

Keijsers, Bart; Diris, Bart; Kole, Erik

Working Paper

Cyclicalities in Losses on Bank Loans

Tinbergen Institute Discussion Paper, No. 15-050/III

Provided in Cooperation with:

Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Keijsers, Bart; Diris, Bart; Kole, Erik (2015) : Cyclicalities in Losses on Bank Loans, Tinbergen Institute Discussion Paper, No. 15-050/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:

<https://hdl.handle.net/10419/111729>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

TI 2015-050/III
Tinbergen Institute Discussion Paper



Cyclicalities in Losses on Bank Loans

Bart Keijsers

Bart Diris

Erik Kole

Erasmus School of Economics, Erasmus University Rotterdam, the Netherlands, and Tinbergen Institute, the Netherlands.

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at <http://www.tinbergen.nl>

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: <http://www.dsf.nl/>

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Cyclicalities in Losses on Bank Loans *

Bart Keijsers[†] Bart Diris
Erik Kole

Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam
Tinbergen Institute, Erasmus University Rotterdam

May 4, 2015

Abstract

Cyclicalities in the losses of bank loans is important for bank risk management. Because loans have a different risk profile than bonds, evidence of cyclicalities in bond losses need not apply to loans. Based on unique data we show that the default rate and loss given default of bank loans share a cyclical component, related to the business cycle. We infer this cycle by a new model that distinguishes loans with large and small losses, and links them to the default rate and macro variables. The loss distributions within the groups stay constant, but the fraction of loans with large losses increases during downturns. Our model implies substantial time-variation in banks' capital reserves, and helps predicting the losses.

Keywords: Loss-given-default, default rates, credit risk, capital requirements, dynamic factor models

JEL classification: C32, C58, G21, G33

*The authors thank NIBC Bank, in particular Michel van Beest, for providing access to the PECDC data and helpful comments. We thank Europlace Institute of Finance for financial support. We would like to thank participants at the Financial Risks International Forum Paris 2015, the PECDC General Members Meeting The Hague 2014, the ESEM Toulouse 2014, the CEF Oslo 2014, the NESG Tilburg 2014, and seminar participants at Erasmus University Rotterdam. The opinions expressed in this article are the authors' own and do not reflect the view of NIBC Bank or the PECDC.

[†]Corresponding author. Address: Burg. Oudlaan 50, Room H08-11, P.O. Box 1738, 3000DR Rotterdam, The Netherlands, Tel. +31 10 408 86 59. E-mail addresses keijsers@ese.eur.nl, diris@ese.eur.nl, kole@ese.eur.nl

1 Introduction

We propose a new model and use it to analyze a unique sample of defaulted bank loans. The central part of our model is a latent component that links the loss given default, the default rate and business cycle variables. This setup allows us to analyze whether such a component shows cyclical behavior, how the variables depend on it and how it is related to the macroeconomy. Recent advances in the risk management of bank loans, such as the stress tests for the banking sector, highlight that the risks on bank loans should not be investigated in isolation but in relation to the macroeconomic environment. As stated in the Basel II Accord, risk measures should “reflect economic downturn conditions where necessary to capture the relevant risks” (BCBS, 2005). Our model is exactly tailored to these kind of analyses.

We think that our research is interesting for two reasons. First, research on bank loans is scarce because data on defaulted bank loans are not easily available and typically constitute small samples (Grunert and Weber, 2009). Second, the properties like cyclicity of bank loan losses might differ from the more commonly studied bond losses. Banks monitor their loans more closely than bond owners, which influences both the default rate and the loss given default. Bank loans are often more senior than other forms of credit and are more often backed by collateral, which reduces the loss given default. Banks can postpone the sale of a borrower’s assets until a favorable economic state, hoping to receive a higher price. These effects can make the default rate and the loss given default less cyclical and less interrelated.

We base our research on default data from the Global Credit Data Consortium.¹ Several banks formed this consortium to pool anonymized information on their defaulted loans for research on credit risk. Currently, the consortium counts 47 banks, both inside and outside Europe. Each member has access to a subsample of this pooled database. Our study uses approximately 22,000 defaults over the period 2003–2010, based on the large and representative proportion of the database to which the Dutch NIBC Bank has access. As a consequence, our research is based on a larger sample and cross section than existing research such as Grunert and Weber (2009), Calabrese and Zenga (2010) and Hartmann-Wendels et al. (2014). They either use a smaller sample or focus on a single country or loan type.

¹In March 2015, the consortium changed its name to Global Credit Data Consortium. Its former name was Pan European Credit Data Consortium.

The loss of a portfolio of loans is typically split into three quantities: the default rate, the loss given default and the exposure at default. The first two elements are usually treated as outcomes of random processes, whereas the third is taken as given. An initial inspection of our data shows that the loss given default on bank loans differs in an important aspect from bonds. Though both distributions are bimodal, as either most of the loan is recovered or fully lost, the loss given default on bank loans can exceed 100% or fall below 0%. In the first case, the bank loses more than the initial loan, for example because of principal advances (the bank lends an additional amount to the borrower for recovery). In the second case, the bank recovers more than the initial loan, for example because it is entitled to penalty fees, additional interest or because of principal advances that are also recovered.

To capture the time and cross-sectional variation in, and dependence between default rates and loss given default, we construct a model that consists of four components. The first and central component is a latent factor that follows an autoregressive process and which we interpret as the credit cycle. The second component consists of a Bernoulli random variable for the default of a loan. In the third component we model the loss given default as a mixture of two normal distributions that differ in their means. The low-mean distribution corresponds to good loans with a high recovery rate, whereas the high-mean distribution relates to bad loans with a low recovery rate.² The loan being good or bad is determined by another latent Bernoulli variable. The parameters of both Bernoulli variables can vary in relation to the latent factor, or according to characteristics of the loans, such as seniority, security, and the industry to which the company belongs. Fourth, we add macroeconomic variables, because research on bond defaults has found a relation between the credit cycle and the state of the economy, see e.g. Allen and Saunders (2003), Pesaran et al. (2006), Duffie et al. (2007), Azizpour et al. (2010) and Creal et al. (2014).

Our model is a state space model with nonlinear and non-Gaussian measurement equations. Because it is not a standard linear Gaussian state space model, we can neither use the Kalman filter to infer the latent process, nor use straightforward maximum likelihood estimation to determine the parameters of our model. Instead, we derive how the simulation-based methods of Jungbacker and Koopman (2007) can be used to infer the latent process, and the Expectation Maximization

²The distinction between good and bad loans is also exploited in Knaup and Wagner (2012). They use this distinction to derive a bank's credit risk indicator.

algorithm of Dempster et al. (1977) to estimate the parameters.

Our results show that the default rate, the loss given default and the macro variables share a common component. This component shows cyclical behavior that leads to default rates that fluctuate between 0.2% and 7%, while loss given default fluctuates between 14% and 29%. High values for the common component indicate a bad credit environment with high default rates and high values for loss given default, and an economic downturn characterized by falling growth rates of GDP and industrial production, and an increasing unemployment rate. Interestingly, the credit cycle that we infer leads the unemployment rate by four quarters.

The time-variation in the loss given default is driven by time-variation in the probability of a defaulted loan being good or bad. We do not find evidence that the average loss given default for either good or bad loans varies over time. When the credit cycle deteriorates the fraction of loans for which most is lost increases, but the LGD conditional on a defaulted loan being good or bad does not vary. Monitoring should therefore concentrate on determining the loan type.

We use our model to determine the capital reserve required for a fictional loan portfolio as in Miu and Ozdemir (2006). We calculate the economic capital as the difference between the portfolio loss with a cumulative probability of 99.9% and the expected loss. From peak to bottom of the cycle, the economic capital increases from 0.15% to 2.23% of the total value of the loan portfolios, an increase of a factor 15. This increase shows the importance of incorporating cyclicity in risk management models. We also show that our model can reduce the uncertainty in LGD predictions. Because resolving loan defaults can take a couple of years, the macro variables in our model help predicting LGD.

Our findings contribute to the literature on credit risk in three ways. First our study shows that the losses on bank loans have a cyclical component that influences both their default rate and loss given default, and is related to the macroeconomy. Altman et al. (2005), Allen and Saunders (2003) and Schuermann (2004) document such a component for bonds. The loss given default for a typical loan is much lower than for a typical bond, but fluctuations have the same magnitude. We conclude that the cyclicity of loan losses resembles that of bonds despite the arguments for less cyclicity.

Second, we develop a new model that captures the unique properties of bank loans, and deviates from existing models. First, Creal et al. (2014) and Bruche and González-Aguado (2010) use

a standard Beta distribution which is bounded between zero and one for the loss given default, while we propose a mixture of normal distributions. Second, in the model by Bruche and González-Aguado (2010) the default rates of bonds and their loss given default jointly depend on a latent Markov chain, whereas we use an autoregressive process. Though the switches in a Markov chain can also give rise to a credit cycle, our inferred process can be more easily linked to macroeconomic variables. Third, our model is more general than Calabrese (2014a), who only models the loss given default. Her model accommodates a mixture of good and bad loans similar to our model, but the mixture probability is constant, whereas we explicitly model its cyclical behavior and link to the default rate and macro variables.

Third, our application shows the flexibility and the added value of our model for risk management. We show how characteristics of the loan or the borrower such as its security, size or industry influence the LGD and default rate. Our model can easily accommodate other characteristics that banks may have on their borrowers to improve their assessments of the risk on bank loans.

2 Data

In 2004, several banks cooperated to establish (what was later named) the Global Credit Data Consortium, a cross border initiative to help measure credit risk to support statistical research for the advanced internal ratings-based approach (IRB) under Basel II. The members pool their resolved defaults to create a large anonymous database. A resolved default is a default that is no longer in the recovery process and thus the final loss given default (LGD) is known. Every member gets access to part of the database, depending on its contribution. We have access to the subset available to NIBC, which contains 46,628 counterparties and 92,797 loans.³ Details such as the default and resolution date are available, as well as loan characteristics such as seniority, security, asset class and industry. We investigate the behavior of LGD for these groups separately. The fraction of the total database available varies per asset class, but overall the NIBC subset represents a large proportion of the Global Credit Data database.

In case of a default, the lender can incur losses, because the borrower is unable to meet its

³Members receive a new version semi-annually. Our sample is a subset of the June 2014 version.

obligations. The LGD is the amount lost as a fraction of the exposure at default (EAD). The LGD in the database is the economic LGD, defined as a sum of cash flows or payments discounted to the default date. We follow industry practice by applying a discount rate that is a combination of the risk free rate and a spread over it.

The default rate (DR) gives the number of defaulted loans as a fraction of the number of loans at the start of the year. Whereas the Global Credit Data Consortium was founded to pool observed defaults, not default rates, they expanded to include an observed DR database in 2009.⁴ The DR database contains default rates per asset class and industry, which we match to the LGD observations of the groups.

2.1 Sample Selection

We apply filters to the LGD dataset, following mostly NIBC's internal policy, to exclude non-representative observations. For details, see appendix A.

The LGD on bank loans can fall outside the interval between 0 (no loss) and 1 (a total loss) due to principal advances, legal costs or penalty fees. A principal advance is an additional amount loaned to aid the recovery of the defaulted borrower. If none of it is paid back, the losses are larger than EAD and LGD is larger than 1. If on the other hand the full debt is recovered, including penalty fees, legal costs and principal advances, the amount received during recovery is larger than EAD and the LGD is negative. We restrict the LGD between -0.5 and 1.5 , similar to Höcht and Zagst (2007) and Hartmann-Wendels et al. (2014). Figure 1(a) presents the empirical LGD distribution, and shows that we can not ignore this, because over 10% of the LGDs lie outside the $[0, 1]$ interval.

We restrict our analysis to the period 2003–2010. The LGD database for resolved defaults contains details of defaults from 1983 to 2014. Figure 2(a) shows the average LGD per year. The number of defaults in the database is small until the early 2000s and the average LGD is noisy because of it. The first defaults have been submitted by the banks in 2005. Not all banks might have databases with all relevant details of many years ago and most observations in the years before 2000 are the substantial losses with a long workout period still in the books.

The workout period is the main difference between bonds and bank loans. Bond holders directly

⁴Members receive a new version annually. We use the June 2013 version.

observe a drop in value as trading continues and the price is discounted by the expected recovery rate. For defaulted bank loans, a recovery process starts that should lead to debt repayment. When no more payments can be obtained, the default is resolved and the recovery process ends. The period from the default date to the resolved date is called the workout period. Most defaults are resolved within one to three years after default, but figure 1(b) shows that the recovery process can last more than five years.

Table I shows that the LGD is significantly higher for longer workout periods, which explains the high average LGD in figure 2(a) before 2003. The higher LGD for loans with longer workout periods is partly explained by discounting. The cash flows are discounted over a longer workout period, thus reducing the recovery and increasing the LGD. Additionally, the workout period is an indication of how hard it is to recover the outstanding debt. If the recovery takes time, it can be due to issues with restructuring or selling of the assets. If demand for an asset is high, it will be sold or restructured faster and its value will be higher.

The database only contains resolved defaults, for which the recovery process has ended, and therefore, by definition, the later years of the database (2011 to 2014) only contain defaults with shorter workout periods. Because a shorter workout period is related to a smaller LGD, the LGD is underestimated in the final years. The average LGD and number of defaults in 2011 is small compared to the previous years, see figure 2(a). Therefore, we restrict our analysis to the period 2003–2010.

Figure 2(b) shows the yearly default rate. In general, the default rate is relatively small with values mostly around 1%. The default rate increases during the financial crises, peaking in 2009 at a default rate of 2.2%, more than twice the default rate in the period 2003–2007. The figure shows that the total number of loans is large in 2003 and increases over time. This is mostly because the number of participating banks increases as well. To match the time period of the LGD dataset, we use the period 2003–2010.

The LGD sample after applying the sample selection consists of 22,080 observations of mostly European defaults, one of the most comprehensive datasets for bank loan LGD studied thus far. Grunert and Weber (2009) summarize the empirical studies on bank loan recovery rates. The largest dataset they found studies 5,782 observations over the period 1992–1995. More recently, Calabrese and Zenga (2010), Calabrese (2014a) and Calabrese (2014b) study a portfolio of 149,378

Italian bank loan recovery rates resolved in 1999 and Hartmann-Wendels et al. (2014) consider 14,322 defaulted German lease contracts from mainly 2001–2009. However, these studies focus on defaults from a single country or a single type whereas our dataset is more extensive.

[Figure 1 about here.]

[Table 1 about here.]

[Figure 2 about here.]

2.2 Sample Characteristics

In this section, we discuss the empirical LGD distribution, the pattern over time and differences across loan characteristics for our sample.

It is a stylized fact that LGD follows a bimodal distribution with most observations close to 0 or 1, see for example Schuermann (2004). In most cases, there is either no or a full loss on the default. Figure 1(a) shows that this also holds for our sample. By far most losses are close to 0, but there is an additional peak at 1. The data is limited to the interval -0.5 to 1.5 . Still, 12.52% of the observations are outside the $[0, 1]$ interval.

In our analysis, defaults are aggregated per quarter to have both a sufficient number of time periods and a sufficient number of observations per period. An advantage of aggregation by quarter is that it matches the frequency of macroeconomic variables. Figure 2(a) shows the average LGD for defaults per quarter. The LGD starts with a relatively large value in 2003 and gradually decreases until 2007. From 2007, the average LGD increases due to the financial crisis. The level is back at its pre-crisis average in 2009. We observe the same pattern for the default rate in figure 2(b).

Figure 3 provides a more in-depth view of the time-variation of the LGD. It shows how the empirical distribution varies from quarter to quarter for the period 2003–2010. All quarters display the bimodal shape with peaks at 0 and 1. The increased number of defaults due to the financial crisis is visible, as well as the increase of the height of the peak around a LGD of 1 for the period 2007–2009. The large proportion of full losses explains the large average LGD in those years. Our modeling framework exploits both the bimodality of and the time-variation in the LGD.

Summary statistics of the sample and subsets based on loan characteristics are presented in table II. As expected, the LGD for unsecured loans is on average larger than for secured loans

because the former are not backed by collateral or a guarantor. Also, the average LGD is larger for subordinated loans than for senior loans. For some groups not many defaults are available. Therefore, we limit our analysis of groups to those with on average at least 100 observations per quarter.

Table II shows that unimodality is rejected by Hartigan and Hartigan's (1985) dip test for the full sample as well as for subsamples, unless the number of defaults is small. The fraction of defaults with an LGD larger than 0.5 is reported to illustrate the close relation with the average LGD.

Because we want to compare multiple means and the LGD is not normally distributed, we use the Kruskal-Wallis (KW) test to test for differences in location of multiple distributions. The KW test is a nonparametric test based on ranks. It tests for the null hypothesis of identical distributions against the alternative of at least two distributions differing in location.

A KW test on the selected groups shows a significant difference between the senior secured and senior unsecured loans, with a p -value of 0.000. Even though the absolute difference between the average of SME and large corporate seems small, the p -value of the KW test is 0.000, strongly rejecting the null hypothesis of equal distributions. For the industries, the LGD for financials is significantly larger than for the industrials and consumer staples industries.

[Figure 3 about here.]

[Table 2 about here.]

2.3 Macroeconomic Variables

Allen and Saunders (2003), Pesaran et al. (2006), Duffie et al. (2007), Azizpour et al. (2010), Creal et al. (2014) and others show that bond defaults are related to the business cycle. We include macroeconomic variables to analyze this behavior for bank loans. We consider the same set of variables as Creal et al. (2014) to represent the state of the economy: the gross domestic product (GDP), industrial production (IP) and the unemployment rate (UR). The series included are the growth compared to the same quarter in the previous year, seasonally adjusted. To match the mostly European default dataset, we use macro variables of European OECD countries or the European union.

