DYNAMIC MODELS OF ASSET RETURNS AND MORTGAGE DEFAULT

Xi Chen

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research in the University of North Carolina at Chapel Hill

Chapel Hill
2017

Approved by:

Eric Ghysels

Chuanshu Ji

Jonathan Hill

Vidyadhar Kulkarni

Vladas Pipiras

ABSTRACT

XI CHEN: Dynamic Models of Asset Returns and Mortgage Default
(Under the direction of Eric Ghysels and Chuanshu Ji)

This dissertation consists of three chapters. The first chapter builds a new series of dynamic copula models and studies the influence of macro variables on the dependence between assets. The second chapter develops a dynamic logistics regression model and investigates how systematic risk affects mortgage default. The third chapter uses the frailty model developed in chapter 2 to explore spatial dependence between commercial and residential mortgage risk. In all three chapters, we extend the generalized autoregressive score (GAS) models proposed in Creal, Koopman and Lucas (2013*a*).

In the first chapter, we propose a series of dynamic copula models with a short- and long run component specification, inspired by the mixed data sampling (MIDAS) component structure applied to univariate GARCH models in Engle, Ghysels and Sohn (2013) and multivariate GARCH models in Colacito, Engle and Ghysels (2011). In particular, we extend the framework of MIDAS to dynamic copulas. In the framework of GAS models, we combine macro variables of low frequency with asset returns of high frequency, and investigate the influence of low frequency macro variables on the dependence between asset returns. Our data consists of stock portfolios and a bond. We assess the new class of models with these data and find that an extra component enhances the model with more volatility. Moreover, the macro variables with MIDAS work as a proxy for the market condition, and allow that the macro environment affects how dependence parameter reacts to innovations. With these two flexibilities, the model performance is consistently improved through our empirical applications.

In the second chapter, we design a new dynamic logistic regression model to track

systematic risk of mortgages. Specifically, we match default rates in multiple dimensions by extending the GAS models. Our data consists of commercial mortgages in the U.S. retail market from 1997 to 2013. An empirical analysis of these data suggests the influence of origination month and the originator preference on default rates. To model the effects of these variables, we group mortgages by these two variables and allow latent factors to vary by groups. Compared with GAS models using a single factor, our multi-factor models feature improved empirical fits. To the best of our knowledge, this is the first attempt that uses observation-driven models to predict mortgage defaults. We show that the new class of models has better tractability compared with parameter-driven models. For instance, although our dataset has more than two million records, and our most complex model incorporates up to 15 frailty factors, the estimation process only takes two minutes using a standard desktop computer.

In the third chapter, we use the frailty model developed in chapter 2 to explore spatial dependence between commercial and residential mortgage risk. Our dataset contains 1.6 million records of commercial mortgages and 140 million records of residential mortgages in the U.S. market. The time range of these records is between January 1999 and March 2016. Our empirical analysis demonstrates strong spatial dependence between commercial defaults and residential default in multiple respects. First, we apply Granger causality tests to the empirical default rates of commercial mortgages and residential mortgages in 10 main MSA areas, and the test results in 9 areas reveal a significant lead and lag relationship of the two mortgage markets. Second, we test the causal relation among the frailty factors that explain systematic risk of commercial mortgage and residential mortgage, and provide strong evidence on the close correlations between the residential and commercial mortgage markets. Last but not least, we show that residential PD is a good explanatory variable in predicting default of commercial mortgages in adjacent area, and this prediction power also implies that local residential market drives the commercial market. To the best of our knowledge, this is the first paper exploring the spatial dependence between commercial mortgage default and residential mortgage default.

ACKNOWLEDGEMENTS

First and foremost, I would like to offer my deepest gratitude to my advisors, Professors Eric Ghysels and Chuanshu Ji. It is my fortune to have them as my academic advisors. Their ingenious thoughts and deep insight have inspired me so much and their knowledge of finance and statistics is a huge resource for my research. They are always on my back and support me whenever needed. Besides research, they are also my personal advisors. They cares about my career and always gives me advice, from which I will benefit for the rest of my life. In a word, I learnt so much from Professors Eric Ghysels and Chuanshu Ji, and without them I would not have such a wonderful experience as a Tar Heel.

I would also like to show my deep appreciation for other committee members: Professor Jonathan Hill, Professor Vidyadhar Kulkarni, and Professor Vladas Pipiras. They provide invaluable helps in my graduate study and advice for my dissertation. Professor Jonathan Hill read my dissertation carefully and provided numerous useful comments and feedbacks. I learnt stochastic models in operations research and market dynamics from Professor Kulkarni. Professor Pipiras taught me advanced probability. I am extremely grateful to them

Lastly and most importantly, I would like to thank my parents and wife. They always love, believe and support me whatever happens and sacrifice a lot for me. I hope to make them proud of me for my accomplishment. I dedicate this dissertation to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: COMPONENT DYNAMIC COPULA MODELS WITH MIDAS

1.1 Introduction

Measuring temporal dependence between financial assets is a key ingredient to risk hedging, asset pricing, portfolio choices, to name only a few. For example, hedging ratios dynamically adjust to the varying dependence between financial assets. Likewise, the pricing of structured products such as CDO's critically relies on the dependence between the underlying financial assets.

To model temporal dependence between asset returns, two main methods have been developed in the literature. One is multivariate GARCH models and the other one is copula-based models. This chapter focuses on the latter. Copula-based models allow researchers to model the marginal distribution and dependence structure separately. This property provides a flexible framework to model multivariate time series and recently increasing attention has been devoted to conditional copulas when modeling dynamic dependence of financial assets. Patton (2006) provided the theoretical foundation for conditional copulas. He used a combination of GARCH models and copulas to model Deutsche Mark and Yen jointly. The dependence parameters of the copulas are driven by autoregressive processes. Guegan and Zhang (2010) compared two dynamic copula models and proposed statistical tests based on conditional copulas. Fengler and Okhrin (2012) utilized realized variance to model the dependence between daily stock returns, and the dynamic of the copula is driven by a HAR (Corsi, 2009) process - which is a MIDAS specification with step functions. To parameterize dynamic copulas in non-Gaussian settings, Creal, Koopman and Lucas (2013*a*) proposed to use the scores of log likelihood functions as the innovation term, and named the models as Generalized

Autoregressive Score Models (GAS).

In this chapter, we propose a series of dynamic copula models with a short- and long run component specification, inspired by the mixed data sampling (MIDAS) component structure applied to univariate GARCH models in Engle, Ghysels and Sohn (2013) and multivariate GARCH models in Colacito, Engle and Ghysels (2011). Hence, the purpose of this chapter is to extend the framework of MIDAS to dynamic copulas. In the framework of GAS models, we combine macro variables of low frequency with asset returns of high frequency, and investigate the influence of low frequency macro variables on the dependence between asset returns. Our data consists of stock portfolios and a bond. The stock data are the daily returns on five industry portfolios. The bond data are the daily returns of a 10-year Treasury bond. We assess the new class of models with these data and find that an extra component enhances the model with more volatility. Moreover, the macro variables with MIDAS work as a proxy for the market condition, and allow that the macro environment affects how dependence parameter reacts to innovations. With these two flexibilities, the model performance are consistently improved through our empirical applications.

The rest of this article is organized as follows. In the next section, we review the literature. Section 1.3 states the model formulation. In section 1.4, we describe the details of estimation. Section 1.5 discusses empirical applications. The last section provides concluding remarks.

1.2 Background

To motivate the theoretical framework, it is useful to review the literature on copulas and related areas. We first introduce the theoretical foundation of static and dynamic copulas. Next, we focus on various parameterization methods for dynamic copulas. At last, we briefly cover component models in the context of volatility forecasting.

We start with the introduction of static copulas. In particular, consider a multivariate random variable $\mathbf{Y} = [Y_1, ..., Y_n]$. Let F be the joint cumulative distribution function (CDF) for $\mathbf{Y}$, and F_i be the CDF for Y_i. By Sklar's theorem (Sklar 1959), there exists a copula function $C(\cdot) : [0,1]^n \to [0,1]$, mapping the marginal distributions of Y_i to the joint distribution through:

$$F(\mathbf{y}) = C(F_1(y_1), ..., F_n(y_n)|\rho) \tag{1.2.1}$$

where ρ is the dependence parameter of interest. Accordingly, the joint probability density function (PDF) can be represented as the product of copula density $c(\cdot)$ and marginal PDF f_i:

$$f(\mathbf{y}) = c(U_1, ..., U_n|\rho) \prod_{i=1}^{n} f_i(y_i) \tag{1.2.2}$$

where $U_i = F_i(y_i)$. Note that it is usually assumed that $c(\cdot)$ and f_i share no common parameters. In the case of (1.2.2), ρ does not appear in f_i.

If we take log of both sides of equation (1.2.2), the product of $c(U_1, ..., U_n|\rho)$ and $f_i(y_i)$ transforms to the sum of $log(c(U_1, ..., U_n|\rho))$ and $log(f_i(y_i))$. This transformation motivates a two-stage estimation: The first stage estimates the parameters of marginal distributions with the likelihoods only involving $log(f_i(y_i))$, and the second stage estimates the parameters of $c(\cdot)$ with the likelihoods only containing $log(c(U_1, ..., U_n|\rho))$. This two-stage estimation greatly reduces the computation cost for estimation, because the parameters of f_i and $c(\cdot)$ can be estimated in separate optimizations. Comparing with a joint estimation of all parameters, this two-stage estimation may entail some efficiency loss. However, as the numbers of parameters increase with problem sizes, the two-stage estimation may be the only feasible estimation method in practice.

While Skalar's theorem motivates the application of static copula models, Patton (2006) further establishes the theoretical foundation of dynamic copula models. Suppose $\mathbf{Y}_t = [Y_{1,t}, ..., Y_{n,t}]$ is a multivariate stochastic process and $\mathcal{F}_{t-1}$ is the information set up to time $t-1$. Patton (2006) showed that the conditional CDF $F(.|\mathcal{F}_{t-1})$ can be decomposed into the conditional marginal CDF $F_i(.|\mathcal{F}_{t-1})$ and conditional copula $C(.|\mathcal{F}_{t-1})$ as what follows:

$$F(\mathbf{y}_t|\mathcal{F}_{t-1}) = C(F_1(y_{1,t}|\mathcal{F}_{t-1}), ..., F_n(y_{n,t}|\mathcal{F}_{t-1})|\mathcal{F}_{t-1}) \tag{1.2.3}$$

$$f(\mathbf{y}_t|\mathcal{F}_{t-1}) = c(U_{1,t}, ..., U_{n,t}|\rho_t, \mathcal{F}_{t-1}) \prod_{i=1}^{n} f_i(y_{i,t}|\mathcal{F}_{t-1}) \tag{1.2.4}$$

where ρ_t is a dynamic dependence parameter and changes with time.

In Patton (2006) both Gaussian and non-Gaussian copulas are constructed to model the dependence between Deutsche mark and Yen. In the Gaussian case, ρ_t has the following dynamic,

$$\begin{aligned} \rho_t &= \Lambda_1(\delta_t) \\ \delta_t &= \omega + \beta\delta_{t-1} + \frac{\alpha}{10}\sum_{k=1}^{10} \Phi^{-1}(U_{1,t-k})\Phi^{-1}(U_{2,t-k}) \end{aligned}$$

where $\Lambda_1(\cdot)$ is a link function to make sure that ρ_t is between -1 and 1. δ_t is the transformed dependence parameter. $\Phi^{-1}(\cdot)$ is the inverse CDF function of normal distribution and $U_{i,t-k} = F_i(y_{i,t-k}|\mathcal{F}_{t-1})$. This dynamic is of autoregressive form, and it is driven by a lagged part and an innovation part. For the non-Gaussian case, symmetrized Joe-Clayton (SJC) copula is used and Patton (2006) suggested the following dynamic for the tail dependence parameter ρ_t:

$$\begin{aligned} \rho_t &= \Lambda_2(\delta_t) \\ \delta_t &= \omega + \beta\delta_{t-1} + \frac{\alpha}{10}\sum_{k=1}^{10} |U_{1,t-k} - U_{2,t-k}| \end{aligned}$$

Here $\Lambda_2(x)$ is a link function to ensure ρ_t lies in its domain. Besides, Patton (2006) used an ARMA-GARCH model for the marginal distribution of assets returns.

Since the seminal work of Patton (2006), various dynamic copulas have been proposed, and most of them focus on the parameterization of dependence parameters. Heinen and Valdesogo (2009) suggested using DCC framework to model the dependence parameter. Christoffersen et al. (2012) adapted the DCC framework to reduce the computational complexity. While all these preceding models are observation driven, Hafner and Manner (2012) proposed a parameter driven model with a latent stochastic process. Hafner and Reznikova (2010) also developed a semi-parametric approach to model the dependence parameter as a smooth function of time. Structural breaks were used to model the dependence parameter in Dias and Embrechts (2002) and Manner and Candelon (2010).

Many of these papers assume that the dynamics of dependence parameters are of autoregressive form that contains a lagged term and an innovation term. Between these two terms, the choice of the innovation term is crucial and depends on the functional forms of copulas. For example, Patton (2006) used cross products and differences in the Gaussian and non-Gaussian cases respectively. In the latter case, differences are used because the interpretation of cross products is not clear with non-Gaussian distributions. To formulate the dynamics of parameters in general settings, Creal, Koopman and Lucas (2013*a*) and Harvey (2013) proposed GAS models, which use the scores of log likelihood functions as the innovation term. These researchers assumed δ_t -the transformed dynamic parameter- of the following form:

$$\delta_t = \omega + \sum_{i=1}^{p} A_i s_{t-i} + \sum_{j=1}^{q} B_j \delta_{t-j}$$

Since δ_t may be a vector, all the terms here are of appropriate dimensions. A_i and B_j are coefficients of the innovation term and lagged term. It is further assumed that s_t is the scaled score of the likelihood function, as shown below:

$$s_t = S_t \cdot \nabla_t$$
$$\nabla_t = \frac{\partial \ln f(y_t|\delta_t, \mathcal{F}_{t-1}; \theta)}{\partial \delta_t}$$
$$S_t = S(t, \delta_t, \mathcal{F}_{t-1}; \theta)$$

where ∇_t is the score of the log likelihood function and S_t is a matrix function to scale the score. Several choices for the scaling matrix S_t are proposed: It can be the inverse information matrix, the "square root" of the inverse information matrix, or an identity matrix, as displayed below:

$$S_t = \mathcal{I}_{t|t-1}^{-1}, \quad \mathcal{I}_{t|t-1} = E_{t-1}[\nabla_t \nabla_{t-1}']$$

or

$$S_t = \mathcal{J}_{t|t-1}^{-1}, \quad \mathcal{J}'_{t|t-1}\mathcal{J}'_{t|t-1} = \mathcal{I}_{t|t-1}^{-1}$$

or

$$S_t = I$$

where I is an identity matrix. While GAS models apply to general problems involving time varying parameters, they are of special importance to copula modeling. A large number of copulas are constructed from non-Gaussian settings, and it is hard to find an innovation term. GAS models have been applied to dynamic copulas in Oh and Patton (2013), Salvatierra and Patton (2014) and Patton (2012). In this chapter we also use this framework and compare its performance with the model driven by cross products.

Component models have also attracted considerable attention in the literature when modeling volatility and correlation of asset returns. In Engle and Lee (1999), they proposed a GARCH model driven by two components and could be seen as a restricted GARCH(2,2) model. In Engle and Rangel (2008), a multiplicative component GARCH model was proposed and they related return volatility to macroeconomics. Similarly, Colacito, Engle and Ghysels (2011) developed an additive component models named DCC-MIDAS, separating long term and short term components. In this chapter, we extend the frameworks of GARCH-MIDAS and DCC-MIDAS to dynamic copula modeling.

1.3 Model Formulation

The purpose of this section is to introduce a new series of dynamic copula models. In a first subsection, we provide some preliminaries and describes two benchmark models in the literature. The second subsection introduces the structures of new dynamic copula models.

1.3.1 Notation and Preliminaries

To set up models, consider a bivariate stochastic process $y_t = (y_{1,t}, y_{2,t})$. We assume the marginal distribution of $y_{i,t}$ follows the GARCH-MIDAS framework in Engle, Ghysels and Sohn (2013). Specifically, the dynamic of $y_{i,t}$ is as follows:

$$y_{i,t} = \mu_i + \sqrt{\tau_{i,t} g_{i,t}} \epsilon_{i,t} \tag{1.3.1}$$

$$g_{i,t} = (1 - \alpha_i - \beta_i) + \alpha_i \frac{(y_{i,t-1} - \mu_i)^2}{\tau_{i,t}} + \beta_i g_{i,t-1} \tag{1.3.2}$$

$$\tau_{i,t} = m + \theta_i \sum_{k=1}^{K} \psi_k(\omega_{i,1}, \omega_{i,2}) X_{t-k} \tag{1.3.3}$$

where μ_i is the constant mean of $y_{i,t}$. The volatility dynamic for $y_{i,t}$ has two components. The short term component g_t assumes an autoregressive form shown in (1.3.2). The long term component τ_t is driven by a weighted sum of X_{t-k}, and X_{t-k} could be external information or derived from $y_{i,j}$ $(j \leq t-1)$. The weight $\psi_k(\omega_{i,1}, \omega_{i,2})$ is determined by parameter $\omega_{i,1}$ and $\omega_{i,2}$. Denote the information set up to time $t-1$ as $\mathcal{F}_{t-1}$, and we assume $\epsilon_{i,t} | \mathcal{F}_{t-1} \sim F_i(\cdot)$.

For the joint distribution, we assume that $\epsilon_t = (\epsilon_{1,t}, \epsilon_{2,t})$ is generated by a dynamic copula, i.e.:

$$F(\epsilon_{1,t}, \epsilon_{2,t}|\rho_t, \theta, \mathcal{F}_{t-1}) = C(F_1(\epsilon_{1,t}), F_2(\epsilon_{2,t})|\rho_t, \nu, \mathcal{F}_{t-1})$$

where $C(\cdot, \cdot)$ is a bivariate copula and ρ_t is the dependence parameter of interest. ν includes static parameters. As will be discussed in the estimation section, we choose t copula for the joint distribution and skewed t distribution for the marginal distribution of $\epsilon_{i,t}$. Note that the dependence parameter ρ_t for t copula lies in the interval $[-1, 1]$.

Regarding the parameterization of ρ_t, two benchmark models arise in the literature since we choose t copula for the joint distribution. One is the "cross product" model used for Gaussian copula in Patton (2006) and t copula in Christoffersen et al. (2012), with the dynamics as below:

$$\rho_t = \Lambda_1(\delta_t) \tag{1.3.4}$$

$$\delta_t = \omega + \beta\delta_{t-1} + \alpha_1 \Psi^{-1}(U_{1,t-1})\Psi^{-1}(U_{2,t-1}) \tag{1.3.5}$$

$$\Lambda_1(x) = \frac{1 - exp(x)}{1 + exp(x)} \tag{1.3.6}$$

where $\Psi^{-1}(\cdot)$ is the inverse CDF of t distribution and $U_{i,t-1} = F_i(\epsilon_{i,t-1})$.[1] We call models based on this dynamic as PROD models in short of "cross product". Another benchmark model is GAS model as discussed in section 1.2. Assuming one lag period, we have the following dynamics:

$$\delta_t = \omega + \beta\delta_{t-1} + \alpha_1 s_{t-1} \tag{1.3.7}$$

where s_{t-1} is the score of the likelihood function with respect to δ_{t-1}.

1.3.2 A New Class of Component Dynamic Copula Models

In this subsection, we introduce the new class of dynamic copula models. At first, we propose an additive component model based on GAS model, motivated by the for-

[1] The degree of freedom and skewness of this t distribution are the same as the ones in the t copula we specified in the preceding paragraph, i.e.: $C(\cdot, \cdot|\rho_t, \nu, (\mathcal{F})_\sqcup$.

mulation of DCC-MIDAS model. We add into equation (1.3.7) a time varying intercept that is the moving average of lagged δ_t as below:

$$\delta_t = \frac{\omega}{k} \sum_{i=1}^{k} \delta_{t-i} + \beta \delta_{t-1} + \alpha_1 s_{t-1} \tag{1.3.8}$$

This model is a natural extension of DCC-MIDAS model to dynamic copulas. The idea underlying DCC-MIDAS is extracting two components from the daily correlations: one short term component from the daily innovation and a long term component driven by the moving average of realized correlations. Similar logic applies here with two notable differences. One difference is that we use scores rather than cross products, because scores incorporate more information from the functional form of t-distribution than cross products. The other difference is the construction of long term component. While DCC-MIDAS model uses realized correlation for the dynamic intercept, we use the fitted dependence parameter to simplify computation. Since most of the times, there is no closed form relation between realized correlation and the dependence parameter of the copula. We call this new additive component model as GAS-ADD model.

Similarly, we can also take an average of k lagged innovations to form a long term innovation, yielding another component model as follows:

$$\delta_t = \omega + \beta \delta_{t-1} + \alpha_1 s_{t-1} + \frac{\alpha_2}{k} \sum_{i=1}^{k} s_{t-i} \tag{1.3.9}$$

This model shares similar properties with DCC model. It could be seen as a specialized DCC model using scores as innovation with parameter restrictions. So we call it GAS-DCC model.

GAS-DCC and GAS-ADD models decompose the daily innovations of dependence parameters into long and short term components; moreover, we also want to extract the influence of macro variables on the dependence parameters. Among various ways to include macro variables, we choose to follow the GARCH-MIDAS framework in Engle, Ghysels and Sohn (2013), and incorporate macro variables multiplicatively:

$$\delta_t = \omega + \beta\delta_{t-1} + \frac{\alpha_1}{\tau_t} s_{t-1} \tag{1.3.10}$$

$$\begin{aligned} log(\tau_t) = m &+ \theta_1 1_{(\sum_{k=1}^{K} \psi_k(\omega_1,\omega_2)X_{t-k})<0} \sum_{k=1}^{K} \psi_k(\omega_1,\omega_2)X_{t-k} \\ &+ \theta_2 1_{(\sum_{k=1}^{K} \psi_k(\omega_1,\omega_2)X_{t-k})>0} \sum_{k=1}^{K} \psi_k(\omega_1,\omega_2)X_{t-k} \end{aligned} \tag{1.3.11}$$

where X_{t-k} is a macroeconomic variable. $\psi_k(\omega_1, \omega_2)$ is the weight assigned by MIDAS polynomial to X_{t-k} with parameters ω_1 and ω_2. The above formulation has a natural interpretation in GARCH model, since asset returns tend to react differently to news depending on the macroeconomic environment. Now τ_t influences ρ_t similarly but in a nonlinear way, due to the existence of link function. Note that in Engle, Ghysels and Sohn (2013), the intercept in equation (1.3.2) is specified as $1 - \alpha - \beta$. However, that specification is based on the assumption of Gaussian distribution that no longer holds here. So we use a separate parameter ω. We call this model GAS-MIDAS thereafter.

