A Two-Factor Model for PD and LGD Correlation

Jiří Witzany

University of Economics, Prague[1]

Abstract: *The paper proposes a two-factor model to capture retail portfolio probability of default (PD) and loss given default (LGD) parameters and in particular their mutual correlation. We argue that the standard one-factor models standing behind the Basel II formula and used by a number of studies cannot capture well the correlation between PD and LGD on a large (asymptotic) portfolio. Parameters of the proposed model are estimated using the Markov Chain Monte Carlo (MCMC) method on a sample of real banking data. The results confirm positive stand-alone PD and LGD correlations and indicate a positive mutual PD x LGD correlation. The estimated Bayesian MCMC distributions of the parameters show that the stand alone correlations are strongly significant with a lower significance of the mutual correlation probably due to a too short observed time period.*

1. Introduction

The credit risk Basle II capital requirement (*C*) is set equal to the difference between the unexpected (*UL*) and expected credit loss (*EL*), calculated on the account level as $C = UL\text{-}EL = (UDR\text{-}PD)\cdot LGD\cdot EAD$, where *PD* is a bank's estimate of the expected default rate, $UDR=UDR(PD)$ a specific regulatory function estimating unexpected default rate from the *PD* parameter, *LGD* a bank's estimate the expected percentage loss conditional upon default, and *EAD* an estimate of the expected exposure of the loan at default. The regulatory approach (BCBS, 2006 or CRD, 2006) applies the Vasicek (1987) formula to get the unexpected default rate that is based on a single factor asymptotic model. The default of each obligor is assumed to depend on a systematic factor and on an idiosyncratic factor. The idiosyncratic factors are assumed to be diversified away in a large portfolio and to obtain a quantile of the overall frequency of default we just need the quantile of the systematic factor transformed by an appropriate formula. On the other hand, the *LGD* parameter is required very vaguely by the

[1] The research has been supported by the Czech Science Foundation grant no. 402/09/0380 "Correlation in Financial Risk Management" and by the grant no. 402/09/0732 "Market Risk and Financial Derivatives"

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

regulation to reflect downturn economic conditions (BCBS, 2005) and may be simply calculated just as a long term default weighted average under relatively normal circumstances. The regulation where LGD is not sufficiently stressed and where the formula does not reflect a possible correlation between the rate defaults and the level of losses given default has been criticized by many practitioners and researchers (see Altman, 2004, Schuermann, 2004).

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 could significantly underestimate the unexpected loss on the targeted confidence probability level (99.9%) and in the considered time horizon (one year). Nevertheless, the empirical studies were done only on bond default data and, as far as we know, there is no study confirming positive correlation on retail data. Some authors (see e.g. Frye, 2000ab, Pykhtin, 2003, Dullmann and Trapp, 2004, Tasche, 2004,Gupton, 2005, Kim, 2006, or Witzany, 2009ab) have proposed alternative unexpected loss formulas incorporating the impact of recovery risk variation.

The empirical studies either estimate just the sample correlation given observed PD and LGD time series or use a single systematic factor model fitted to empirical data (Frye, 2000b). The Frye model assumes that LGD depends on the same systematic factor as PD and on an idiosyncratic factor independent on the PD idiosyncratic factors. We argue in Section 2 that this model can be reasonably fitted to empirical data only if the number of defaults is low. If the number of defaults is large then the LGD idiosyncratic factors diversify away and we end up in the situation when the large portfolio PD and LGD depend only on the one systematic factor. Consequently the correlation will be close to one whatever the estimated parameters of the transformation functions are. Hence it is natural to extend the model with a second systematic factor that is assumed to impact the LGD itself. The two-factor model has not been used in the literature to our knowledge. The model and the proposed MCMC estimation procedure are described in Section 2. In Section 3 we use default and recovery data on a retail portfolio obtained from a large Czech bank to test the proposed methodology. The results are summarized and final remarks are made in Section 4.

2. One and Two Factor Models

We are firstly going to give an overview of the single systematic factor models of Frye (2000a, 2000b), Pykhtin (2003), Tasche (2004), and others that have been in a general form described in Witzany (2009b). Let us consider a (percentage) loss function of a given receivable in a time horizon. We assume that it is an increasing function of one systemic factor X and of a vector $\vec{\zeta}$ of idiosyncratic factors $L = L(X, \vec{\zeta})$. The single systematic factor X captures macroeconomic or other systemic influences that may develop in time while the vector of idiosyncratic factors $\vec{\zeta}$ reflects specificities of each individual obligor in a portfolio. Hence the impact of $\vec{\zeta}$ is diversified away in a large (asymptotic) portfolio while X remains as the only time dependent risk factor. The future loss on a large portfolio can be modeled as $E[L \mid X]$ (see Gordy, 2003 for details). Since we assume that L is increasing in X the problem to find quantiles of $E[L \mid X]$ reduces to calculation of the quantiles of X. If x is the desired (e.g. for 99.9%) quantile of X then $UL = E[L \mid X = x]$. This is a clear advantage of the single factor approach compared to a multi-factor approach where we work with a vector $\vec{X}$ of systemic factors instead of one factor X and the determination of quantiles of $E[L \mid \vec{X}]$ becomes more complex. It is natural to define the event of default by the condition $L > 0$ and to decompose the unexpected loss into two parts corresponding to the default rate and the loss given default:

$$E[L \mid X = x] = E[L \mid L > 0, X = x] \cdot P[L > 0 \mid X = x].$$

Tasche, Frye, and Pykhtin One-factor Models

The simplest version of a single-factor model is probably the model proposed by the Tasche (2004). The loss function $L = L(X, \zeta)$ is driven by one standard-normally distributed factor $Y = \sqrt{\rho}X + \sqrt{1-\rho}\zeta$ where X and ζ are independent standard-normally distributed, and ρ is the systematic factor loading, i.e. the correlation between $Y(a)$ and $Y(b)$ for different receivables a and b. If L is assumed to have a cumulative probability distribution function $F_L : [0,1] \to [0,1]$ then we may express the loss function in the form

$$L(X, \zeta) = F_L^*(\Phi(\sqrt{\rho}X + \sqrt{1-\rho}\zeta)) \text{ or just } L(Y) = F_L^*(\Phi(Y))$$

where $F_L^*(z) = \inf\{l : F_L(l) \geq z\}$ is the generalized inverse of F_L.

A more natural model has been proposed by Frye (2000a, 2000b). It may be described as follows. Let

$$Y_1 = \sqrt{\rho_1} X + \sqrt{1-\rho_1}\zeta_1 \text{and}$$

(1) $$Y_2 = \sqrt{\rho_2} X + \sqrt{1-\rho_2}\zeta_2$$

be two standard-normally distributed factors with one systematic and two independent idiosyncratic factors. The systematic factor loadings (i.e. stand alone PD and LGD correlations) ρ_1 and ρ_2 may be in general different. The first factor Y_1 drives the default in the model while the second factor Y_2 is assumed to drive losses in case of default. I.e. there is a default threshold y_D and a nonnegative non-decreasing function G so that the loss function can be expresses as:

(2) $$L(X,\zeta_1,\zeta_2) = \begin{cases} 0 \text{ if } Y_1 \le y_D, \\ G(Y_2) \text{ otherwise.} \end{cases}$$

If F_G is the distribution function of the random variable $G(Y_2)$ then the loss given default may be again expressed as $LGD(Y_2) = G(Y_2) = F_G^*(\Phi(Y_2))$.

