Estimating LGD Correlation

Jiří Witzany

University of Economics, Prague[1]

Abstract: *The paper proposes a new method to estimate correlation of account level Basle II Loss Given Default (LGD). The correlation determines the probability distribution of portfolio level LGD in the context of a copula model which is used to stress the LGD parameter as well as to estimate the LGD discount rate and other parameters. Given historical LGD observations we apply the maximum likelihood method to find the best estimated correlation parameter. The method is applied and analyzed on a real large data set of unsecured retail account level LGDs and the corresponding monthly series of the average LGDs. The correlation estimate comes relatively close to the PD regulatory correlation. It is also tested for stability using the bootstrapping method and used in an efficient formula to estimate ex ante one-year stressed LGD, i.e. one-year LGD quantiles on any reasonable probability level.*

1 Introduction

The Basle II regulatory formula (see Basle, 2006) aims to provide a sufficiently robust estimate of unexpected losses on banking credit exposures that should be covered by the capital. The capital requirement (C) is set equal to the difference between the unexpected (UL) and expected credit loss (EL), calculated for each receivable as $C = UL\text{-}EL = (UDR\text{-}PD)\cdot LGD\cdot EAD$, where PD is the expected default rate, $UDR=UDR(PD)$ a specific regulatory function estimating unexpected default rate from the PD parameter, LGD the expected percentage loss conditional upon default, and EAD the expected exposure of the receivable at default.

[1] The research has been supported by the Czech Science Foundation grant no. 402/06/0890 “Credit Risk and Credit Derivatives” and by the grant no. 402/09/0732 “Market Risk and Financial Derivatives”

Keywords: credit risk, recovery rate, loss given default, correlation, regulatory capital
JEL: G21, G28, C14

The regulatory approach (BCBS, 2006 or CRD,2006) is very specific regarding unexpected default rate applying the Vasicek (1987) formula that is generally considered to be sufficiently robust. On the other hand the *LGD* parameter is specified very vaguely by the regulation to reflect downturn economic conditions but may be also calculated just as a long term default weighted average under relatively normal circumstances. This deficiency has been criticized by many practitioners and researchers.

It has been empirically shown in a series of papers by Altman et al. (2004), Gupton et al. (2000), Frye (2000b, 2003), Acharya et al. (2007), etc. that there is not only a significant systemic variation of recovery rates but moreover a negative correlation between frequencies of default and recovery rates, or equivalently a positive correlation between frequencies of default and losses given default. Consequently the regulatory formula may significantly underestimate the unexpected loss on the targeted confidence probability level (99.9%) and in the considered time horizon (one year). Some authors (see e.g. Frye, 2000ab, Dullmann and Trapp, 2004, Pykhtin, 2003, Tasche, 2004, or Witzany, 2009ab) have proposed alternative unexpected loss formulas incorporating the impact of recovery risk variation. The unexpected recovery risk is also important for determination of the recovery cash flows discount rate in line with the regulatory requirements. Witzany (2009c) proposes a methodology to estimate the discount rate and the unexpected recovery risk but the empirical study uses an expertly set correlation at the level of 10%.

The aim of this paper is to propose and test on real banking data an estimation methodology for the *LGD* correlation. Section 2 outlines the *LGD* asymptotic model and the corresponding *LGD* correlation estimation methodology. The empirical results are then described in Section 3.

2 The LGD Model and the Estimation Method

The model proposed in Witzany (2009bc) can be summarized as follows: We assume that account level identically distributed loss given default rate LGD_j are normalized (see also Gupton, 2005 or Kim, 2006) to $Y_j = N^{-1}(Q(LGD_j))$ where Q is the account level LGD cdf and N denotes the standardized normal cdf. We use the one-factor copula model for Y_j and LGD_j

$$(1)\qquad \begin{aligned} &Y_j = \sqrt{\rho}\cdot S + \sqrt{1-\rho}\cdot W_j, \text{ i.e.} \\ &LGD_j = Q^{-1}(N(\sqrt{\rho}\cdot S + \sqrt{1-\rho}\cdot W_j)) \end{aligned}$$