GAS-MIDAS model lets macro variables influence short term innovations. Alternatively, we can also weight macro variables by the long term innovations in the GAS-DCC model. In particular, we divide α_2 in equation (1.3.9) by macro variables and have the following dynamics:

$$\delta_t = \omega + \beta\delta_{t-1} + \alpha_1 s_{t-1} + \frac{\alpha_2}{k\tau_t} \sum_{i=1}^{k} s_{t-i} \tag{1.3.12}$$

$$\begin{aligned} log(\tau_t) = m &+ \theta_1 1_{\sum_{k=1}^{K} \psi_k(\omega_1,\omega_2)X_{t-k}<0} \sum_{k=1}^{K} \psi_k(\omega_1,\omega_2)X_{t-k} \\ &+ \theta_2 1_{\sum_{k=1}^{K} \psi_k(\omega_1,\omega_2)X_{t-k}>0} \sum_{k=1}^{K} \psi_k(\omega_1,\omega_2)X_{t-k} \end{aligned} \tag{1.3.13}$$

The idea here is weighting long term innovation with long term influence of macro variables, and we call this model GAS-DCC-MIDAS. Through equation (1.3.8) to equation (1.3.13), we propose three new component models for dynamic copulas. We will compare the performance of these new models with the benchmark models by empirical applications in section 1.5.

1.4 Estimation

To estimate the parameters of these copula models, we apply the two-stage method discussed in section 1.2. In the first stage, we estimate the univariate GARCH-MIDAS models with quasi maximum likelihood method, and normalize the dependent variables using fitted standard errors. Then we fit a marginal distribution for each of the normalized variables. In the second stage, we fit the copula model with standard maximum likelihood method. This two-stage estimation is generally applied in literature and makes the computation much easier than joint estimation. In the following paragraphs, we discuss the details of this estimation method.

There are numerous choices for the marginal distribution, such as normal distribution, the standardized t distribution (Bollerslev (1987)), and the skewed t distribution (Patton (2004)). We use the skewed t distribution for its flexibility. This distribution has two shape parameters controlling its skewness and tail thickness. A skewness parameter, $\lambda \in (-1, 1)$, describes the degree of asymmetry, and a degrees of freedom parameter, $\nu \in (2, \infty)$, measures the tail thickness. If $\lambda = 0$, we have the standardized Student's t distribution. When $\nu \to \infty$, we have skewed normal distribution. If $\nu \to \infty$ and $\lambda = 0$, we recover a standard normal distribution. All these flexibilities make skewed t distribution a good choice to model univariate variables. For further results on this distribution, refer to Hansen (1994) and Jondeau and Rockinger (2006).

We choose t copula for the multivariate modeling because of its capability to incorporate various dependence structures. First, it has a degrees of freedom parameter ν^c controlling the tail thickness. Second, by varying the dependence parameter ρ, t copula can model data of both negative and positive correlation. If ρ equals to one (minus one), we have perfectly positively (negatively) correlated data series. This property is important since the data in empirical applications exhibits both positive and negative correlations. Not all copulas have such flexibility. For example, Clayton and a number of other Archimedean copulas can only model positively or negatively correlated data. For further results on these copulas, refer to Joe (2014) and Nelsen (2007).

For the MIDAS component containing macroeconomic variables, we use the same variable for both univariate modeling and multivariate modeling. This consistency ensures the conditional copula a valid one. For discussions on the validity of conditional copulas, see Patton (2006). To select the number of lag periods for the MIDAS components, we follow the profiling method discussed in Engle, Ghysels and Sohn (2013) and Colacito, Engle and Ghysels (2011). For the MIDAS polynomial, we choose a Beta weighting scheme of the following form:

$$\psi_k(\omega_1, \omega_2) = \frac{(k/K)^{\omega_1 - 1}(1 - k/K)^{\omega_2 - 1}}{\sum_{j=1}^{K}(k/K)^{\omega_1 - 1}(1 - k/K)^{\omega_2 - 1}}$$

Beta weighting scheme offers flexible shape for the MIDAS filters. It can provide both decreasing and increasing schemes. Moreover, it can also offer a hump shaped weighting shape limited to be unimodal. Besides, there are other weighting schemes available. See Ghysels, Sinko and Valkanov (2007) for a further discussion on the choices of weighting schemes.

Besides, for the GAS-ADD model, we choose $k = 22$ for the time varying intercept. For GAS-DCC model, we choose $k = 5$ for the long term innovation. These lagging periods are picked by the profiling likelihood methods and the clear interpretation of being monthly and weekly averages.

1.5 Empirical Application

1.5.1 Data and Variables

In this section, we use the new class of models to investigate the dependence between stocks and bonds. The bond data are the daily returns of a 10-year Treasury bond. The stock data are the daily returns of five industry portfolios compiled by Kenneth French, which could be downloaded from his web page. The five industries include consumer goods, manufacturing, high tech, health and others. The time range of the data is from November 30, 1985 to December 30, 2013, with 7042 observations. Because of the similar patterns across these five industries, we mainly use the pair of manufacturing industry and 10-year treasury bond as an example. If we mention stock, we mean the stock portfolio of manufacturing industry. This applies to all the figure examples in the following paragraphs.

We use monthly growth rate of industrial production (IP) in U.S. as the macro variable.[2] For univariate modeling, we compute the quarterly rolling average of the IP rates and apply a MIDAS polynomial with the quarterly average. Specifically, if X_t is the variable in the MIDAS polynomial as in equation (1.3.3) and X'_t is the monthly IP rate, then $X_t = (X'_t + X'_{t-1} + X'_{t-2})/3$. For each day, we look back for 16 months, i.e.: $K = 16$ in the MIDAS polynomial. Therefore, there are actually two filters smoothing the macro variables. Similar applications of filters can be found in Engle, Ghysels and Sohn (2013).

For multivariate modeling, we use the first order difference of the IP growth rates and take the quarterly average of the difference to smooth the data. Unlike the univariate modeling, we only assume a flat weight for the MIDAS filter. We make these transformations based on empirical investigations. Both models of the raw rates and

[2] The IP rates are calculated year over year.

differences are tested, and the latter one shows a better performance. For the number of lag periods, empirical tests favor a short window of three months. This short window size is also supported by the volatile fluctuation of the realized correlations, which is shown in the the upper panel of Figure 1.1. Clearly, the correlations have many spikes, even if they are calculated on a quarterly basis. Meanwhile, the IP growth rates in the lower panel of Figure 1.1 change relatively slow. It is hard to relate the change of correlations today to the variation of IP growth rates one year ago. Therefore, we only look back for three months. For such short lag period, the difference between flat weights and uneven weights becomes negligible, so we select flat weights for computational simplicity.

1.5.2 Results

Before discussing numerical results, let us examine some figures to have a general impression of these models. Figure 1.2 plots the quarterly realized correlations of the stock and bond, and the correlations exhibit strong temporal variations. These variations support the applications of dynamic copulas, because no static copula can produce such volatile patterns. Figure 1.3 further plots the quarterly realized correlations along with the implied correlations of GAS model. It shows that the implied correlations closely follow the realized ones while the realized ones have wilder fluctuations. In Figure 1.4, we reduce the sampling window of realized correlations to one month and plot the monthly realized correlations with the implied correlations of GAS model. Now the implied correlations seem to be a long term component of the monthly realized correlation. These two figures convey that GAS model is highly persistent and similar patterns can also be observed with GAS-based models through Figure 1.5 to Figure 1.7.

All the estimation results are presented in Table 1 to 5. The last rows of these tables contain the likelihoods of the six models. By comparing these likelihoods, we make a number of observations on the model performances. First, GAS-based models have higher likelihoods than PROD model driven by cross products. Moreover, GAS-DCC and GAS-ADD models are better than GAS model, and GAS-DCC-MIDAS model is better than GAS-DCC model in terms of likelihood. Since GAS and GAS-ADD models are nested, we can apply likelihood ratio tests to compare these two models. It can be shown that the difference is statistically significant. Similar arguments apply to the pairs of GAS/GAS-DCC and GAS-DCC/GAS-DCC-MIDAS. By contrast, GAS-MIDAS model does not offer much more than the GAS model. This may suggest that it is better to weight long term variations by macro variables than short term variations.

Now let us turn our attention to interpreting parameter estimates. The first rows of these tables report the estimates of α_1, reflecting the influence of daily innovations on dependence parameter. For all the six models, this parameter is significant and has a positive sign as expected. Furthermore, the estimates of GAS-DCC, GAS-DCC-MIDAS and GAS-ADD models have much bigger values than the one of GAS model. This may imply that the former models can offer more volatilities than the latter one.

The second rows of these tables refer to the estimates of α_2. This parameter measures the influence of long term (weekly) innovations on the dependence parameter of GAS-DCC and GAS-DCC-MIDAS models. The estimates of both models are significant and have negative signs. These negative signs seem to remove part of the daily innovations accumulated in the past week, and make GAS-DCC and GAS-DCC-MIDAS models more volatile than other models. Figure 1.6 plots the implied correlations of GAS and GAS-DCC models. GAS-DCC model has a thicker curve than GAS model, and this also conveys that the former model has more volatility than the latter.

Besides α_1 and α_2, θ_1 and θ_2 also affect how innovations integrate into the dependence parameters. For most of the five industries, these two parameters are significant. In GAS-DCC-MIDAS model, θ_1 are positive and θ_2 are negative for all five industries. This means that $(\theta_1 X_t 1_{X_t<0} + \theta_2 X_t 1_{X_t\geq 0}) < 0$ and $\tau_t < 1$ always. When the absolute value of X_t is higher, $\theta_1 X_t 1_{X_t<0} + \theta_2 X_t 1_{X_t\geq 0}$ is smaller and $|\alpha_2/\tau_t|$ is larger. A large $|\alpha_2/\tau_t|$ removes more weekly innovations from the dependence parameter, providing more fluctuations to the model. Figure 1.5 also demonstrates that the implied correlations of GAS-DCC-MIDAS model are more volatile than the ones of GAS-DCC model. That's probably why GAS-DCC-MIDAS model has a better performance than the GAS-DCC model. θ_1 and θ_2 also appear in GAS-MIDAS model. Since GAS-MIDAS model is not much different from GAS model in terms of likelihoods, we skip interpreting these two parameters for GAS-MIDAS model.

While all the parameters mentioned above measure the loading of innovation terms, β measures the influence of lagged terms. For all the six models except GAS-ADD model, β are fairly close to one, and GAS based models have higher values than PROD model. These estimates are consistent with the results in other papers also using GAS models, say Patton (2012). GAS-ADD model is a special case, since its β is lower than 80 percent. But if we add up β and ω, we find that the sums are also close to one in all industries. So GAS-ADD model transfers part of the weight on the lagged implied correlation to the intercept that is the long term component in the model. This model still yields highly persistent implied correlations.

ν measures the tail thickness of the t copula. The larger the ν value is, the thinner the tail is. Compare the ν value of GAS-DCC/GAS-DCC-MIDAS/GAS-ADD model with that of PROD/GAS model, we find that the former model has a higher value than the latter, i.e.: the former model has a thinner tail than the latter, and less extreme events happen with the former models. This comparison conveys that we reduce the tail thickness by providing more accurate estimates of the dependence parameter.

Some further comparisons can be drawn between GAS-ADD and GAS-DCC models. Comparing the coefficients and likelihoods of these two models, we find that all these outputs are rather similar: the values of likelihoods are almost identical; the estimates of α_1 and ν are alike; the loadings on lagged terms are both close to one. This similarity can be further confirmed by examining the implied correlations of these two models in Figure 1.7, and the two curves of implied correlations are almost the same. These "coincidences" reveal the interconnection between GAS-DCC and GAS-ADD models and are possibly caused by the high persistence of GAS models. Since β is approximately one, it influences the model equivalently by adding a time varying intercept or a weekly innovation.

1.6 Conclusions

In this chapter, we propose a novel class of dynamic copula models and extract long and short term components from the dependence parameter. An extra component adds flexibility to the model, and most of the empirical applications show improved prediction. Moreover, we introduce the MIDAS framework to dynamic copulas and combine daily returns with monthly updated macro variables. Specifically, we use the macro variable as a proxy for the market condition, and allow that the market condition affects how dependence parameter reacts to innovations. We find that introducing macro variables adds more volatility to the dependence parameter, and therefore improved the model performance consistently through the empirical applications.

Table 1.1: Estimation Results for Consumer Goods

	PROD	GAS	GAS-DCC	GAS-MIDAS	GAS-DCC-MIDAS	GAS-ADD
α_1	0.028	0.100	0.294	0.132	0.242	0.227
	(5.40)	(5.58)	(19.59)	(4.95)	(5.37)	(6.53)
α_2	-	-	-0.212	-	-0.110	-
	-	-	(-49.40)	-	(-2.56)	-
β	0.981	0.998	0.999	0.998	0.998	0.851
	(293.82)	(939.56)	(1277.40)	(366.50)	(1251.10)	(22.44)
θ_1	-	-	-	-0.823	0.435	-
	-	-	-	(-11.15)	(2.92)	-
θ_2	-	-	-	0.489	-0.667	-
	-	-	-	(2.73)	(-3.61)	-
ν	7.69	8.25	8.83	8.39	8.87	8.80
	(9.55)	(8.54)	(85.68)	(431.68)	(8.42)	(8.96)
ω	0.00	0.00	0.00	0.00	0.00	0.15
	(-2.97)	(0.03)	(0.04)	(0.13)	(0.87)	(3.86)
$logL$	575	578	587	580	593	587

Notes: This table reports the estimates for the dynamic copula models with 10-year Treasury bond and the stock portfolio of the consumer goods industry. T-statistics are in parentheses.

Table 1.2: Estimation Results for Manufacturing

	PROD	GAS	GAS-DCC	GAS-MIDAS	GAS-DCC-MIDAS	GAS-ADD
α_1	0.028	0.114	0.300	0.154	0.246	0.242
	(6.01)	(4.58)	(10.97)	(33.95)	(6.12)	(7.68)
α_2	-	-	-0.211	-	-0.098	-
	-	-	(-7.83)	-	(-2.96)	-
β	0.982	0.997	0.998	0.998	0.999	0.841
	(350.64)	(569.77)	(1118.20)	(752.08)	(1437.60)	(21.10)
θ_1	-	-	-	-1.399	0.943	-
	-	-	-	(-67.46)	(6.48)	-
θ_2	-	-	-	0.411	-0.670	-
	-	-	-	(25.89)	(-4.56)	-
ν	9.31	9.88	10.55	10.07	10.31	10.53
	(8.32)	(26.53)	(524.23)	(1872.30)	(7.93)	(123.70)
ω	0.00	0.00	0.00	0.00	0.00	0.15
	(-3.64)	(-0.23)	(-0.22)	(-0.07)	(1.01)	(3.90)
$logL$	583	594	604	597	617	605

Notes: This table reports the estimates for the dynamic copula models with 10-year Treasury bond and the stock portfolio of the manufacturing industry. T-statistics are in parentheses.

Table 1.3: Estimation Results for Health

	PROD	GAS	GAS-DCC	GAS-MIDAS	GAS-DCC-MIDAS	GAS-ADD
α_1	0.027	0.114	0.300	0.154	0.246	0.242
	(5.86)	(4.58)	(10.97)	(33.95)	(6.12)	(7.68)
α_2	-	-	-0.211	-	-0.098	-
	-	-	(-7.83)	-	(-2.96)	-
β	0.982	0.998	0.998	0.998	0.998	0.899
	(332.80)	(1047.70)	(1358.20)	(1267.70)	(1283.00)	(7.80)
θ_1	-	-	-	-0.864	0.297	-
	-	-	-	(-1.79)	(2.06)	-
θ_2	-	-	-	0.217	-0.035	-
	-	-	-	(0.58)	(-0.15)	-
ν	8.55	8.74	8.97	8.85	8.98	8.92
	(9.05)	(8.64)	(8.43)	(8.39)	(8.30)	(0.42)
ω	0.00	0.00	0.00	0.00	0.00	0.10
	(-3.18)	(0.30)	(0.35)	(0.18)	(-0.18)	(0.90)
$logL$	513	520	525	522	527	524

Notes: This table reports the estimates for the dynamic copula models with 10-year Treasury bond and the stock portfolio of the health industry. T-statistics are in parentheses.

Table 1.4: Estimation Results for HiTec

	PROD	GAS	GAS-DCC	GAS-MIDAS	GAS-DCC-MIDAS	GAS-ADD
α_1	0.028	0.083	0.211	0.101	0.183	0.177
	(6.04)	(6.29)	(4.41)	(5.45)	(4.44)	(4.16)
α_2	-	-	-0.135	-	-0.083	-
	-	-	(-2.75)	-	(-2.24)	-
β	0.981	0.998	0.998	0.998	0.998	0.875
	(329.57)	(1151.00)	(1292.80)	(1229.00)	(1430.80)	(17.04)
θ_1	-	-	-	-0.890	0.789	-
	-	-	-	(-1.53)	(4.23)	-
θ_2	-	-	-	0.056	-0.198	-
	-	-	-	(0.19)	(-0.99)	-
ν	9.35	9.83	10.42	9.90	10.54	10.39
	(8.27)	(123.55)	(6.88)	(7.73)	(7.44)	(18.56)
ω	0.00	0.00	0.00	0.00	0.00	0.12
	(-2.95)	(0.13)	(0.12)	(-0.59)	(0.14)	(2.40)
$logL$	440	445	449	446	453	450

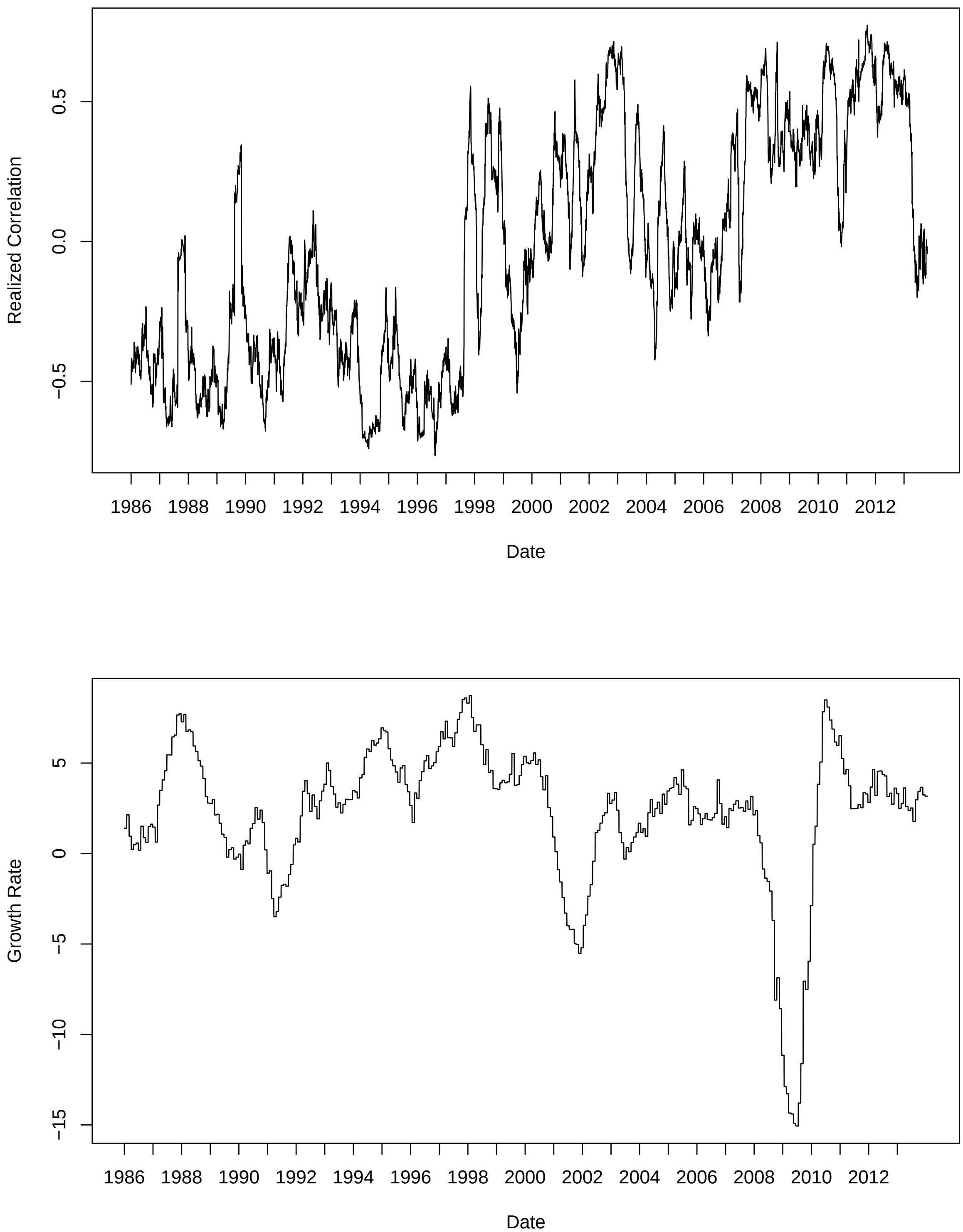
Notes: This table reports the estimates for the dynamic copula models with 10-year Treasury bond and the stock portfolio of the HiTec industry. T-statistics are in parentheses.