3 Model specification

We propose a ‘mixed-measurement’ model (Creal et al., 2014), where the observations can follow different distributions, but depend on the single latent factor α_t . The latent factor follows an AR(1) process,

$$\alpha_{t+1} = \gamma + \rho\alpha_t + \eta_t, \quad (1)$$

with $\eta_t \sim N(0, \omega^2)$. The initial state α_1 follows the unconditional distribution of the latent process, $\alpha_1 \sim N(\gamma/(1 - \rho), \omega^2/(1 - \rho^2))$.

3.1 Loss given default

Based on the empirical distribution in figure 1(a), we propose a mixture of two normals for the LGD and define distributions 0 and 1 as the distributions for good and bad loans,⁵

$$y_{it}^1 \sim \begin{cases} N(\mu_{j0}, \sigma_j^2) & \text{if } s_{it} = 0 \text{ (good loan),} \\ N(\mu_{j1}, \sigma_j^2) & \text{if } s_{it} = 1 \text{ (bad loan),} \end{cases} \quad (2)$$

with y_{it}^1 the LGD of loan i that defaulted at time t for $i = 1, \dots, N^1$, $j = 1, \dots, J$ and $t = 1, \dots, T$. We treat s_{it} as the unobserved state that is 1 if loan i at time t is a bad loan and 0 otherwise. The probability that LGD is a bad loan varies across loan characteristics, such as industry or seniority, and across time. We define the sets of loans belonging to the categories of a loan characteristic as C_j for $j = 1, \dots, J$. For example, we have the categories large corporate and SME for loan characteristic asset class. If the loan i defaulted at time t belongs to the category C_j , then the probability of a bad loan is

$$P(s_{it} = 1 | i \in C_j) = p_{jt} = \Lambda(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t), \quad (3)$$

where $\Lambda(x) = \exp(x)/(1 + \exp(x))$ is the logistic function. The model has one factor α_t , such that differences between groups are due to coefficients β_{j0}^1 and β_{j1}^1 .

The distribution can change in three ways, influencing the average LGD: (i) a change in the

⁵In section 7.3, we consider a mixture of Student’s t distributions.

mixture probability $P(s_{it} = 1)$, (ii) a change in the mean of good loans μ_{j0} and/or (iii) a change in the mean of bad loans μ_{j1} . Most LGDs are (close to) 0 or 1, see figure 1(a), so we do not expect the means to vary much. This is supported by figure 3, where the modes stay at 0 and 1, but the (relative) height of the peaks varies over time. Therefore, we propose that a larger (smaller) average LGD in a time period is caused by an increase (decrease) in the proportion of bad to good loans. We examine the alternative of time-varying means in section 7.2. We restrict the variance to be equal across the mixture components for identification of the modes and $\mu_{j0} < \mu_{j1}$ to interpret good and bad loans.

3.2 Default rate

The default of loan i at time t follows a Bernoulli distribution. This implies that the number of defaults in period t is a realization of a binomial distribution, as in Bruche and González-Aguado (2010). The distribution depends on the latent signal α_t through the probability of default q_{it} ,

$$y_{it}^d \sim \text{Binomial}(L_{it}, q_{it}), \quad (4)$$

$$q_{it} = \Lambda \left(\beta_{i0}^d + \beta_{i1}^d \alpha_t \right), \quad (5)$$

with y_{it}^d the number of defaults and L_{it} the number of loans of group i at time t for $i = 1, \dots, N^d$ and $t = 1, \dots, T$. If it is available from the DR database, we use the group specific default rate to match the J groups in the LGD component, otherwise we use the full sample default rate. Hence, N^d is either 1 or J , which is for example three for industries.

The defaults and loans are observed yearly, not quarterly. We set the third quarter equal to the yearly observation, because this is approximately the middle of year, and define the other quarters as missing.⁶ Using the same method, Bernanke et al. (1997) construct a monthly time series from a quarterly observed variable.

⁶If we set the second quarter equal to the yearly observation, we get similar results.

3.3 Macroeconomic variables

To relate the latent factor to the state of the economy, we add macroeconomic variables to the model,

$$\mathbf{y}_t^m = \boldsymbol{\beta}_0^m + \boldsymbol{\beta}_1^m \alpha_t + \boldsymbol{\nu}_t, \quad (6)$$

where \mathbf{y}_t^m is the $N^m \times 1$ observation of the macro variables at time t and $\boldsymbol{\nu}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ for $t = 1, \dots, T$. The macro variables are standardized to have zero mean and unit variance, such that we can easily compare the relation with the latent factor across the macro variables.

3.4 Missing values and identification

We do not observe multiple defaults per loan. We treat the loans for which the default date is not in period t as missing during that quarter. For each loan, we have one observed and $T - 1$ missing values. One of the advantages of a state space model is its ability to easily handle missing values. The densities are cross-sectionally independent given α_t , such that

$$\log p(\mathbf{y}_t | \alpha_t) = \sum_{i=1}^N \delta_{it} \log p_i(y_{it} | \alpha_t), \quad (7)$$

where $\mathbf{y}_t = (y_{1t}, \dots, y_{Nt})'$, $N = N^l + N^d + N^m$ is the number of observations and δ_{it} is an indicator function which is 1 if y_{it} is observed and 0 otherwise. Therefore, only the observed values determine the loglikelihood. The loglikelihood consists of the sum of the model components. The different components are given in appendix C.1.

Without restrictions, the latent factor $\boldsymbol{\alpha}$ and the coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are not identified. For identification of $\boldsymbol{\beta}_0$, we set the intercept $\gamma = 0$ in equation (1). To make sure $\boldsymbol{\beta}_1$ is identified, we standardize the signal variance $\omega^2 = 1$. Finally, we restrict one element of $\boldsymbol{\beta}_1$ to be positive to identify the sign of the signal.

4 Estimation

Estimation of the parameters, denoted by $\boldsymbol{\theta}$, is done using maximum likelihood. Analytical solutions are not available and direct numerical optimization is infeasible, due to the dimensionality of the optimization problem. Because it would be possible to optimize for the parameters $\boldsymbol{\theta}$ if $\boldsymbol{\alpha}$ were known and vice versa, we employ the Expectation Maximization (EM) algorithm, introduced by Dempster et al. (1977) and developed for state space models by Shumway and Stoffer (1982) and Watson and Engle (1983). The algorithm is a well-known iterative procedure consisting of repeating two steps, which is proven to increase the loglikelihood for every iteration.

The m -th iteration of the EM algorithm is

1. *E-step*: Given the estimate of the m -th iteration $\boldsymbol{\theta}^{(m)}$, take the expectation of the complete data loglikelihood $\ell_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{S}, \boldsymbol{\alpha})$,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = \mathbb{E}_{\boldsymbol{\theta}^{(m)}} [\ell_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{S}, \boldsymbol{\alpha})]. \quad (8)$$

Evaluating the expected value of the complete data loglikelihood implies that we need expected values for the states \mathbf{S} and the latent signal $\boldsymbol{\alpha}$ given the observed LGD, defaults and macro variables. Because the mixed-measurement model is a nonlinear non-Gaussian state space model, methods for linear Gaussian state space models like the Kalman filter are invalid. Following Jungbacker and Koopman (2007), we therefore apply importance sampling to get a smoothed estimate of the expected value, variance and autocovariance of $\boldsymbol{\alpha}$, the probability of a bad loan $P(s_{it} = 1|\mathbf{Y}, \boldsymbol{\alpha})$ and its cross-product. We draw from an approximating Gaussian state space model as importance density. Appendix B provides an outline of the method. The expected loglikelihood (8) is derived in appendix C.2 and C.3. Derivations for mode estimation of $\boldsymbol{\alpha}$, used to get the approximating Gaussian state space model, are in appendix C.4. We set the number of replications $R = 1000$ and employ four antithetic variables in the importance sampling algorithm. Increasing R does not impact the results.

2. *M-step*: Obtain a new estimate $\boldsymbol{\theta}^{(m+1)}$ by maximizing the expected loglikelihood with respect

to $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}^{(m+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}). \quad (9)$$

Solving equation (9) involves the method of maximum likelihood. We use analytical solutions for the parameters if possible because the EM algorithm is already an iterative optimization. That is, we numerically optimize ρ , β_{j0}^l and β_{j1}^l for $j = 1, \dots, J$, and β_{i0}^d and β_{i1}^d for $i = 1, \dots, N^d$, and use analytical solutions for the other parameters. The analytical solutions are conditional on the other parameters, which makes the two-step procedure an ECM algorithm (Meng and Rubin, 1993).

These steps are repeated until the stopping criterion is met. If the loglikelihood increase after the m -th step, denoted by $\ell(\boldsymbol{\theta}^{(m)}|\mathbf{Y}) - \ell(\boldsymbol{\theta}^{(m-1)}|\mathbf{Y})$, is smaller than $\epsilon = 10^{-3}$, we switch to direct numerical optimization of the loglikelihood until the loglikelihood increase is less than $\epsilon = 10^{-6}$. Increasing the precision does not impact results.

Following Ho et al. (2012), we initialize the EM algorithm by 2-means clustering with random starting values. Then, starting values for μ_{j0} and μ_{j1} are the sample mean of the two clusters and σ_j^2 their average variance. We set $\Lambda(\beta_{j0}^l)$ equal to the group proportions from 2-means clustering and $\Lambda(\beta_{i0}^d)$ to the average default rate. Finally, $\beta_{j1}^l = 1$ for all $j = 1, \dots, J$, $\beta_{i1}^d = 1$ for all $i = 1, \dots, N^d$, and the factor is initialized at zero.

5 Results

5.1 LGD and DR

First, we discuss results for the model without cross-sectional variation, i.e. we do not use different factors or coefficients for groups such as industries or asset classes. The LGD parameter estimates are presented in the first column of table III.

The parameter estimates clearly distinguish two distributions. The estimate for the mean of a good loan is 0.072 and for the mean of a bad loan 0.828. The estimates for the means confirm our interpretation of the components as the distributions of good and bad loans and captures the stylized fact that most LGDs are either close to 0 or 1. The mean for bad loans is not exactly 1,

because of the observations between 0 and 1.⁷

We cannot directly compare the sensitivities towards the factor of LGD and the defaults via the coefficients β_1^l and β_1^d because of the nonlinearity of the logistic function. Instead, we compare the average marginal effect of the signal, given by the average of the first derivative of the probability function with respect to the signal α_t , $1/T \sum_{t=1}^T \partial p_t / \partial \alpha_t$. We present these effects in panel F of table III.

The coefficients β_1^l and β_1^d are both significantly positive, which means that the probabilities of a bad loan p_t and of a default q_t move in the same direction. The significant effect of the factor is strengthened by a large average marginal effect of 0.041 for p_t , which means that a rise of the factor by one standard deviation increases the probability of a bad loan by 4.1%. It indicates that the factor has a stronger effect on the probability of a bad loan than for the default probability, where the effect is 1.2%, so p_t fluctuates more than q_t . They follow the same pattern over time, but at a different level, see figure 4. This is in line with research on losses on bonds, where DR and LGD are time-varying through a common cyclical component.

The factor underlying the probability of a bad loan is presented in figure 4(a). Due to the monotonicity of the logistic transformation, interpreting the factor and its coefficients is straightforward. The positive estimate for β_1^l means that an increase of the factor corresponds with an increase in the ex ante probability of a bad loan and a default.

The estimated factor resembles the average LGD. The first few years are characterized by a downward trend until 2007. In 2007 the level increases, after which in 2009 it decreases slightly. It differs however, for a couple of reasons. First, the factor is a combination of the LGD, the default rates and macroeconomic variables and is estimated using all three sources of information. Second, the factor is a smoothed estimate, which means that it is conditional on all information of the complete sample period. It is not simply the average of the LGD at the particular point in time, but contains information from the preceding and following observations.

Figure 5 shows the fit of the mixture for two quarters. The difference between the two panels is the ex ante probability of a bad loan p_t , which is larger in the second quarter of 2008, such that

⁷To fit the observations between the modes, the means are shrunk to 0.5. The shrinkage is stronger for μ_1 because the number of observations with an LGD near 1 is smaller than the number of defaults with an LGD near 0, see figure 5. The estimates of the means are closer to 0 and 1 if we replace the mixture of normals by a mixture of Student's t distributions, see section 7.3.

the relative height of the mode for bad loans compared to good loans is higher. The relatively high mode for bad loans corresponds with the relatively large fraction of high LGD observations in the second quarter of 2008. Further, for both quarters, the location of the distribution of good loans captures the large peak at 0 and the distribution of bad loans fits the high LGD observations. It captures the stylized fact and the changes across time match what we observe in the empirical distribution.

Similar to losses on bonds, the defaults show cyclical behavior. Higher default rates are accompanied by a higher probability of a bad loan, hence aggravating the loss during bad times. This time-variation should be taken into account. The claim that LGD estimates should “reflect economic downturn conditions where necessary to capture the relevant risks” (BCBS, 2005) is mostly motivated by research on bonds. We provide evidence that it holds for bank loans as well.

[Table 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

5.2 Relation with Macroeconomic Variables

Research on losses on bonds reports a link with macroeconomic variables, see e.g. Allen and Saunders (2003). Here, we investigate this link for losses on bank loans. The coefficients in panel D of table III indicate a relation between the state of the economy and credit conditions. They are significantly different from zero and have the expected sign. GDP and IP are negatively related, whereas the UR is positively related to the factor. Because a high factor implies more bad loans and a high default rate, this finding implies that both the number of defaults and the proportion of bad loans increase when the economy is in a bad state.

The first two columns of table III and figure 6 show that including the macro variables alters the factor only slightly. The factor of the model with macro variables in figure 6 is almost identical to the factor of the model without macro variables: the correlation between both factors is 0.996. Further, the coefficients of the LGD and DR only vary slightly. The macro variables support the shape of the factor we find, but do not drive the results.

The model includes contemporaneous macro variables, but the actual relation between the economy and the credit conditions may exhibit leading or lagging behavior. On the one hand, it could be that if the economy deteriorates, it takes a few months before companies are affected and go into default. On the other hand, the default of many companies could turn the economy into distress. The workout period further distorts the relation. The LGD are grouped by default date, but are a combination of cash flows in the recovery period, which depend on the state of the economy during the recovery period. The workout period can be less than a year or take up to five years.

Figure 7 presents the correlation of the factor for the model with default rate and macro variables with the macroeconomic variables for different leads and lags. The correlations with both GDP and IP show that the factor is contemporaneously related to the state of the economy. The unemployment rate is strongly related to the factor lagged three or four periods, in line with the other correlations because UR lags the state of the economy. Figure 7 confirms the significant relation we find from the estimates in table III, but also shows that the link between macro variables and credit conditions is more complicated than indicated by the model.

[Figure 6 about here.]

[Figure 7 about here.]

5.3 Loan Characteristics

We expect differences in credit conditions across loan characteristics, such as seniority and industry. We examine two possibilities: (i) one factor is underlying all groups, but the parameters vary depending on the characteristics, or (ii) every category has a different underlying factor. The first model type implies that the underlying credit conditions change in the same way for all industries, but the sensitivities towards it can vary. The second model type allows for different credit cycles per group. Under this model type, it could be that industry A is in distress, whereas the credit conditions in industry B are not irregular. We investigate how related the group-specific credit conditions are. If there are differences, banks can exploit this to diversify their portfolio. The characteristics we look into are security, seniority, asset class and industry, but only select categories with on average at least 100 observations per quarter. For the different asset classes and industries,

we use group-specific default rates. Macro variables are included to compare the relation with the business cycle.

The parameter estimates for the first model type, with a single common factor for all groups, are presented in the first column in tables IV–VI. The parameter estimates for the second model type, with a different latent factor per group, are presented in the remaining columns. The main differences in LGD across groups are the coefficients β_{j0}^1 and β_{j1}^1 , which determine the relation between the factor and the probability of a bad loan. The identification of the distributions of good and bad loans holds for all subsamples. The means for good loans are estimated between 0.05 and 0.09 and for bad loans between 0.79 and 0.86.

First, consider the difference between senior secured and unsecured loans in table IV. The intercept β_{j0}^1 is more negative for senior secured loans than for unsecured loans which means that the average LGD is smaller for secured loans. The average marginal effect is almost twice as high for senior secured loans than for unsecured loans, which implies a higher sensitivity for the time-variation in the latent factor. This is reflected in the high estimate for β_{j1}^1 for senior secured loans. Figure 8(a) shows that the pattern over time is much alike for the for the senior secured and unsecured factors. The correlation between the factors is 0.84. The finding that senior secured defaults are more time-varying than unsecured defaults is in contrast with the findings of Araten et al. (2004). An explanation is that the values of the securities backing the loan are cyclical. For example, demand for the collateral may vary depending on the state of the economy and its value therefore changes substantially over time.

Second, we consider the asset classes large corporate (LC) and small and medium enterprises (SME), where we have group-specific default rates. Table V presents the parameter estimates. If we consider different factors per group, the mean marginal effect for the LGD is approximately the same. However, the factor for SME is more connected with macroeconomic conditions. In particular the difference in the relation with UR is substantial, see panel D of table V. Given the estimates of β_{j0}^1 , β_{j1}^1 and the average marginal effect in the model with one common factor, the LGD for SME loans is slightly higher and more sensitive to changes in the factor. In contrast, the mean marginal effect for the default rate of LC is much higher than for SME. The estimate for β_{j0}^d is higher for LC than for SME, which implies that bank loans on LC default relatively more often.

The time-variation of credit conditions for asset classes LC and SME differs in pattern and

in size. The correlation between the factors for LC and SME in figure 8(b) is only 0.59. The LGD for SME is slightly more time-varying than for LC, but the default rate for LC is much more time-varying than for SME.

Third, we study the difference in cyclicalities across industries, for which we also have the industry-specific default rates. Consumer staples (CS) should be an industry with relatively stable credit conditions over time, because it produces goods such as food and household supplies. Demand will exist, independent of the economic situation. On the other hand, financials (FIN) are expected to be volatile, especially because the time period includes the financial crisis of 2007–2009.

