Testing Observable and Latent Risk Factor Models for Systematic Credit Risk on a Large Loan Level Data Set of Residential Mortgages

Felix Jose Farias Fueyo

Supervisor: (Bart) B.J.L. Keijsers MSc

Erasmus University

December 14, 2017

Abstract

This paper compares the performance of two models for probability of default of residential mortgages, namely a two-step logistic regression model with observable systematic risk factor and a mixed-measurement observation driven generalized autoregressive score (GAS) model. The data set is of residential mortgages set available by the Federal National Mortgage Association. The comparison of the models is realized in two different settings, namely a setting where the data set covers a complete economic cycle and a setting where the data set only covers parts of the economic cycle. Specifically, this paper focuses on the implications of restrictions in the length of loan level observations for the estimation of the two models. This is specially of interest since the length of loan level data sets might not fully cover economic cycles. Thus, resulting in biased model estimates. The results obtained in this paper indicate that the length of the data set plays a significant role defining the maximum likelihood estimates of both models. Furthermore, it is shown that the mix-measurement observation driven GAS model provides a very flexible framework capable of incorporating all futures of the two-step logit model with observable risk factor. By comparing the out-of-sample forecasting performance of both models it is shown that the GAS model can perform significantly better in periods of economic distress. This paper provides a comprehensive and detailed description of how the mix-measurement observation driven GAS model can be applied to estimate the probability of default of residential mortgages. Thus, providing financial institutions with an additional and rather accurate tool to estimate regulatory capital of retail portfolios.

Key words: capital requirements, probability of default, hazard rate models, latent risk factors, observable risk factors, maximum likelihood estimation, point-in-time probability of default, retail portfolios, residential mortgages, generalized autoregressive score models.

Acknowledgements

Few are the occasions in a lifetime when it is possible to demonstrate those who have offered their unconditional support how much they are appreciated. Considering the conclusion of this master program as an achievement following the efforts of seven years of work from the moment I left my hometown in Mexico, I can only thank to each and every single person that supported me during the last seven years. Specially, to my parents, who have done an immense sacrifice by letting me go so far from home early in my life, to my siblings, who I love so much, to my uncles Claus and Socorro, who have provided me with their unconditional support throughout all these years, to the foundation of the Aloisiuskolleg, for giving me a life-changing opportunity, specially to Dr. Christopher Haep, who I believe is an extraordinary human being, and to Andreas Unger, who believed in me and opened for me the doors into the German education system. Without these and many more persons I would not be where I am today. Also, I sincerely thank my supervisor, Bart Keijsers. Without his valuable comments and insights the quality of the work here presented would have been a different one. Finally, I thank Open Source Investor Services B.V. for their support throughout the period in which I wrote this thesis.

Table of Contents

1 Introduction

After the economic crisis of 2008, credit risk modeling has become increasingly important for large financial institutions. Especially the modeling of two variables is of interest: probability of default (PD) and loss given default (LGD). These two terms are largely related to each other. Probability of default is the financial term assigned to the probability that a credit borrower will be unwilling or unable to repay its debt in full or on time and loss given default is the share of an asset that is lost if a borrower defaults. Under Basel II guidelines, banks are allowed to use their own estimated risk parameters — probability of default, loss given default, exposure at default (EAD), maturity (M)- for the purpose of calculating regulatory capital. Regulatory capital is the the amount of capital a bank or other financial institution has to hold as required by its financial regulator. It is usually expressed as a capital adequacy ratio of equity that must be held as a percentage of risk-weighted assets.

When a bank uses its own estimated risk parameters to determine regulatory capital, then it is said the bank applies an internal ratings-based (IRB) approach to capital requirements for credit risk. In order to use this approach, a bank must take two major steps: (1) categorize their exposures into various asset classes as defined by the Basel II accord and (2) estimate the risk parameters. Each banking exposure is categorized into one of the following asset classes: corporate, sovereign, bank, retail and equity. Retail exposures are further categorized into residential mortgage, qualifying revolving exposure and other retail. The research presented in this paper focuses on the modeling of probabilities of default from residential mortgages.

Modeling a borrower's probability of default plays an important role in credit risk management. In the majority of cases, the information available to estimate PD can be divided in two categories, namely macroeconomic information and borrower specific information. Macroeconomic information includes all macroeconomic variables that could potentially influence a borrower's ability to pay back its debt, such as GDP, unemployment rates, house price indexes, inflation rates, etc. Obliger specific information refers to the characteristics of the borrower and the loan such as the income of the borrower at the moment of loan origination, the interest on the loan, size of the loan, etc.

Depending on the information used to estimate PD, the estimates are commonly referred to as through-the-cycle (TTC) or point-int-time (PIT) PD. A PIT PD is an estimate that the obligor will default over a particular time horizon considering the current macroeconomic conditions as well as time-varying or static obligor specific information. The main purpose of a PIT PD is to determine a borrower's probability of default at a specific point in an economic cycle. A TTC PD is an estimate that the obligor will default over a particular time horizon considering the current obligor specific information, irrespective of the current state of the economy.

While the interest in modeling of credit risk has increased after the current financial crisis, one of the earliest analyses of probability of default or hazard rates on residential or corporate loans that could be found is a paper by Deng (1997). Deng estimates a Cox proportional hazard rate model with competing risks and time-varying covariates using a data set of 1,489,732 observations on single family mortgage loans issued between 1976 and 1983. Deng concludes that the loan-to-value (LTV) of residential mortgages is an important risk driver for prepayments and defaults. More resent applications of hazard rate models to investigate the probability of default of retail loans includes the work of Donchev (2008) and Bellotti & Crook (2014).

Alongside models with observable systematic (macroeconomic) risk factors as the ones described by Koyluoglu & Hickman (1998) and Hamerle et al. (2003), other models have introduced the possibility of the PD being affected by latent risk factors. This is the case of the models applied in the work of Lee et al. (2016), Creal et al. (2014) and Keijsers et al. (2017). The main difference between these two modeling approaches is based in the information used to estimate the probabilities of default. On the one hand, observable risk factor models assume that the necessary information to estimate PD is observable and readily accessible. On the other hand, unobservable risk factor models assume the estimation of accurate PD requires the use of information that is not completely captured by observable risk factors and, therefore, is latent. Fitting latent risk factor models generally requires the use of simulation methods to compute integrals and is more computational intensive. Observable risk factor models are calibrated without the need of simulation and are less computational intensive.

In the after math of the financial crisis a special effort has been set to create databases that could improve the understanding of retail portfolios. An example is the initiative set by the European Data Warehouse, which is the first central data warehouse in Europe for collecting, validating and making available for download detailed, standardized and asset class specific loan level data for asset-backed securities transactions. However, many of these initiatives are quite recent and, in the case of loan level data from residential mortgages, the available databases tend to be rather short. A problem arising from short data sets is that they might not span whole economic cycles and, therefore, model estimates obtained using this data can be biased. For example, if a model is calibrated using data corresponding to a recession or crisis, such as the crisis of 2008, then the parameter estimates will present an upward bias due to the large amount of defaults observed within the sample. Recently published data from Fannie Mae and Fannie Mac offer the possibility to analyze long loan level data sets. The database of Fannie Mae is composed by loan level observations of single family residential mortgages from 2000 to 2016.

This paper complements and extends the available literature on PD modeling for residential mortgages by comparing two alternative models and by evaluating the bias introduced by the length of the data set on the estimates generated by the models. In the literature review for this paper, no research has been found that explicitly investigates the bias introduced by the length of loan level data sets. The models compared are inspired by the two-step logistic regression model with latent factor presented by Lee et al. (2016) and the mix-measurement generalized autoregressive score model presented by Creal et al. (2014). The first model considered is a simplification of the model by Lee et al., where it is assumed that all systematic risk can be captured by a single observable systematic risk factor. Therefore, the model observed is a two-step logistic regression PD model with observable risk factor, where the second step transforms TTC PDs to PIT PDs. The second model consists of an adaptation of the mix-measurement observation driven dynamic factor model from Creal et al. (2014) to the context of PD modeling of residential mortgages.

In their paper, Creal et al. use the mix-measurement GAS model to estimate rating transition probabilities of a set of corporates. Hence, the difference in the model here presented and the model by Creal et al. is based on the use of a binomial distribution to model transitions to default as opposed to using a multinomial distribution to model rating transitions. Overall, the mix-measurement GAS model is considered a non-parsimonious model, since it involves the estimation of a large amount of parameters, a set of latent factors and the joint modeling of macroeconomic variables and default transitions.

The interest on using the model proposed by Creal et al. is based on its practical applicability. While most latent factor models require the use of time-consuming simulation procedures to generate estimation results, Creal et al. propose a latent factor modeling framework that does not require the use of simulation since the likelihood function of the model is known in closed form. Given the special focus of this research on the practical applicability of the models, the possibility of estimating a latent factor model via a maximum likelihood procedure is considered an advantage against other latent factor models whose calibration relies on simulations procedures.

The main difference between Creal et al. latent factor model and other latent factor models is that the current parameters as defined by Creal et al. are observation driven and, therefore, deterministic functions of lagged dependent variables as well as contemporaneous and lagged exogenous variables. In this setting, parameters evolve randomly over time but are perfectly predictable one-step-ahead given past information. A well-known example of observation driven models is the generalized autoregressive conditional heteroscedasticity (GARCH) model. A through distinction between observation- and parameter-driven models is provided by Cox et al. (1981).

All in all, the research here presented aims to answer two questions. First, it addresses the question regarding estimation bias introduce by the length of the loan level data set and how this bias affects the forecasting performance of probability of default models. Specially, the potential bias being introduced due to the length of the data set not covering complete economic cycles is investigated. The second question addresses the efficiency or accuracy gains from using models that are known to be more computational intensive and less intuitive, such as latent factor models, against models that are less computational intensive and whose factors do not need interpretation.

In order to address the above mentioned research questions the following procedure is proposed.

First, the properties of the data set of residential mortgages at hand are presented. The data set is obtained from Fannie Mae's database and consists of quarterly observations from 23,902,488 unique single-family loans over the period of 2000 to 2016. Second, the models are introduced in depth in a methodology section. In this section, the details of the models are explained, including their estimation procedure and their underlying assumptions. Third, a simulation study is undertaken to analyze the potential bias arising from calibrating PD models to data sets covering incomplete economic cycles. A simulation study is of relevance for the research here proposed, since by knowing the properties of the data generating properties more detailed theoretical results can be obtained. In the simulation study defaults are modeled following two different approaches. Fourth, in section 4.2 an empirical analysis of the models is undertaken to analyze the potential bias arising from calibrating PD models to data sets covering incomplete economic cycles. To investigate this bias both models are calibrated using in-sample windows with different lengths. Then, the calibration results are compared to determine the effect of the length of loan level data on the estimation of PD models. Fifth, to investigate the added value introduced by complexity in PD models, the observable and unobservable risk factor models presented in this paper are compared by evaluating their out-of-sample forecasting accuracy. Sixth, a set of robustness checks are performed to confirm the obtained results are free of bias. Finally, a conclusion regarding the importance of the sample's length as well as the added value of using latent factors while modeling panels of residential-mortgages is provided.

2 Data

The data set consists of quarterly loan level observations for 23,902,488 unique single-family loans over the period of 2000 to 2016. This data set is set to disposal by Fannie Mae and can be easily downloaded in their website under subscription. Within the data set two types of loan level data are accessible, namely acquisition and performance data. Acquisition data concerns borrower and loan characteristics at the moment of origination, for example, original loan-to-value, borrower credit score, debt-to-income ratio, etc. Performance data concerns, amongst other things, the status of a loan over time, for example, loan age, remaining months to maturity, current interest rate, current loan delinquency status, etc.

While the data provides observations for a variety of loan-specific covariates, in some cases the amount of missing quarterly observations in the database is very large. This is specially the case for performance variables such as the current unpaid principal balance of loans per quarter. The reason for the lack of observations is unknown. Therefore, the number of loan-specific covariates taken into account in the construction of the models has been reduced by removing covariates with less than 80% quarterly observations. For example, in the case of the current unpaid principal balance more than 50% of the observations are missing. Only an amount of 20% missing observations is allowed in order to avoid quarterly distributions with large densities in the mean. Missing values for variables that satisfy the 80% threshold are handled as follows. First, two median values for the variable are computed. The first median value corresponds to non-defaulting loans only and the second median value corresponds to defaulting loans only. Second, at a given quarter missing values are populated by either of both medians depending if the loan, for which the variable is not observed, defaults or not. Table C.1 in Appendix C provides and overview of the remaining variables available and used for modeling as well as a description of what their values represent.

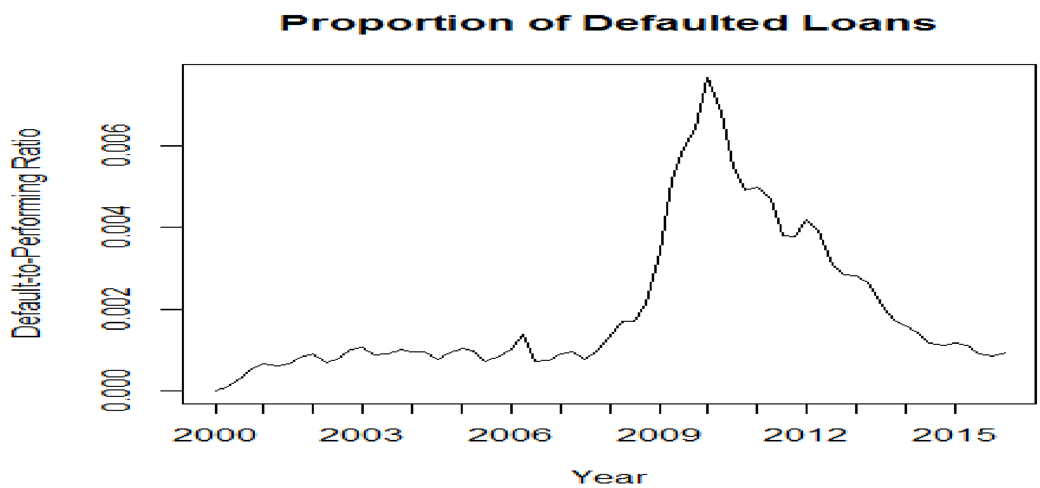
Additionally to the variables provided by Fannie Mae one indicator variables is constructed, namely the default indicator. The default indicator for mortgage i at time t takes either the value one if the borrower defaults in the next period or the value zero if the borrower's status is performing on the next period, i.e.,

$$Y_{i,t} = \begin{cases} 1, & \text{if mortgage } i \text{ in time } t+1 \text{ defaults} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Therefore, the default indicator indicates whether a borrower defaults in the next quarter. Given the interest in obtaining the relationship between the value of a performance or acquisition variable in quarter t with a default event in quarter $t+1$, this is considered an appropriate definition for the default indicator.

An overview of the amount of mortgage observations per year is provided in table C.2. This table shows the aggregated amount of quarterly performing and non-performing loan observations per year. It also provides the default-to-observation ratio, also called default rate, for each year. In total the data set consists of 348,478,251 loan observations from which 808,045 are default observations. A clearer view of the evolution of transitions from performing to default can be seen in Figure 1. As it could have been expected, the default rate drastically increases at the beginning of 2008, reaches a maximum in the aftermath of the financial crisis and decreases afterwards to reach a pre-crisis default rate level. Table 1 provides summary statistics of relevant performance and acquisition variables. As it can be seen some acquisition variables in the data set at hand have a large number of missing observations.

Although the number of missing observations is small relative to the total number of unique loans in the data set, lack of observations could introduce a bias if the missing observations all correspond to defaulted loans. In the same table, it can be seen that the missing observations are spread between loans that do not default and loans that default at some point in time. This problematic is avoided by filling missing values as previously described. By observing the mean values for defaulting and non-defaulting loans it can be seen that the data set at hand is in line with other data sets investigated in similar literature (see Table 2 in Lee et al. (2016)). More precisely, it can be seen that OLTV and DTI are larger for defaulting loans than for non-defaulting loans, and the FICO score is higher for non-defaulting loans compared to defaulting loans.



This figure shows the time variation in the default rate (D/N) for every quarter between 2000 and 2016.