Table 1.5: Estimation Results for Other

	PROD	GAS	GAS-DCC	GAS-MIDAS	GAS-DCC-MIDAS	GAS-ADD
α_1	0.031	0.129	0.310	0.166	0.284	0.233
	(6.45)	(3.96)	(7.09)	(7.05)	(6.48)	(6.99)
α_2	-	-	-0.207	-	-0.135	-
	-	-	(-4.41)	-	(-3.33)	-
β	0.979	0.997	0.998	0.997	0.998	0.891
	(332.55)	(630.02)	(1160.10)	(1044.40)	(1301.80)	(27.79)
θ_1	-	-	-	-0.982	0.677	-
	-	-	-	(-2.18)	(5.54)	-
θ_2	-	-	-	0.262	-0.444	-
	-	-	-	(0.90)	(-5.05)	-
ν	7.33	7.70	8.16	7.83	8.44	8.08
	(10.58)	(383.07)	(8.85)	(9.52)	(8.88)	(8.94)
ω	0.00	0.00	0.00	0.00	0.00	0.11
	(-4.51)	(-0.39)	(-0.25)	(-0.64)	(-1.12)	(3.31)
$logL$	740	760	770	763	777	769

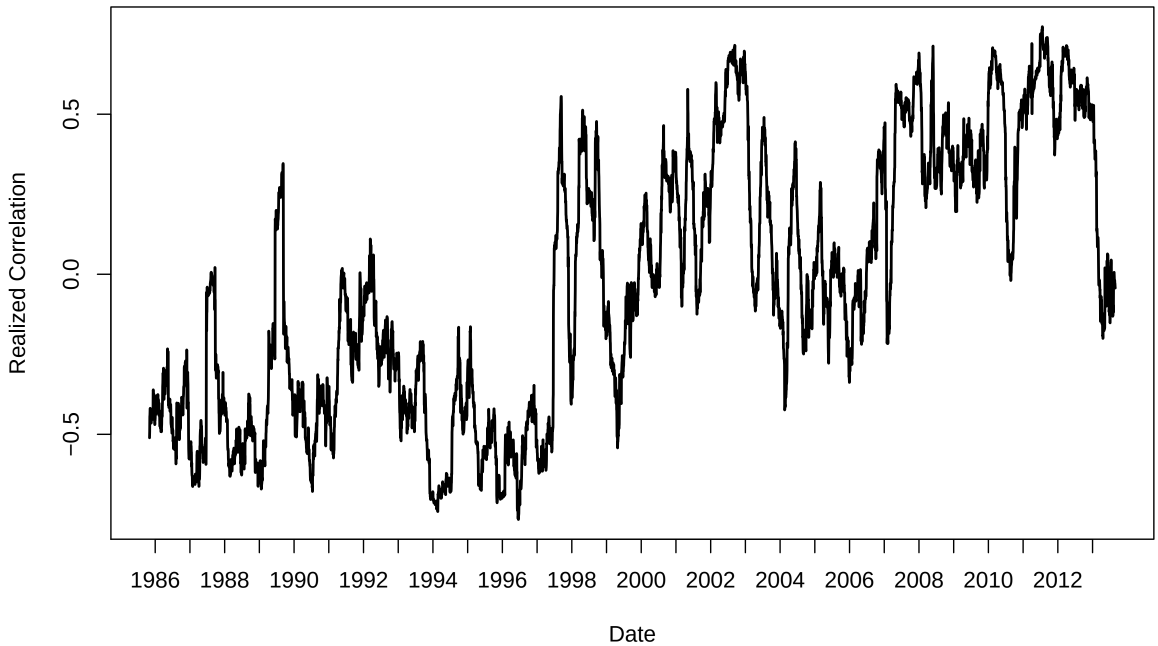
Notes: This table reports the estimates for the dynamic copula models with 10-year Treasury bond and the stock portfolio of the other industry. T-statistics are in parentheses.

Figure 1.1: Industrial Production and Realized Correlations



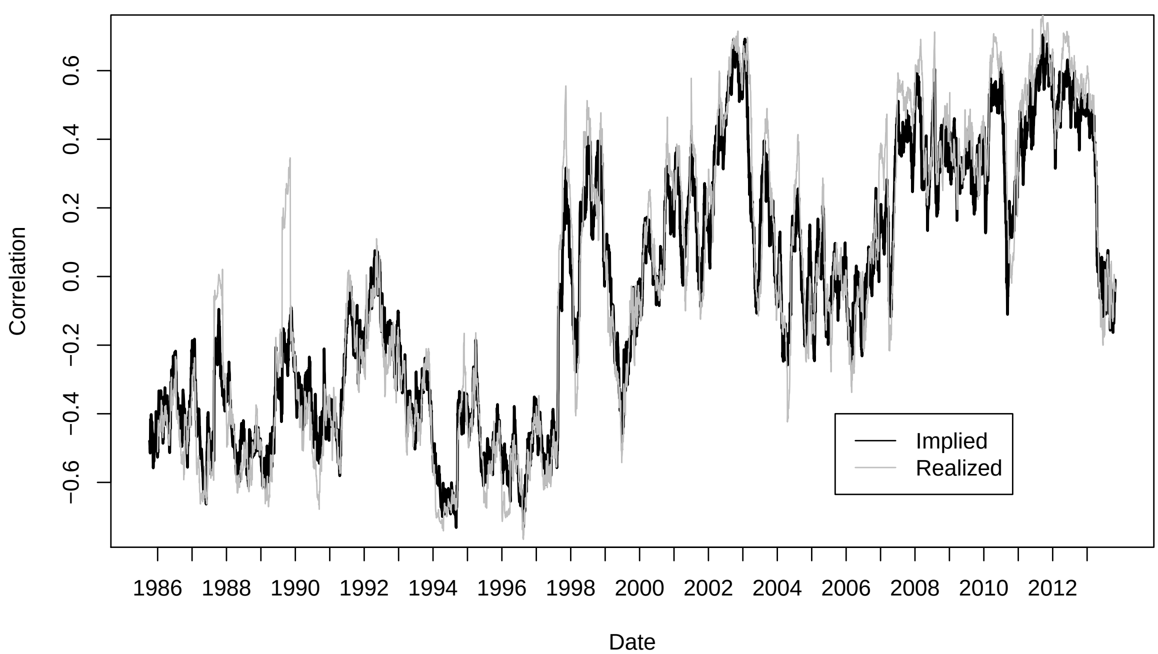
Notes: The upper panel shows the quarterly realized correlations of 10-year Treasury bond and the stock portfolio of manufacturing industry. The lower panel presents the monthly growth rates of industrial production in U.S.

Figure 1.2: The Quarterly Realized Correlations of the Stock and Bond



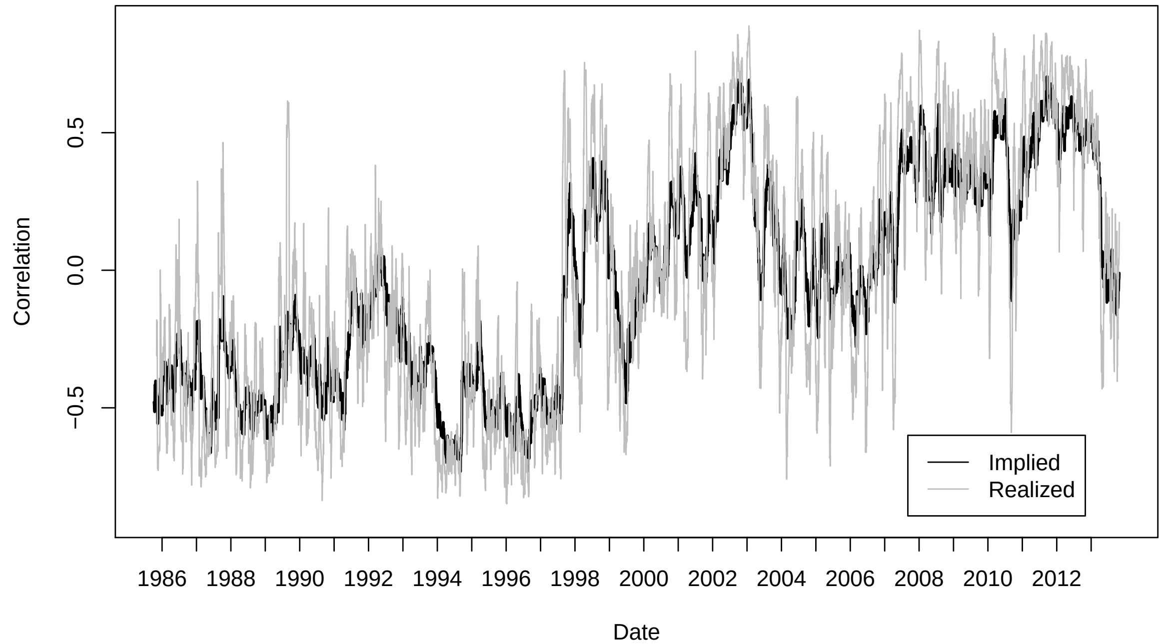
Notes: This picture reports the quarterly realized correlations of 10-year Treasury bond and the stock portfolio of manufacturing industry. The realized correlations are calculated on a rolling basis.

Figure 1.3: The Quarterly Realized Correlation and Implied Correlations



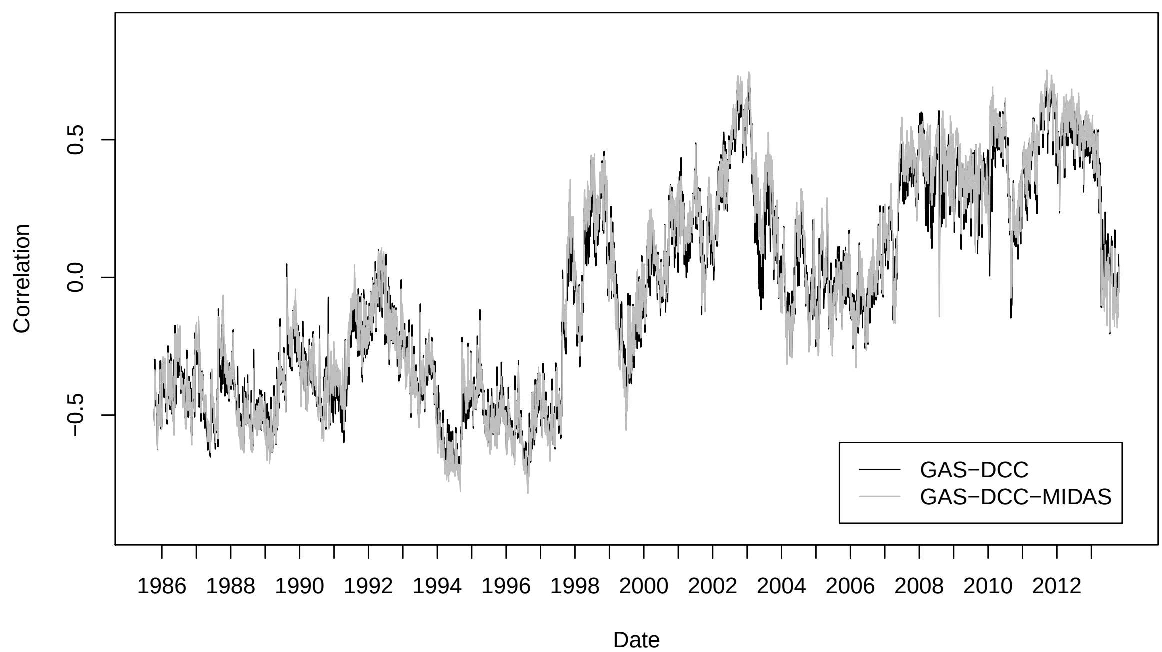
Notes: This picture reports the correlations of 10-year Treasury bond and the stock portfolio of manufacturing industry. The dark line shows the implied correlations of GAS model. The light line represents the quarterly realized correlations calculated on a rolling basis.

Figure 1.4: The Monthly Realized Correlations and the Implied Correlations



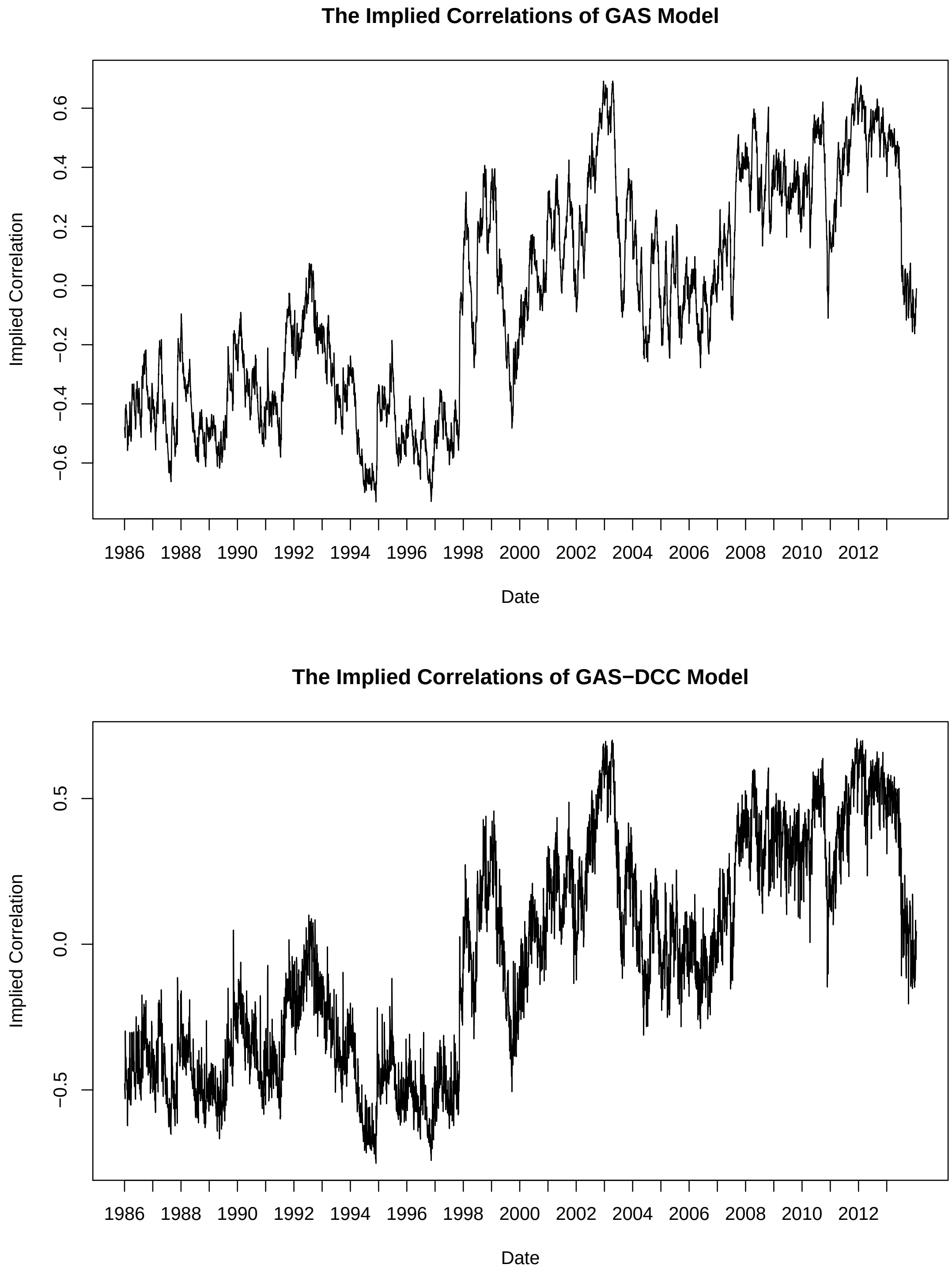
Notes: This picture reports the correlations of 10-year Treasury bond and the stock portfolio of manufacturing industry. The dark line shows the implied correlations of GAS model. The light line represents the monthly realized correlations calculated on a rolling basis.

Figure 1.5: The Implied Correlations of GAS-DCC and GAS-DCC-MIDAS Models



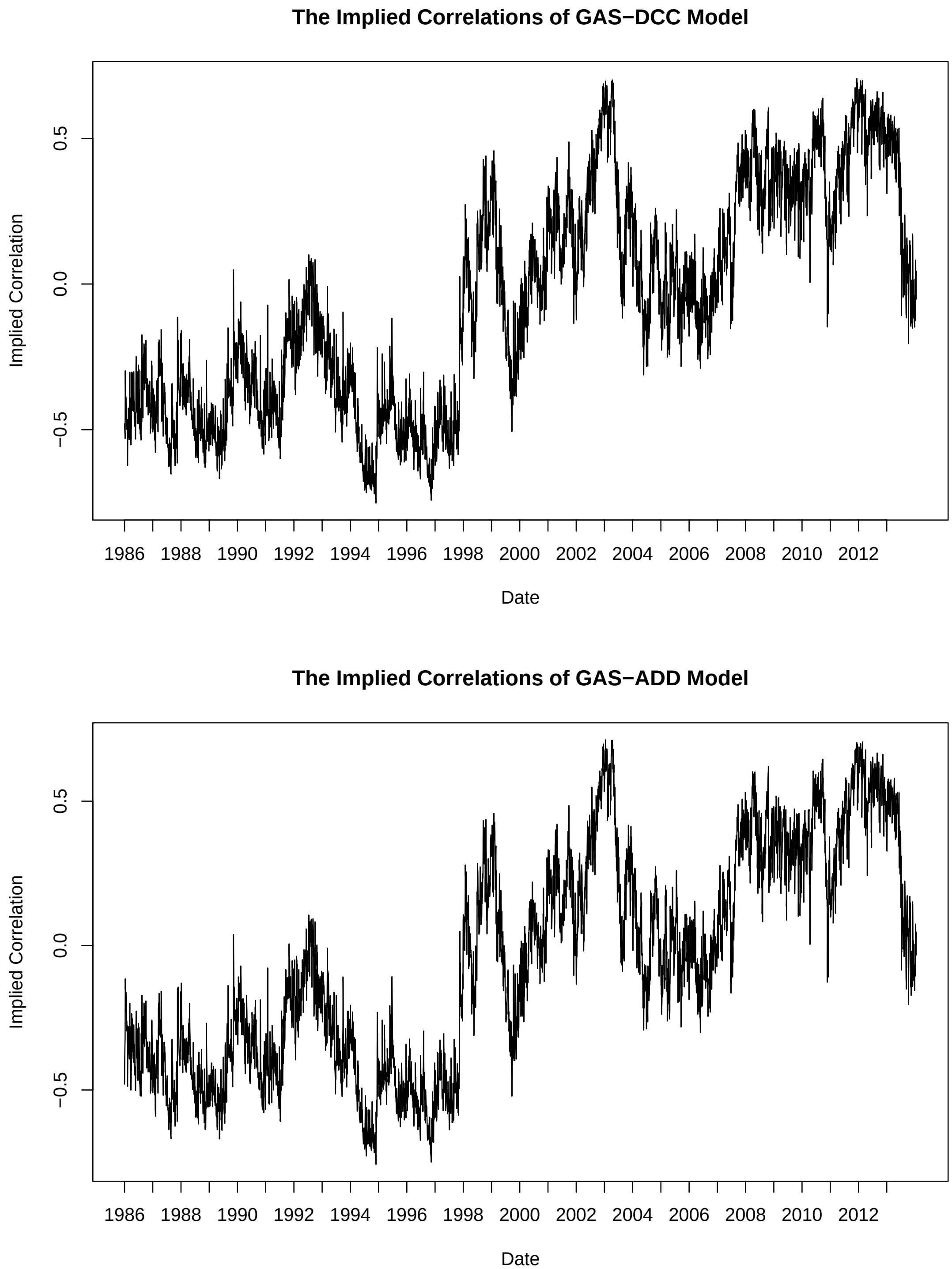
Notes: This picture reports the correlations of 10-year Treasury bond and the stock portfolio of manufacturing industry. The dark line shows the implied correlations of GAS-DCC model. The light line represents the implied correlations of GAS-DCC-MIDAS model.

Figure 1.6: The Implied Correlations of GAS and GAS-DCC Models



Notes: This picture reports the correlations of 10-year Treasury bond and the stock portfolio of manufacturing industry. The upper panel shows the implied correlations of GAS model. The lower panel presents the implied correlations of GAS-DCC model.

Figure 1.7: The Implied Correlations of GAS-DCC and GAS-ADD Model



Notes: This picture reports the correlations of 10-year Treasury bond and the stock portfolio of manufacturing industry. The upper panel shows the implied correlations of GAS-DCC model. The lower panel presents the implied correlations of GAS-ADD model.

CHAPTER 2: FRAILTY MODELS FOR COMMERCIAL MORTGAGES

2.1 Introduction

Credit risk affects virtually all aspects of financial activities. It is an important factor in pricing financial products and has profound influence on risk management. Moreover, policy makers and regulators pay special attention to credit risk when they design economic policies and regulatory frameworks.

There is an extensive literature on the measurement and management of credit risk. Researchers have used various approaches to model the credit risk of corporate debts, mortgages and derivatives. In general, these models can be divided into two categories – structural models and reduced-form models. This article presents a reduced-form model that contains frailty factors to predict mortgage default.

Reduced-form models rely on two sources of information to explain credit risk. One source is the financial information of borrowers, which is employed to track the idiosyncratic part of the credit risk. The other source comes from macro variables, which approximates the systematic part. Failure to explain all the systematic risk would introduce biases when estimating risk measures. Das et al. (2007) – while studying corporate bond defaults – provided evidence that macro variables alone were not enough to explain all the systematic risk. They further demonstrated that the lack of explanatory variables underestimates value-at-risk. To produce unbiased estimates, they proposed frailty models for corporate bond credit risk, to account for the unexplained part of systematic risk.

Frailty models can be classified into two types - using a characterization put forward by Cox et al. (1981): parameter-driven models and observation-driven models. Both types of models have been used to predict credit risk. For example, parameter-driven models have been used to track the credit risk of corporate debts in Duffie et al. (2009) and Koopman and Lucas (2008), and to forecast mortgage default in Kau, Keenan and Li (2011). Meanwhile, using observation-driven models, Creal, Koopman and Lucas (2013*a*) and Creal et al. (2014) investigated corporate defaults. The predictability of latent factors is a main feature of observation-driven models. This feature indicates that current period frailty factors can be computed using only past information. In contrast, the computation of frailty factors in parameter-driven models not only requires past information but also future and current information. Inference of parameter-driven models therefore generally requires simulation, which is time consuming with large data sets. The estimation of observation-driven models is in comparison rather straightforward.

In this chapter, we develop a novel framework to model systematic risk of mortgages. Specifically, we match default rates in multiple dimensions by extending the generalized autoregressive score (GAS) models proposed in Creal, Koopman and Lucas (2013*a*). Our data consists of commercial mortgages in the U.S. multifamily market from 1997 to 2013. We construct a series of models and employ multiple tests to demonstrate the advantages of our framework.

To the best of our knowledge, this is the first attempt that uses observation-driven models to predict mortgage defaults. We show that the new class of models we propose has better tractability compared with parameter-driven models. For instance, although our dataset has more than two million records, and our most complex model incorporates up to 15 frailty factors, the estimation process only takes two minutes using a standard desktop computer. Compare this with for example Kau, Keenan and Smurov (2006) who employed parameter-driven models to predict mortgage defaults using a small data set. Their method requires simulations and is very time consuming and therefore practically infeasible when using large data sets typically encountered in practice.

In addition to the merit of tractability, our multi-factor formulation is able to match the default rates of mortgages in multiple dimensions. While a single latent factor suffices to match the temporal fluctuation of default rates, it may not capture the variation of default rates along other dimensions. For example, our empirical analysis provides evidence that the origination month and the originator preference influence the default rates of commercial mortgages. To model the influence of these variables, we group mortgages by these two variables and allow latent factors to vary by groups. Compared with GAS models using a single factor, our multi-factor frailty models feature improved empirical fits - even out-of-sample.

The remainder of this chapter is organized as follows. The next section reviews the literature. Section 3 formulates our dynamic frailty models. In section 4, we discuss empirical applications. The last section offers some concluding remarks.

2.2 Literature Review

To motivate our theoretical framework and demonstrate the merit of frailty models, it is useful to review the literature on credit risk modeling. We first review the studies that used either structural or reduced-form approaches to model the credit risk of corporate debt. Next, we focus on two types of reduced-form models augmented by frailty factors: observation-driven models and parameter-driven models. The bulk of the literature focuses on corporate debt - due to the fact that such data is most readily available. Since our chapter deals with commercial mortgages, we review the relatively small literature covering the current state of the art models for mortgage credit risks.

