March 2020

Essays on Mortgage Portfolio Diversification

Timothy Patrick Dombrowski
Louisiana State University and Agricultural and Mechanical College

ESSAYS ON MORTGAGE PORTFOLIO DIVERSIFICATION

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Finance

by
Timothy Patrick Dombrowski
A.A., Pasco-Hernando State College, 2012
B.A., Saint Leo University, 2015
May 2020

Acknowledgments

I would first like to thank my advisor and committee chair, Professor R. Kelley Pace, for incredible support and mentorship throughout my studies and the process of preparing this dissertation. I would like to thank my other committee members: Professors Rajesh P. Narayanan, V. Carlos Slawson Jr., Dek Terrell, and the Dean's Representative, Norman Massel, for an immense amount of insightful comments and feedback throughout the process of writing this dissertation. I would also like to thank Junbo Wang, Jason Brown, Nuno Mota, John Clapp, George Comer, David Harrison, Michael LaCour-Little, Moussa Diop, Lok Man Michelle Tong, Cai Liu, Sofia Dermisi, Dimuthu Ratnadiwakara, Shuang Zhu, David Suleiman, and numerous others who have contributed valuable remarks and suggestions across a variety of academic presentations and discussions.

Much of this work would not have been possible without the support of the E.J. Ourso College of Business at Louisiana State University. Particularly, the faculty and students of the Departments of Finance and Economics have contributed to this research through our day-to-day interactions and conversations. Lastly, I would also like to thank all of my friends and family who have provided unbounded support and encouragement throughout the entirety of my life.

Table of Contents

List of Tables

List of Figures

Abstract

Portfolios of mortgage loans played an important role in the Great Recession and continue to compose a material part of bank assets. The distribution of mortgage portfolio returns, and consequently, the risk of these portfolios, is quite distinct even from other fixed income asset classes. This dissertation contains three essays, each aiming to analyze a specific component of risk in mortgage portfolios and role of geographical diversification in reducing this risk.

The first essay investigates how cross-sectional dependence in the underlying properties flows through to the loan returns, and thus, the risk of the portfolio. In addition to demonstrating this relationship theoretically, this essay demonstrates how the spatially dependent structure of the underlying housing returns is revealed in the mortgage market by a shock to the default rate. The resulting increase in the asset correlations reduces the effectiveness of any geographical diversification present in the portfolio.

Even when the distribution of mortgage returns is known, the ability to reduce portfolio risk through geographical diversification can be limited due to the concentration of mortgage debt in major metropolitan areas. The second essay aims to model this geographical concentration for various partitions of the mortgage market and examine the role this has on limiting investors' ability to diversify risk. This is accomplished by fitting the empirical regularity from regional science known as the rank-size rule to measure this concentration.

The third and final essay in this dissertation focuses on modeling the mortgage default decision and imputing unobserved factors that may bias the estimated impact of observed factors such as the loan-to-value ratio. As alluded to in the first essay, the default rate, or probability of default ex-ante, is an important determinant of the observed correlation across mortgage returns. This essay develops a ridge regression model, which is tuned to maximize out-of-sample predictive performance using cross-validation, that imputes these unobserved factors while preventing model overfitting.

Chapter 1. Introduction

A oft-referenced topic that arises in discussions recalling the 2008 financial crisis is the credit ratings, or the evaluation of risk, for mortgage-backed securities (MBS). These financial products effectively bundle a large number of mortgages (loans that are secured by real estate assets) into a portfolio and sell off the cash flow rights to the portfolio's proceeds. With assets such as these, modeling the distribution of returns or measuring portfolio risk can be challenging. This dissertation aims to examine three important considerations for evaluating the risk of a mortgage portfolio and the impact of geographical diversification.

In the first essay, Chapter 2, the multivariate return structure of mortgage portfolios is examined, and the correlation across these returns is modeled as a function of the default rate and the correlation of the underlying housing returns. The censored variable framework that is used demonstrates how periods of strong economic growth and low default rates coincide with small observed correlations across mortgage returns and low risk for diversified portfolios. However, a shock to the default rate reveals the latent correlations of the underlying housing, and subsequently, the risk reduction from geographical diversification diminishes due to the revealed correlations from the recovered collateral.

This essay the models the cross-sectional dependence – correlation – of the underlying housing using local house price indices (HPIs) from the Federal Housing Finance Agency (FHFA). This dependence is partitioned into a strong (macro) component and a weak (spatial) component that fades as the distance between the properties increases. This dependence, and its spatial structure, often remains hidden during strong economic periods and low default rates. However, a default shock can have a disproportionately large impact on portfolio risk. The use of Value at Risk (VaR) in Section 2.4 shows the compounding effects from the increased risk of the individual assets and the increased correlation across asset returns. A simulation exercise in Section 2.5 demonstrates how the estimated housing correlations flow through to mortgage portfolio risk.

Although this relationship between portfolio risk and the underlying collateral exists for other fixed-income portfolios, the setting provided by residential mortgages is convenient since the value of the collateral is simply the value of the property. Unlike other fixed-income assets, such as corporate bonds where the collateral is the firm's assets, residential mortgages have readily available estimates of their collateral. The FHFA has made HPIs available for a wide range of geography levels and frequencies (Bogin, Doerner, and Larson, 2018). Additionally, residential properties exhibit substantial degrees of comovement (Fischer, Füss, and Stehle, 2019) and tend to have spillover effects, particularly with distressed assets (Lin, Rosenblatt, and Yao, 2009; Daneshvary, Clauretie, and Kader, 2011; Hartley, 2014). Thus, understanding the spatial structure of housing return dependence and how the return structure flows through to mortgage portfolios is important for modeling the return distribution at the portfolio level.

With multivariate normal asset returns such as in the traditional Markowitz (1952) model, the portfolio variance for a set of weights (w) is given by (1).

$$\sigma_p^2 = w' \cdot \Sigma \cdot w \tag{1}$$

In this case, the covariance matrix (Σ) for asset returns defines the diversification potential. One can think of diversification as the process of tuning the portfolio weights of w to achieve the investor's desired risk/return balance. If the weight vector w contains an element for every city or metropolitan statistical area (MSA) represented in the portfolio, then this process of geographically diversifying a portfolio may be constrained if the underlying assets (mortgages) are concentrated in just a small number of cities. This limiting effect on the diversification of mortgage portfolios that is induced by large geographical concentration of mortgage debt in major metropolitan areas is the focus of the second essay, Chapter 3.

When it comes to mortgage debt, this large degree of geographical concentration results from not just a larger quantity of loans originating in more populous cities, but also the price differential between dense urban areas and rural locations. In addition to these compounding effects of the quantity and size of new mortgage originations, D'Acunto and Rossi (2019) find that since 2011, the largest mortgage lenders have decreased the approval rate of small and medium-size loans, and increased those of large loans. This essay aims to examine the degree to which the resulting concentration of mortgage debt can limit the ability of investors to diversify a portfolio of these assets.

The examination of this limiting effect to portfolio diversification is further complicated once the institutional details of the U.S. mortgage market are factored in. For example, a report by the Urban Institute (Dec. 2019) shows that since being placed under governmental conservatorship in 2008, the government sponsored enterprises (GSEs) have guaranteed roughly 50% of all new mortgage originations. Combined with other governmental programs such as Federal Housing Administration (FHA) and Veteran Affairs (VA) mortgages, roughly 60–80% of new mortgages have their risk borne by taxpayers. This partitioning of the mortgage market with geographically disperse programs, leaves the remaining private mortgage market to be quite heavily concentrated, as will be demonstrated in Chapter 3.

Since the private-label MBS market dried up in 2008, most private market loans have been held as portfolio loans on bank balance sheets. In recent years (2018–2019), this private-label MBS market has seen a small recovery (roughly 2% of 2019 Q1–Q3 originations as per the Urban Institute report); however, this is still a small share of the private mortgage market as portfolio loans have accounting for at least a 30% share since 2014. Since most of the exposure to this highly concentrated private mortgage market is held by systemically important financial institutions, understanding the differences between the geographically disperse portfolios of the GSE securities and highly concentrated portfolios of non-conforming loans can be an important factor when measuring portfolio risk.

Coinciding with the resurfacing of the private-label MBS market in recent years, the Trump administration, along with the FHFA, has expressed interest in returning the GSEs back to the private market (Ackerman and Davidson, 2019b). This transition, which has begun with the Treasury Department and FHFA agreeing in late 2019 to allow the companies to begin retaining more earnings in preparation for the move (Ackerman and Davidson, 2019a), can have a substantial impact on the breakdown between the public and private mortgage markets. Based on the numbers from the Urban Institute report, this move would more than double the size of the private mortgage market.

If the GSEs are considered as a part of the private mortgage market during the pre-crisis period,[1] then the Urban Institute report shows more than 95% of first-lien mortgages originating into the private market at the peak in 2005–2006. Following the explicit guarantee of the GSEs and collapse of the private-label securitization market, this proportion fell to just 16% in 2008. The years following the crisis (2008–2013) saw roughly a 20% / 80% split between the private and public market shares. In 2014, the share of portfolio loans roughly doubled resulting in approximately a 30% / 70% breakdown for the 2014–2018 period. The most recent percentages for 2019 Q1–Q3 show another shift towards the private market with the respective shares at 39% and 61%. With these numbers, a reclassification of the GSEs to the private market would shock these percentages to 81% and 19%, respectively for the private and public markets.

Regardless of the breakdown of the private/public mortgage markets, a critical component of any analysis of risk in a mortgage portfolio is the prediction of default risk. The modeling of mortgage default has been studied across a rich literature; however, the limitations and reliability can often result in challenges to drawing conclusions. The final essay in Chapter 4 aims to address two potential omissions in mortgage default models and proposes an econometric method to attempt to impute these unobserved factors and correct the potential biases.

[1] Although Fannie Mae and Freddie Mac did not have an explicit guarantee that the U.S. government would bail them out in the event of financial difficulties, the GSEs did benefit from a direct line of credit to the U.S. Treasury. This access to low-cost debt and the general perception that the U.S. government would not allow these entities to fail are effectively known as an "implicit" guarantee. This guarantee became explicit in September 2008 when the GSEs were placed into conservatorship.

The first of these potential omissions is unobserved dynamics in borrower or loan characteristics. For example, if a borrower reports their income or credit score on a mortgage application, the use of this origination value in a dynamic default model implicitly assumes that the value remains constant over time. In other words, without addressing the omitted dynamics, one make the assumption that the borrower's credit score or income does not change over time. Left unaddressed, this can bias the estimated effect of other observed dynamic variables if they are correlated with the omitted variable. For example, if a borrower's income is correlated with the property value, then using the property value (either explicitly, or implicitly through the loan-to-value, LTV, ratio) will capture some of the explanatory power of the income dynamics.