Figure 1: Evolution of the Default Rate Over Time

Following the standard in the literature, systematic risk from general macroeconomic conditions is taken account for by using four macroeconomic indicators that have been obtained from the FRED database at the Federal Reserve Bank of St. Louis. The indicators are the quarterly change in industrial production growth (quarterly), quarterly change in the unemployment rate, quarterly change in real GDP (quarterly) and quarterly change in the house price index (quarterly). The length of all four variables span from 2000 to 2016 and, therefore, all four series capture the effects of the financial crises of 2008. Furthermore, the macroeconomic variables have been standardized previous to using them for model calibration by subtracting the time series mean and dividing by the standard deviation. Tables C.1 and 1 provide a description and summary statistics for the macroeconomic variables used in this research.

Table 1: Summary Statistics of Loan Level Covariates and Macroeconomic Variables

	Non-Default			Default			Total							
Variable	**Mean**	**STD**	**NA's**	**Mean**	**STD**	**NA's**	**Mean**	**STD**	**Min**	**Q1**	**Median**	**Q3**	**Max**	**NA's**
Acquisition Variable														
FICO Score	740	53	108,079	686	57	6,340	738	54	300	702	750	782	850	114,419
Original Unpaid Balance	197,412	109,776	0	178,431	95,799	0	196,800	109,387.20	4,000	116,000	173,000	253,000	1,403,000	0
OLTV	73	16	11	78	13	0	73	16	1	65	78	80	97	11
Original Interest Rate	5.60	1	6	6.33	1	0	6	1	2	5	6	7	14	6
DTI	34	12	500,233	40	12	31,170	34.3	12	1	26	34	42	64	531,403
Performance Variable														
Current Interest Rate	6.01	1	0	6.15	1	0	6.02	1	2.12	5.53	5.94	6.44	11.98	0
Unemployment Rate							6.39	1.81	3.90	5.00	5.80	7.85	10.10	0
GDP							1.87	2.52	-8.20	0.80	2.15	3.25	7.80	0
Industrial Production Growth							0.91	4.29	-15.10	0.37	2.13	2.86	7.15	0
HPI							156.00	27.79	101.70	139.70	149.40	177.20	206.30	0

This table gives descriptive statistics of loan level covariates and macroeconomic variables. The covariates are observed either once per loan at the moment of origination (acquisition variables) or once every quarter for each loan (performance variables). Summary statistics from three different samples are presented, namely a sample of defaulting loans, a sample of non-defaulting loans and a sample with all loans. Statistics for performance variables are obtained as averages of quarterly statistics.

3 Methodology

In this section two models to estimate the probability of default of residential mortgages are presented. Detailed descriptions of their calibration procedures as well as the necessary mathematical derivations are provided in the appendix.

3.1 Discrete Hazard Rate Models and Multi-Period Logit Models

While the probability of default from residential mortgages has not been as widely modeled as the probability of default for corporate loans, in the literature reviewed for this paper there seems to be a widely used standard model to estimate PD of residential mortgages, namely the discrete hazard rate model from Cox & Oakes (1984). The application of this model to the estimation of probability of default is illustrated by, for example, Deng (1997), Bellotti & Crook (2014) and Ambrose & Capone (2000). The model specification is as follows. Given a pool of loans is observed for a time interval $t = 1, ..., T$, let t_i be the discrete time value at which loan i defaults. Then for every loan in the pool, $i \in N$, define the probability mass function of failure as $f(t; \theta, x_{it})$, where θ represents the vector of parameters of f and x represents a vector of loan specific information at time t. The survival function, $S(t; \theta, x_{it})$, and the hazard rate function $\phi(t; \theta, x_{it})$ are defined as

$$S(t; \theta, x_{it}) = 1 - \sum_{j<t} f(j; \theta, x_{ij}) \tag{2}$$

and

$$\phi(t; \theta, x_{it}) = P(t < t_i < t + \Delta t | t_i > t) = \frac{f(t; \theta, x_{it})}{S(t; \theta, x_{it})} \tag{3}$$

where equation (3) follows from Bayes rule. The survivor function gives the probability of surviving up to time t and the hazard function gives the probability of failure at t conditional on surviving to t.

Besides the discrete hazard rate model presented above, probability of default has been also widely estimated by logistic or probit regression models. At first glance hazard rate and logit models might seem different. Specially, hazard rate models might seem better for modeling random variables that can only take a certain value once as it is the case for a binary indicator that takes upon the value one when the borrower defaults and zero otherwise. However, Shumway (2001) demonstrates that a multi-period logit model is equivalent to a discrete-time hazard model with hazard function $\phi(t; \theta, x_{it}) = F(t; \theta, x_{it})$, where F is the cumulative density function that corresponds to the probability mass function of failure, $f(t; \theta, x_{it})$. Furthermore, it is shown that, assuming the covariates in the model capture all cross-sectional and time-series heterogeneity in the loans, the discrete hazard rate model can provide consistent and unbiased estimators of θ.

The parsimonious model proposed in this paper is a two-step observable-factor model that incorporates the concepts of through-the-cycle and point-in-time probability of default. Similar models have been proposed in the literature as it is the model by Lee et al. (2016). In the first step, a TTC PD is estimated by means of logistic regression. More precisely, for each loan let Y_{it} denote the default indicator for mortgage i at time t, which takes either the value of one for default or the value of zero for non-default. The first step of the model consists of determining the TTC PD of the loans by calibrating the following model

$$\begin{aligned} p_{it}^{TTC} &= P(Y_{it} = 1 | x_{it-1}) = \Lambda(\beta_0 + \sum_j \beta_j x_{ijt-1} + \epsilon_{i,t}) \\ &= \Lambda(h_{it|t-1}^{TTC}), \end{aligned} \tag{4}$$

where $\epsilon \sim$ Logisitc$(0,1)$ and Λ is the logistic function.

Given the estimate of the TTC PD, the PIT PD is obtained by incorporating the macroeconomic conditions at a specific point in time t by means of the observable risk factors m_t as follows

$$p_{it}^{PIT} = P(Y_{it} = 1|x_{t-1}, f_t) = \Lambda(\alpha h_{it}^{TTC} + \lambda' m_t). \tag{5}$$

This model can be estimated using the method of maximum likelihood as shown in detailed in Appendix A.1.

3.2 Observation Driven Mixed-Measurement Dynamic Factor Model

The observable factor model presented above takes the explicit assumption that time-series heterogeneity and the cross-sectional heterogeneity among loan observations is fully captured by the loan specific covariates and the observable risk factor. However, this assumption can be relaxed by incorporating the possibility of latent risk factors having an effect in the loan's probability of default. Creal et al. (2014) develop an observation driven mixed-measurement dynamic factor model that incorporates latent risks factors into the estimation of probability of default. The main reasons why this specific model is considered as an alternative to the observable factor model is that while the estimation procedure of both models is based on the maximization of a closed form likelihood function, the model by Creal et al. allows the identification of several latent components of credit risk. The availability of a closed form likelihood function leads to a simple procedure for likelihood maximization and, compared to other models, it avoids the need for simulation based methods to evaluate the likelihood.

In their paper Creal et al. investigate corporate credit rating transition probabilities for 7,505 companies. The application of the observation driven mixed-measurement dynamic factor model in this paper differs in three ways. First, the model is used to estimate the probability of default from residential mortgages as opposed to the estimation of rating transition probabilities. As it is shown by Creal et al. (2014), when modeling corporate rating transitions it is necessary to take into account that the rating assigned to a corporate in the future is chosen from a set of different possible ratings. Credit ratings classify corporates depending on their credit worthiness. For example, Moody's Analytics divides the credit worthiness of corporates in ten classifications, Aaa (prime) being the best possible rating and C (in default) being the worst possible rating. Hence, Creal et al. (2014) model corporate credit rating transitions using a multinomial distribution. Probability of default modeling from residential mortgages differs to the modeling of rating transitions, because a residential mortgage can only be in one of two possible states, namely either performing or default. Therefore, the mix-measurement GAS model presented in this paper uses a bernoulli distribution to model probabilities of default from residential mortgages. Second, the data set of residential mortgages at hand is much larger than the corporate ratings data set used by Creal et al. to fit the model. Third, the application of the model in this paper incorporates borrower specific information into the estimation procedure.

The unobservable factor model for the analysis of probability of default proposed in this paper is an adaptation of their model to the framework of residential mortgages. Start by considering the vector $y_t = (y_t^m, y_t^r)'$, where y_t^m denotes the N_m-dimensional vector formed by the macroeconomic variables incorporated in the model and y_t^r represents the default indicator for a residential mortgage. The general specification for the measurement density of any element in the vector is represented by

$$y_{it} \sim p_i(y_{it}|f_t, F_{t-1}; \Theta), \qquad \text{for} \qquad i = 1, ..., N, \quad t = 1, ..., T, \tag{6}$$

where f_t is vector of unobserved factors or time-varying parameters, $F_t = \{y_1, ..., y_t\}$ is the set of past and concurrent observations at time t, and Θ is a vector of static unknown parameters. The vector of latent factors is further specified by letting its dynamics be captured by the following autoregressive moving average process

$$f_{t+1} = \omega + \sum_{i=1}^{p} A_i s_{t-i+1} + \sum_{j=1}^{q} B_j f_{t-j+1}, \qquad t = 1, ..., T \tag{7}$$

where ω is an M x 1 vector of constants and coefficients A_i and B_j are M x M parameter matrices for $i = 1, ..., p$ and $j = 1, ..., q$. The innovation s_t is set equal to the score of the log-density $p(y_t|f_t, F_t; \Theta)$ for $t = 1, ..., T$. This specification for the dynamics of the latent factor corresponds to the specification of generalized autoregressive score models (GAS).

GAS models are introduced by Creal et al. (2013). The motivation for using these type of models is that they have the advantages of observation-driven models and at the same time allow to exploit the complete density structure rather than means and higher moments only, because the score depends on the complete density, and not only on the first- or second-order moments of the random variable y_t. Therefore, GAS models exploit higher-order moments by incorporating the score function of the predictive model density into the estimation procedure. In particular a scaled score s_t is used as a component in the autoregressive equation for the latent factors. By exploiting the full density, the GAS model introduces new transformations of the data that can be used to update the time-varying parameter f_t. The scale score is defined as

$$s_t = S_t \nabla_t, \qquad \text{where} \qquad \nabla_t = \frac{\partial log(p(y_t|f_t, F_{t-1}; \Theta))}{\partial f_t}. \tag{8}$$

The matrix S_t is a M x M positive definite scaling matrix known at time t and ∇_t is the derivative of the log-likelihood function with respect to f_t evaluated at y_t. As it is indicated by Creal et al. (2013), in many situations it is natural to consider a form of the scaling matrix that depends on the variance of the score. As it is described by Ardia et al. (2016), the standard in the literature is to set the scaling matrix to a power $\gamma > 0$ of the inverse of the information matrix of f_t to account for the variance of ∇_t. More precisely,

$$S_t(f_t) = \mathrm{I}_t(f_t)^{-\gamma}, \qquad \text{with} \qquad \mathrm{I}_t(f_t) = \mathrm{E}_{t-1}[\nabla_t \nabla_t'], \tag{9}$$

where the expectation is taken with respect to the conditional distribution of y_t given y_{t-1}.

In their paper, Creal et al. (2014) set $\gamma = \frac{1}{2}$ and construct the scaling matrix by using the eigendecomposition of Fisher information matrix. In the analysis performed for this paper, the same approach has been followed. Therefore, the scaling matrix is defined as

$$S_t = U_t \Sigma_t^{-\frac{1}{2}} U_t', \tag{10}$$

where U_t is a matrix composed by the eigenvectors corresponding to the nonzero eigenvalues of Fisher information matrix and Σ_t is the diagonal matrix composed by the nonzero eigenvalues of Fisher information matrix. Ardia et al. (2016) elaborates on an important property of the score series, namely that whatever the choice of γ, s_t is a martingale difference with respect to the distribution of y_t given $y_{1:t-1}$. This implies that given the information up to a specific point-in-time, the mean of the score series is equal to zero conditional on all previous information. This property of the score series becomes specially useful while determining the unconditional mean of the latent factor process f_t as shown in appendix A.2.1.

Given the general framework of the observation-driven model, it is possible to define a model that can specifically be used to estimate the PD of residential mortgages. The measurement densities of the model are given by

$$y_t^m \sim N(\mu(f_t), \Sigma_m) \tag{11}$$

$$y_t^r \sim Logit(\pi_i(f_t)) \tag{12}$$

where the N_m x 1 mean vector μ_t is a function of f_t, the variance covariance matrix Σ_m is fixed over time and the probability $\pi_i(f_t)$ concerns borrower i and its transition from performing to default during period t as a function of f_t.

The dynamic specification for the parameters in the model are

$$\mu_t = Z^m f_t, \tag{13}$$

where the N_m x M matrix of factor loadings Z^m relates f_t to μ_t, and

$$\pi_{it} = P[y_{it}^r = 1 | x_{it}, f_t; \Theta] = \frac{exp(\beta_0 + \sum_j \beta_j x_{ijt-1} - Z^{r'} f_t)}{1 + exp(\beta_0 + \sum_j \beta_j x_{ijt-1} - Z^{r'} f_t)}, \tag{14}$$

where the probability of default depends on borrower specific characteristics, x_{it}, and is related to the time-varying factor f_t via the vector of factor loadings Z^r. It is important to note that as the macroeconomic variables have been standardized before they enter the analysis, a constant term has not been included in 13.

A precise derivation of the score, s_t, and the contributions to the likelihood function from the logit and normal components of the model as well as a detailed description of the algorithmic procedure to estimate the probability of default are provided in Appendix A.2. The derived score for the logit component is

$$\nabla_t^r = \sum_i^{L_t} (y_{it}^r - \pi_{it}) Z^r \tag{15}$$

As it can be seen, the score of the logit component provides an insight into how the values of the latent factors are generated. The first element of the equation represents the difference between the observed default rate and the default rate generated using a specific selection of parameter values. The second element scales the value of the score. Similarly, by observing the equation for the score of the normal component

$$\nabla_t^m = Z^{m\prime} \Sigma_m^{-1} (y_t^m - \mu_t) \tag{16}$$

it can be seen that the same logic applies. The score of the normal component is the standardized deviation of the mean implied by the parameters with respect to the realized values of the normal random variables.

Overall, the values of the score of the mix-measurement model can be interpreted as a measure of the fit of the data generating process implied by the estimated parameters to the true data generating process. In other words, as indicated in Creal et al. (2013), the quantity s_t indicates the direction to update the vector of parameters from f_t to f_{t+1}, acting as a steepest ascent algorithm for improving the model fit given the current parameter position. By presenting the factor in its recursive solution

$$f_t = \sum_{i=0}^{t-1} B^i w + \sum_{i=0}^{t-1} B^i s_{t-1-i} + B^t f_1 \tag{17}$$

it can be seen that the factor should be interpreted as a measure that captures the information that is not being captured by the observable factors. Hence, given the context of this research, the factor should represent either macroeconomic or default information that is not incorporated in the model.

4 Effect of the Loan Panel's Length on the Models' Calibration

As it has been mentioned, this research addresses two questions. The first one concerns the effect of the sample's length on parameter estimates of observable and unobservable risk factor models for probability of default. The two models presented in this paper can be calibrated by means of maximum likelihood estimation. Hence, the objective is to investigate if maximum likelihood estimates can be significantly biased by the length of the sample at hand. However, in theory the bias of a parameter estimate can only be determined if its actual value is known.

Therefore, to investigate the potential bias introduced by the length of loan panels two steps are taken. First, a preliminary analysis on the effect of the sample length on maximum likelihood estimates is performed in a controlled simulation study, where the probabilities of default are generated by two different data generating processes. Preliminary conclusions are derived in the simulation study. Second, the effect of the data length in the model calibration is investigated in an empirical setting by fitting the probability of default models to non-simulated samples of increasing windows and identifying differences and trends in the resulting parameter values.

The second research question addresses the difference in forecasting performance generated by assuming all time-series dependence can be captured by observable risk factors against the assumption that some time-series dependence is latent and can't be captured by observable risk factors. This question is targeted by analyzing the out-of-sample performance of the observable and unobservable risk factor models in section 5.

4.1 Simulation Study

The main purpose of performing a simulation study in this context is to investigate the potential estimation bias introduced by the length of the loan panel. Specially, it is of interest to investigate the resulting estimation bias generated by a data set that does not covers full economic cycles. The simulation study is performed in two simplified versions of the global economy. In the first case, the default rates are generated by a logistic regression model with a single observable risk factor. In the second case, default rates are generated by a logistic regression model with a generalized autoregressive score component and an observable risk factor.