Research on credit risk prediction has a long history, dating back to Beaver (1968) and Altman (1968). Since then, numerous models have been developed to assess risk factors and predict default events. Following the classification in Altman et al. (2005), credit risk models can be categorized into two groups: structural-form models and reduced-form models. A majority of structural-form models are based on the framework proposed by Black and Scholes (1973) and Merton (1974), which utilizes the principle of option pricing. This principle assumes that a corporation will default when its assets drop to a sufficiently low level relative to its liabilities. As a result, all the relevant credit risk elements are functions of structural characteristics of the corporation, such as asset volatility and capital structure. Similar structural-form models have been proposed in Black and Cox (1976), Jones, Mason and Rosenfeld (1984), Fischer, Heinkel and Zechner (1989), Shimko, Tejima and Van Deventer (1993), Leland (1994), Longstaff and Schwartz (1995), Nielsen, Saà-Requejo and Santa-Clara (2001), and Hui, Lo and Tsang (2003). Both macro variables and financial information of borrowers have been used in these models.[1]

While structural-form models assess credit risk elements such as probability of default and recovery rates through an implied process of the value of a corporation, reduced-form models impose separate and explicit assumptions on default probabilities and recovery rates. Specifically, reduced-form models assume credit risk elements not only relate to the structural features of the firm, but also depend on macro variables and financial information of borrowers. This line of research started with Beaver (1968) and Altman (1968) who used discriminant analysis as the main tool.[2] Since Ohlson (1980) and Zmijewski (1984), binary response models like logit and probit regressions have been introduced into credit risk modeling. Most of these models estimate single-period default probabilities or credit scores. Some recent studies began to extend the prediction horizon to multiple periods using multiple logit models (see for example, Campbell, Hilscher and

[1] Structural-form models have also been widely used by financial corporations, such as Moody's KMV, a leading provider of quantitative credit analysis tools. For a complete description of Moody's KMV method, see Crosbie and Bohn (2003).

[2] These models are often called "the first generation of reduced-form models" (Duan, Sun and Wang 2012).

Szilagyi 2008).

All the reduced-form models mentioned above are based on the assumption that macro variables are enough to explain all the systematic risk. However, Das et al. (2007) provide evidence that this assumption can be violated. In particular, Duffie et al. (2009) introduce the notion of common latent factors - or so called frailty factors - into the default intensity of their proportional hazard models for corporate debt credit risk. They showed that failure to control for these latent factors caused downward biases in calculating value-at-risk. In similar contexts, Duan and Fulop (2013) proposed frailty models with forward intensity methods and attempted to overcome the potential computational burden induced by models in Duffie (2009).

Koopman and Lucas (2008) considered frailty methods with formulations other than proportional hazard models. In particular, they added latent dynamic factors to logistic regressions and developed a non-Gaussian multivariate state-space model to predict corporate default. Koopman, Lucas and Schwaab (2012) extend the framework of Koopman and Lucas (2008) by jointly modeling macro variables with default events, and also include industry effects. Schwaab, Koopman and Lucas (2015) consider a framework similar to the one in Koopman, Lucas and Schwaab (2012) adding regional effects to model default outcomes. All the models involving latent factors discussed so far are parameter-driven models. Estimation of such models generally involves simulation-based algorithms, such as particle filtering and importance sampling. Typical applications of these algorithms are time consuming and therefore impractical to implement with large data sets.

Observation-driven models have also been used to model credit risk. In particular, Creal, Koopman and Lucas (2013*a*) designed a class of observation-driven models - called GAS models - and applied them to Moody's credit rating data. GAS models use scaled scores of the likelihood function to update the dynamics of latent factors. Using these models, Creal et al. (2014) jointly model macro variables and default outcomes with data of mixed-measurement and mixed-frequency.[3]

While the bulk of the literature covers corporate debt credit risk, there are a few papers studying mortgage defaults. Frailty factors play an important role in modeling mortgages and are called "baseline hazards" in the literature. Currently, non-parametric methods are used in modeling frailty factors. Among all non-parametric methods, the flexible methods proposed in Han and Hausman (1990), Sueyoshi (1992) and McCall (1996) are popular choices (for an application, see Deng, Quigley and Order 2000). Shared frailty, used in Follain, Ondrich and Sinha (1997), is an alternative to flexible methods. Kau, Keenan and Li (2011) extended Follain, Ondrich and Sinha (1997) by considering regional effects. Finally, Kau, Keenan and Smurov (2004) and Kau, Keenan and Smurov (2006) employed parameter-driven models to capture baseline hazards of residential mortgages.

2.3 Model Formulation

We develop in this section a novel class of models for systematic credit risk of mortgages, extending the generalized autoregressive score (GAS) models proposed in Creal, Koopman and Lucas (2013*a*).

To set up the model, consider a set of n_t mortgages and let the vector of default status be denoted as $y_t = [y_{1,t}, ..., y_{n_t,t}]$, where n_t is the number of mortgages at time t. The element y_{it} of this vector, is a binary variable referring to the default status of mortgage i at time t; it equals to one if a default happens at time t, and zero if the borrower can make timely payment. Let us further denote the default probability as

[3] If one applies GAS models to binary logistic regressions and chooses the intercept as the dynamic parameter, then score of the dynamic intercept is actually the generalized residual defined in Gouriéroux et al. (1987). In this context, GAS models reduce to the generalized autoregressive moving average models proposed in Shephard (1995) and Benjamin, Rigby and Stasinopoulos (2003).

$\pi_{i,t}$, i.e. $\pi_{i,t} = P(y_{i,t} = 1)$. To relate the default probability with covariates, we adopt a binary logit model:

$$\pi_{i,t} = exp(\mu_{i,t})/(1 + exp(\mu_{i,t}))$$

where $\mu_{i,t}$ is transformed default probability. It is defined as a linear combination of covariates and a latent factor:

$$\mu_{i,t} = x'_{i,t}\beta + f_{i,t}$$

where the vector $x_{i,t}$ consists of a series of observed variables, such as financial information of the mortgage, β is a coefficient vector and constant across all mortgages, and $f_{i,t}$ is a latent factor or factors - as clarified later. It is important to emphasize one interpretation of the above equation. The presence of $f_{i,t}$ can be viewed as the *intercept* to the default intensity. Hence, a latent factor or set of factors affect the baseline defaults. More specifically, because fixed income products have different credit quality, their exposure to common sources of systematic risk varies. To model this feature, researchers usually group fixed-income products according to certain criteria, and allow latent factors to vary across different groups. Rating classes are a popular criterion in grouping corporate debts. For example, Creal et al. (2014) grouped corporate debts by rating classes, and their models have three factors based on multiple classes. Schwaab, Koopman and Lucas (2015) grouped corporate debts not only by rating class, but also by region and industry. However, the grouping criteria are not obvious for mortgages, as there is no rating class for mortgages. Researchers have to define their own criteria. For instance, in Kau, Keenan and Li (2011), the authors use locations of mortgages as the criterion to define latent factors.

Our empirical analysis suggests that mortgage default rates vary with origination month and originator preference. Based on this evidence, we choose these two properties of mortgages as our grouping criteria. Specifically, we group mortgages of consecutive origination months in the same group and use c_i^1 to denote the group number of mortgage

i according to origination month. Meanwhile, mortgages of similar originator preference are grouped together and c_i^2 represents the group number of mortgage i according to the grouping criterion of originator.

We assume a common latent factor for mortgages in the same group and allow latent factors to vary across different groups. Since mortgages in different groups have distinct latent factors, every grouping criterion corresponds to a separate set of latent factors. Combining this assumption with our two grouping criteria, we decompose $f_{i,t}$, the latent factor specification for mortgage i at time t, into two parts:

$$f_{i,t} = f^1_{c_i^1,t} + f^2_{c_i^2,t}$$

where $f^1_{c_i^1,t}$ represents the set of frailty factors related to "origination month", and $f^2_{c_i^2,t}$ stands for the set of frailty factors relevant to "originator preference". Generally, if n grouping criteria are identified then $f_{i,t}$ is decomposed into n parts.

We further assume that the latent factors have an autoregressive form, namely:

$$f^k_{c_i^k,t} = \theta_k f^k_{c_i^k,t-1} + \alpha_k s^k_{c_i^k,t-1}, k \in \{1, 2\}$$

The time subscript to the innovation $s^k_{c_i^k,t-1}$ indicates that it is computed using time $t-1$ information. The choice of innovation is crucial in updating the dynamics of frailty factors, and use generalized residuals, as dubbed by Gouriéroux et al. (1987). Specifically, we characterize the innovations as:

$$\begin{aligned} s^k_{c_i^k,t-1} &= \bar{y}^k_{c_i^k,t-1} - \hat{y}^k_{c_i^k,t-1} \\ \bar{y}^k_{c_i^k,t-1} &= \sum_{j=1}^{n_{t-1}} y_{j,t-1} 1_{c_j^k = c_i^k} \\ \hat{y}^k_{c_i^k,t-1} &= \sum_{j=1}^{n_{t-1}} \hat{\pi}_{j,t-1} 1_{c_j^k = c_i^k} \end{aligned}$$

where $\bar{y}^k_{c_i^k,t-1}$ is the empirical default rate for group c_i^k at time $t-1$, and $\hat{y}^k_{c_i^k,t-1}$ is the fitted default rate for group c_i^k at time $t-1$. $1_{c_j^k = c_i^k}$ is an indicator variable; it equals to one if mortgage j is in group c_i^k and zero otherwise. Using this indicator variable, we include only information from group c_i^k to update the frailty factors related to mortgage i. Likewise, $\hat{\pi}_{j,t-1}$ is the estimated default probability (PD) for mortgage j. Intuitively, our innovation term is the difference between empirical default probability and fitted default probability. The innovation therefore measures the distance between models and data. If the innovation term is positive/negative, then the empirical PD is larger/smaller than fitted one.

Finally, it is important to note that the coefficients of the frailty factors, namely θ_k and α_k, are the same across all groups. This restriction enables us to keep the model parsimonious and tractable. Practically speaking, our data set is large and we have 15 groups to characterize the heterogeneity across mortgages. Without this restriction, the model would become intractable, and more than 30 parameters need to be estimated in a non-linear setting. In our model, the sign of the coefficient α_k determines the properties of frailty factor. A positive sign implies that the fitted PD will be adjusted towards the empirical one whenever there exists a mismatch between the model default predictions and the data.

2.4 Empirical Applications

2.4.1 Data and Variables

Our mortgage data contains 2,207,588 records of commercial mortgages in the U.S. multifamily market. The records start in July 1996 and end in the middle of 2013, with none of the mortgage pre-dating the start of the sample. There are 18561 distinct mortgages and all of them are 10-year-balloon mortgages.

As discussed earlier, default risk of mortgages consists of idiosyncratic risk and systematic risk. To explain idiosyncratic risk, we use mortgage age, debt service coverage ratio, and an indicator variable reflecting servicers' warning. For systematic risk, we employ lagged values of default rates and a single frailty factor or multiple frailty factors, such as origination month frailty and originator frailty, to track the temporal fluctuation of default rates. In the following paragraphs, we define each of these variables and discuss their influence on default risk.

Mortgage age is defined as the number of months passed since the initiation of mortgages. It has been widely used in the literature to explain the trend of mortgage default. Most of the papers we reviewed discussed only fully-amortizing mortgages, which have zero balance at maturity date. Since the payment due at maturity is small compared to the principal, the default barely happens at maturity. However, our mortgages are 10-year-balloon mortgages, which are partially-amortizing. That is, the borrowers need to make a balloon payment at maturity, which is relatively large compared to the principal. Generally, this payment is funded by refinancing. Due to the potential failure of refinancing, borrowers of balloon mortgages are more likely to default at the maturity date than those of fully-amortizing mortgages.

Figure 2.1 provides evidence of high default rates at maturity for balloon mortgages. In the figure, default rate rises dramatically from the 120^{th} month (maturity month), peaks at the level of 0.03 in the 123^{th} month, then decreases sharply right after. This peak clearly shows the influence of balloon payment on default rates. Default could also happen before maturity because borrowers may not be able to make monthly payments. Figure 2.1 also demonstrates this phenomenon by showing that the number of defaults gradually increases in the first 51 months and then declines until the maturity date. In light of these two peaks on the curve, we design a piecewise linear function to capture the influence of mortgage age on default rates. This function is called the age function and has the following form:

$$
\begin{aligned}
age_1 &= min(age, 51)\\
age_2 &= max(min(120-51, age-51))\\
age_3 &= min(max(age-120, 0), 3)\\
age_4 &= max(0, age-123)\\
Age\_function &= age_1\beta_1 + age_2\beta_2 + age_3\beta_3 + age_4\beta_4
\end{aligned}
$$

where age is short for mortgage age in the formulas above.

Debt Service Coverage Ratio (DSCR) is another crucial variable in modeling mortgage default. It is defined as the ratio between net operating income and current debt obligations. DSCR larger one indicates borrowers have enough cash flow to make monthly payments. Otherwise, borrowers may default. We use the original value of DSCR without any transformation.

We use a dummy variable indicating whether servicers pay special attention to mortgages. We name this indicator variable as SW. If SW is one, then the servicer may consider the mortgage at high default risk. Otherwise, the situation of the mortgage is normal. SW enters into our models without transformations.

After describing variables for idiosyncratic risk, we turn our attention to variables tracking systematic risk. The first variable we use is the lagged value of default rates, denoted as lagged_PD. Since our models divides mortgage data into several groups, we compute lagged_PD for each group. We use logit transformation of lagged_PD here, because logistic regression is utilized in our models. The transformed variable is defined in the following way:

$$lagged\_PD' = log((lagged\_PD)/(1 - lagged\_PD))$$

Our empirical analysis shows this transformation significantly improves our model performance, so we use $lagged\_PD'$ instead of the original values.

The second set of variables we use are frailty factors, which capture the unexplained part of systematic risk. We develop three formulations for frailty factors. These formulations differ in the criterion to group the mortgages, and therefore, in the number of frailty factors. Since we use data of monthly frequency, all of these frailty factors are updated every month. The first formulation is called “single frailty”. In this formulation, all mortgages belong to one group, which indicates a common factor for all mortgages. The second formulation is called “origination frailty’, in which we divide mortgage data into several groups by their origination months. Accordingly, we have a separate frailty factor for each origination group. The third formulation is called “originator frailty”, where we group mortgages by their originators and obtain a multi-factor formulation.

2.4.2 Estimation

We build six models to examine the influence of the variables specified in the previous subsection on mortgage default rates. Table 2.1 lists the variables used in each model. The first three models, Static I, II and III, only contain static variables, such as $Age\_function$ and $DSCR$. The next three models are dynamic models which include both static variables and various frailty factors. In Table 2.1 the models appear from specific to general, i.e. each model is nested in the model on its right. For example, Static I is nested in Static II, because the former uses original values of age while the latter employs a flexible age function. Similarly, Dynamic III includes both origination frailty and originator frailty, while Dynamic II only contains origination frailty.

For the dynamic models, grouping of mortgages is a key step in constructing frailty factors. For “single frailty”, no grouping is needed. However, when using “origination frailty” and “originator frailty”, we need to carefully consider the groupings. On the one hand, a small group size is desired to ensure the similarity in credit quality among mortgages. On the other hand, a group that is too small may produce imprecise estimates, since mortgage default is a rare event. The group size should be large enough to produce smooth estimates of empirical default rates, which are inputs to our frailty factors. Therefore, to pick a proper group size, we have to strike a balance between controlling biases and producing smooth estimates.

For origination frailty, we choose two years as our window size to group mortgages. Ideally, mortgages originated in the same month should form a group, because mortgages initiated in the same month have less variation in credit quality than mortgages initiated in a relatively long period. However, the number of mortgages initiated in one month is too small to produce a smooth estimate of default rates. The two-year window size is chosen based on empirical tests. By comparing a number of alternative window sizes from six months to four years, we find that the window size of six months produces non-smooth default curves, and the window size of four years groups mortgages of different qualities together. The two-year window size appears to strike a balance as it not only controls the variation in mortgage quality but also keeps estimates of default rates smooth. Since the time range of our data set is 18 years, we have nine origination groups in total.

For originator frailty, we group mortgages based on the performance of their originators in the past. For each originator, we calculate the differences between empirical default rates and fitted default rates, the later implied by Dynamic II. All originators are then ranked by their differences, and originators of similar differences are grouped together. Similar to determining the number of origination groups, we also try to keep the balance between variation and sample size. As a result, five thresholds are specified, and originators with differences in between two adjacent thresholds fall into one group. We denote the difference as Δ and display the thresholds in Table 2.2. In total, we have 6 originator groups. When default rates is the main information used to assess originator preferences in this chapter, our modeling framework also works with alternative grouping criteria enhanced by more information.

We assume conditional independence for dependent variables in all the models and use standard maximum likelihood methods to estimate parameters. The data set has more than two million records, and the most sophisticated model has more than 15 factors which require updating every month. To accelerate the estimation process, we derive analytic gradients and to obtain good initial values for the optimization, we estimate the models sequentially from specific to general. As the models are nested, we select the initial values for the estimation based on the final estimates of the restricted model. Following these procedures, we can complete the estimation for our most complex model in two minutes with a standard desktop computer.

2.4.3 Results

Table 2.3 reports the estimation results for both static and dynamic models. All parameters are highly significant and have the expected signs. Except for the Static I model, we use a piecewise-linear function to capture the influence of mortgage age on default rates in all models. In particular, since our data set consists of balloon mortgages, a piece-wise linear function is designed to accommodate the effects of balloon payments on mortgage default. We find that the coefficients of age_1 and age_3 are positive, and the coefficients of age_2 and age_4 are negative. This is consistent with the trend of empirical default rates shown in Figure 2.1. Our results provide evidence that the effect of DSCR on mortgage default is negative as expected, since borrowers with higher income are less likely to default.

The lower part of Table 2.3 presents parameter estimates related to latent factors. Note that Dynamic III has four parameters, while Dynamic I and II only have two parameters. In Dynamic I, α and θ are the autoregressive and innovation coefficients of the single-frailty factor. In dynamic II and III, α_1 and θ_1 are coefficients for the origination factors, and α_2 and θ_2 are coefficients for the originator factors. The positive estimates of α, α_1, and α_2 indicate that fitted default rates are adjusted towards empirical default rates when there are mismatches between these two rates.

The last row of Table 2.3 reports the log-likelihood of each model. As we can see, this number increases from the leftmost model to the rightmost model, and the smallest difference between any two likelihoods is over 1000. Since Dynamic III has the highest log-likelihood and the first five models are nested in it, these numbers suggest Dynamic III has the best sample fit. In the following paragraphs, we use a series of diagnostic plots to demonstrate the advantages of Dynamic III.

To begin with, Figure 2.2 illustrates the improvement from Static I to Static II. As we know, Static II augments Static I by introducing a piecewise-linear age function. In Figure 2.2, the empirical default rates by mortgage age are represented by a solid line. The default curve implied by Static II, shown by a dark dashed line, follows the empirical default curve closely. By contrast, the default curve generated by Static I, drawn by a light dashed line, barely tracks the trend of the empirical default curve. This comparison is not surprising at all, because a constant coefficient cannot fit the non-linear effect of age on default.

While Figure 2.2 shows the advantages of the age function, Figure 2.3 and 2.4 demonstrate the benefits of tracking systematic risk with $lagged\_PD'$. In these two figures, default rates are computed by exposure months, and the empirical default curve exhibits significant default clustering around 2005 and 2011. Unfortunately, as shown in Figure 2.3, the curves generated by Static I and II cannot approximate the default clusters. In Figure 2.4 the curve implied by Static III roughly follows the empirical default curve and its clustering patterns. These figures convey that $lagged\_PD'$ is a good proxy for systematic risk. However, the mismatches between the empirical curve and the fitted curve are still not negligible.

Considering the mismatches in Figure 2.4, we further include single-frailty in Dynamic I. Namely, we allow a separate dynamic intercept for each month, and the intercept is the same for all mortgages. Figure 2.5 displays the fitted curve by Dynamic I along with the empirical curve. The fitted curve matches the empirical one not only in the period of default clustering (2005 and 2011), but also in almost all other periods of the data set. The closeness of these two curves shows that mismatches are largely corrected by the dynamic intercept.

These results convey that a dynamic intercept would suffice if we only want to track the default rates in the dimension of exposure month. However, inconsistency appears again when we investigate whether fitted default rates follow empirical ones along the dimension of origination month. To show this inconsistency, we group mortgages by their origination months and plot default rates in Figure 2.6. We observe a noticeable gap for mortgages initiated after 2008. While the empirical default rate drops to zero for these mortgages, the fitted default rate increases until 2011 and remains above 0.008 in the next two years. Besides, the mismatches for mortgages initiated before 2004 are not negligible, either. Clearly, the model with a single dynamic intercept overestimates the default risk for mortgages initiated after 2008 and yields rough predictions for mortgages initiated before 2004.

To narrow the gaps in Figure 2.6, we allow separate intercepts for different origination groups in Dynamic II. Specifically, we group mortgages by origination with a two-year window size and allow each group to have its own intercept (the details appeared in the model specification section). Figure 2.7 evidently shows the improvement after allowing a separate intercept for each origination group. The fitted default rates by Dynamic II sharply decline after 2008 and remains at zero thereafter. Additionally, the fitted curve also closely tracks the empirical one before 2004.

In Dynamic II, we fit default rates along dimensions of not only exposure month but also origination month. As known, originator preference also influences default rates significantly. We proceed to explore this fitting dimension and report our findings in Figure 2.8 and 2.9. In Figure 2.8, we re-group the data by originator group and plot the empirical curve with the fitted curve of Dynamic II. The gap between these two curves illustrate that Dynamic II cannot model the variation of default rates caused by originator preference.

Although Figure 2.8 shows mismatches of default rates at the level of originator group, it is still not clear whether there are systematic mismatches within each group. To examine the possibility of systematic deviation, we divide each originator group into several sub-groups by mortgage age, and plot the fitted and empirical rates in Figure 2.9. Since each group corresponds to an empirical curve and a fitted curve, we plot 12 curves in Figure 2.9. These 12 curves confirm the existence of systematic deviation within each originator group. For the first three groups, these curves imply that the fitted rates are lower than the empirical rates for almost all mortgage ages. For the last three groups, these curves suggest an over-estimation of default rates across nearly all mortgage ages.

Motivated by Figure 2.8 and 2.9, we supplement origination frailty factors with originator frailty factors in Dynamic III. Figure 2.10 and 2.11 support our modification. In Figure 2.10, the empirical curve largely overlaps the fitted curve by Dynamic III. This overlapping shows the reduction of gaps by employing extra frailty factors. Furthermore, Figure 2.11 plots six sets of overlapping curves. This indicates that the systematic deviations shown in Figure 2.9 are also corrected.