A second potential omission for many mortgage default models is the unobserved heterogeneity across borrowers. Although some differences between borrowers can be controlled for explicitly in a predictive model, (ex. credit scores, income, etc.), unobservable factors such as borrowers attitudes towards default can be difficult to capture statistically. This unobserved heterogeneity across borrowers, both static and dynamic, is imputed in a penalized ridge regression model, which is tuned using 10-fold cross-validation to maximize out of sample predictive performance.

In their totality, these three essays combine to make a contribution to the modeling of risk and geographical diversification for mortgage portfolios. The first essays quantifies the diversification potential given the underlying multivariate distribution of housing returns. The second essay examines the ability for investors or stakeholders to achieve this potential diversification by modeling the geographical concentration of mortgage debt and its role in limiting the ability to diversify risk. The final essay focuses on the mortgage default decision and provides an econometric methodology for imputing unobserved heterogeneity and dynamics for default models.

Chapter 2. Mortgage Return Dependence

Since real estate constitutes one of the largest asset classes, loans secured by real estate (mortgages) also represent one of the largest fixed income asset classes. Usually such mortgages are held in a pool or a portfolio. Diversification provides at least one reason to hold a portfolio of mortgages and this raises the question of how much diversification can one obtain? Certainly, the poor performance of many of such mortgage portfolios during the Great Recession raises questions about the extent of possible diversification.

In this manuscript, we aim to answer this question at the finest scale possible. At the individual level, a loan effectively represents a censored random variable where performing loans yield a constant return and non-performing loans yield a variable return. The diversification potential across a portfolio or pool of mortgage loans would therefore depend on the correlations among censored random variables. These correlations between loan returns may be different than the correlations between the underlying properties or borrowers.

The nature of correlation between censored and truncated variables has a long history in psychology (Birnbaum, Paulson, and Andrews, 1950; Aitkin, 1964; Muthén, 1990). For example, Aitken motivates the issue by the situation where educational institutions screen individuals on aptitude tests, but only measure achievement on admitted students. The correlation between aptitude and achievement for the population (true latent correlation) is higher than for the selected group.

Although it might seem difficult to see how this would work in a loan context, we can consider the extremes to obtain intuition. To start, if all loans had no risk and their payoffs were constant, no diversification would be possible since the correlation between constants is zero. At the other extreme, if all loans defaulted and the lenders foreclosed, the correlation between loan returns would just equal the correlation among the future property returns. In other words, a set of mortgage loans where all have been foreclosed becomes a portfolio of properties.

In the case of mortgage loans, the latent correlation is equal to that of the underlying real estate values. Properties show high levels of spatial (Pace, 1997) and, more broadly, cross-sectional dependence (Pesaran, 2007). In this paper, we make a distinction between spatial and cross-sectional dependence that partitions correlations into diverisifiable and systematic portions. This is analogous to the distinction between weak and strong cross-sectional dependence in Chudik, Pesaran, and Tosetti (2011). In Section 2.2, we find roughly a 1:2 ratio of spatial to broader cross-sectional dependence in empirical HPI return correlations.

Regional science, urban economics, and spatial econometrics have documented the many ways that variables at one location can affect variables in other locations (Kuethe et al., 2008; LeSage and Pace, 2009; Hoogstra et al., 2017). Considering more intermediate cases for defaults, the extremes suggest that as the risk of loans increase (default rates rise) the correlations between the returns on loans will rise and the diversification potential will fall.

The goal of this chapter is to (1) document aspects of the spatial and cross-sectional dependence across borrowers and properties; (2) show how correlations among loan returns vary with default rate given the underlying correlations among properties; and (3) illustrate how much of a difference this makes for the diversification across mortgage loans in a portfolio.

The rest of this chapter is structured as follows: Section 2.1 describes the data used to document substantial dependence across housing returns in Section 2.2. This is followed with a theoretical model in Section 2.3, which relates the latent dependence with that of mortgage returns. Section 2.4 examines the implications for portfolio risk, and Section 2.5 combines these results into a simulation exercise to demonstrate the effect across portfolios of various sizes. Section 2.6 concludes the chapter.

2.1. Data

As follows from portfolio theory, larger correlations across asset returns limit the diversification potential for portfolios of these assets. This section will provide a brief discussion of the empirical data that will be used to document substantial dependence across housing returns. Since individual house values are only observed in the event of a sale, we make use of annual house price indices (HPIs) at the five-digit ZIP code level from the FHFA. These indices begin as early as 1975 and are available through 2017. For ease of presentation, we restrict the analyses to the 20 CS-MSAs.[2] We then use the 2010 ZIP Code Tabulation Area to Metropolitan and Micropolitan Statistical Areas Relationship File from the U.S. Census to identify the five-digit ZIP codes within each MSA.

Table 2.1 presents some statistics regarding the five-digit ZIP codes and populations within these MSAs. For our factor models in Section 2.2.1, we use national time series of the unemployment rate from the Bureau of Labor Statistics and the average 30-year fixed-rate mortgage from Freddie Mac. For our spatial analysis, the distances between ZIP codes are computed using the Haversine formula (Sinnott, 1984)[3] on the geographic coordinates for each combination, which are obtained from the 2017 U.S. Census Gazetteer Files.

2.2. Empirical Dependence

As is demonstrated theoretically in Section 2.3, mortgage loan returns are more highly correlated during periods with larger default rates. This correlation can be attributed to the larger portion of defaulting loans revealing the underlying housing correlations. In this section, we document substantial correlations across house price returns at the five-digit ZIP code level. Tools from spatial statistics are then employed to show how these returns covary as a function of the separation distance.

Although the market values for individual houses are only observed in the event of a sale, Bailey, Muth, and Nourse (1963) introduced a repeat-sales framework to construct price indices for housing markets. This methodology was further developed by Case and Shiller (1987, 1989) and later applied to various geography levels down to the census tract level by Bogin, Doerner, and Larson (2018) and made available by the FHFA. We use log returns of their annual HPIs at the five-digit ZIP code level.

[2] The 20 CS-MSAs are listed in full in Table A.1 of Appendix A.

[3] More specifically, the MATLAB function, lldistkm.m, by M. Sohrabinia available online at https://www.mathworks.com/matlabcentral/fileexchange/38812-latlon-distance.

Table 2.1. Summary Statistics for the 20 CS-MSAs

MSA	2010 HHs	ZIPs	100HH+	HPIs	>3T
ATL	2,165,495	227	208	196	195
BOS	1,883,206	286	277	257	257
CHA	737,775	90	81	71	71
CHI	3,797,247	414	386	362	361
CLE	955,756	118	108	101	101
DAL	2,502,075	283	270	230	230
DEN	1,078,837	147	131	110	110
DET	1,886,537	232	221	216	206
LV	840,343	73	69	54	54
LA	4,493,983	385	366	347	345
MIA	2,464,417	186	183	173	171
MIN	1,354,973	227	212	207	206
NY	7,527,752	919	863	734	732
PHX	1,798,501	164	158	140	139
POR	925,076	131	124	120	120
SD	1,164,786	108	100	87	86
SF	1,741,999	182	172	140	138
SEA	1,463,295	171	159	141	141
TPA	1,353,158	136	133	127	127
DC	2,213,752	352	300	281	280
ALL	42,348,963	4,831	4,521	4,094	4,070
USA	123,365,608	25,659	23,011	16,540	N/A

The 2010 HHs is the total number of households per the 2010 U.S. Census. ZIPs is the total number of five-digit ZIP codes. 100HH+ counts the ZIPs with more than 100 households. HPIs counts the ZIP codes that match to a HPI in the FHFA data. Finally, >3T removes a minimal number of ZIP codes to ensure at least 4 overlapping observations for every combination. Note: ZIP code 92672 contains 15,632 households in LA and 76 households in SD, so it is assigned to the LA MSA.

Since ZIP code combinations within MSAs are included in our dataset, these are likely to be highly correlated and may account for the larger mean (0.52) when compared to MSA-level mean of 0.29 in Cotter, Gabriel, and Roll (2014). To provide some feel for typical values of house price return correlations across various geographies, Table 2.2 presents some average correlations within and across MSAs. Table 2.3 presents selected correlations for MSAs,[4] and ZIP codes both within- and across the respective MSAs.

[4] Table B.1 in Appendix B expands Panel A of Table 2.3 to include all 20 CS-MSAs.

Table 2.2. Average Housing Return Correlations

MSA	Within-MSA	Across-MSA	Difference
ATL	0.6834	0.5157	0.1677
BOS	0.8008	0.4928	0.3080
CHA	0.5885	0.4080	0.1806
CHI	0.6920	0.5325	0.1595
CLE	0.5339	0.4025	0.1314
DAL	0.5490	0.2968	0.2522
DEN	0.6989	0.3605	0.3384
DET	0.7536	0.4756	0.2779
LV	0.9088	0.6303	0.2784
LA	0.8267	0.5308	0.2960
MIA	0.7953	0.5695	0.2258
MIN	0.7709	0.5939	0.1771
NY	0.7441	0.4941	0.2500
PHX	0.8630	0.5824	0.2806
POR	0.8113	0.4553	0.3560
SD	0.8102	0.5641	0.2461
SF	0.7789	0.5461	0.2327
SEA	0.8174	0.5085	0.3089
TPA	0.8299	0.6142	0.2157
DC	0.7700	0.5603	0.2096
ALL	0.7469	0.5062	0.2407
N	628,508	7,651,907	8,280,415

Average pairwise housing return correlations at the ZIP code level partitioned by combinations within- and across-MSAs. Averages are presented for each CS-MSA and the full sample along with the total number of combinations in each partition.

2.2.1. Weak vs. Strong Cross-Sectional Dependence

The high level of correlation among housing returns for cities separated by thousands of miles suggests that more than just spatial dependence is at work. Chudik, Pesaran, and Tosetti (2011) distinguish between weak (spatial) and strong (macro) cross-sectional dependence in the context of panel models. These definitions relate to the asymptotic behavior of the largest eigenvalue of the covariance matrix as the cross-sectional dimension (N) increases. If the largest eigenvalue converges to a constant value as N tends to infinity, this suggests the absence of any strong cross-sectional dependence and that only weak depen-

Table 2.3. Selected Housing Return Correlations

Panel A: Average House Price Correlations Across MSAs

MSA:	BOS	DAL	LA	MIN	NY	SD	SF
BOS	0.80						
DAL	0.32	0.55					
LA	0.47	0.20	0.83				
MIN	0.63	0.42	0.55	0.77			
NY	0.69	0.22	0.55	0.59	0.74		
SD	0.55	0.24	0.78	0.63	0.55	0.81	
SF	0.52	0.29	0.73	0.60	0.51	0.74	0.78

Panel B: Selected ZIPs Across MSAs

ZIP:	02176	76148	90230	55423	07066	92104	94523
02176	1.00						
76148	0.24	1.00					
90230	0.38	0.04	1.00				
55423	0.37	0.35	0.59	1.00			
07066	0.73	0.06	0.43	0.24	1.00		
92104	0.48	0.08	0.36	0.54	-0.16	1.00	
94523	0.45	0.21	0.79	0.66	0.48	0.40	1.00

Panel C: Selected ZIPs Within NY MSA

ZIP:	10467	11209	11214	11226	11229	11235	11385
10467	1.00						
11209	0.48	1.00					
11214	0.53	0.83	1.00				
11226	0.42	0.62	0.68	1.00			
11229	0.63	0.87	0.91	0.64	1.00		
11235	0.59	0.86	0.88	0.71	0.90	1.00	
11385	0.62	0.84	0.81	0.62	0.89	0.85	1.00

Panel A presents the average pairwise correlations across seven MSA combinations. Panel B presents the correlations for select ZIP codes in each of the seven MSAs, and Panel C presents correlations across the seven most populous ZIP codes in the NY MSA. Each of the ZIP codes in Panel B was listed as one of America's 50 Highest Demand ZIP Codes Of 2016 by Forbes.

dence remains. In the context of portfolios of mortgages, this would suggest some limitations to the diversification benefits of increasing the N, or the number of loans in the portfolio.