The Pykhtin (2003) model in a sense unifies the two models. Let Y_1 be the driver of default as in the Frye model, i.e.

$$Y_1 = \sqrt{\rho_1} X + \sqrt{1-\rho_1}\zeta_1.$$

On the other hand let

(3) $$Y_2 = \sqrt{\rho_2} X + \sqrt{1-\rho_2}(\sqrt{\omega}\zeta_1 + \sqrt{1-\omega}\zeta_2)$$

be the driver of loss given default incorporating not only the systemic factor X and the idiosyncratic factor ζ_1 shared with Y_1, but also a new idiosyncratic factor ζ_2. Here ρ_1 and ρ_2 are the systematic factor loadings while ω determines the impact of the specific default factor ζ_1 on the LGD process. The loss function $L(X,\zeta_1,\zeta_2)$ is expressed by (2) as in the Frye's model. The approach enables us to model the fact that the loss in case of obligor's default is determined not only by the value of the assets and the obligor's specific financial situation at the time of default but also by a workout/bankruptcy specific process.

The three single-factor models can be used to model separately the PD and LGD, or the account level correlation between the event of default and subsequent loss given default. The models are however inappropriate to model correlation of a large (asymptotic) portfolio probability of default $PD = P[L > 0 | X] = g(X)$ and the loss given default $LGD = E[L | L > 0, X] = h(X)$, depending on the single systematic factor X. Consequently the models are not also appropriate to model the distribution of overall portfolio losses as $g(X) \cdot h(X)$. Note that if the functions g and h of one random variable X can be reasonably well approximated by linear functions then the model correlation between PD and LGD is obviously close to 1 whatever are the slopes of the two functions, i.e. whatever are the sensitivities to the systematic factor X estimated from given data.

Let us consider for example the Frye model. The first function $g(x)$ is just the Vasicek's formula employed by Basel II (see Witzany, 2009a):

$$(4) \qquad g(x) = P[\sqrt{\rho_1}x + \sqrt{1-\rho_1}\zeta_1 > y_D \,|\, x] = \Phi\left(\frac{\sqrt{\rho_1}x - y_D}{\sqrt{1-\rho_1}}\right),$$

where $y_D = -\Phi^{-1}(PD_0)$ is given by the overall expected probability of default PD_0 and Φ denotes as usual the cumulative standard normal distribution. The parameter PD_0 may be obtained as an average of a PD time series while the correlation parameter ρ_1 is usually estimated maximizing the likelihood given the observations. Figure 1 indicates that the functions can be reasonably approximated by a linear function around the origin for different pairs of PD_0 and ρ_1.

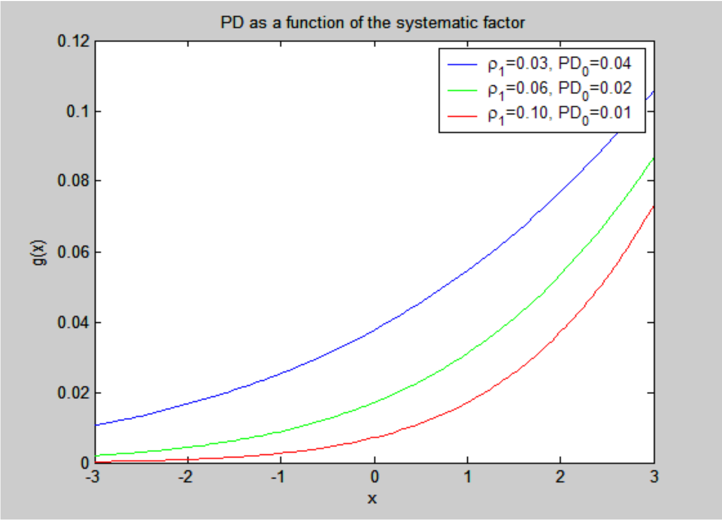

Figure 1. Unexpected probability of default (4) as a function of the systematic factor

To express the function

(5) $$h(x)=E\left[G\left(\sqrt{\rho_2}x+\sqrt{1-\rho_2}\zeta_2\right)\mid x\right]=\int_{-\infty}^{\infty}G(\sqrt{\rho_2}x+\sqrt{1-\rho_2}\cdot w))\,\phi(w)dw$$

we need to specify $G=Q^{-1}\circ\Phi$, i.e. the account level cumulative LGD distribution function Q. It is shown in Witzany (2009d), where the correlation ρ_2 is estimated from a time series of observed average LGDs that defaulted at given time (e.g. on monthly basis), that the function is in practice also near linear. Consequently the implied correlation between $g(X)$ and $h(X)$ will be indeed in general close to one. Specifically, the correlation can be estimated using numerical integration or simply with a Monte Carlo simulation generating a large number of pairs of $g(X)$ and $h(X)$ drawing X from $N(0,1)$ and calculating the sample correlation. For example when we have used the parameters $\rho_1=0.03, PD_0=0.04$ for g and the empirical LGD distribution Q as described in the next section with $\rho_2=0.03$ we have obtained 0.9806 model correlation between the modeled $PD=g(X)$ and $LGD=h(X)$ (based on 100,000 Monte Carlo scenarios). The parameters used are close to the empirical estimations, but the implied correlation is far from the observed sample correlation between the original PD and LGD series that is somewhere between 0.1 and 0.2.

The problem is neither solved in the Tasche nor in the Pykhtin model since the asymptotic portfolio PDs and LGDs are still functions of one systematic factors, although the models capture better the correlation between the event of default and account level LGD. It is interesting to note that Frye (2000b) empirically fits the data on defaults and recoveries from a Moody’s bond database using the single factor model. He uses the asymptotic formula (4) to fit the PD correlation and to imply the systematic factors and then fits the non-asymptotic LGD model with an assumption of recovery rates normality. The limited number of averaged idiosyncratic factors creates a second “systematic” factor allowing to estimate the remaining coefficients. The model however does not yield a reasonable estimate of the mutual PD and LGD correlation and becomes inconsistent as the number of defaults in the individual observed periods becomes large. In fact the portfolio PD/LGD correlation implied by the model depends on the number of defaults.