2.5 Conclusions

To the best of our knowledge we are the first to use observation-driven frailty models to predict commercial real estate mortgage defaults. In particular, we introduced a class of frailty models to track the variations of mortgage default rates in multiple dimensions. Our frailty factors track origination and originator characteristics. The frailty factors enable our models to track the variation of default rates in three dimensions: exposure month, origination month and originator group. In our empirical application, a series of models were constructed to investigate the effects of frailty factors on default rates. We tested performance of the models using a data set with more than two million records. The models exhibit computational advantages when using the large data set. Since our data set consists of balloon mortgages, a piece-wise linear function is designed to accommodate the effects of balloon payments on mortgage default. Financial information of borrowers is also used to track idiosyncratic risk of mortgages. Diagnostic plots and statistic test results demonstrate the superior performance of our dynamic models. While we only explore three fitting dimensions, our framework can be easily generalized by including more latent factors and fitting default rates in more dimensions.

Table 2.1: Components of Static and Dynamic Models

	Static I	Static II	Static III	Dynamic I	Dynamic II	Dynamic III
Age	X					
Age_function		X	X	X	X	X
DSCR	X	X	X	X	X	X
SW	X	X	X	X	X	X
$Lagged\_PD'$			X	X	X	X
Exposure Frailty				X		
Origination Frailty					X	X
Originator Frailty						X

Notes: This table describes components of each models in empirical applications. X means the variable is a part of the model.

Table 2.2: The Grouping Criterion for Originator Frailty

Group	Difference
1	$\Delta > 0.02$
2	$0.01 < \Delta \leq 0.02$
3	$0 < \Delta \leq 0.01$
4	$-0.005 < \Delta \leq 0$
5	$-0.01 < \Delta \leq -0.005$
6	$\Delta \leq -0.01$

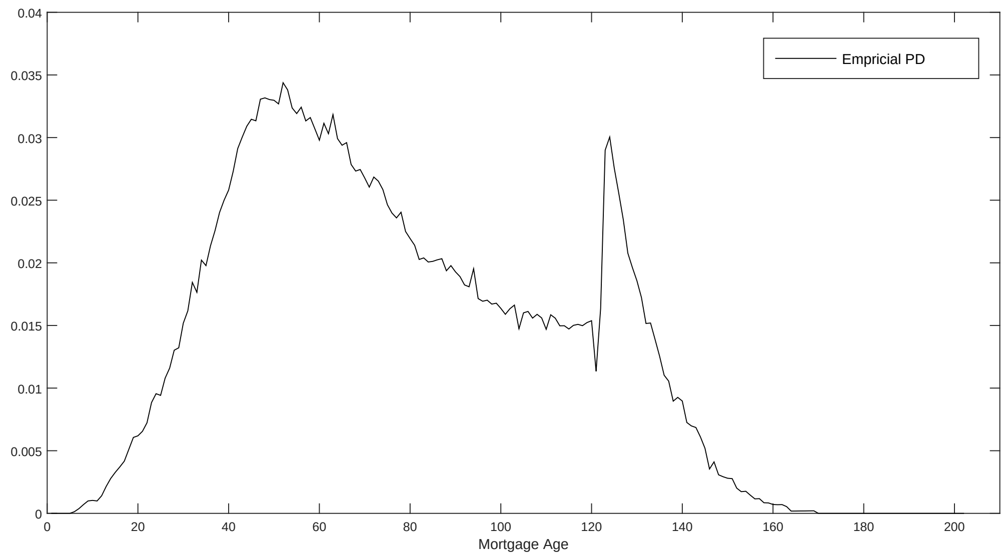
Notes: The table specifies the thresholds used in grouping originators.

Table 2.3: Estimates for Static and Dynamic Models

	Static I	Static II	Static III	Dynamic I	Dynamic II	Dynamic III
Age	0.0014	-	-	-	-	-
	(2.8)					
Age_1	-	0.0585	0.0429	0.0466	0.0407	0.0256
		(56)	(48)	(60)	(38)	(32)
Age_2	-	-0.017	-0.0178	-0.0207	-0.0207	-0.0220
		(-37)	(-50)	(-55)	(-45)	(-38)
Age_3	-	0.4965	0.4447	0.4253	0.362	0.3577
		(52)	(65)	(74)	(40)	(35)
Age_4	-	-0.1014	-0.1128	-0.1129	-0.1032	-0.1018
		(-51)	(-39)	(-34)	(-28)	(-33)
$DSCR$	-1.673	-1.628	-1.553	-1.530	-1.499	-1.491
	(-190)	(-157)	(-126)	(-139)	(-114)	(-120)
SW	-3.32	-3.2176	-3.29	-3.324	-3.329	-3.329
	(-147)	(-83)	(-77)	(-99)	(-80)	(-88)
$Intercept$	-1.7083	-3.86	1.37	-	-	-
	(-16)	(-8)	(-7.5)			
α	-	-	-	49.5	-	-
				(28)		
θ	-	-	-	0.992	-	-
				(672)		
α_1	-	-	-	-	28.9	20.21
					(117)	(86)
θ_1	-	-	-	-	0.987	1.02
					(1320)	(803)
α_2	-	-	-	-	-	7.13
						(92)
θ_2	-	-	-	-	-	0.995
						(772)
$logL$	-164524	-156510	-148925	-148137	-146523	-145392

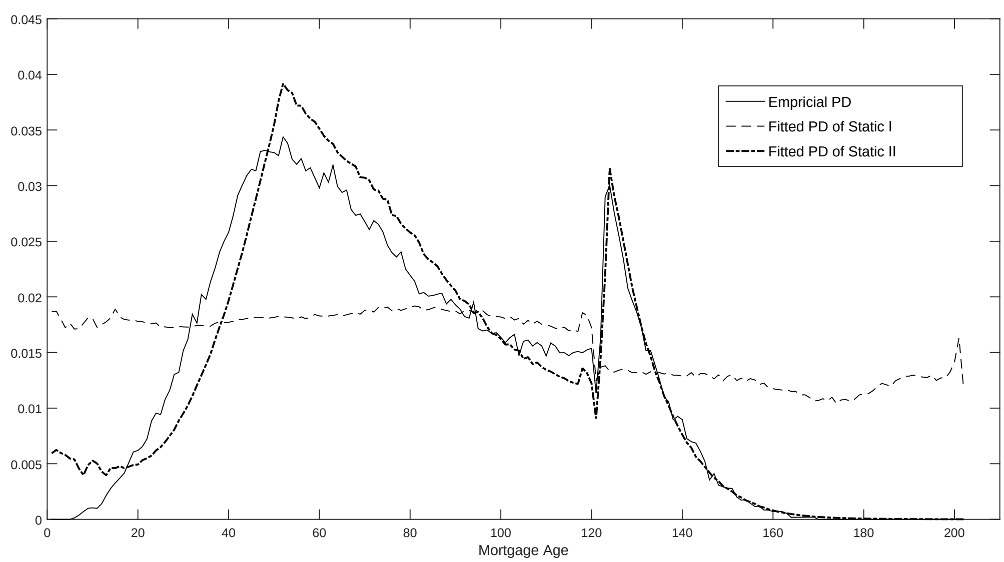
Notes: This table reports the estimates for static and dynamic models. T-statistics are in parentheses.

Figure 2.1: Empirical Default Rates (PD) by Mortgage Age



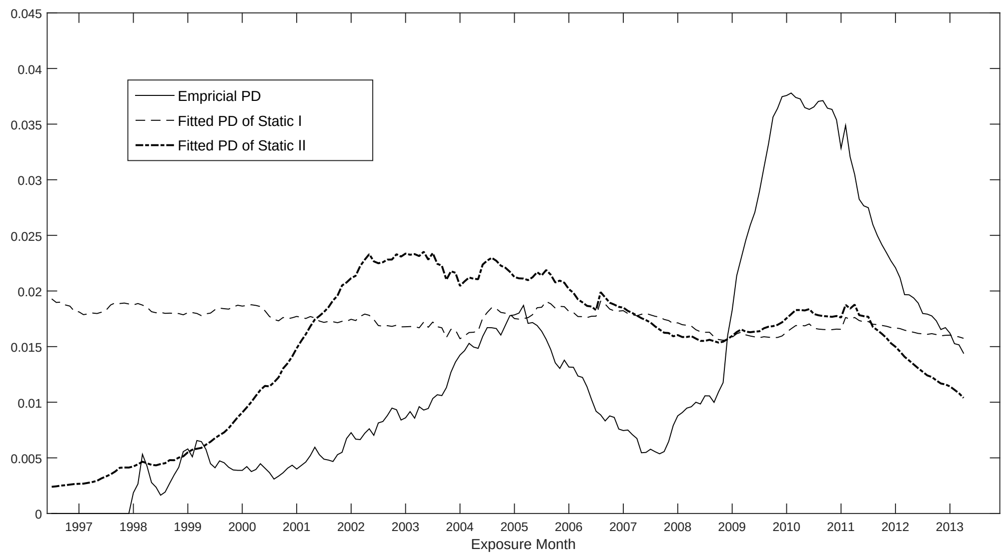
Notes: The picture shows empirical default rates of mortgages grouped by mortgage age. The horizontal axis is mortgage age in months and the vertical axis is default rates.

Figure 2.2: Default Rates by Mortgage Age of Static I and Static II



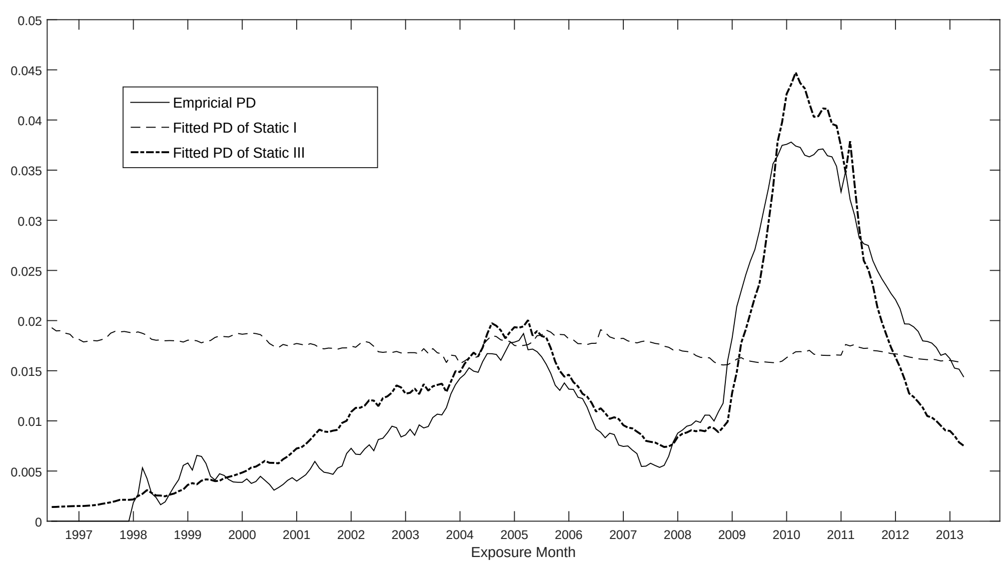
Notes: The horizontal axis is mortgage age in months. The solid line shows empirical default rates. The light dashed line represents fitted default rates of Static I, which uses mortgage age, DSCR and SW. The dark dashed line refers to fitted default rates of Static II. Variables of this model consist of mortgage age function, DSCR and SW.

Figure 2.3: Default Rates by Exposure Month of Static I and Static II



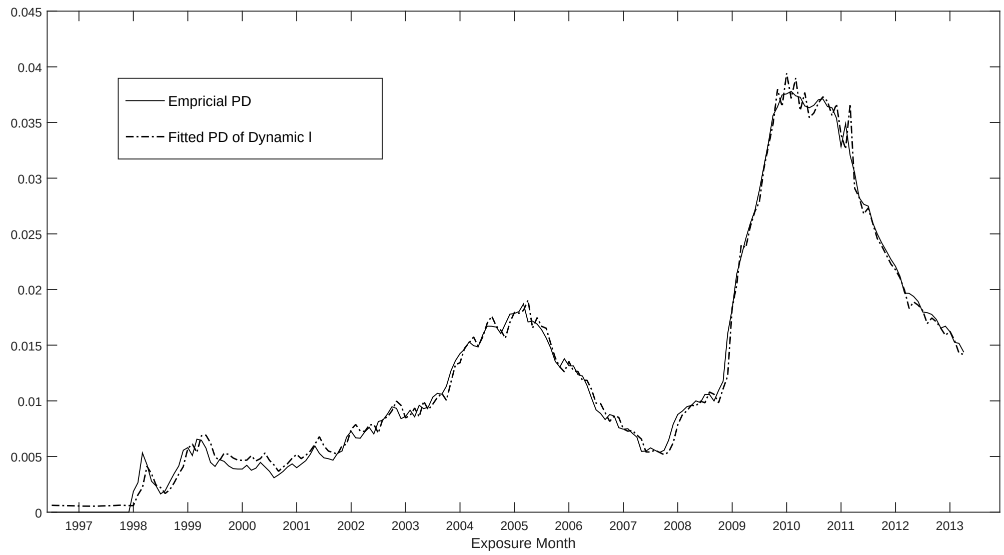
Notes: The picture reports default rates of mortgages grouped by exposure month. The horizontal axis is exposure month. The solid line shows empirical default rates. The light dashed line represents fitted default rates of Static I, which uses mortgage age, DSCR and SW. The dark dashed line refers to fitted default rates of Static II. Variables of this model consist of mortgage age function, DSCR and SW.

Figure 2.4: Default Rates by Exposure Month of Static I and Static III



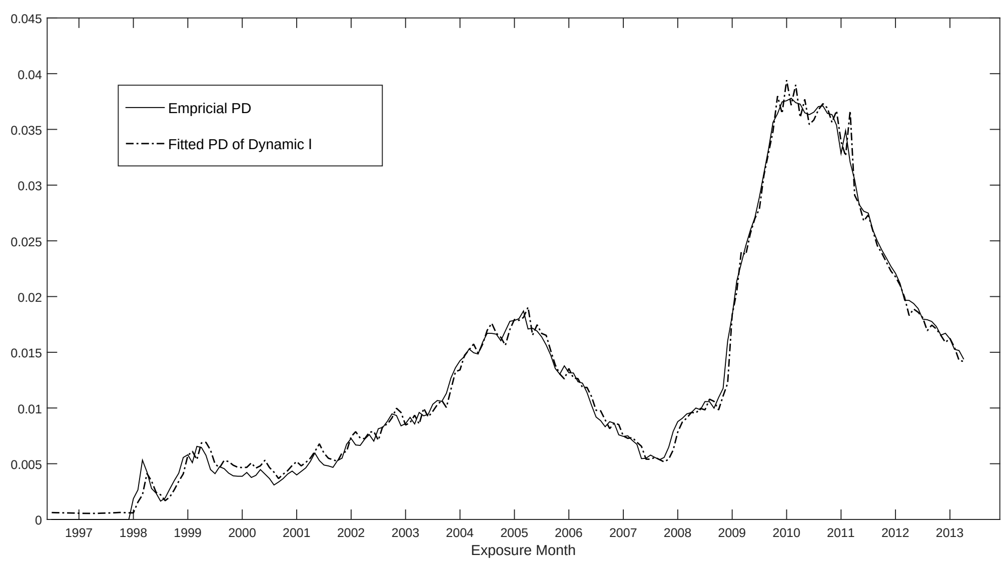
Notes: The picture reports default rates of mortgages grouped by exposure month. The horizontal axis is exposure month. The solid line shows empirical default rates. The light dashed line represents fitted default rates of static model one, which uses mortgage age, DSCR and SW. The dark dashed line refers to fitted default rates of static model three. Variables of this model consist of mortgage age function, DSCR, SW, and lagged_PD.

Figure 2.5: Default Rates by Exposure Month of Dynamic I



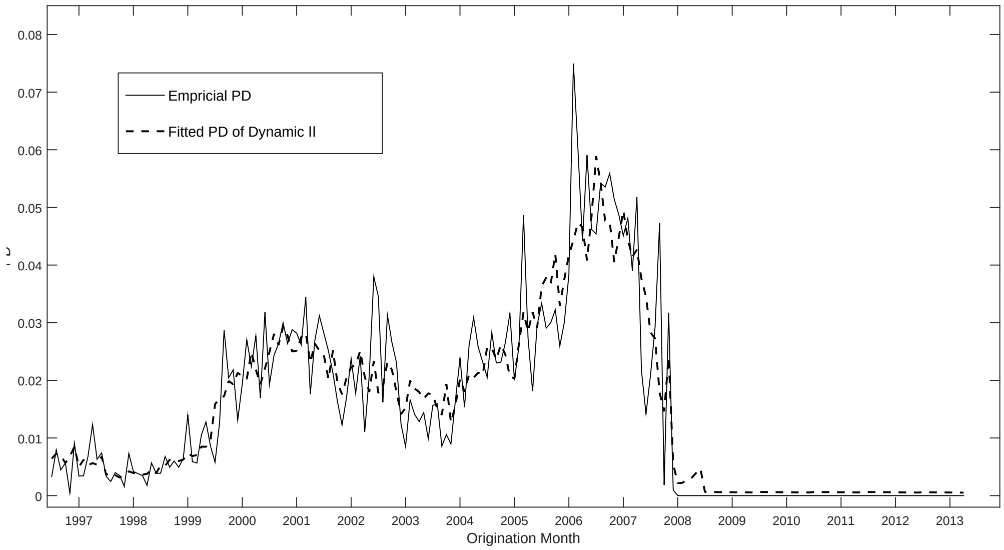
Notes: The picture reports default rates of mortgages grouped by exposure month. The horizontal axis is exposure month. The solid line shows empirical default rates. The dashed line represents fitted default rates of Dynamic I. Variables of this model include mortgage age function, DSCR, SW, lagged_PD, and single-frailty factor.

Figure 2.6: Default Rates by Origination Month of Dynamic I



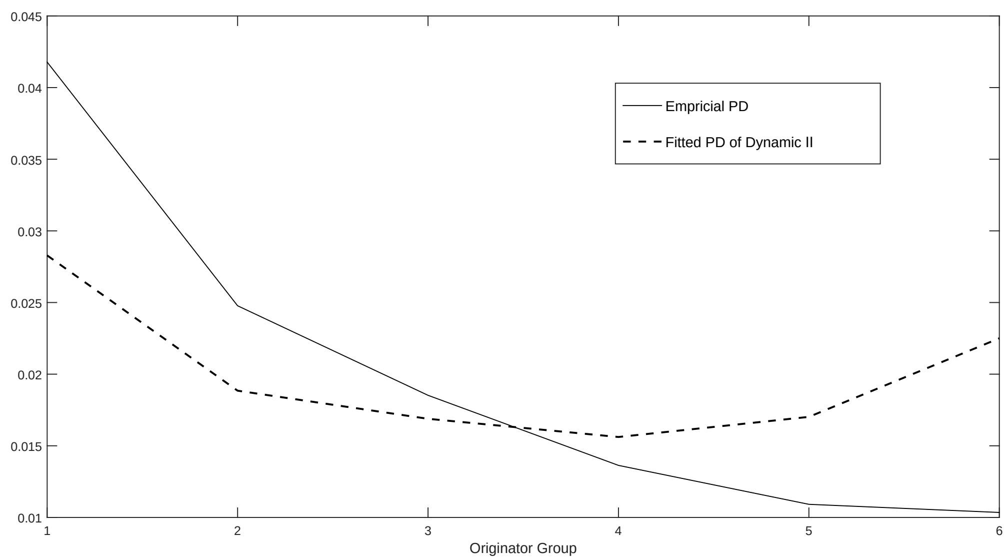
Notes: The picture reports default rates of mortgages grouped by origination month. The horizontal axis is origination month. The solid line shows empirical default rates. The dashed line represents fitted default rates of Dynamic I. Variables of this model include mortgage age function, DSCR, SW, lagged_PD, and single-frailty factor.

Figure 2.7: Default Rates by Origination Month of Dynamic II



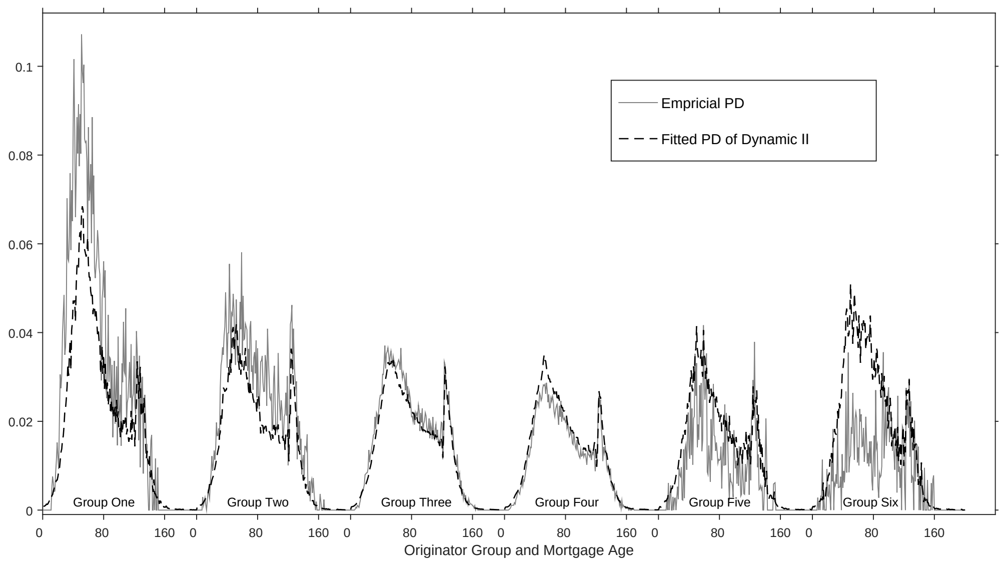
Notes: The picture reports default rates of mortgages grouped by origination month. The horizontal axis is origination month. The solid line shows empirical default rates. The dashed line represents fitted default rates of Dynamic II. Variables in this model include mortgage age function, DSCR, SW, lagged_PD, and origination month frailty factor.

Figure 2.8: Default Rates by Originator Group of Dynamic II



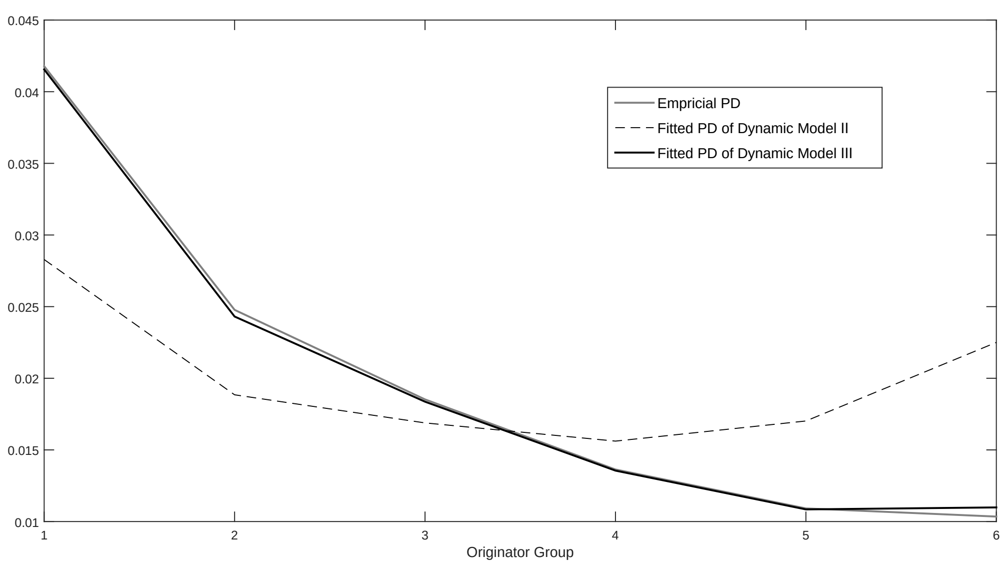
Notes: The picture reports default rates of mortgages grouped by originator. The horizontal axis in the figure is originator group. The solid line shows empirical default rates. The dashed line represents fitted default rates of Dynamic II. Variables in this model include mortgage age function, DSCR, SW, lagged_PD, and origination month frailty factor.

