OTTO-VON-GUERICKE-UNIVERSITY MAGDEBURG
FACULTY OF ECONOMICS AND MANAGEMENT

Simulation and Estimation of Loss Given Default

Stefan Hlawatsch • Sebastian Ostrowski

FEMM Working Paper No. 10, March 2010

F E M M

Faculty of Economics and Management Magdeburg

Working Paper Series

Otto-von-Guericke-University Magdeburg
Faculty of Economics and Management
P.O. Box 4120
39016 Magdeburg, Germany
http://www.ww.uni-magdeburg.de/

Simulation and Estimation of Loss Given Default*

Stefan Hlawatsch and Sebastian Ostrowski**

March 2010

Abstract

The aim of our paper is the development of an adequate estimation model for the loss given default, which incorporates the empirically observed bimodality and bounded nature of the distribution. Therefore we introduce an adjusted Expectation Maximization algorithm to estimate the parameters of a univariate mixture distribution, consisting of two beta distributions. Subsequently these estimations are compared with the Maximum Likelihood estimators to test the efficiency and accuracy of both algorithms. Furthermore we analyze our derived estimation model with estimation models proposed in the literature on a synthesized loan portfolio. The simulated loan portfolio consists of possibly loss-influencing parameters that are merged with loss given default observations via a quasi-random approach. Our results show that our proposed model exhibits more accurate loss given default estimators than the benchmark models for different simulated data sets comprising obligor-specific parameters with either high predictive power or low predictive power for the loss given default.

Keywords Bimodality · EM Algorithm · Loss Given Default · Maximum Likelihood · Mixture Distribution · Portfolio Simulation

J.E.L. Classification C01 · C13 · C15 · C16 · C5

*The authors are very grateful to Waltraud Kahle and Peter Reichling for their comments and suggestions.

**Both authors are from Otto-von-Guericke-University Magdeburg, Faculty of Economics and Management, Department of Banking and Finance, Postfach 4120, 39016 Magdeburg, Germany, phone +49 391 67-12256, e-mail Stefan.Hlawatsch@ovgu.de

1 Introduction

The general idea of our paper is the development of an estimation model for the loss given default (LGD). The LGD "means the ratio of the loss on an exposure due to the default of a counterparty to the amount outstanding at default."[1] This ratio is required to be estimated by banks following the Internal Ratings-Based Approach for their retail portfolios. Here the estimated LGDs are important for the capital requirement of the institute due to incurred credit risk. Furthermore the pricing process of credits is also highly influenced by an accurate estimation of possible losses.

Empirical analyses have shown that the distribution of the LGD often exhibits a bimodal shape.[2] This characteristic is intuitively reasonable when considering two possible developments of a defaulted loan. First the obligor may recover and continues her contractual repayment covenants. This results in a very small loss amount basically driven by administrative costs. Secondly the obligor may not recover which results generally in a higher loss amount. The mentioned bimodality in the density of the LGD distribution leads to difficulties in the development of possible estimation models. Established approaches do not consider this bimodal shape and we could not find any estimation procedure that overcomes this difficulty. This scientific gap shall be closed by our paper where we do not only account for the bounded character of the LGD distribution but also for its bimodality which enhances the estimation of future LGDs.

The approach we propose is to approximate the LGD distribution by a mixture of two beta distributions and furthermore regress the estimated transformed distribution values of the LGD on certain obligor-specific parameters. We obtain the needed parameters of the mixture by either applying a Maximum Likelihood estimation or an adjusted version of the Expectation-Maximization algorithm. The whole analysis and comparison with certain benchmark models is based on synthesized loan portfolios simulated by Monte Carlo techniques. We therefore generated two different data sets of loan portfolios with different explanatory power of the obligor-specific parameters within each portfolio. The results show that our approach could outperform the benchmark models in both accuracy and lower deviation of the estimated values from the realized ones for both data sets simulated. So our model provides potential for enhancements in credit pricing models.

The further proceeding in our paper is as follows: After a literature review of credit risk simulations and LGD estimation models in Section 2 we proceed with the description of the data set simulations in Section 3. Section 4 describes our estimation model in detail where the results of the carried out analyses are presented in Section 5. Section 6 concludes the paper.

[1] See Article 4 (27) Directive 2006/48/EC.

[2] See for example the mentioned literature in Appendix 1.

2 Literature Review

This section is divided into two parts, one regarding simulation approaches for portfolio credit risks and the second part recapitulates current literature concerning LGD estimations.

2.1 Literature Review of Credit Risk Simulation

Up to our knowledge, the recent literature about simulations of credit risk is driven by the estimation of either rare loss events or risk measures, as e.g. expected shortfall, value-at-risk or expected loss. However, the main focus is put on the sensitivity of the risk measures for changes in the default probability or dependencies between different risk factors and no studies could be found that address the development of estimation models based on simulated credit portfolios.

Gordy (1998) compares J.P. Morgan's CreditMetrics and Credit Suisse Financial Product's CreditRisk$^+$ via the underlying mathematical structure and highlights the differences between both models in functional form, distributional assumptions and estimation precision. Both models assume a constant LGD, focus on the default probability and the correlation between the defaults. The simulation incorporates three aspects of a credit portfolio: credit quality, total number of obligors in the portfolio and the distribution of dollar outstandings across the obligors within one rating grade. As a result he finds that both models perform similar on average quality portfolios and both models demand higher capital on lower quality portfolios. Furthermore an independence on the distribution of loan sizes in the portfolio could be found and both models are highly sensitive to the average default correlations in the portfolio.

A simulation study on the comparison of expected shortfall and value-at-risk for the allocation of capital in credit portfolios is carried out by Kalkbrener et al. (2004). They employ a multi-factor model for specifying the default correlations and an importance sampling algorithm is introduced for Merton-type models to efficiently compute the value-at-risk and expected shortfall of credit portfolios. The LGD is assumed to be constant at 100 percent for each obligor. Results of the study are the insight that expected shortfall is a better risk measure than the value-at-risk and by applying the importance sampling algorithm the variance and computation time of the simulation are strongly reduced.

Another simulation study in the context of credit risk is taken out by Jobst and Zenios (2005). They simulate risk-free interest rates and credit spreads to determine prices of credit risk sensitive securities. Here the LGD is assumed to be either uniformly distributed or constant at a level of 100 percent. Thus the credit spread almost solely depends on the

rating and therefore on the default probability of the security. This approach results in a consistent pricing approach for these kind of securities.

Kang and Shahabuddin (2005) simulate the dependency between systematic and unsystematic risk factors via a t-copula and transform this copula into a probability of default for each obligor. The LGD is modelled by a linear function and results in a uniformly distribution among the obligors. They also present an important sampling algorithm for the estimation of multi-factor portfolio credit risk for the t-copula model. A similar approach is taken by Bassamboo et al. (2008), who derive sharp asymptotics for portfolio credit risk that illustrate the implications of extremal dependence among obligors.

Glasserman et al. (2008) employ an important sampling technique within a Gaussian copula model for the estimation of portfolio credit risk as a rare event study. The LGD is assumed to be bounded to the top and can be stochastic. However, the focus of the paper is put on the important sampling algorithms and not on a certain LGD-modelling approach.

2.2 Literature Review of LGD Estimation Models

The second part of the literature review describes different LGD estimation models as for example implied market models and workout LGD models, where some of them are used as benchmarks in the simulation part of this paper.

One of the most famous models is the LossCalc model introduced by Moody's KMV.[3] The general idea for estimating the recovery rate is to apply a multivariate linear regression model including certain risk factors, e.g., industry factors and macroeconomic factors, and include transformed risk factors resulting from "mini-models". These mini-models try to take the dependencies between some influencing factors into consideration, i.e. two or more LGD-influencing factors that offer a certain relationship are transformed into one variable which is afterwards included in the overall regression model. The first step preceding the regression is the transformation of recovery rates into an approximately normally distributed random variable to remove the bounded characteristic of this variable. This transformation will be reversed after the regression and yields again a bounded estimator for the recovery rate.

Another estimation model, proposed by Glößner et al. (2006), consists of two steps, namely a scoring and a calibration step. The scoring step includes the estimation of a score using collateralizations, haircuts, expected exposure at default of the loan and recovery rates of the uncollateralized exposure. The score itself can be interpreted as a recovery rate of the total loan but is only used for relative ordering in this case. For computing the

[3] See Gupton and Stein (2005).

distribution of the loss rate depending on the score, this loss rate is approximated by the aggregated exposure of the portfolio up to the score value and the average loss rate of the portfolio.

An estimation model based on a linear regression in connection with a logit transformation of the LGD and a time consideration by using lag-variables within the regression is presented in Hamerle et al. (2006). The dependent variable in the regression is the transformed LGD and the independent variables are credit specific and macroeconomic variables representing potential systematic sources of risk. These systematic risk-variables are considered with a time-lag. Furthermore two unsystematic factors are included to account for credit-specific risk and for correlation between credits which is not considered by the macroeconomic factors.

Peter (2006) and Appasamy et al. (2008) propose a kind of multi-step approach. Both define possible scenarios after default of a credit, e.g. cure, restructuring or liquidation, and compute a scenario probability-weighted LGD. Either average scenario LGDs or credit specific scenario LGDs can be applied. Possible ways of computing scenario probabilities might be a logistic regression model or applying a Markov chain.

The assumption of a beta distributed LGD is taken up by Huang and Oosterlee (2008), where they apply different kinds of beta regressions for modeling the LGD. In short, the beta regressions are used to estimate the mean of the underlying distribution depending on certain LGD-influencing factors. Furthermore, depending on the model, other covariates are estimated by linear regression or Maximum Likelihood estimations. It is showed that the beta regressions provide a good way in modeling LGDs and finally, the portfolio loss distribution could efficiently be approximated.

Bastos (2009) presents two models for the LGD estimation. The first approach is called fractional response regression and is based on either a log functional relationship or a log-log functional relationship between the LGD and the independent variables to ensure the bounded nature of the dependent variable. The second approach is a so-called regression tree, which is a kind of clustering of the LGDs in homogeneous subgroups. This clustering is grounded on explanatory variables in the way that the group-inherent variance of the LGD is as small as possible.

The approaches proposed in Hamerle et al. (2006), Gupton and Stein (2005), and Bastos (2009) serve as benchmark models for comparison with our approach derived in Section 4. The exact implementation of the mentioned models is described in Section 4.4.

3 Simulation Model

3.1 Simulation of LGD

This first part introduces the LGD simulation and tries to reason why a mixture distribution framework is a more appropriate setting when referring to empirical LGDs. As shown in the literature review, a bimodal distribution is mostly assumed. This assumption can be justified by the following idea.

In the first case the obligor fails to meet the payment commitment for a specific date and, therefore, is defaulted according to the Basel II capital requirements.[4] Nevertheless in most cases the outstanding payment is acquitted later and so the obligor has recovered. Thus the loss of the debtor is limited to internal costs, e.g. administration and personnel costs. In the second case the obligor defaults at all and so the loss equals the amount outstanding less proceeds of collateral recoveries.

Bimodal distributions are constructed by mixing canonical distributions under some certain assumptions. Since the LGD is in general a value between zero and one, it is reasonable to use distributions with a bounded support, e.g. we use beta distributions on the interval [0,1]. The density of a random variable X following a general beta distribution over the interval [a,b] reads as:

$$\begin{aligned} f(x) &= \frac{1}{B(a,b,p,q)} \cdot (x-a)^{p-1} \cdot (b-x)^{q-1} \\ &\text{where} \\ B(a,b,p,q) &= \frac{\Gamma(p) \cdot \Gamma(q)}{\Gamma(p+q)} \cdot (b-a)^{p+q-1}, \end{aligned} \tag{1}$$

and $\Gamma(.)$ denotes the Gamma function.