Proposed Two-Factor Model

The conclusion is that we need to propose an extended model and the natural choice we plan to investigate is the model defined by the equations:

$$(6)\qquad \begin{aligned} &Y_1 = \sqrt{\rho_1} X_1 + \sqrt{1-\rho_1}\zeta_1 \text{ and} \\ &Y_2 = \sqrt{\rho_2}\left(\omega X_1 + \sqrt{1-\omega^2} X_2\right) + \sqrt{1-\rho_2}\zeta_2 \end{aligned},$$

where X_1 and X_2 are two independent standardized normal variables. As above all the factors $X_1, X_2, \zeta_1, \zeta_2$ are iid N(0,1). The parameter ρ_1 is again the loading (stand alone PD correlation) of the systematic factor X_1 in the default driver Y_1 and ρ_2 is the loading of the systematic factor in the LGD driver Y_2. However the LGD systematic factor $\omega X_1 + \sqrt{1-\omega^2} X_2$ is now composed of the PD systematic factor and an additional systematic factor X_2 characterizing possibly changing economic conditions during the recovery period, changes in the organization and efficiency of the workout process, etc. The model (copula) correlation between the PD systematic factor and the LGD systematic factor is parameterized by ω. We allow $\omega \in [-1,1]$ as the correlation could be in general negative. The correlation of the transformed portfolio PD and LGD given parameters of the model must be again calculated numerically, but we expect a value close or at least proportional to ω.

In practice we need to formulate the model in terms of time series. We observe a times series of default rates $\text{pd}(t), t = 1,...,T$ and a time series of loss given default rates $\text{lgd}(t), t = 1,...,T$. The PDs and LGDs are measured over certain time periods (e.g. quarterly, monthly, or annually) and on a large product portfolio, e.g. of consumer loans, or mortgages etc. According to our model the defaults and account level LGDs are driven by certain latent factors in the form of (6) that change over time. The latent systematic factors $x_1(t), t = 1,...,T$ and $x_2(t), t = 1,...,T$ determine the observed portfolio PD and LGD values through the functions g and h given by (4) and (5), while the independent idiosyncratic factors ζ_1, ζ_2 diversify away, i.e.

$$(7)\qquad \begin{aligned} &\text{pd}(t) = g(x_1(t)), t = 1,...,T, \\ &\text{lgd}(t) = h(\omega x_1(t) + \sqrt{1-\omega^2} x_2(t)), t = 1,...,T. \end{aligned}$$

The function g depends on the unknown correlation parameter ρ_1 and the default level y_D. The function h itself depends on ρ_2 and on the account level LGD distribution. Finally we have to make an assumption on the latent time series. The series are by definition Gaussian, with mean zero and variance one. We assume stationarity but we have to admit a possible autocorrelation since the systematic factors are supposed to represent some sort of macroeconomic variables. In general, we will assume ARMA(p,q) specification for the both time series, i.e.

$$(8) \qquad x_k(t) = \sum_{i=1}^{p} \alpha_{k,i} x_k(t-i) + \sum_{j=0}^{q} \beta_{k,j} u_k(t-j),\ t = 1,\ldots,T,\ k = 1,2$$

where $\alpha_{k,i}, \beta_{k,j}$ are the ARMA coefficients and $u_k(t), t = 1,..,T, k = 1,2$ iid N(0,1) innovations (with pre-sample values set equal to the mean values, i.e. $0 = x_k(0) = x_k(-1) = \cdots$ and $0 = u_k(0) = u_k(-1) = \cdots$).

Estimation of the Model

Our goal is to estimate the key model parameters ρ_1, ρ_2 and ω based on the description above and given certain observed PD and LGD time series $\mathrm{pd}(t), t = 1,\ldots,T$ and $\mathrm{lgd}(t), t = 1,\ldots,T$. In order to calculate a goodness of fit of the model we also need to specify the long-term LGD distribution function G, the default level y_D, the two ARMA(p,q) processes, and in fact also the iid N(0,1) innovations $u_k(t), t = 1,..,T, k = 1,2$. Given all that let us express the likelihood of the observed data $L\left(\langle \mathrm{pd}(t), \mathrm{lgd}(t)\rangle \mid \rho_1, \rho_2, \omega, G, y_D, \alpha_{k,i}, \beta_{k,j}\right)$. For the sake of brevity we are not going to list always all the conditional parameters $\rho_1, \rho_2, \omega, G, y_D, \alpha_{k,i}, \beta_{k,j}$ in the likelihood function. Following the argument of Witzany (2009d) we just need to express the likelihoods $L\left(\mathrm{pd}(t) \mid \mathrm{pd}(\tau), \tau < t\right)$ and $L\left(\mathrm{lgd}(t) \mid \mathrm{pd}(\tau), \tau \le t; \mathrm{lgd}(\tau), \tau < t\right)$ since the total likelihood can be decomposed as

$$L\left(\langle \mathrm{pd}(t), \mathrm{lgd}(t)\rangle\right) = \prod_{t=1}^{T} L\left(\mathrm{pd}(t) \mid \mathrm{pd}(\tau), \tau < t\right) \cdot L\left(\mathrm{lgd}(t) \mid \mathrm{pd}(\tau), \tau \le t; \mathrm{lgd}(\tau), \tau < t\right).$$

Here we are using the model assumption according to which $\mathrm{pd}(t)$ may depend only on the previous PD values while $\mathrm{lgd}(t)$ generally depends on the previous LGD values and on the PD values up to the time t. Note that the residuals $u_k(t)$ are implied by the observed data and the model parameters. They have the standard normal distribution and moreover are independent on the previous or in fact any other values. Thus to get the likelihood function we just need to look how the residuals are transformed to the observed PD and LGD values. According to (7) and (8)

(9) $\mathrm{pd}(t) = f(u_1(t)) = g(x_1(t)) = g\left(\beta_{1,0}u_1(t) + \cdots\right).$

Note that the expression (8) for $x_1(t)$ all the factors, except of $\beta_{1,0}u_1(t)$, are determined by $\mathrm{pd}(\tau), \tau < t$ and thus behave like constants. Consequently differentiating (9) and applying the chain rule

$$L\left(\mathrm{pd}(t) \mid \mathrm{pd}(\tau), \tau < t\right) = \frac{\phi(u_1(t))}{f'(u_1(t))} = \frac{\phi(u_1(t))}{\beta_{1,0}g'(x_1(t))}.$$

Similarly we may proceed to get the conditional likelihood of

$$\mathrm{lgd}(t) = \tilde{f}(u_2(t)) = h(\omega x_1(t) + \sqrt{1-\omega^2}x_2(t)),\ x_2(t) = \beta_{2,0}u_2(t) + \cdots.$$

The only difference is that we also have to differentiate the argument of h with respect to x_2, and then x_2 with respect to u_2:

$$L\left(\mathrm{lgd}(t) \mid \mathrm{lgd}(\tau), \tau < t; \mathrm{pd}(t), \tau \le t\right) = \frac{\phi(u_2(t))}{\tilde{f}'(u_2(t))} = \frac{\phi(u_2(t))}{\sqrt{1-\omega^2}\beta_{2,0}h'(\omega x_1(t) + \sqrt{1-\omega^2}x_2(t))}.$$