The two-stage approach of Bailey, Holly, and Pesaran (2016) for spatio-temporal analysis incorporates a test for weak cross-sectional dependence using the CD statistic developed in Pesaran (2004, 2015). Although the standard CD statistic is defined for balanced panels with fixed N and T, Pesaran (2004) provides an extension for unbalanced panels, which is the variant that we apply to our housing return panel. This statistic, presented in (2), involves a weighted average of the pairwise correlations ($\hat{\rho}_{ij}$), where the weights, $\sqrt{T_{ij}}$, are the square root of the number of periods with overlapping observations for units i and j.

$$CD = \sqrt{\frac{2}{N(N-1)}} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sqrt{T_{ij}} \hat{\rho}_{ij} \right) \tag{2}$$

Rejection of the null hypothesis in this first step suggests the presence of (semi-)strong dependence, which can be removed by using residuals from factor models. Bailey, Holly, and Pesaran (2016) suggest two approaches for implementing these factor models. The first regresses housing return time series for each ZIP code on the cross-sectional averages at the national or regional levels. Alternatively, a principal components approach can be taken using the panel of housing returns. Sufficiently reducing the test statistic to the point where the null is no longer rejected suggests that only weak dependence remains. At this point, modeling of the spatial structure commences with the defactored observations.

In addition to the CD statistic, Bailey, Kapetanios, and Pesaran (2016) develop a measure of the degree (or strength) of cross-sectional dependence known as the exponent of cross-sectional dependence. The α measure estimates the rate of increase in the largest eigenvalue of the covariance matrix as $N \to \infty$. Their bias-corrected estimator (3) consistently estimates this exponent for $\alpha > 1/2$. The interpretation of this measure is that larger α relates with higher degrees of cross-sectional dependence.

$$\mathring{\alpha} = 1 + \frac{1}{2}\frac{\ln \hat{\sigma}_{\bar{x}}^2}{\ln N} - \frac{\ln \hat{\mu}_v^2}{2 \ln N} - \frac{\hat{c}_N}{2[N \ln N]\hat{\sigma}_{\bar{x}}^2} \tag{3}$$

The estimation of α in (3) requires the use of the consistent estimators $\hat{\sigma}_{\bar{x}}^2$, $\hat{\mu}_v^2$, and $\hat{c}_N$. The first estimator (4) is simply a measure of the variation in the cross-sectional averages.

$$\hat{\sigma}_{\bar{x}}^2 = \frac{1}{T}\sum_{t=1}^{T}(\bar{x}_t - \bar{x})^2 \tag{4}$$

The estimator in (5) uses $\bar{x}_t(\mathbf{c}_p)$, which are the cross-sectional averages of only the ZIP codes with a significant factor loading in the time series regression on the cross-sectional averages. Following from Bailey, Kapetanios, and Pesaran (2016), we use the critical values suggested in Holm (1979).

$$\hat{\mu}_v^2 = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[\bar{x}_t(\mathbf{c}_p) - \bar{x}(\mathbf{c}_p)\right]^2} \tag{5}$$

The final estimator (6) makes use of $\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}^2$, where $\hat{u}_{it} = x_{it} - \hat{\delta}_i\bar{x}_t$, which is the observation less the variation explained by the cross-sectional average.

$$\hat{c}_N = \frac{1}{N}\sum_{i=1}^{N}\hat{\sigma_i^2} \tag{6}$$

In Column (1) of Table 2.4, we calculate these statistics for the raw housing returns as well as the average pairwise correlation across the panel. Column (2) presents the similar results after standardizing each time series. As evident from the large CD statistics, these panels exhibit substantial, strong cross-sectional dependence.

To reduce this strong dependence, we estimate factor models to remove variation from common, macroeconomic factors. Our first factor, which follows from Bailey, Holly, and Pesaran (2016), is the time series of cross-sectional averages from the housing return panel. After regressing each housing return time series on an intercept and the cross-sectional average, we test the residuals for cross-sectional dependence and produce Column (3) of Table 2.4. After defactoring with the cross-sectional averages, this greatly reduces the CD statistic down to 77.95.

Table 2.4. Measures of Cross-Sectional Dependence in Housing Return Factor Models

Panel A: Measures of Cross-Sectional Dependence							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Average Correlation	0.5245	0.5245	0.0019	-0.0003	-0.0007	-0.0004	-0.0008
CD Statistic	16096.7050	16096.7050	77.9481	11.8758	1.6385	10.6535	-0.1431
Exponent of CSD	0.9893	0.9910	0.6743	0.6588	0.6517	0.6491	0.6710
Controls:							
Cross-Sectional Average	N	N	Y	Y	Y	Y	Y
Current Unemployment	N	N	N	Y	N	Y	N
Lagged Unemployment	N	N	N	N	Y	N	Y
Current Interest Rates	N	N	N	Y	N	N	Y
Lagged Interest Rates	N	N	N	N	Y	Y	N
Panel B: Spherical Semivariogram Parameters							
Range (km)	2,833	893	829	772	778	769	781
Nugget	0.0020	0.1938	0.3946	0.4459	0.4421	0.4472	0.4400
Sill	0.0043	0.4903	0.9794	0.9760	0.9787	0.9732	0.9812

Panel A presents various measures of cross-sectional dependence across a variety of factor model specifications. These measures include the average pairwise correlation, CD test statistic calculated as in (2), and the exponent of cross-sectional dependence calculated as in (3). Panel B includes the parameters from a fitted spherical semivariogram model for each panel of housing returns. The results are presented for raw housing returns (1), standardized returns (2), and residuals from various factor regressions on the cross-sectional averages, unemployment rates, and interest rates.

In effort to remove other potential common factors, we include some additional macroeconomic factors that are often related to housing, such as unemployment rates and interest rates. Columns (4)–(7) of Table 2.4 present the cross-sectional dependence statistics for the residuals from these three-factor models using combinations of current and lagged unemployment and interest rates.

The CD statistic is sufficiently lowered to insignificance for two of the specifications: the three-factor models with lagged unemployment rates and either current ($CD = -0.14$) or lagged ($CD = 1.64$) interest rates. Since the exponent of cross-sectional dependence is lower with lagged interest rates ($\mathring{\alpha} = 0.6517 < 0.6710$), we select this specification to defactor the housing returns in the subsequent spatial analysis.

2.2.2. Spatial Dependence

In Section 2.2.1, the strong cross-sectional dependence was modeled using three factors: a market factor and lagged unemployment and interest rates. This section focuses on modeling the weak cross-sectional dependence by applying tools from spatial statistics on the factor model residuals, specifically from Column (5) in Table 2.4. Some common tools to describe the spatial structure of a process are the variogram, covariogram, and correlogram. A (semi-) variogram fits (half) the average squared difference between observations as a function of a measure of the distance between the points. A covariogram, and its normalized form, a correlogram, fit the covariance and correlation as a function of the distance measure.

Prior literature has considered both separation over time, yielding an autocorrelogram (Gourieroux and Jasiak, 2002), and space (Dubin, 1998), which yields a spatial cross-correlogram. Dubin, Pace, and Thibodeau (1999) discuss correlograms and semivariograms within the context of real estate values and fit empirical housing data to various functional forms including the spherical model (7), which we use to fit an empirical semivariogram of housing returns.[5]

[5] For a more detailed summary of variogram concepts and valid variogram models, see Anselin (2016) and Smith (2020).

The parameters a, s, and g refer to the nugget, sill, and range of the model, which respectively refer to the function value at the origin, the long-range semivariance, and distance at which the semivariance reaches the sill.

$$\gamma(k;a,s,g) = \begin{cases} 0 & k = 0 \\ a + (s-a)\left(\dfrac{3k}{2g} - \dfrac{k^3}{2g^3}\right) & 0 < k \leq g \\ s & k > g \end{cases} \tag{7}$$

For a cross-section of housing returns, an empirical semivariogram partitions the observations into K bins based on a distance metric (km between ZIP codes) and computes (8), which is equal to half of the average squared difference between housing returns across all combinations within each bin, k.

$$\hat{\gamma}(k) = \frac{1}{2|N_k|} \sum_{i,j \in k} (r_i - r_j)^2 \tag{8}$$

To take full advantage of the panel of data, the variogram measure is expanded to average across the squared differences over time in addition to being averaged at the cross-sectional level. Equation (9) calculates this where $T(i,j)$ is the number of periods with observed housing returns in both locations i and j, and $\bar{T}_k$ is a vector of the mean lengths of the time series in each bin k. This transforms the roughly 8 million combinations per year into the empirical semivariogram in Figure 2.1 where $K = 100$ equal sized bins.

$$\hat{\gamma}(k) = \frac{1}{2 \cdot (N_k \cdot \bar{T}_k)} \sum_{i,j \in k} \sum_{t \in T(i,j)} (r_{i,t} - r_{j,t})^2 \tag{9}$$

Panel B of Table 2.4 presents the fitted semivariogram parameters from the raw housing return panel, standardized returns, and defactored housing returns from the various factor models described in Section 2.2.1. Figure 2.1 plots the empirical semivariogram and fitted spherical model for the defactored housing returns from the three-factor model (Column (5) in Table 2.4).