From Figure 1 it can be seen that during the financial crisis of 2008, default rates increased in average by a factor of six and during recovery periods the default rate is higher by a factor of three and a half than the default rates on stable states of the economy. Moreover, it can be seen that during the stable state of the economy default rates have a low variance and recovery periods are approximately two thirds longer than the time it takes for a financial crisis to achieve its worst. In the following simulation study, these characteristics are taken into account to generate an economic cycle that is closed to what is observed in reality.

The results of the simulation exercise for the two models provide insights into the potential biases encountered when the model parameters are estimated via maximum likelihood. In the simulation of the logistic regression model with single systematic risk factor, the risk factor drives the probabilities of default. Therefore, there is a direct link between the state of the economy and the probabilities of default. The results of the simulation and estimation exercise for each model indicate towards different results. In the case of the logistic regression model with observable risk factor the parameter estimates obtained from calibrating the model with data corresponding to a recovery period tend to be less biased than the parameter estimates obtained from calibrating the model to data sets consisting of several components of an economic cycle. In the case of the logistic GAS model, it is not possible to derive a conclusion regarding the effect of the sample's length on the bias of the parameter estimates. While for some parameters it is possible to observe a trend in the change of their bias as the sample incorporates more components of the economic cycle, for the majority of the parameters it is not possible to explain the variation on the bias based on the sample's length.

The conclusions of the simulation exercise are supported by the results in section 4.1.3, In this section, it is attempted to determine the explanatory power of the sample's length on the parameter estimates by fitting a set of linear regression to the simulated data and the obtained bias for each parameter of both models. The results indicate that the variation in the bias of the parameter estimates from the logistic regression model can be explained by the length of the sample. Specially, by the amount of different economic states included in the data set used to calibrate the model. For the logistic GAS model, the results indicate that for the majority of the parameters, it is not possible to explain the variation in the estimates' bias by the amount of different economic states included in the data set used to calibrate the model. Overall, the results of the simulation exercises indicate the number of economic states encompassed by the data set has different effects on the maximum likelihood estimation of each model.

4.1.1 Logistic Regression with Single Systematic Risk Factor

A simplified version of the world is considered, where the default rate of a pool of loans is assumed to depend uniquely on one observable macroeconomic factor. The macroeconomic factor is modeled by a mixture of normal distributions with different means and variances. Specifically, the economy is assumed to be on either a crisis, a recovery or a stable growth. The described data generating process for probability of default is

$$y_t \sim B(p_t, n_t), \qquad p_t \sim \Lambda\big(x_t\big) = \frac{\tau}{1+e^{-x_t}}, \qquad x_t \sim \begin{cases} N(\mu_s, \sigma_s), & \text{if stable state} \\ N(\mu_c, \sigma_c), & \text{if crisis state} \\ N(\mu_r, \sigma_r), & \text{if recovery state} \end{cases}$$

where the performing loans at time t, n_t, follow a binomial distribution with probability of default p_t. Hence, $y_{i,t}$ takes value one if loan i defaults at time t and zero otherwise.

As it can be seen in figure 1, from 2000 to 2016 the default-to-performing ratio from the available pool of loans reached a maximum of 0.008 during the pick of one of the strongest housing crises in history. Note that the available data set is one of the largest data sets containing default information of American residential mortgages. Hence, given the nature of the housing crisis and the size of the data set, it is assumed that default rates above the 0.008 level are assumed to be implausible. Therefore, the maximum value of the logistic function, τ, as defined by Verhulst (1838), is set to be seven times higher than the default rate in a stable state of the economy, which is approximately the maximum factor by which default rates differ in a good and bad states. Let $\tau = 0.008$, then the average default rate during a stable state, a crisis and a recovery are set to be 0.001, 0.007 and 0.0045 respectively. For the risk factor, this implies that its average values during the three states of the economy are $\mu_c = \log 0.142$, $\mu_s = \log 7$ and $\mu_{cr} = \log 0.777$. Following the same approach as for the mean of x_t, appropriate values for the standard deviations are inferred for each state of the economy, namely, $\sigma_s = 0.5$ and $\sigma_c = \sigma_r = 0.55$.

Furthermore, the data shown in figure 1 is also used to determine a realistic value for the time-length of financial crises versus the time-length of recoveries and stable periods. In total, two occurrences for each state of the economy are simulated. Then, a whole economic cycle is constructed, where it is assumed that in an economy a crisis can only occur after a period of stable growth and a recovery can only occur after a period of crisis. Figure B.1 shows one simulated complete economic cycle, which consists of a path for the risk factor, x_t, and the implied values from applying a logistic transformation. An economic cycle is said to be complete when it is composed by two periods of stable economic growth, two periods of crisis and two periods of recovery.

The aim of this simulation exercise is to obtain an initial overview of the characteristics from the model parameters when the time-length of the panel covers incomplete economic cycles. Hence, in this setting, it is assumed that the the data generating process is known such that the results are not affected by model selection bias. The only unknown are the model parameters. Parameter estimates are obtained by means of maximum likelihood estimation. In the context of the model at hand the estimation procedure consists of providing a solution to the following maximization problem

$$\hat{\beta} = \arg\max_{\beta} \sum_{t=1}^{T} \log \left(\frac{\tau}{1+e^{-\beta x_t}} \right)^{y_t} \left(1 - \frac{\tau}{1+e^{-\beta x_t}} \right)^{1-y_t}.$$

The value of the objective function is maximized by means of Broyden–Fletcher–Goldfarb–Shanno (BFGS) iterative maximization procedure.

Table 2 shows the bias and standard deviation from maximum likelihood estimates obtained from 10,000 different paths, where the cross-section size is set to $N_t = 1,000$. Additionally, the table contains the results of testing the null hypothesis that the mean of the bias distribution of β obtained from fitting the model to complete economic cycles is the equal to the mean of the bias distribution of β obtained from fitting the model to incomplete economic cycles. Incomplete economic cycles are obtained as subsets of the simulated complete economic cycles similar to the one presented in figure B.1. The test results are easily interpretable, namely all means of the bias distributions obtained using incomplete cycles to calibrate the model are statistically different to the mean of the bias distribution obtained using complete cycles to fit the model. For the research question addressed in this paper, it implies that, ceteris paribus, the amount of economic scenarios included in the data set has an impact in the model's estimation and can play an important role in the practical applicability of the model.

Table 2: Bias of Parameters from the Logistic Regression Model

	[2C, 2R, 2SG]	[2C, 2R]	[2C, 2SG]	[2R, 2SG]	[2C]	[2R]	[2SG]
Bias(β)	-1.48	0.14	-1.57	-2.30	0.22	-0.03	-3.56
	(0.044)	(0.118)	(0.045)	(0.101)	(0.053)	(0.392)	(0.035)
p-value		0.000	0.000	0.000	0.000	0.000	0.000

This table shows the bias from maximum likelihood estimates and, in parenthesis, the standard deviation of the estimated values for β. C, R and SG refer to the different states included in a time-series. For example, [2C, 2R, 2SG] indicates the observations within the time-series cover two economic crises, two recoveries and two period of stable growth. For each scenario 10,000 economic cycles have been simulated. Below the bias estimates, p-values are provided. The p-values correspond to a test on the mean of the parameter distributions with the alternative hypothesis that the true difference in means is not equal to zero.

An advantage of parsimonious models such as the logistic regression model is that the robustness of the calibration results can be investigated by determining if the obtained parameter values, or in this case average bias values are in line with the expectations. In this case, it can be seen that the bias in the parameter estimates actually indicate what could be expected, namely the bias on β tends to be positive when the model is fitted to periods of economic distress and negative when the model is fitted to periods of stable growth. However, from all calibration exercises, the scenario with the lowest absolute bias is the scenario where the model is calibrated with data from a recovery period. This indicates that on average the parameter estimates obtained from a recovery period are more accurate. A possible explanation for this result is that the parameter estimates obtained using datasets covering more elements of the economic cycle are bias towards the default rates of the extremes within the simulated cycles, namely either stable growth or crisis. Overall, this is a plausible argument in favor of using data from a recovery period when calibrating a model that is intended to be used for modeling defaults throughout a complete cycle.

4.1.2 Logisitc Regression with GAS Component

Given the approach taken in this paper to specify the dynamics of the latent factors' dynamics as following a generalized autoregressive score processes, it is of interest to observe the finite sample properties of maximum likelihood estimates for logistic regression models with GAS components when the length of the simulated data set differs. Following the example of the previous section the main aim of this simulation exercise is to replicate the characteristics of the data in figure 1 and to incorporate the idea of a changing economy via a macroeconomic risk factor. The data generating process for probability of default is defined as follows

$$y_t \sim B(p_t, n_t), \qquad p_t \sim \Lambda(p_t^*) = \frac{\tau}{1 + e^{-p_t^*}}, \qquad p_t^* = \omega + \alpha f_t + \beta x_t,$$

where the factor f_t follows GAS process defined by the following equations

$$f_t = \lambda s_{t-1} + \delta f_{t-1}, \qquad s_t = S_t \frac{\partial l(y_t|f_t)}{\partial f_t}, \qquad S_t = I.$$

The inclusion of a generalized autoregressive score component into the simulation procedure requires the incorporation of the initial value of the GAS process, f_1, into the parameter set. The choice of the parameters and their corresponding constraints follows the common approach in the GAS literature (see for example Ardia et al. (2016)). For the specific model here considered two restrictions are set on the parameter values. First, the spectral radius of δ is constraint to be less than one in order to avoid an explosive pattern for f_t. Second, λ is restricted to be positive such that the signal coming from the conditional score is not distort. The coefficient λ, that premultiplies the scaled score s_t, controls for the impact of s_t to f_t. Specially, as it has been mentioned in section 3 the quantity s_t indicates the direction to update the vector of parameters from f_t to f_{t+1} acting as a steepest ascent algorithm for improving the model fit given the current parameter position. Hence, λ can be interpreted as the step of the update, and needs to be designed in a way to do not distort the signal coming from s_t.

Using straightforward algebra and the assumptions on the independence of observations given the information up to time t - 1, the maximization problem to be solved can be easily derived. Furthermore, given that the dynamics of f_t are driven by an autoregressive component and a score component, to simulate the factor across time it is necessary to compute a functional form for the score of the log-likelihood at time t. The log-likelihood function and the score function of the GAS model presented above are derived in Appendix B.2.

Similar to the previous section, the mean and variance of the macroeconomic factor, x_t, have been chosen such that the generated default rates capture the characteristics of the transition rates in figure 1. An instance of the resulting Monte Carlo paths is shown in Figure B.2. A total of 10,000 paths have been generated. For each path binary observations are drawn from a binomial distribution with time varying success probabilities. Table 3 shows the bias and standard deviation of 10,000 maximum likelihood estimates for Θ computed using the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) constrained optimization algorithm. The p-values provided in the table indicate the result of testing the null hypothesis that the mean of the bias distribution of a parameter obtained using a complete economic cycle to fit the model is equal to the mean of the bias distribution obtained using an incomplete cycle.

Overall, compared to the simulation results obtained for the logistic regression model, the results presented in this section are more cumbersome to interpret. It is worth noticing that the results obtained for the logistic regression model still hold in setting of the logistic GAS model, namely the bias of β is the lowest when the model is calibrated with data corresponding to a recovery period and the bias of β resulting from the estimation with different samples are intuitive. From the remaining estimates, it is worth noticing that the bias of the factor's intercept is relatively large and no pattern can be recognized from the difference in the bias obtained with different samples. While this result is rather unexpected, it is considered a characteristic of the maximum likelihood estimates for ω given the current model specification.

For the majority of the cases, the p-values for the test of significant difference between the mean of the bias distribution obtained from calibrating the model to a complete cycle relative to the means of the bias distributions obtained from calibrating the the model to incomplete cycles indicate that the null hypothesis of equal bias in the parameters is rejected. Similar to the results obtained for the logistic regression model, this implies that the amount of economic scenarios included in the

Table 3: Bias of Parameters from the Logistic GAS Model

	[2C, 2R, 2SG]	[2C, 2R]	[2C, 2SG]	[2R, 2SG]	[2C]	[2R]	[2SG]
Bias(f_1)	0.004 (0.065)	0.086 (0.104)	0.034 (0.44)	−0.019 (0.353)	0.296 (0.133)	0.0523 (0.046)	0.016 (0.005)
p-value		0.68	0.82	0.000	0.000	0.22	0.81
Bias(ω)	35.04 (43.43)	42.25 (33.79)	106.54 (22.20)	22.67 (21.77)	18.31 (6.71)	36.30 (15.55)	26.75 (1.21)
p-value		0.000	0.000	0.000	0.000	0.000	0.000
Bias(α)	−0.034 (0.255)	−0.041 (0.071)	−0.009 (0.648)	−0.030 (0.573)	0.004 (0.119)	−0.054 (0.0303)	−0.059 (0.0003)
p-value		0.63	0.10	0.000	0.50	0.09	0.09
Bias(β)	−0.416 (10.05)	−2.029 (9.726)	−29.39 (6.481)	−3.602 (7.623)	0.904 (2.876)	−0.276 (6.426)	53.232 (2.942)
p-value		0.000	0.000	0.000	0.000	0.000	0.000
Bias(λ)	0.070 (0.211)	0.005 (0.029)	0.0382 (0.228)	0.013 (0.214)	−0.001 (0.002)	0.001 (0.004)	0.0007 (0.0001)
p-value		0.09	0.04	0.01	0.000	0.000	0.000
Bias(δ)	−0.481 (0.711)	−0.526 (0.710)	−0.528 (0.713)	−0.522 (0.740)	−0.525 (0.755)	−0.528 (0.700)	−0.519 (0.710)
p-value		0.18	0.04	0.05	0.06	0.04	0.04

This table shows the bias from maximum likelihood estimates for the logistic model with a GAS component. In parenthesis, the standard deviation of the estimated values for $\theta = \{\alpha, \beta, \lambda, \delta, \omega, f_1\}$ are given. C, R and SG refer to the different states included in a time-series. For example, [2C, 2R, 2SG] indicates the observations within the time-series cover two economic crises, two recoveries and two period of stable growth. In each case, 10,000 economic cycles have been simulated.

data set has an impact in the model's estimation and can play an important role in the practical applicability of the model. Although in the case of the logistic regression model it has been possible to identify the effects of the bias in the parameters in the default rates estimated by the model, in the case of the logistic GAS model the effect of the bias of specific parameter estimates in the outcome of the model is more difficult to determine. This task becomes specially cumbersome if the bias of each parameter is assumed not to be independent from the others. This supports the need of an empirical study to determine the implications of the data set not covering complete economic cycles on the practical applicability of the model. The results of such an empirical study are presented in section 5, where the forecasting performance of a mix-measurement GAS model is compared with the forecasting performance of the two-step logistic regression model.

4.1.3 Quantifying the Bias

In the two previous sections, the bias of GAS and logit parameter estimates for varying sample lengths have been computed. This provides us with an overview of the expected bias in the model parameters arising from calibrating the models to data sets that partially cover economic cycles. In this section, the expected bias on each parameter is quantified by fitting linear regressions. The estimated bias obtained in the simulation study are used as dependent variables. A special effort is set to quantify the expected bias in maximum likelihood estimates for both models when the data set does not cover complete economic cycles. Therefore, dummy variables are used as explanatory variables. These variables indicate whether a data set covers a crisis, a stable growth period or a recovery.

As it has been mentioned in the previous section, the simulation consisted of generating 10,000 paths for each of the seven scenarios considered. Therefore, each linear regression model is fitted to 70,000 data points. The linear model fitted using the obtained bias estimates from the simulation exercise as dependent variables and the described variables as independent variables is the following

$$\text{Bias}(\theta) = \alpha + \beta_1 D_{\text{SG}} + \beta_2 D_{\text{R}} + \beta_3 D_{\text{C}} + \epsilon,$$

where $\theta = \beta$ for the logistic regression model and $\theta = \{\alpha, \beta, \lambda, \delta, \omega, f_1\}$ for the logistic GAS model, D_{C} indicates if the data sets covers a crisis, D_{SG} indicates if the data set covers a stable growth period, D_{R} indicates if the data set covers a recovery period and ϵ is an error term.