Finally we get

(10) $L\left(\langle \mathrm{pd}(t), \mathrm{lgd}(t) \rangle\right) = \prod_{t=1}^{T} \frac{\phi(u_1(t))}{\beta_{1,0}g'(x_1(t))} \prod_{t=1}^{T} \frac{\phi(u_2(t))}{\sqrt{1-\omega^2}\beta_{2,0}h'(\omega x_1(t) + \sqrt{1-\omega^2}x_2(t))}.$

The likelihood function is conditional upon the parameters $\rho_1, \rho_2, \omega, G, y_D, \alpha_{k,i}, \beta_{k,j}$. To make the estimation feasible we need to specify the account level LGD distribution. Our model assumes that there is a general (through the cycle) distribution of account level LGD. Observed average LGDs at different time period are lower or higher than their mean due to changing economic conditions. Our dataset needs to contain not only a time series of average $\mathrm{lgd}(t), t = 1, \ldots, T$ but also all the corresponding account level observations $\{lgd(a) \mid a \in A\}$. Thus the average loss given default rates can be expressed as

$lgd(t) = \frac{1}{|A(t)|}\sum_{a \in A(t)} lgd(a)$ where $A(t) = \{a \in A \mid t(a) = t\}$ is the set of accounts that defaulted in the time period t. The observed distribution can be fitted by a beta distribution as or by a kernel smoothed empirical distribution due to an irregular shape of our dataset as discussed in Witzany (2009d). We assume G to be given by any of those methods. Similarly we have not only the default rate time series $\mathrm{pd}(t), t=1,\dots,T$ but also the information on the number of non-defaulted observed accounts $N(t)$ at the beginning of each period t as well as the number of observed accounts $n(t)$ that defaulted during the period t. Hence $\mathrm{pd}(t)$ is the default rate $\frac{n(t)}{N(t)}$ observed over the period t. The overall long-term probability of default is estimated as the observation weighted average of $\mathrm{pd}(t), t=1,\dots,T$, i.e. $PD_0 = \sum_{t=1}^{T} n(t) \Big/ \sum_{t=1}^{T} N(t)$. Finally we set the default level $y_D = -\Phi^{-1}(PD_0)$ as used in (4).

To estimate the key parameters $\rho_1, \rho_2, \omega, \alpha_{k,i}, \beta_{k,j}$ maximizing the likelihood function (10) or applying the MCMC described below we still need to specify the two ARMA models for the PD and LGD systematic factors. Our goal is not to spend unnecessary energy on the latent time series model identification, but we have to keep in mind that the likelihood function (10) is based on the assumption that the residuals $u_k(t)$ are independent. To preserve consistency of the model we need to take into account any significant autocorrelation detected in the series. Since the functions g and h are nearly linear we will specify the ARMA parameters analyzing the time series $\mathrm{pd}(t), t=1,\dots,T$ and $\mathrm{lgd}(t), t=1,\dots,T$. Having parsimony in mind we certainly try to keep the number of ARMA coefficients at a minimum.

Having specified the two ARMA models we may finally estimate the correlations ρ_1, ρ_2, ω and a number of ARMA parameters $\alpha_{k,i}, \beta_{k,j}$, probably two or more, numerically maximizing the likelihood function (10). The maximization is not only computationally demanding but also potentially unstable since the search algorithm may get stuck in a local extreme of the numerically evaluated likelihood function. Moreover the MLE method does not give any confidence intervals of the estimated parameters. Bootstrapping applied Witzany (2009d) also becomes difficult due to a more complex structure of the PD and LGD input data. In this situation we propose to the Markov Chain Monte Carlo sampling that gives us not only point estimates but also Bayesian marginal distributions of all the parameters being estimated.

Markov Chain Monte Carlo

The Bayesian MCMC sampling algorithm has become a strong and frequently used tool to estimate complex models with multidimensional parameters, including latent parameters and state variables. Examples are financial stochastic models with jumps, stochastic volatility, with complex correlation structure, or switching-regime processes. For a more complete treatment of MCMC methods and applications we refer reader for example to Johannes, Polson (2003) or Lynch (2010).

MCMC provides a method to sample from multivariate densities that are not easy to sample from, by breaking these densities down into more manageable univariate or lower dimensional multivariate densities. To estimate a vector of parameters $\Theta = \left(\theta_1, ..., \theta_k\right)$ the Gibbs sampler works according to the following generic procedure:

0. Assign a vector of initial values to $\Theta^0 = \left(\theta_1^0, ..., \theta_k^0\right)$ and set $j = 0$.

1. Set $j = j + 1$.

2. Sample $\theta_1^j \sim p(\theta_1 \mid \theta_2^{j-1}, ..., \theta_k^{j-1})$.

3. Sample $\theta_2^j \sim p(\theta_2 \mid \theta_1^j, \theta_3^{j-1}, ..., \theta_k^{j-1})$.

$\vdots$

k+1. Sample $\theta_k^j \sim p(\theta_k \mid \theta_1^j, \theta_2^j, ..., \theta_{k-1}^j)$ and return to step 1.

According to the Clifford-Hammersley theorem the conditional distributions $p(\theta_l \mid \theta_1, ..., \theta_{l-1}, \theta_{l+1}, ..., \theta_k)$ fully characterize the joint distribution $p\left(\Theta\right)$ and moreover under mild conditions the Gibbs sampler distribution converges to the target joint distribution (Johannes, Polson, 2003).

The conditional probabilities are typically obtained applying the Bayes theorem to the likelihood function and a prior density, e.g.

(11) $p\left(\theta_1 \mid \theta_2^{j-1}, ..., \theta_k^{j-1}\right) \propto L\left(\text{data} \mid \theta_1, \theta_2^{j-1}, ..., \theta_k^{j-1}\right) \cdot \text{prior}\left(\theta_1\right)$.

We will use uninformative priors, so for simplicity, further on we assume that $\text{prior}\left(\theta_i\right) = 1$. In order to use the Gibbs sampler the right hand side of the proportional relationship needs to be normalized, i.e. we need to integrate the right hand side with respect to θ_1 conditional on

$\theta_2^{j-1},...,\theta_k^{j-1}$. If the integration of the right hand side is not analytically possible (which will be our case) then the Metropolis-Hastings algorithm can be used. It is based on the rejection sampling algorithm. For example in step 2 the idea is firstly to sample a new proposal value of θ_1^j and then accept it or reject it (i.e. reset $\theta_1^j := \theta_1^{j-1}$) so that we rather move to parameters with higher corresponding likelihood values.