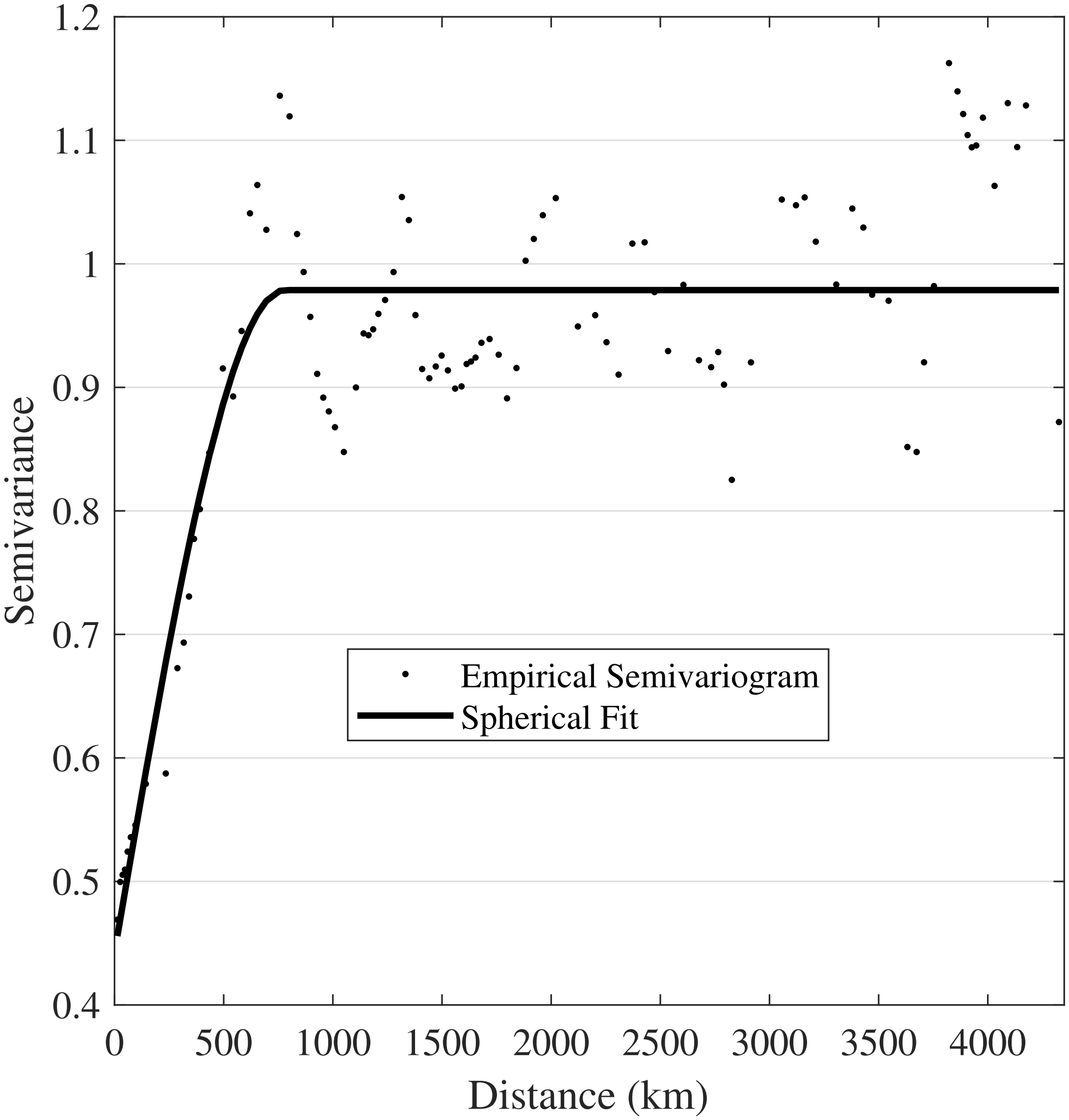

Figure 2.1. Empirical Semivariogram for Housing Returns

Note: Empirical semivariogram with 100 bins and spherical fitted curve for an unbalanced panel of 4,070 defactored housing return time series. Fitted model parameters, a, s, and k in (7), respectively refer to a nugget of 0.44, a sill of 0.98, and range of 778 km.

An intuitive extension of the variogram concept is its theoretical relation with the analogous concepts of covariograms and correlograms. Equation (4.1.7) of **?** is reproduced in (10), which relates the covariogram measure, $C(k)$, to the process variance, σ^2, and the variogram measure, $\gamma(k)$. Since we use standardized residuals with unit variance, the spherical semivariogram can be transformed to its implied covariogram, which is equivalent to the correlogram following from the unit variance.

$$C(k) = \sigma^2 - \gamma(k) \tag{10}$$

The semivariogram in Figure 2.1 suggests that spatial dependence alone might account for nearly a 0.5 correlation for properties located close to each other. At a distance of 200 km (roughly Los Angeles to San Diego), this spatial correlation is down to 0.35. For a distance of 450 km (roughly Los Angeles to Las Vegas), the correlation is down to 0.14, and by the range of 778 km (just beyond the distance between Los Angeles and San Francisco), the correlation drops to 0 and the spatial dependence fades.

However, these tools from spatial statistics are only modeling the weak component of the housing return dependence. The strong dependence that was removed in the factor regressions must be replaced in order to account for dependence imposed by macroeconomic factors. After using the spherical variogram model to estimate the spatial correlation given a specific separation distance, we can replace the strong dependence that was removed in the factor regressions. This provides a framework for simulating asset returns with both strong and weak cross-sectional dependence, which is the focus of next section.

2.3. Theory

In this section, we model the relation between the risk of a loan portfolio with the dependence in the underlying collateral. When considering the return of an individual mortgage and the underlying property value, appreciation (or depreciation) in the latter is censored so long as the borrower remains current on the loan. In other words, performing loans yield constant returns, which do not covary with other loan returns. Alternatively, if the borrower defaults and the lender forecloses, the loan return is equal to the return on the value of the property securing the loan. Thus, intermediate cases suggest that the risk of a portfolio of mortgages involves correlations across censored random variables.

The statistical foundations for our derivation of the correlation between censored variables comes from the work on the closely related topic of truncation. Early work on truncation (see Birnbaum (1950) and Aitkin (1964)) has roots in psychology and the effect of truncation in educational admissions. Our derivation of the correlation across censored variables is adapted from the example in Chapter 46 of Kotz, Balakrishnan, and Johnson (2000), which focuses of the bivariate truncated normal distribution.[6]

To begin, consider a one-period loan with principal, L, and coupon rate, c. The price of the underlying collateral equals 1 at origination and has a value after one period of $V = 1 + r$ where r represents the return on the asset. At maturity of the loan, the borrower either defaults ($d = 1$) or repays the loan ($d = 0$). If the borrower repays $L(1 + c)$, this yields a constant return of c. If the borrower defaults, the lender receives the value of the collateral, which yields a variable return of $V/L - 1$ or $L^{-1}(1 + r) - 1$.

If $V < L(1 - \kappa)$ where κ represents potentially heterogeneous frictions to default,[7] then the borrower defaults. The equivalent condition in (11) defines this dichotomy in terms of the housing return where $h = L(1 - \kappa) - 1$. This default threshold combines both frictions and leverage. We assume it is constant across borrowers and will tune it to vary the default rate at the portfolio level as the paper progresses.

$$d = \begin{cases} 1, & \text{if } r < h \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

At $t = 0$, the lender has a portfolio of two loans: one risky loan that defaults as in (11), and a loan which has previously defaulted at some $t < 0$. The return on the risky loan is given

[6] A more detailed derivation relating censoring with truncation is provided in Appendix C.

[7] A parameter of $\kappa = 0$ would represent the case of ruthless default.

in (12), which is censored from above by performance of the loan. If the borrower defaults, the loan return is equal to a linear transformation of the collateral return.

$$R_1 = \begin{cases} r_1^* & \text{if } r_1 < h \\ \\ c & \text{otherwise} \end{cases} \tag{12}$$

$$r_1^* = L^{-1}(1 + r_1) - 1 \tag{13}$$

For the defaulted loan, the initial loss is a sunk cost and future returns are equal to the returns on the underlying collateral. Thus,

$$R_2 = r_2 \tag{14}$$

For simplicity, we assume that the housing returns follow a bivariate normal distribution with a correlation of ρ. Since correlations are invariant to the addition of constants and factors of proportionality, we let r_1^* and r_2 each have zero mean and unit variance. Finally, we assume a zero coupon rate for the loans to further simplify the solution. This yields a correlation equal to that of a censored random variable ($\tilde{r}_1 = d \cdot r_1^*$) and the uncensored return of the defaulted loan, r_2.

$$\text{corr}(R_1, R_2) = \text{corr}(\tilde{r}_1, r_2) \tag{15}$$

The correlation between the risky loan and defaulted loan is defined by (16). Unlike truncation, censoring of one variable does not affect the distribution of the other; thus, the normalized return of the previously defaulted loan results in the latter terms of both the numerator and denominator dropping out, yielding (17).

$$\begin{aligned} \text{corr}(\tilde{r}_1, r_2) &= \frac{\mathbb{E}[\tilde{r}_1 r_2] - \mathbb{E}[\tilde{r}_1]\mathbb{E}[r_2]}{\sqrt{\text{var}(\tilde{r}_1)}\sqrt{\text{var}(r_2)}} && (16) \\ &= \frac{\mathbb{E}[\tilde{r}_1 r_2]}{\sqrt{\text{var}(\tilde{r}_1)}} && (17) \end{aligned}$$

Appendix C derives the relevant univariate and bivariate moments, including the correlation, for this case of right censoring on the standard normal distribution. This yields the moments for $\tilde{r}_1$ given by (18)–(20), where $\phi(\cdot)$ and $\Phi(\cdot)$ refer to the normal pdf and cdf, respectively.

$$\mathbb{E}[\tilde{r}_1] = -\phi(h) \tag{18}$$

$$\mathbb{E}[\tilde{r}_1^2] = \Phi(h) - h\phi(h) \tag{19}$$

$$\text{var}(\tilde{r}_1) = \Phi(h) - h\phi(h) - \phi(h)^2 \tag{20}$$

The joint expectation is solved using the law of total expectation.

$$\mathbb{E}[\tilde{r}_1 r_2] = \mathbb{E}_{\tilde{r}_1}\left[\mathbb{E}[\tilde{r}_1 r_2 | \tilde{r}_1]\right] \tag{21}$$

Since $\tilde{r}_1$ is a constant within the inner expectation, it factors out yielding (22).

$$\mathbb{E}[\tilde{r}_1 r_2] = \mathbb{E}_{\tilde{r}_1}\left[\tilde{r}_1 \cdot \mathbb{E}[r_2 | \tilde{r}_1]\right] \tag{22}$$

The conditional expectation of r_2 is identical to that of the truncated distribution in (143) of Table C.1 in Appendix C, which yields (23).

$$\mathbb{E}[\tilde{r}_1 r_2] = \mathbb{E}_{\tilde{r}_1}\left[\rho \cdot \tilde{r}_1^2\right] \tag{23}$$

After factoring out ρ and substituting in (19), we obtain (24).

$$\mathbb{E}[\tilde{r}_1 r_2] = \rho \cdot (\Phi(h) - h\phi(h)) \tag{24}$$

Finally, we substitute (24) and (20) into (17) and then (15) and obtain the closed form solution for the correlation, (25).

$$\tilde{\rho} = \text{corr}(R_1, R_2) = \rho \cdot \frac{\Phi(h) - h\phi(h)}{\sqrt{\Phi(h) - h\phi(h) - \phi(h)^2}} \tag{25}$$

From this solution, the underlying latent correlation is reduced by a factor equal to a nonlinear function of the default cutoff parameter.

$$\frac{\tilde{\rho}}{\rho} = \frac{\Phi(h) - h\phi(h)}{\sqrt{\Phi(h) - h\phi(h) - \phi(h)^2}} = \lambda(h) \leq 1 \tag{26}$$

For some numerical intuition, if we evaluate (26) at $h = 0$, we have $\Phi(0) = 0.5$ and $\phi(0) \approxeq 0.4$. Using these values, the loan correlation is approximately 6/7 of the latent correlation. This suggests that the first 50% of defaults reveal over 85% of the underlying housing return correlation. Figure 2.2 depicts this nonlinearity for three cases of latent correlations, $\rho = 0.1, 0.5, 0.9$. In regards to the sensitivity to the probability of default, the figure shows that the loan return correlations are at their most sensitive for the lowest default rates.