Table 4 presents the ordinary least squares estimates for the parameters as well as the adjusted R-squared of the regression. Below the coefficient estimates, the p-value resulting from a significance test on the coefficient estimates is provided. It is worth noticing that the R-squared of the regression with the bias of β as dependent variables is relatively high compared to the R-squared of the regression with the bias of other parameter estimates as dependent variables. This could have been expected from the simulation exercise, where it is possible to identify a clear pattern on the change of bias for β depending on the cycle components used to estimate the logistic model or the logistic GAS model. The R-squared of the other regressions is relatively low. The interpretation of these results is straight forward, namely the effect of the data covering different elements of an economic cycle has a major impact on the coefficient assigned to the macroeconomic variable. For the empirical analysis that follows this implies that for a logistic regression model as well as a GAS model, it should be expected to see variation in some parameters due to the length of the data set. However, for the case of the GAS model, it is not straight forward to determine in what extent the difference in available information within the data sets can affect the accuracy or overall performance of the model.

Table 4: Ordinary Least Squares Estimates

	α	D_{C}	D_{SG}	D_{R}
Logistic Regression Model				
Bias(β)	-0.94 (0.000)	1.03 (0.000)	-2.08 (0.000)	0.55 (0.000)
R-squared$_\beta$ = 0.91				
Logistic GAS Model				
Bias(f_1)	-0.37 (0.000)	-0.16 (0.000)	0.07 (0.000)	-0.04 (0.000)
R-squared$_{f_1}$ = 0.15				
Bias(ω)	9.13 (0.000)	50.06 (0.000)	28.92 (0.000)	9 (0.000)
R-squared$_\omega$ = 0.31				
Bias(α)	-0.50 (0.000)	-0.01 (0.29)	0.04 (0.002)	0.04 (0.000)
R-squared$_\alpha$ = 0.001				
Bias(β)	31.30 (0.000)	3.69 (0.000)	-30.74 (0.000)	-18.95 (0.000)
R-squared$_\beta$ = 0.19				
Bias(λ)	-0.57 (0.000)	0.04 (0.002)	0.03 (0.010)	0.03 (0.007)
R-squared$_\lambda$ = 0.001				
Bias(δ)	-0.54 (0.000)	-0.54 (0.169)	-0.54 (0.274)	-0.54 (0.147)
R-squared$_\delta$ = 0.0003				

This table shows ordinary least squares estimates for the linear regression with the estimated bias as dependent variables and the properties of the data set as independent variables. P-values for the test of significance in the coefficients are provided in parenthesis. Additionally, the R-squared from each regression is provided.

4.2 Empirical Analysis

The estimation procedures of the observable and unobservable risk factor models presented in section 3 are described in appendix A. The procedure consists of finding parameter values that maximize the log-likelihood function of the models. However, for large data sets this can result in a very computational intensive task. In the setting of this research, the original data set of 23,902,488 loans, available by Fannie Mae database, resulted impossible to use as sample for the calibration of the models. Therefore the following analysis on the properties of observable and unobservable risk factor models is performed using a sub-sample with 10,000 active loans selected uniformly at random without replacement from the complete pool of loans. In section 6.1, a robustness check is performed where it is shown that the subsample used in this empirical analysis captures the dynamics and the mean of default series from larger subsamples. Therefore, results obtained using the 10,000 subsample are considered representative of the results that would have been obtained by using the complete sample.

The purpose of this paper is to analyze the potential model misspecification arising from fitting a probability of default model to a data set of limited length. Since both models use macroeconomic data in their calibration procedures, it is specially of interest to determine the effect of the data not covering complete economic cycles. In order to investigate potential differences in maximum likelihood estimates arising from a change in the observation period, both models are fitted to samples of increasing lengths. The first step taken towards investigating this matter, is fitting the models to samples of increasing length starting with a sample spanning two years from 2016 to 2014 and then successively increasing the size of the sample by two years until the complete length of the available data is used in the calibration.

The covariates included in the model have been tested to determine if there is no high multi-collinearity among them. The collinearity between variables has been measured by means of the variance inflation factor (VIF). The VIF for each covariate is computed in two steps. First, an ordinary least squares regression is ran where one variable, say variable i, is regressed on the other variables, a constant term and an error term. Second, the VIF is computed as follows

$$\text{VIF} = \frac{1}{1 - R_i^2},$$

where R_i^2 is the coefficient of determination of the regression in step one, with covariate i as dependent variable and all other covariates as independent variables. As suggested in Neter (2004), as a rule of thumb if the variance inflation factors is above ten, then multicollinearity is high. The results in table 5 presents variance inflation factors for the available set of covariates computed using the 10,000 subsample of active loans. As it can be seen collinearity among the covariates is not high. Therefore, in this case the behavior of the parameters as the size increases is not affected by multi-collinearity.

Table 5: Variance Inflation Factors

	Model: $y = \alpha + \sum_i^v \beta_i x_i$				
N_s	UB	BCS	OLTV	DTI	CIR
10,000	1.148	1.148	1.057	1.030	1.234
	GDP	UR	IPG	HPI	
	1.133	2.147	2.299	2.147	

This table shows the variance inflation factors for the loan level covariates original unpaid balance (UB), borrower credit score (BCS), original loan-to-value (OLTV), debt-to-income (DTI) and current interest rate (CIR) and the macro-variables gross domestic product (GDP), unemployment rate (UR), industrial production growth (IPG) and house price index (HPI). N_l indicates the total amount of unique loans composing the sub-sample.

4.2.1 Logistic Regression Model

The results of fitting the observable risk factor model to the saples of increasing length are presented in table 6. Overall, the estimated parameters indicate the length of the loan panel plays a fundamental role in determining the values for parameter estimates. This can be recognized by analyzing the estimated parameters for each of the model's components. Starting by analyzing the estimated intercepts, it can be seen that the role of this parameter is to assign an appropriate level to the PD generated by the model. Lower intercept values generate lower default probability estimates by the model and vice versa. It is interesting to notice that the intercept is always significant. Therefore, regardless of the time span, the model always defines a level that is in line with the information contained in the data. This is specially visible by observing the change in the intercept values as the sample length increases to incorporate a period of economic recovery (slightly high default rates), a crisis (high default rates) and a period of stable growth (low default rates).

By comparing the intercept estimate for the 2014-2016 time span with that of 2003-2016 it is visible that, although both are the highest among the intercept, the later one incorporates the possibility of a crisis occurring and, therefore, takes a lower value than the former one. This is due to the fact that the smaller sample does not include nor an economic crisis or a recovery. Therefore, the model calibration with this sample does not incorporate the higher default rates observed during this period. Similarly, the intercept estimates obtained for the other time spans incorporate the information contained in the data sets by changing in accordance to the default rate levels.

Overall, the intercept values obtained are in line with the results from the simulation exercise in the previous section, where the estimates generated vary depending on the amount of economic states captured by the data set. Besides the intercept, only the parameter estimates of the borrower's credit score and current interest rate are significantly different from zero at the 5% level for all time spans. It is noticeable that the estimated parameters for these two variables behave the same as the estimates for the intercept.

Furthermore, the obtained parameter estimates for the macro-variables provide an interesting view of the added value from the through-the-cycle to point-in-time transformation that is commonly applied in the modeling of probabilities of default from mortgages. As it can be seen the parameter estimates for α are all significant at the 1% level and close to one. Values of α are interpreted as a volatility correction for the probabilities of default generated by the model. Values above or below one increase or decrease the value of $h_{it|t-1}^{TTC}$ by some percentage respectively. Thus, increasing or decreasing the variation in the estimated PD. It can be observed that the estimated values for α tend to increase as the sample size increases and take values above one for samples that include the financial crisis.

Regarding the parameter estimates for macro-variables, again a distinction can be done between data sets that include several economic states and those that do not can be done. While models calibrated with data sets not covering the financial crisis generate insignificant coefficients for macro-variables, models calibrated with data sets that include a financial crisis and a recovery generate significant coefficients for the same variables. This implies that the classification of mortgages as default or non-default in an economy without different states only benefits from loan level covariates that capture idiosyncratic risks. However, as the sample length increases to include different economic states and, therefore, some systematic behavior in defaults, the role of macro-variables in determining the PD becomes more important. In summary, the model calibrated with a longer data set incorporates the possibility of loans behaving in a systematic manner via the significant macroeconomic variables, while models calibrated with a shorter data set ignore this possibility.

4.2.2 Mix-Measurement Gas Model

For the analysis of the mix-measurement observation driven GAS model, the set of macro-variables included has been limited to the unemployment rate and the house price index. The selection of these two macro-variables from the complete set of available variables is based on the results obtained from the observable factor model, where only these two variables have shown to be significant at the 1% level in at least one of the time spans. Furthermore, it is also observed that

Table 6: Observable and Unobservable Factor Model Calibration with Increasing Time Spans

Two-Step Logit Model

Time Span	β_0	β_1	β_2	β_3	β_4	β_5	α	λ_1	λ_2	λ_3	λ_4
2014 - 2016	-6.615^{***} (0.284)	-0.699^{***} (0.160)	0.467^{*} (0.249)	-0.267 (0.222)	0.036 (0.206)	0.428^{**} (0.218)	0.961^{***} (0.153)	0.318 (0.722)	2.068 (4.898)	2.570 (4.518)	-0.693 (0.887)
2012 - 2016	-6.239^{***} (0.144)	-0.529^{***} (0.095)	0.274^{**} (0.139)	0.110 (0.124)	0.083 (0.121)	0.617^{***} (0.134)	0.978^{***} (0.032)	0.321 (0.254)	-1.234 (0.847)	-0.697 (0.828)	-0.39 (0.588)
2010 - 2016	-6.011^{***} (0.0970)	-0.426^{***} (0.066)	0.349^{**} (0.095)	0.276^{**} (0.078)	0.208^{**} (0.074)	0.812^{***} (0.095)	1.039^{***} (0.022)	0.128 (0.118)	-0.251 (0.331)	0.141 (0.323)	-0.030 (0.291)
2008 - 2016	-6.149^{***} (0.086)	-0.502^{***} 0.051)	0.413^{**} (0.078)	0.367^{**} (0.062)	0.273^{**} (0.058)	0.809^{***} (0.082)	1.054^{***} (0.018)	0.150^{**} (0.070)	-0.183 (0.233)	0.258^{*} (0.143)	-0.046 (0.048)
2006 - 2016	-6.317^{***} (0.083)	-0.505^{***} (0.049)	0.391^{**} (0.075)	0.365^{**} (0.060)	0.262^{**} (0.058)	0.733^{***} (0.082)	1.039^{***} (0.016)	0.137^{**} (0.068)	-0.475^{***} (0.175)	0.209 (0.134)	-0.050 (0.048)
2003 - 2016	-6.504^{***} (0.081)	-0.498^{***} (0.049)	0.368^{**} (0.072)	0.369^{**} (0.058)	0.280^{**} (0.059)	0.651^{***} (0.078)	1.053^{***} (0.013)	-0.038 (0.058)	0.062 (0.133)	0.722^{***} (0.098)	-0.003 (0.046)

Logit-Normal GAS One Factor Model

Time Span	β_1	β_2	β_3	β_4	β_5	Z^m
2014 - 2016	0.075^{***} (0.009)	-0.072 (0.059)	-0.041 (0.207)	0.192 (0.204)	1.105^{***} (0.243)	0.846 (0.941)
2012 - 2016	0.037^{***} (0.010)	-0.047 (0.032)	-0.064 (0.207)	0.111 (0.204)	0.874^{***} (0.203)	0.710^{**} (0.441)
2010 - 2016	0.020^{***} (0.003)	0.002 (0.001)	-0.063 (0.207)	-0.012 (0.204)	0.502^{***} (0.100)	0.650^{**} (0.241)
2008 - 2016	0.029^{***} (0.003)	0.018 (0.011)	-0.057 (0.207)	-0.052 (0.204)	0.278^{***} (0.020)	0.925^{**} (0.441)
2006 - 2016	0.027^{***} (0.007)	0.025 (0.019)	-0.052 (0.207)	-0.051 (0.204)	0.228^{***} (0.041)	0.187^{***} (0.051)
2003 - 2016	0.020^{***} (0.009)	0.029 (0.015)	-0.042 (0.207)	-0.046 (0.204)	0.174^{***} (0.030)	0.338^{***} (0.101)

This table shows the the maximum likelihood estimates for the observable and unobservable risk factor model as the sample length increases. Standard errors of the estimates are given in parenthesis. ***, ** and * indicate the null hypothesis that a parameter's trues value is equal to zero is rejected at the 0.001, 0.05 or 0.10 significance level respectively. In the observable factor model, λ_1, λ_2, λ_3 and λ_4 are the coefficients of GDP, HPI, UR and IPG respectively, while β_1, β_2, β_3, β_4, β_5 are the coefficients of the borrower credit score, OLTV, DTI, UPB and current interest rate. In the unobservable factor model βs are assigned equally and Z^m is the factor loading assigned to the macroeconomic latent factor.

in none of the cases more than two macro-variables are significant for one time span. Therefore, the inclusion of more than two macro-variables can result in an over-fitted model. As it has been mentioned above, the mix-measurement model consists of a logit and a normal component. In this setting, the normal component is composed by two macro-variables, which are assumed to follow a normal distribution with time-varying means as it is indicated in equation (11). In the logit component of the model, a probability of default is defined for each borrower at every point in time by introducing the loan level covariates as exogenous variables. The logit and normal components are linked by one or several factors. Following the approach taken by Creal et al., the dynamic specification of the factors is provided by equation (7) with $p = q = 1$.

In the study realized for this paper, models with different combinations of factors have been considered. Table 7 presents the maximum value of the log-likelihood function attained when applying the BFGS-algorithm for different model specifications as well as the corresponding Akaike information criterion (AIC) and Bayesian information criterion (BIC). The results indicate that a model with one macroeconomic factor and one frailty factor best fits the data. This result is also observed by Creal et al..

Table 7: Likelihoods and Information Criteria

	(1,0)	(1,1)	(2,0)
log-Lik	-2282.756	-2142.96	-2283.69
AIC	4593.51	4321.924	4603.39
BIC	4620.55	4356.697	4638.17

The table contains the log-likelihood values and information criteria for alternative model specifications. Each model contains different number of macroeconomic (m) and credit or frailty risk (r) factors, which are ordered as (m, r),

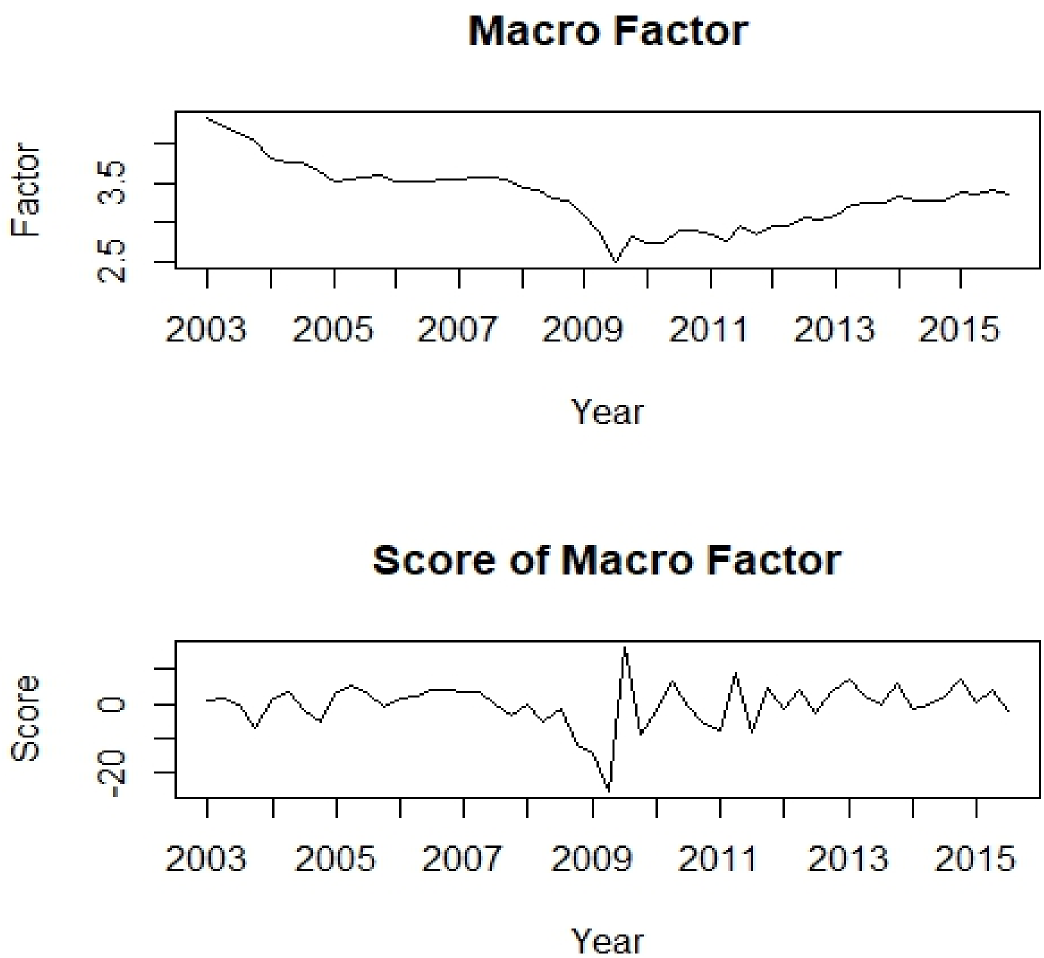
While measures of fit in general can provide a good indication towards the superiority of one model against another, it is also important to test the appropriateness of a model by observing its ability to generate the input data. Therefore, the two best performing models, namely (1,0) and (1,1), are scrutinized further by comparing the probability of default, factor and score series generated by the calibrated models. As a first step, the factor and score series of the models are analyzed. These series are investigated by determining if they are centered around their corresponding long term means, derived in appendix A.2.1.