Specifically, step 1 is replaced with a two step procedure:

1. A. Draw θ_1^j from a proposal density $q(\theta_1 \mid \theta_1^{j-1},\theta_2^{j-1},...,\theta_k^{j-1})$,

   B. Accept θ_1^j with probability

(12) $$\alpha = \min\left(\frac{p\left(\theta_1^j \mid \theta_2^{j-1},...,\theta_k^{j-1}\right) q\left(\theta_1^{j-1} \mid \theta_1^j,\theta_2^{j-1},...,\theta_k^{j-1}\right)}{p\left(\theta_1^{j-1} \mid \theta_2^{j-1},...,\theta_k^{j-1}\right) q\left(\theta_1^j \mid \theta_1^{j-1},\theta_2^{j-1},...,\theta_k^{j-1}\right)},1 \right).$$

It is again shown (see Johannes, Polson, 2003) that under certain mild conditions the limiting distribution is the joint distribution $p\left(\Theta\right)$ of the parameter vector. Note that the limiting distribution does not depend on the proposal density nor on the starting parameter values. It only makes the algorithm more or less efficient. A popular proposal density is random walk, i.e. sampling

(13) $$\theta_1^j \sim \theta_1^{j-1} + N(0,c).$$

The algorithm is then called random walk Metropolis-Hastings. The proposal density is in this case symmetric, i.e. the probability of going from θ_1^{j-1} to θ_1^j is the same as the probability of going from θ_1^j to θ_1^{j-1} (fixing the other parameters), and so the second part of the fraction in the formula for α in step 1B cancels out. Moreover the normalizing constant on the right hand of (11) cancels out in the fraction (12) as well and so the acceptance or rejection is driven by the likelihood ratio

$$R = \frac{L\left(\text{data} \mid \theta_1^j,\theta_2^{j-1},...,\theta_k^{j-1}\right)}{L\left(\text{data} \mid \theta_1^{j-1},\theta_2^{j-1},...,\theta_k^{j-1}\right)}.$$

In practice step 1B is implemented by sampling a $u \sim U(0,1)$ from the uniform distribution and accepting θ_1^j if and only if $u < R$.

In our particular case the estimated parameters are $\Theta = \left(\rho_1,\rho_2,\omega,\alpha_{k,i},\beta_{k,j}\right)$ where $\alpha_{k,i},\beta_{k,j}$ stand for a number of parameters depending on specification of the ARMA processes. As indicated above we use the random walk Metropolis-Hastings algorithm based on the likelihood function (10). The random walk variances are set to optimize the algorithm performance. The resulting empirical distribution approximating the true distribution of Θ can be used to estimate the key parameters ρ_1, ρ_2, ω from the marginal Bayesian distributions as the means, medians, or maximum likelihood values, and to analyze the confidence intervals.

Unexpected Loss Implied by the Model

Once the parameters, based on the historical PD and LGD values, have been estimated, the model may be used to obtain the distribution of the future losses. I.e. given $\mathrm{pd}(t), t = 1,...,T$ and $\mathrm{lgd}(t), t = 1,...,T$ we want to find the distribution of the relative loss $\mathrm{pd}(T+1)\cdot\mathrm{lgd}(T+1)$ in the future time period. Given the estimated parameters we know the functions g and h specified by (4) and (5), the systematic factors $x_1(t), x_2(t)$ and the residuals $u_1(t), u_2(t), t = 1,...,T$. To simulate a future loss rate

(14) $$\mathrm{pd}(T+1)\cdot\mathrm{lgd}(T+1) = g(x_1(T+1))\cdot h(\omega x_1(T+1) + \sqrt{1-\omega^2}\, x_2(T+1))$$

we just need to sample iid $u_1(T+1), u_2(T+1) \sim N(0,1)$ and calculate $x_1(T+1), x_2(T+1)$ according to the ARMA specification (8). A Monte Carlo simulation then yields a distribution of losses which can be used to estimate unexpected loss $\widehat{UL}(\alpha)$ on a given probability level α .

If the time step is shorter than one year, e.g. one month, and we need to estimate the one year unexpected loss, then we perform a simulation of the two processes 12 steps ahead up to $T+12$ and approximate the one-year horizon loss distribution by the distribution of the simulated cumulative values $\sum_{m=1}^{12} \mathrm{pd}(T+m)\cdot\mathrm{lgd}(T+m)$. Of course, at the end we are interested in comparing the estimated one-year horizon unexpected loss $\widehat{UL}(0.999)$ with the Basel II unexpected loss given by the regulatory formula that implicitly assumes zero PD x LGD correlation.

3. Empirical Study

Data description

We are going to apply the methodology described above on a dataset containing default and loss given default information on a portfolio of unsecured retail loans obtained from a large Czech bank. The dataset covers 57 months over the period 2002-2008, i.e. $T = 57$ and we will work with monthly periods. The number of accounts in the portfolio goes from 250 000 up to more than 700 000 accounts at the end of the observed period. We are given the numbers of non-defaulted accounts $N(t)$ at the beginning and the number of defaulted accounts at the end of each month, $n(t), t = 1,...,T$. The development of the monthly observed rate of default time series, i.e. $\mathrm{pd}(t) = n(t) / N(t), t = 1,...,T$, is shown in Figure 2 and its basic descriptive statistics in Table 1. The long term annualized probability of default is calculated as the observation weighted average of the monthly time series $PD_0 = \dfrac{91\,202}{25\,572\,087} = 0.36\%$ [2].

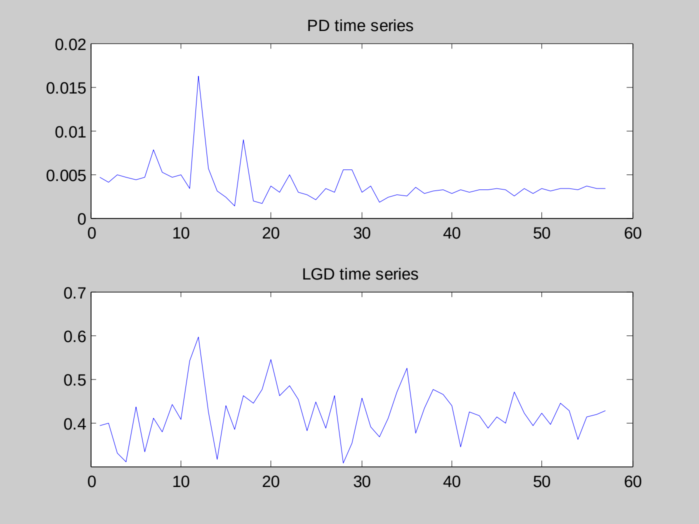

Figure 2. The monthly PD (upper chart) and average monthly LGD time series (lower chart)

PD time series		LGD time series	
Num	57	Num	57
Max	0.0162	Max	0.5963

[2] The average monthly default rate corresponds to the annualized default rate $4.28\% \doteq 12 \cdot 0.36\%$.