with independent standardized normal systematic factor S and account-specific idiosyncratic factor W_j. For a large portfolio of receivables that defaulted at the same time t and have been recovered during the same period the systematic factor could be kept fixed at S but the idiosyncratic factor varies over all possible values according to its distribution. Hence the asymptotic average portfolio loss rate conditional upon S can be approximated by

$$(2)\qquad LGD = H(S) = E\left[Q^{-1}(N(\sqrt{\rho}S + \sqrt{1-\rho}\cdot W)) \mid S\right] = \int_{-\infty}^{\infty} Q^{-1}(N(\sqrt{\rho}S + \sqrt{1-\rho}\cdot w))\,\phi(w)dw$$

where ϕ is the standardized normal pdf. Once we know the distribution function Q and the correlation ρ we also know the transformation function H and so the distribution of LGD transforming $S \sim N(0,1)$ by H. Note that the function H is increasing with positive first derivative for regular distributions Q.

We will use the Gaussian copula model to estimate the correlation ρ given a data set of observed defaulted accounts A. We assume that for each $a \in A$ we are given the month of default (alternatively quarter, year, or another time unit) as an ordinal $t(a) \in \{t_0, t_0+1, ..., t_1\}$ and the realized loss given default $lgd(a)$ calculated as 1 minus the discounted recovery cash flows (see Witzany, 2009c). The realized values are expected to be distributed in the interval $[0,1]$ but we admit also values less than 0 and larger than 1. Moreover we assume that there is an unobserved time series of the systematic factors $s(t)$ for $t \in \{t_0, t_0+1, ..., t_1\}$ and the independent idiosyncratic factors $w(a)$ corresponding to (1) for every $a \in A$. We admit certain autocorrelation in the time series $s(t)$, which would not be surprising, but we assume that the series is weakly independent, i.e. that $s(t)$ and $s(t+h)$ are almost independent for any sufficiently large h. This assumption holds for the models like AR(1) or AR(n) that we will use. Consequently we may apply the law of large numbers, in particular we may assume that the empirical distribution of $\{\sqrt{\rho}s(t(a)) + \sqrt{1-\rho}w(a) \mid a \in A\}$ approximates well the standardized normal distribution for a large enough dataset A.

The first task is to estimate the cdf Q for the account level LGDs. Based on the assumptions above the empirical distribution $\{lgd(a) \mid a \in A\}$ approximates well the theoretical distribution Q. We will get Q in two ways:

a. As a fitted parametric Beta distribution.

b. As a normal kernel smoothed distribution obtained from the empirical distribution.

The beta distribution determined by its minimum A, maximum B (normally 0 and 1), and coefficients α and β is recommended by many authors (see e.g. Schuermann, 2004 or Gupton, 2005). If all the observations were in the interval $(0,1)$ then α and β could be fitted using the maximum likelihood method. However as we will see in Section 3 there could be outliers with very low (negative) and very high (above 1) LGDs. If we set the Beta distribution parameters A and B approximately at the observed minimal and maximal value then the fitted distribution may appear unrealistically flat. On the other hand we cannot use maximum likelihood if any of the observed values falls outside of the interval (A,B). Hence we will set A and B at appropriate quantiles of the empirical distribution (e.g.1% and 99%) and to fit the parameters α and β to the first two moments, i.e. to the sample mean μ and standard deviation σ:

$$\alpha = \tilde{\mu}\left(\frac{\tilde{\mu}(1-\tilde{\mu})}{\tilde{\sigma}^2} - 1\right), \beta = (1-\tilde{\mu})\left(\frac{\tilde{\mu}(1-\tilde{\mu})}{\tilde{\sigma}^2} - 1\right), \text{ where } \tilde{\mu} = \frac{\mu - A}{B - A}, \tilde{\sigma} = \frac{\sigma}{B - A}.$$

Once we specify the account level LGD distribution Q we may proceed to estimation of ρ based on (2) and the maximum likelihood method. In addition we need to assume that the observed vintages $A(t) = \{a \in A \mid t(a) = t\}$ are sufficiently large so that the observed average values $lgd(t) = \frac{1}{|A(t)|}\sum_{a \in A(t)} lgd(a)$ follow (approximately) the asymptotic distribution determined by (2).