In the spirit of providing a useful approximation, such as the popular Rule of 72 for estimating the duration or interest rate to double one's investment, we provide a simplification of this closed form solution for the loan return correlation, we apply the natural logarithm to both sides of (26), which yields (27). The linear approximation of this relation, (28), is estimated using OLS on 1,000 equal spaced increments over the range of $h \in [-3, 0]$, which corresponds to a domain of $\Phi(h) \in (0, 0.5)$ or probability of default ranging from 0 to 50%. This transforms the solution into a exponential function of the default rate, (29), and fits with an R^2 of 0.9914 compared to the exact solution from (26).

$$\ln\left(\frac{\tilde{\rho}}{\rho}\right) = \ln(\Phi(h) - h\phi(h)) - 0.5 \cdot \ln(\Phi(h) - h\phi(h) - \phi(h)^2) \tag{27}$$

$$\approxeq 0.2070 + 0.3336 \cdot \ln(\Phi(h)) \tag{28}$$

$$\implies \frac{\tilde{\rho}}{\rho} \approxeq \underbrace{\exp(0.2070)}_{=1.23} \cdot \Phi(h)^{0.3336} \tag{29}$$

To further simplify this approximation, the coefficient on $\ln(\Phi(h))$ is rounded to 1/3 to obtain the cube-root form in (30). This adjustment maintains the high correlation (0.9928) with the true solution in (26) and improves the simplicity of its functional form.

$$\frac{\tilde{\rho}}{\rho} \approxeq 1.23 \cdot \sqrt[3]{\Phi(h)} \tag{30}$$

Figure 2.3 depicts this approximation in reference to the true scaling factor over the domain of $\Phi(h) \in (0, 0.3)$. As would be suggested by the high R^2 from the estimation, this approximation closely fits to the true solution over the most relevant domain (< 30% default).

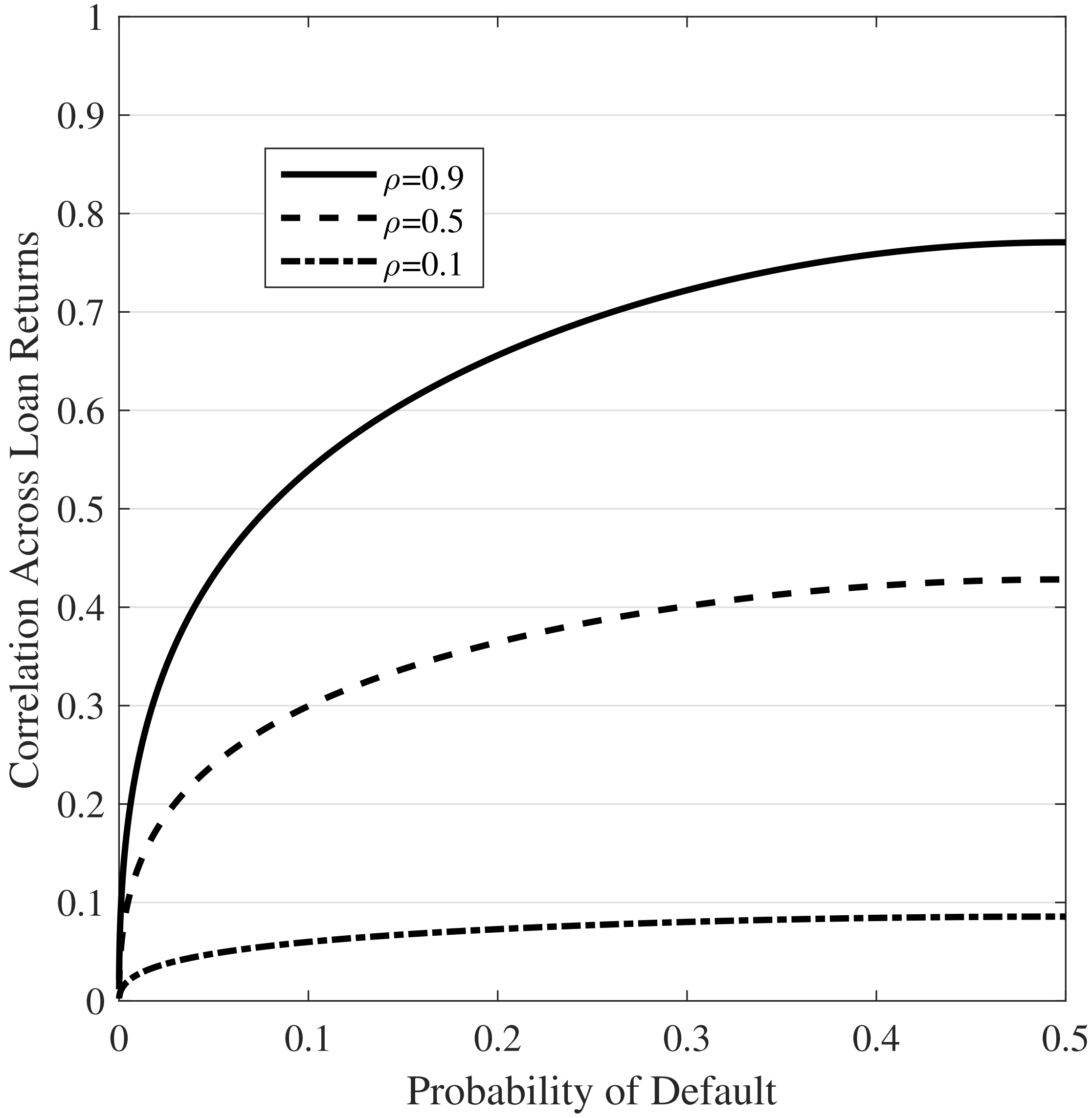

Figure 2.2. Loan Return Correlations and the Default Rate

Note: Loan return correlations across various default rates and underlying asset correlations. Correlations are for a two-asset portfolio with one loan defaulting with 100% certainty and the other with a likelihood of being censored equal to 1 minus the probability of default. Each curve reflects a specific underlying housing return correlation.

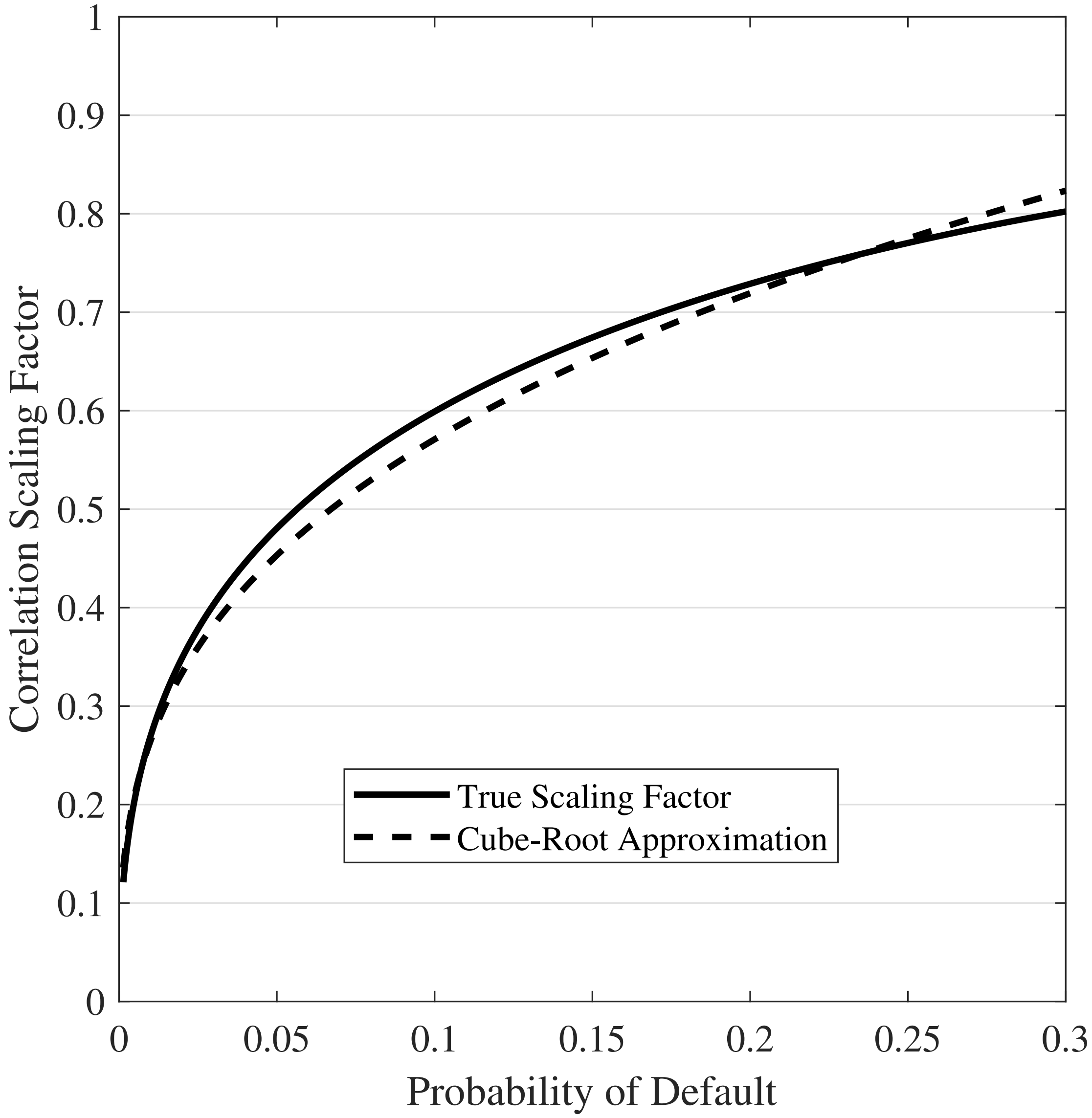

Figure 2.3. Cube-Root Approximation for Loan Return Correlations

Note: Plot depicting the cube-root approximation (dashed line) from (30) and the true correlation scaling factor (solid line) from (26) over the interval of $\Phi(h) \in (0, 0.3)$.