Figure 2.9: Default Rates by Originator Group and Mortgage Age of Dynamic II



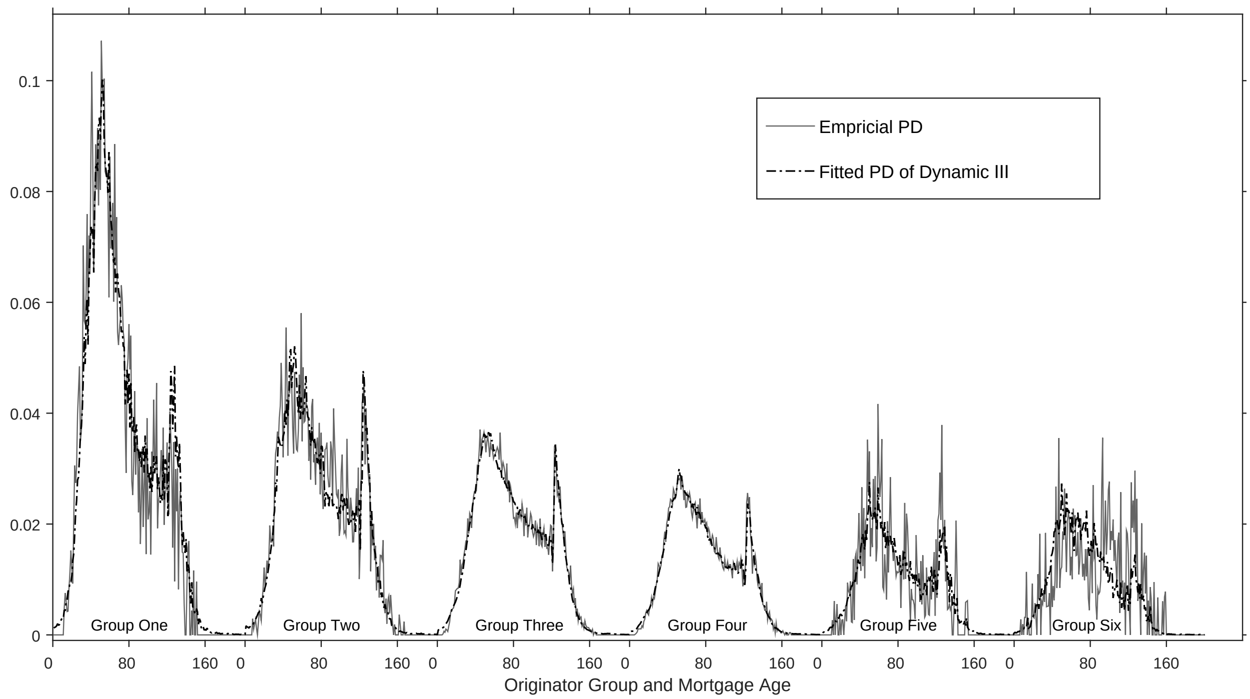
Notes: The picture reports default rates of mortgages grouped by originator and mortgage age. The horizontal axis in the figure is originator group and mortgage age. The light line shows empirical default rates. The dark line represents fitted default rates of Dynamic II. Variables in this model include mortgage age function, DSCR, SW, lagged_PD, and origination month frailty.

Figure 2.10: Default Rates by Originator Group of Dynamic II and Dynamic III



Notes: The picture reports default rates of mortgages grouped by originator. The horizontal axis is originator group. The light line shows empirical default rates. The dark line is fitted default rates of Dynamic III. Variables in this model include mortgage age function, DSCR, SW, lagged_PD, origination frailty factor, and originator frailty factor. The dashed line pertains to fitted default rates of Dynamic II, which include mortgage age function, DSCR, SW, lagged_PD, and origination month frailty factor.

Figure 2.11: Default Rates by Originator Group and Mortgage Age of Dynamic II and Dynamic III



Notes: The picture reports default rates of mortgages grouped by originator and mortgage. The horizontal axis is originator group and mortgage age. The light line shows empirical default rates. The dark line represents fitted default rates of Dynamic III. Variables in this model include mortgage age function, DSCR, SW, lagged_PD, origination frailty factor, and originator frailty factor.

CHAPTER 3: COMMERCIAL AND RESIDENTIAL MORTGAGE DEFAULTS: SPATIAL DEPENDENCE WITH FRAILTY

3.1 Introduction

Both the academic literature and the industry separate commercial and residential mortgage risks. In academia there is an extensive literature pertaining to residential mortgage design and risks, and there is a somewhat less voluminous literature treating commercial mortgages. In industry there is an organizational silos structure. Financial institutions usually have a residential real estate unit that is operating completely independently of the commercial mortgage department.

In reality, defaults in residential mortgages and commercial ones are interrelated. When a business initiates major layoffs, it will have ripple effects on the housing as well as retail sector in adjacent locations. Cities with major economic declines, such as Detroit, are prime examples revealing the interconnections between the default risks across households and businesses.

Researchers have used various approaches to model the credit risk of corporate debt, mortgages and derivatives. In general, these models can be divided into two categories - structural models and reduced-form models. This paper presents a reduced-form model that contains frailty factors to predict commercial mortgage default. Das et al. (2007) – while studying corporate bond defaults – provided evidence that macro variables alone were not enough to explain all the systematic risk. They further demonstrated that the lack of explanatory variables underestimates value-at-risk. To produce unbiased estimates, they proposed frailty models for corporate bond credit risk, to account for the unexplained part of systematic risk. Building on our earlier work, Chen, Ghysels and

Telfeyan (2016), we propose a new class of spatial frailty models linking commercial and residential mortgage risk.

In this paper we use the aforementioned frailty model setup to explore spatial dependence between commercial and residential mortgage risk. Our commercial mortgage data contains 1,601,617 records of commercial mortgages in the U.S. retail industry. The records start in January 1999 and end in January 2016. There are 16,719 distinct mortgages and all of them are 10-year-balloon mortgages. Our residential mortgage data consists of 140 million records of single family mortgages in the U.S. market. The earliest records were reported in January 1999 and the latest ones were documented in March 2016. There are 3.2 million distinct mortgages, and all of them are fully amortizing and fixed rate.

To the best of our knowledge, this is the first paper exploring the spatial dependence between commercial mortgage default and residential mortgage default. Our empirical analysis demonstrates strong spatial dependence between commercial default and residential default in multiple respects. First, we apply Granger causality tests to the default rates of commercial mortgages and residential mortgages in the top 10 MSA areas, and the test results reveal a significant lead and lag relationship for the two mortgage markets in 9 areas. Second, we test the causal relation among the frailty factors that explain systematic risk of commercial mortgage and residential mortgage, and provide strong evidence on the close correlations between the residential and commercial mortgage markets. Last but not least, we show that residential PD is a good explanatory variable in predicting commercial mortgage default in the same area, and this prediction power also implies that local residential market drives the commercial market.

The high tractability of our frailty models plays a core part in exploring the causal relations of big data sets. While our computational tool is only a plain laptop, the estimation process usually finishes in minutes even with datasets of millions of records. For example, our largest residential dataset contains 31 million records, and the estimation for frailty models finishes in less than 5 minutes. Moreover, for a complicated commercial mortgage models with more than 20 frailty factors and millions of records, the estimation completes in a half minute.

The remainder of this paper is organized as follows. The next section reviews the background of frailty models. Section 3 formulates our dynamic frailty models. In section 4, we discuss empirical applications. The last section offers some concluding remarks.

3.2 Background

In this section, we review studies that are relevant to spatial frailty models. We first take a look at the older studies that used reduced-form approaches to model the credit risk of corporate debts and mortgages. Then, we review recent papers that utilize observation-driven model and those used parameter-driven models. Next, we summarize the mortgage literature exploring spatial information.

Research on credit risk prediction has a long history, dating back to Beaver (1968) and Altman (1968). Since then, numerous models have been developed to assess risk factors and predict default events. Following the classification in Altman et al. (2005), credit risk models are categorized into two groups: structural-form models and reduced-form models. While structural-form methods assess credit risk elements through an implied value process of fixed income products, reduced-form models impose separate and explicit assumptions on risk measures like default probabilities and recovery rates. Specifically, reduced-form models assume credit risk elements are not only related to the structural features of the firm, but also depend on macro variables and financial information of borrowers. For a further discussion on the comparison between structural-form models and reduced-form models, please refer to Jarrow and Protter (2004). In this paper, we only review papers that used reduced-form models because we focus on reduced-form models.

Research on reduced-form methods started with Beaver (1968) and Altman (1968) who used discriminant analysis as the main tool.[1] Since Ohlson (1980) and Zmijewski (1984), binary response models like logit and probit regressions have been introduced into credit risk modeling. Most of these models estimate single-period default probabilities or credit scores. Some recent studies began to extend the prediction horizon to multiple periods using multiple logit models (for example, see Campbell, Hilscher and Szilagyi 2008). These reduced-form models assume that macro variables are able to explain all the systematic risk. However, Das et al. (2007) stated that this assumption could be violated and provided evidence on it. Following Das et al. (2007), Duffie et al. (2009) introduced the notion of common latent factors - or so called frailty factors - into the default intensity of proportional hazard models for corporate debt when evaluating credit risk. They showed that failure to control for these latent factors could cause downward biases in calculating value-at-risk. Since the seminal work of Das et al. (2007), increasing attention has been paid to credit risk modeling with frailty factors. Koopman and Lucas (2008) added latent dynamic factors to logistic regressions and developed a non-Gaussian multivariate state-space model to predict corporate default. Koopman, Lucas and Schwaab (2012) and Schwaab, Koopman and Lucas (2015) extended the framework of Koopman and Lucas (2008) by jointly modeling macro variables and default events, by including industry effects ad regional effects.

Following the definitions in Cox et al. (1981), frailty models can be divided into two types: parameter-driven models and observation-driven models. All the models mentioned in the last paragraph fall into the category of parameter-driven models. Meanwhile, some researchers use observation-driven to model credit risks. For example, Creal, Koopman and Lucas (2013*b*) designed a class of observation-driven models - called GAS

[1] These models are often called “the first generation of reduced-form models” (Duan, Sun and Wang 2012).

models - and applied them to Moody's credit rating data. GAS models use scaled scores of the likelihood function to update the dynamics of latent factors. Using GAS models, Creal et al. (2014) jointly modeled macro variables and default outcomes with data of mixed-measurement and mixed-frequency.[2] While the aforementioned work developed frailty models for corporate debts, Chen, Ghysels and Telfeyan (2016) further extended the GAS framework to model commercial mortgages and matched the default risk of commercial mortgages in multiple dimensions. In terms of computational complexity, observation driven models have evident advantages over parameter-driven models, because the estimation process of the latter usually requires intensive simulation, which is not needed by the former.

When designing the structures for frailty factors, researchers rarely assume a single frailty factor for all assets in the portfolio. In tradition, they group corporate debts or mortgages based on certain criteria and assume separate latent factors for each group. Popular grouping criteria are rating classes, industries, origination time, originators, and locations. Two main reasons justify using separate latent factors for each group. One reason is that fixed income products have different credit quality, and their exposure to common sources of systematic risk varies. Therefore, a separate latent factor for each group captures the variation of exposure for different products. The other reason is that fixed income products may absorb different business shock in the same business cycle, depending on factors like regional economy or industry performance. Accordingly, separate factors can explain the varying risk of different business environment.

As discussed earlier, researchers usually associated the frailty factors with key risk factors, such as rating classes and industry types. Among various risk factors, the locations of mortgages are of special importance, because revenues of underlying properties largely

[2] If one applies GAS models to binary logistic regressions and chooses the intercept as the dynamic parameter, then the score of the dynamic intercept is actually the generalized residual defined in Gouriéroux et al. (1987). In this context, GAS models reduce to the generalized autoregressive moving average models proposed in Shephard (1995) and Benjamin, Rigby and Stasinopoulos (2003).

depend on the local economy. Considering the importance of locations, researchers have developed various spatial frailty models. Deng, Pavlov and Yang (2005) proposed a space-varying coefficient model for residential mortgages and the space-varying coefficients actually play the role of spatial latent factors. Since their models require a three-stage maximum likelihood estimation, the computational burden limit further applications of their models. Non-parametric methods are also explored to capture the unobservable spatial variations of mortgage default. Based on Follain, Ondrich and Sinha (1997), Kau, Keenan and Li (2011) introduced MSA-level frailty and demonstrated that MSA-level frailty has significant effects on mortgage terminations risks. Before the introduction of frailty models, location dummies are the main tools to explain spatial variations of mortgage default (for example, Yildirim 2008).

In addition to location, local macroeconomic variables are also used to explain the spatial variations of mortgage default. Popular variables are local divorce rates, unemployment rates and house price index. For instance, An et al. (2013) used local home price appreciation to predict commercial mortgage default and concluded that local home price appreciation is a powerful explanatory variable for commercial default. To our best knowledge, this paper is the first attempt that links the markets of commercial real estate to residential real estate.

3.3 Model Formulation

In this section, we introduce the model formulations to prediction mortgages default and discuss the specifics in modeling residential mortgages and commercial mortgages.

To set up the model, consider a set of n_t mortgages and let the vector of default status be denoted as $y_t = [y_{1,t}, ..., y_{n_t,t}]$, where n_t is the number of mortgages at time t. The element $y_{i,t}$ of this vector, is a binary variable referring to the default status of mortgage i at time t; it equals to one if a default happens at time t, and zero if the borrower can make timely payment. Let us further denote the default probability as $\pi_{i,t}$, i.e. $\pi_{i,t} = P(y_{i,t} = 1)$. To relate the default probability with covariates, we adopt a binary logit model:

$$\pi_{i,t} = exp(\mu_{i,t})/(1 + exp(\mu_{i,t}))$$

where $\mu_{i,t}$ is a transformed default probability. It is defined as a linear combination of covariates and a latent factor:

$$\mu_{i,t} = x'_{i,t}\beta + f_{i,t}.$$

where the vector $x_{i,t}$ consists of a series of observed variables, such as financial information of the mortgage, β is a coefficient vector and constant across all mortgages, and $f_{i,t}$ is a latent factor or factors - as clarified later. It is important to emphasize one interpretation of the above equation. The presence of $f_{i,t}$ can be viewed as the *intercept* of the default intensity and captures the unexplained systematic risk.

As is discussed in Section 3.2, the exposure of mortgages to systematic risk are largely influenced by two factors: locations and credit quality of mortgages. Researchers group mortgages based on these two criteria and assume latent factors differ across groups. In the following paragraphs, we explain our grouping criteria for commercial mortgages and residential mortgages in details.

For commercial mortgages, we choose origination months and locations as the grouping criteria. Origination months are used as a proxy for the credit quality of mortgages. In Chen, Ghysels and Telfeyan (2016), they found that default rates of commercial mortgages vary significantly with the origination month and concluded that origination month is a good proxy for credit quality of mortgages. Since we found similar patterns in the empirical analysis, we adopt this criterion. Specifically, we put mortgages of consecutive origination months in the same group and use c_i^1 to denote the group number of mortgage i according to the criterion of origination month. Meanwhile, mortgages of adjacent areas are grouped together and c_i^2 represents the group number of mortgage i according to the grouping criterion of locations.

We assume mortgages in the same group share one common latent factor and allow latent factors to vary across different groups. Since mortgages in different groups have distinct latent factors, every grouping criterion corresponds to a separate set of latent factors. Combining this assumption with our two grouping criteria, we decompose the latent factor specification for mortgage i at time t, $f_{i,t}$, into two parts:

$$f_{i,t} = f^1_{c^1_i,t} + f^2_{c^2_i,t}$$

where $f^1_{c^1_i,t}$ represents the set of frailty factors related to "origination month", and $f^2_{c^2_i,t}$ stands for the set of frailty factors relevant to "locations". Generally, if n grouping criteria are identified then $f_{i,t}$ is decomposed into n parts.

We further assume that the latent factors have an autoregressive form, namely:

$$f^k_{c^k_i,t} = \theta_k f^k_{c^k_i,t-1} + \alpha_k s^k_{c^k_i,t-1}, k \in \{1,2\}$$

The time subscript to the innovation $s^k_{c^k_i,t-1}$ indicates that it is computed using information at time $t-1$. The choice of innovation is crucial in updating the dynamics of frailty factors, and use generalized residuals, as dubbed by Gouriéroux et al. (1987). Specifically, we characterize the innovations as:

$$\begin{aligned}
s^k_{c^k_i,t-1} &= \bar{y}^k_{c^k_i,t-1} - \hat{y}^k_{c^k_i,t-1} \\
\bar{y}^k_{c^k_i,t-1} &= \frac{1}{n_{t-1}} \sum_{j=1}^{n_{t-1}} y_{j,t-1} 1_{c^k_j = c^k_i} \\
\hat{y}^k_{c^k_i,t-1} &= \frac{1}{n_{t-1}} \sum_{j=1}^{n_{t-1}} \hat{\pi}_{j,t-1} 1_{c^k_j = c^k_i}
\end{aligned}$$

where $\bar{y}^k_{c^k_i,t-1}$ is the empirical default rate for group c^k_i at time $t-1$, and $\hat{y}^k_{c^k_i,t-1}$ is the fitted default rate for group c^k_i at time $t-1$. $1_{c^k_j=c^k_i}$ is an indicator variable; it equals to one if mortgage j is in group c^k_i and zero otherwise. Using this indicator variable, we include only information from group c_i^k to update the frailty factors related to mortgage i. Likewise, $\hat{\pi}_{j,t-1}$ is the estimated default probability (PD) for mortgage j. Intuitively, our innovation term is the difference between empirical default probability and fitted default probability. The innovation therefore measures the distance between models and data. If the innovation term is positive(negative), then the empirical PD is larger(smaller) than fitted one.

When modeling commercial mortgages, we assume a single set of parameters for mortgages in different areas. An alternative formulation is allowing separate set of parameters for mortgages in different areas and performing individual estimation for each area. Obviously, the latter formulation is more flexible than the former one. However, the latter formulation divides the full sample into small groups and may not yield smooth estimates for empirical PD, which restricts its application to commercial mortgages. In contrary, our residential dataset has 1.04 billion records and has no concerns of sample size. Therefore, we adopt the latter formulation to model residential mortgages and follow a two-step procedure. In the first step, we divide the sample of residential mortgages into several sub-samples by locations. Each spatial group assumes a separate set of parameters, including β, θ, α. In the second step, we further divide borrowers in the same spatial group into several subgroups by their FICO scores. We allow each subgroup a separate latent factor. Note that the dynamics of this FICO frailty is updated in the same way as frailty factors of commercial mortgages as we discussed above. The FICO frailty factor assume an autoregressive form, and the innovation terms of FICO factors measure the differences between the empirical PD and fitted PD of the same spatial group.

3.4 Empirical Applications

3.4.1 Data

In this paper, we utilize two datasets to study the spatial dependence between the default of commercial mortgages and residential mortgages. One dataset with commercial mortgages is provided by Trepp LLC [3], and the other with residential mortgages is published by Freddie Mac. For commercial mortgages, the dataset contains 1,601,617 records of mortgages in the U.S. retail market. The records started in January 1999 and ended in January 2016. The origination time of the mortgages is between January 1995 and December 2015. There are 16,719 distinct mortgages, and all of them are 10-year-balloon mortgages. Table 3.1 provides detailed information on the top 10 MSA areas we focus on. Among these MSA areas, the Los Angeles area has the most number of mortgages, which are 878, and the Anaheim area has the least number of mortgages, which are 289. In total, these 10 MSA areas include almost 30 percent observations of the entire dataset.

For residential mortgages, the original dataset contains 1.04 billion records of residential mortgages in the U.S. market. There are 22.5 million distinct mortgages and all of them are fixed-rate fully amortizing mortgages. The data we used in estimation only consists of mortgages in the top 10 MSA areas, which contains 0.14 billion records and 3.2 million distinct mortgages. Our sample contains approximately 14 percent of the original dataset published by Freddie Mac. The time period of this dataset is from February 1999 to March 2016, and the origination time of the mortgages is between the first quarter of 1999 and the third quarter of 2015. Table 3.2 provides information of these mortgages in the 10 MSA areas: the Chicago area has the most number of records and mortgages, amounting to 31 million and 0.74 million respectively; the Anaheim area has the least numbers of records and mortgages, which are 3 million and 0.09 million

[3] which is a leading provider of real estate data

respectively; for other regions, the number of records is generally between 10 million and 20 million, and the count of mortgages is between 0.2 million and 0.5 million.

3.4.2 Variables for Commercial Mortgages

As discussed earlier, the default risk of commercial mortgages consists of idiosyncratic risk and systematic risk. To explain idiosyncratic risk, we use mortgage age and debt service coverage ratio. For systematic risk, we use lagged values of default rates and a single frailty factor or multiple frailty factors, such as origination month frailty and spatial frailty, to track the temporal fluctuation of default rates. In the following paragraphs, we define each of these variables and discuss their influence on default risk.

Mortgage age is defined as the number of months passed since the initiation of mortgages. It has been widely used in the literature to explain the trend of mortgage default, which is usually approximated by a continuous function. The selection of this function largely depends on the amortization types of mortgages. Our commercial mortgages data are all 10-year-balloon mortgages and are therefore partially amortizing. In the case of partially-amortizing mortgages, usually there are two default clusterings observed in the life of mortgages: one happens around the fifth year and the other occurs at the maturity. The former is due to deteriorating financial situations of the borrowers, and the latter is caused by failures of refinancing. Since we found similar double default clusterings in our data, we design a piecewise linear function to capture the influence of mortgage age on default rates. This function is called the age function and has the following form:

$$\begin{aligned}
age_1 &= min(age, 56) \\
age_2 &= max(min(120-56, age-56), 0) \\
age_3 &= min(max(age-120, 0), 3) \\
age_4 &= max(0, age-123) \\
Age_function &= age_1\beta_1 + age_2\beta_2 + age_3\beta_3 + age_4\beta_4
\end{aligned}$$

where age is short for mortgage age in the formulas above.