The generation of the density function of a random variable following a multi-mode distribution resulting from a mixture of m random variables Υ_i with given density $f_j(\upsilon_i), j = 1, \ldots, m$ is just the convex combination of the densities of Υ_i. The same holds for the distribution function:

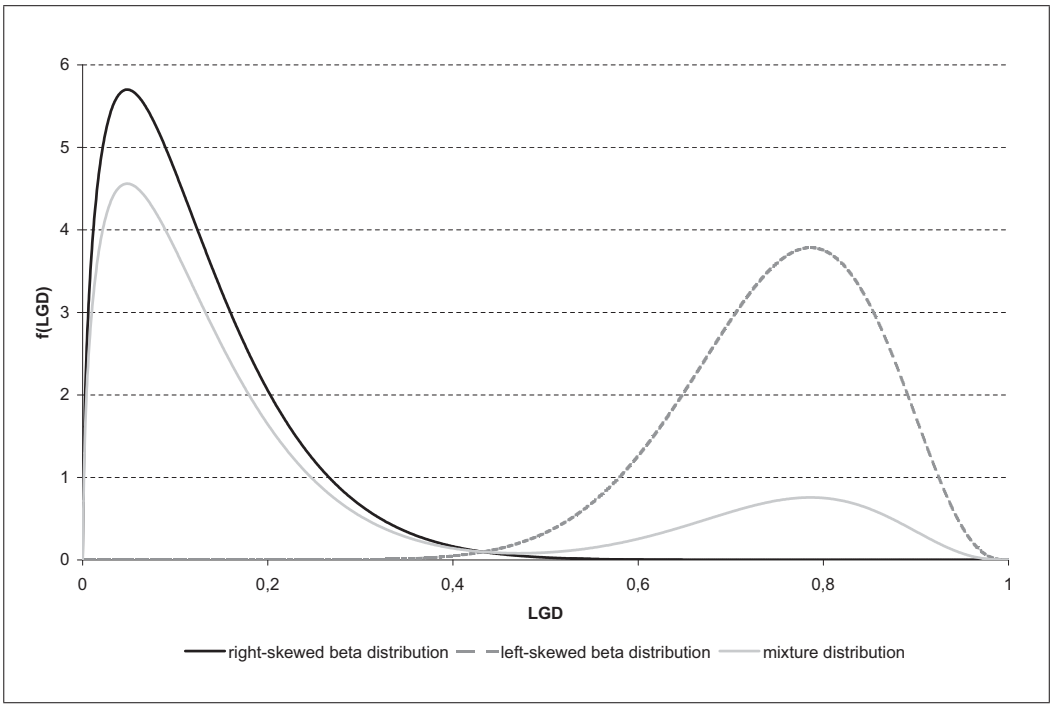
$$\begin{aligned} f(\upsilon_i) &= \sum_{j=1}^{m} \omega_j \cdot f_j(\upsilon_i) \\ F(\upsilon_i) &= \sum_{j=1}^{m} \omega_j \cdot F_j(\upsilon_i) \\ &\text{s.t.} \\ &\sum_{j=1}^{m} \omega_j = 1,\ \omega_j \geq 0\ \forall\ j = 1, \ldots, m \end{aligned} \tag{2}$$

[4] Possible default reasons are mentioned in Directive 2006/48/EC Annex VII Part 4 Nr. 44.

In our case Υ_i is assumed to be beta distributed and $m = 2$. Thus Figure 1 shows a possible density function of a mixed distribution resulting from the densities of two beta distributed random variables Υ_1 and Υ_2 over the interval [0,1].

Figure 1: Mixture density of two beta distributions

The figure shows a possible density function of a mixed distribution resulting from the densities of two beta distributed random variables over the interval [0,1] with $\omega_1 = 0.8$, $p_1 = 1.5$, $q_1 = 11$, $p_2 = 12$ and $q_2 = 4$.



The simulation of the LGD is the first step for the simulation of a loan portfolio. Generally this part of our approach can be divided into four steps. At first two beta distributions over the interval[5] [0,1] are chosen and, subsequently, the mixture distribution is constructed. Henceforth n random realizations are drawn from the mixture distribution and finally the whole procedure is done k-times. As a result we get k loan portfolios with each containing n obligors characterized by the corresponding LGD.

Beta distributions are defined by the two parameters, here denoted as p and q, which are no location and scale parameter itself but in combination. The expected value, the mode, denoted as $\gamma(.)$, and the variance of a beta distributed random variable Υ over [0,1] are

[5] We chose the interval [0,1] for computational purposes since every LGD distribution can be transformed in a way that [0,1] is sufficient.

calculated by:

$$\mathrm{E}(\Upsilon) = \frac{p}{p+q} \tag{3}$$

$$\gamma(\Upsilon) = \frac{p-1}{p+q-2} \tag{4}$$

$$\mathrm{Var}(\Upsilon) = \frac{p \cdot q}{(p+q+1) \cdot (p+q)^2} \tag{5}$$

Since we want to construct a bimodal distribution from the two beta distributed random variables Υ_1 and Υ_2, where the density of Υ_1 is assumed to be right-skewed and the density of Υ_2 is assumed to be left-skewed, with corresponding parameters p_1, q_1 and p_2, q_2, resp., the parameters should fulfill certain assumptions.

First each parameter p and q should be greater than one since otherwise the density of these distributions do not exhibit any mode. Furthermore we assume that the expected values $\mathrm{E}(\Upsilon_1)$ and $\mathrm{E}(\Upsilon_2)$ are contained in different intervals to ensure a bimodal density of the mixture distribution and we also assume that $\mathrm{Var}(\Upsilon_1)$ and $\mathrm{Var}(\Upsilon_2)$ are contained in different intervals.[6] $\mathrm{E}(\Upsilon_1)$ should be contained in the interval [0.059,0.3] and $\mathrm{E}(\Upsilon_2)$ should lay in [0.7,0.941] where the upper bound of the first interval and the lower bound of the second one are empirically justified.[7] $\mathrm{Var}(\Upsilon_1)$ is contained in the interval $\left[0.003, \frac{(1-\mathrm{E}(\Upsilon_1))\cdot(\mathrm{E}(\Upsilon_1))^2}{1+\mathrm{E}(\Upsilon_1)}\right)$ and $\mathrm{Var}(\Upsilon_2)$ is contained in the interval $\left[0.003, \frac{\mathrm{E}(\Upsilon_2)\cdot(1-\mathrm{E}(\Upsilon_2))^2}{2-\mathrm{E}(\Upsilon_2)}\right)$. For each portfolio simulated we first need to generate one random mixture distribution by randomly choosing two beta distributions. Therefore, two parameters for each distribution have to be selected by chance. The parameters $\mathrm{E}(\Upsilon_1)$, $\mathrm{E}(\Upsilon_2)$ are at first drawn from the described intervals and the parameters $\mathrm{Var}(\Upsilon_1)$ and $\mathrm{Var}(\Upsilon_2)$ are drawn afterwards. These values are used subsequently for the computation of the parameters p and q of each distribution. Assuming p and q greater than one, the upper bound of the variance interval depends on a given expected value. The upper bound of $\mathrm{Var}(\Upsilon_1)$ is derived here, whereas the upper bound of $\mathrm{Var}(\Upsilon_2)$ is derived in Appendix 2.

Since Υ_1 is assumed to be right-skewed, it follows that q is greater than p. We therefore rearrange Equation (3) for q and substitute the value in Equation (5).

$$\begin{aligned}\mathrm{Var}(\Upsilon_1) &= \frac{\left(\frac{p}{\mathrm{E}(\Upsilon_1)} - p\right) \cdot p}{\left(\frac{p}{\mathrm{E}(\Upsilon_1)} - p + p + 1\right) \cdot \left(\frac{p}{\mathrm{E}(\Upsilon_1)} - p + p\right)^2} \\ &= \frac{(1-\mathrm{E}(\Upsilon_1)) \cdot (\mathrm{E}(\Upsilon_1))^2}{p + \mathrm{E}(\Upsilon_1)}\end{aligned} \tag{6}$$

[6] A more reasonable parameter for a bimodal distribution might be the modes but due to computational simplifications the expected values are chosen instead.

[7] In Appendix 1 an overview of different studies about LGD distributions with respect to different time spans, industrial branches and financial contracts is given. For our intervals the maximum and minimum values of these studies are taken.

Furthermore, the upper bound for the variance is given if p equals its minimal possible value, namely one, and so we get:

$$\mathrm{Var}(\Upsilon_1) < \frac{(1 - \mathrm{E}(\Upsilon_1)) \cdot (\mathrm{E}(\Upsilon_1))^2}{1 + \mathrm{E}(\Upsilon_1)} \tag{7}$$

The lower bounds of the variance intervals are determined simultaneously with the lower bound of the expected value interval in case of Υ_1 and the upper bound of the expected value interval in case of Υ_2.

The idea behind the determination of the lower bound of the variance interval is that we want to avoid extremely peaked densities since these densities would result in a very narrow interval for the realizations, namely the LGD in our case. However, such an LGD distribution is not observed in reality and so it seems not suitable for our simulation. Given an expected value, the ratio between the parameters p and q remains constant.[8] In Appendix 4 it is shown that the variance of a beta distributed random variable decreases for increasing parameters p and q given a constant ratio between these parameters. This implies that, given a certain expected value, the variance is higher for lower parameter values p and q. However a lower bound for these parameters is given by one. Hence a minimal value for the variance interval corresponds to a minimal expected value. In Table 1 certain combinations of a minimal variance given an expected value are provided. In the whole explanation above, a right-skewed distribution was assumed. For the left-skewed case the parameters p and q change their roles and we assume an upper bound for the expected value, but the derivation remains the same.

Table 1: Combinations of variances and expected values
The table shows different expected values and corresponding maximum variances (used as the lower bound of the variance intervals).

variance	expected value (right-skewed)	expected value (left-skewed)
0.001	0.033	0.967
0.002	0.047	0.953
0.003	0.059	0.941
0.004	0.068	0.932
0.005	0.076	0.924

For our purpose a minimum variance of 0.003 is chosen for the interval. A lower variance would result in very peaked densities and a higher variance implies that the minimal expected value for the right-skewed density is above six percent, which is not very meaningful for recovered loans.