The estimate of coefficient β_{j1}^l in table VI is largest for FIN, which induces the high mean marginal effect of 0.049. As expected, the LGD for FIN is most sensitive to changes in the factor. The estimate of β_{j0}^l is smaller than that of the other industries. Hence, the probability of a bad loan is in general smaller for FIN, but more sensitive to economic conditions. The time-variation explains the significantly larger average LGD over the full sample in section 2.2.

If we consider industry-specific factors, the mean marginal effect is smallest for CS, both for the LGD and the default rate. The estimates for the coefficients β_{j1}^l and β_{j1}^d are both approximately 0.05 in the second column of table VI. If we allow for a single factor, the mean marginal effects are larger, but still small compared to FIN.

For industrials (IND), the LGD is less time-varying compared to other industries. The estimate of β_{j1}^l for the model with a single factor is smaller than for CS. If all industries have a different factor, the time-varying effect is stronger for IND. On the other hand, the default rate is sensitive to credit conditions. The default rate has the highest mean marginal effect for IND, slightly higher than for FIN, in both the model with an industry-specific factor and the model with a single factor.

The difference between industries is further illustrated by the relation with the macroeconomic variables. Panel D of table VI shows that the factor underlying the industries FIN and IND is closer related to the macroeconomic variables than for CS. The coefficients for GDP and IP are estimated almost twice as high for FIN and IND. The credit conditions of CS are less related to the macroeconomic variables than for the other industries.

Figure 8(c) presents the single factor and the industry-specific factors. The industry factors move in the same direction over time, as the recent crisis hit all of the considered industries. But they are far from identical, with correlations from 0.52 between the factors of CS and IND to 0.81

between those of FIN and IND. The response of the factor of CS is lagged, but stronger than for the other industries. The increase of the factor of CS is larger than for other industries, but the mean marginal effect on the probability of a bad loan is only 0.007. This is due to the small estimate for the coefficient β_{j1}^1 .

The results indicate that it is important to consider the portfolio composition of defaults. We find that there is not a single credit cycle. The credit conditions across loan characteristics do share a common component, but clear differences exist. The probability of a bad loan determines most variation across groups, in terms of level and time-variation. Senior secured loans vary more over time, whereas unsecured loans have a higher average probability of a bad loan. Especially the time-variation across industries is important for banks focusing on a small set of sectors. Financials are sensitive to macro conditions, while consumer staples are more stable over time.

Banks gain a more in-depth view of the risk of the loan portfolio and how sensitive it is to macro conditions by taking the loan characteristics into account. For example, they can anticipate industry-specific time-variations and adjust their risk parameters accordingly to more accurately estimate the loan-specific risk. Further, they can diversify some of their time-varying risk by investing in different industries or asset classes.

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Figure 8 about here.]

5.4 Relation with Financial Variables

The latent factor underlies both the default rates and the loss given default. Therefore, we interpret the factor as a measure for credit conditions and we expect it to have a relation with other financial variables. We examine this by adding financial variables to the vector \mathbf{y}_t^m in equation (6). We select the long-term interest rate (LIR), the credit spread (CRS) and the yield spread (YLS). The LIR is the yield on the Euro area 10-years government bond, the CRS is the difference between the yield on the IBOXX European corporate 10+-years BBB-rated bond index and the LIR, and the

YLS is the difference between the LIR and the yield on the 3-month Euro Interbank Offered Rate (Euribor).

The last column of table III presents the estimates of the model including the financial variables. The estimates in panel E of table III indicate a significant positive relation between the factor and the financial variables. In particular, we observe a strong connection with the credit spread, given the estimate of 0.578 for β_1^m . This strong connection is not surprising, because the credit spread represents the market's expectation of credit risk, in terms of default rate and LGD.

Further, adding the financial variables does not affect the relation of the LGD, defaults and macro variables with the factor. The latent factor is virtually unchanged with a correlation of 0.997 with the factor from the model without financial variables.

The results validate our interpretation of the factor for credit conditions given the strong relation with the credit spread.

6 Applications in Risk Management

Banks can apply the model to assess their risks. The model can be used in a stress testing exercise or to formulate a downturn LGD (see e.g. Calabrese, 2014a). Below, we illustrate its use to determine economic capital and show how we can predict future credit conditions.

6.1 Economic Capital

Economic capital is an internal risk measure used by banks. It represents the amount of capital that the bank should hold in order to remain solvent, accounting for unexpected losses due to their exposure to risks. Given an estimate for the latent factor, we simulate realizations of defaults and LGD. In particular, it yields the economic capital, which is computed as the difference between the loss at a particular quantile in the right tail, usually at 99.9%, and the expected loss. For a given loan portfolio, based on simulated losses, we get the corresponding loss distribution.

We draw 50,000 times for every time period a portfolio of 2,000 loans, each with an exposure at default (EAD) of €1, similar to the portfolio considered by Miu and Ozdemir (2006) in their simulation exercise. Further, we consider the latent factor distributed as the one inferred in section 5.1, from the model including macro variables but without differences per group, such that we do

not have to make assumptions on the portfolio composition over industry and asset class.

The results show that changes in credit conditions can have severe consequences for the portfolio loss. First, figure 9(a) shows the loss distribution for two quarters. The two distributions are clearly different, only due to a difference in credit conditions, given by the level of the latent factor. In the fourth quarter of 2005, when the factor is low (see figure 4(a)), most losses are (close to) 0 and barely exceed 0.25% of the loan amount. On the other hand, in the second quarter of 2008, when the factor is high, the losses are more dispersed and almost always larger than 1% of the loan amount.

Figure 9(b) presents the loss distribution over time, and confirms the vast differences in loss distribution. The expected loss is mostly between 0% and 0.25% of the loan amount, but increases to 2% in 2008. The entire 95% confidence interval of losses in the second and fourth quarter of 2008 is larger than the maximum loss of the 95% interval for the period up to 2008, and 2010.

Finally, figure 9(c) presents the economic capital at 99.9% over time and shows that the right tail becomes fatter during bad credit conditions. The economic capital varies much over time with a maximum of 2.23% of the loan amount, approximately 15 times the minimum of 0.15%. Ignoring the fluctuation induces large uncovered potential losses.

An advantage of our model is its easy adaptation to a specific portfolio, the composition of loans over asset class, industry or other characteristics, because we can distinguish between different groups as in section 5.3. The loss distribution and economic capital varies per portfolio and sampling from a more tailored portfolio yields a more accurate loss distribution. Using a tailored model yields insight in the bank's risk and how it changes by adjusting the strategy, moving in or out a particular sector.

[Figure 9 about here.]

6.2 Prediction

The previous section shows that we can calculate the economic capital based on simulations for the estimation period, using information on the resolved defaults. It takes some time before all defaults are resolved and the economic LGD is observed. Due to the workout period, the estimation period excludes the most recent years, see section 2.1. We would like to have information on the losses on

the unresolved defaults that occurred in the recent years, and predict future losses. Banks can form an expectation of the write-offs for unresolved defaults and construct scenarios for future credit conditions. In this section, we propose two methods to predict future credit conditions, such that we can determine the economic capital out-of-sample.

The first method predicts the factor based on the autoregressive process in equation (1), such that the factor predicted h periods ahead is given by $\alpha_{T+h|T} = \rho^h \alpha_{T|T}$. A disadvantage of this method is that it ignores the available information in the recent years. The prediction of the factor at time $T+h$, for forecast horizon $h > 0$, is only based on the information in the estimation period, the period up to time T .

A second method uses the information available in the out-of-sample period. For example, we can use the macroeconomic information to update the prediction of the latent factor. An advantage of using macro variables is that they are reported quarterly, whereas credit data such as the default rate is usually only available at the end of the year. Another advantage is that if we only use the macro variables, the model reduces to a linear Gaussian state space model and we can apply straightforward methods such as the Kalman filter to update the prediction.

The method involves updating the prediction of the factor at time $T+h$ to get a filtered estimate, based on the information up to time $T+h$. The factor estimate from the autoregressive process at each one-step ahead forecast is adjusted based on the prediction error for the macroeconomic variables, given the relation in equation (6).

We compare both prediction methods by considering the out-of-sample period 2011–2013, for which macro information is available. Figure 10(a) shows that using the macro information strongly decreases the variance of the forecasted factor. For period $T+1$, the first quarter of 2011, the variance of the filtered factor is only 56% of the variance of the predicted factor, based on information up to and including 2010, time T . Further, the credit conditions are forecasted to be worse, whereas we cannot infer anything on the direction of the factor from the prediction based on information without out-of-sample macro information. The prediction of the first method is not far from the long-run average partly because the factor is already close to 0 at the end of the in-sample period. The prediction starts at the level of the factor at time T and converges to the long-run average with the rate of the AR coefficient ρ as the forecast horizon increases.

Figure 10(b) illustrates the difference in terms of economic capital for a portfolio of 2,000 loans,

each with an EAD of €1, simulated 50,000 times. The economic capital at 99.9% increases quickly with the forecast horizon if the macro information is ignored, due to the added uncertainty. On the other hand, if the macro information is taken into account, the economic capital does not explode. The variance of the filtered estimate does not increase with the forecast horizon, such that the difference in economic capital is only due to a change in the level of the factor.

Alternatively, we could have included financial variables, such as the credit spread, yield spread and long-term interest rate from section 5.4. Macro variables are reported with a lag, whereas the financial variables are available instantaneously and can be used to predict in real time. Further, we could include lagged macro variables such that the information we need to predict the credit conditions at time $T + 1$ is available at time T .

[Figure 10 about here.]

7 Alternative Specifications

The current model for LGD proposes time-variation in the probability of a bad loan in a mixture of normals. In this section, we challenge the proposed model by considering alternative specifications. First, we test whether the distribution is time-varying and consider time-varying means to introduce time-variation into the mixture. Second, we check whether a mixture of Student's t distributions provides a better fit.

7.1 No Time-Variation

To test whether the time-variation is present in our sample, we estimate a mixture of normals on the full set of LGD observations using a standard EM algorithm for mixtures.

The results in table VII confirm that the probability of a bad loan is time-varying. The large difference in loglikelihood provides significant evidence of time-variation in our LGD sample. Even if we account for the larger number of parameters, the model where the probability of a bad loan can vary over time provides a better fit in terms of BIC. Hence, this time-variation should be taken into account when constructing LGD estimates.

[Table 7 about here.]

7.2 Time-Variation in the Mean

The model for the LGD includes a time-varying probability of a bad loan. Here, we introduce time-variation through the mean of a good or bad loan. We replace equations (2) and (3) by a mixture of normals with $P(s_{it} = 1) = p$ and $\mu_{0t} = \mu_0 + \beta_1^1 \alpha_t$ and/or $\mu_{1t} = \mu_1 + \beta_1^1 \alpha_t$. To estimate the parameters and the latent factor, we combine the state space methods for dynamic linear models of Shumway and Stoffer (1991) and the univariate treatment of Durbin and Koopman (2012). The univariate treatment can be applied because the observations are independent conditional on the latent factor.

First, table VII shows that the parameter estimates for β_1^1 are very small. This indicates that the time-varying aspect in the means is not important. Second, table VII shows that the models with time-varying mean are barely able to improve the fit in terms of loglikelihood, and are even worse compared to the model without time-variation when accounting for the increased number of parameters. Third, figure 6 shows that models with (one of) the means time-varying underestimate the observed time-variation in the average LGD. The difference with the sample average is almost 0.1 on a scale of 0 to 1 during very good or bad times. An error of this size on the LGD leads to under- or overestimation of the expected losses on a portfolio, especially during times when it matters most. Only the model with time-varying probability of a bad loan is able to replicate the pattern of the average LGD over time.

These results confirm that the time-variation is due to changes in the probability of a bad loan, and not due to shifts of the location of the distributions.

[Figure 11 about here.]

7.3 Mixture of Student's t Distributions

To test whether a different distribution than the normal provides a better fit, we replace the mixture of normals in equation (2) by a mixture of Student's t distributions. The probability of a bad loan remains time-varying. To estimate the parameters and the latent factor, we use the ECM algorithm of Basso et al. (2010)⁸ in combination with the importance sampling methods described in appendix B.

⁸The ECM algorithm is implemented using a Matlab translation of the R package by Prates et al. (2011).

The last column of table VII presents the parameter estimates. The peaks of the empirical distribution are better identified compared to the mixture of normals with means equal to 0.03 and 0.99. The fit is improved, as reflected in a higher loglikelihood. Figure 12 shows the fit for the same quarters as figure 5 and clearly illustrates that the empirical distribution is described better by a mixture of Student's t distributions. Not only the location, but also the peakedness of the modes is captured. The improved identification could imply a more accurate estimate of the probability of a bad loan and the latent factor. However, figure 13 shows that the factor is largely unaffected by changes in distribution. The correlation between the factor with a mixture of normals and the factor from the model with a mixture of Student's t distributions is 0.99.

Using the model with the Student's t distribution comes with some disadvantages. Due to the large peakedness at 0 and 1 and the observations between the modes, the degrees of freedom are estimated close to 1 for good loans and even below 1 for bad loans. None of the moments are defined for a distribution with less than 1 degree of freedom, which makes it difficult to interpret μ_0 and μ_1 . Further, a default observed with an LGD larger than 1 could be identified as a good loan by the model, although it clearly is not because more than the full exposure is lost. Figure 14(b) shows that this happens because the smoothed probabilities of a bad loan $\hat{\pi}_{it} = P(s_{it} = 1 | y_{it}^1, \alpha_t)$ decrease (increase) for LGD values larger (smaller) than the mean of the distribution of bad (good) loans due to the fat tails. None of the issues occur if we use the mixture of normals, see figure 14(a).

The model with the mixture of normals is preferred over the mixture of Student's t distributions. Even though the model with the fat-tailed distribution provides a better fit, it is difficult to interpret due to the low degrees of freedom and because some loans are obviously misidentified by the model. Further, the gains in terms of identification of the latent signal are limited, because the factor is very similar to a specification with a mixture of normals.

[Figure 12 about here.]

[Figure 13 about here.]

[Figure 14 about here.]

8 Conclusion

The loss given default and the default rate on bank loans are both cyclical. We infer a common underlying factor that is a measure for the credit conditions and related to the business cycle. The time-variation in the loss given default is explained by changes in the probability of a bad loan. Banks should take this into account when determining the risk parameters.

We propose a model that describes the stylized facts of the loss given default on bank loans well. It captures the bimodal shape of the empirical distribution and provides an interpretation of the components, by explicitly modeling the extremes of no and full loss. It is flexible enough to include the differences across loan characteristics that we find. Further, the model has applications in risk management, such as the calculation of the economic capital and the prediction of future credit conditions.

References

- Allen, L. and Saunders, A. (2003). A survey of cyclical effects in credit risk measurement models. BIS Working Paper 126, Bank of International Settlements, Basel, Switzerland.
- Altman, E., Brady, B., Resti, A., and Sironi, A. (2005). The Link between Default and Recovery Rates: Theory, Empirical Evidence, and Implications. *The Journal of Business*, 78(6):2203–2228.
- Araten, M., Jacobs, Jr., M., and Varshney, P. (2004). Measuring LGD on Commercial Loans: An 18-Year Internal Study. *The RMA Journal*, pages 28–35.
- Azizpour, S., Giesecke, K., and Schwenkler, G. (2010). Exploring the Sources of Default Clustering. Technical report, Stanford University working paper series.
- Basel Committee on Banking Supervision (2005). Guidance on Paragraph 468 of the Framework Document. *Bank for International Settlements*.
- Basso, R. M., Lachos, V. H., Cabral, C. R. B., and Ghosh, P. (2010). Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics & Data Analysis*, 54(12):2926–2941.
- Bernanke, B. S., Gertler, M., and Watson, M. W. (1997). Systematic Monetary Policy and the Effects of Oil Price Shocks. *Brookings Papers on Economic Activity*, (1):91–157.
- Bruche, M. and González-Aguado, C. (2010). Recovery rates, default probabilities, and the credit cycle. *Journal of Banking & Finance*, 34(4):754–764.
- Calabrese, R. (2014a). Downturn Loss Given Default: Mixture distribution estimation. *European Journal of Operational Research*, 237(1):271–277.
- Calabrese, R. (2014b). Predicting bank loan recovery rates with a mixed continuous-discrete model. *Applied Stochastic Models in Business and Industry*, 30(2):99–114.
- Calabrese, R. and Zenga, M. (2010). Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking & Finance*, 34(5):903–911.

- Creal, D., Schwaab, B., Koopman, S. J., and Lucas, A. (2014). Observation Driven Mixed-Measurement Dynamic Factor Models with an Application to Credit Risk. *Review of Economics and Statistics*, 96(5):898–915.
- De Jong, P. and Shephard, N. (1995). The simulation smoother for time series models. *Biometrika*, 82(2):339–350.
- 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*, 39(1):1–38.
- Duffie, D., Saita, L., and Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83(3):635–665.
- Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.
- Grunert, J. and Weber, M. (2009). Recovery rates of commercial lending: Empirical evidence for German companies. *Journal of Banking & Finance*, 33(3):505–513.
- Hartigan, J. A. and Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics*, 13(1):70–84.
- Hartmann-Wendels, T., Miller, P., and Töws, E. (2014). Loss given default for leasing: Parametric and nonparametric estimations. *Journal of Banking & Finance*, 40:364–375.
- Ho, H. J., Pyne, S., and Lin, T. I. (2012). Maximum likelihood inference for mixtures of skew Student- t -normal distributions through practical EM-type algorithms. *Statistics and Computing*, 22(1):287–299.
- Höcht, S. and Zagst, R. (2007). Loan Recovery Determinants - A Pan-European Study. *Working Paper*.
- Jungbacker, B. and Koopman, S. J. (2007). Monte Carlo estimation for nonlinear non-Gaussian state space models. *Biometrika*, 94(4):827–839.
- Knaup, M. and Wagner, W. (2012). A Market-Based Measure of Credit Portfolio Quality and Bank’s Performance During the Subprime Crisis. *Management Science*, 58(8):1423–1437.

- Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278.
- Miu, P. and Ozdemir, B. (2006). Basel Requirement of Downturn LGD: Modeling and Estimating PD & LGD Correlations. *Journal of Credit Risk*, 2(2):43–68.
- Pesaran, M. H., Schuermann, T., Treutler, B.-J., and Weiner, S. M. (2006). Macroeconomic Dynamics and Credit Risk: A Global Perspective. *Journal of Money, Credit and Banking*, 38(5):1211–1261.
- Prates, M. O., Lachos, V. H., and Cabral, C. R. B. (2011). mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions. *R package version 0.2-9*.
- Schuermann, T. (2004). What Do We Know About Loss Given Default? In Shimko, D., editor, *Credit Risk Models and Management*, chapter 9. Risk Books, London, 2nd edition.
- Shumway, R. H. and Stoffer, D. S. (1982). An Approach to Time Series Smoothing and Forecasting using the EM Algorithm. *Journal of Time Series Analysis*, 3(4):253–264.
- Shumway, R. H. and Stoffer, D. S. (1991). Dynamic Linear Models with Switching. *Journal of the American Statistical Association*, 86(415):763–769.
- Watson, M. W. and Engle, R. F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23(3):385–400.

Appendix A Data Filter

Following Höcht and Zagst (2007), who also use data from the Global Credit Data Consortium, and NIBC's internal policy, we apply the following filters to the LGD database.

- $EAD \geq \text{€}100,000$. The paper focuses on loans where there has been an actual (possible) loss, so EAD should be at least larger than 0. Furthermore, there are some extreme LGD values in the database for small EAD. To account for this noise, loans with EAD smaller than € 100,000 are excluded.
- $-10\% < ((CF + CO) - (EAD - EAR))/(EAD + PA) < 10\%$, where CF cash flows, CO charge-offs and PA principal advances. The cash flows that make up the LGD should be plausible, because they are the major building blocks of the LGD. A way of checking this is by looking at under-/overpayments. The difference between the EAD and the exposure at resolution (EAR), where resolution is the moment where the default is resolved, should be close to the sum of the cash flows and charge-offs. The cash flow is the money coming in and the charge-off is the acknowledgement of a loss in the balance sheet, because the exposure is expected not to be repaid. Both reduce the exposure and should explain the difference between EAD and EAR. There might be an under- or overpayment, resulting in a difference. To exclude implausible cash flows, these loans are excluded when they are more than or equal to 10% of the EAD and principal advances (PA). The 10% is a choice of the Global Credit Data Consortium.
- $-0.5 \leq LGD \leq 1.5$. Although theoretically, LGD is expected between 0 and 1, it is possible to have an LGD outside this range, e.g. due to principal advances or a profit on the sale of assets. Abnormally high or low values are excluded. They are implausible and influence LGD statistics too much.
- *No government guarantees*. The database contains loans with special guarantees from the government. Most of the loans are subordinated, but due to the guarantee, the average of the subordinated LGD is lower than expected. Because the loans are very different from others with the same seniority and to prevent underestimation of the subordinated LGD, these loans are excluded from the dataset.

Some consortium members also filter for high principle advances ratios, which is the sum of the principal advances divided by the EAD. Even though high ratios are plausible, they are considered to influence the data too much and therefore exclude loans with ratios larger than 100%. NIBC does include these loans, because they are supposed to contain valuable information and the influence of outliers is mitigated because they cap their LGD to 1.5. The data shows that the principal advances ratio does not exceed 100%, so applying the filter does not affect the data and is therefore not considered.

Appendix B Importance Sampling

We outline the simulation based method of importance sampling, which we use to evaluate the non-Gaussian state space model. For more information on importance sampling for state space models, see for example Durbin and Koopman (2012).

Consider the following nonlinear non-Gaussian state space model with a linear and Gaussian signal,

$$\mathbf{y}_t \sim p(\mathbf{y}_t | \alpha_t), \quad (10)$$

$$\alpha_{t+1} = \rho \alpha_t + \eta_t, \quad (11)$$

with $\eta_t \sim \text{NID}(0, \omega^2)$ for $t = 1, \dots, T$, where \mathbf{y}_t is an $N \times 1$ observation vector and α_t the signal at time t . For notational convenience, we express the state space model in matrix form. We stack the observations into an $N \times T$ observation matrix $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_T)'$ and $T \times 1$ signal vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_T)'$ such that we have

$$\mathbf{Y} \sim p(\mathbf{Y} | \boldsymbol{\alpha}), \quad (12)$$

$$\boldsymbol{\alpha} \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{\Psi}). \quad (13)$$

The method of importance sampling is a way of evaluating integrals by means of simulation. It can be difficult or infeasible to sample directly from $p(\boldsymbol{\alpha} | \mathbf{Y})$, which is the case for non-Gaussian state space models. Therefore, an importance density $g(\boldsymbol{\alpha} | \mathbf{Y})$ is used to approximate the $p(\boldsymbol{\alpha} | \mathbf{Y})$ from which it is easier to sample. In particular, consider the evaluation of the expected value of

the function $x(\boldsymbol{\alpha})$,

$$\bar{x} = E[x(\boldsymbol{\alpha})|\mathbf{Y}] = \int x(\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\mathbf{Y})d\boldsymbol{\alpha} = \int x(\boldsymbol{\alpha})\frac{p(\boldsymbol{\alpha}|\mathbf{Y})}{g(\boldsymbol{\alpha}|\mathbf{Y})}g(\boldsymbol{\alpha}|\mathbf{Y})d\boldsymbol{\alpha} = E_g\left[x(\boldsymbol{\alpha})\frac{p(\boldsymbol{\alpha}|\mathbf{Y})}{g(\boldsymbol{\alpha}|\mathbf{Y})}\right]. \quad (14)$$

For a non-Gaussian state space model with Gaussian signal, this can be rewritten into

$$\bar{x} = \frac{E_g[x(\boldsymbol{\alpha})w(\boldsymbol{\alpha}, \mathbf{Y})]}{E_g[w(\boldsymbol{\alpha}, \mathbf{Y})]}, \quad (15)$$

$$w(\boldsymbol{\alpha}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\alpha})}{g(\mathbf{Y}|\boldsymbol{\alpha})}, \quad (16)$$

which contains densities that are easy to sample from. Then \bar{x} is estimated by replacing the expectations with its sample estimates.

The function to be estimated $x(\boldsymbol{\alpha})$ can be any function of $\boldsymbol{\alpha}$. For example, the mean is estimated by setting $x(\boldsymbol{\alpha}) = \boldsymbol{\alpha}$. For the estimation of the likelihood $L(\boldsymbol{\theta}|\mathbf{Y}) = p(\mathbf{Y}|\boldsymbol{\theta})$ we have

$$L(\boldsymbol{\theta}|\mathbf{Y}) = \int \frac{p(\boldsymbol{\alpha}, \mathbf{Y})}{g(\boldsymbol{\alpha}|\mathbf{Y})}g(\boldsymbol{\alpha}|\mathbf{Y})d\boldsymbol{\alpha} = g(\mathbf{Y}) \int \frac{p(\boldsymbol{\alpha}, \mathbf{Y})}{g(\boldsymbol{\alpha}, \mathbf{Y})}g(\boldsymbol{\alpha}|\mathbf{Y})d\boldsymbol{\alpha} = L_g(\boldsymbol{\theta}|\mathbf{Y})E_g[w(\boldsymbol{\alpha}, \mathbf{Y})], \quad (17)$$

where $L_g(\boldsymbol{\theta}|\mathbf{Y}) = g(\mathbf{Y})$ is the likelihood of the approximating Gaussian model. This is estimated by the sample analog $\hat{L}_g(\boldsymbol{\theta})\bar{w}$, with $\bar{w} = (1/R)\sum_{r=1}^R w(\boldsymbol{\alpha}^{(r)}, \mathbf{Y})$ where $\boldsymbol{\alpha}^{(r)}$, $r = 1, \dots, R$, are independent draws from $g(\boldsymbol{\alpha}|\mathbf{Y})$, using the simulation smoother. Its log version is $\log \hat{L}(\boldsymbol{\theta}|\mathbf{Y}) = \log \hat{L}_g(\boldsymbol{\theta}|\mathbf{Y}) + \log \bar{w}$.

B.1 Mode Estimation

The importance density $g(\boldsymbol{\alpha}|\mathbf{Y})$ must be chosen such that it is easy to sample from and approximates the target density well. If the importance density does not share the support of the target density, the estimation will be inaccurate. An example of a suitable importance density is to take a Gaussian density that has the same mean and variance as the target density.

It is possible to sample from $p(\boldsymbol{\alpha}|\mathbf{Y})$ for a Gaussian state space model using the simulation smoother developed by De Jong and Shephard (1995). Therefore, we would like to get a Gaussian model that approximates the non-Gaussian model, defined by equations (10) and (11).

The approximating Gaussian model can be obtained by mode estimation. It is a

Newton-Raphson procedure to get the mode of signal $\boldsymbol{\alpha}$ for a non-Gaussian state space model. The procedure of mode estimation is outlined below, including how it results into an approximating Gaussian state space model.

Given an initial guess \mathbf{g} for the mode of $\boldsymbol{\alpha}$, for example based on knowledge of the data, we have the following Newton-Raphson procedure to get a new estimate of the mode,

$$\mathbf{g}^+ = \mathbf{g} - (\ddot{p}(\boldsymbol{\alpha}|\mathbf{Y})|_{\boldsymbol{\alpha}=\mathbf{g}})^{-1}\dot{p}(\boldsymbol{\alpha}|\mathbf{Y})|_{\boldsymbol{\alpha}=\mathbf{g}}, \quad (18)$$

with $\dot{p}(\cdot|\cdot) = \partial \log p(\cdot|\cdot)/\partial \boldsymbol{\alpha}$, a $T \times 1$ vector, and $\ddot{p}(\cdot|\cdot) = \partial^2 \log p(\cdot|\cdot)/\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'$, a $T \times T$ matrix. We cannot directly apply the procedure because $p(\boldsymbol{\alpha}|\mathbf{Y})$ is unknown, but Bayes' rule enables us to rewrite the smoothed log density as

$$\log p(\boldsymbol{\alpha}|\mathbf{Y}) = \log p(\mathbf{Y}|\boldsymbol{\alpha}) + \log p(\boldsymbol{\alpha}) - \log p(\mathbf{Y}), \quad (19)$$

where $\log p(\mathbf{Y}|\boldsymbol{\alpha}) = \sum_{t=1}^T \log p(\mathbf{y}_t|\alpha_t) = \sum_{t=1}^T \sum_{i=1}^N \log p_i(y_{it}|\alpha_t)$, $p(\boldsymbol{\alpha})$ is given in equation (13) and the last term does not depend on $\boldsymbol{\alpha}$ and can thus be left unspecified. The distribution $p_i(y_{it}|\alpha_t)$ may vary over i , so observations are allowed have different distributions. We get

$$\dot{p}(\boldsymbol{\alpha}|\mathbf{Y}) = \dot{p}(\mathbf{Y}|\boldsymbol{\alpha}) - \boldsymbol{\Psi}^{-1}(\boldsymbol{\alpha} - \boldsymbol{\mu}), \quad (20)$$

$$\ddot{p}(\boldsymbol{\alpha}|\mathbf{Y}) = \ddot{p}(\mathbf{Y}|\boldsymbol{\alpha}) - \boldsymbol{\Psi}^{-1}, \quad (21)$$

where $\dot{p}(\mathbf{Y}|\boldsymbol{\alpha}) = (\dot{p}_1(\mathbf{y}_1|\alpha_1), \dots, \dot{p}_T(\mathbf{y}_T|\alpha_T))'$ and $\ddot{p}(\mathbf{Y}|\boldsymbol{\alpha}) = \text{diag}(\ddot{p}_1(\mathbf{y}_1|\alpha_1), \dots, \ddot{p}_T(\mathbf{y}_T|\alpha_T))$, with $\dot{p}_t(\cdot|\cdot) = \partial \log p(\cdot|\cdot)/\partial \alpha_t$ and $\ddot{p}_t(\cdot|\cdot) = \partial^2 \log p(\cdot|\cdot)/\partial \alpha_t \partial \alpha_t'$. If we plug in the expressions (20) and (21) in equation (18), we get

$$\begin{aligned} \mathbf{g}^+ &= \mathbf{g} - (\ddot{p}(\mathbf{Y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\mathbf{g}} - \boldsymbol{\Psi}^{-1})^{-1} (\dot{p}(\mathbf{Y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\mathbf{g}} - \boldsymbol{\Psi}^{-1}(\boldsymbol{\alpha} - \boldsymbol{\mu})) \\ &= (\boldsymbol{\Psi}^{-1} + \mathbf{A}^{-1})^{-1} (\mathbf{A}^{-1}\mathbf{z} + \boldsymbol{\Psi}^{-1}\boldsymbol{\mu}), \end{aligned} \quad (22)$$

$$\mathbf{z} = \mathbf{g} + \mathbf{A}\dot{p}(\mathbf{Y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\mathbf{g}}, \quad (23)$$

$$\mathbf{A} = -(\ddot{p}(\mathbf{Y}|\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\mathbf{g}})^{-1}, \quad (24)$$

where $\mathbf{z} = (z_1, \dots, z_T)'$ a $T \times 1$ vector and $\mathbf{A} = \text{diag}(A_1, \dots, A_T)$ a $T \times T$ matrix.

It can be shown that equation (22) is the output from the Kalman filter and smoother for a linear Gaussian model with ‘observation’ vector \mathbf{z} and ‘variance’ matrix \mathbf{A} . From mode estimation, we have thus obtained the following approximating Gaussian model,

$$z_t = \alpha_t + u_t, \tag{25}$$

$$\alpha_{t+1} = \rho\alpha_t + \eta_t, \tag{26}$$

where $u_t \sim \text{NID}(0, A_t)$ and $\eta_t \sim \text{NID}(0, \omega^2)$ for $t = 1, \dots, T$, with z_t and A_t defined in equations (23) and (24). The Newton-Raphson procedure described above is equivalent to repeatedly applying the Kalman filter and smoother to this model. The density $p(\boldsymbol{\alpha}|\mathbf{z})$ from the model is Gaussian and approximates the non-Gaussian target model well, because it has the same mean and variance. Therefore, the density $p(\boldsymbol{\alpha}|\mathbf{z})$ from equations (25) and (26) is suitable as an importance density.

Appendix C EM Equations

C.1 Observed Data Loglikelihood

The likelihood of the observed data \mathbf{Y} , which includes the LGD, the defaults and the macro variables, conditional on the latent factor $\boldsymbol{\alpha}$ and the parameters $\boldsymbol{\theta}$ is given by the product of the densities,

$$L(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{t=1}^T \prod_{i=1}^N (p_i(y_{it}|\boldsymbol{\alpha}))^{\delta_{it}}, \tag{27}$$

where δ_{it} is 1 if y_{it} is observed and 0 if it is missing or unobserved and $N = N^l + N^d + N^m$ the total number of observations. The conditional loglikelihood is then given by

$$\ell(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\alpha}) = \sum_{t=1}^T \sum_{i=1}^N \delta_{it} \log p_i(y_{it}|\alpha_t), \quad (28)$$

$$\log p(y_{it}^l|\alpha_t) = \sum_{j=1}^J \zeta_{ij} \log \left((1 - p_{jt})\phi_{j0}(y_{it}^l) + p_{jt}\phi_{j1}(y_{it}^l) \right), \quad (29)$$

$$\log p(y_{it}^d|\alpha_t) = \log \begin{pmatrix} L_{it} \\ y_{it}^d \end{pmatrix} + y_{it}^d \log(q_{it}) + (L_{it} - y_{it}^d) \log(1 - q_{it}), \quad (30)$$

$$\log p(\mathbf{y}_t^m|\alpha_t) = -\frac{N^m}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{y}_t^m - \boldsymbol{\beta}_0^m - \boldsymbol{\beta}_1^m \alpha_t)' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t^m - \boldsymbol{\beta}_0^m - \boldsymbol{\beta}_1^m \alpha_t), \quad (31)$$

where ζ_{ij} is 1 if loan i belongs to group j and 0 otherwise, and $\phi_{jk}(\cdot)$ is the normal density function with mean μ_{jk} and variance σ_j^2 given that $s_{it} = k$, for $k = 0, 1$. Groups are defined by the characteristics of the loans, for example industry, country or seniority.

The observed data loglikelihood is obtained by integrating out the stochastic latent factor out of the joint density of the observations and this latent factor,

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \int p(\mathbf{Y}, \boldsymbol{\alpha}|\boldsymbol{\theta}) d\boldsymbol{\alpha} = \int p(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}|\boldsymbol{\theta}) d\boldsymbol{\alpha}. \quad (32)$$

This observed data loglikelihood has no closed form expression because $\boldsymbol{\alpha}$ enters the likelihood non-linearly. The likelihood will be evaluated using the importance sampling methods of appendix B.