Debt Service Coverage Ratio (DSCR) is another crucial variable in modeling mortgage default. It is defined as the ratio between net operating income and current debt obligations. DSCR larger than one indicates borrowers have enough cash flow to make monthly payments. Otherwise, borrowers may default. We use the original value of DSCR without any transformation.

Besides variables describing idiosyncratic risk of mortgages, we have another set of variables to approximate systematic risk. The first variable we use is the lagged value of commercial mortgage default rates, denoted as lagged_PD. Since our models divide mortgage data into several groups, we compute lagged_PD for each group. Instead of using the original value, a logit transformation is applied to lagged_PD here, because logistic regression is utilized in our models. The transformed variable is defined in the following way:

$$lagged\_PD' = log((lagged\_PD)/(1 - lagged\_PD))$$

An empirical analysis strongly favors this transformed value over the original value, so we use $lagged\_PD'$ instead of the original values.

The lagged values of residential default rates are used to explain the systematic risk of commercial mortgages as well. To investigate the spatial dependence between commercial default and residential default, the residential PD is matched with commercial data in the same MSA area. We consider four lag orders: one month, three months, six months and twelve months, and transform the lagged residential PD with logit function to improve empirical fits.

The second set of variables we use for systematic risk are frailty factors, which capture the unexplained part of systematic risk. We develop three formulations for frailty factors. These formulations differ in the criterion to group the mortgages, and therefore, in the number of frailty factors. Since we use data of monthly frequency, all of these frailty factors are updated every month. The first formulation is denoted "single frailty". In this formulation, all mortgages belong to one group, which indicates a common factor for all mortgages. The second formulation is called "spatial frailty', in which we divide mortgage data into several groups by their locations. Accordingly, we have a separate frailty factor for each spatial group. The third formulation is called "origination frailty", and group mortgages by their origination time, yielding a multi-factor formulation.

3.4.3 Variables for Residential Mortgages

To model the default risk of residential mortgages, we introduce another set of variables. For the part of idiosyncratic risk, we employ debt to income ratios, flag of first time home buyers, loan purposes, and occupancy status to model risk. For the part of systematic risk, lagged PD of residential default and frailty factors are used.

Debt to income ratio is extensively used in the literature to model the idiosyncratic risk of borrowers. It is defined as the ratio between the monthly payment of mortgage and the borrower's income. If DTI is larger than one, then it implies that the borrower cannot cover the monthly payment with his income. DTI measures the financial capacity of borrowers and is similar to DSCR for commercial mortgages. A notable difference between DTI and DSCR is the frequency of data availability: DTI is only measured at the origination time of mortgages, and DSCR is updated dynamically based on the most recent financial statements of borrowers. Therefore, DTI is a static variable and works as a fixed intercept actually. By contrast, DSCR is a dynamic variable and reflects the current financial status of borrowers.

Flag of first time home buyer is a dummy variable, with one indicating a first time home buyer and 0 not. Potentially, a first time home buyer may have higher default risk than other borrowers, because they do not have enough savings as others. Similarly, loan purpose is also a dummy variable. This variable is 0 if the borrower use the money to buy a house, and 1 for refinance purposes. Occupancy status reflects how borrowers use the house and has three possible values: occupied, second home and investment. Due to the three categories, this variable is modeled as two dummy variables. One dummy variable indicates whether the house is a second home or not and another one determines if the house is an investment. Accordingly, "occupied" served as a base case and is absorbed in the intercept.

Comparing with the categorical variables mentioned above, FICO scores provide a continuous measure for the financial situation of borrowers. While most researchers use FICO scores as a static variable to reflect the baseline risk, we relate FICO scores to frailty factors and track the time-varying systematic risk. Specifically, we group borrowers by their FICO scores and allow borrowers in different groups to have separate latent factors. We denote these frailty factors as "FICO" frailty. Additionally, single frailty, which assumes only one latent factor for all mortgages, is also introduced to compare with the FICO frailty.

3.4.4 Estimation

For commercial mortgages, we build six models to examine the influence of the variables specified in Section 3.4.2 on mortgage default rates. Table 3.3 lists the variables used in each model. The first two models, Static I and II, only contain static variables, such as $Age\_function$ and $DSCR$. The next four models are dynamic models with both static variables and various frailty factors. In Table 3.3 the models appear from specific to general, i.e. each model is nested in the model on its right. For example, Static I is nested in Static II, because the former uses all variables of the latter except lagged residential PD. Likewise, Dynamic IX nests Dynamic II and Dynamic III, because the last two models only contain one of two frailty factors in Dynamic IX.

For residential mortgages, three models are developed to investigate the influence of various risk factors. Table 3.4 details the specification of each model. The first model, Static I, only contains static variables. The other two models are dynamic models, containing both static variables and frailty factors. Note that these models are nested from left to right, similar to the specifications for commercial mortgages in Table 3.3.

For dynamic models, grouping of mortgages is a key step in constructing frailty factors. For "single frailty", no grouping is needed. However, when using "origination frailty" and "spatial frailty", we need to carefully consider the groupings. On the one hand, a small group size is desired to ensure the similarity in credit quality among mortgages. On the other hand, a group that is too small may produce imprecise estimates, since mortgage default is a rare event. The group size should be large enough to produce smooth estimates of empirical default rates, which are inputs to our frailty factors. Therefore, to pick a proper group size, we have to strike a balance between controlling biases and producing smooth estimates.

While grouping mortgages spatially, there are various delineations available. Popular choices are census regions, census divisions, states, metropolitan statistical areas, and counties[4]. Among these choices, we select MSA as the grouping criteria based on the considerations discussed in the last paragraph. First, using census regions/divisions as the criterion may group areas with different default trends together. For example, using census divisions would group the Boston area and New York area in the same Northeast division, implying that these two areas share the same frailty factor. However, the real estate market in these two areas exhibited completely different trend in history. Second, using smaller delineations, like states, may divide one cross-state market into two parts, which is not desirable. Take the area of New York City and Jersey City as an example. This area will be split into two separate groups by their states, while the markets in these two cities are highly correlated. Compared with criteria mentioned above, MSA delineation has obvious advantages. While delineating MSA areas, social-economic ties between areas are the top consideration. Using MSA as the grouping criterion, neither will we group unrelated areas together, like Boston and New York, nor will we break closely a cross-state market, like New York City and Jersey City. Moreover, the population requirement for MSA areas ensure that there are enough samples in one group to make inference for frailty factors. On the contrary, county based groupings are often too small to make reliable inference. Hence, we choose MSA as our grouping criterion for spatial frailty.

While using MSA as the grouping criterion, we also try to balance sample size and homogeneous sample. There are over 900 MSA areas, and a lot of them do not have enough sample to generate smooth estimates of PD. To find MSA areas with enough samples, we rank all the MSA areas by their numbers of distinct commercial mortgages in a descending order. Then we allow the mortgages in each of the top 10 MSA areas to have their own group. For the rest of MSA areas, we use one group to keep the model

[4] For further information on these delineations, please refer to the website of U.S. Census Bureau: https://www.census.gov

parsimonious. In summary, we have 11 spatial groups: 10 groups for the top 10 MSA areas and 1 group for all other areas.

For origination frailty, we choose two years as our window size to group mortgages. Ideally, mortgages originated in the same month should form a group, because mortgages initiated in the same month have less variation in credit quality than mortgages initiated in a relatively long period. However, the number of mortgages initiated in one month is too small to produce a smooth estimate of default rates. The two-year window size is chosen based on empirical tests. By comparing a number of alternative window sizes from six months to four years, we find that the window size of six months produces non-smooth default curves, and the window size of four years groups mortgages of different qualities together. The two-year window size appears to strike a balance as it not only controls the variation in mortgage quality but also keeps estimates of default rates smooth.

Unlike the limited sample size of commercial mortgages, residential mortgages have much larger sample size. Table 3.2 shows the large volume of residential mortgages in the top 10 MSA areas, and Figure 3.1 further demonstrates that these MSA areas have enough sample size to produce smooth estimates of empirical PD. In light of the large sample size, we decide to estimate the models of residential mortgages separately for each of the 10 MSA areas, rather than using a joint estimation that we do for commercial mortgages. For mortgages in each MSA areas, we further classify them into two groups by their FICO scores. One group contains mortgages with FICO scores less than 640, which is generally considered as subprime mortgages. The other group consists of all the other mortgages(prime mortgages). We assume a separate latent factor for each group and hence have two frailty factors for each MSA area.

We assume conditional independence for dependent variables in all the models and use standard maximum likelihood methods to estimate parameters.

3.4.5 Results

Table 3.5 reports the estimation results for commercial mortgages models. All parameters are highly significant and have the expected signs. As anticipated, our results provide evidence that the effect of DSCR on mortgage default is negative as expected, since borrowers with higher income are less likely to default. The signs of Age 1, Age 2, Age 3, and Age 4 are positive, negative, positive and negative respectively, and this sign pattern is consistent with the double default clustering observed in previous empirical analysis.

Moreover, the significant estimates of lagged residential PD demonstrate that residential default is a good predictors for commercial default, and provide evidence of close correlation between these two markets. These parameters are not only significant in static models, but also significant in dynamic models, even after adding various frailty factors. This conveys that residential PD indeed contains information that is not in the commercial market, and indicates the intuition that residential market leads commercial market.

Among these four lagged residential PD, 3-month lagged PD seems to have the most explanatory power for commercial default, because it has the only positive sign and the largest absolute value. By contrast, 12 month lagged PD has the smallest absolute value and its value decreases to -0.02 in Dynamic IX, including both spatial and origination frailty. This small value may suggest that 12 month is too long to use residential PD to predict commercial PD.

The last three rows present statistics measuring model performance. Likelihood increases from the left to right, indicating that the introduction of residential PD and frailty factors has improved the fitting performance. The difference among likelihood of nested models are also significant when we perform likelihood ration tests. AIC and BIC decrease from left to right, which demonstrates it is worth including residential PD and frailty factors even after penalizing for extra model complexity.

Table 3.6 to 3.15 document the estimation results for the models of residential mortgages. Most of the parameters are highly significant and have the expected signs. The signs of DTI are positive, which is consistent with the motivation that borrowers with less income are of high risk. The signs of flag of first time home buyers are mixed. In static models, they are positive across all the MSA areas, while in dynamic models they convert to negative. The positive signs of occupancy status convey that if borrowers have enough money to purchase houses as their second homes or investments, they are less likely to default.

Likelihood, AIC and BIC are listed in the last three rows. From left to right, Likelihood increases while BIC and AIC decreases. This demonstrates consistent improvements of model performance in using frailty factors. Moreover, the superior performance of Dynamic II over Dynamic I illustrates the advantage of separating borrowers into prime borrowers and subprime borrowers.

Figure 3.1 shows the curves of empirical PD of mortgages in the 10 MSA areas. In each grid, a red curve represents the commercial PD, a green curve shows the residential PD of prime borrowers and a blue curve displays the residential PD of subprime borrowers. We can observe a clear pattern of lead and lag relationship among many of these MSA areas. Take Los Angeles in the top left corner as an example: the blue curve of subprime mortgages rises at first during financial crisis, and then the green curve of prime mortgages follows, after roughly 3 or 4 months of steady growth by the blue and green curves, the red curve representing commercial default finally starts surging. This lead-lag pattern clearly illustrates the phenomenon that the subprime market leads the prime market and the residential market leads the commercial market in the Los Angeles area while the market worsen. Similar lead-lag relationship also happens while the market recovers: the blue and green curves begin decreasing more than 1 year earlier than the red curve.

The Los Angeles area is not the only example demonstrating the lead-lag relationship between two local real estate markets. Similar patterns can be observed in most other areas except Anaheim and New York. For these two exceptions, limited and biased sample may prevent us from relating residential and commercial markets together. For Anaheim, there are only 289 commercial mortgages in the sample, and this sample size may be too small to estimate a reliable empirical PD, which can explain the highly volatile curves observed in the bottom left corner. Regarding the New York area, a possibly biased residential sample may obscure our observation: the subprime sample only started from 2005, which implies a low participation rate of Freddie Mac in this local market.

To examine whether the lead-lag relationship shown in Figure 3.1 is statistically significant, we perform a Granger test between the commercial PD and residential PD in the same area. Specifically, we use residential PD as the explanatory variables, and test the causal relationships with 4 lag orders: 3 months, 6 months, 12 months and 24 months. Since these time series of PD are non-stationary, we perform tests on the first order differences. The test results are detailed in Table 3.16. In particular, the upper panel lists the p-values of tests between commercial default and prime default, and the lower one reports the p-values between commercial default and subprime default. P-values in both panels further confirm the lead-lag relationship. For the causal relationship between commercial default and prime default, 9 out of 10 areas exhibit significant causal relations with 5 percent rejection level. The only outlier is Anaheim, which may suffer from a limited sample size problem as discussed above. Among these 9 significant areas, 5 of them show strong causal relationships for all four lag orders. Moreover, 8 out of 10 areas show strong causal relationships between subprime default and commercial default for at least one lag order, and 4 of these areas are significant for at least three lag orders. The only two areas demonstrate no significant correlation are New York and Anaheim, and the possible causes have been discussed in the preceding paragraph.

To further explore the spatial dependence between commercial default and residential default, we proceed to test the causal relations among the fitted frailty of dynamic models. Specifically, we test the commercial frailty factors of Dynamic III and the residential frailty factors of Dynamic II. Comparing with empirical PD, frailty factors are free from idiosyncratic risk that has been explained by the variables in the model. In this sense, tests on frailty factor reveal more accurate information about the spatial dependence of two real estate markets, since information unrelated to systematic risk is extracted.

Figure 3.2 shows the curves of commercial frailty and residential frailty in the 10 MSA areas. In each grid, there are three curves: the red for commercial frailty, the blue for subprime frailty, and the green for prime frailty. Note that these curves are not in the same scale of PD due to the logistic transformation applied in modeling. Similar to the PD plots, these frailty curves exhibit lead-lag relationships: the blue and green curves representing residential frailty rose earlier than the red curve in the financial crisis, and declined before the red curve when the market rebounded. Hence, these plots illustrate the close association between the two real estate markets.

Besides the visual examinations of frailty curves, we perform Granger tests on the frailty factors. Similar to the tests on PD, we use residential frailty as explanatory variables, and test the causal relation on 4 lag orders: 3 months, 6 months, 12 months and 24 months. We perform tests on the first order difference of the frailty factors due to non-stationarity. When interpreting the p-values, we adopt 0.05 as the threshold for significance. Table 3.17 lists the test results on frailty factors, including an upper panel and a lower panel. While the upper one consists of test results between commercial frailty and prime frailty, the lower panel contains the test p-values between commercial frailty and subprime frailty. For the commercial-prime causal relation, 8 out of 10 areas exhibit significant causal relations, and 6 of the significant areas shows strong causal relations with at least three lag orders. The two outlier with insignificant results are again New York and Anaheim. With regard to the commercial-subprime causal tests, 8 out of 10 areas exhibit significant causal relations, and 4 of the significant areas show strong causal relations with all four lag orders. The only two areas show no significant relations are Houston and Dallas. However, when we further inspect the p-values for these areas, we find that the smallest ones are 0.06 for both areas. These values are fairly close to the significant level of 0.05 and are actually significant with a rejection level of 0.10. In conclusion, the test results with frailty factors provide strong evidence for the causal relationship between commercial and residential markets.

3.5 Conclusion

To the best of our knowledge, we are the first to investigate the spatial dependence between commercial and residential mortgage default. We utilize numerous methods to test the causal relationship of these two mortgage markets and obtain positive results. Using Granger causality tests, we demonstrate the lead-lag relationship between residential PD and commercial PD in almost all of the 10 main MSA areas. Moreover, we further perform tests on the fitted frailty using the dynamic models of mortgage risk, and affirmed the significant causal relationship of these two markets. In addition, it provides strong evidence for the causal relation that lagged residential PD offers significant explanatory power in modeling commercial mortgages default. Our model formulation also exhibit evident computational advantages in fitting complex models with big datasets, since all estimations finish in minutes.

Table 3.1: The Summary Statistics of Commercial Mortgages in the Top 10 MSA Areas

MSA Area	MSA Code	Mortgages Count	Percent of Count	Record Count	Percent of Records
Los Angeles-Long Beach-Glendale, CA	31084	878	0.053	85,927	0.054
New York-Jersey City-White Plains, NY-NJ	35614	551	0.033	51,952	0.032
Houston-The Woodlands-Sugar Land, TX	26420	530	0.032	49,494	0.031
Atlanta-Sandy Springs-Roswell, GA	12060	527	0.032	46,955	0.029
Phoenix-Mesa-Scottsdale, AZ	38060	473	0.028	46,162	0.029
Dallas-Plano-Irving, TX	19124	426	0.025	40,526	0.025
Riverside-San Bernardino-Ontario, CA	40140	350	0.021	33,378	0.021
Chicago-Naperville-Arlington Heights, IL	16974	346	0.021	33,234	0.021
Anaheim-Santa Ana-Irvine, CA	11244	289	0.017	28,350	0.018
Total	-	4,370	0.26	415,978	0.26

Notes: This table reports the summary statistics of commercial mortgages in the top 10 MSA areas.

Table 3.2: The Summary Statistics of Residential Mortgages in the Top 10 MSA Areas

MSA Area	MSA Code	Mortgages Count	Percent of Count	Record Count	Percent of Records
Los Angeles-Long Beach-Glendale, CA	31084	529,556	0.024	20,289,045	0.019
New York-Jersey City-White Plains, NY-NJ	35614	148,806	0.007	5,094,835	0.005
Houston-The Woodlands-Sugar Land, TX	26420	318,909	0.014	16,000,693	0.015
Atlanta-Sandy Springs-Roswell, GA	12060	417,504	0.019	21,243,365	0.020
Phoenix-Mesa-Scottsdale, AZ	38060	397,192	0.018	17,372,578	0.017
Dallas-Plano-Irving, TX	19124	280,734	0.012	13,172,318	0.013
Riverside-San Bernardino-Ontario, CA	40140	286,448	0.013	12,240,812	0.012
Chicago-Naperville-Arlington Heights, IL	16974	748,038	0.033	31,284,067	0.030
Anaheim-Santa Ana-Irvine, CA	11244	85,293	0.004	2,298,556	0.002
Total	-	3,212,480	0.14	138,996,269	0.13

Notes: This table reports the summary statistics of residential mortgages in the top 10 MSA areas.

Table 3.3: Components of Static and Dynamic Models for Commercial Mortgages

Variable	Static I	Static II	Dynamic I	Dynamic II	Dynamic III	Dynamic IX
DSCR	X	X	X	X	X	X
Age 1	X	X	X	X	X	X
Age 2	X	X	X	X	X	X
Age 3	X	X	X	X	X	X
Age 4	X	X	X	X	X	X
Commercial Lagged PD	X	X	X	X	X	X
1 Month Lagged Residential PD		X	X	X	X	X
3 Month Lagged Residential PD		X	X	X	X	X
6 Month Lagged Residential PD		X	X	X	X	X
12 Month Lagged Residential PD		X	X	X	X	X
Singe Frailty			X			
Origination Frailty				X		X
Spatial Frailty					X	X
Log Likelihood	X	X	X	X	X	X
AIC	X	X	X	X	X	X
BIC	X	X	X	X	X	X

Notes: This table describes components of each models for commercial mortgages in the empirical applications. X means the variable is a part of the model.

Table 3.4: Components of Static and Dynamic Models for Residential Mortgages

Variable	Static I	Dynamic I	Dynamic II
DTI	X	X	X
Flag of First Time Home Buyer	X	X	X
Loan Purpose (Non_purchase)	X	X	X
Occupancy Status (Second Home)	X	X	X
Occupancy Status (Investment)	X	X	X
Single Frailty		X	
FICO Frailty			X

Notes: This table describes the components of each models for residential mortgages in the empirical applications. X means the variable is a part of the model.

Table 3.5: Estimates of the Static and Dynamic Models for Commercial Mortgages

Variable	Static I	Static II	Dynamic I	Dynamic II	Dynamic III	Dynamic IX
DSCR	-1.28	-1.28	-1.30	-1.34	-1.31	-1.34
Age 1	0.03	0.03	0.03	0.01	0.03	0.01
Age 2	-0.05	-0.05	-0.04	-0.03	-0.04	-0.03
Age 3	0.40	0.40	0.40	0.40	0.40	0.41
Age 4	-0.10	-0.10	-0.09	-0.13	-0.09	-0.13
Commercial Lagged PD	0.75	0.81	0.96	0.85	0.62	0.63
1 Month Lagged Residential PD	-	-0.20	-0.44	-0.52	-0.45	-0.41
3 Month Lagged Residential PD	-	0.47	0.57	0.71	0.60	0.47
6 Month Lagged Residential PD	-	-0.19	-0.13	-0.24	-0.18	-0.16
12 Month Lagged Residential PD	-	-0.14	0.05	-0.08	-0.05	-0.02
θ_1	-	-	0.82	0.96	0.96	0.97
α_1	-	-	40.97	9.38	22.23	9.16
θ_2	-	-	-	-	-	0.98
α_2	-	-	-	-	-	18.79
Log Likelihood	-96803	-96690	-96541	-94338	-94127	-93972
AIC	193620	193402	193107	188701	188278	187971
BIC	193706	193537	193254	188848	188425	188143

Notes: This table reports the estimates of the static and dynamic Models for commercial mortgages.

Table 3.6: Estimates of Static and Dynamic Models in the Los Angeles-Long Beach-Glendale Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-6.06	-	-
DTI	0.04	0.04	0.04
Flag of First Time Home Buyer	0.08	-0.17	-0.16
Loan Purpose (Non purchase)	0.15	0	-0.1
Occupancy Status (Second Home)	-0.66	-0.84	-0.81
Occupancy Status (Investment)	-0.56	-0.67	-0.62
θ_1	-	0.997	0.997
α_1	-	65	20
Log Likelihood	-1301891	-1181395	-1164016
AIC	2603795	2362805	2328046
BIC	2603884	2362909	2328150

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the Los Angeles-Long Beach-Glendale area.

Table 3.7: Estimates of Static and Dynamic Models in the New York-Jersey City-White Plains Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-7.13	-	-
DTI	0.04	0.04	0.03
Flag of First Time Home Buyer	0.01	0.01	-0.03
Loan Purpose (Non purchase)	0.65	0.46	0.38
Occupancy Status (Second Home)	-0.75	-0.8	-0.64
Occupancy Status (Investment)	-0.54	-0.47	-0.4
θ_1	-	0.996	0.996
α_1	-	165	165
Log Likelihood	-167287	-162773	-154864
AIC	334587	325561	309743
BIC	334667	325655	309837

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the New York-Jersey City-White Plains area.

Table 3.8: Estimates of Static and Dynamic Models in the Houston-The Woodlands-Sugar Land Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-5.72	-	-
DTI	0.03	0.03	0.02
Flag of First Time Home Buyer	0.15	0.11	0.06
Loan Purpose (Non purchase)	0.03	-0.07	-0.11
Occupancy Status (Second Home)	0.18	0.1	0.27
Occupancy Status (Investment)	-0.11	-0.19	-0.02
θ_1	-	0.995	0.995
α_1	-	129	129
Log Likelihood	-821494	-805630	-764713
AIC	1643000	1611274	1529440
BIC	1643087	1611376	1529542

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the Houston-The Woodlands-Sugar Land area.