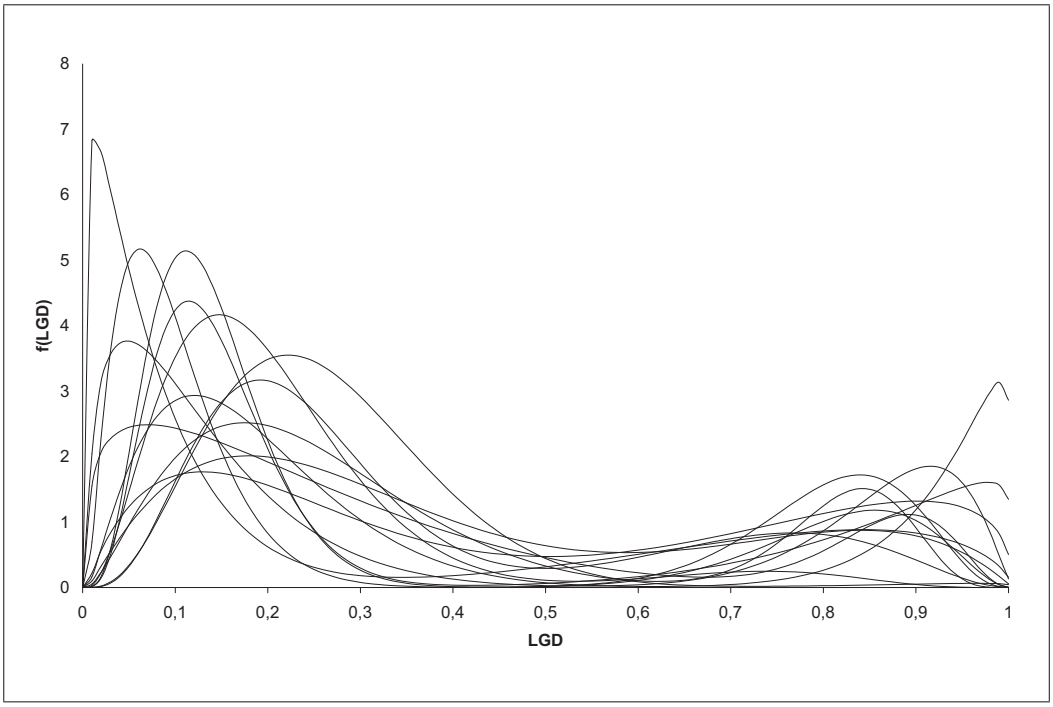
[8] See Appendix 3.

Finally the weight ω_1 is randomly selected from the open interval (0.5,1), where ω_1 can be interpreted as a kind of probability of recovery and $1 - \omega_1$ as a kind of write-off probability. Thus we implicitly assume that the probability of recovery is always larger than the write-off probability.[9]

An illustration of possible mixture densities, which result according to the approach described above, is provided in Figure 2.

Figure 2: Possible mixture densities of two beta distributions

The figure shows possible mixed density functions of two beta distributed random variables over the interval [0,1] according to the approach described earlier.



After creating the mixture distribution, the portfolio generation requires that n realizations from this underlying distribution have to be drawn. In general there are two widely accepted methods to draw realizations from a given distribution in a Monte Carlo simulation. The most famous one is the Inversion Method. Due to the definition of a beta distributed random variable it is not possible to obtain the inverse of the distribution function analytically and this is also true for mixture distributions consisting of beta distributions. Nevertheless we apply this method numerically by dividing the interval [0,1] into 10,001 equidistant points. For each point, η_i, we compute the value of the mixture distribution function which equals the cumulated probability up to this point according to Equation (2). Subsequently we draw n realizations ϱ from a [0,1]-uniformly distributed

[9] All drawn values are, of course, assumed to be uniformly distributed over the respective interval.

random variable and obtain n simulated LGDs (ψ_i) as a solution of:

$$\psi_i = F^{-1}\left(\arg\max_{F(\eta_i)}(\varrho - F(\eta_i)) > 0\right) \tag{8}$$

The second widely accepted method is the Acceptance and Rejection Method where the general idea is to find a target probability distribution T, with given density function $t(x)$, from which random realizations can be easily drawn.[10] Furthermore the target's density should be close to the density of the examined distribution, in our case the mixture distribution. However there is no "simple" distribution which density is close to a bimodal density. Hence when using well-known distributions as the target the algorithm becomes inefficient since there is a great rejection probability resulting from the difference between the two densities. Therefore the numerical Inversion Method described above seems more efficient and reasonable.

3.2 Simulation of Obligor's Parameters

The second part of this section describes the modelling of the obligor-specific parameters. Therefore we use four LGD-influencing variables. Due to descriptive ease we will constrain the simulation of the obligor-specific parameters to fictitious parameters named A, B, C, and D. However at the end of this section possible LGD influencing parameters are provided. Via usage of a copula we will also include dependencies between three of the parameters. After the modelling, Section 3.3 will provide the merger of the LGD and the simulated parameter values.

Parameter A is assumed to be beta distributed with $p = q = 5$. Since A is truncated on the [0,1]-interval it can be interpreted as a kind of ratio, e.g. a financial statement ratio or another credit-influencing ratio. Furthermore we imply a positive causal relation between A and the LGD. The second parameter, B, is normally distributed with $\mu = 0.05$ and $\sigma = 0.2$. This parameter can be interpreted as a profit ratio, e.g. return on investment or return on equity. Thus we imply a negative causal relation between B and the LGD. Parameter C is binary distributed where the cut-off point between the two classes is 0.7, i.e. the probability of an observation of the first class is 70 per cent. This parameter can be interpreted as a kind of status variable, e.g. repeated default, existence of collateralization or classification into junior or senior debt. We imply a negative causal relation between C and the LGD. The last parameter, D, is beta distributed with $p = 2$ and $q = 10$. Like the first parameter, D can be interpreted as a kind of ratio, whereas the causal relation between D and the LGD is assumed to be negative.

[10] The Acceptance and Rejection Method was introduced by von Neumann (1951). For further information see Gentle (2003), pp. 113–125.

Regarding the correlation between all parameters, we assume that parameter A has no causal coherence to any other parameter and thus we assume independency between this parameter and the others. So this means that we can independently draw from the parameter-underlying distribution by using the Inversion Method. The causal coherence between the parameters B, C and D is assumed to be positive between each of them. For the construction of this dependency, we choose a copula approach.

A copula can be, roughly speaking, regarded as a function that joins or couples multivariate distribution functions to their one-dimensional marginal distribution functions and whose one-dimensional margins are uniform.[11] More precisely,[12]

Definition 1. *A function* $C : [0,1]^n \to [0,1]$ *is a n-copula if it enjoys the following properties:*

- $\forall\ u \in [0,1] : C(1,\ldots,1,u,1,\ldots,1) = u$,
- $\forall\ u_i \in [0,1] : C(u_1,\ldots,u_n) = 0$ *if* $\exists\ j \in \{1,\ldots,n\} : u_j = 0$,
- C *is grounded and n-increasing.*

Since the aim of our model is not the analysis of asymmetric dependencies in the parameters, we apply a Gaussian copula and not an Archimedean copula. When using this simulation approach with real historical data one has to be cautious in choosing the appropriate dependency structure, i.e. a Gaussian copula may not be the best fitting dependency model for other parameters.[13]

Now we introduce the construction principle for our case of a three-dimensional Gaussian copula. First we need the correlation matrix ρ which has to be at least positive semidefinite according to the definition of a copula. Our general assumption is a positive dependency between each variable which, however, does not imply the positiv semidefiniteness of ρ. We construct intervals for each correlation parameter between two variables within the copula since we do not want to have a fixed correlation matrix for all simulated portfolios. Therefore we will draw single correlation parameters from the intervals for each portfolio subject to the constraint that the resulting correlation matrix is at least positive semidefinite. Furthermore we assume that the dependency between parameters B and C and the dependency between C and D are at least as high as the dependency between parameters B and D. Since we demand intervals as large as possible, a non-linear optimization problem according to Appendix 5 has to be solved, resulting in the following correlation matrix, where the first row denotes the dependencies of B, the second row

[11] See Nelsen (2006), p. 7.

[12] See Malervergne and Sornette (2006), pp. 103-104.

[13] A general overview of existing copula types is given in, e.g., Nelsen (2006).

denotes the dependencies of C and the last row denotes the dependencies of D.

$$\rho = \begin{pmatrix} 1 & \rho_{B,C} & \rho_{B,D} \\ & 1 & \rho_{C,D} \\ & & 1 \end{pmatrix} = \begin{pmatrix} 1 & [0.5, 0.75] & [0.125, 0.5] \\ & 1 & [0.5, 0.75] \\ & & 1 \end{pmatrix} \tag{9}$$

The next step in the copula computation is the transformation of three independently standard-normally distributed random variables $Z_i,\ i = 1, 2, 3$ into three dependent standard-normally distributed random variables $X_i,\ i = 1, 2, 3$ using the Cholesky decomposition of ρ. Hence the dependent variables are given by:

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= \rho_{B,C} \cdot Z_1 + \sqrt{1 - \rho_{B,C}^2} \cdot Z_2 \\ X_3 &= \rho_{B,D} \cdot Z_1 + \frac{\rho_{C,D} - \rho_{B,D} \cdot \rho_{B,C}}{\sqrt{1 - \rho_{B,C}^2}} \cdot Z_2 \\ &\quad + \sqrt{1 - \rho_{B,D}^2 - (\rho_{C,D} - \rho_{B,D} \cdot \rho_{B,C})^2} \cdot Z_3 \end{aligned} \tag{10}$$

In the next step the three dependent variables are plugged in the distribution function of the standard normal distribution and thus we get three dependent uniformly distributed random variables $U_i,\ i = 1, 2, 3$. Finally the U_i's are used as the input for the Inversion Method to generate observations from the marginal distributions underlying the copula, e.g. U_3 is used as the input for the inverse beta distribution to generate observations for parameter D.[14]

After this simulation procedure we obtained k portfolios each comprising n observations of four obligor parameters. The final step for the portfolio creation is the merger of the four parameters with the simulated LGD value in a systematic manner described in the next part.

3.3 Merging of LGD and Parameters

This part finishes the portfolio creation by combining the four parameters and the LGD into one obligor-specific defaulted loan. The parameters will be matched in a quasi-random procedure where the dependency structure between each parameter and the LGD mentioned in the prior part is applied. However, we include a stochastic component in this

[14] A further step is needed to generate the copula. The computation of the joint distribution involves creating the weighted sum of the inverse of the marginal distributions. Nevertheless we do not need aggregated observations and a meaningful interpretation of this aggregate observation is not given for our case.

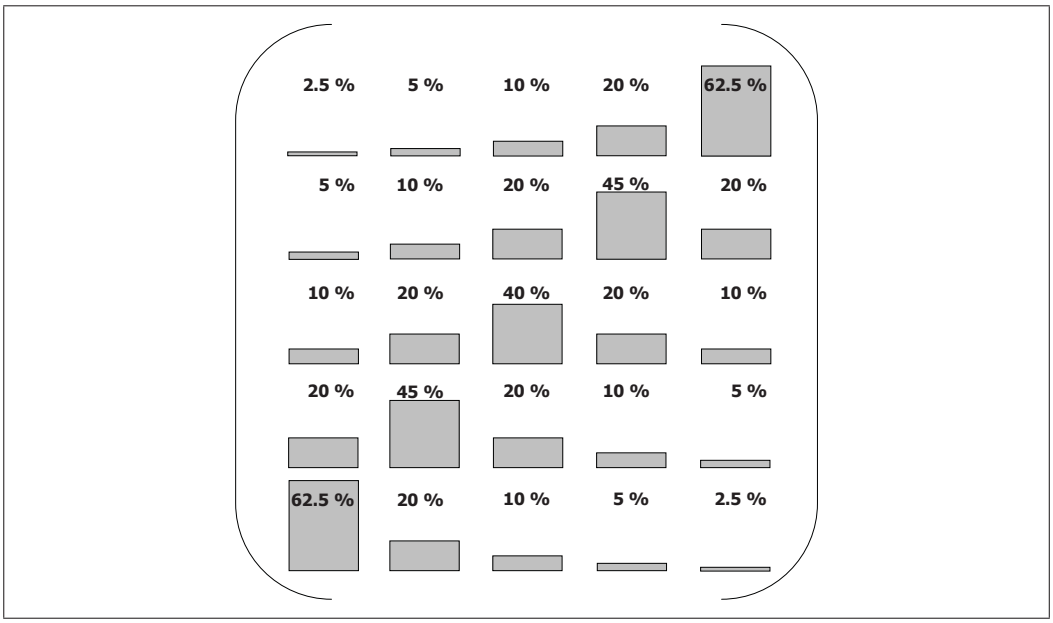
matching to obtain a more realistic credit portfolio. The detailed procedure is described in the following.