C.2 Complete Data Loglikelihood

The joint density of the model is given by

$$p(\mathbf{Y}, \mathbf{S}, \boldsymbol{\alpha}) = p(\alpha_1) \prod_{t=2}^T p(\alpha_t|\alpha_{t-1}) \prod_{t=1}^T \prod_{i=1}^N \delta_{it} p(y_{it}, s_{it}|\alpha_t), \quad (33)$$

where we have for the joint density of the observed LGD and the unobserved states

$$\begin{aligned}
p(y_{it}^1, s_{it} | \alpha_t) &= p(y_{it}^1 | \alpha_t, s_{it}) p(s_{it}) = \prod_{j=1}^J \left\{ (p_{jt} \phi_{j1})^{s_{it}} ((1 - p_{jt}) \phi_{j0})^{1-s_{it}} \right\}^{\zeta_{ij}} \\
&= \prod_{j=1}^J \left\{ \left(\frac{\exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)}{1 + \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)} \phi_{j1} \right)^{s_{it}} \left(\frac{1}{1 + \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)} \phi_{j0} \right)^{1-s_{it}} \right\}^{\zeta_{ij}}. \quad (34)
\end{aligned}$$

Further, we have a Gaussian signal following an AR(1) process. This means that the complete data loglikelihood for the parameter vector $\boldsymbol{\theta}$ is

$$\ell_c(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{S}, \boldsymbol{\alpha}) = p(\alpha_1) + \sum_{t=2}^T \log p(\alpha_t | \alpha_{t-1}) + \sum_{t=1}^T \sum_{i=1}^N \delta_{it} \log p_i(y_{it}, s_{it} | \alpha_t), \quad (35)$$

$$\log p(\alpha_1) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(P_1) - \frac{1}{2P_1} (\alpha_1 - a_1)^2, \quad (36)$$

$$\log p(\alpha_t | \alpha_{t-1}) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\omega^2) - \frac{1}{2\omega^2} (\alpha_t - \rho \alpha_{t-1})^2, \quad (37)$$

$$\begin{aligned}
\log p(y_{it}^1, s_{it} | \alpha_t) &= \sum_{j=1}^J \zeta_{ij} \left\{ s_{it} (\beta_{j0}^1 + \beta_{j1}^1 \alpha_t) - \log \left(1 + \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t) \right) \right. \\
&\quad + (1 - s_{it}) \left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_j^2) - \frac{1}{2\sigma_j^2} (y_{it}^1 - \mu_{j0})^2 \right) \\
&\quad \left. + s_{it} \left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_j^2) - \frac{1}{2\sigma_j^2} (y_{it}^1 - \mu_{j1})^2 \right) \right\}, \quad (38)
\end{aligned}$$

and $\log p(y_{it}^d, s_{it} | \alpha_t) = \log p(y_{it}^d | \alpha_t)$ and $\log p(\mathbf{y}_t^m, s_{it} | \alpha_t) = \log p(\mathbf{y}_t^m | \alpha_t)$ given in equations (30) and (31).

C.3 Expected Loglikelihood

The expected loglikelihood, also known as the Q -function, given the m -th step estimate for $\boldsymbol{\theta}^{(m)}$ for our model is given by

$$\begin{aligned}
Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}\right) &= \text{constant} - \frac{1}{2} \log\left(\frac{1}{1-\rho^2}\right) - \frac{1-\rho^2}{2}(P_{1|T} + \alpha_{1|T}^2) \\
&\quad - \frac{1}{2}(\hat{e}_{00} - 2\rho\hat{e}_{10} + \rho^2\hat{e}_{11}) \\
&\quad + \sum_{t=1}^T \sum_{i=1}^{N^1} \delta_{it} \left\{ \sum_{j=1}^J \zeta_{ij} \left\{ \hat{\pi}_{it} \beta_{j0}^1 + \beta_{j1}^1 \mathbf{E}[s_{it} \alpha_t | \mathbf{Y}] \right\} \right. \\
&\quad - \mathbf{E}_{\boldsymbol{\alpha}|\mathbf{Y}} \left[\log \left(1 + \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t) \right) \right] \\
&\quad + (1 - \hat{\pi}_{it}) \left(-\frac{1}{2} \log(\sigma_j^2) - \frac{1}{2\sigma_j^2} (y_{it}^1 - \mu_{j0})^2 \right) \\
&\quad \left. + \hat{\pi}_{it} \left(-\frac{1}{2} \log(\sigma_j^2) - \frac{1}{2\sigma_j^2} (y_{it}^1 - \mu_{j1})^2 \right) \right\} \Bigg\} \\
&\quad + \sum_{t=1}^T \sum_{i=N^1+1}^{N^1+N^d} \delta_{it} \left\{ y_{it}^d (\beta_{i0}^d + \beta_{i1}^d \alpha_t) \right. \\
&\quad \left. - L_{it} \mathbf{E}_{\boldsymbol{\alpha}|\mathbf{Y}} \left[\log \left(1 + \exp(\beta_{i0}^d + \beta_{i1}^d \alpha_t) \right) \right] \right\} \\
&\quad + \sum_{t=1}^T \left\{ -\frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \text{tr} \{ \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}_1^m P_{t|T} (\boldsymbol{\beta}_1^m)' \} \right. \\
&\quad \left. - \frac{1}{2} \text{tr} \{ \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t^m - \boldsymbol{\beta}_0^m - \boldsymbol{\beta}_1^m \alpha_{t|T}) (\mathbf{y}_t^m - \boldsymbol{\beta}_0^m - \boldsymbol{\beta}_1^m \alpha_{t|T})' \} \right\},
\end{aligned} \tag{39}$$

where the constant does not depend on any of the latent variables or parameters and we set the parameters $a_1 = 0$, $P_1 = 1/(1 - \rho^2)$ and $\omega^2 = 1$ for identification, see section 3.4. Further,

$$\hat{e}_{10} = \sum_{t=2}^T P_{t-1,t|T} + \alpha_{t-1|T} \alpha_{t|T}, \tag{40}$$

$$\hat{e}_{11} = \sum_{t=2}^T P_{t-1|T} + \alpha_{t-1|T}^2 = \sum_{t=1}^{T-1} P_{t|T} + \alpha_{t|T}^2, \tag{41}$$

with $\alpha_{t|T} = \mathbf{E}[\alpha_t | \mathbf{Y}]$ the smoothed factor and $P_{t|T} = \text{Var}(\alpha_t | \mathbf{Y})$ and $P_{t,t-1|T} = \text{Cov}(\alpha_t, \alpha_{t-1} | \mathbf{Y})$ its variance and autocovariance. The probability of the states s_{it} depends on the mean and variance of the mixture components and the ex ante mixture probability p_{jt} , which is a function of α_t . Therefore, the posterior mixture probabilities $\hat{\pi}_{it} = \mathbf{E}[s_{it} | y_{it}^1, \alpha_t] = P(s_{it} = 1 | y_{it}^1, \alpha_t)$ and the

expectation of the cross-product of the states and the signal $E[s_{it}\alpha_t|y_{it}^1]$ are computed using the law of iterated expectations: $E[s_{it}\alpha_t|y_{it}^1] = E[E[s_{it}\alpha_t, y_{it}^1|\alpha_t]|y_{it}^1]$. The expected values are calculated using importance sampling, using that for the expectation of a function of the states and the latent factor $x(\mathbf{S}, \boldsymbol{\alpha})$ conditional on the observed data, we have

$$\begin{aligned} E[x(\mathbf{S}, \boldsymbol{\alpha})] &= \int \int x(\mathbf{S}, \boldsymbol{\alpha}) p(\mathbf{S}, \boldsymbol{\alpha} | \mathbf{Y}) d\mathbf{S} d\boldsymbol{\alpha} \\ &= \int \int x(\mathbf{S}, \boldsymbol{\alpha}) p(\mathbf{S} | \mathbf{Y}, \boldsymbol{\alpha}) p(\boldsymbol{\alpha} | \mathbf{Y}) d\mathbf{S} d\boldsymbol{\alpha} \\ &= \int \left(\int x(\mathbf{S}, \boldsymbol{\alpha}) p(\mathbf{S} | \mathbf{Y}, \boldsymbol{\alpha}) d\mathbf{S} \right) p(\boldsymbol{\alpha} | \mathbf{Y}) d\boldsymbol{\alpha}. \end{aligned} \quad (42)$$

Using moments of the log-normal distribution, a first order Taylor approximation $E[\log(X)] \approx \log(E[X]) - \frac{1}{2} \text{Var}(X)/(E[X])^2$ and define $\theta_{it} = \beta_{i0}^d + \beta_{i1}^d \alpha_t$ for notational convenience, we get

$$\begin{aligned} E_{\boldsymbol{\alpha} | \mathbf{Y}} \left[\log \left(1 + \exp(\beta_{i0}^d + \beta_{i1}^d \alpha_t) \right) \right] &\approx \log \left(E_{\boldsymbol{\alpha} | \mathbf{Y}} [1 + \exp(\theta_{it})] \right) - \frac{1}{2} \frac{\text{Var}(1 + \exp(\theta_{it}))}{\left(E_{\boldsymbol{\alpha} | \mathbf{Y}} [1 + \exp(\theta_{it})] \right)^2} \\ &= \log \left(1 + \exp(\theta_{i,t|T} + \frac{1}{2} \text{Var}(\theta_{i,t|T})) \right) \\ &\quad - \frac{1}{2} \frac{\exp(2\theta_{i,t|T} + \text{Var}(\theta_{i,t|T}))}{\left(1 + \exp(\theta_{i,t|T} + \frac{1}{2} \text{Var}(\theta_{i,t|T})) \right)^2} \\ &\quad \times (\exp(\text{Var}(\theta_{i,t|T})) - 1) \\ &= \log \left(1 + \exp(\beta_{i0}^d + \beta_{i1}^d \alpha_{t|T} + \frac{1}{2} (\beta_{i1}^d)^2 P_{t|T}) \right) \\ &\quad - \frac{1}{2} \frac{\exp(2(\beta_{i0}^d + \beta_{i1}^d \alpha_{t|T}) + (\beta_{i1}^d)^2 P_{t|T})}{\left(1 + \exp(\beta_{i0}^d + \beta_{i1}^d \alpha_{t|T} + \frac{1}{2} (\beta_{i1}^d)^2 P_{t|T}) \right)^2} \\ &\quad \times \left(\exp((\beta_{i1}^d)^2 P_{t|T}) - 1 \right), \end{aligned} \quad (43)$$

and $E_{\boldsymbol{\alpha} | \mathbf{Y}} \left[\log \left(1 + \exp(\beta_{j0}^l + \beta_{j1}^l \alpha_t) \right) \right]$ defined similarly.

C.4 Mode Estimation - Derivatives

For the mode estimation algorithm we need to derive the first and second derivative of the distribution of the observed variable conditional on the signal. The log of the density of y_{it} given α_t is given in equations (29)–(31). We rewrite the loglikelihood for the LGD observations in equation

(29) as

$$\begin{aligned} \log p(y_{it}^1 | \alpha_t) &= \sum_{j=1}^J \zeta_{ij} \left\{ -\log(1 + \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)) \right. \\ &\quad \left. + \log \left(\exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t) \phi_{j1}(y_{it}^1) + \phi_{j0}(y_{it}^1) \right) \right\}. \end{aligned} \quad (44)$$

The first derivative is

$$\dot{p}_t(\mathbf{y}_t | \alpha_t) = \sum_{i=1}^N \dot{p}_t(y_{it} | \alpha_t), \quad (45)$$

$$\begin{aligned} \dot{p}_t(y_{it}^1 | \alpha_t) &= \sum_{j=1}^J \zeta_{ij} \left\{ -\frac{\beta_{j1}^1 \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)}{1 + \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)} \right. \\ &\quad \left. + \frac{\beta_{j1}^1 \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t) \phi_{j1}(y_{it}^1)}{\exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t) \phi_{j1}(y_{it}^1) + \phi_{j0}(y_{it}^1)} \right\}, \end{aligned} \quad (46)$$

$$\dot{p}_t(y_{it}^d | \alpha_t) = y_{it}^d \beta_{i1}^d - \beta_{i1}^d L_{it} \frac{\exp(\beta_{i0}^d + \beta_{i1}^d \alpha_t)}{1 + \exp(\beta_{i0}^d + \beta_{i1}^d \alpha_t)}, \quad (47)$$

$$\dot{p}_t(\mathbf{y}_t^m | \alpha_t) = (\boldsymbol{\beta}_1^m)' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t^m - \boldsymbol{\beta}_0^m - \boldsymbol{\beta}_1^m \alpha_t), \quad (48)$$

and the second derivative is

$$\ddot{p}_t(\mathbf{y}_t | \alpha_t) = \sum_{i=1}^N \ddot{p}_t(y_{it} | \alpha_t), \quad (49)$$

$$\begin{aligned} \ddot{p}_t(y_{it}^1 | \alpha_t) &= \sum_{j=1}^J \zeta_{ij} \left\{ -\frac{(\beta_{j1}^1)^2 \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)}{\left(1 + \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)\right)^2} \right. \\ &\quad \left. + \frac{(\beta_{j1}^1)^2 \exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t) \phi_{j0}(y_{it}^1) \phi_{j1}(y_{it}^1)}{\left(\exp(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t) \phi_{j1}(y_{it}^1) + \phi_{j0}(y_{it}^1)\right)^2} \right\}, \end{aligned} \quad (50)$$

$$\ddot{p}_t(y_{it}^d | \alpha_t) = (\beta_{i1}^d)^2 L_{it} \frac{\exp(\beta_{i0}^d + \beta_{i1}^d \alpha_t)}{(1 + \exp(\beta_{i0}^d + \beta_{i1}^d \alpha_t))^2}, \quad (51)$$

$$\ddot{p}_t(\mathbf{y}_t^m | \alpha_t) = -(\boldsymbol{\beta}_1^m)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}_1^m, \quad (52)$$

where we use $\frac{\partial}{\partial x} [\exp(ax)/(c + b \exp(ax))] = \exp(ax)ac/(c + b \exp(ax))^2$.

C.5 Maximum Likelihood Estimates

The maximum likelihood estimator (MLE) of the means of the normal distributions in the mixture of normals conditional on the other parameters are

$$\hat{\mu}_{j0} = \frac{\sum_{t=1}^T \sum_{i=1}^{N^1} \delta_{it} \zeta_{ij} (1 - \hat{\pi}_{it}) y_{it}^1}{\sum_{t=1}^T \sum_{i=1}^{N^1} \delta_{it} \zeta_{ij} (1 - \hat{\pi}_{it})}, \quad (53)$$

$$\hat{\mu}_{j1} = \frac{\sum_{t=1}^T \sum_{i=1}^{N^1} \delta_{it} \zeta_{ij} \hat{\pi}_{it} y_{it}^1}{\sum_{t=1}^T \sum_{i=1}^{N^1} \delta_{it} \zeta_{ij} \hat{\pi}_{it}}, \quad (54)$$

for all $j = 1, \dots, J$. The conditional MLE for the variance of the normal distributions is

$$\hat{\sigma}_j^2 = \frac{1}{N^1} \sum_{t=1}^T \sum_{i=1}^{N^1} \delta_{it} \left\{ \zeta_{ij} \left\{ (1 - \hat{\pi}_{it}) (y_{it}^1 - \mu_{j0})^2 + \hat{\pi}_{it} (y_{it}^1 - \mu_{j1})^2 \right\} \right\}. \quad (55)$$

The conditional MLE of the parameters for the macroeconomic variable are

$$\hat{\beta}_0^m = \frac{1}{T} \sum_{t=1}^T \mathbf{y}_t^m - \beta_1^m \alpha_{t|T}, \quad (56)$$

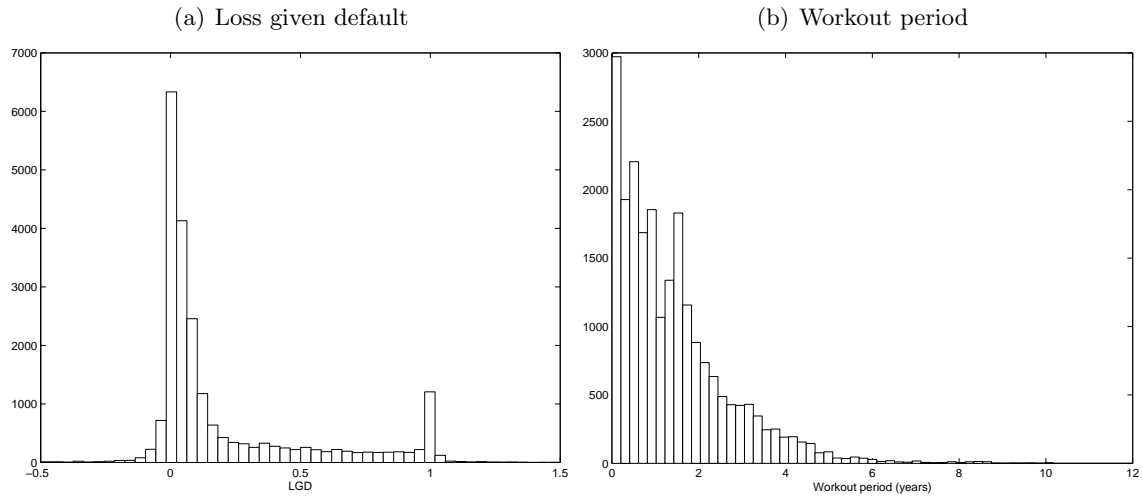
$$\hat{\beta}_1^m = \frac{1}{\hat{e}_0} \sum_{t=1}^T (\mathbf{y}_t^m - \beta_0^m) \alpha_{t|T}, \quad (57)$$

$$\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^T \left\{ \beta_1^m P_{t|T} (\beta_1^m)' + (\mathbf{y}_t^m - \beta_0^m - \beta_1^m \alpha_{t|T}) (\mathbf{y}_t^m - \beta_0^m - \beta_1^m \alpha_{t|T})' \right\}, \quad (58)$$

where $\hat{e}_0 = \sum_{t=1}^T P_{t|T} + \alpha_{t|T}^2$.

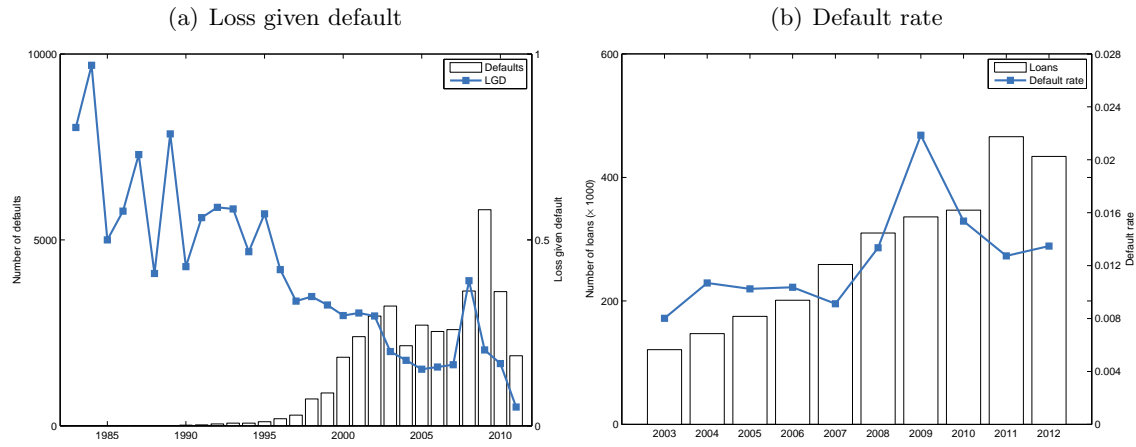
For the other parameters, the MLE cannot be solved analytically and need to be optimized numerically. We split the parameter space into independent subspaces over which we maximize. Hence, we optimize the expected loglikelihood (39) for AR coefficient ρ in the state equation, coefficients β_{j0}^1 and β_{j1}^1 and β_{i0}^d and β_{i1}^d separately.