Min	0.0014	Min	0.3065
Mean	0.0038	Mean	0.4209
Median	0.0033	Median	0.4211
Range	0.0148	Range	0.2898
Std	0.0021	Std	0.0567

Table 1. Descriptive statistics of the annualized monthly PD and of the LGD time series

Regarding LGD we are just given a representative subsample of the 91 202 observed defaults. The original full database contains a recovery cash flows and other specific information on each of the accounts. Because of size the exported dataset is limited only to 4 000 randomly selected defaulted cases. The monthly number of defaults with observed LGD ranges from 39 to 108 which is not optimal but can be still considered as sufficient to obtain the monthly sample average values with a reasonable precision. The recovery rates and the complementary LGDs are calculated by the standard formula

$$(15) \qquad \mathrm{rr}(a) = \frac{1}{EAD(a)} \sum_{i=1}^{t(a)} \frac{RR(a,i)}{(1+r)^{i/12}}, \quad \mathrm{lgd}(a) = 1 - \mathrm{rr}(a),$$

where $RR(a,i)$ are the monthly recovery cash flows (net of recovery costs) recorded from the time of default of an account a, r is a discount rate set by the bank (usually the product interest rate), and $EAD(a)$ the exposure at default. The average exposure in the portfolio is around 50 000 CZK. Our empirical study however focuses only on observed relative LGDs. The development of the observed monthly average LGD time series, i.e. $\mathrm{lgd}(t), t = 1,...,T$, is shown in Figure 2 and its basic descriptive statistic**s** again in Table 1.

The banking sector generally did not systematically recovery data and lacks sufficiently long LGD time series. A particular problem is the fact that in-house recovery processes usually take a long time and so to calculate the ultimate recovery rates according to (15) we need 36 months or even longer recovery history (from the time of default). If we want to use data on relatively recent defaults some sort of extrapolation must to be applied (see Witzany et al, 2010). In order to get almost 5 year long time series (as required by Basel II) we have limited the recovery process length to 36 months, i.e. $t(a) \le 36$ in (15), and moreover extrapolated using the technique described in Rychnovsky (2009). The histogram of the observed account level LGDs is shown in Figure 3 and the basic descriptive statistics in Table 2.

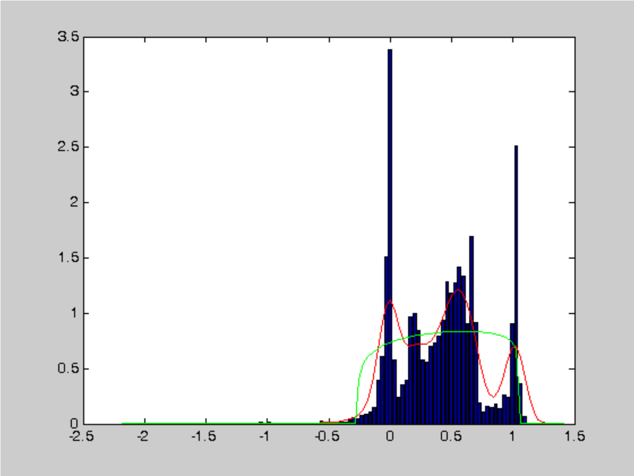

Figure 3. The histogram of observed LGDs, fitted beta, and kernel smoothed distributions.

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

Table 2. Descriptive statistics of the account level 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 distribution is not even bimodal, but rather tri-modal. The medium mode should be however accounted to the extrapolation procedure which tends to assign average values to accounts with only partial recovery history. 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. Figure 3 shows the $\cap$ - shaped beta distribution fitted the first two moments of the data. Since it deviates significantly from the observed distribution we will use the (tri-modal) kernel smoothed empirical distribution Q_k , as in Witzany (2009d) calculated in the MATLAB application using the *ksdensity* function (the optimal normal kernel widths has been set to $u = 0.0753$).

In order to specify the ARMA(p,q) models we need to inspect the autocorrelation and partial autocorrelation charts of the PD and LGD series. In both cases we also apply the Ljung-Box Q-test. The ACF and PACF patterns of the PD series in Figure 4 indicate that it is an MA(5) with lag 5 significant nonzero autocorrelation only. This also confirmed by Ljung-Box Q-test which shows lag 5 autocorrelation as the first that significantly differs from zero (on the 95% probability level).

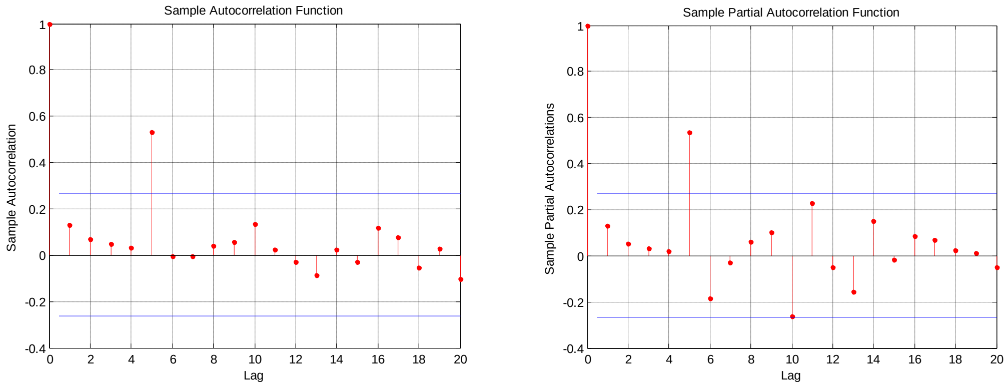

Figure 4. The PD time series sample and partial autocorrelation functions

Regarding the LGD time series the ACF and PACF functions (Figure 5) indicate lag 6 autocorrelation nonzero rather with the AR(6) pattern. The Ljung-Box Q-test however does not show any autocorrelation significantly different from zero (lag 6 p-value equals to 0.1386). Having parsimony in mind we choose the simplest ARMA(0,0) specification. However due to flexibility of the MCMC estimation approach we will be also able to test and compare the result in case of AR(6) specification (only with lag 6 coefficient nonzero). Although the ACF and PACF functions do not exhibit strong lag 1 autocorrelations we will also look at AR(1) models for both PD and LGD systematic factors. The reason is that the systematic factors are supposed to represent different states of the economy fluctuating over the cycles. Thus the monthly values should show a significant persistence. Low significance of the lag 1 autocorrelations might be explained by a too short observed time series that unfortunately does not go through the cycles.