First we sort the simulated observations of each parameter, including the LGD, and create quintiles. For the parameters in the copula we apply the sorting procedure and quintile creation to the copula values instead of the simulated observations of the marginal distributions. Parameter A will be matched separately to the LGD since there is no dependency to the other parameters. The matching procedure is done by means of a so called "double-stochastic matrix". Due to the disposition of each parameter into quintiles, we will match each quintile of the obligor's parameter to an LGD quintile according to the specific dependency structure. For the ease of comprehension the matching procedure will be explained for parameters with a negative causal relation to the LGD, e.g. B.

Initially we need a five-by-five matrix M describing our double-stochastic matrix since we want to match quintiles. Each element of M, $m_{i,j},\ i,j = 1,\ldots,5$, describes the probability that an observation of the j-th quintile of (in this example) parameter B is matched to an observation of the i-th quintile of the LGD. For completeness reasons of the matching, the row sum has to equal one for all rows, i.e., $\sum_{j=1}^{5} m_{i,j} = 1\ \forall\ i$, as well as for the column sum, i.e., $\sum_{i=1}^{5} m_{i,j} = 1\ \forall\ j$. A linear program can be solved for the generation of M and is described in Appendix 6. The resulting matrix for our case is presented in Figure 3 where the probability for each $m_{i,j}$ is given.

Figure 3: Matrix M for the matching process

The figure shows the outcome of the linear program according to Appendix 6 presented as the matrix M.



After generation of the matrix, we pick the first observation of the first LGD-quintile. Furthermore a random number N_r from a uniform-[0,1]-distribution is drawn and compared with the cumulated probabilities of the first row of M. If, e.g., $N_r = 0.425$, the first LGD-observation is matched with an observation of parameter B of the fifth quintile. This observation is randomly chosen from the fifth quintile and both matched values are deleted from the set of observations. The procedure is repeated n times for k portfolios resulting in $n \cdot k$ pairs of observations. An obstacle that may occur during the matching process might be a chosen parameter B quintile which is already empty. In this case, the next plausible quintile is chosen.

The whole procedure described for parameter B is done for the copula values instead of the realizations of the marginal distributions within the copula. After matching the copula with the LGD, we replace the copula value with the respective observations of the marginal distributions. One has to consider that for a different dependency structure, e.g. for parameter A, the matrix M is mirrored around the third column. Finally we obtain k matrices with each consisting of five columns (LGD and four obligor parameters) and n rows. Thus each matrix describes a loan portfolio consisting of n obligors.

As can be seen in Figure 3 the matching probabilities for the assumed relationship between the LGD and the obligor-specific parameters are rather small in the sense that, e.g. for the second, third and fourth quintile the probability for the correct classification assumed is below 50 percent, namely 45 percent for the second and fourth quintile and 40 percent for the third quintile. This matching structure allows for many outliers, which finally leads to a less predictive relationship between the obligor-specific parameters and the LGD. Thus we will refer to this resulting data set as the "bad data set" in the further sections of this paper. To enhance the general validity of our upcoming estimation model we create a second data set according to the same process described above. However, this data set exhibits a quite higher predictive relationship due to an increase of the probabilities for the correct classification assumed and is in the further proceeding referred to as "good data set".[15] So both data sets can be used to verify the performance quality of our model and the benchmark models developed in the next Section.

This finishes the simulation part of our paper. We will use the simulated portfolios in the following sections for analyzing different LGD-estimation models. Our simulation approach is different to all existing methods described in the literature to model credit portfolios where the idea is focused on a kind of macroeconomic description of a loan portfolio in the sense of using a superordinated systematic risk factor. Instead our approach is more focused on the idea of the microstructure of the loan portfolio in the sense of using obligor-specific properties. Therefore banks are able to apply this approach to

[15] The corresponding matrix for the matching process for the good data set can be found in Appendix 7.

get any synthesized number of credit portfolios using historical data to approximate the parameter-underlying distributions. This is especially important if the size of the real credit portfolio is large enough for obtaining the distributions and dependencies of the obligor's parameters but too small for a decomposition of the portfolio for further analyses like modeling, validating and stress testing. Additionally the described approach is adjustable for simulating similar problem sets, e.g., probability of default estimations, rating function derivations or pricing models.

As mentioned before, the portfolio consists of fictitious obligor-specific parameters. A possible LGD-influencing factor might be the rating of the obligor. The rating can be considered as a measure for the default probability of the obligor. Several studies found an empirical evidence for a positive relation between default probability and LGD.[16] Another LGD-influencing factor is represented by the seniority of a loan. The relationship between seniority and LGD is assumed to be negative since it is obvious that senior loans are served prior to junior loans in case of a bankruptcy. Also the industry and macroeconomic factors that are influenced by the business cycle might affect the LGD as shown in Schuermann (2005).

4 Estimation Model and Benchmark Models

This part introduces the estimation model applied. As our LGD-estimation shall account for the bimodality of the LGD-distribution, we initially need to estimate the parameters of the mixture distribution. Since we assume the LGD-distribution being a mixture of two beta distributions, two commonly used algorithms are introduced for the estimation of the two parameters, p and q, of each beta distribution and the weight ω_1. We furthermore analyze the efficiency and reliability of both algorithms since up to our knowledge this analysis has not been done for this special case of mixture distributions.

4.1 Maximum Likelihood Algorithm

The well-known Maximum Likelihood (ML) algorithm can be used to compute the parameters and weights of a mixture distribution but as we will see shortly the problem may be analytically intractable. The likelihood function for a mixture distribution consisting of two distributions reads as:

$$L_{\mathrm{ML}} = \prod_{i=1}^{n} f(\upsilon_i) = \prod_{i=1}^{n} \left(\omega_1 \cdot f_1(\upsilon_i) + (1-\omega_1) \cdot f_2(\upsilon_i)\right) \tag{11}$$

[16] See e.g. Bakshi et al. (2001) and Jokivuolle and Peura (2003). For a detailed literature overview on this topic see Altman (2006).

and thus the log likelihood function reads as:

$$\log L_{\mathrm{ML}} \quad = \quad \sum_{i=1}^{n} \log\left(\omega_1 \cdot f_1(\upsilon_i) + (1-\omega_1) \cdot f_2(\upsilon_i)\right). \tag{12}$$

As we assume a mixture consisting of two beta distributed random variables, an analytical solution of the maximization is not possible. Therefore a numerical procedure has to be used where the iterative step of the enhancement of the parameter vector θ reads as:

$$\theta^{(k+1)} \quad = \quad \theta^{(k)} - \left(\frac{\partial^2 \log L_{\mathrm{ML}}}{\partial\theta\partial\theta'}\right)^{-1} \cdot \frac{\partial \log L_{\mathrm{ML}}}{\partial\theta^{(k)}} \tag{13}$$

This approach is the well-known Newton-Raphson method and is considered as a fast and reliable approximation algorithm. Thus we need the first and second (partial) derivatives of the log likelihood function for the gradient and the Hessian inverse. The iterative procedure is repeated until the relative change in the log likelihood function does not exceed a predefined level of tolerance.

4.2 Expectation Maximization Algorithm

The Expectation Maximization (EM) algorithm introduced by Dempster et al. (1977) mainly comprises of two steps, the so called E-step and the M-step. First it is assumed that a data set Υ is observed and contains observations described by some distribution. The data set Υ is assumed to be an uncomplete data set. Therefore the existence of a complete data set $\Lambda = (\Upsilon, \Xi)$ is assumed, where Ξ is an additional data set containing the missing information. Ξ is assumed to be random and unobservable. Furthermore a distribution for Λ is assumed as well, which is just the joint distribution of (Υ, Ξ). Hence the density reads as:

$$f(\lambda|\Theta) = f(\upsilon, \xi|\Theta) = f(\xi|\upsilon, \Theta) \cdot f(\upsilon|\Theta), \tag{14}$$

where Θ is the set of distribution parameters that are to be estimated. Finally the complete data likelihood function denoted as L_{EM} can be set:

$$L_{\mathrm{EM}} = \prod_{i=1}^{n} f(\upsilon_i, \xi_i|\Theta). \tag{15}$$

In our case of a mixture of two beta distributions, the density of the incomplete data set Υ reads as:

$$f(\upsilon_i|\Theta) = \sum_{j=1}^{2} \omega_j \cdot f_j(\upsilon_i|\theta_j), \tag{16}$$

where $\Theta = (\omega_1, \omega_2, \theta_1 (= p_1, q_1), \theta_2 (= p_2, q_2))$ and f_j denotes the density of the right-skewed beta distribution (j=1) or the left-skewed beta distribution (j=2). Thus the log likelihood function of the incomplete data set results in:

$$\log L_{\mathrm{EM}} = \prod_{i=1}^{n} f(\upsilon_i, \xi_i|\Theta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{2} \omega_j \cdot f_j(\upsilon_i|\theta_j) \right). \tag{17}$$

Since the optimization of the logarithm of a sum is pretty difficult, the above mentioned concept about the existence of additional data simplifies the optimization problem. In our case the additional data set Ξ is defined as:

$$\xi_i = \begin{cases} 1 & \text{if observation } i \text{ belongs to the first (right-skewed) beta distribution} \\ 2 & \text{if observation } i \text{ belongs to the second (left-skewed) beta distribution.} \end{cases} \tag{18}$$

If the additional data Ξ is known, the log likelihood rearranges to:

$$\log L_{\mathrm{EM}} = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{2} \mathbb{1}_{\{\xi_i = j\}} \cdot \omega_j \cdot f_j(\upsilon_i|\theta_j) \right), \tag{19}$$

where still a logarithm of a sum is observable but for every observation just one of the summands remains. The reason is that the observations under the additional information are from mutually exclusive densities, i.e. one observation can only be generated by one density. Thus the logarithm of the remaining summand is just the logarithm of a product.