As it can be seen from figures 2 and 3 the score series generated by both models are centered around their theoretical mean, zero. However, results obtained for the factor series of the two factor model indicate that the series are not centered around their theoretical mean. This is not the case for the one factor model. Additionally, while for both models the macro factor seems to be interpretable, the frailty factor is simply a concave function. This is a clear indicator that the two factor model is possibly resulting in an overfitted model.

As a final sanity check, the probability of default series generated by each model is compared. Figure 4 on page 25 compares the estimated default rates of the one factor and two factor models with the actual default rate series. It is easily observable that the default rate series generated by both models capture the level of the actual default rates appropriately in most of the years. However, the default rates estimated by the two factor model fail to capture the high default rates observed during periods of economic distress or in tail scenarios. In fact, the model estimates default rates that are half the size of the default rates observed. On the other hand, the one factor model seems to better fit the data by appropriately estimating the larger default rates during the financial crisis.

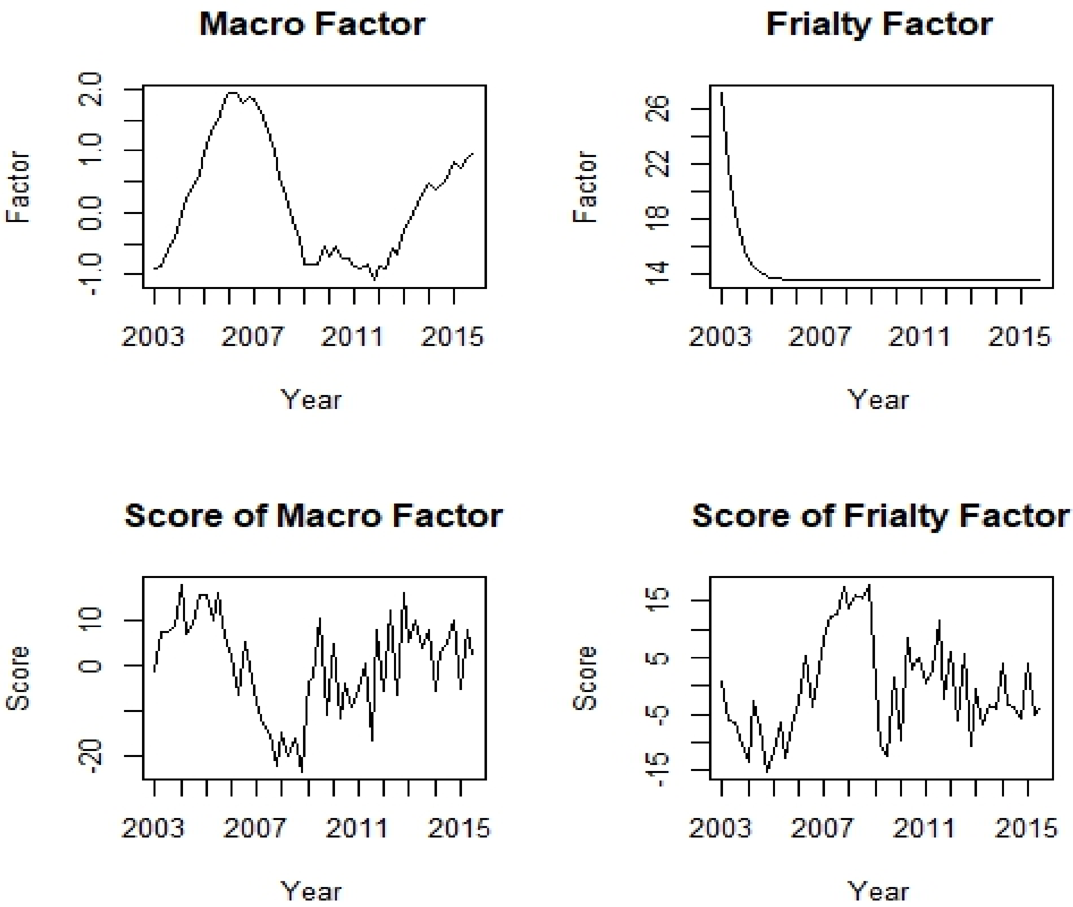
All in all, it is observed that both models can successfully estimate default rates that fit the actual data. Based on the measures of fit, the two factor model seems to be a better candidate to fit the loan level data at hand. However, by analyzing the specific theoretical properties of the model and the estimated default rates, it can be argued that the one factor model provides a better fit to the data at hand. Therefore, further analyses involving the mix-measurement GAS model are performed using one macroeconomic factor only.

At this point it is worth mentioning that the results presented above have been obtained by using the same specification for the log-odds as indicated in equation (14), but without the intercept coefficient. When calibrating the model with the log-odds specified as in equation (14) unreasonable estimation results where obtained whenever the intercept was kept in the equation. Therefore, the restriction $\beta_0 = 0$ in equation (14) is imposed.



This figure shows the resulting score and factor series from calibrating the mix-measurement one factor GAS model to the complete sample.

Figure 2: Score and Factor of One Factor Model



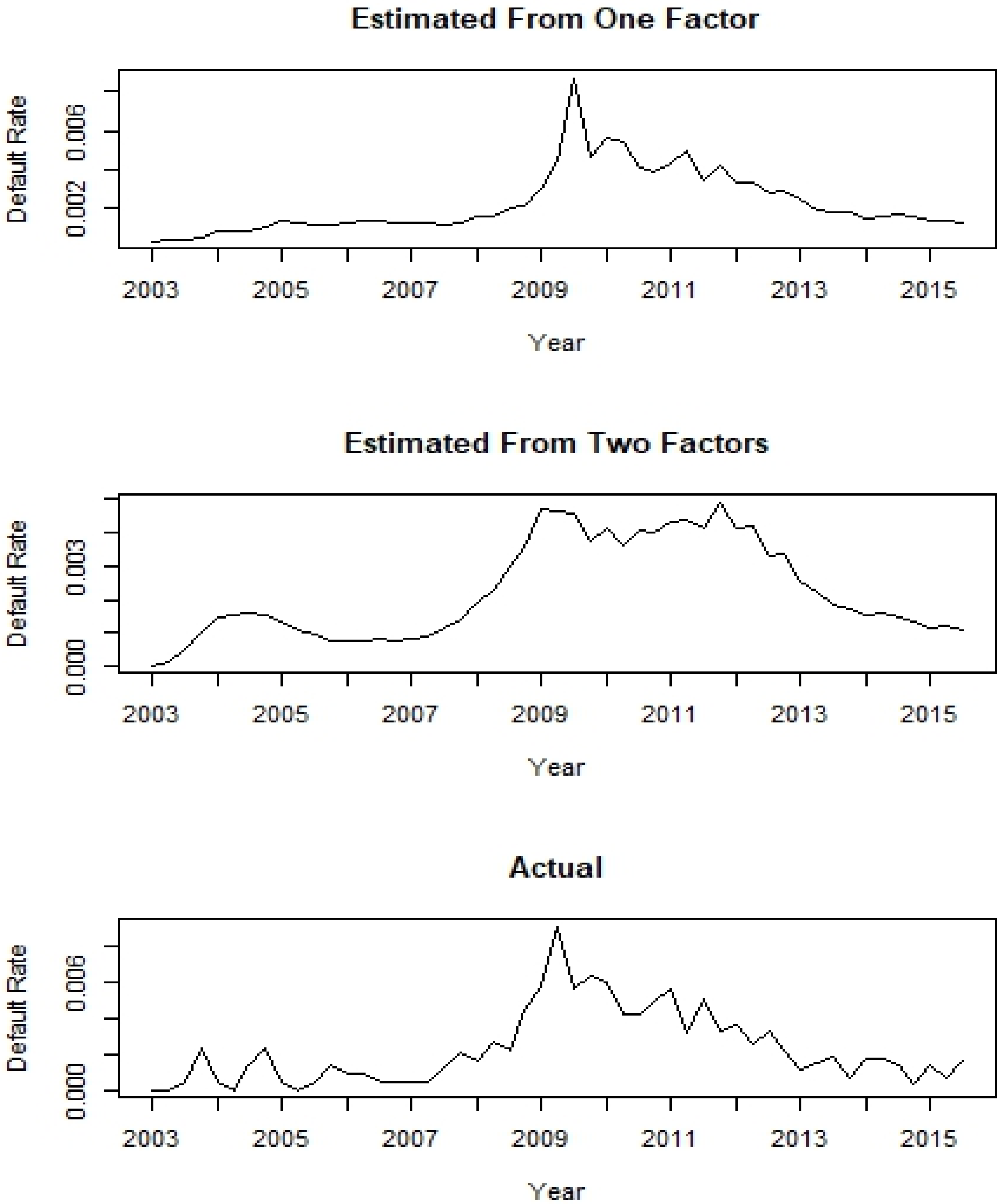
This figure shows the resulting score and factor series from calibrating a mix-measurement GAS model with one macroeconomic factor and one frailty to the complete sample.

Figure 3: Scores and Factors of Two Factor Model

Table 8: Unconditional Mean of Factor Series

	2003-2016	2006-2016	2008-2016	2010-2016	2012-2016	2014-2016
$E[f_t]$	17.80	31.09	6.24	1.91	8.89	7.37

This table shows the unconditional mean of the resulting factor series from calibrating the logit-normal one factor GAS model to samples of increasing length.



This figure shows three default rate series. The series on the top is generated with teh mix-measurement GAS model with one macroeconomic latent factor. The series in the middle is generated with the mix-measurement GAS model with one macroeconomic latent factor and one frailty latent factor. The series at the bottom is the observed default rate per quarter in the 10,000 loan sample used to calibrate the models.

Figure 4: Comparison of Default Rate Series

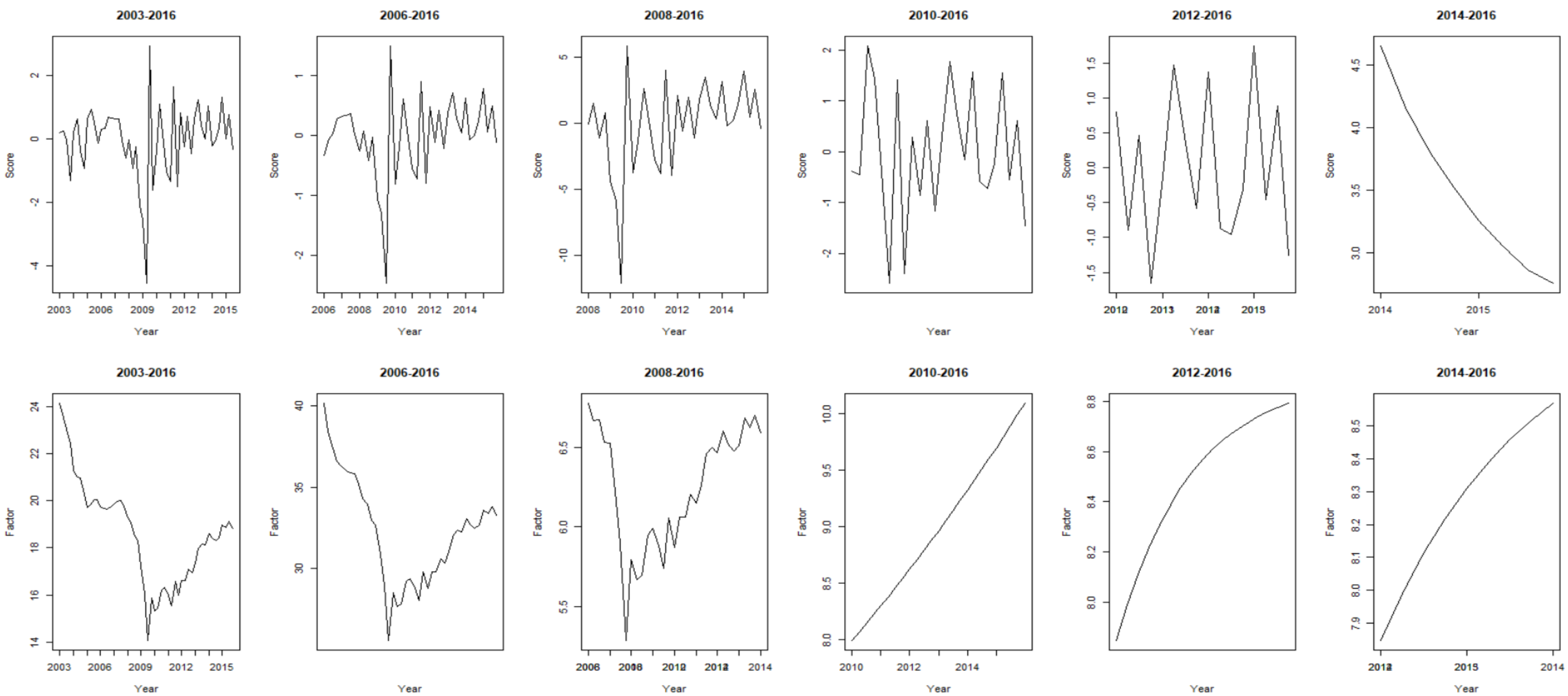
The results of fitting the latent risk factor model to samples of increasing length are presented in table 6. In order to check the correctness of the results generated by the maximum likelihood estimation, the presence of two theoretical properties of the model have been checked in every occasion. Firstly, it has been checked that the score series generated using the obtained estimates is centered around its unconditional mean, zero. Secondly, it has been checked that the factor series is centered around its unconditional mean as derived in appendix A.2.1.

Figure 5 shows the resulting score and factor series from calibrating the mix-measurement one factor GAS model introduced in section 3 to samples of increasing length. As it can be seen from the figure, with exception of the sample covering the 2014-2016 period only, the score series are centered around the theoretical mean zero. The estimated unconditional means of the factor series are presented in table 8. Comparing the estimated unconditional means with the factor series in figure 5, it can be seen that only the factor series resulting from calibrating the logit-normal one factor GAS model to samples covering the crisis are centered around their theoretical mean. This result can be interpreted as a misspecification issue arising from fitting a mix-measurement factor model that aims to identify a systematic behavior in mortgage defaults when there is none. Hence, in periods when there is no systematic default risk the calibration of the model results in a misspecified factor series. However, the factor series obtained from calibrating the model to data sets covering the economic crisis are indeed centered around their theoretical mean.

All in all, these results indicate the factor should be interpreted as a macroeconomic latent factor that captures systematic risk in defaults of residential mortgages. Comparing these results with the results obtained by Creal et al. (2014), it is worth noticing that while they found frailty factors to be important to capture non-systematic default risks, in this case it is concluded that the frailty factor does not improve the model's ability to generate accurate default rates, specially not during periods of economic distress. A possible explanation for the difference in the results observed in this paper is that the loan level covariates included in the log-odds equation capture the information provided by the frailty factor.

As it can be seen from the results, the parameter estimates from the observable and latent factor models are very different. Specially, the above mentioned pattern on the changes of the coefficient estimates for the borrower credit score and the current interest rate are not observed in the estimates obtained for the logit-normal one factor GAS model. This is an indicator that both models incorporate the additional information provided by the increasing length of the sample in a different manner. However, both models indicate towards the same results, namely the borrower's credit score and the current interest rate are the most powerful explanatory variables. Furthermore, it is seen from the results for the latent risk factor model that the loading on the factor is not significantly different from zero if the data set is short and does not incorporate an economic crisis. All in all, calibrating the observable and unobservable factor models to the data sets of increasing length has provided an insight into the effect of a panel's length in the maximum likelihood estimates. While the changes in parameters of the observable factor model follow a more intuitive logic, the changes in the maximum likelihood estimates of the unobservable factor model do not follow a very intuitive pattern.