2.4. Portfolio Implications

To demonstrate the implications for holding a portfolio of loans with returns that are correlated collateral returns, we first examine the expected return and risk of the two-loan portfolio described previously in Section 2.3. We let the weight of the risky loan equal w, and thus the defaulted loan has a weight of $1 - w$.

$$\mathbf{w} = \begin{bmatrix} w & 1-w \end{bmatrix}^T \tag{31}$$

Using (12), (14), and (18), the expected returns are given by (32).

$$\mathbf{r} = \begin{bmatrix} c - \phi(h) & 0 \end{bmatrix}^T \tag{32}$$

From (20) and (24), we construct the covariance matrix, (33).

$$\Sigma = \begin{bmatrix} \Phi(h) - h\phi(h) - \phi(h)^2 & \rho \cdot (\Phi(h) - h\phi(h)) \\ \rho \cdot (\Phi(h) - h\phi(h)) & 1 \end{bmatrix} \tag{33}$$

Using the traditional formulas for portfolio expected return and risk, (34) and (36), we obtain the closed form solutions in (35) and (37).

$$\begin{aligned} \mathbb{E}[r_p] &= \mathbf{w}^T \cdot \mathbf{r} && (34) \\ &= w(c - \phi(h)) && (35) \\ \sigma_p^2 &= \mathbf{w}^T \cdot \Sigma \cdot \mathbf{w} && (36) \\ &= w^2(\Phi(h) - h\phi(h) - \phi(h)^2) \\ &+ 2\rho w(1-w)(\Phi(h) - h\phi(h)) + (1-w)^2 && (37) \end{aligned}$$

From this simple two-loan portfolio, we see that an economic shock increasing the underlying default parameter, h, will influence a portfolio in multiple ways. First, the expected return falls for each asset. Additionally, the variance of each asset increases, and this leads to an increased correlation across the loan returns. A Value at Risk (VaR) measure consolidates these impacts into a single risk measure, (38). This effectively measures the portfolio loss where losses beyond this value occur with probability α.

$$\text{VaR}(h;\alpha) = z(1-\alpha) \cdot \sqrt{\sigma_p^2} - \mathbb{E}[r_p] \tag{38}$$

Assuming the distribution of portfolio returns is normally distributed, then $z(1-\alpha) = \Phi^{-1}(1-\alpha)$, which suggests that the 1% worst outcomes produce returns roughly 2.33 standard deviations less than the expected portfolio return. From this measure, we can show how a macroeconomic shock increasing default rates will percolate through to risk of a portfolio in a single measure that combines the effects on the expected returns, variances, and correlations.

To better demonstrate these different effects, we compute the derivative of this VaR measure with respect to the theoretical default rate, $\Phi(h)$. The first step, (39), uses the chain rule to first take the derivative with respect to h before transforming to units of the default rate, Φ.

$$\frac{d\text{VaR}}{d\Phi} = \frac{d\text{VaR}}{dh} \cdot \frac{dh}{d\Phi} \tag{39}$$

After substituting (38) into (39), the result, (40), decomposes the derivative into three important pieces: $d\mathbb{E}[r_p]/dh$, $d\sigma_p^2/dh$, and $dh/d\Phi$.

$$\frac{d\text{VaR}}{d\Phi} = \left(z(1-\alpha) \cdot \frac{1}{2} \left(\sigma_p^2\right)^{-1/2} \frac{d\sigma_p^2}{dh} - \frac{d\mathbb{E}[r_p]}{dh} \right) \cdot \frac{dh}{d\Phi} \tag{40}$$

The first derivative, $d\sigma_p^2/dh$, captures the effect of a shock to the default rate on the variance of the portfolio. From the variance in (37), this includes the change in the variance of the risky asset (first term) and the effect on the covariance (second term). Over the most relevant domain ($<$ 30% default), as the default rate increases, the variance of the risky asset and the covariance both increase, which combine to increase the variance across portfolio returns.

The second derivative, $d\mathbb{E}[r_p]/dh$, demonstrates the effect of the shock on the expected return of the portfolio, and the final term simply converts the derivative units to the theoretical default rate instead of the underlying default cutoff parameter. Intuitively, as default rates rise, the expected portfolio return decreases in the relevant domain, and since $\Phi(h)$

is the normal cdf, the last derivative is equal to $\phi(h)^{-1}$, which is strictly positive over the domain of (0,1).

If we consider an economic shock that increases the default rate, then the expected return for each asset decreases, the variance of the each asset increases, and the correlation across the returns in the portfolio increases. The VaR framework demonstrates how these effects compound on each other to increase the riskiness of a portfolio of mortgages when default rates rise (Figure 2.4).

2.5. Simulations

As demonstrated in Section 2.3, the latent correlations across housing returns flow through to the loans secured by those properties as a function of the default rate. The empirical housing return correlations estimated in Section 2.2 show that these underlying correlations are certainly non-trivial and likely to play a major role in the risk of a mortgage portfolio in periods with high default rates. In this section, we build off of these foundations and conduct a simulation exercise, which extends the two-loan example of Section 2.3 to portfolios of various sizes and uses parameters based off of empirical estimates in Section 2.2.

We begin by simulating 100,000 independent draws for 4,070 random variables from a standard normal distribution, each corresponding with a specific ZIP code. To mimic the defactored housing returns, a positive-definite correlation matrix is generated by applying the spherical semivariogram model to the distance matrix, D, and estimating the theoretical, positive-definite correlation matrix.

After imposing these correlations on the simulated variables, we simulate 100,000 draws for each of the three factors from the factor regressions and impose the empirical correlations across the factors. These simulated factors are then multiplied by the respective factor loadings for each ZIP code and added to the simulated variables, which are scaled by the model RMSEs (mean RMSE = 0.69). This procedure effectively adds back the strong cross-sectional dependence that was removed in the factor regressions.

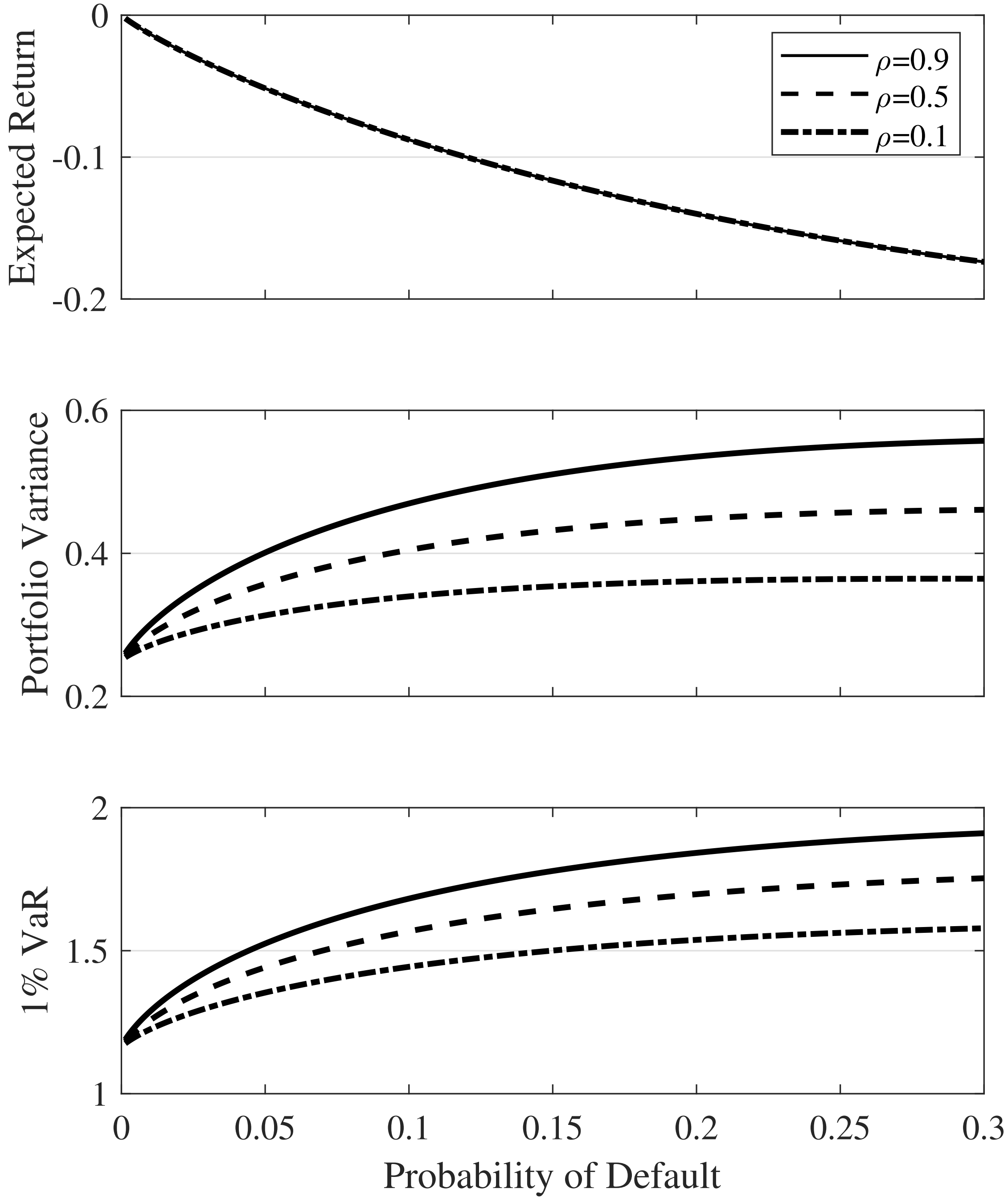

Figure 2.4. Portfolio Risk and the Default Rate

Note: Simple example of the relation between the distribution of portfolio returns across various risk levels and latent correlations. The expected return of the portfolio decreases the same regardless of the underlying housing correlations; however, the variance increases more with larger latent correlations and thus amplifies the effect of an economic shock to the default risk on a VaR measure.

These simulated variables now reflect similar correlations as the empirical housing returns. However, these represent the normalized housing returns. To match the empirical distributions, we scale each simulated variable by the respective standard deviation and add the mean return. This results in simulated variables with similar means, variances, and correlations as the empirical housing return panel.

With these simulated asset returns, we construct portfolios of various sizes ($N = 5, 50, 500$) by randomly sampling from the simulated variables with replacement, which effectively assumes there is sufficient mortgage debt secured by properties in each ZIP code. As in Section 2.3, we let the underlying default cutoff vary such that the portfolio default rates vary from 0–30% default. For the performing loans, we assume a constant 5% coupon rate. The standard deviation of the equal weighted portfolio returns is then estimated where $\mathbf{w} = 1/N \cdot \iota$, ι is an N-element column vector of ones, and Σ is the $N \times N$ covariance matrix of the simulated portfolio loan returns.

$$\sigma_p = (\mathbf{w}^T \cdot \Sigma \cdot \mathbf{w})^{1/2} \tag{41}$$

Given the simulated default choices for a given cutoff h, the portfolio standard deviation is computed as in (41). Figure 2.5 presents this measure of portfolio risk across various default rates and portfolio sizes. Intuitively, as the default rate rises, portfolio risk increases. Similarly, portfolios with more loans have less risk; however, the marginal benefit is decreasing in size. The reduction in risk when going from 5 to 50 loans is substantially larger than the additional reduction from increasing to 500 loans.