The problem that the additional data set is not known still remains. However it is assumed that the additional data is random and so by taking the conditional expectation of the complete data log likelihood function with respect to the additional data Ξ (known as the E-Step) we end up with a function only depending on the observable data Υ and the parameters Θ that are to be estimated. For optimizing this function, we need both initial values for the density parameters and the distribution of the additional data for taking the expectation. Therefore the conditional expectation of the k-th iteration reads as:

$$\begin{aligned} \mathrm{E}_{\Xi}\left[\log L_{\mathrm{EM}}|\Theta^{(k-1)}\right] &= \mathrm{E}_{\Xi}\Bigg[\sum_{i=1}^{n} \log \Big(\mathbb{1}_{\{\xi_i^{(k-1)}=1\}} \cdot \omega_1^{(k)} \cdot f_1(\upsilon_i|\theta_1^{(k)}) \\ &\quad + \mathbb{1}_{\{\xi_i^{(k-1)}=2\}} \cdot (1-\omega_1^{(k)}) \cdot f_2(\upsilon_i|\theta_2^{(k)}) \Big) \Bigg] \\ &= \sum_{i=1}^{n} \Big(\mathrm{E}_{\Xi}(\mathbb{1}_{\{\xi_i^{(k-1)}=1\}}) \cdot (\log f_1(\upsilon_i, \theta_1^{(k)}) + \log \omega_1^{(k)}) \\ &\quad + \mathrm{E}_{\Xi}(\mathbb{1}_{\{\xi_i^{(k-1)}=2\}}) \cdot \Big(\log f_2(\upsilon_i, \theta_2^{(k)}) + \log(1-\omega_1^{(k)}) \Big) \Big) \end{aligned} \tag{20}$$

Since the expectation is just taken with respect to Ξ the indicator functions have to be

considered only. So the expectations result in:

$$\begin{aligned}
\mathrm{E}_{\Xi}\left(\mathbb{1}_{\{\xi_i^{(k-1)}=1\}}\right) &= 1\cdot\frac{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}+0\cdot\frac{\omega_2^{(k-1)}\cdot f_2(\upsilon_i,\theta_2^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}\\
&= \frac{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}\\
\text{and analogously}&\\
\mathrm{E}_{\Xi}\left(\mathbb{1}_{\{\xi_i^{(k-1)}=2\}}\right) &= \frac{\omega_2^{(k-1)}\cdot f_2(\upsilon_i,\theta_2^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})} \qquad (21)
\end{aligned}$$

Inserting (21) in (20) results in:

$$\begin{aligned}
\mathrm{E}_{\Xi}\left[\log L_{\mathrm{EM}}|\Theta^{(k-1)}\right] &= \sum_{i=1}^{n}\left(\frac{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}\cdot\left(\log f_1(\upsilon_i,\theta_1^{(k)})+\log\omega_1^{(k)}\right)\right.\\
&\quad+\frac{\omega_2^{(k-1)}\cdot f_2(\upsilon_i,\theta_2^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}\cdot\log f_2(\upsilon_i,\theta_2^{(k)})\\
&\quad\left.+\frac{\omega_2^{(k-1)}\cdot f_2(\upsilon_i,\theta_2^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}\cdot\log\left(1-\omega_1^{(k)}\right)\right) \qquad (22)
\end{aligned}$$

The M-step requires the optimization of $\Theta^{(k)}$, which includes the parameters $p_1^{(k)}, q_1^{(k)}, p_2^{(k)}, q_2^{(k)}$ and $\omega_1^{(k)}$. The resulting estimators are subsequently used again for the next iteration of the EM algorithm. Below we show the derivation of the optimal estimator for $\omega_1^{(k)}$, which is just the sum of the posterior probabilities over all observations divided by the number of observations.

$$\begin{aligned}
0\overset{!}{=}\frac{\partial\mathrm{E}_{\Xi}\left[\log L_{\mathrm{EM}}|\Theta^{(k-1)}\right]}{\partial\omega_1^{(k)}} &= \sum_{i=1}^{n}\left(\frac{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}\cdot\frac{1}{\omega_1^{(k)}}\right.\\
&\quad\left.-\frac{\omega_2^{(k-1)}\cdot f_2(\upsilon_i,\theta_2^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}\cdot\frac{1}{(1-\omega_1^{(k)})}\right)\\
\sum_{i=1}^{n}\frac{\omega_2^{(k-1)}\cdot f_2(\upsilon_i,\theta_2^{(k-1)})\cdot\omega_1^{(k)}}{f(\upsilon_i,\Theta^{(k-1)})\cdot(1-\omega_1^{(k)})\cdot\omega_1^{(k)}} &= \sum_{i=1}^{n}\frac{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})\cdot(1-\omega_1^{(k)})}{f(\upsilon_i,\Theta^{(k-1)})\cdot(1-\omega_1^{(k)})\cdot\omega_1^{(k)}}\\
\sum_{i=1}^{n}\frac{\omega_2^{(k-1)}\cdot f_2(\upsilon_i,\theta_2^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})} &= \sum_{i=1}^{n}\frac{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})\cdot\omega_1^{(k)}}\\
&\quad-\sum_{i=1}^{n}\frac{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}\\
\sum_{i=1}^{n}\frac{\overbrace{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})+\omega_2^{(k-1)}\cdot f_2(\upsilon_i,\theta_2^{(k-1)})}^{=f(\upsilon_i,\Theta^{(k-1)})}}{f(\upsilon_i,\Theta^{(k-1)})} &= \sum_{i=1}^{n}\frac{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})\cdot\omega_1^{(k)}}\\
\omega_1^{(k)} &= \frac{\sum_{i=1}^{n}\frac{\omega_1^{(k-1)}\cdot f_1(\upsilon_i,\theta_1^{(k-1)})}{f(\upsilon_i,\Theta^{(k-1)})}}{n} \qquad (23)
\end{aligned}$$

The estimators for the parameters p and q of each distribution in the mixture can be derived analogously. However the derivation of these estimators is not that straightforward since it here involves the derivative of the beta density. Beckman and Tietjen (1978) and others show that it is quite complicated to solve the likelihood equations. We therefore consider the Method of Moments (MM) estimators which read as:

$$\begin{aligned} p &= \overline{v} \cdot \left(\frac{\overline{v}}{s_v^2} \cdot (1 - \overline{v}) - 1 \right) \\ q &= (1 - \overline{v}) \cdot \left(\frac{\overline{v}}{s_v^2} \cdot (1 - \overline{v}) - 1 \right), \end{aligned} \tag{24}$$

where s^2 denotes the sample variance and $\overline{v}$ denotes the sample mean. Due to our modification of using the MM estimators for the parameters p and q, we only obtain an adjusted EM (aEM) algorithm instead of the pure EM algorithm. We avoid the usage of numerical derivatives as in the ML algorithm due to computational effort.

The general idea of our simulation is to obtain 1,000 credit portfolios each containing 10,000 creditors with individual LGDs and obligor-specific parameters. We had to repeat the above mentioned algorithms 1,385 times to achieve the required number of portfolios, where the ML algorithm stopped 385 times due to non-convergence and in 197 cases the aEM algorithm did not converge. However, in these cases the ML algorithm did not converge as well. In consequence we can state that in our case the aEM algorithm has better convergence properties than the ML algorithm.

In a next step we look at the reliability of our estimators by comparing the mean squared errors (MSE) over all portfolios for the parameters of the mixture distributions. Table 2 shows the results of this reliability analysis.

Table 2: Parameter reliability analysis

The table provides the mean squared errors of each parameter estimated in the mixture model, namely the two parameters of each beta distribution and the weight.

Algorithm	MSE of p_1	MSE of q_1	MSE of p_2	MSE of q_2	MSE of ω_1
ML	0.0055	0.1583	1.9490	0.2153	$1.7968 \cdot 10^{-4}$
aEM	0.0158	0.3768	10.8126	0.3426	$1.7934 \cdot 10^{-4}$

It can clearly be seen that the ML estimators are much more accurate in all parameters than the corresponding aEM estimators except for the estimation of the weighting factor. However the difference between both MSEs is negligible. Resuming the analysis we obtain that the ML algorithm provides more accuracy than the aEM algorithm but at the expense of the convergency properties.[17] Due to the relative small values of the MSEs for each

[17] We omit a further statistical analysis due to the clear difference in the MSE values and the number of observations. A statistical test on equality of the parameter values would lead to similar results.

parameter, the resulting LGD estimation based on these estimated parameters will only slightly deviate from the "true" LGD values underlying the simulation.

4.3 Estimation of LGDs with Obligor-Specific Parameters

Now we will explain our complete estimation model in more detail. It generally consists of five steps with small deviations depending on the estimation approach. First the distribution parameters of the mixture distribution, i.e. p_1, q_1, p_2, q_2 and ω_1, are estimated via either the ML or aEM algorithm. However we only apply the ML algorithm due to the better accuracy in the parameter estimation. The next step requires the computation of the estimated distribution values for both beta distributions over all LGDs (compare Equation (1)). There are two different approaches pursued in the third step. One requires the computation of the mixture distribution values followed by a regression of the transformed mixture distribution values on the obligor-specific parameters. The transformation is done due to the bounded nature of the dependent variable and will be discussed in more detail below. The second alternative carried out in this step requires the regression of the transformed single distribution values on the obligor-specific parameters for each beta distribution. Subsequently both estimators of the regressions are weighted with estimated ω. The fourth step is equal for both alternatives and contains the re-transformation of the estimators into an again bounded estimator for the distribution value of the mixture. Finally the LGDs are estimated via the Inversion Method described at the end of Section 3.1.

Now the single steps are described in more detail. The first step was described in detail in the Sections 4.1 and 4.2. By using the estimated parameter values, the required estimated distribution values can be computed in the second step. The third step first involves the transformation of the estimated distribution values of either the mixture or the single beta distribution values. Two different transformation types are used, namely the logit transformation and the normal transformation. The logit transformation reads as:

$$\begin{aligned} y_i^{\log} &= \log \frac{F(\upsilon_i)}{1 - F(\upsilon_i)} \\ \text{and accordingly} & \\ y_i^{j,\log} &= \log \frac{F_j(\upsilon_i)}{1 - F_j(\upsilon_i)}, \quad j = 1, 2, \end{aligned} \tag{25}$$

and the normal transformation reads as:

$$\begin{aligned} y_i^{\text{norm}} &= \Phi^{-1}(\mu, \sigma, F(\upsilon_i)) \\ \text{and accordingly} & \\ y_i^{j,\text{norm}} &= \Phi^{-1}(\mu, \sigma, F_j(\upsilon_i)), \quad j = 1, 2. \end{aligned} \tag{26}$$

The argument υ_i is the i-th observation of the realized LGDs, $F(.)$ denotes the mixture distribution and $F_j(.)$ denotes the single beta distribution. The parameters μ and σ are the mean and the standard deviation of the mixture distribution which act as the location and scale parameter in the normal distribution and Φ^{-1} denotes the inverse distribution function of a standard normally distributed random variable. After the transformation a linear regression of the estimated distribution values for either the mixture distribution or each beta distribution on the obligor-specific parameters is carried out.

In the fourth step, the estimators are either re-transformed in case of the regression involving the mixture distribution or the ω-weighted estimators from each beta distribution are re-transformed. The final step comprises the computation of the LGD via the inverse of the estimated mixture distribution values by applying the Inversion Method according to Equation (8). For future LGD estimations one would skip the first two steps and would just insert the obligor-specific parameters into the regression model of the third step. Afterwards one would proceed with the last two steps.

4.4 Benchmark Models

We will now introduce the before mentioned benchmark models in the way we will apply them in our analysis. Starting with the approach of Hamerle et al. (2006) we first transform the bounded LGDs into an unbounded variable y^{Ham} by the logit transformation:

$$y_i^{\text{Ham}} = \log \frac{\text{LGD}_i}{1 - \text{LGD}_i}. \tag{27}$$

Afterwards an OLS regression is applied to the transformed LGDs. Since we do not have any time consideration, the panel regression described by Hamerle et al. (2006) is reduced to an OLS regression in our case. The dependent variable y^{ham} is retransformed into an estimation of the LGD.