To express the likelihood function let us start with the first observed vintage level loss given default $l_0 = lgd(t_0)$. The likelihood of the observation is given by the corresponding density function value of the random variable $LGD = H(S)$ with $S \sim N(0,1)$. Let $s_0 = H^{-1}(l_0)$ then using the chain rule the likelihood of the single observation is $L(l_0) = \frac{\phi(s_0)}{H'(s_0)}$. The same holds for the second month observation $l_1 = lgd(t_0 + 1)$ but for the joint likelihood calculation we have to take a possible autocorrelation into account. Let us assume that the systematic factors $s_i = H^{-1}(l_i), l_i = lgd(t_0 + i)$ follow the AR(1) process, i.e. $s_i = c_1 \cdot s_{i-1} + c_2 \cdot u_i$ where $u_i \sim N(0,1)$ are iid and $c_1^2 + c_2^2 = 1$. The coefficient c_1 may be estimated as the time series autocorrelation and $c_2 = \sqrt{1-c_1^2}$. To express $L(l_0, l_1) = L(l_0) \cdot L(l_1 \mid l_0)$ we need to use $L(l_1 \mid l_0) = \dfrac{\phi(u_1)}{c_2 \cdot H'(s_1)}$ again applying the chain rule on $H(c_1 \cdot s_0 + c_2 \cdot U_1)$ where $U_1 \sim N(0,1)$ and s_0 is fixed. Since $L(l_i \mid l_0, ..., l_{i-1}) = L(l_i \mid l_{i-1})$ we may continue for $i = 1, ..., n = t_1 - t_0$ to get

(3) $$L\left(\langle l_i \rangle_{i=0}^n\right) = L(l_0) \cdot \Pi_{i=1}^n L(l_i \mid l_{i-1}) = \frac{\phi(s_0)}{H'(s_0)} \cdot \Pi_{i=1}^n \frac{\phi(u_i)}{c_2 H'(u_i)}, \text{where } u_i = \frac{s_i - c_1 s_{i-1}}{c_2}.$$

The log-likelihood function $\log L\left(\langle l_i \rangle_{i=0}^n\right)$ now may be maximized with respect to the correlation parameter ρ that enters into Q. Since we admit an arbitrary (smoothed) empirical distribution Q the integral (2) and the inverse function must be evaluated empirically and we need to use a numerically efficient maximization algorithm (implemented e.g. in Matlab). To get a standard error estimation of the parameter $\hat{\rho}$ we can use the bootstrapping technique on the dataset A making sure that size of the bootstrapped vintages remains unchanged.

The remaining theoretical question is how to use the correlation to estimate one-year horizon unexpected LGD in case the estimation is based on shorter time interval, e.g. monthly series. The random variable we want to model can be in that case expressed as

(4) $$LGD_{1Y} = \frac{1}{12} \sum_{i=1}^{12} H(S_{n+i})$$

provided the number of defaults in individual months is stable and $S_{n+1}, ..., S_{n+12}$ are the next twelve months unknown systematic factors. One conservative approach is to set all the factors equal to a quantile $N^{-1}(x)$, e.g. $N^{-1}(0.95)$, but the resulting stressed LGD is clearly larger than the 95% quantile of (4) since we are disregarding the partial independence of S_{n+i}. A fully precise approach would be to estimate the quantile empirically (e.g. using Monte Carlo simulation) based on the relationship $S_{n+i} = c_1 S_{n+i-1} + c_2 U_{n+i}$ with iid $U_{n+i} \sim N(0,1)$ and $i = 1, ..., n$. We will observe that the function H is "almost" linear for reasonable values of the systematic factor. Consequently a practical approach standing in terms of precision somewhere in between the two approaches described above would be to take the function H out of the sum on the right hand side of (4) and calculate the desired quantile of