If we compare these results with some real mortgage-backed securities, many of these securities contain more than 1,000 loans with a few even exceeding 10,000 loans. These findings suggest that these larger MBS deals might only be marginally more diversified than the smaller deals with only several hundred loans. However, diversification in reality depends on more dimensions than just geographical space. Variation across loan terms, vintage, and borrower characteristics are examples of some other dimensions that might provide incentives to further grow and diversify a portfolio. Additionally, due to economics of scale and the presence of fixed costs when constructing a portfolio of mortgages, larger portfolios provide benefits beyond just diversification.

2.6. Discussion

In this chapter, we investigate the impact of underlying housing return correlations on the risk of a mortgage portfolio. Individual mortgage returns can be viewed as censored random variables where defaulting loans reveal the latent housing return correlations. This results in the distribution of mortgage portfolio returns as a function of both the underlying house price correlations and the default realizations. At one extreme, the ideal scenario of 0% default yields a constant return with no risk. At the other extreme, a 100% default rate transforms the loan portfolio into a portfolio of houses and the return correlations are equal to the underlying house price return correlations. Thus, low default rates imply high censoring which masks the latent asset correlations.

The theoretical model developed in Section 2.3 relates the correlation across loan returns to the underlying house price return correlations and the likelihood of censoring ($1-$probability of default). When considering a portfolio of these loans, the interaction between the default rate, the underlying housing co-movement, and portfolio theory increase the complexity of modeling the risk in a mortgage portfolio. Since underwriting standards usually strive to keep default rates low, and do so the vast majority of the time, historical correlations of returns do not provide a very good idea of the behavior of portfolios under severe economic conditions. Under poor economic conditions where default rates rise, we how that loan returns fall, correlations among loan returns increase, and the potential to diversify away risk is reduced.

To contextualize the non-linearity of this relationship, let's consider the lowest pre-crisis delinquency rate from Federal Reserve Economic Data (FRED) (1.41% in 2004Q4) to the peak (11.54% in 2010Q1).[8] Using our cube-root approximation in (30) and Figure 2.3, the

[8] Although the 30+ day delinquency rates from FRED may be too broad to directly reflect a hard default as in the model, the relative increase may provide a reasonable comparison across

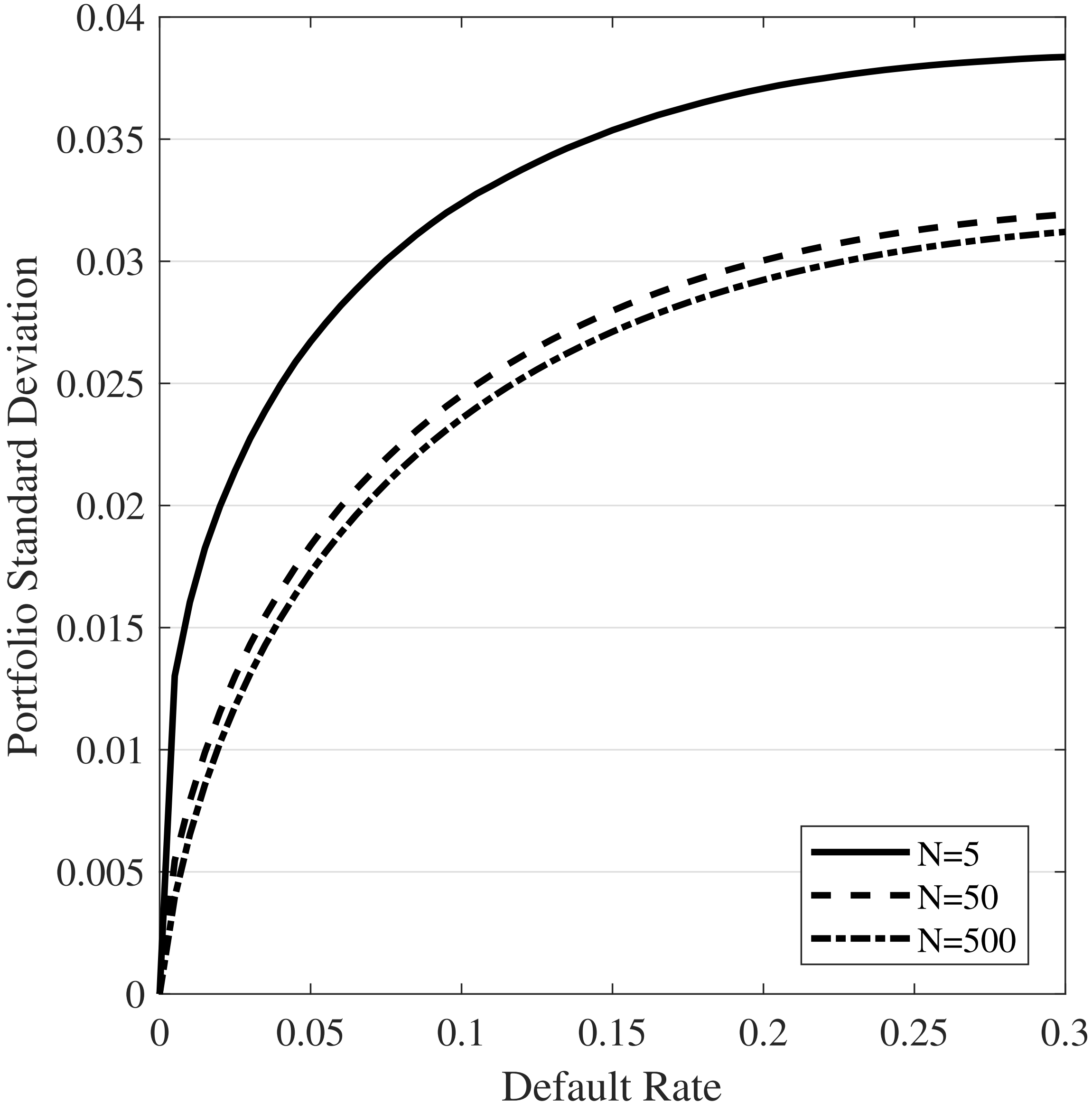

Figure 2.5. Simulated Effects of Portfolio Size on Risk

Note: Standard deviation across 100,000 draws of simulated portfolio returns. Simulated returns are scaled and correlated so as to match the empirical distributions in means, varaiances, and all pairwise correlations. Each curve plotted represents the standard deviation across randomly selected portfolios of each size in each simulation.

proportion of the underlying housing correlations that is revealed in a mortgage portfolio would effectively double from 29.7% to 59.9% at the peak.

In the case of real estate loans or mortgages, the large degree of dependence in the underlying collateral (house price returns) has both strong (macro) and weak (spatial) aspects. In Section 2.2, we documented substantial correlations across house price returns at the five-digit ZIP code level. The empirical correlations suggest that two properties within the same MSA are roughly 50% more correlated than two properties secured by properties in different MSAs. The impact of this spatial dependence appears to fade to zero around 778 km or roughly the distance between Los Angeles and San Francisco.

Incorporating the FRED delinquency rates with the empirical housing correlations, the estimated mortgage return correlation for properties in Los Angeles and San Diego would increase from 0.23 to 0.46. However, for geographically separate MSAs such as Los Angeles and New York City, the increase from 0.16 to 0.33 is proportionally similar, but smaller in magnitude due to the lack of spatial dependence.

Section 2.5 demonstrates the relation between the default rate of a portfolio and the variance across simulated mortgage returns tuned to the empirical parameters defining the underlying covariance structure. Intuitively, as the number of loans in the portfolio grows, the risk is reduced; however, this reduction in risk is substantially larger when going from 5 to 50 loans than it is when going from 50 to 500 loans.

Again, to contextualize with the FRED delinquency rates, the standard deviation in Figure 2.5 across 5-loan portfolios is quite large even for low default rates, but still increases by 83% from 0.018 at 1.41% default to 0.033 at 11.54%. This sensitivity of portfolio risk to the default rate increases with portfolio size as the 50-loan portfolios increase by 160% (from 0.010 to 0.026) and the 500-loan portfolios increased by 190% (from 0.009 to 0.025). These

regimes. Potentially offsetting this attenuation is the underestimation of the true theoretical relation in Figure 2.3. Additionally, the FRED rates only consider loans that are still accruing interest; thus, previously defaulted loans may still be exposed in a mortgage portfolio and contribute to the asset co-movement.

findings suggest a drastic difference in mortgage portfolio return behaviors in good times and bad.

To conclude, our findings quantify the diversification potential for portfolios of mortgages where the underlying collateral exhibit substantial cross-sectional dependence. Future potential for extensions of this work may include the examination of the ability to achieve this potential diversification through the concentration of mortgage debt across space. Further, one might extend this investigation to asset classes beyond mortgages and real estate by incorporating an MGARCH approach as in Heaney and Sriananthakumar (2012).

Chapter 3. Geographical Concentration of Mortgages

In the context of mortgage portfolios, where loans are secured by the underlying real estate assets, geographical diversification is an attractive approach for risk management. However, the concentration of mortgage debt in major metropolitan areas can result in limitations to the amount of diversification that stakeholders can attain. Stakeholders in the mortgage market are somewhat more involved than for some other asset classes. For example, according to the U.S. Department of Housing and Urban Development, the FHA guarantees approximately 12% of newly originated mortgage debt as of 2018, and the risk of these mortgages is borne by U.S. taxpayers.

Additionally, another 42% of mortgage debt is guaranteed and securitized by Fannie Mae and Freddie Mac, which are currently under government conservatorship and classified as GSEs. As per the Urban Institute (Dec. 2019), these governmental programs, along with Veteran Affairs (VA) mortgages, combine for more than a 60% share of new originations. The remaining 40% of private market loans are held as portfolio loans or packaged into non-agency mortgage-backed securities. Since the 2008 housing crisis when the private securitization market effectively vanished, the vast majority of the private market loans have been held on the balance sheets of systemically important financial institutions. Of these bank-owned mortgages, approximately 50% of the aggregate balance comes from jumbo mortgages, which have principal amounts greater than the GSE's conforming loan limits (CLLs) and are ineligible to be purchased by the GSEs.

The varying regulatory requirements across these governmental programs, such as minimum credit scores or the CLLs, result in differential degrees of geographical concentration across the various programs and classifications. Since these programs generally capture a relatively disperse set of mortgages, the remaining private market tends to be quite heavily concentrated. For example, although the conforming mortgage market (loans with balances under the CLLs) has only a modest degree of geographical concentration, the residual jumbo mortgage market exhibits a substantial degree of concentration.