Finally, although the results for both models differ to a large extent, it can be observed that the sign of parameter estimates that are significantly different from zero are intuitive on both models. For example, higher FICO scores contribute to a lower probability of default, while higher debt to income, original loan to value, unpaid balance or current interest rate contribute to higher default probabilities. Overall, the results indicate that the forecasting performance of both models is probably largely attributed to the explanatory power of the FICO score and the current interest rate, and to the way each model incorporates macroeconomic information to predict default scenarios.



This figure shows the resulting score and factor series from calibrating the mix-measurement one factor GAS model to samples of increasing length.

Figure 5: Estimated Score and Factor Series for Growing Sample Lengths

5 Observable or Latent Risk Factor Model

As it has been seen in the previous section, maximum likelihood parameter estimates from probability of default models can largely depend on the length of the data set. Specially, the proportions of economic cycles covered by the length of the data can influence the value and significance of the estimated parameters. While it has already been discussed how observable and latent factor models incorporate new information in their maximum likelihood estimates, it has not been discussed yet how these models perform in changing economic environments. Therefore, in this section the performance of both models is tested in two different scenarios. In the first scenario, the models are fitted to a data set of stable economic growth, spanning the years 2003 to 2007. Then, the calibrated models are used to forecast the following two years of economic crisis out-of-sample. In the second scenario, the models are fitted to a data set spanning the years 2003 to 2014. Thus, containing a period of stable economic growth, an economic crisis and a recovery. Again, the calibrated models are used to forecast the following two years of out-of-sample. The main purpose of this analysis is to investigate how observable and unobservable factor models incorporate the available information in the parameter estimates of the model and, therefore, in the estimated probabilities of default.

The comparison between both models is performed by observing the accuracy of forecasts at different intervals, when the models are fit to in-sample periods of changing lengths. For the two-step logistic regression model, two measures of probability of default have been introduced. While the first measure depends on loan-specific covariates, the second measure incorporates exposure to systematic risk by sizing the effect of an observable macroeconomic variable in the probability of default. This PD is consider the point-in-time PD. For the mix-measurement GAS model, the procedure is similar. However, exposures to systematic risk are instantaneously captured by the latent risk factors entering the equations of the parameters in the model. Therefore, the PD estimates obtained from the unobservable risk factor model should be considered as point-in-time. Hence, the comparison of the models is realized by comparing the accuracy of estimated point-in-time PD.

The loan level and macro variables included in the calibration of the model remains the same as in the previous section. However, this time variables with parameter values that are not significantly different from zero at the 1% level are not included in the forecasting models. To evaluate the performance of the probability of default models, the predictive accuracy and sensitivity of the models are determined and compared by means of the mean absolute deviation (MAD) and the area under the curve (AUC) statistics respectively. The AUC measures the area under the receiver operating characteristic (ROC) curve.

The AUC statistic is widely used in the literature to display the performance of binary classification methods such as the probability of default models, where loans are classified as either defaulting or non-defaulting (see for example Bradley (1997), Cleves & Rock (2002) or Ferri et al. (2002)). When plotted, the ROC curve shows the sensitivity (the proportion of correctly classified positive observations) and specificity (the proportion of correctly classified negative observations) as the output threshold is moved over the range of all possible values. The threshold determines whether the estimated probabilities of default should be considered an indicator for a loan defaulting or not defaulting in a specific quarter. Since the area of the ROC curve increases as sensitivity and specificity increase, a higher AUC means a better classification.

Robin et al. (2011) accentuate the importance of performing AUC comparisons by means of statistical analysis. They argue that even small differences in AUCs can be significant if ROC curves are strongly correlated, and without statistical testing two AUCs can be incorrectly labeled as similar. Similarly, as shown by Hanczar et al. (2010), large differences in AUCs can result non significant in small samples. Therefore, in order to obtain a more robust performance comparison between the forecasts generated by the observable and unobservable risk factor models, the bootstrap based statistical test by Hanley & McNeil (1983) is applied.

The test as described by Robin et al. is the following. First, start by defining Z as

$$Z = \frac{\theta_1 - \theta_2}{\mathrm{sd}(\theta_1 - \theta_2)}$$

where θ_1 and θ_2 are the two AUCs being compared. Second, the standard deviation of the difference between AUCs is computed with N bootstrap replicates. In each replicate r, the original measurements are resampled with replacement; both new ROC curves corresponding to this new sample are built. Third, the resampled AUCs $\theta_{1,r}$ and $\theta_{2,r}$ and their difference are computed. Finally, the standard deviation of the difference is computed. Given that Z approximately follows a normal distribution, one or two-tailed p-values are calculated accordingly.

In a case study, Robin et al. show that the bootstrapping based test from Hanley & McNeil can provide more robust results compared to other statistical tests such as the test by Venkatraman & Begg (1996). This is due to the importance given by these tests on comparing the differences in very specific regions of the ROC curve instead of giving the same importance to all differences across the complete ROC curve. Furthermore, Robin et al. accentuate that since the test by Hanley & McNeil depends on the variance of the difference in AUCs and the variance itself depends on the covariance of the ROC curves, strongly correlated ROC curves can have similar AUC values and still be significantly different. Therefore, it is important to assess the results of the test critically.

While the AUC can provide an insight into the ability of the model to correctly classify the loans within the panel, it does not provide an aggregated overview of the performance of the model. The mean absolute deviation (MAD) on the other hand can provide an overview of the performance of the model at an aggregate level. This becomes specially important at the moment of determining accurate loss forecasts for a portfolio of loans. Therefore, the forecasting error at an aggregate level is measured by the MAD given as

$$MAD = \frac{1}{T} \sum_{t=1}^{T} |r_t - \hat{r}_t|, \tag{18}$$

where r_t is the realized default rate and $\hat{r}_t$ is the predicted default rate. Table 9 presents the obtained AUC and MAD statistics from the predictions generated by the observable and unobservable risk factor models for their corresponding eight upcoming quarters.

All in all, the forecasting performance comparison between the two-step logistic regression model and the logit-normal one factor GAS model indicate the following. First, given the results from the statistical test on the difference of the estimated AUCs, it is clear that neither model dominates the other. As mentioned above, the AUC statistic measures the ability of a model to correctly classify the borrowers in defaulting or not defaulting. Given that both models have been calibrated using the same loan level covariates, it is plausible that both models performed very similar in terms of their classification ability. However, at an aggregate level, the computed MAD statistics indicate that the latent factor model can outperform the observable factor model, specially while forecasting default rates under changing economic conditions as shown by its lower MAD during the economic crisis.

Table 9: Area Under the Curve and Mean Absolute Deviation

	2008 Q1	2008 Q2	2008 Q3	2008 Q4	2009 Q1	2009 Q2	2009 Q3	2009 Q4
AUC (obs.)	0.60	0.73	0.84	0.82	0.74	0.80	0.77	0.84
AUC (unobs.)	0.59	0.82	0.82	0.84	0.68	0.69	0.75	0.73
Hanley & McNeil	0.92	0.43	0.802	0.11	0.27	0.01	0.66	0.00
	2014 Q1	2014 Q2	2014 Q3	2014 Q4	2015 Q1	2015 Q2	2015 Q3	2015 Q4
AUC (Obs.)	0.84	0.53	0.80	0.74	0.89	0.82	0.90	0.93
AUC (Unobs.)	0.90	0.61	0.80	0.75	0.87	0.82	0.91	0.94
Hanley & McNeil	0.19	0.17	0.79	0.77	0.00	0.30	0.02	0.29
	2007-2009	2014-2016						
MAD (obs.)	0.0037	0.0007						
MAD (unobs.)	0.0029	0.0006						

This table shows the calculated area under the ROC curve and mean absolute deviation for the out-of-sample predictions generated by the observable and unobservable factor models. The predictive performance of the models is compared by means of an statistical tests. The results are given below the AUC and MAD statistics respectively. Hanley & McNeil's tests the null hypothesis that the true difference in AUC is equal to zero.

6 Robustness Checks

In this section, two robustness checks are performed. The checks are performed to determine how changes in the methodology undertaken in this paper could affect the conclusions. Firstly, the potential biases in the results introduced by using a 10,000 loan subsample instead of the complete sample is analyze. In order to determine the added value of using larger samples, the default rates of nine samples of increasing sizes are compared. The comparison indicates the default rates of the 10,000 loan subsample are representative of the default rates observed in the complete sample. Secondly, the out-of-sample performance of the mix-measurement GAS model is checked for robustness by choosing a different definition for the scaling matrix in equation (10), namely $S_t = I$.

6.1 Subsamples

The empirical analysis presented in the previous section has been performed using a subset of 10,000 randomly selected loans. In this section, it is argued that the results obtained by analyzing this subset can be regarded as being representative of the results that would have been obtained by using the complete loan panel. In order to demonstrate this, the properties of nine sub-samples of increasing sizes are investigated. The smallest sub-sample consists of the same 10,000 loans used for the analysis presented above. The remaining sub-samples are constructed of 25,000, 50,000, 75,000, 100,000, 200,000, 300,000, 400,000 and 500,000 randomly selected loans respectively.

When working with sub-samples of the original data set it is important that the selected observations are representative of the complete sample. Specially, given the focus of this paper on the modeling of probability of default, it is essential that the default rates of any constructed sub-sample resembles the default rates observed in the complete sample. Figure D.3 shows the default rates of the nine different sub-samples. The sub-samples have been constructed by selecting loans from the complete pool of loans originated between 2000 and 2016 using a uniform random selection procedure without replacement. As it can be seen the time-series resemble the default rates of the complete pool of loans shown in Figure 1. Specially, it can be noticed that the constructed sub-samples maintain the same default rate levels and capture the strong increase and slow decrease in default rates during and after the crisis of 2008. This is a clear indicator that by using either of the sub-samples the maximum likelihood estimation is based on default series with the same properties and, therefore, the results obtained with the 10,000 loan sub-sample can be considered representative of the results that would be obtained by using the complete sample.

6.2 Scaling Matrix

The scaling matrix, S_t, is an important component of the mix-measurement GAS model presented in section 3. As it is mentioned by Creal et al. (2013), different choices for the scaling matrix S_t result in different GAS models, whose statistical and empirical properties can differ. This supports the need to perform a robustness check to determine to what extent the results obtained in the previous section can change by changing the choice of the scaling matrix.

Hence, the performance of the two-step logistic regression model with observable systematic risk factor is compared again with the performance of the mix-measurement one factor GAS model, but this time the scaling matrix is set equal to the identity matrix and there is no scaling. This choice for the scaling matrix is common in the application of GAS models. As an example, the paper by Ardia et al. performs a simulation and forecasting exercise using this specification for the scaling matrix. In table D.3 the out-of-sample forecasting performance of the latent factor model with no scaling is compared with the performance of the logistic regression model with observable systematic risk factors. As it should be expected, the obtained AUC for the latent factor model remains unchanged. Recall that the AUC statistic provides an insight into the ability of the model to correctly classify a borrower as defaulting or not defaulting. Since all borrowers are affected by the latent factor in the same way, changes in the factor series should only have an effect on the model's performance at the aggregate level. The mean absolute deviation statistic measures the ability of the model to generate default rates that are similar to the default rates observed over a certain period of time. Therefore, this metric measures the performance of the model at an aggregate level. As it can be seen from the table, the choice of the scaling matrix can have a big impact on the forecasting performance of the model over certain periods.

In this case, applying the mix-measurement GAS model with no scaling to forecast default rates during the crises has resulted in default rate forecasts that are much higher than the default rates observed in the data. The inaccuracy in the generated default rates is reflected by the high MAD during this period. The accuracy of forecasts generated by the model with identity scaling matrix for the period after the financial crisis is very similar to the accuracy of forecasts generated by the model with the scaling matrix defined as in equation (10). Overall, these results indicate that the choice for the scaling matrix does play an important role on the out-of-sample forecasting performance of the mix-measurement one factor GAS model as implemented in this paper. Furthermore, the scaling matrix proposed by Creal et al. (2014) seems to be a good choice to account for a sudden systematic behavior in borrower defaults.

7 Conclusion

In this paper, two research questions have been investigated. The first question addresses the potential bias introduced by the data set of residential mortgages not spanning different states in the economic cycle. The second question addresses the potential accuracy gains of relaxing the assumption that all systematic risk can be captured by macroeconomic variables such as the unemployment rate, house price index, gross domestic product or industrial production growth. In order to investigate the first research question two steps have been taken. First, a simulation study has been performed to determine the bias in maximum likelihood estimates when model selection bias is controlled for. Second, the two models discussed in this paper have been fitted to samples of increasing sizes that covered one to several states of the economy.

In order to investigate the second research question, the forecasting accuracy of an observable factor model and a latent factor model have been compared in two settings. In the first setting the models are calibrated with observation from 2003 to 2007 and used to forecast PD during the financial crisis. In the second setting the models are calibrated with observations from 2003 to 2014 and used to forecast PD for the eight upcoming quarters. The latent factor model is based on the mix-measurement observation driven dynamic factor model introduced by Creal et al. (2014) and the two-step observable factor model is based on the model introduced by Lee et al. (2016).

The implementation of the latent factor model in this paper differs from the implementation on the paper by Creal et al. in three ways. First, Creal et al. apply the model to a data set of macroeconomic indicators, corporate rating transitions and losses at default. In this paper, the model has been applied to a data set of macroeconomic indicators and loan level observations from residential mortgages. Second, the model introduced by Creal et al. does not incorporate company specific information to estimate a company's PD. In this paper, the loan level data provided by Fannie Mae is used to determine borrower specific PD. Third, Creal et al. model transition probabilities using a multinomial distribution. In this paper, default probabilities are modeled using a binomial distribution. In the literature review, no research paper could be found that has implemented the mix-measure observation driven dynamic factor model as proposed by Creal et al. in the setting of PD modeling of residential mortgages. Therefore, the mathematical derivations in this paper open the possibility of applying the recently developed GAS models to model probabilities of default and calculating regulatory capital of retail portfolios.

Overall, the results of the simulation study and the empirical analysis indicate towards clear differences in the way the observable and latent factor models presented in this paper incorporate the additional information provided by the length of the data set. On the one hand, it has been observed that the observable factor model incorporates the additional information by means of intuitive changes in the maximum likelihood estimates. Specially, the intercept in equation (4) decreased to incorporate the higher default rates observed during the crisis. On the other hand, the changes in maximum likelihood estimates for the latent factor model are not as intuitive as the changes in the estimates of the observable factor model. Nonetheless, the maximum likelihood estimates of both models vary as the length of the loan panel increases.

All in all, it is concluded that for observable factor models as the one presented in this paper, the length of the data does play an important role determining the value of the parameter estimates. Specially, it has been shown that the changes in parameter estimates can be explained by changes in macroeconomic conditions. However, for the latent factor model it has not been possible to link the changes in parameter values to changes in macroeconomic conditions directly. Hence, given the empirical and simulation results for both models, it can be concluded that large data sets do not necessarily improve the quality of the models. In fact, the length of the data set used to calibrate the model should be determined depending on the specific interests of the researcher. If the interest is to build a model that can generate accurate forecasts during a specific economic state, then such a state should be present in the data set. However, if the interest of the researcher is to calibrate a model with the least amount of bias, then the data should contain good, bad and intermediary economic conditions in equal amounts. This is comparable to calibrating the model to a period of economic recovery.

Given the unintuitive change in parameter estimates of the latent factor model as the panel's length increases, it could have been expected that the model would not out perform the observable risk factor model while prediction defaults out-of-sample. However, the results of their predictive performance indicate towards two results. First, the explanatory power of significant covariates is used by both models in the same way to classify borrowers as defaulting or non-defaulting. This has been shown by comparing the out-of-sample classification power of the models using the area under the curve and statistical test proposed by Hanley & McNeil (1983). Furthermore, by comparing the computed MADs of the observable and latent risk factor models it is concluded that the latent factor model can perform relatively better than the observable risk factor model while forecasting default rates of a loan panel during states of economic distress. Hence, although the two-step logistic regression model can be considered more intuitive, the mix-measurement observation driven GAS model has proven to provide the same classification power by means of the inclusion of loan level information in equation (14) while performing better at an aggregate level.