(5) $$\frac{1}{12}\sum_{i=1}^{12} S_{n+i} = c_1^{12} S_n + \sum_{i=1}^{12} d_i U_{n+i}, \text{ where } d_i = \frac{c_2}{12}\sum_{k=i}^{12} c_1^{k-i}.$$

The standard deviation σ_{1Y} of the sum on the right hand side (5) can be easily calculated, $\sigma_{1Y} = \sqrt{\sum_{i=1}^{12} d_i^2}$, since U_{n+i} are independent $N(0,1)$. To calculate the quantile the first term on the right hand side can be neglected and in fact it should be set equal to 0 as the forward looking *LGD* estimations need to be based on a long term *LGD* average, i.e. zero systematic factor. Consequently we may estimate the unexpected *LGD* on the probability level x simply as $H^{-1}(\sigma_{1Y} \cdot N^{-1}(x))$.

3 Empirical Results

We have obtained an LGD data set of 4 000 defaulted unsecured retail loans from a large Czech retail bank. The loans defaulted in a recent period of 57 months ($t_0 = 1$ and $t_1 = 57$). The data set contains account level information on net discounted monthly recovery cash flows as well as some basic application and behavior explanatory variables. Ultimate recovery is achieved by a sale of receivable, write-off or after 36 months. Since the data have been observed shortly after the end of the observation period many of the recoveries remain uncompleted, for accounts defaulting in month t there are in fact at most $58 - t$ monthly recoveries. This is a typical situation which needs to be resolved somehow in practice. Banks first of all do not have sufficiently long historical data; secondly the recent yet incomplete data contain important information regarding recent trends. The possible extrapolation techniques including survival time analysis methods are studied in Charamza, Rychnovsky, and Witzany (2009). For the sake of our study we will use logistic regression based extrapolation of the ultimate recovery rates $rr36(a)$ and work with $lgd(a) = 1 - rr36(a)$. The recoveries and LGDs are relative to the exposures at default and the averages are default (not exposure) weighted. At the end of the section we will also discuss some alternatives to this approach. The histogram of the observed LGDs is shown on Figure 1 and the descriptive statistics in Table 1.

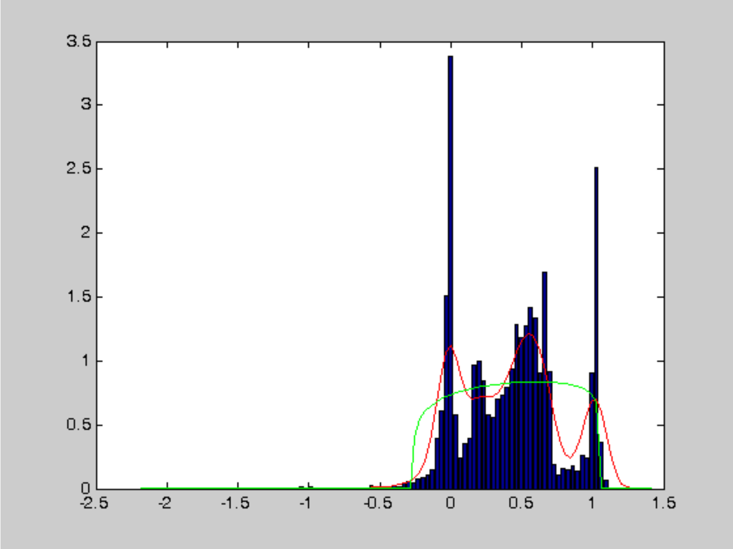

Figure 1. The histogram of observed LGDs, fitted beta, and kernel smoothed distributions.

Num	4000
Max	1.2726
Min	-2.0301
Mean	0.4173
Median	0.4467
Range	3.3028
Std	0.3609

Table 1. Descriptive statistics of the LGD data set

The histogram shows that the real data rather deviate from our expectation of the LGD distribution, i.e. a beta distribution on the interval $[0,1]$. The high values (up to 127%) correspond to situations when there are relatively significant recovery costs but no actual recovery amounts collected. On the other hand the negative observed LGDs (down to -203%) are realized when the debtors decide to pay all the obligation including late fees and sanction interest with discounted total significantly exceeding the initial exposure at default. To fit the beta distribution we have used the 1% empirical quantile $A = -0.26$ and the 99% quantile $B = 1.046$. The beta distribution Q_b fitted to the first two moments with $\alpha = 11.12$ and $\beta = 3.88$ is shown in Figure 1. It appears that better result should be obtained with the normal kernel smoothed empirical distribution Q_k calculated in the Matlab application using the *ksdensity* function (see Figure 2).