Figure 1. Empirical distribution LGD and workout period



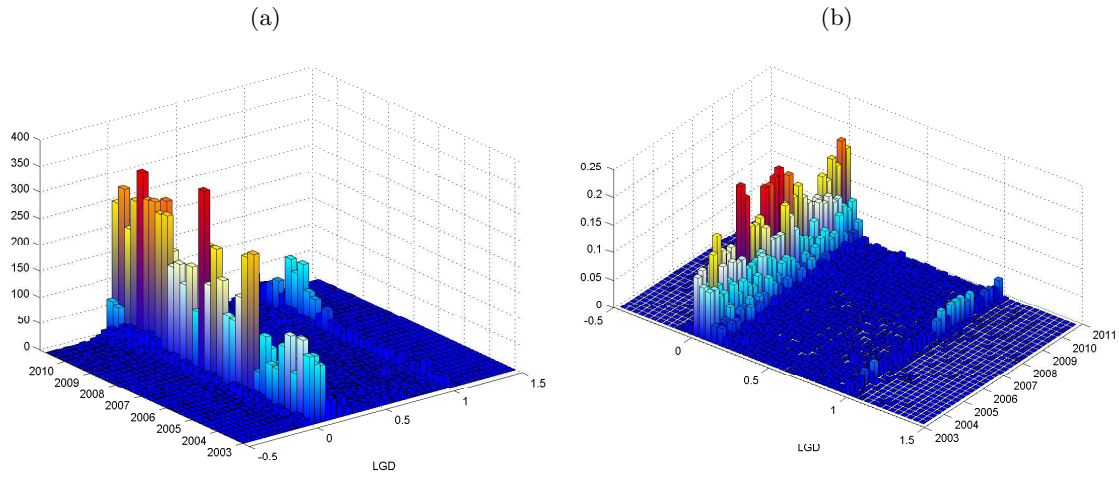
The figures show the empirical distribution of the loss given default (a) and the workout period (b) for the defaults from the period 2003–2010, after applying the data filter in appendix A.

Figure 2. Default data time series



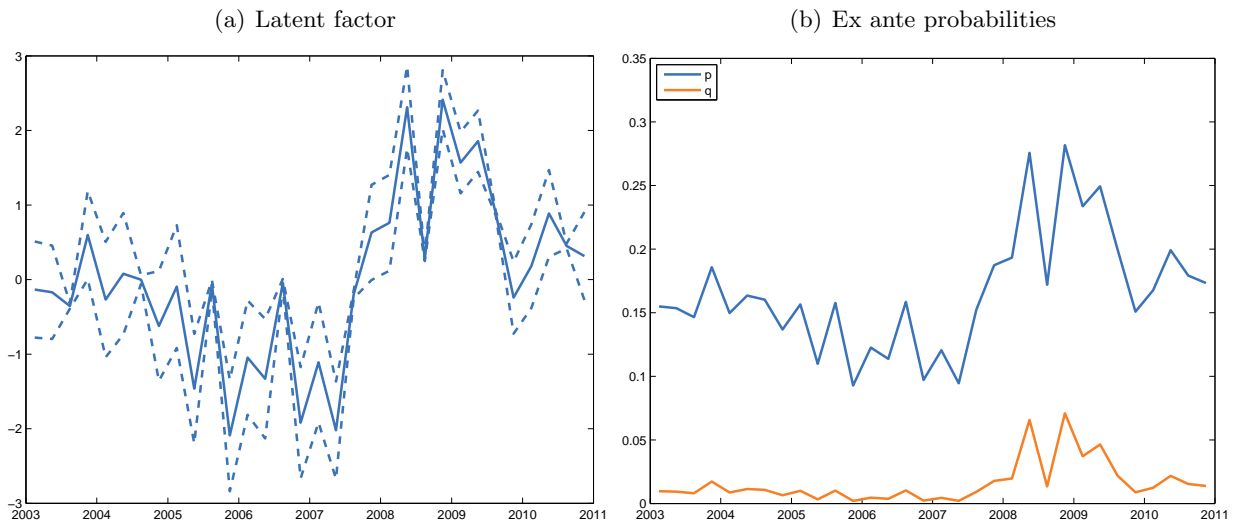
Panel a presents the average loss given default and the number of observations per year for the period 1983–2011 from the Global Credit Data LGD database, after applying the data filter in appendix A. Panel b presents the number of loans and the observed default rate per year for the period 2003–2012 from the Global Credit Data DR database.

Figure 3. Empirical distribution LGD over time



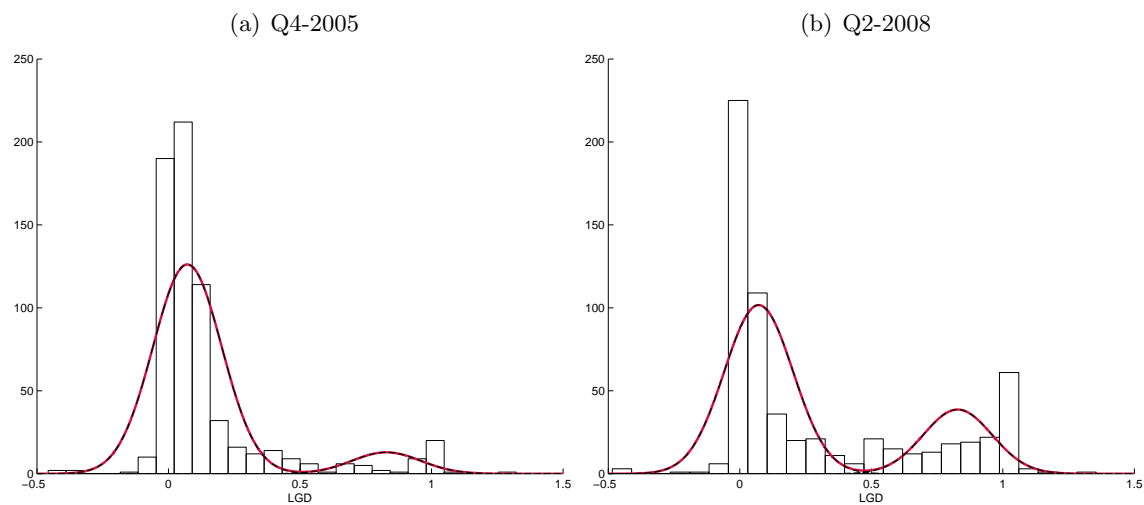
Panel a presents the empirical distribution of the LGD per quarter for the period 2003–2010 after applying the data filter in appendix A. Panel b presents the standardized empirical distribution, where every quarter is divided by the number of observations per period such that the distributions are comparable across time. It is rotated by 90 degrees compared to panel a.

Figure 4. Factor and ex ante probabilities



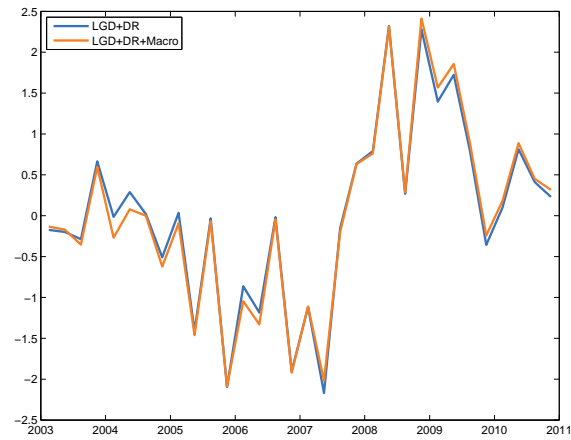
Panel a presents the smoothed factor α (solid line) with 95% confidence bounds (dashed lines) for the general model, without cross-sectional variation but including default rates and macroeconomic variables. Panel b presents the ex ante probabilities, defined as $\Lambda(\beta_0 + \beta_1 \alpha)$, where $\Lambda(x) = \exp(x)/(1 + \exp(x))$ is the logistic function. They are based on the smoothed factor α from panel a and the estimates from table III for the parameters β_0^l and β_1^l for the probability of bad loan p_t , and β_0^d and β_1^d for the default rate q_t .

Figure 5. Mixture fit



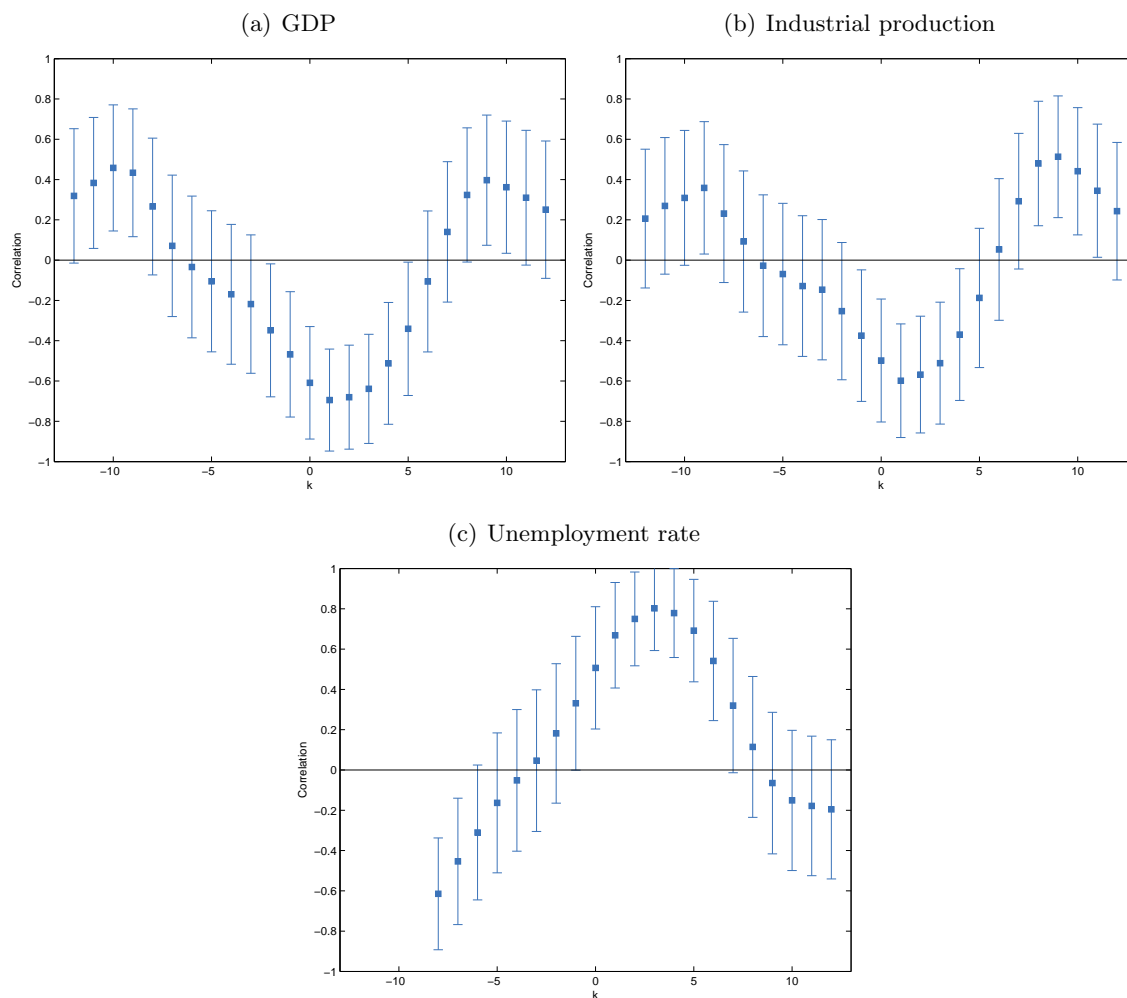
The figures present the fit of the mixture of normals for the fourth quarter of 2005 (a) and the second quarter of 2008 (b) for the model without cross-sectional variation but with default rates and macro variables.

Figure 6. Latent factor with and without macro variables



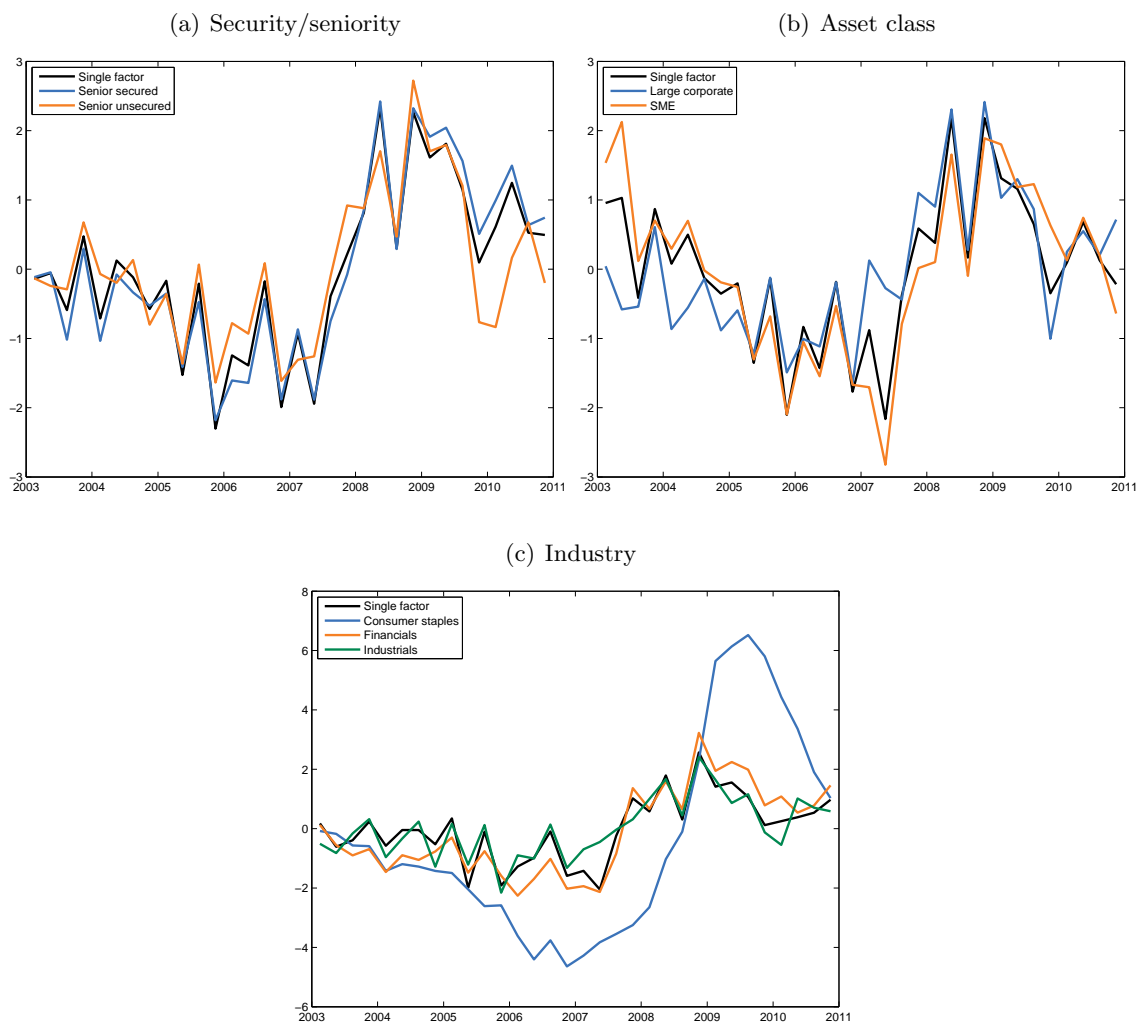
The figure presents the smoothed latent factor α , for the model with (orange line) and without (blue line) the macroeconomic variables GDP, industrial production and unemployment rate.