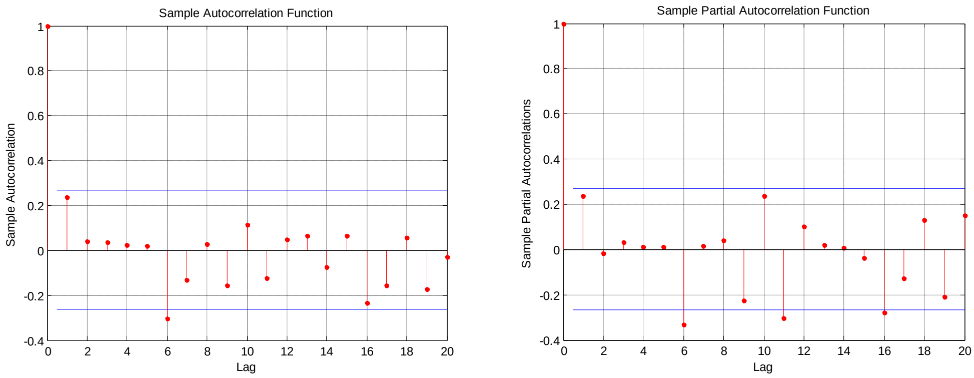

Figure 5. The LGD time series sample and partial autocorrelation functions

MCMC estimation

At this point we are ready to estimate the correlation coefficients ρ_1, ρ_2 and ω applying the model and the MCMC procedure described in Section 2. We run the MCMC procedure firstly 1000 times to obtain approximate initial values for the parameters and then 5000 times to get the desired parameter distributions. The initial values for ρ_1, ρ_2 were set at 0.04 and 0.03, for ω at 0.06, and for the PD systematic factor MA(6) coefficient $\beta_{1,6}$ at 0.44. The Metropolis-Hastings random walk steps (13) were set to 1%, 2%, 12%, and 10% for ρ_1, ρ_2, ω and $\beta_{1,6}$ respectively in order to achieve recommended acceptance rate between 30% and 40% (see Lynch, 2010). Note that the lag 0 coefficient is always calculated by $\beta_{1,0} = \sqrt{1-\beta_{1,6}^2}$. The results are summarized in Table 3. The average parameter estimates $\hat{\rho}_1 = 0.0189$ and $\hat{\rho}_2 = 0.0248$ are significant even on the 99% probability level, however the average estimate for $\hat{\omega} = 0.0775$ indicates that the PD x LGD correlation is positive but we cannot confirm that is differs from zero even on the 95% probability level.

Par.	Mean	Std	Median	1% quant.	5% quant.	95% quant.
ρ_1	0.0189	0.0038	0.0184	0.0119	0.0136	0.0258

ρ_2	0.0248	0.0049	0.0243	0.0156	0.0179	0.0339
ω	0.0775	0.1284	0.0801	-0.2204	-0.1304	0.2847
$\beta_{1,6}$	0.4113	0.0841	0.4214	0.1754	0.2532	0.5325

Table 3. Distribution of the estimated parameters (ARMA(0,6) + ARMA(0,0))

The MCMC method allows us to inspect the Bayesian parameter distributions and convergence of the simulation in more detail. For example Figure 6 shows a good convergence of the parameters ρ_1 and ω. We can see that the simulated values of ρ_1 remain positive while the simulated values of ω are often negative. This is more precisely reflected by the smoothed marginal densities of the two parameters shown in Figure 7**.** The empirical densities also allow us to find the modes of the empirical marginal distributions which are very close but not identical to the averages of the parameter estimates.

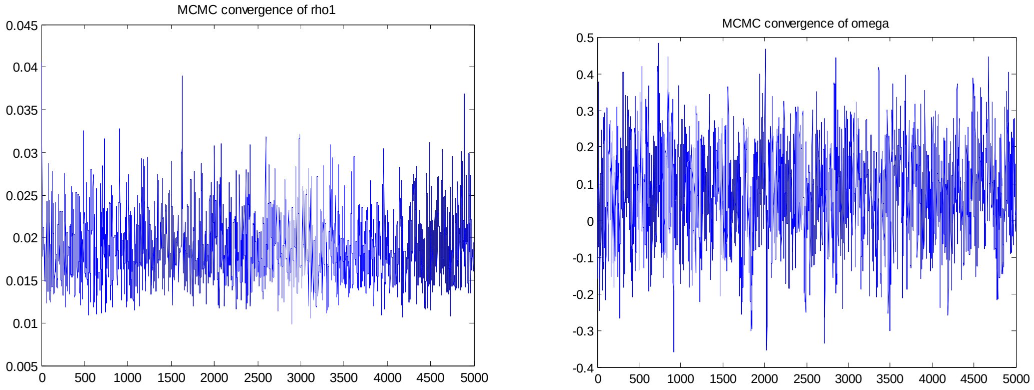

Figure 6. Simulated MCMC values of the parameters ρ_1 and ω

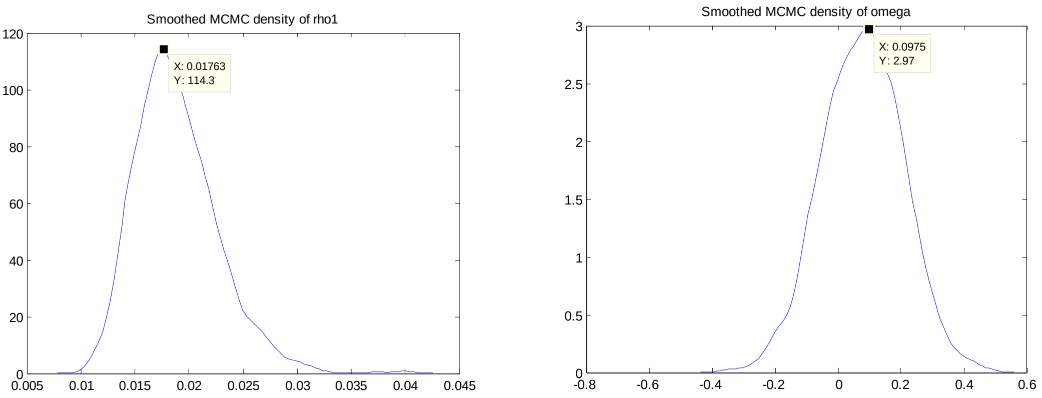

Figure 7. Kernel smoothed MCMC densities for the parameters ρ_1 and ω

We have also run the MCMC procedure for alternative ARMA specifications of the LGD systematic factor AR(6) or MA(6), or for AR(1) specification for both PD and LGD factors. The differences, in terms averages and confidence intervals of the estimated parameters, were relatively negligible. Table 4 shows the estimated parameters if both the PD and LGD systematic factors are specified as AR(1).

Par.	**Mean**	**Std**	**Median**	**1% quant.**	**5% quant.**	**95% quant.**
ρ_1	0.0205	0.0046	0.0197	0.0128	0.0145	0.0293
ρ_2	0.0260	0.0059	0.0251	0.0165	0.0186	0.0369
ω	0.1053	0.1234	0.1033	-0.1627	-0.0982	0.3120
$\alpha_{1,1}$	0.2350	0.1398	0.2297	-0.0873	0.0080	0.4716
$\alpha_{2,1}$	0.2436	0.1394	0.2450	-0.0976	0.0113	0.4786

Table 4. Distribution of the estimated parameters (ARMA(1,0) + ARMA(1,0))

It appears that the correlations ρ_1, ρ_2 estimates essentially do not depend on the ARMA specification, in case of ω there is, depending of the model, a variation which is not however surprising due to low significance of the parameter. This can be also visually demonstrated by looking on the two dimensional plot of the estimated MCMC correlation and an ARMA parameter, e.g. ω and $\beta_{1,6}$ shown on Figure 8**.** The figure indicates that the conditional distribution of ω if does not change too much if $\beta_{1,6}$ is set equal to a fixed value (around its mean).