Next we need to analyze the time series of the average monthly $lgd(t)$ shown on Figure 2. The figure as well as the descriptive statistics (Table 2) shows that the variation of monthly portfolio level LGDs is much smaller than the variation of account level LGDs. The number of accounts in monthly vintages ranges from 39 to 108 which is not optimal but can be still considered as sufficient with respect to the asymptotic model.

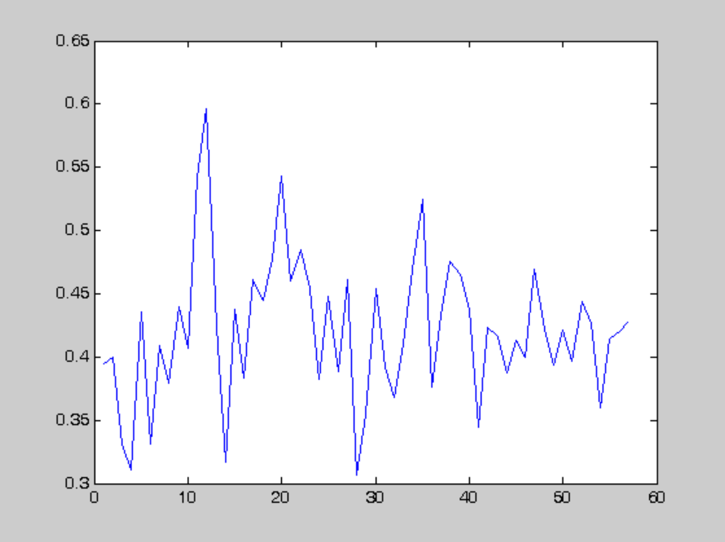

Figure 2. The vintage LGD time series.

Num	57
Max	0.5963
Min	0.3065
Mean	0.4209
Median	0.4211
Range	0.2898
Std	0.0567

Table 2. Descriptive statistics of the vintage LGD time series

Given the account level distributions we may for any given correlation ρ evaluate the transformation function H according to (2). Figure 3 shows the function for different correlation values, alternatively for the beta and the empirical distribution.

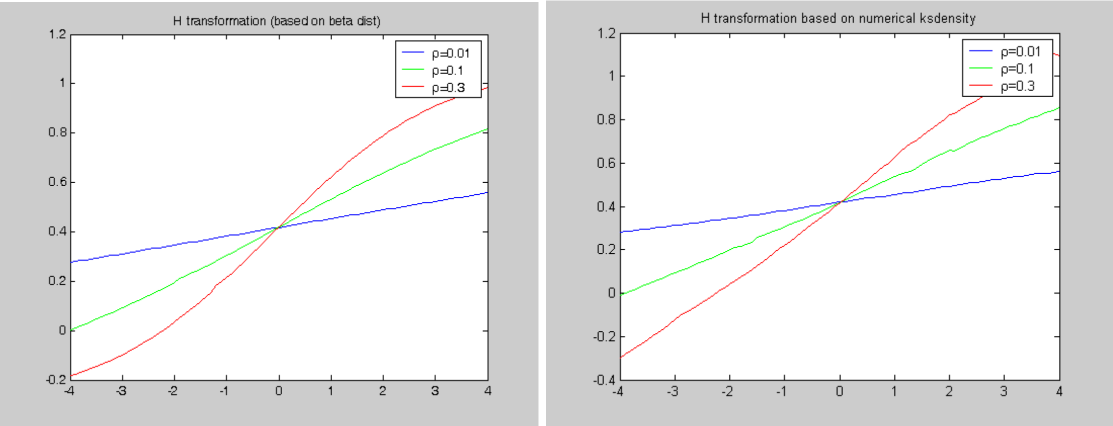

Figure 3. The H transformation for the beta and the empirical distribution and different correlation values