The second benchmark model is in the style of the LossCalc model by Moody's KMV described by Gupton and Stein (2005). The assumption underlying the model is an LGD described by just one beta distribution. Therefore a transformation of the LGD is done by normalizing the LGD via a beta transformation:

$$y_i^{\text{KMV}} \quad = \quad \Phi^{-1}(\mu, \sigma, \text{Beta}(\text{LGD}_i)), \tag{28}$$

where μ and σ are the estimators of the mean and the standard deviation of the beta distribution. Analogously to the fist benchmark model an OLS regression is applied. The

retransformation is done by:

$$\begin{aligned} \widehat{\text{LGD}}_i &= \text{Beta}^{-1}(p, q, \Phi(\widehat{y}_i)), \\ \text{where} & \\ p &= \frac{\mu^2 \cdot (1-\mu)}{\sigma^2} - \mu \qquad q = \frac{1}{\mu} - 1. \end{aligned} \tag{29}$$

As a third benchmark model we choose the regression tree model applied to recovery rates proposed by Bastos (2009). Regression trees are nonparametric and nonlinear predictive models and were first introduced by Breiman et al. (1998). The general idea is a clustering according to certain splitting rules of the credit portfolio according to the obligor-specific parameters into more homogeneous, mutually exclusive subportfolios. Logical if-then conditions lead to a decomposition of a root node into daughter nodes by a recursive search algorithm.[18] There are certain splitting rules applicable, where Bastos (2009) refers to the so-called "maximum deviance reduction" rule. The referring decision measure of this rule is denoted as DR and reads as:

$$DR = \sigma_T - \frac{\mu_{T1}}{\mu_T} \cdot \sigma_{T1} - \frac{\mu_{T2}}{\mu_T} \cdot \sigma_{T2}, \tag{30}$$

where σ and μ denote the standard deviation and the mean of the nodes named in the subscript, T denotes the parent node and $T1$ and $T2$ are the observation set of the daughter nodes. In our case we implemented the model via the MATLAB routine `classregtree` where the initial tree is reduced to a certain level (deepness of the tree) by the command `prune`. The best level is the one that produces the smallest tree that is within one standard error of the minimum-cost subtree.

5 Results

We first define the following abbreviations:

- HO^{log}: mixture distribution model with logit transformation
- HO^{norm}: mixture distribution model with normal transformation
- Bas: benchmark model with the approach according to Bastos (2009)
- KMV: benchmark model with the approach according to Gupton and Stein (2005)
- Ham: benchmark model with the approach according to Hamerle et al. (2006).

[18] We omit the technical details of the procedure since it is not in the focus of our paper. For more details see Breiman et al. (1998).

We will refer to these abbreviations in the following parts. The models where the regression is done on the single beta distribution values are not analyzed in more detail since the analysis showed that these models could not exhibit a constant sufficient performance over all portfolios. This may be due to the kind of data, which we cannot check since no real data are available. It may also be due to the kind of regression and transformation process.

The complete analysis is done for both the good and the bad data set.[19] First the mean absolute deviations (MAD) of the estimated LGDs from the realized LGDs are computed for each portfolio and each modelling approach. The minimum, the maximum and the mean MAD for the bad data set are presented in Table 3.

Table 3: Results for the bad data set
The table shows the mean MAD and the standard deviation of the MAD (std MAD) over all portfolios, the minimum MAD, and the maximum MAD for our two models and the three benchmark models for the bad data set. All results are multiplied with a factor 100 so that they can be interpreted as percentage points.

model	mean MAD	std MAD	min MAD	max MAD
HO^{log}	12.8501	3.5467	3.0184	19.2286
HO^{norm}	12.8707	3.5467	3.0286	19.2307
Bas	13.8004	3.7838	3.5422	22.7381
KMV	15.0221	4.0899	3.0297	23.2934
Ham	14.0952	4.1663	3.0290	21.7918

It can be seen that our models are at least one percentage point better than the benchmark models with respect to the mean MAD, which results in an accuracy improvement of at least 6.94 percent (HO^{norm} vs. Bas) up to 11.28 percent (HO^{log} vs. KMV). Furthermore the standard deviations of the MAD of our models are also smaller than the ones of the benchmark models. The relative differences between the standard deviations of the MAD between our models and the benchmark models reach from 6.27 percent ($HO^{norm/log}$ vs. Bas) to 14.87 percent ($HO^{norm/log}$ vs. Ham). This result is additionally emphasized by the lower maximal values of the MADs.

The results for the good data set are presented in Table 4. In case of the good data set it is obvious that our approaches are still better than the models according to Gupton and Stein (2005) and Hamerle et al. (2006) with respect to the mean MAD. Both approaches offer mean MADs that are at least 16.37 percent lower than the mean MAD of the model proposed by Hamerle et al. (2006). In comparison to KMV our both approaches even

[19] It has to be noted that the KMV model did not work properly in 55 cases of all portfolios. This problem arose since the small values obtained after the transformation could not be processed by the computer. These portfolios were only removed for the analysis of the KMV model.

Table 4: Results for the good data set

The table shows the mean MAD and the standard deviation of the MAD (std MAD) over all portfolios, the minimum MAD, and the maximum MAD for our two models and the three benchmark models for the good data set. All results are multiplied with a factor 100 so that they can be interpreted as percentage points.

model	mean MAD	std MAD	min MAD	max MAD
HO^{log}	8.0929	1.7000	2.0748	11.6390
HO^{norm}	8.0730	1.6723	2.0884	11.6298
Bas	7.5675	1.4896	2.1293	11.9463
KMV	10.8753	2.7582	2.0873	17.8138
Ham	9.9110	2.5897	2.0988	17.3259

posses a mean MAD that is more than 22 percent smaller. When comparing the standard deviations of the MADs and choosing the standard deviation of HO^{log} as the benchmark, we are still more than 34 percent smaller than the model of Hamerle et al. (2006) and more than 38 percent smaller than the KMV model. Thus our models offer a considerably better accuracy in the LGD estimation than both just mentioned benchmark models.

Regarding the model proposed by Bastos (2009) the mean and the standard deviation of the MAD are lower than those of our models. However our models are better in case of considering the extreme values of the MADs. It has to be mentioned that there are some critical points regarding regression trees. One problem is the rigidity of the approach. Since it is a kind of clustering model the continuity of the influencing parameters vanishes and is reduced to a binary decision rule. This means that depending on the branching structure and deepness of the tree an outlier in just one influencing factor may lead to completely misspecified LGDs which may also lead to counterintuitive interpretations. This is especially important when having a data set that offers a weak predictive power of the influencing factors on the LGD. A second disadvantage of the regression tree model compared to a straight-forward regression approach is the inflexibility. This means that regular regression approaches can use nearly every functional dependency between the explanatory variables and the dependent variable and even between the explanatory variables themselves.

We exemplarily choose a portfolio to illustrate the regression results and the corresponding estimated LGD distributions. This is done for the good data set as well as for the bad data set. In Table 5 the regression coefficients, their significance and the coefficient of determination for our models, the KMV model and the model according to Hamerle et al. (2006) are presented. The regression coefficients are obtained from the linear regression $\hat{y}_i = b_0 + b_1 \cdot A_i + b_2 \cdot B_i + b_3 \cdot C_i + b_4 \cdot D_i$, where A, B, C, and D are the obligor-specific parameters and $\hat{y}$ is either the transformed realized LGD or the transformed distribution

value of the realized LGD. Since the regression tree model proposed by Bastos (2009) does not exhibit any functional relationship we could not obtain any regression coefficients nor a coefficient of determination and is therefore excluded from the table. The regression tree for the good data set of the exemplarily chosen portfolio can be found in Appendix 8.

Table 5: Estimation results for the randomly chosen portfolio
The table shows the regression coefficients for the four regression models according to the equation $\hat{y}_i = b_0 + b_1 \cdot A_i + b_2 \cdot B_i + b_3 \cdot C_i + b_4 \cdot D_i$. Furthermore the coefficient of determination is given as well. A double asterisk indicates significance on the one percent level and a single asterisk indicates significance on the five percent level.

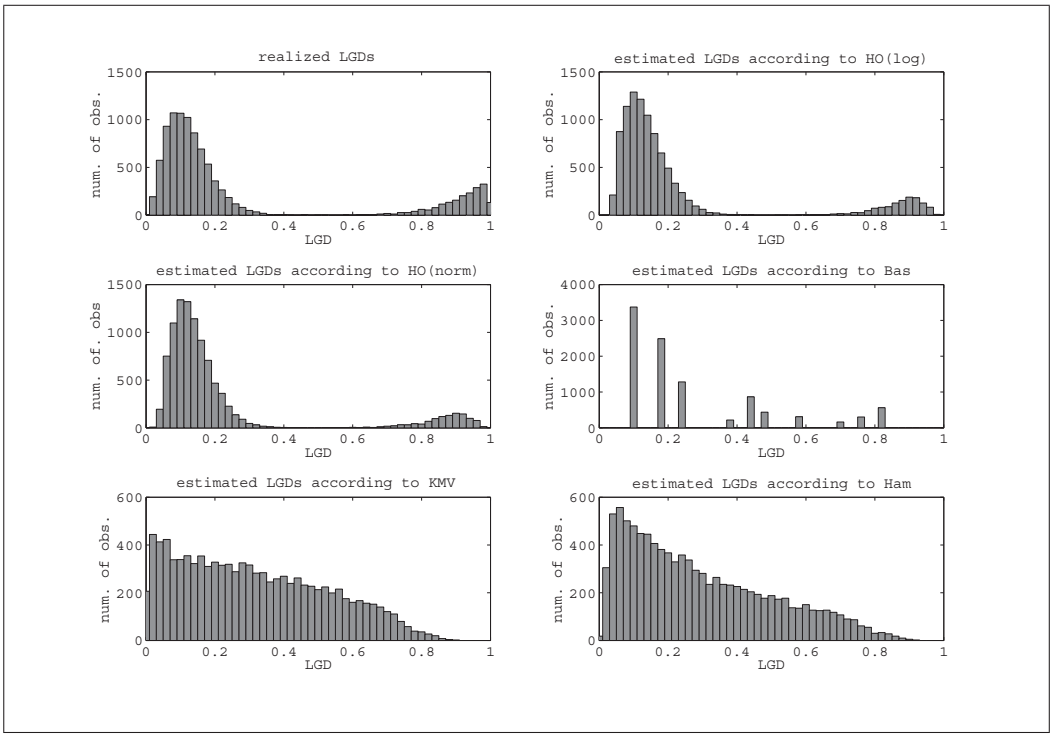
	HO^{log}	HO^{norm}	KMV	Ham
	bad data set			
b_0	-2.0267^{**}	-0.0901^{**}	0.2392^{**}	-3.4135^{**}
b_1	5.4244^{**}	0.9783^{**}	1.0960^{**}	5.7675^{**}
b_2	-1.7129^{**}	-0.3128^{**}	-0.3564^{**}	-1.8646^{**}
b_3	-0.3075^{**}	-0.0569^{**}	-0.0225^{**}	-0.1507^{**}
b_4	-3.1045^{**}	-0.5615^{**}	-0.5727^{**}	-3.0556^{**}
R^2	0.4536	0.4834	0.3361	0.3641
	good data set			
b_0	-2.4484^{**}	-0.1733^{**}	0.0746^{**}	-4.2057^{**}
b_1	6.2471^{**}	1.1377^{**}	1.3636^{**}	7.0808^{**}
b_2	-1.8225^{**}	-0.3263^{**}	-0.3975^{**}	-2.0710^{**}
b_3	-0.3636^{**}	-0.0643^{**}	0.0191^{*}	0.0224
b_4	-2.9174^{**}	-0.5245^{**}	-0.4529^{**}	-2.5056^{**}
R^2	0.6667	0.7081	0.4919	0.5328

The table shows that all regression coefficients are highly significant for all models regarding the bad data set. The influence direction of the obligor-specific parameters (indicated by the sign of the coefficients) is the same for all models and the same to the one proposed by our simulation. Furthermore the coefficients of determination are higher for our approaches compared to the benchmark models but they all are below 0.5. Regarding the good data set the regression coefficients are again always significant and exhibit the assumed direction of influence of the obligor-specific parameters on the LGD for our approaches. This is also true for both benchmark models regarding the parameters A, B, and D. However parameter C exhibits the contrary direction of influence on the LGD but is not significant for the model according to Hamerle et al. (2006). The coefficients of determination have increased for all models whereas our models still offer higher coefficients of determination.