Table 3.9: Estimates of Static and Dynamic Models in the Atlanta-Sandy Springs-Roswell Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-5.36	-	-
DTI	0.03	0.03	0.03
Flag of First Time Home Buyer	0.28	0.21	0.2
Loan Purpose (Non purchase)	0.18	0.07	0.06
Occupancy Status (Second Home)	0.39	0.2	0.22
Occupancy Status (Investment)	0.4	0.34	0.46
θ_1	-	0.995	0.995
α_1	-	62	62
Log Likelihood	-1731684	-1640351	-1594627
AIC	3463380	3280716	3189269
BIC	3463469	3280820	3189373

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the Atlanta-Sandy Springs-Roswell area.

Table 3.10: Estimates of Static and Dynamic Models in the Phoenix-Mesa-Scottsdale Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-5.57	-	-
DTI	0.04	0.04	0.04
Flag of First Time Home Buyer	0.22	0.11	0.1
Loan Purpose (Non purchase)	0.3	0.17	0.13
Occupancy Status (Second Home)	-0.73	-0.96	-0.89
Occupancy Status (Investment)	-0.36	-0.53	-0.45
θ_1	-	0.995	0.995
α_1	-	45	45
Log Likelihood	-1525105	-1352255	-1336063
AIC	3050222	2704524	2672140
BIC	3050310	2704627	2672242

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the Phoenix-Mesa-Scottsdale area.

Table 3.11: Estimates of Static and Dynamic Models in the Dallas-Plano-Irving Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-5.87	-	-
DTI	0.03	0.03	0.03
Flag of First Time Home Buyer	0.09	0.04	0.03
Loan Purpose (Non purchase)	0.09	-0.04	-0.07
Occupancy Status (Second Home)	0.15	0.03	0.15
Occupancy Status (Investment)	-0.25	-0.35	-0.17
θ_1	-	0.994	0.994
α_1	-	120	120
Log Likelihood	-643649	-628828	-602833
AIC	1287309	1257670	1205679
BIC	1287396	1257771	1205780

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the Dallas-Plano-Irving area

Table 3.12: Estimates of Static and Dynamic Models in the Riverside-San Bernardino-Ontario Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-5.27	-	-
DTI	0.04	0.03	0.03
Flag of First Time Home Buyer	0.17	-0.02	-0.02
Loan Purpose (Non purchase)	0.34	0.26	0.2
Occupancy Status (Second Home)	-0.81	-0.99	-0.92
Occupancy Status (Investment)	-0.7	-0.89	-0.82
θ_1	-	0.996	0.996
α_1	-	34	34
Log Likelihood	-1307491	-1151182	-1137759
AIC	2614994	2302378	2275532
BIC	2615080	2302478	2275632

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the Riverside-San Bernardino-Ontario area.

Table 3.13: Estimates of Static and Dynamic Models in the Chicago-Naperville-Arlington Heights Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-5.18	-	-
DTI	0.04	0.03	0.03
Flag of First Time Home Buyer	0.12	-0.1	-0.08
Loan Purpose (Non purchase)	0.11	-0.13	-0.14
Occupancy Status (Second Home)	-0.42	-0.67	-0.6
Occupancy Status (Investment)	0.25	0.11	0.19
θ_1	-	0.994	0.994
α_1	-	44	44
Log Likelihood	-3262711	-3009556	-2926986
AIC	6525435	6019125	5853987
BIC	6525526	6019232	5854093

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the Chicago-Naperville-Arlington Heights area

Table 3.14: Estimates of Static and Dynamic Models in the Anaheim-Santa Ana-Irvine Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-10.53	-	-
DTI	0.07	0.07	0.08
Flag of First Time Home Buyer	1.11	1.23	2.32
Loan Purpose (Non purchase)	0.84	0.38	1.75
Occupancy Status (Second Home)	0.68	0.59	0.91
Occupancy Status (Investment)	-0.96	-0.71	-0.58
θ_1	-	0.999	0.999
α_1	-	335	335
Log Likelihood	-16029	-14767	-14979
AIC	32071	29548	29971
BIC	32147	29637	30060

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the Anaheim-Santa Ana-Irvine Area.

Table 3.15: Estimates of Static and Dynamic Models in the Washington-Arlington-Alexandria Area

Variable	Static I	Dynamic I	Dynamic II
Intercept	-6.91	-	-
DTI	0.05	0.05	0.05
Flag of First Time Home Buyer	0.2	0.06	0.06
Loan Purpose (Non purchase)	0.56	0.42	0.26
Occupancy Status (Second Home)	-0.48	-0.61	-0.49
Occupancy Status (Investment)	-0.25	-0.38	-0.27
θ_1	-	0.998	0.997
α_1	-	92	21
Log Likelihood	-1018039	-947005	-906952
AIC	2036090	1894024	1813919
BIC	2036178	1894127	1814022

Notes: This table reports the estimates for static and dynamic models of residential mortgages in the Washington-Arlington-Alexandria area.

Table 3.16: Granger Causality Tests between Commercial PD and Residential PD in Main MSA Areas

MSA Area	Borrowers	3 Month	6 Month	12 Month	24 Month
Los Angeles-Long Beach-Glendale, CA	Prime Borrowers	$<$0.01	$<$0.01	0.01	$<$0.01
New York-Jersey City-White Plains, NY-NJ	Prime Borrowers	0.77	0.49	0.56	$<$0.01
Houston-The Woodlands-Sugar Land, TX	Prime Borrowers	0.43	0.04	0.06	0.04
Atlanta-Sandy Springs-Roswell, GA	Prime Borrowers	$<$0.01	$<$0.01	$<$0.01	$<$0.01
Phoenix-Mesa-Scottsdale, AZ	Prime Borrowers	$<$0.01	$<$0.01	$<$0.01	$<$0.01
Dallas-Plano-Irving, TX	Prime Borrowers	$<$0.01	0.01	0.02	0.02
Riverside-San Bernardino-Ontario, CA	Prime Borrowers	$<$0.01	$<$0.01	$<$0.01	$<$0.01
Chicago-Naperville-Arlington Heights, IL	Prime Borrowers	0.11	0.22	0.01	0.21
Anaheim-Santa Ana-Irvine, CA	Prime Borrowers	0.09	0.25	0.26	0.44
Washington-Arlington-Alexandria, DC-VA-MD-WV	Prime Borrowers	0.06	0.18	0.13	$<$0.01
Los Angeles-Long Beach-Glendale, CA	Subprime Borrowers	0.04	0.03	0.07	$<$0.01
New York-Jersey City-White Plains, NY-NJ	Subprime Borrowers	0.88	0.49	0.4	0.08
Houston-The Woodlands-Sugar Land, TX	Subprime Borrowers	0.71	0.1	0.03	0.07
Atlanta-Sandy Springs-Roswell, GA	Subprime Borrowers	$<$0.01	$<$0.01	$<$0.01	$<$0.01
Phoenix-Mesa-Scottsdale, AZ	Subprime Borrowers	$<$0.01	$<$0.01	$<$0.01	$<$0.01
Dallas-Plano-Irving, TX	Subprime Borrowers	$<$0.01	$<$0.01	$<$0.01	0.08
Riverside-San Bernardino-Ontario, CA	Subprime Borrowers	$<$0.01	$<$0.01	$<$0.01	$<$0.01
Chicago-Naperville-Arlington Heights, IL	Subprime Borrowers	0.18	0.19	$<$0.01	0.04
Anaheim-Santa Ana-Irvine, CA	Subprime Borrowers	0.36	0.73	0.85	0.08
Washington-Arlington-Alexandria, DC-VA-MD-WV	Subprime Borrowers	0.73	0.79	0.09	$<$0.01

Notes: This table reports the p-values of Granger causality tests between commercial PD and residential PD in main MSA areas.

Table 3.17: Granger Causality Tests between Commercial Frailty and Residential Frailty in Main MSA Areas

MSA Area	Borrowers	3 Month	6 Month	12 Month	24 Month
Los Angeles-Long Beach-Glendale, CA	Prime Borrowers	<0.01	<0.01	<0.01	<0.01
New York-Jersey City-White Plains, NY-NJ	Prime Borrowers	0.68	0.9	0.99	1
Houston-The Woodlands-Sugar Land, TX	Prime Borrowers	0.28	0.02	0.03	0.2
Atlanta-Sandy Springs-Roswell, GA	Prime Borrowers	<0.01	<0.01	<0.01	<0.01
Phoenix-Mesa-Scottsdale, AZ	Prime Borrowers	0.07	<0.01	<0.01	<0.01
Dallas-Plano-Irving, TX	Prime Borrowers	<0.01	<0.01	<0.01	0.16
Riverside-San Bernardino-Ontario, CA	Prime Borrowers	<0.01	<0.01	<0.01	<0.01
Chicago-Naperville-Arlington Heights, IL	Prime Borrowers	<0.01	<0.01	<0.01	<0.01
Anaheim-Santa Ana-Irvine, CA	Prime Borrowers	0.34	0.46	0.29	0.25
Washington-Arlington-Alexandria, DC-VA-MD-WV	Prime Borrowers	0.09	0.05	<0.01	<0.01
Los Angeles-Long Beach-Glendale, CA	Subprime Borrowers	<0.01	<0.01	<0.01	<0.01
New York-Jersey City-White Plains, NY-NJ	Subprime Borrowers	0.64	0.32	0.4	0.04
Houston-The Woodlands-Sugar Land, TX	Subprime Borrowers	0.1	0.06	0.12	0.1
Atlanta-Sandy Springs-Roswell, GA	Subprime Borrowers	<0.01	<0.01	<0.01	<0.01
Phoenix-Mesa-Scottsdale, AZ	Subprime Borrowers	<0.01	<0.01	<0.01	<0.01
Dallas-Plano-Irving, TX	Subprime Borrowers	0.06	0.31	0.46	0.06
Riverside-San Bernardino-Ontario, CA	Subprime Borrowers	<0.01	<0.01	<0.01	<0.01
Chicago-Naperville-Arlington Heights, IL	Subprime Borrowers	0.04	0.05	<0.01	0.02
Anaheim-Santa Ana-Irvine, CA	Subprime Borrowers	0.01	0.07	0.14	0.33
Washington-Arlington-Alexandria, DC-VA-MD-WV	Subprime Borrowers	0.2	0.23	<0.01	<0.01

Notes: This table reports the p-values of Granger causality tests between commercial frailty and residential frailty in main MSA areas.

Figure 3.1: Empirical Default Rates of Commercial and Residential Mortgages in the Top 10 MSA Areas

Notes: The picture shows empirical default rates of mortgages in the top 10 MSA areas. The horizontal axis is date and the vertical axis is default rates.

Figure 3.2: Frailty Factors of Commercial and Residential Mortgages in the Top 10 MSA Areas

Notes: The picture shows frailty factors of mortgages in the top 10 MSA areas. The horizontal axis is date and the vertical axis is frailty factor.

BIBLIOGRAPHY

Altman, Edward I. 1968. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." *Journal of Finance*, 23(4): 589–609.

Altman, Edward I, Brooks Brady, Andrea Resti, and Andrea Sironi. 2005. "The link between default and recovery rates: Theory, empirical evidence, and implications." *Journal of Business*, 78(6): 2203.

An, Xudong, Yongheng Deng, Joseph B Nichols, and Anthony B Sanders. 2013. "Local traits and securitized commercial mortgage default." *The Journal of Real Estate Finance and Economics*, 47(4): 787–813.

Beaver, William H. 1968. "Alternative accounting measures as predictors of failure." *Accounting Review*, 43(1): 113–122.

Benjamin, Michael A, Robert A Rigby, and D Mikis Stasinopoulos. 2003. "Generalized autoregressive moving average models." *Journal of the American Statistical Association*, 98(461): 214–223.

Black, Fischer, and John C Cox. 1976. "Valuing corporate securities: Some effects of bond indenture provisions." *Journal of Finance*, 31(2): 351–367.

Black, Fischer, and Myron Scholes. 1973. "The pricing of options and corporate liabilities." *The Journal of Political Economy*, 81(3): 637–654.

Bollerslev, Tim. 1987. "A conditionally heteroskedastic time series model for speculative prices and rates of return." *The Review of Economics and Statistics*, 69(3): 542–547.

Campbell, John Y, Jens Hilscher, and Jan Szilagyi. 2008. "In search of distress risk." *The Journal of finance*, 63(6): 2899–2939.

Chen, Xi, Eric Ghysels, and Roland Telfeyan. 2016. "Frailty Models for Commercial Mortgages." *Journal of Fixed Income*, 26: 16–31.

Christoffersen, Peter, Vihang Errunza, Kris Jacobs, and Hugues Langlois. 2012. "Is the potential for international diversification disappearing? A dynamic copula approach." *Review of Financial Studies*, 25(12): 3711–3751.

Colacito, Riccardo, Robert F Engle, and Eric Ghysels. 2011. "A component model for dynamic correlations." *Journal of Econometrics*, 164(1): 45–59.

Corsi, Fulvio. 2009. "A simple approximate long-memory model of realized volatility." *Journal of Financial Econometrics*, 7(2): 174–196.

Cox, David R, Gudmundur Gudmundsson, Georg Lindgren, Lennart Bondesson, Erik Harsaae, Petter Laake, Katarina Juselius, and Steffen L Lauritzen. 1981. "Statistical analysis of time series: some recent developments [with discussion and reply]." *Scandinavian Journal of Statistics*, 93–115.

Creal, Drew, Bernd Schwaab, Siem Jan Koopman, and Andr Lucas. 2014. "Observation-Driven mixed-measurement dynamic factor models with an application to credit risk." *Review of Economics and Statistics*, 96(5): 898–915.

Creal, Drew, Siem Jan Koopman, and André Lucas. 2013*a*. "Generalized autoregressive score models with applications." *Journal of Applied Econometrics*, 28(5): 777–795.

Creal, Drew, Siem Jan Koopman, and Andr Lucas. 2013*b*. "Generalized autoregressive score models with applications." *Journal of Applied Econometrics*, 28(5): 777–795.

Crosbie, Peter, and Jeff Bohn. 2003. "Modeling default risk."

Das, Sanjiv R, Darrell Duffie, Nikunj Kapadia, and Leandro Saita. 2007. "Common failings: How corporate defaults are correlated." *Journal of Finance*, 62(1): 93–117.

Deng, Yongheng, Andrey D Pavlov, and Lihong Yang. 2005. "Spatial heterogeneity in mortgage terminations by refinance, sale and default." *Real Estate Economics*, 33(4): 739–764.

Deng, Yongheng, John M Quigley, and Robert Order. 2000. "Mortgage terminations, heterogeneity and the exercise of mortgage options." *Econometrica*, 68(2): 275–307.

Dias, Alexandra, and Paul Embrechts. 2002. "Change-point analysis for dependence structures in finance and insurance." *Novos Rumos em Estatstica(Ed. C. Carvalho, F. Brilhante and F. Rosado), Sociedade Portuguesa de Estatstica, Lisbon*, 69–86.

Duan, Jin-Chuan, and Andras Fulop. 2013. "Multiperiod corporate default prediction with the partially-conditioned forward intensity." *Available at SSRN 2151174*.

Duan, Jin-Chuan, Jie Sun, and Tao Wang. 2012. "Multiperiod corporate default prediction: A forward intensity approach." *Journal of Econometrics*, 170(1): 191–209.

Duffie, Darrell, Andreas Eckner, Guillaume Horel, and Leandro Saita. 2009. "Frailty correlated default." *Journal of Finance*, 64(5): 2089–2123.

Engle, Robert F, and Gary Lee. 1999. "A long-run and short-run component model of stock return volatility." *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive WJ Granger*, 475–497.

Engle, Robert F, and Jose Gonzalo Rangel. 2008. "The spline-GARCH model for low-frequency volatility and its global macroeconomic causes." *Review of Financial Studies*, 21(3): 1187–1222.

Engle, Robert F, Eric Ghysels, and Bumjean Sohn. 2013. "Stock market volatility and macroeconomic fundamentals." *Review of Economics and Statistics*, 95(3): 776–797.

Fengler, Matthias R, and Ostap Okhrin. 2012. "Realized copula." SFB 649 Discussion Paper.

Fischer, Edwin O, Robert Heinkel, and Josef Zechner. 1989. "Dynamic capital structure choice: Theory and tests." *Journal of Finance*, 44(1): 19–40.

Follain, James R, Jan Ondrich, and Gyan P Sinha. 1997. "Ruthless prepayment? Evidence from multifamily mortgages." *Journal of Urban Economics*, 41(1): 78–101.

Ghysels, Eric, Arthur Sinko, and Rossen Valkanov. 2007. "MIDAS regressions: Further results and new directions." *Econometric Reviews*, 26(1): 53–90.

Gouriéroux, Christian, Alain Monfort, Eric Renault, and Alain Trognon. 1987. "Generalised residuals." *Journal of Econometrics*, 34(1): 5–32.

Guegan, Dominique, and Jing Zhang. 2010. "Change analysis of a dynamic copula for measuring dependence in multivariate financial data." *Quantitative Finance*, 10(4): 421–430.

Hafner, Christian M, and Hans Manner. 2012. "Dynamic stochastic copula models: Estimation, inference and applications." *Journal of Applied Econometrics*, 27(2): 269–295.

Hafner, Christian M, and Olga Reznikova. 2010. "Efficient estimation of a semiparametric dynamic copula model." *Computational Statistics & Data Analysis*, 54(11): 2609–2627.

Han, Aaron, and Jerry Hausman. 1990. "Flexible parametric estimation of duration and competing risk models." *Journal of Applied Econometrics.*, 5(1): 1–28.

Hansen, Bruce E. 1994. "Autoregressive conditional density estimation." *International Economic Review*, 35(3): 705–730.

Harvey, Andrew C. 2013. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series.* Cambridge University Press.

Heinen, Andréas, and Alfonso Valdesogo. 2009. *Asymmetric CAPM dependence for large dimensions: The canonical vine autoregressive model.* CORE.

Hui, Cho-Hoi, Chi-Fai Lo, and Shun-Wai Tsang. 2003. "Pricing corporate bonds with dynamic default barriers." *Journal of Risk*, 5(3): 17–37.

Jarrow, Robert A, and Philip Protter. 2004. "Structural vs reduced form models: a new information based perspective." *Journal of investment management*, 2(2): 1–10.

Joe, Harry. 2014. *Dependence Modeling with Copulas.* CRC Press.

Jondeau, Eric, and Michael Rockinger. 2006. "The copula-garch model of conditional dependencies: An international stock market application." *Journal of International Money and Finance*, 25(5): 827–853.

Jones, E Philip, Scott P Mason, and Eric Rosenfeld. 1984. "Contingent claims analysis of corporate capital structures: An empirical investigation." *Journal of Finance*, 39(3): 611–625.

Kau, James B, Donald C Keenan, and Alexey A Smurov. 2004. "Reduced-form mortgage valuation." *University of Georgia, Georgia.*

Kau, James B, Donald C Keenan, and Alexey A Smurov. 2006. "Reduced form mortgage pricing as an alternative to option-pricing models." *The Journal of Real Estate Finance and Economics*, 33(3): 183–196.

Kau, James B, Donald C Keenan, and Xiaowei Li. 2011. "An analysis of mortgage termination risks: a shared frailty approach with MSA-level random effects." *The Journal of Real Estate Finance and Economics*, 42(1): 51–67.

Koopman, Siem Jan, and Andr Lucas. 2008. "A non-Gaussian panel time series model for estimating and decomposing default risk." *Journal of Business & Economic Statistics*, 26(4): 510–525.

Koopman, Siem Jan, Andr Lucas, and Bernd Schwaab. 2012. "Dynamic factor models with macro, frailty, and industry effects for US default counts: the credit crisis of 2008." *Journal of Business & Economic Statistics*, 30(4): 521–532.

Leland, Hayne E. 1994. "Corporate debt value, bond covenants, and optimal capital structure." *Journal of Finance*, 49(4): 1213–1252.

Longstaff, Francis A, and Eduardo S Schwartz. 1995. "A simple approach to valuing risky fixed and floating rate debt." *Journal of Finance*, 50(3): 789–819.

Manner, Hans, and Bertrand Candelon. 2010. "Testing for asset market linkages: A new approach based on time-varying copulas." *Pacific Economic Review*, 15(3): 364–384.

McCall, Brian P. 1996. "Unemployment insurance rules, joblessness, and part-time work." *Econometrica: Journal of the Econometric Society*, 64(3): 647–682.

Merton, Robert C. 1974. "On the pricing of corporate debt: The risk structure of interest rates*." *Journal of Finance*, 29(2): 449–470.

Nelsen, Roger B. 2007. *An introduction to copulas.* Springer Science & Business Media.

Nielsen, Lars Tyge, Jesus Saà-Requejo, and Pedro Santa-Clara. 2001. *Default risk and interest rate risk: The term structure of default spreads.* INSEAD.

Oh, Dong Hwan, and Andrew J Patton. 2013. "Time-varying systemic risk: Evidence from a dynamic copula model of cds spreads." *Economic Research Initiatives at Duke (ERID) Working Paper.*

Ohlson, James A. 1980. "Financial ratios and the probabilistic prediction of bankruptcy." *Journal of Accounting Research*, 18(1): 109–131.

Patton, Andrew J. 2004. "On the out-of-sample importance of skewness and asymmetric dependence for asset allocation." *Journal of Financial Econometrics*, 2(1): 130–168.

Patton, Andrew J. 2006. "Modelling asymmetric exchange rate dependence." *International Economic Review*, 47(2): 527–556.

Patton, Andrew J. 2012. "A review of copula models for economic time series." *Journal of Multivariate Analysis*, 110: 4–18.

Salvatierra, Irving De Lira, and Andrew J Patton. 2014. "Dynamic copula models and high frequency data." *Journal of Empirical Finance*, 30: 120–135.

Schwaab, Bernd, Siem Jan Koopman, and Andr Lucas. 2015. "Global credit risk: World, country and industry factors." Tinbergen Institute Discussion Paper Report.

Shephard, Neil. 1995. "Generalized linear autoregressions."

Shimko, David C, Naohiko Tejima, and Donald R Van Deventer. 1993. "The pricing of risky debt when interest rates are stochastic." *The Journal of Fixed Income*, 3(2): 58–65.

Sklar, M. 1959. *Fonctions de répartition à n dimensions et leurs marges.* Université Paris 8.

Sueyoshi, Glenn T. 1992. "Semiparametric proportional hazards estimation of competing risks models with time-varying covariates." *Journal of Econometrics*, 51(1): 25–58.

Yildirim, Yildiray. 2008. "Estimating default probabilities of CMBS loans with clustering and heavy censoring." *The Journal of Real Estate Finance and Economics*, 37(2): 93–111.

Zmijewski, Mark E. 1984. "Methodological issues related to the estimation of financial distress prediction models." *Journal of Accounting Research*, 22(3): 59–82.