Note that since the function $H(S)$ is "almost" linear, at least for systematic factor S values in the interval $[-2,2]$, and as S is standardized normal the variable $H(S)$ is also "almost" normal. This appears to be surprisingly even more true for the kernel smoothed empirical distribution Q_k. Since the standard deviation of $H(S)$ equals approximately to the slope of H we are essentially seeking the correlation ρ such that the slope of the corresponding transformation $H = H_\rho$ equals approximately to the observed standard deviation of the observed monthly LGDs. Of course this not exactly how the maximum likelihood runs but we may check the relationship for consistency when the calculation is done.

Before we run the maximum likelihood estimation we may look on autocorrelation of the LGD time series (see Figure 3). The autocorrelation for lags larger than 1 do not appear significant, hence we will use the AR(1) model for the systematic factors as described in Section 2.

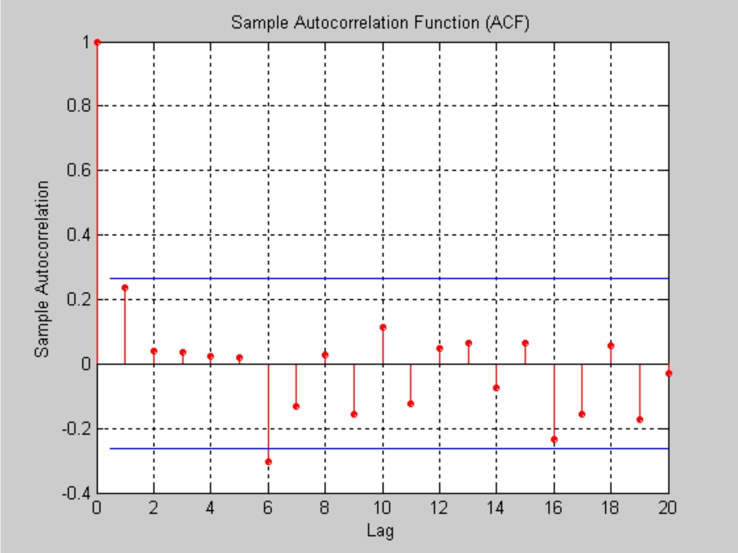

Figure 3. Autocorrelation of the monthly vintage LGD time series

Finally we ran the maximum likelihood estimation and bootstrap the sample 100 times for both types of distributions. We get the estimations $\hat{\rho}_b = 4.48\%$ with s.e.=0.65% based on the beta distribution and $\hat{\rho}_k = 3.9\%$ with s.e.=0.8%. To apply the consistency check mentioned above we may estimate the slope of H_ρ e.g. in the case of the empirical distribution by $H'_\rho(0) = 6.3\%$ which is indeed close to the standard deviation of the LGD time series (Table 2). Finally we conclude that the estimation technique is relatively stable even without significant dependence on the shape of the account level LGD distribution.

Let us calculate the 95% probability level stressed one-year average LGD based on the model and empirical distribution based correlation $\hat{\rho}_k = 3.9\%$. The simplest approach is to set $ULGD_1 = H(1.65) = 53.76\%$ which is 12% more compared to the long term LGD average (41.73%). Secondly let us use the formula (5) to calculate the standard deviation σ_{1Y} of the average forward looking twelve months systematic factor. The lag 1 autocorrelation of the historical systematic factors turns out to be $c_1 = 0.2353$. Using (5)we get that $\sigma_{1Y} = 0.3583$ and the second estimate $ULGD_2 = H(0.3583 \cdot 1.65) = 45.83\%$ turns out to be much lower than the first estimate. Finally we have simulated possible LGD_{1Y} values based on (4) and obtained a third estimate of the 95% quantile $ULGD_3 = 45.7\%$ that is as expected very close to the second simulation based estimate.