All in all, in this paper it has been shown that the results of maximum likelihood estimation can largely depend on the length of the data set and the assumptions on how non-idiosyncratic risk should be incorporated in the modeling procedure. On the one hand, the parameter estimates from the two-step logistic regression model have shown to change in a rather intuitive manner as the length of the the data set increases to incorporate periods of increasing and decreasing systematic risk. On the other hand, the parameter estimates from the logit-normal GAS model change in a less intuitive manner as the length of the data set increases. However, from the comparison of the out-of-sample forecasting performance of both models, it has been shown that the logit-normal one factor GAS model can perform significantly better when forecasting periods with changing macroeconomic conditions or perform at least as good as the two-step logistic regression model when forecasting periods with non-changing macroeconomic conditions. Therefore, even though the two-step logistic regression model can result to be more intuitive, if the objective is the accurate prediction of probabilities of default when the data set at hand does not incorporate several parts of an economic cycle, then the logit-normal one factor GAS model should be considered more adequate than the two-step logistic regression model.

8 Further Research

As it has been seen in this paper, mix-measurement observation driven GAS models are a rather new and flexible type of models that can be relatively easily adapted or extended. The flexibility in the choice of the underlying distributions creates the opportunity for mix-measurement models to be applied in a large variety of settings. In this case, the model has been used to model probabilities of default for residential mortgages. However, it would be interesting to research the application of this model to other type of problems relevant to the financial industry.

In the context of the research here presented, it would be interesting to test the results obtained in this for the mix-measurement GAS factor model by comparing the latent factors resulting from two different calibration exercises. In this first case, the model can be calibrated by incorporating the loan level information as indicated in section 3. In the second case, no loan level covariates are included in the model. Then, of interest would be to compare the results from the estimated latent factors. Given the results presented in this paper, it could be expected that at least one of the factors captures the systematic default risk. However, if for the model not including loan level covariates a second factor happens to be relevant to capture other latent non-systematic default risk, then it would be of interest to determine how this factor is linked to the borrower specific information that is not included in the model.

Furthermore, as it has been seen in the simulation exercise for the mix-measurement dynamic factor model presented above, the estimate for the intercept, ω, have shown to be highly biased in at least one of the cases considered in section 3. In the empirical analysis from section 4 is mentioned that the parameter estimates obtained from including an intercept in the log-odds equation are implausible. By removing the intercept from the equation, i.e. by assuming it is equal to zero, it has been possible to obtain reasonable parameter estimates. Given the results from the simulation exercise, it would be interesting to perform the same simulation exercise without including ω in the equation for p_t^*. If the bias from the parameter estimates obtained from this simulation results to be lower than that of the parameter estimates from the simulation performed in this paper, then this could be an indicator that including an intercept in these type of models can lead to misspecification.

Regarding the forecasting performance of the two-step logistic regression and the logit-normal one factor GAS models, it has been shown that both models can perform very similar while forecasting periods of non-changing economic conditions, but they can also perform rather differently while forecasting periods of changing economic conditions. In this paper, the analysis of the forecasting performance of the models has been limited to these two settings. In the first setting, the models where calibrated with a data from periods of economic stability. In the second setting, the models where fitted to a larger data set covering more parts of the economic cycle. In conclusion, the logit-normal one factor GAS model has been considered a better model due to its ability to provide more accurate forecasts in periods of changing macroeconomic conditions. However, the performance comparison across the models could easily be extended in order to obtain a more robust conclusion on whether either model dominates the other. This comparison could be extended, for example, by calibrating the models with a data set covering a complete economic cycle and then use the results to forecast a crisis, a stable growth period or a recovery out-of-sample.

A Maximum Likelihood Estimation

A.1 Logistic Regression with Observable Systematic Risk Factor

Under the assumption that the covariates and the observable risk factors capture all cross-sectional and time-series heterogeneity and the assumption that the default indicator follows a Bernoulli distribution, the likelihood function of $Y = (Y_{N_1}, Y_{N_2}, ..., Y_{N_T})$ takes the following functional form

$$P(Y|x, f) = \prod_{t=1}^{T} \prod_{i=1}^{N_t} (p_{i,t|t-1})^{y_{i,t}} (1 - p_{i,t|t-1})^{1-y_{i,t}},$$

where $p_{i,t|t-1}$ is given in equation (5) and N_t denotes the number of mortgages at time $t = 1, ..., T$. The maximization procedure is as follows

1. Define set of initial parameter values Θ^0 for the parameters in equation (4).
2. Obtain the through-the-cycle probability of default for each loan at every point in time, $P(Y_{it} = |x_{t-1}, \Theta^0)$, and compute the log-likelihood contribution of the loan observations at time t.
3. Apply quasi-Newton optimization procedure to find the maximum likelihood estimates Θ^*.
4. Compute h_{it}^{TTC*} and repeat the maximization procedure for equation (5).

A.2 Mix-measurement Observation Driven GAS Model

The maximum likelihood estimation procedure for the observation driven mixed-measurement dynamic factor model is more involved than the procedure for the multi-period logit model with observable risk factors. The aim is to obtain a solution to the maximization problem

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{t=1}^{T} \log p(y_t | f_t, F_{t-1}; \Theta),$$

where $y_t = (y_t^m, y_t^r)'$. Following the procedure by Creal et al., a factor model structure is assumed in which all $y_{i,t}$'s at time t are cross-sectionally independent conditional on f_t and the information set F_{t-1}. The maximization algorithm is as follows:

1. Define set of initial parameters values Θ^0 for the parameters in equations (11), (7), (13) and (14). f_1 is considered to be within the set of initial parameters. The unconditional mean is derived in the following section of this appendix. In order to decrease the computational intensity of the model and to avoid an explosive pattern for f_t the following constraints on the parameters are imposed: $\sum_j^q B_{mm,j} \neq 1$ for m = {1, ..., M}, B_j and A_i are diagonal for all $i = \{1, ..., p\}$ and $j = \{1, ..., q\}$. $\omega = 0$ and the non-zero columns of Z^m are lower triangular with ones in the diagonal.
2. Given y_1 and f_1, compute $s_1 = S_1 \nabla_1$. If S_t is obtained using the eigendecomposition of Fisher information matrix, then compute the information matrix and gradient of the mixed-measurement model.
3. Compute the likelihood at time $t = 1$, i.e. compute
$$\log p(y_1^m | f_t, F_{t-1}; \Theta) + \log p(y_1^r | f_t, F_{t-1}; \Theta).$$
4. Compute f_2.
5. Redo steps 1 to 3 until time $t = T$.
6. Use the scoring algorithm with quasi-Newton optimization method to find the maximum likelihood estimates Θ^*

To perform the maximization procedure it is necessary to compute functional forms for the likelihood contributions of the macroeconomic component and the logistic component as well as their corresponding score and Fisher information matrix.

A.2.1 Unconditional Mean of f_t and Covariance Stationarity

For the adaptive sequence $\{s_t, F_t\}_{\infty}^{\infty}$ on a probability space $(\Omega, F, \mathbb{P})$, s_t is said to be a martingale difference sequence if the unconditional mean of the sequence is smaller than infinity, $\mathrm{E}[s_t] < \infty$, and its conditional mean with respect to the filtration up to time $t-1$ is equal to zero, $\mathrm{E}[s_t|F_{t-1}] = 0$. The unconditional mean of the latent factor process, f_t, can then be obtained using the properties of the martingale difference as follows

$$E[f_{t+1}] = \omega + \sum_{i}^{q} A_i E[s_{t-j+1}] + \sum_{j}^{q} B_j E[f_{t-j+1}]$$

$$\mu = (\mathrm{I} - \sum_{i=1}^{q} B_j)^{-1}\omega,$$

where the second equality follows from the fact that the scale score is a function of past observations, factors and unknown parameters, i.e. $\mathrm{E}[s_t] = \mathrm{E}[S_t \nabla_t] = \mathrm{E}[s_t|F_{t-1}]$. This results indicates that for the unconditional mean to exist the M x M matrix $(\mathrm{I} - \sum_{i=1}^{q} B_j)$ must be invertible. By definition a square matrix is invertible if and only if its determinant is nonzero. Straight forward algebra demonstrates that the determinant is zero if at least one of the diagonal entries in the matrix $\sum_{i=1}^{q} B_j$ is equal to one. Therefore, for the first moment to be constant it must hold $\sum_{i=1}^{q} B_{mm,j} \neq 1$ for any $m = \{1, ..., M\}$.

A.2.2 Log-likelihood Contribution: Normal Component

Since the macroeconomic variables with normal density are observable at every quarter, their contribution to the log-likelihood at time t can be obtained in multivariate form as follows

$$\log\, p(y_t^m|\mu_t, \Sigma_m) = \log\, \frac{1}{\sqrt{|2\pi\Sigma_m|}} \exp\Big(-\frac{1}{2}(y_t^m - \mu_t)'\Sigma_m^{-1}(y_t^m - \mu_t)\Big)$$

$$= \mathrm{const} - 0.5 \log\, |\Sigma_m| - 0.5(y_t^m - \mu_t)'\Sigma_m^{-1}(y_t^m - \mu_t),$$

where y_t^m is an M x 1 column vector with observations of the four macroeconomic variables at time t.

A.2.3 Score and Fisher Information Matrix: Normal Component

In the derivation for the gradient of the normal component a set of matrix differentiation results are applied. Specially, the following two equations are used. Given an m x n matrix A, an m x 1 vector y and an n x 1 vector x,

$$\frac{\partial y'x}{\partial y} = \frac{\partial x'y}{\partial y} = x \qquad \text{and} \qquad \frac{\partial x'Ax}{\partial x} = (A + A')x.$$

Proofs for these equations can be found in Barnes (2006) and Petersen et al. (2008). The gradient of the normal component is derived as follows

$$\nabla_t^m = \frac{\partial \log\, p(y_t^m|\mu_t, \Sigma_m;\Theta)}{\partial f_t} = -0.5\frac{\partial}{\partial f_t}(y_t^m - \mu_t)'\Sigma_m^{-1}(y_t^m - \mu_t)$$

$$= -0.5\frac{\partial}{\partial f_t}y_t^{m'}\Sigma_m^{-1}y_t^m + 0.5\frac{\partial}{\partial f_t}(Z^m f_t)'\Sigma_m^{-1}y_t^m + 0.5\frac{\partial}{\partial f_t}y_t^{m'}\Sigma_m^{-1}(Z^m f_t) - 0.5\frac{\partial}{\partial f_t}(Z^m f_t)'\Sigma_m^{-1}(Z^m f_t)$$

$$= 0.5Z^{m\prime}\Sigma_m^{-1}y_t^m + 0.5Z^{m\prime}\Sigma_m^{-1}y_t^m - (Z^m f_t)'\Sigma_m^{-1}(Z^m)$$

$$= Z^{m\prime}\Sigma_m^{-1}(y_t^m - \mu_t)$$

To obtain Fisher information matrix recall that under suitable smoothness conditions on the density of y_t^m, which are satisfied by the exponential families, the information matrix is equal to

$$\mathrm{I}_t^m = \mathrm{E}_{t-1}\Big[\nabla_t^m \nabla_t^{m'}\Big] = -\mathrm{E}_{t-1}\Big[\frac{\partial^2 \log\, p(y_t^m|\mu_t, \Sigma_m)}{\partial^2 f_t}\Big]. \tag{19}$$

A proof for this statement is given by Lehmann & Casella (2006). Using this result, it is straight forward to compute the information matrix

$$\mathrm{I}_t^m = \mathrm{I}^m = Z^{m'}\Sigma_m^{-1}Z^m$$

A.2.4 Log-likelihood Contribution: Logit Component

The log-likelihood contribution of the logit component at time t under the assumption of cross-sectionally independence given f_t and F_{t-1} is obtained as follows

$$\log\, p(y_t^r|f_t, F_{t-1};\Theta) = \sum_{i=1}^{L_t}\log\, p(y_{it}^r|\pi_{it}) = \sum_{i=1}^{L_t}\log\, (\pi_{it})^{y_{it}}(1-\pi_{it})^{1-y_{it}}$$

$$= \sum_{i=1}^{L_t}\log\,(1-\pi_{it}) + y_{it}\log\,\frac{\pi_{it}}{1-\pi_{it}}$$

$$= \sum_{i=1}^{L_t} y_{it}(\beta_0 + \sum_j \beta_j x_{ijt-1} - Z^{r'}f_t) - \log\,(1+\exp(\beta_0 + \sum_j \beta_j x_{ijt-1} - Z^{r'}f_t)),$$

where L_t is equal to the number of performing or defaulting loans at time t.

A.2.5 Score and Fisher Information Matrix: Logit Component

The gradient of the logit component is obtained as follows

$$\nabla_t^r = \sum_{i=1}^{L_t}\frac{\partial}{\partial f_t}y_{it}^r(\beta_0 + \sum_j \beta_j x_{ijt-1} - Z^{r'}f_t) - \sum_{i=1}^{L_t}\frac{\partial}{\partial f_t}\log\,(1+\exp(\beta_0 + \sum_j \beta_j x_{ijt-1} - Z^{r'}f_t))$$

$$= \sum_{i=1}^{L_t}(\pi_{it} - y_{it}^r)Z^r$$

Fisher information matrix for the for the density of y_t^r is derived under the assumption that given the information up to time $t-1$, y_{it}^r and y_{lt}^r, with $i, l \in 1, ..., L_t$ and $i \neq j$, are independent. The calculations are as follows

$$\mathrm{I}_t^r = \mathrm{E}_{t-1}\Big[\nabla_t^r \nabla_t^{r'}\Big] = \mathrm{E}_{t-1}\Big[\sum_{i=1}^{L_t}(y_{it}^r - \pi_{it})Z^r \sum_{i=1}^{L_t}(y_{it}^r - \pi_{it})Z^{r'}\Big]$$

$$= Z^r Z^{r'} \ \mathrm{E}_{t-1}\Big[\sum_i^{L_t} y_{it} \sum_j^{L_t} y_{jt} - 2\sum_i^{L_t} \pi_{it} \sum_i^{L_t} y_{it} + \sum_i^{L_t} \pi_{it} \sum_i^{L_t} \pi_{it}\Big]$$

$$= Z^r Z^{r'} \Big[\ \mathrm{E}_{t-1}\Big[\sum_i^{L_t} y_{it}^2\Big] + \ \mathrm{E}_{t-1}\Big[\sum_i^{L_t} y_{it} \sum_{j \neq i}^{L_t} y_{jt}\Big] - (\sum_i^{L_t} \pi_{it})^2\Big]$$

$$= Z^r Z^{r'} \Big[\sum_i^{L_t} \pi_{it} + \sum_i^{L_t} \pi_{it} \sum_{j \neq i}^{L_t} \pi_{jt} - (\sum_i^{L_t} \pi_{it})^2\Big]$$

where the third equality follows from

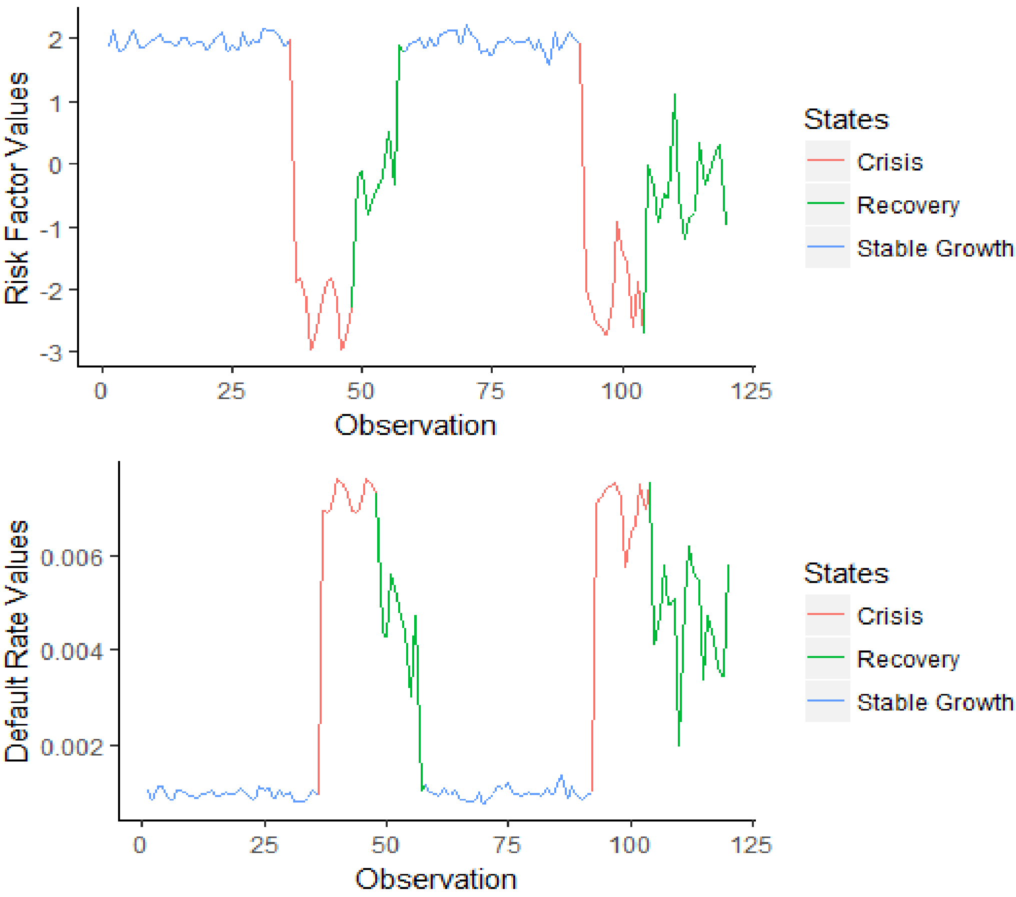
$$Var(y_{it}) = E[y_{it}^2] - E[y_{it}]^2$$

$$\pi_{it}(1 - \pi_{it}) = E[y_{it}^2] - \pi_{it}^2$$

$$\pi_{it} = E[y_{it}^2]$$

B Simulation Results

B.1 Logistic Regression with Single Systematic Risk Factor



This figure shows the simulation of an economic cycle. Sections in blue correspond to stable economic growth, sections in red correspond to an economic crisis and sections in green correspond to an economic recovery. In total the cycle consists of 120 observations, which correspond to two periods of stable growth, two economic crises and two periods of economic recovery.