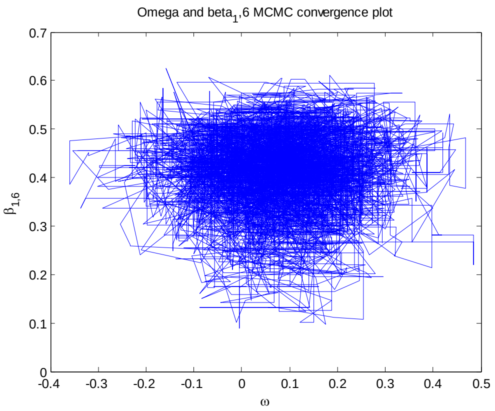

Figure 8. Joint MCMC convergence of ω and $\beta_{1,6}$

Unexpected Loss Estimation

Finally, we are able to simulate the distribution of losses during the next 12 months conditional on the last observed PD and LGD values as described at the end of Section 2. To simulate the PD and LGD processes 12 months ahead we firstly use the average estimated parameters from Table 3. The smoothed simulated distribution based on 100 000 simulations is shown in Figure 9. The unexpected loss on 99.9% probability level, i.e the 99.9% quantile of the simulated distribution $\widehat{UL}(0.999) = 3.01\%$ turns out to relatively low compared to the expected loss, i.e. the mean of the simulated losses $\widehat{EL} = 1.95\%$. Consequently $\widehat{UL}(0.999) - \widehat{EL} = 1.06\%$.

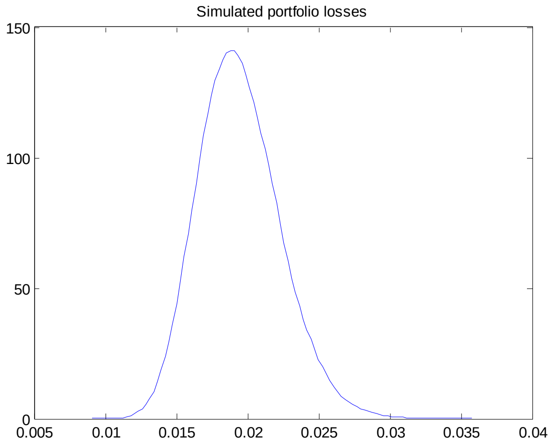

Figure 9. Smoothed density of simulate relative portfolio losses

More conservative estimate is obtained if we use 95% quantile values of the correlation parameters from Table 3. The resulting difference between the unexpected and expected loss is $\widehat{UL}(0.999) - \widehat{EL} = 1.46\%$. The same calculation in case of the AR(1) model (Table 4) gives a higher unexpected loss estimate $\widehat{UL}(0.999) - \widehat{EL} = 2.23\%$ that might be explained by stronger autocorrelations of the systematic factors. In those cases we have incorporated a sufficient estimation margin of conservatism into our unexpected loss calculation as required by Basel II. The uncertainty of the parameters can be in a Bayesian approach incorporated directly into the unexpected loss Monte Carlo simulation by sampling simultaneously the parameters from the MCMC previously estimated distributions.

Basel II PD correlation, in case of unsecured retail loans, is calculated by the regulatory formula (BCBS, 2006):

(16) $$\rho_{reg} = 0.03\frac{1-e^{-35p}}{1-e^{-35}} + 0.16\frac{e^{-35p}-e^{-35}}{1-e^{-35}} = 5.906\%$$

with p set equal to the annualized observed monthly probability of default $PD = PD_0 \cdot 12 = 4.28\%$. The unexpected default rate given by formula (4) is then $UDR = 15.91\%$ and the relative unexpected loss

$$UL_{BII} = (UDR - PD) \cdot LGD = 4.85\%$$

where $LGD = 41.73\%$ is the average observed account level loss given default. The difference between our estimated unexpected loss is explained by the relatively low PD correlation and by the fact that the one year loss is generated by a 12 month low series where the risk diversifies due to a low autocorrelation. To demonstrate importance of the LGD correlation and PD x LGD correlation let us use the model (14), but just for a single one-year period with $\rho_1 = \rho_{reg} = 5.906\%$, $\rho_2 = 2.48\%$, and $\omega = 7.75\%$. Then the unexpected minus expected loss $\widehat{UL}(0.999) - \widehat{EL} = 7.07\%$ is 2.22% more than the regulatory capital. This would in practice mean a more than 45% higher capital requirement.

4 Conclusions

The goal of the paper was to formulate a two-factor model for PD and LGD that, if observed on a large portfolio over time, show significant and according to many empirical studies systematic variability. We have proposed to estimate parameters of the model by the flexible MCMC sampling approach. The model with estimated parameters allows analyzing the distribution of future losses, in particular the unexpected losses on the regulatory 99.9% level to be compared with the Basel II capital requirement. We have tested the model on retail banking data. Since there is a general problem with longer LGD time series in the banking sector we had to work only with 57 months long series that does not go through the cycle and so cannot capture fully the relationship between the two series. The resulting estimated relatively low PD correlation $\hat{\rho}_1 = 0.0189$ and LGD correlation $\hat{\rho}_2 = 0.0248$ turn out to be significant on the 99% probability level, however the estimate of the PD x LGD correlation $\hat{\omega} = 0.0775$ only indicates that the correlation is positive. The estimate has a low significance probably due to the too short observed time period. In our opinion a good estimate would require 15 or more years of data covering good and bad times of the economy.

Our calculation of unexpected losses on one hand side shows that the regulatory formula strongly overestimates the unexpected loss based on our model parameters and a 12 one-month periods simulation. On the other if the unexpected loss is estimated in a single one-year horizon with $\rho_1 = \rho_{reg} = 5.906\%$ equal to the regulatory value and with the estimated LGD and PD x LGD correlations then the regulatory formula turns out to underestimate significantly the modeled unexpected loss. Our conclusion is that regulators and researchers should continue studying of the LGD and mutual PD x LGD correlations, which are not sufficiently captured by the regulatory formula, but may cause significant additional unexpected losses. We believe that the presented model and the proposed estimation methodology contribute to this effort.

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

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

Johannes, M. S. and Polson, N. (2003), "MCMC Methods for Continuous-Time Financial Econometrics", Working Paper, SSRN

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

Lynch S.M. (2010), "Introduction to Applied Bayesian Statistics for Social Scientists", Springer, p. 364

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

Rychnovsky M. (2009), "Mathematical Models of LGD", Diploma Thesis, Charles University – Faculty of Mathematics and Physics, April

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

Witzany J. (2009d), "Estimating LGD Correlation", IES Working Paper No. 21/2009

Witzany, J., Rychnovský, M., Charamza, P. (2010), " Survival Analysis in LGD Modeling " IES Working Paper 2/2010. IES FSV. Charles University.