Remark: We have pointed out at the beginning of this section that the ultimate LGDs have been extrapolated from incomplete data. If we look only on accounts with completed 36 months recovery rates, i.e. on accounts that defaulted in month 1 to 21=57-36, then we obtain a slightly different distribution (see Figure 4).

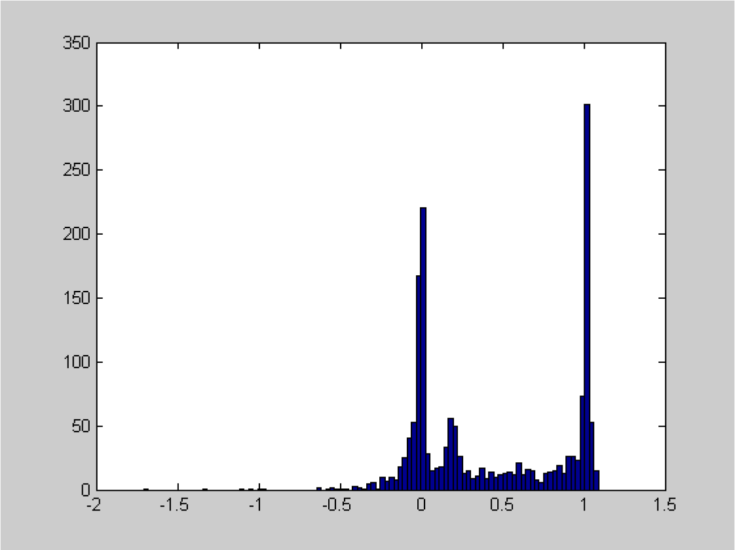

Figure 4. The histogram of completed LGDs after 36 months

Notice that the distribution differs from the one on Figure 1 where there is a significant hump in the middle. This is probably caused by the logistic regression based extrapolation which tends to the average values. The reason why we did not limit ourselves only to those data is that the time series becomes too short (only 21 months with 1651 accounts) and the estimation becomes unreliable. However running the beta distribution density based estimation we obtained $\hat{\rho} = 4.12\%$ with s.e.=1%. We have also investigate the possibility using just partial, e.g. 12 month, but realized recoveries. The correlation estimate came out slightly higher but the basic account level distribution appears to have a sinificantly different shape (Figure 4). It seems that large repayments causing LGD to be close to zero happen mostly in later phases of the recovery process. Consequently we also had to reject this alternative approach.

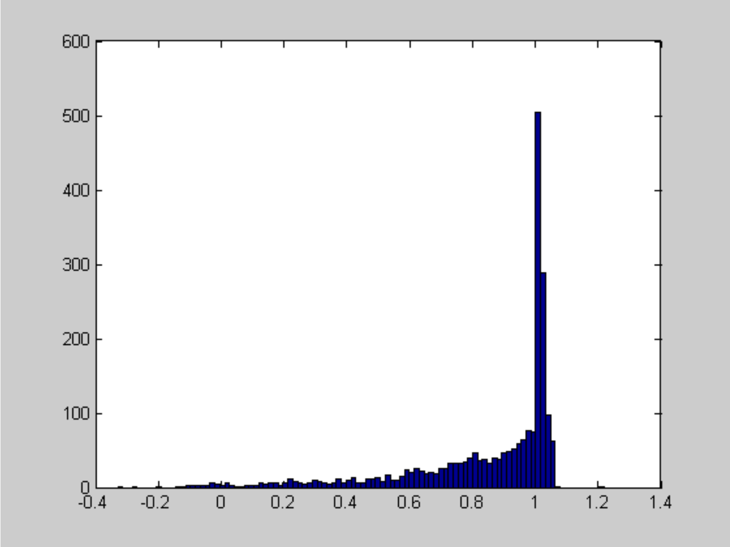

Figure 5. Distribution of LGD based on 12 months recoveries , i.e. $1 - rr12$

4 Conclusions

The proposed LGD correlation methodology applied to a relatively large sample of defaulted unsecured retail loans led to a relatively stable correlation estimates at about $\hat{\rho} = 3.9\%$. The result is surprisingly close to the regulatory correlation (see BCBS, 2006) entering the unexpected probability of default formula that is 3% for revolving loans and

$$\rho_{reg} = 0.03 \frac{1 - e^{-35p}}{1 - e^{-35}} + 0.16 \frac{e^{-35p} - e^{-35}}{1 - e^{-35}}$$

for "other" retail loans (other than mortgages and revolving loans) where p is the probability of default. Disregarding the peculiarity of the formula if we use the default probability of 4% indicated by the bank we get $\rho_{reg} = 6.21\%$.