Figure 7. Correlation between factor and macroeconomic variables



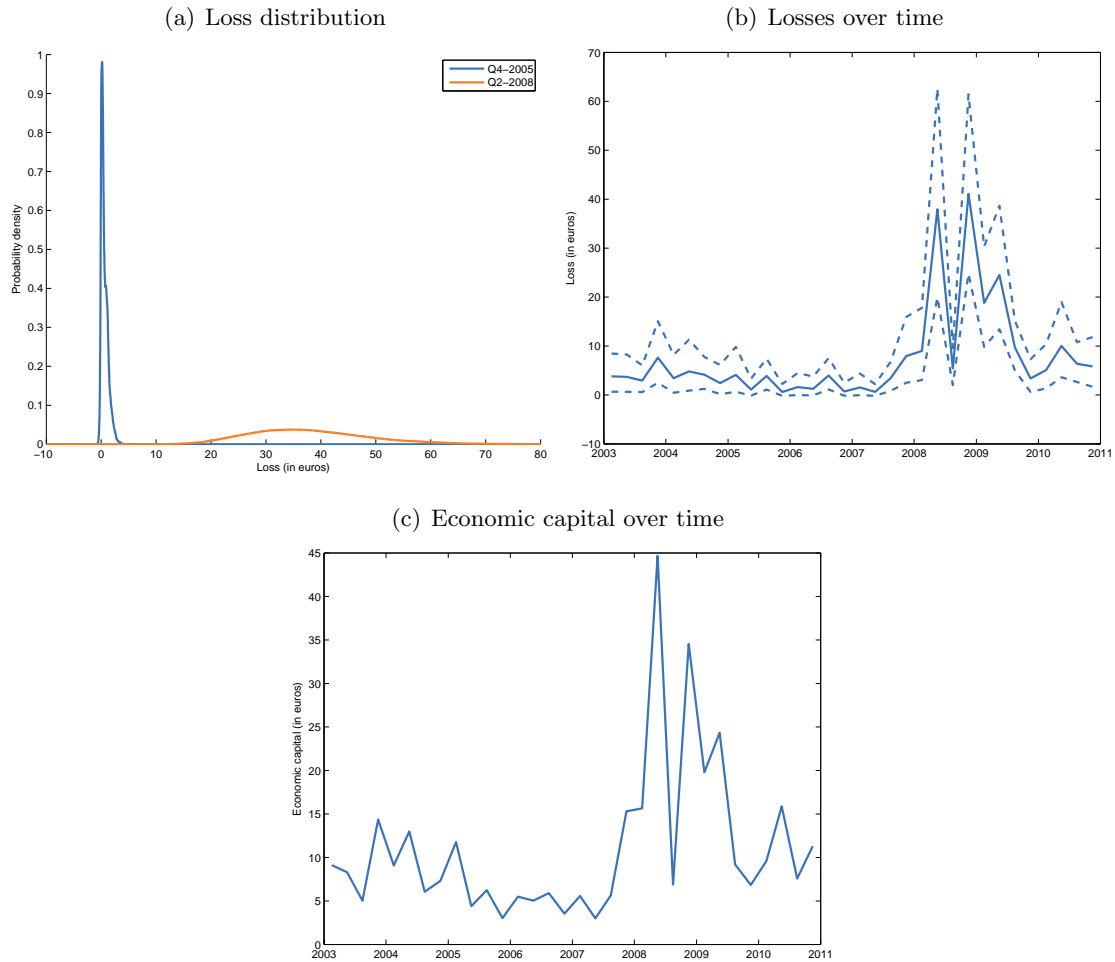
The figures present the correlation of α_t , the latent factor for the model with default rate and macroeconomic variables, including 95% confidence intervals, with y_{t+k}^m , the macroeconomic variables GDP (a), industrial production (b) and unemployment rate (c), all in difference to the same period in the previous year, for different leads and lags, k . The x -axis presents the leads and lags of the macro variable, k , such that positive k reflects the correlation between the factor and future macro conditions.

Figure 8. Factors per loan characteristic



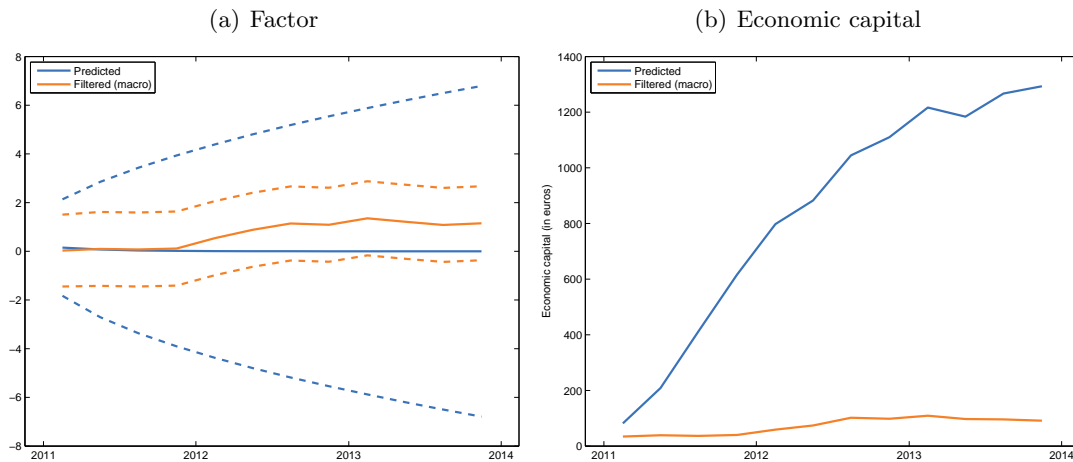
The figures present the factor per loan characteristic, where we have (i) a single factor underlying all categories with different coefficient or (ii) a different factor for category. Panel a shows the factor per seniority and security. Panel b shows the factor per asset classes large corporate (LC) and small and medium enterprises (SME). Panel c shows the factor per industries consumer staples (CS), financials (FIN) and industrials (IND).

Figure 9. Loss simulation



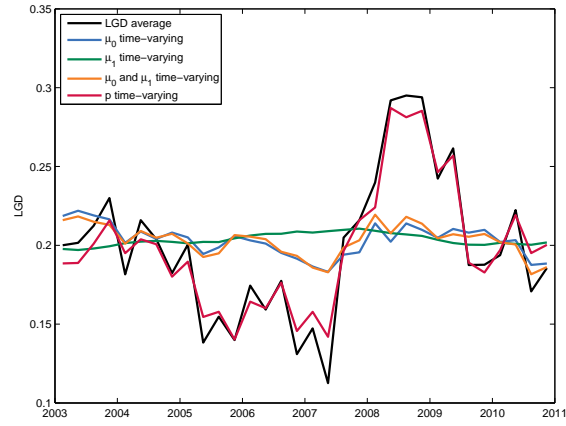
Panel a presents the loss distribution from simulating 50,000 times a portfolio of 2,000 loans, each with an EAD of €1, of the model defined by equations (1)–(6) for the fourth quarter of 2005 (blue line) and the second quarter of 2008 (orange line). Panel b presents the expected loss (solid line), including a 95% confidence interval (dashed lines), and panel c the economic capital at 99.9%.

Figure 10. Prediction



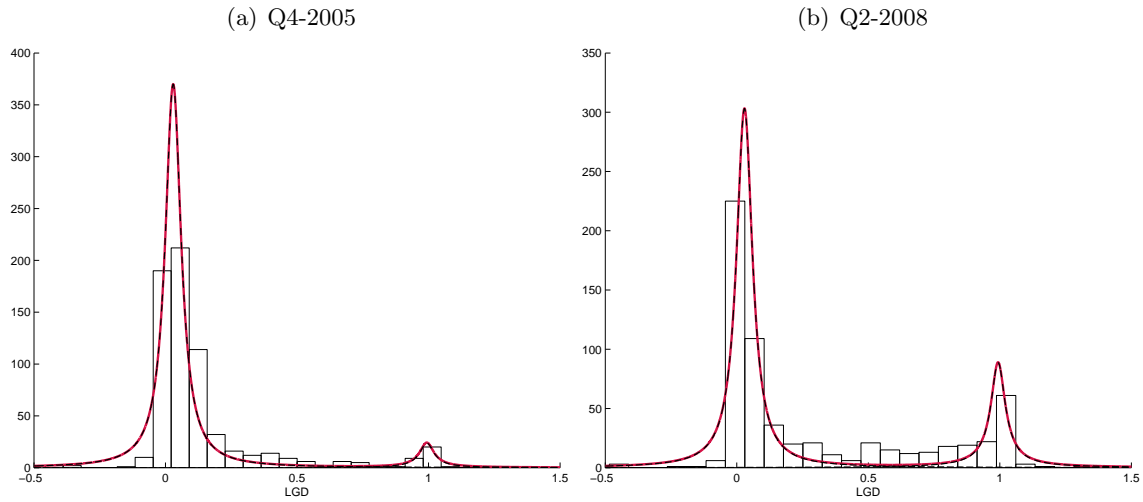
Panel a presents the predicted factor (solid line) for the period 2011–2013, including a 95% confidence interval (dashed lines), given the information up to and including 2010, time T , (blue line) and given the information up to and including 2010, time T , plus the macro variables at time $T+h$, where h is the forecast horizon (orange line). Panel b presents the economic capital at 99.9%, based on simulating 50,000 times a portfolio of 2,000 loans, each with an EAD of €1, given the predicted and macro-filtered factor in panel a.

Figure 11. Implied average LGD



The figure presents the average LGD per quarter and the average LGD implied by the estimation of the model defined by equations (1)–(3), with time-variation assumed in the probability of a bad loan p , mean of a good loan μ_0 and/or mean of a bad loan μ_1 .

Figure 12. Mixture fit: mixture of Student's t distributions



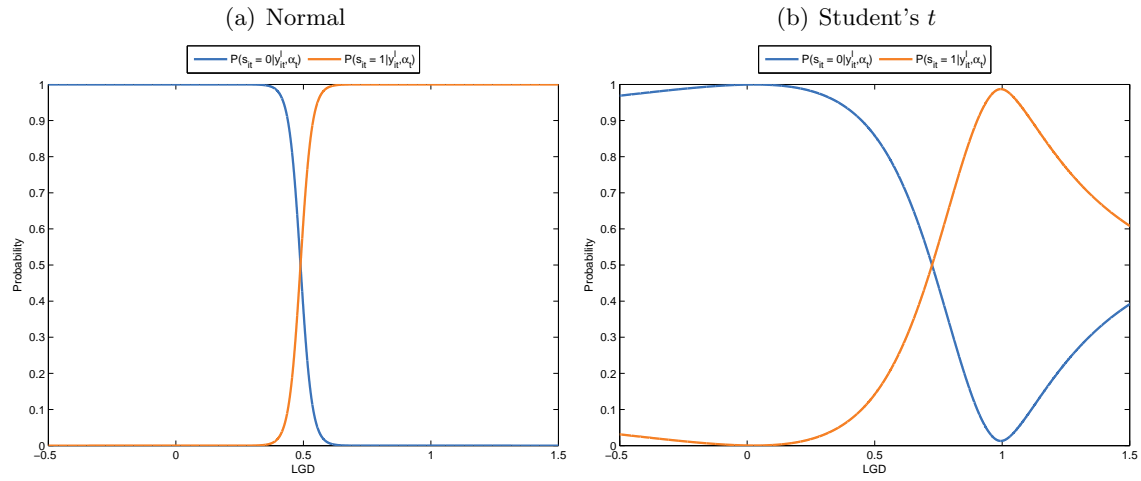
The figures present the fit for the fourth quarter of 2005 (a) and the second quarter of 2008 (b) for the model defined by equations (1)–(3), without the default rates and macro variables, with the mixture of normals replaced by a mixture of Student's t distributions.

Figure 13. Latent factor: normal versus Student's t distribution



The figure presents the latent factor for the model defined by equations (1)–(3), without the default rates and macro variables, with a mixture of normals (blue line) and with a mixture of Student's t distributions (orange line) for the LGD.

Figure 14. Smoothed state probability: normal versus Student's t distribution



The figures present the smoothed state probabilities $P(s_{it} = 0 | y_{it}^1, \alpha_t)$ (blue line) and $P(s_{it} = 1 | y_{it}^1, \alpha_t)$ (orange line) for the model defined by equations (1)–(3), with a mixture of normals (a) and with a mixture of Student's t distributions (b) for the LGD. The probabilities are for the first quarter of 2003.

Table I. LGD versus workout period

The table presents the number of defaults and the average LGD for different workout periods from the period 2003–2010, after applying the data filter in appendix A.

| Workout period (years) | Defaults | Average LGD |
|------------------------|----------|-------------|
| 0-1 | 10,464 | 0.119 |
| 1-2 | 6,258 | 0.232 |
| 2-3 | 2,794 | 0.284 |
| 3-5 | 2,208 | 0.383 |
| >5 | 356 | 0.432 |

Table II. Summary statistics

The table presents the number of defaults, the average, the fraction of defaults with an LGD larger than 0.5 and the p -value of the Hartigan and Hartigan's (1985) dip statistic (HDS) using 500 bootstraps, to test the null hypothesis of a unimodal distribution versus the alternative of a multimodal distribution, for different subsets of the 2003–2010 sample after applying the data filter in appendix A. Groups with on average at least 100 defaults per quarter, indicated by a *, are selected for analysis with our model in section 5.3.

| Group | Defaults | Average | Fraction LGD > 0.5 | HDS p -value |
|---------------------------------|----------|---------|-----------------------|-------------------|
| Total | 22,080 | 0.204 | 0.170 | 0.000 |
| Panel A: Seniority and security | | | | |
| Senior unsecured* | 12,011 | 0.222 | 0.191 | 0.000 |
| Senior secured* | 9,723 | 0.175 | 0.138 | 0.000 |
| Subordinated | 236 | 0.427 | 0.419 | 0.000 |
| Subordinated Secured | 110 | 0.289 | 0.255 | 0.002 |
| Panel B: Asset class | | | | |
| SME* | 12,028 | 0.193 | 0.164 | 0.000 |
| Large Corporate* | 6,496 | 0.199 | 0.159 | 0.000 |
| Real Estate Finance | 2,068 | 0.326 | 0.284 | 0.000 |
| Aircraft Finance | 556 | 0.088 | 0.045 | 0.000 |
| Shipping Finance | 331 | 0.077 | 0.054 | 0.100 |
| Project Finance | 302 | 0.177 | 0.132 | 0.002 |
| Banks | 276 | 0.286 | 0.286 | 0.000 |
| Public Services | 23 | 0.246 | 0.174 | 0.234 |
| Panel C: Industries | | | | |
| Industrials* | 6,944 | 0.178 | 0.150 | 0.000 |
| Financials* | 4,629 | 0.217 | 0.178 | 0.000 |
| Consumer Staples* | 3,232 | 0.186 | 0.162 | 0.000 |
| Unknown | 2,817 | 0.309 | 0.279 | 0.000 |
| Information Technology | 1,384 | 0.188 | 0.155 | 0.000 |
| Consumer Discretionary | 1,089 | 0.196 | 0.128 | 0.034 |
| Other | 606 | 0.147 | 0.102 | 0.000 |
| Telecommunication Services | 410 | 0.203 | 0.183 | 0.304 |
| Utilities | 391 | 0.145 | 0.079 | 0.280 |
| Health Care | 366 | 0.123 | 0.082 | 0.086 |
| Materials | 212 | 0.147 | 0.127 | 0.534 |

Table III. Parameter estimates

The table presents the parameter estimates for the model defined by equations (1)–(6). The standard errors are in parentheses next to the estimates. Panel A presents the parameter estimate of the factor component, the AR coefficient ρ . Panel B presents the parameter estimates of the LGD components, a mixture of two normals with the same variance σ^2 , for good (μ_0) and bad (μ_1) loans, where $\mu_0 < \mu_1$. The probability of a bad loan is given by $p_t = \Lambda(\beta_0^l + \beta_1^l \alpha_t)$, where $\Lambda(x) = \exp(x)/(1 + \exp(x))$ is the logistic function. Panel C presents parameter estimates of the default rate component where the number of defaults follows a binomial distribution with default probability $q_t = \Lambda(\beta_0^d + \beta_1^d \alpha_t)$. Panel D presents the parameter estimates of the macroeconomic component, the intercepts β_0 and the coefficients β_1 , with the variables gross domestic product (GDP), industrial production (IP) and unemployment rate (UR), all in difference to the same period in the previous year and standardized to have zero mean and unit variance. Panel E presents the estimates of the the intercepts β_0 and the coefficients β_1 of the financial component, with the variables long-term interest rate (LIR), credit spread (CRS) and the yield spread (YLS). Panel F presents the mean marginal effects, defined as the average over the marginal effects $\partial\Lambda(\beta_0 + \beta_1\alpha_t)/\partial\beta_1$, for all $t = 1, \dots, T$, for the probability of a bad loan p and the probability of default q . Finally, the bottom of the table presents the loglikelihood and the number of observations, given by the sum of the LGD, default rate, macroeconomic and financial observations.

| Parameter | LGD + DR + Macro | | LGD + DR | | LGD + DR + Macro + Financial | |
|--------------------------------|---------------------|---------|----------|---------|------------------------------------|---------|
| Panel A: Factor | | | | | | |
| ρ | 0.484 | (0.162) | 0.449 | (0.167) | 0.524 | (0.156) |
| Panel B: Loss given default | | | | | | |
| μ_0 | 0.072 | (0.001) | 0.072 | (0.001) | 0.072 | (0.001) |
| μ_1 | 0.828 | (0.002) | 0.828 | (0.002) | 0.828 | (0.002) |
| σ | 0.131 | (0.001) | 0.131 | (0.001) | 0.131 | (0.001) |
| β_0^l | -1.656 | (0.102) | -1.652 | (0.100) | -1.657 | (0.107) |
| β_1^l | 0.299 | (0.044) | 0.311 | (0.045) | 0.291 | (0.043) |
| Panel C: Default rate | | | | | | |
| β_0^d | -4.526 | (0.284) | -4.545 | (0.310) | -4.505 | (0.279) |
| β_1^d | 0.809 | (0.181) | 0.931 | (0.248) | 0.743 | (0.149) |
| Panel D: Macro variables | | | | | | |
| β_0^{GDP} | -0.005 | (0.218) | | | -0.003 | (0.226) |
| β_1^{GDP} | -0.499 | (0.070) | | | -0.491 | (0.067) |
| β_0^{IP} | -0.004 | (0.204) | | | -0.002 | (0.210) |
| β_1^{IP} | -0.408 | (0.059) | | | -0.403 | (0.057) |
| β_0^{UR} | 0.004 | (0.205) | | | 0.003 | (0.211) |
| β_1^{UR} | 0.415 | (0.060) | | | 0.411 | (0.058) |
| Panel E: Financial variables | | | | | | |
| β_0^{LIR} | | | | | 0.002 | (0.194) |
| β_1^{LIR} | | | | | 0.293 | (0.043) |
| β_0^{CRS} | | | | | 0.004 | (0.243) |
| β_1^{CRS} | | | | | 0.579 | (0.075) |
| β_0^{YLS} | | | | | 0.001 | (0.176) |
| β_1^{YLS} | | | | | 0.092 | (0.014) |
| Panel F: Mean marginal effects | | | | | | |
| p | 0.041 | | 0.042 | | 0.039 | |
| q | 0.012 | | 0.015 | | 0.011 | |
| Loglikelihood | 3,884 | | 3,940 | | 3,825 | |
| Observations | 22,184 | | 22,088 | | 22,280 | |