Figures 4 and 5 present the distributions of the estimated LGDs in comparison to the distribution of the realized LGDs for the exemplarily chosen portfolio.

Figure 4: Histogram for the exemplarily chosen portfolio and the bad data set

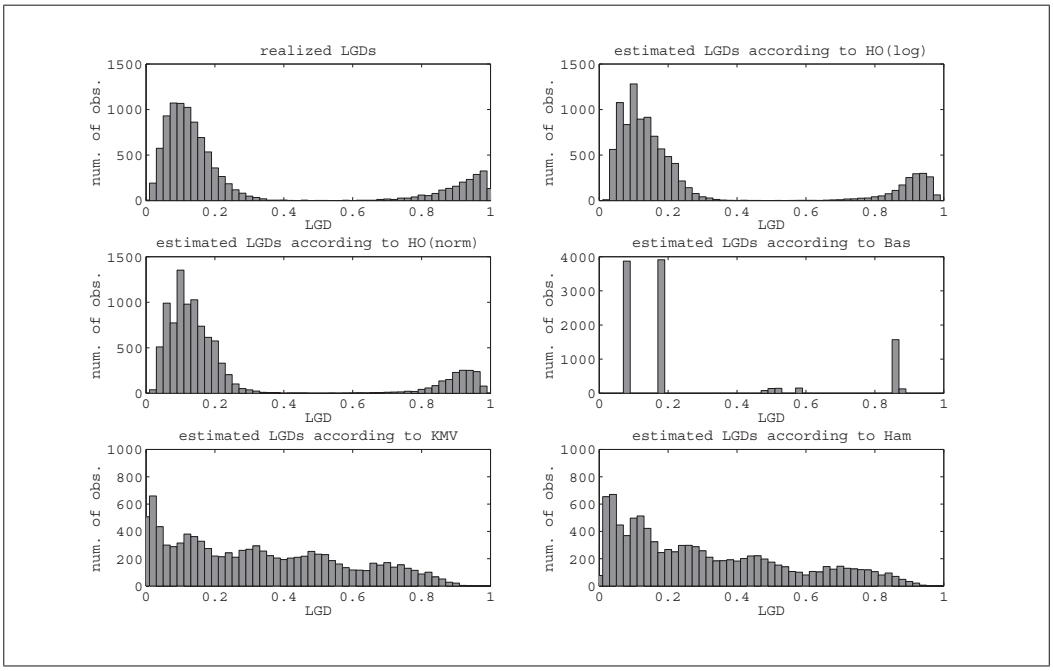
The figure shows the distributions of the realized and estimated LGDs of the bad data set.



It can clearly be seen that using our approaches the assumed bimodality of the realized LGDs is maintained. The regression tree model proposed by Bastos (2009) exhibits only a few number of classes (i.e. classes in the histogram) compared to the other models. This is not surprising since there are only a few decision nodes in the optimal tree. Of course the number of decision nodes can be enlarged and a bimodal structure can be obtained but the resulting tree will be more complex and can grow exponentially which results in an exponential growth of the number of decision pathes and makes the model inoperative for practical purposes. The models according to Gupton and Stein (2005) and Hamerle et al. (2006) produce similar results whereas in both cases no bimodal structure occurs and a lot of observations can be found in the interval [0.3,0.7] where only very few realized LGDs are present. The results regarding the good data set are similar and need no further explanation.

Figure 5: Histogram for the exemplarily chosen portfolio and the good data set

The figure shows the distributions of the realized and estimated LGDs of the good data set.



6 Conclusion

The aim of the paper was to develop an LGD estimation model that incorporates the two main distribution characteristics of the loss given default, namely the bounded character and the bimodality. Although the bounded character is accounted for in the recent literature the bimodality, however, has not been incorporated in any model up to our knowledge. Here we fill this gap by introducing an estimation model that is able to handle the bimodal structure of the distribution.

We introduced a mixture distribution approach comprised of two beta distributions and applied both the Maximum Likelihood algorithm and an adjusted version of the Expectation Maximization algorithm to estimate the distribution parameters of the mixture. We have chosen a mixture of two beta distributions since they are quite flexible in approximating different distribution shapes and account for the bounded character as well. However these advantages have to be paid by the pretty difficult derivative characteristics of the incorporated Gamma function. Here the Maximum Likelihood algorithm can only be used numerically and for parts of the estimations done by the Expectation Maximization algorithm the Method of Moments estimators were used instead. It was shown that the Maximum Likelihood algorithm was more accurate in parts of the parameter estimations but could not compete with the convergence characteristics of the adjusted Expectation Maximization algorithm.

The parameter estimators obtained by one of the mentioned algorithms are subsequently used for the estimation of the distribution values of the realized LGDs, which serve as the dependent variable within the regression model afterwards. In comparison to some benchmark models described in the literature our approach clearly outperformed these benchmark models according to the accuracy, deviations of the estimated LGDs and the shape of the resulting estimator's distribution especially accounting for the bimodality. There is still some space for developments of our approach in the regression model since we just used a linear regression for simplifications and our model can be extended to a multi-modal approach which enlarges the field of applications.

The whole analysis regarding our model and the benchmark models was carried out on synthesized loan portfolios each comprising of 10,000 obligors. Therefore we simulate obligor-specific parameters which are either independent from all others or are part of a certain dependency structure modelled via a Gaussian copula. The simulated obligor-specific parameters are matched with the realized LGDs via a quasi-random approach. This proceeding may of course reduce the validity of our results but we try to counteract this possible shortcoming by simulating two data sets of credit portfolios with different predictive power of the obligor-specific parameters on the LGDs within each portfolio.

In all circumstances it could be shown that our approach is superior to the benchmark models. Due to these advantages obtained, our approach is especially preferable when pricing credits. Here not only unbiased estimates are required but also precision for each single estimation. Given this accuracy in our estimation procedure one can now enhance the integration of the LGD estimation in models where the LGD is assumed to be an exogenous influencing factor, especially regarding the interface of probability of default and LGD.

References

Altman, E. I. (2006), 'Default Recovery Rates and LGD in Credit Risk Modeling and Practice: An Updated Review of the Literature and Empirical Evidence'. Working Paper, New York University, Stern School of Business.

Appasamy, B., Dörr, U., Ebel, H. and Stützle, E. A. (2008), 'LGD-Schätzung im Retailgeschäft am Beispiel Automobilfinanzierung', *Zeitschrift für das gesamte Kreditwesen* (5), 206–209.

Araten, M., Jacobs, M. and Varshney, P. (2004), 'Measuring LGD on Commercial Loans: An 18-Year Internal Study', *The RMA Journal* **86**, 28–35.

Bakshi, G., Madan, D. and Zhang, F. (2001), 'Understanding the Role of Recovery in Default Risk Models: Empirical Comparisons and Implied Recovery Rates'. Finance

and Economics Discussion Series, Federal Reserve Board of Governors, Washington D.C.

Bassamboo, A., Juneja, S. and Zeevi, A. (2008), 'Portfolio Credit Risk with Extremal Dependence: Asymptotic Analysis and Efficient Simulation', *Operations Research* **56**(3), 593–606.

Bastos, J. A. (2009), 'Forcasting bank loans loss-given-default'. Working Paper.

Beckman, R. J. and Tietjen, G. L. (1978), 'Maximum Likelihood Estimation for the Beta Distribution', *Journal of Statistical Computation and Simulation* **7**(3), 253–258.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1998), *Classification and Regression Trees*, Chapman & Hall/CRC.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society B* **39**, 1–38.

Dermine, J. and de Carvalho, C. N. (2005), 'Bank Loan Losses-Given-Default, A Case Study'. Working Paper.

Felsovalyi, A. and Hurt, L. (1998), 'Measuring Loss on Latin American Defaulted Bank Loans: A 27-Year Study of 27 Countries', *Journal of Lending and Credit Risk Management* .

Gentle, J. E. (2003), *Random Number Generation and Monte Carlo Methods*, 2 edn, New York: Springer.

Glasserman, P., Kang, W. and Shahabuddin, P. (2008), 'Fast Simulation of Multifactor Portfolio Credit Risk', *Operations Research* **56**(5), 1200–1217.

Glößner, P., Steinbauer, A. and Ivanova, V. (2006), 'Internal LGD Estimation in Practice', *WILMOTT Magazine* (1), 86–91.

Gordy, M. B. (1998), 'A Comparative Anatomy of Credit Risk Models'. Working Paper.

Gupton, G. M. and Stein, R. M. (2005), 'LossCalc V2: Dynamic Prediction of LGD', *Moody's Investors Service* pp. 1–44.

Hamerle, A., Knapp, M. and Wildenauer, N. (2006), Modelling Loss Given Default: A "Point in Time"-Approach, *in* B. Engelmann and R. Rauhmeier, eds, 'The Basel II Risk Parameters; Estimation, Validation, and Stress Testing', Springerlink, Berlin, pp. 127–142.

Hartmann-Wendels, T. and Honal, M. (2006), 'Do Economic Downturns Have an Impact on the Loss Given Default of Mobile Lease Contracts?'. Working Paper.

Huang, X. and Oosterlee, C. W. (2008), 'Generalized Beta Regression Models for Random Loss-Given-Default'. Working Paper.

Jacobs, M. and Karagozoglu, A. K. (2007), 'Understanding and Predicting Ultimate Loss-Given-Default on Bonds and Loans '. Working Paper.

Jobst, N. J. and Zenios, S. A. (2005), 'On the simulation of portfolios of interest rate and credit risk sensitive securities', *European Journal of Operational Research* **161**, 298–324.

Jokivuolle, E. and Peura, S. (2003), 'Incorporating Collateral Value Uncertainty in Loss Given Default Estimates and Loan-to-value Ratios', *European Financial Management* **9**(3), 299–314.

Kalkbrener, M., Lotter, H. and Overbeck, L. (2004), 'Sensible and Efficient Capital Allocation for Credit Portfolios'. Working Paper.

Kang, W. and Shahabuddin, P. (2005), 'Fast Simulation for Multifactor Portfolio Credit Risk in the t-Copula Model'. Working Paper, Proceedings of the 2005 Winter Simulation Conference.

Malervergne, Y. and Sornette, D. (2006), *Extreme Financial Risks - From Dependence to Risk Management*, Heidelberg: Springer.

Nelsen, R. B. (2006), *An Introduction to Copulas*, 2 edn, New York: Springer.

Peter, C. (2006), Estimating Loss Given Default - Experiences from Banking Practise, *in* B. Engelmann and R. Rauhmeier, eds, 'The Basel II Risk Parameters; Estimation, Validation, and Stress Testing', Springerlink, Berlin, pp. 143–175.

Querci, F. (2005), 'Loss Given Default on a medium-sized Italian bank's loans: an empirical exercise'. Working Paper.

Schuermann, T. (2005), What Do We Know About Loss Given Default?, *in* E. Altman, A. Resti and A. Sironi, eds, 'Recovery Risk, The Next Challenge in Credit Risk Management', London: Risk Books, pp. 3–24.

von Neumann, J. (1951), Various Techniques Used in Connection With Random Digits, *in* 'Applied Math Series', Vol. 12, Pergamon, New York, pp. 36–38.

