Sun, J., and Wang, T. (2012), "Multiperiod corporate default predictionA forward intensity approach," *Journal of Econometrics* 170(1), 191-209.

[8] Duffee, G. R. (1999), "Estimating the price of default risk," *Review of Financial Studies* 12(1), 197-226.

[9] Duffie, D., Eckner, A., Horel, G., and Saita, L. (2009), "Frailty correlated default," *Journal of Finance* 64(5), 2089-2123.

[10] Duffie, D., Saita, L. and Wang, K. (2007), "Multi-period corporate default prediction with stochastic covariates," *Journal of Financial Economics* 83, 635–665.

[11] Finger, C. (1999), "Conditional approaches for CreditMetrics portfolio distributions," *CreditMetrics Monitor* 1, 14-33.

[12] Frey, R. and McNeil, A.J. (2001), "Modelling dependent defaults," ETH, Eidgenössische Technische Hochschule Zürich, Department of Mathematics.

[13] Giesecke, K., Goldberg, L. R., and Ding, X. (2011), "A top-down approach to multiname credit," *Operations Research* 59(2), 283-300.

[14] Giesecke, K., and Weber, S. (2004), "Cyclical correlations, credit contagion, and portfolio losses," *Journal of Banking and Finance* 28(12), 3009-3036.

[15] Hull, J. C., and White, A. D. (2004), "Valuation of a CDO and an n-th to default CDS without Monte Carlo simulation," *Journal of Derivatives* 12(2), 8-23.

[16] Hurd, T., and Kuznetsov, A. (2007), "Affine Markov chain models of multifirm credit migration", *Journal of Credit Risk* 3(1), 3-29.

[17] Jain, P., Raghu, M., and Inderjit S. D. (2010), "Guaranteed rank minimization via singular value projection," *Advances in Neural Information Processing Systems*, 937-945.

[18] Li, D. X. (2000), "On default correlation: A copula function approach," *Journal of Fixed Income* 9(4), 43-54.

[19] RMI-CRI Technical Report (Version 2014 Update 1), Risk Management Institute, National University of Singapore.

[20] Schönbucher, P. J. (2001), "Factor models: Portfolio credit risks when defaults are correlated," *Journal of Risk Finance* 3(1), 45-56.

[21] Schönbucher, P. J. (2005), "Portfolio losses and the term structure of loss transition rates: a new methodology for the pricing of portfolio credit derivatives," Working paper.

[22] Schönbucher, P. J., and Schubert, D. (2001), "Copula-dependent default risk in intensity models," Working paper, Department of Statistics, Bonn University.

[23] Stock, J. H. and Watson, M. W. (2002), "Macroeconomic forecasting using diffusion indexes," *Journal of Business and Economic Statistics* 20(2), 147-162.

[24] Vasicek, O. (1987), "Probability of loss on loan portfolio," *KMV Corporation* 12(6).

[25] Yu, F. (2007), "Correlated default in intensity -based models", *Mathematical Finance* 17(2), 155-173.