Table IV. Parameter estimates: security/seniority

The table presents the parameter estimates for the model defined by equations (1)–(6) for senior secured (1) and senior unsecured (2) loans, where we have (i) a single factor underlying all categories with different coefficient or (ii) a different factor for category. The standard errors are in parentheses next to the estimates. Panel A presents the parameter estimate of the factor component, the AR coefficient ρ . Panel B presents the parameter estimates of the LGD components, a mixture of two normals with the same variance σ_j^2 , for good (μ_{j0}) and bad (μ_{j1}) loans, where $\mu_{j0} < \mu_{j1}$, for all groups $j = 1, \dots, J$. The probability of a bad loan is given by $p_{jt} = \Lambda(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)$, where $\Lambda(x) = \exp(x)/(1 + \exp(x))$ is the logistic function. Panel C presents parameter estimates of the default rate component where the number of defaults follows a binomial distribution with default probability $q_t = \Lambda(\beta_0^d + \beta_1^d \alpha_t)$. Panel D presents the parameter estimates of the macroeconomic component, the intercepts β_0 and the coefficients β_1 , with the variables gross domestic product (GDP), industrial production (IP) and unemployment rate (UR), all in difference to the same period in the previous year and standardized to have zero mean and unit variance. Panel E presents the mean marginal effects, defined as the average over the marginal effects $\partial \Lambda(\beta_0 + \beta_1 \alpha_t) / \partial \beta_1$, for all $t = 1, \dots, T$, for the probabilities of a bad loan p_j and the probability of default q . Finally, the bottom of the table presents the loglikelihood and the number of observations, given by the sum of the LGD, default rate and macroeconomic observations.

| Parameter | Single factor | | Senior secured | | Senior unsecured | |
|--------------------------------|---------------|---------|----------------|---------|------------------|---------|
| Panel A: Factor | | | | | | |
| ρ | 0.529 | (0.157) | 0.624 | (0.144) | 0.478 | (0.184) |
| Panel B: Loss given default | | | | | | |
| μ_{10} | 0.070 | (0.002) | 0.070 | (0.002) | | |
| μ_{11} | 0.766 | (0.004) | 0.765 | (0.004) | | |
| σ_1 | 0.129 | (0.001) | 0.129 | (0.001) | | |
| β_{10}^1 | -1.927 | (0.184) | -1.937 | (0.206) | | |
| β_{11}^1 | 0.504 | (0.075) | 0.452 | (0.071) | | |
| μ_{20} | 0.071 | (0.001) | | | 0.071 | (0.001) |
| μ_{21} | 0.858 | (0.003) | | | 0.858 | (0.003) |
| σ_2 | 0.130 | (0.001) | | | 0.130 | (0.001) |
| β_{20}^1 | -1.501 | (0.067) | | | -1.505 | (0.065) |
| β_{21}^1 | 0.172 | (0.005) | | | 0.182 | (0.040) |
| Panel C: Default rate | | | | | | |
| β_0^d | -4.463 | (0.214) | -4.403 | (0.177) | -4.618 | (0.263) |
| β_1^d | 0.575 | (0.076) | 0.383 | (0.030) | 0.676 | (0.164) |
| Panel D: Macro variables | | | | | | |
| β_0^{GDP} | -0.013 | (0.225) | -0.016 | (0.259) | 0.004 | (0.217) |
| β_1^{GDP} | -0.505 | (0.068) | -0.497 | (0.060) | -0.501 | (0.076) |
| β_0^{IP} | -0.010 | (0.206) | -0.012 | (0.229) | 0.004 | (0.210) |
| β_1^{IP} | -0.395 | (0.055) | -0.384 | (0.049) | -0.454 | (0.070) |
| β_0^{UR} | 0.012 | (0.218) | 0.015 | (0.254) | -0.003 | (0.193) |
| β_1^{UR} | 0.461 | (0.063) | 0.477 | (0.059) | 0.322 | (0.053) |
| Panel E: Mean marginal effects | | | | | | |
| p_1 | 0.058 | | 0.052 | | | |
| p_2 | 0.026 | | | | 0.027 | |
| q | 0.008 | | 0.005 | | 0.008 | |
| Loglikelihood | 4,208 | | 2,317 | | 1,763 | |
| Observations | 21,838 | | 9,827 | | 12,115 | |

Table V. Parameter estimates: asset class

The table presents the parameter estimates for the model defined by equations (1)–(6) for loans of asset classes large corporate (LC, 1) and small and medium enterprises (SME, 2), where we have (i) a single factor underlying all categories with different coefficient or (ii) a different factor for category. The standard errors are in parentheses next to the estimates. Panel A presents the parameter estimate of the factor component, the AR coefficient ρ . Panel B presents the parameter estimates of the LGD components, a mixture of two normals with the same variance σ_j^2 , for good (μ_{j0}) and bad (μ_{j1}) loans, where $\mu_{j0} < \mu_{j1}$, for all groups $j = 1, \dots, J$. The probability of a bad loan is given by $p_{jt} = \Lambda(\beta_{j0}^1 + \beta_{j1}^1 \alpha_t)$, where $\Lambda(x) = \exp(x)/(1 + \exp(x))$ is the logistic function. Panel C presents parameter estimates of the default rate component where the number of defaults follows a binomial distribution with default probability $q_{it} = \Lambda(\beta_{i0}^d + \beta_{i1}^d \alpha_t)$, for all groups $i = 1, \dots, N^d$. Panel D presents the parameter estimates of the macroeconomic component, the intercepts β_0 and the coefficients β_1 , with the variables gross domestic product (GDP), industrial production (IP) and unemployment rate (UR), all in difference to the same period in the previous year and standardized to have zero mean and unit variance. Panel E presents the mean marginal effects, defined as the average over the marginal effects $\partial\Lambda(\beta_0 + \beta_1\alpha_t)/\partial\beta_1$, for all $t = 1, \dots, T$, for the probabilities of a bad loan p_j and the probabilities of default q_i . Finally, the bottom of the table presents the loglikelihood and the number of observations, given by the sum of the LGD, default rate and macroeconomic observations.

| Parameter | Single factor | | Large corporate | | SME | |
|--------------------------------|---------------|---------|-----------------|---------|--------|---------|
| Panel A: Factor | | | | | | |
| ρ | 0.406 | (0.177) | 0.358 | (0.200) | 0.604 | (0.163) |
| Panel B: Loss given default | | | | | | |
| μ_{10} | 0.075 | (0.002) | 0.075 | (0.002) | | |
| μ_{11} | 0.849 | (0.005) | 0.849 | (0.005) | | |
| σ_1 | 0.126 | (0.001) | 0.126 | (0.001) | | |
| β_{10}^1 | -1.778 | (0.094) | -1.788 | (0.094) | | |
| β_{11}^1 | 0.289 | (0.053) | 0.312 | (0.059) | | |
| μ_{20} | 0.062 | (0.001) | | | 0.062 | (0.001) |
| μ_{21} | 0.849 | (0.003) | | | 0.849 | (0.003) |
| σ_2 | 0.124 | (0.001) | | | 0.124 | (0.001) |
| β_{20}^1 | -1.643 | (0.093) | | | -1.635 | (0.123) |
| β_{21}^1 | 0.305 | (0.014) | | | 0.282 | (0.050) |
| Panel C: Default rate | | | | | | |
| β_{10}^d | -2.676 | (0.296) | -2.706 | (0.230) | | |
| β_{11}^d | 0.948 | (0.271) | 0.739 | (0.173) | | |
| β_{20}^d | -6.690 | (0.237) | | | -6.665 | (0.182) |
| β_{21}^d | 0.754 | (0.178) | | | 0.405 | (0.053) |
| Panel D: Macro variables | | | | | | |
| β_0^{GDP} | -0.006 | (0.202) | -0.005 | (0.192) | -0.015 | (0.244) |
| β_1^{GDP} | -0.469 | (0.071) | -0.414 | (0.068) | -0.469 | (0.061) |
| β_0^{IP} | -0.005 | (0.194) | -0.004 | (0.185) | -0.013 | (0.226) |
| β_1^{IP} | -0.388 | (0.060) | -0.328 | (0.055) | -0.397 | (0.053) |
| β_0^{UR} | 0.005 | (0.195) | 0.003 | (0.180) | 0.015 | (0.245) |
| β_1^{UR} | 0.399 | (0.062) | 0.243 | (0.043) | 0.473 | (0.061) |
| Panel E: Mean marginal effects | | | | | | |
| p_1 | 0.036 | | 0.039 | | | |
| p_2 | 0.042 | | | | 0.039 | |
| q_1 | 0.072 | | 0.050 | | | |
| q_2 | 0.001 | | | | 0.001 | |
| Loglikelihood | 4,287 | | 1,435 | | 2,801 | |
| Observations | 18,636 | | 6,600 | | 12,132 | |

Table VI. Parameter estimates: industry

The table presents the parameter estimates for the model defined by equations (1)–(6) for loans of industries consumer staples (CS, 1), financials (FIN, 2) and industrials (IND, 3), where we have (i) a single factor underlying all categories with different coefficient or (ii) a different factor for category. The standard errors are in parentheses next to the estimates. Panel A presents the parameter estimate of the factor component, the AR coefficient ρ . Panel B presents the parameter estimates of the LGD components, a mixture of two normals with the same variance σ_j^2 , for good (μ_{j0}) and bad (μ_{j1}) loans, where $\mu_{j0} < \mu_{j1}$, for all groups $j = 1, \dots, J$. The probability of a bad loan is given by $p_{jt} = \Lambda(\beta_{j0}^l + \beta_{j1}^l \alpha_t)$, where $\Lambda(x) = \exp(x)/(1 + \exp(x))$ is the logistic function. Panel C presents parameter estimates of the default rate component where the number of defaults follows a binomial distribution with default probability $q_{it} = \Lambda(\beta_{i0}^d + \beta_{i1}^d \alpha_t)$, for all groups $i = 1, \dots, N^d$. Panel D presents the parameter estimates of the macroeconomic component, the intercepts β_0 and the coefficients β_1 , with the variables gross domestic product (GDP), industrial production (IP) and unemployment rate (UR), all in difference to the same period in the previous year and standardized to have zero mean and unit variance. Panel E presents the mean marginal effects, defined as the average over the marginal effects $\partial\Lambda(\beta_0 + \beta_1\alpha_t)/\partial\beta_1$, for all $t = 1, \dots, T$, for the probabilities of a bad loan p_j and the probabilities of default q_i . Finally, the bottom of the table presents the loglikelihood and the number of observations, given by the sum of the LGD, default rate and macroeconomic observations.

| Parameter | Single factor | | Consumer staples | | Financials | | Industrials | |
|--------------------------------|---------------|---------|------------------|---------|------------|---------|-------------|---------|
| Panel A: Factor | | | | | | | | |
| ρ | 0.491 | (0.170) | 0.933 | (0.048) | 0.742 | (0.133) | 0.373 | (0.234) |
| Panel B: Loss given default | | | | | | | | |
| μ_{10} | 0.056 | (0.002) | 0.056 | (0.002) | | | | |
| μ_{11} | 0.851 | (0.006) | 0.851 | (0.006) | | | | |
| σ_1 | 0.120 | (0.002) | 0.120 | (0.002) | | | | |
| β_{10}^l | -1.662 | (0.106) | -1.642 | (0.113) | | | | |
| β_{11}^l | 0.276 | (0.056) | 0.052 | (0.017) | | | | |
| μ_{20} | 0.085 | (0.003) | | | 0.084 | (0.003) | | |
| μ_{21} | 0.796 | (0.006) | | | 0.795 | (0.006) | | |
| σ_2 | 0.144 | (0.002) | | | 0.144 | (0.002) | | |
| β_{20}^l | -1.796 | (0.190) | | | -1.824 | (0.265) | | |
| β_{21}^l | 0.539 | (0.049) | | | 0.407 | (0.090) | | |
| μ_{30} | 0.056 | (0.002) | | | | | 0.056 | (0.002) |
| μ_{31} | 0.836 | (0.004) | | | | | 0.836 | (0.004) |
| σ_3 | 0.119 | (0.001) | | | | | 0.119 | (0.001) |
| β_{30}^l | -1.773 | (0.092) | | | | | -1.781 | (0.089) |
| β_{31}^l | 0.248 | (0.012) | | | | | 0.288 | (0.070) |
| Panel C: Default rate | | | | | | | | |
| β_{10}^d | -4.561 | (0.126) | -4.484 | (0.087) | | | | |
| β_{11}^d | 0.346 | (0.035) | 0.044 | (0.000) | | | | |
| β_{20}^d | -5.311 | (0.393) | | | -5.120 | (0.314) | | |
| β_{21}^d | 1.092 | (0.328) | | | 0.481 | (0.061) | | |
| β_{30}^d | -4.268 | (0.301) | | | | | -4.463 | (0.336) |
| β_{31}^d | 0.835 | (0.192) | | | | | 0.937 | (0.352) |
| Panel D: Macro variables | | | | | | | | |
| β_0^{GDP} | -0.016 | (0.219) | -0.104 | (0.495) | -0.062 | (0.315) | -0.001 | (0.198) |
| β_1^{GDP} | -0.498 | (0.070) | -0.248 | (0.013) | -0.452 | (0.048) | -0.467 | (0.076) |
| β_0^{IP} | -0.013 | (0.204) | -0.083 | (0.410) | -0.050 | (0.273) | -0.001 | (0.190) |
| β_1^{IP} | -0.402 | (0.058) | -0.199 | (0.012) | -0.362 | (0.040) | -0.376 | (0.063) |
| β_0^{UR} | 0.014 | (0.207) | 0.123 | (0.574) | 0.056 | (0.292) | 0.001 | (0.185) |
| β_1^{UR} | 0.419 | (0.061) | 0.293 | (0.012) | 0.404 | (0.044) | 0.319 | (0.056) |
| Panel E: Mean marginal effects | | | | | | | | |
| p_1 | 0.037 | | 0.007 | | | | | |
| p_2 | 0.068 | | | | 0.049 | | | |
| p_3 | 0.031 | | | | | | 0.036 | |
| q_1 | 0.004 | | 0.000 | | | | | |
| q_2 | 0.010 | | | | 0.013 | | | |
| q_3 | 0.016 | | | | | | 0.016 | |
| Loglikelihood | 3,145 | | 781 | | 329 | | 1,922 | |
| Observations | 14,925 | | 3,336 | | 4,733 | | 7,048 | |

Table VII. Parameter estimates: LGD model alternatives

The table presents the parameter estimates for the model defined by equations (1)–(3), for alternative versions of the LGD distribution. The standard errors are in parentheses next to the estimates. Panel A presents the parameter estimate of the factor component, the AR coefficient ρ . Panel B presents the parameter estimates of the LGD components, a mixture of two distributions with the same variance σ^2 , for good (μ_0) and bad (μ_1) loans, where $\mu_0 < \mu_1$. The probability of a bad loan is given by p . Finally, the bottom of the table presents the loglikelihood, the Bayesian information criterion (BIC) and the number of LGD observations. The following distributions are considered: (a) a mixture of normals with time-varying probability of a bad loan $p_t = \Lambda(\beta_0^1 + \beta_1^1 \alpha_t)$, where $\Lambda(x) = \exp(x)/(1 + \exp(x))$ is the logistic function; (b) a mixture of normals without time-varying parameters; (c) a mixture of normals with time-varying mean of a good loan $\mu_{0t} = \mu_0 + \beta_1^1 \alpha_t$; (d) a mixture of normals with time-varying mean of a bad loan $\mu_{1t} = \mu_1 + \beta_1^1 \alpha_t$; (e) a mixture of normals with equal time-variation in the means of good and bad loans $\mu_{0t} = \mu_0 + \beta_1^1 \alpha_t$ and $\mu_{1t} = \mu_0 + \beta_1^1 \alpha_t$, such that the location of the distribution is time-varying; (f) a mixture of Student's t distributions with degrees of freedom ν_0 and ν_1 and time-varying probability of a bad loan $p_t = \Lambda(\beta_0^1 + \beta_1^1 \alpha_t)$.

| Parameter | (a) p_t | (b) No time variation | (c) μ_{0t} | (d) μ_{1t} | (e) $\mu_{0t} + \mu_{1t}$ | (f) Student's t |
|-----------------------------|----------------|-----------------------|----------------|----------------|---------------------------|-------------------|
| Panel A: Factor | | | | | | |
| ρ | 0.706 (0.131) | | 0.634 (0.194) | 0.874 (0.086) | 0.631 (0.192) | 0.724 (0.127) |
| Panel B: Loss given default | | | | | | |
| p | | 0.174 (0.003) | 0.173 (0.003) | 0.174 (0.003) | 0.173 (0.003) | |
| μ_0 | 0.072 (0.001) | 0.072 (0.001) | 0.072 (0.005) | 0.072 (0.001) | 0.073 (0.004) | 0.029 (0.000) |
| μ_1 | 0.829 (0.002) | 0.829 (0.003) | 0.830 (0.003) | 0.828 (0.012) | 0.829 (0.005) | 0.994 (0.001) |
| σ | 0.131 (0.001) | 0.131 (0.001) | 0.130 (0.001) | 0.131 (0.001) | 0.130 (0.001) | 0.037 (0.000) |
| ν_0 | | | | | | 1.013 (0.016) |
| ν_1 | | | | | | 0.688 (0.021) |
| β_0^1 | -1.655 (0.148) | | 0.010 (0.002) | 0.011 (0.003) | 0.009 (0.002) | -1.999 (0.171) |
| β_1^1 | 0.261 (0.043) | | | | | 0.284 (0.047) |
| Loglikelihood | 4,043 | 3,834 | 3,893 | 3,864 | 3,880 | 9,350 |
| BIC | -7,706 | -7,627 | -7,406 | -7,348 | -7,380 | -18,299 |
| Observations | 22,080 | 22,080 | 22,080 | 22,080 | 22,080 | 22,080 |