Appendix 1

Table 6: Empirical modes of LGD densities

The table shows the range of modes of LGD densities through different financial contracts, different time spans and different businesses.

time span	left mode	right mode	description	source
1970-2003	$0.10 - 0.30$	$0.70 - 0.90$	Bonds & Loans	Schuermann (2005)
1982-1999	$0.05 - 0.15$	$1.00 - 1.10$	Commercial Loans	Araten et al. (2004)
1993-2004	$0.00 - 0.10$	$0.90 - 1.10$	Vehicle Leasing Contracts	Hartmann-Wendels and Honal (2006)
1993-2004	$0.00 - 0.20$	$0.80 - 1.00$	IT Leasing Contracts	Hartmann-Wendels and Honal (2006)
1995-2000	$0.00 - 0.05$	$0.95 - 1.00$	SME Loans	Dermine and de Carvalho (2005)
1980-2004	$0.00 - 0.10$	$0.90 - 1.00$	Loans	Querci (2005)
1970-1996	$0.00 - 0.15$	$0.95 - 1.00$	Commercial and Industrial Loans	Felsovalyi and Hurt (1998)
1985-2006	$0.00 - 0.05$	$0.95 - 1.00$	Bonds & Loans	Jacobs and Karagozoglu (2007)

Appendix 2

Since Υ_2 is assumed to be left-skewed, it follows that p is greater than q. We therefore rearrange Equation (3) for p and substitute the value in Equation (5).

$$\begin{aligned}\mathrm{Var}(\Upsilon_2) &= \frac{\frac{q\cdot \mathrm{E}(\Upsilon_2)}{1-\mathrm{E}(\Upsilon_2)}\cdot q}{\left(\frac{q\cdot \mathrm{E}(\Upsilon_2)}{1-\mathrm{E}(\Upsilon_2)}+q+1\right)\cdot\left(\frac{q\cdot \mathrm{E}(\Upsilon_2)}{1-\mathrm{E}(\Upsilon_2)}+q\right)^2} \\ &= \frac{\mathrm{E}(\Upsilon_2)\cdot(1-\mathrm{E}(\Upsilon_2))^2}{q+1-\mathrm{E}(\Upsilon_2)}\end{aligned}$$

Furthermore if q equals one we get the upper bound for $\mathrm{Var}(\Upsilon_2)$.

$$\mathrm{Var}(\Upsilon_2) < \frac{\mathrm{E}(\Upsilon_2)\cdot(1-\mathrm{E}(\Upsilon_2))^2}{2-\mathrm{E}(\Upsilon_2)}$$

Appendix 3

Let f be a function of the parameters p and q: $f(p,q)=\frac{p}{q}$. According to Equation (3) it holds that:

$$\mathrm{E}(\Upsilon)=\frac{p}{p+q}.$$

If we substitute p in Equation (3) with $p=f(p,q)\cdot q$ we get:

$$\mathrm{E}(\Upsilon)=\frac{q\cdot f(p,q)}{q\cdot f(p,q)+q}=\frac{f(p,q)}{f(p,q)+1}.$$

If $\mathrm{E}(\Upsilon)$ is given as a constant c the following holds:

$$c=\frac{f(p,q)}{f(p,q)+1}\Rightarrow f(p,q)=\frac{c}{1-c}.$$

Appendix 4

The variance of a beta distributed random variable Υ according to Equation (5) decreases for increasing parameters p and q given a constant ratio between these parameters. The constant ratio can be expressed as $\frac{p}{q}=c$ with c being some constant value. Thus we can

write the variance just depending on q and c as:

$$\begin{aligned} \text{Var}(\Upsilon) &= \frac{c \cdot q^2}{(c \cdot q + q + 1) \cdot (c \cdot q + q)^2} = \frac{c}{(c \cdot q + q + 1) \cdot (c+1)^2} \\ \Rightarrow \frac{\partial \text{Var}(\Upsilon)}{\partial q} &= -\frac{c}{(c \cdot q + q + 1)^2 \cdot (c+1)} < 0 \quad \forall q. \end{aligned}$$

Appendix 5

The general correlation matrix with varying interval bounds is given as:

$$\rho = \begin{pmatrix} 1 & [a,b] & [c,d] \\ & 1 & [e,f] \\ & & 1 \end{pmatrix}$$

Since the matrix has to be at least positive semidefinite, all determinants of the leading principle minors have to be non-negative. The determinant of the first leading principle minor of ρ equals 1, and the determinant of the second principal minor is non-negative for any non-negative correlation. The remaining constraints result in the following optimization problem given by:

$$\begin{aligned} \max z &= (b-a) + (d-c) + (f-e) \\ \text{s.t.} & \\ 0 &\leq 1 - a^2 - c^2 - e^2 + 2 \cdot a \cdot c \cdot e \\ 0 &\leq 1 - a^2 - c^2 - f^2 + 2 \cdot a \cdot c \cdot f \\ 0 &\leq 1 - a^2 - d^2 - e^2 + 2 \cdot a \cdot d \cdot e \\ 0 &\leq 1 - a^2 - d^2 - f^2 + 2 \cdot a \cdot d \cdot f \\ 0 &\leq 1 - b^2 - c^2 - e^2 + 2 \cdot b \cdot c \cdot e \\ 0 &\leq 1 - b^2 - c^2 - f^2 + 2 \cdot b \cdot c \cdot f \\ 0 &\leq 1 - b^2 - d^2 - e^2 + 2 \cdot b \cdot d \cdot e \\ 0 &\leq 1 - b^2 - d^2 - f^2 + 2 \cdot b \cdot d \cdot f \\ 0 &= a - e \\ 0 &= b - f \\ & 0.5 \leq a \leq 0.75,\ 0.75 \leq b \leq 1,\ 0 \leq c \leq 0.25, \\ & 0.25 \leq d \leq 0.5,\ 0.5 \leq e \leq 0.75,\ 0.75 \leq f \leq 1 \end{aligned}$$

Appendix 6

For calculating the double-stochastic matrix, the following linear program has to be solved:

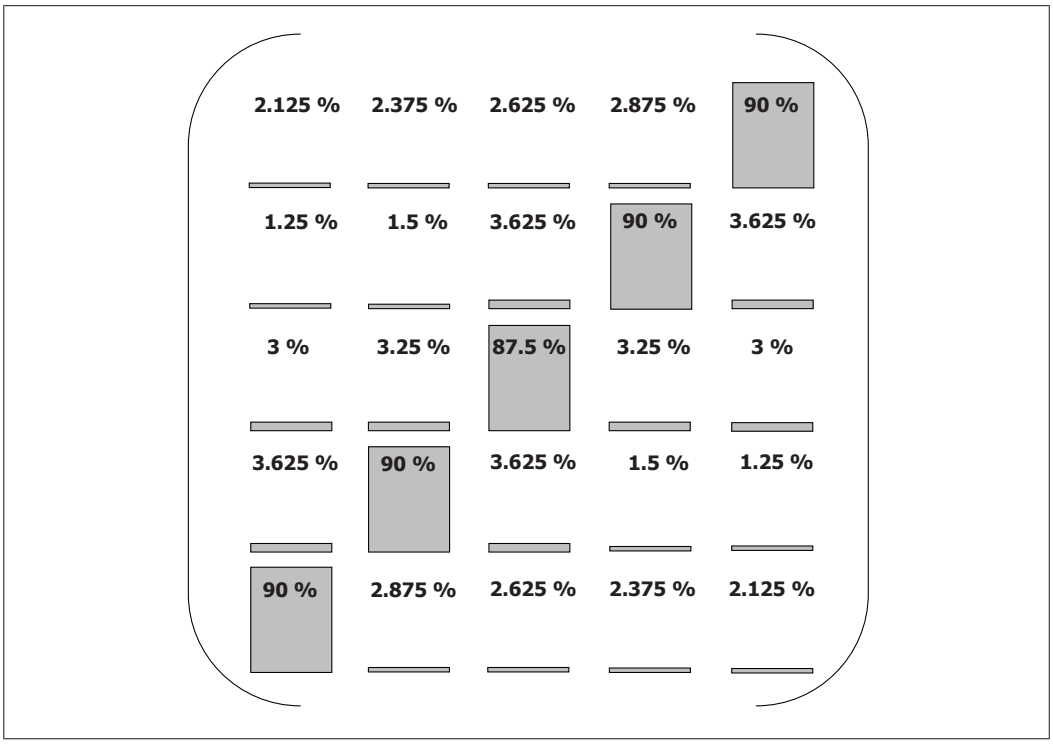
$$
\begin{aligned}
\max z = \quad & \sum_{i=1}^{5}\sum_{j=1}^{5} m_{i,j} \\
\text{s.t.} \quad & \\
1 = \quad & \sum_{j=1}^{5} m_{i,j} \;\forall\; i, \qquad 1 = \quad \sum_{i=1}^{5} m_{i,j} \;\forall\; j \\
0 = \quad & m_{1,5} - m_{5,1}, \qquad 0 = \quad m_{1,4} - m_{5,2} \\
0 = \quad & m_{1,3} - m_{5,3}, \qquad 0 = \quad m_{1,2} - m_{5,2} \\
0 = \quad & m_{2,1} - m_{4,5}, \qquad 0 = \quad m_{2,2} - m_{4,4} \\
0 = \quad & m_{2,3} - m_{4,3}, \qquad 0 = \quad m_{2,4} - m_{4,2} \\
0 = \quad & m_{3,1} - m_{3,5}, \qquad 0 = \quad m_{3,2} - m_{3,4} \\
0 = \quad & m_{2,3} - m_{2,5}, \qquad 0 \geq \quad 2 \cdot m_{3,2} - m_{3,3} \\
0 \geq \quad & 2 \cdot m_{1,1} - m_{1,2}, \quad 0 \geq \quad 2 \cdot m_{1,2} - m_{1,3} \\
0 \geq \quad & 2 \cdot m_{1,3} - m_{1,4}, \quad 0 \geq \quad 2 \cdot m_{1,4} - m_{1,5} \\
0 \geq \quad & 2 \cdot m_{2,1} - m_{2,2}, \quad 0 \geq \quad 2 \cdot m_{2,2} - m_{2,3} \\
0 \geq \quad & 2 \cdot m_{2,3} - m_{2,4}, \quad 0 \geq \quad 2 \cdot m_{3,1} - m_{3,2} \\
0.025 \leq \quad & m_{1,1} \\
0.025 \leq \quad & m_{2,1} \\
0.025 \leq \quad & m_{3,1}
\end{aligned}
$$

First the structure of the goal function is irrelevant due to the constraints structure. Furthermore the third up to the 12th constraint are to ensure the symmetry of M. The constraints 13 up to 21 ensure that the probability of being in one quintile is at least twice the probability of being in the next less probable quintile. The last three constraints ensure a minimal probability of choosing each quintile in the matching process of 2.5 %. Due to this constraint structure we ensure that there are some outliers possible but just to a reasonable extent.

Appendix 7

Figure 6: Matrix M for the matching process for the good data set

The figure shows the outcome of an adjusted linear program similar to the one in Appendix 6 presented as the Matrix M.



Appendix 8

Figure 7: Regression Tree for the exemplarily chosen portfolio and the good data set

The figure shows the regression tree for the exemplarily chosen portfolio of the good data set. The end nodes represent the estimated LGD for this specific branch.