We have also proposed a simplified yet efficient estimation of the stressed 1 year LGD based on monthly LGD series correlation. It can be used to verify that the slightly higher regulatory correlation compared to our estimate $\hat{\rho} = 3.9\%$ nevertheless leads to a significantly higher modeled unexpected LGD.

The correlation estimation procedure should be ideally applied to ultimate realized recoveries. This is in practice almost impossible as the recoveries of recent defaults usually remain uncompleted. Further research should be made regarding the impact of various extrapolation methods. Last but not least a research on PD x LGD correlation in the context of the proposed methodology should follow.

Literature

Acharya, Viral, V., S. Bharath and A. Srinivasan (2007), “Does Industry-wide Distress Affect Defaulted Firms? – Evidence from Creditor Recoveries,” Journal of Financial Economics 85(3):787–821.

Altman E., Resti A., Sironi A. (2004), “Default Recovery Rates in Credit Risk Modelling: A Review of the Literature and Empirical Evidence”, Economic Notes by Banca dei Paschi di Siena SpA, vol.33, no. 2-2004, pp. 183-208

BCBS (2005), Basel Committee on Banking Supervision, Guidance on Paragraph 468 of the Framework Document.

BCBS (2006), Basel Committee on Banking Supervision, **“**International Convergence of Capital Measurement and Capital Standards, A Revised Framework – Comprehensive Version”, Bank for International Settlements

Charamza P., Rychnovsky M., and Witzany J (2009), “Applying Cox Regression to LGD estimations”, Working Paper

CRD (2006), Directive 2006/48/EC of the European Parliament and the Council of 14 June 2006 relating to the taking up and pursuit of the business of credit institutions (recast).

Dullman, K. and M. Trapp (2004), “Systematic Risk in Recovery Rates – An Empirical Analysis of U.S. Corporate Credit Exposures”, EFWA Basel Paper.

Frye, J. (2000a), “Collateral Damage”, RISK 13(4), 91–94.

Frye, J. (2000b), “Depressing recoveries”, RISK 13(11), 106–111.

Frye, J. (2003), “A false sense of security, RISK 16(8), 63–67.

G. Gupton, D. Gates, and L. Carty (2000), “Bank loan losses given default”, Moody’s Global Credit Research, Special Comment.

Gupton G.M. (2005), „Advancing Loss Given Default Prediction Models: How the Quiet Have Quickened”, Economic Notes by Banca dei Paschi di Siena SpA, vol.34, no. 2-2005, pp. 185-230

Kim J., Kim H. (2006), „Loss Given Default Modelling under the Asymptotic Single Risk Factor Assumption”, Working Paper, MPRA

Pykhtin, M. (2003), “Unexpected recovery risk”, Risk, Vol 16, No 8. pp. 74-78.

Schuermann T. (2004), “What Do We Know About Loss Given Default”, Credit Risk Models and Management, 2nd Edition, London, Risk Books

Tasche, Dirk. (2004), “The single risk factor approach to capital charges in case of correlated loss given default rates”, Working paper, Deutsche Bundesbank, February 2004.

Vasicek O. (1987), “Probability of Loss on a Loan Portfolio,” KMV Working Paper

Witzany J. (2009a), “Basle II Capital Requirements Sensitivity to the Definition of Default” Icfai University Journal of Financial Risk Management, Vol. VI, No. 1, pp. 55-75, March

Witzany J. (2009b), “Loss, Default, and Loss Given Default Modeling”, IES Working Paper No. 9/2009

Witzany J. (2009c), “Unexpected Recovery Risk and LGD Discount Rate Determination”, European Finance and Accounting Journal, No. 1/2009