Figure B.1: Simulation of Economic Cycle

B.2 Logistic Regression with GAS Component: Log-Likelihood and Score Function

The log-likelihood at time $T = t$ is derived as follows

$$\log\, p(y_t|f_t, F_{t-1}; \Theta) = \log\, (p_t)^{y_t}(1-p_t)^{1-y_t} = \log\, (1-p_t) + y_t \log\, \left(\frac{p_t}{1-p_t}\right)$$

$$= log(1-p_t) + y_t \log\, \left(\frac{\tau}{1-\tau+e^{-p_t}}\right)$$

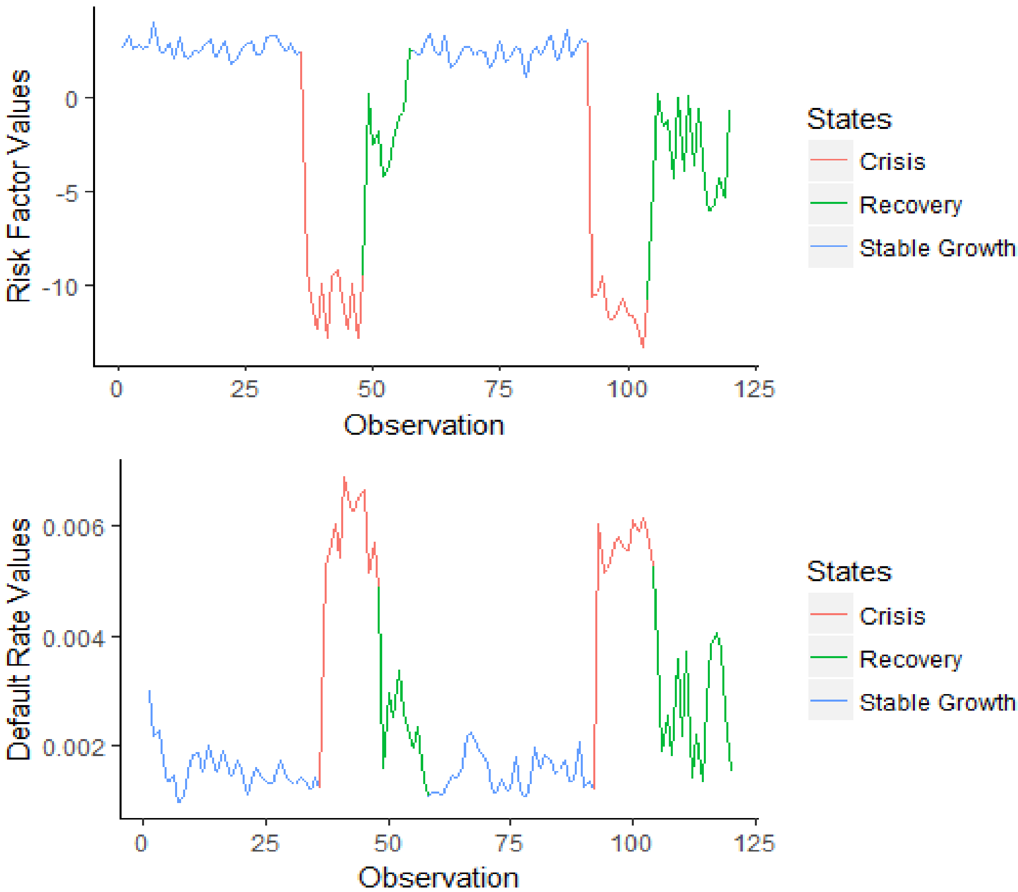
$$p = log(1+(1-\tau)e^{p}) - log(1+e^{p}) + y_t \log\, (\tau) - y_t \log\, (1-\tau+e^{-p})$$

The corresponding score function is derived as follows

$$\frac{\partial \log\, p(y_t|f_t, F_{t-1}; \Theta)}{\partial f_t} = \frac{\partial}{\partial f_t}\log\, (1+(1-\tau)e^{p}) - \frac{\partial}{\partial f_t}\log\, (1+e^{p}) + \frac{\partial}{\partial f_t} y_t \log\, (\tau) - \frac{\partial}{\partial f_t} y_t \log\, (1-\tau+e^{-p})$$

$$= \left(\frac{(1-\tau)e^{p}}{1+(1-\tau)e^{p}} - \frac{e^{p}}{1+e^{p}} + y_t \frac{e^{-p}}{1-\tau+e^{-p}}\right)\alpha$$

B.3 Logistic Regression with GAS Component



This figure shows the simulation of an economic cycle using a logistic regression with a GAS component and an observable risk factor. Sections in red correspond to stable economic growth, sections in green correspond to an economic crisis and sections in blue correspond to an economic recovery. In total there are 120 observations, which correspond to two periods of stable growth, two economic crises and two periods of economic recovery.

Figure B.2: Simulation of Economic Cycle II

C Data Properties

Table C.1: Loan-Specific Covariates and Macroeconomic Variables

Variable (Covariate)	**Type**	**Description**
Current Interest Rate	Performance	The interest rate on a mortgage loan in effect for the periodic installment due.
Original Unpaid Principal Balance (UPB)	Acquisition	The original amount of the mortgage loan as indicated by the mortgage documents.
Original Loan-to-Value (OLTV)	Acquisition	A ratio calculated at the time of origination for a mortgage loan. The Original LTV reflects the loan-to-value ratio of the loan amount secured by a mortgaged property on the origination date of the underlying mortgage loan.
Debt-to-Income	Acquisition	A ratio calculated at origination derived by dividing the borrower's total monthly obligations (including housing expense) by his or her stable monthly income
Borrower Credit Score	Acquisition	The "classic" FICO score developed by Fair Isaac Corporation.
House Price Index (HPI)	-	House price trends at various geographic levels obtained from the website of the Federal Housing Finance Agency.
Unemployment Rate	-	Quarterly percentage change in seasonality adjusted unemployment rate for persons aged 15-64 in the United States.
Gross Domestic Product (GDP)	-	Quarterly percentage change in seasonality adjusted real GDP from the United States
Industry Production Rate	-	Quarterly percentage change in seasonally adjusted total industry production excluding construction for the United States
Default Indicator	-	An indicator whether a loan is not performing on next quarter, given that it is performing on the previous quarter.

This table provides descriptions of loan level covariates and macroeconomic variables that are relevant for the calibration of the models presented in this paper.

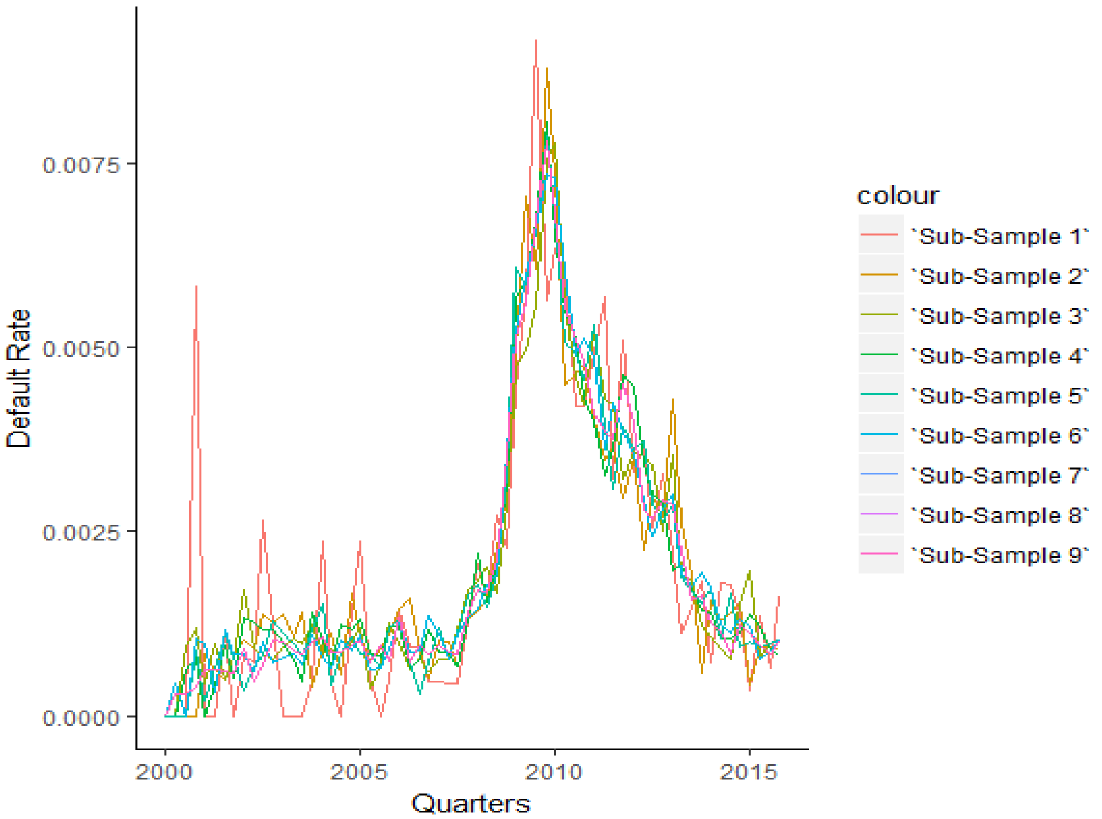
Table C.2: Aggregation of Observations at Yearly Level

Years	$\mathbf{N}_t$	$\mathbf{D}_t$	$\mathbf{D}_t/\mathbf{N}_t$
2000	1,666,781	601	0.04%
2001	6,928,747	4,876	0.07%
2002	13,230,878	11,263	0.09%
2003	18,317,844	17,436	0.10%
2004	25,701,544	23,955	0.09%
2005	20,759,994	18,412	0.09%
2006	21,618,846	20,213	0.09%
2007	22,854,307	23,029	0.10%
2008	25,179,023	56,362	0.22%
2009	27,517,040	165,847	0.60%
2010	28,039,994	155,361	0.55%
2011	27,431,470	112,606	0.41%
2012	26,723,489	84,572	0.32%
2013	26,287,889	52,771	0.20%
2014	27,332,051	33,442	0.12%
2015	28,898,354	27,299	0.09%
Total	348,478,251	808,045	0.23%

This table shows the number of mortgage observations (N), the number of defaults (D) and the default rate (D/N) per observation year. Yearly figures are obtained as aggregates of quarterly observations.

D Robustness Check

D.1 Subsamples



This figure shows the default-to-performing ratio at each quarter for nine sub-sampled data sets with increasing number of loans. Sub-samples 1 to 9 consist of 10,000, 25,000, 50,000, 75,000, 100,000, 200,000, 300,000, 400,000 and 500,000 unique loans randomly sampled from the complete set of 23,902,488 single-family loans respectively.

Figure D.3: Sub-sampled Data sets

D.2 Scaling Matrix

Table D.3: Area Under the Curve and Mean Absolute Deviation II

	2008 Q1	2008 Q2	2008 Q3	2008 Q4	2009 Q1	2009 Q2	2009 Q3	2009 Q4
AUC (obs.)	0.60	0.73	0.84	0.82	0.74	0.80	0.77	0.84
AUC (unobs.)	0.59	0.82	0.82	0.84	0.68	0.69	0.75	0.73
Hanley & McNeil	0.92	0.43	0.802	0.11	0.27	0.01	0.66	0.00
	2014 Q1	2014 Q2	2014 Q3	2014 Q4	2015 Q1	2015 Q2	2015 Q3	2015 Q4
AUC (Obs.)	0.84	0.53	0.80	0.74	0.89	0.82	0.90	0.93
AUC (Unobs.)	0.90	0.61	0.80	0.75	0.87	0.82	0.91	0.94
Hanley & McNeil	0.19	0.17	0.79	0.77	0.00	0.30	0.02	0.29
	2007-2009	2014-2016						
MAD (obs.)	0.0037	0.0007						
MAD (unobs.)	0.0494	0.0006						

This table shows the calculated area under the ROC curve and mean absolute deviation for the out-of-sample predictions generated by the observable and unobservable factor models. The scaling matrix of the mix-measurement GAS model is set to the identity matrix. The predictive performance of the models is compared by means of an statistical tests. The results are given below the AUC and MAD statistics respectively. Hanley & McNeil's tests the null hypothesis that the true difference in AUC is equal to zero.

References

Ambrose, B. W., & Capone, C. A. (2000). The hazard rates of first and second defaults. *The Journal of Real Estate Finance and Economics*, *20*(3), 275–293.

Ardia, D., Boudt, K., & Catania, L. (2016). Generalized autoregressive score models in r: The gas package.

Barnes, R. J. (2006). Matrix differentiation.

Bellotti, T., & Crook, J. (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society*, *65*(3), 340–350.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145–1159.

Cleves, M. A., & Rock, L. (2002). From the help desk: Comparing areas under receiver operating characteristic curves from two or more probit or logit models. *Stata J*, *2*(3), 301–13.

Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., ... Lauritzen, S. L. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 93–115.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data* (Vol. 21). CRC Press.

Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, *28*(5), 777–795.

Creal, D., Schwaab, B., Koopman, S. J., & 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.

Deng, Y. (1997). Mortgage termination: An empirical hazard model with a stochastic term structure. *The Journal of Real Estate Finance and Economics*, *14*(3), 309–331.

Donchev, T. (2008). Modeling defaults in residential mortgage backed securities: An intensity based approach.

Ferri, C., Flach, P., & Hernández-Orallo, J. (2002). Learning decision trees using the area under the roc curve. In *Icml* (Vol. 2, pp. 139–146).

Hamerle, A., Liebig, T., & Rösch, D. (2003). Credit risk factor modeling and the basel ii irb approach.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of roc-related estimates. *Bioinformatics*, *26*(6), 822–830.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*(3), 839–843.

Keijsers, B., Diris, B. F., & Kole, E. (2017). Cyclicality in losses on bank loans.

Koyluoglu, H. U., & Hickman, A. (1998). A generalized framework for credit risk portfolio models. *Manuscript, Oliver Wyman & Company*.

Lee, Y., Rösch, D., & Scheule, H. (2016). Accuracy of mortgage portfolio risk forecasts during financial crises. *European Journal of Operational Research*, *249*(2), 440–456.

Lehmann, E. L., & Casella, G. (2006). *Theory of point estimation.* Springer Science & Business Media.

Neter, J. (2004). *Applied linear statistical models* (Vol. 4).

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, *7*, 15.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, *12*(1), 77.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, *74*(1), 101–124.

Venkatraman, E. S., & Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, *83*(4), 835–848.

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a. *Quetelet*, *10*, 113–121.