As these governmental programs continue to play an increased role in mortgage markets, the implications regarding the risk of these portfolios is an important consideration for the economy. As noted in a recent article by *The Washington Post*, these governmental agencies guarantee nearly $7 trillion in mortgage-related debt, which is 33% more than the years prior to the 2008 housing crisis (Paletta, 2019). This paper examines the use of rank-size relations to parameterize the geographical concentration of mortgage markets and the potential implications of large concentrations for portfolio diversification.

Rank-size relations provide a means to characterize the geographical concentration in mortgage debt. For example, the Zipf distribution (Zipf, 1949) suggests a linear relationship between the natural logarithms of city sizes and their respective ranks when sorted from largest to smallest. The traditional rank-size rule, or Zipf's Law, is a special case where the slope of the log-linear relationship is equal to -1. This gives the simple result that the size of a city multiplied by its rank is constant, or that the second largest city is 1/2 the size of the largest, the third is 1/3 the size of the largest, and so on.

Power law distributions, such as Zipf's, have been a long-standing empirical regularity with various applications including city sizes, firm sizes, wealth, international trade, and word-use frequencies across many languages.[9] Fitting these distributions effectively produces measurements that describe the degree of concentration in a variable. As a result, the geographical distribution of mortgage debt can be simplified down to just one or two parameters.

With the Zipf distribution, the limiting distribution of city sizes follows a power law distribution where the rate of geometric decline is given by the estimated slope parameter. However, when considering larger and more complete sets of cities, Eeckhout (2004) argues that the law of proportionate effect (city growth is independent of absolute size) results in

[9] Although originally credited to Auerbach (1913), the rank-size rule was popularized by Zipf (1949), and has since branched into a vast literature. Some more recent papers include: Brakman, Garretsen, Van Marrewijk, and Van Den Berg (1999); Axtell (2001); Reed (2002); Ioannides and Overman (2003); Klass, Biham, Levy, Malcai, and Solomon (2006); Gabaix (1999, 2011); Piantadosi (2014); Chaney (2018).

a limiting distribution that is log-normal, which is known as Gibrat's Law (Gibrat, 1931). Following some additional debate (Eeckhout, 2009; Levy, 2009), Malevergne, Pisarenko, and Sornette (2011) use the uniformly most powerful unbiased test to compare the two and conclude that the power law hypothesis should be accepted, and that the log-normal hypothesis be rejected.

However, rather than simply testing the hypothesis of city sizes following the traditional rank-size rule, I aim to parameterize the geographic concentration of mortgage debt using data from Black Knight Financial Services (BKFS). One notable drawback of the power law fit of the Zipf distribution is that it is most accurate in the tail and overestimates the size of the largest cities for the U.S. I resolve this concern by adapting the parabolic fractal distribution, which extends the log-linear form of the Zipf distribution to include a quadratic term for $\ln(rank)$. This greatly improves upon the linear fit, particularly for the largest cities.

The interpretation of the of the estimated parameters becomes more complicated with the parabolic fractal distribution. To address this, I orthogonalize the quadratic term from the linear term, which isolates the effects from each part. Thus, the additional fit provided by curvature can be compared directly to the portion that is explained by the linear term from the Zipf distribution. This can also be interpreted in the sense of comparing the power law fit from the linear component with a correction term that remedies the functional misspecification. I find that the relative effect of these two components is fairly constant in explaining the geographical concentration across both populations and the various sectors of the mortgage market.

Although the market shares of each city can be directly measured from the data, these rank-size relations provide a way to parameterize the distributions into just a single measure of concentration. The fitted rank-size curves are then scaled to produce a probability mass function or implied weights based on these estimates. These effectively represent theoretical weights for portfolios constructed by randomly sampling from the implied distributions.

Alternatively, one can think of these weights as capturing the relative exposure of a specific market or variable to local shocks in each individual city.

When considering a portfolio of mortgages, the aggregation from the risk of individual assets to portfolio-level risk involves both a correlation and concentration component. In this manuscript, the focus is on modeling the concentration component using the aforementioned rank-size relations. The correlation across mortgage returns can be decomposed into weak spatial dependence and strong macroeconomic dependence (Dombrowski, Pace, and Narayanan, 2020), which describe the potential for diversification. Given the dependence structure of returns, the concentration component can limit the ability for investors to achieve the suggested diversification potential.

Within real estate, the role of geographical diversification has been studied going back to Corgel and Gay (1987), who focus on diversifying across local economic conditions. More recently, Cheng and Roulac (2007) measure the effectiveness of geographical diversification in real estate investment, and Cotter, Gabriel, and Roll (2014) find that increased market integration lowers the diversification potential for housing investment. Market integration in this case is analogous to the macroeconomic component of housing dependence, which is shown in Section 3.2.2. to lower the limiting effect of high geographical concentration on portfolio risk. This reduction in the ability to diversify risk follows from the lesser potential for diversification that results from the larger correlations.

The remainder of this chapter is structured as follows: Section 3.1 describes the rank-size relationships and how they are estimated, Section 3.2 relates these distributions with portfolio variance and the subsequent implications for diversification, Section 3.3 applies these models to the BKFS mortgage data, and finally, Section 3.4 provides a brief discussion and concludes.

3.1. Rank-Size Relations

The rank-size rule, or Zipf's Law, states that for a ranked set of observations, the size of a given observation (c_i) is inversely proportional to its rank i.

$$c_i \propto \frac{1}{i} \tag{42}$$

In the context of city populations, this suggests that the size of the ith ranked city will be equal to the population of the largest city (c_1) divided by its rank. For the U.S., New York City is the top ranked (largest) city with nearly 20 million people within its metropolitan area as of 2018. The traditional rank-size rule would then suggest that the second ranked city, Los Angeles, should have a population of approximately 10 million. However, this simple estimate leaves much room for improvement since the actual census estimates suggest more than 13 million in 2018.

One way to generalize this rank-size relation is to allow for different rates of geometric decline in the sequence of ranked sizes. This is done by introducing a shape parameter, α, which governs how quickly the sequence declines.

$$c_i \propto \frac{1}{i^\alpha} \tag{43}$$

The shape parameter, α, is effectively a measure of the degree of concentration in a variable. For example, the special case of $\alpha = 0$ produces a result where every observation has equal size. Since $i^0 = 1$, the ranked observations are proportional to a constant, and thus $c_1 = c_2 = \cdots = c_n$. As a result of the ranking procedure, this case acts as a lower bound and any amount of geographical concentration will produce larger estimates of α.

The traditional rank-size rule refers to the scenario where $\alpha = 1$. In this case, the second ranked city is half the size of the top ranked city, the third ranked city is one-third the size of the top ranked city, and so on. If one factors out the size of the top ranked city, what remains is the following sequence: $1/1, 1/2, 1/3, \ldots, 1/n$. If one views this sequence as a set of weights that sum to one, this normalized set of weights $r(x)$ can define a statistical distribution where (44) is the probability mass function.

$$r(x) = \frac{1/x}{1 + 1/2 + 1/3, \ldots, 1/n} \tag{44}$$

In its generalized form, this is known as the Zipf distribution, which has a probability mass function given by (45), where the scaling factor $H_{n,\alpha}$ is a generalized harmonic sum, as in (46).

$$f(x; n, \alpha) = x^{-\alpha} H_{n,\alpha}^{-1} \tag{45}$$

$$H_{n,\alpha} = \sum_{i=1}^{n} i^{-\alpha} \tag{46}$$

The Zipf distribution has some connections to a few other statistical distributions. For example, the Pareto distribution has the same general notion; however, it is defined for continuous variables as opposed to the discrete case of the Zipf distribution. Another related distribution is the zeta distribution, which is the limit of the Zipf distribution as $n \to \infty$. In this case, the sum in (46) becomes infinite, which is known as the Riemann zeta function. An interesting property of this function is the convergence of the Riemann zeta function when $\alpha > 1$. This property will be explored further in Section 3.2; however, the general consequence of this convergence is the existence of an asymptotic bound on the amount of diversification that one can obtain. In other words, at a certain point, including an additional city in the portfolio does not provide any further benefit in regards to lowering portfolio risk.

To provide some additional intuition regarding the concentration parameter, α, and the implied weights from the Zipf distribution, consider a scenario where $n = 400$ and $\alpha = \{0, 1, 2\}$. These cases of α respectively refer to the equal-weighted case, traditional rank-size rule, and a convergent case of the Zipf distribution.

In the equal-weighted case ($\alpha = 0$), the weight assigned to the top ranked site is simply $1/n$ or 0.25%, which is the same for all cities. Alternatively, with the traditional rank-size rule ($\alpha = 1$), the top ranked site receives a weight of 15.2%, the second rank is 7.6%, and by rank 60, the weight has decayed down to the equal-weighted level of 0.25%. In the highly concentrated, convergent case ($\alpha = 2$), the top ranked site has a 60.9% weight, the second rank has 15.2%, and the remaining 398 sites make up the remaining 23.9%.

If we examine the median ranks that indicate how many of the top sites are needed to account for 50% of the weight, the equal-weighted case simply yields 200.5, which suggests that the top 200 sites have half the weight and the bottom 200 have the remaining half. In the $\alpha = 1$ case, the larger concentration gives the top 15 sites a 50% share and the bottom 385 the other half. With $\alpha = 2$, over 50% weight is given to just the top ranked site and the remaining 399 produce less than half of the concentration (39.1%).

3.1.1. Estimation

For an empirical cross-section of data, estimation of the α parameter effectively produces a measurement of the degree of concentration in the variable. This is achieved by fitting the log-linear regression model in (47) where the ordinary least squares coefficient for β_1 is an estimate of $-\alpha$. The variable of interest, x, is ranked from largest to smallest and shifted by 1/2, $r = rank - 1/2$. This rank shift follows from Gabaix and Ibragimov (2011) to reduce small sample bias. Thus, in regards to city populations, the top rank is 0.5 (New York City), followed by 1.5 (Los Angeles), and so on.

$$\ln(x) = \beta_0 + \beta_1 \ln(r) + \varepsilon \tag{47}$$

As an example of the fit provided by the Zipf distribution, Figure 3.1 depicts the fitted curve for the populations across the 50 largest CBSAs from the 2018 intercensal estimates from U.S. Census. With an adjusted R^2 of 0.953 and estimated slope of -0.662, this demonstrates a relatively close fit for a mild degree of concentration among the top 50 metropolitan areas. Since this estimate of α is less than one, there does not appear to be any major concerns regarding diversification within this subset of the top 50 CBSAs.

An observation that one may have regarding this linear fit is the autocorrelation in the residuals. Gabaix (2009) notes that this positive autocorrelation follows from the ranking procedure, and that as a result, the typical OLS standard errors are incorrect. To address this, standard errors that are presented throughout the paper are computed across 10,000 bootstrap iterations.

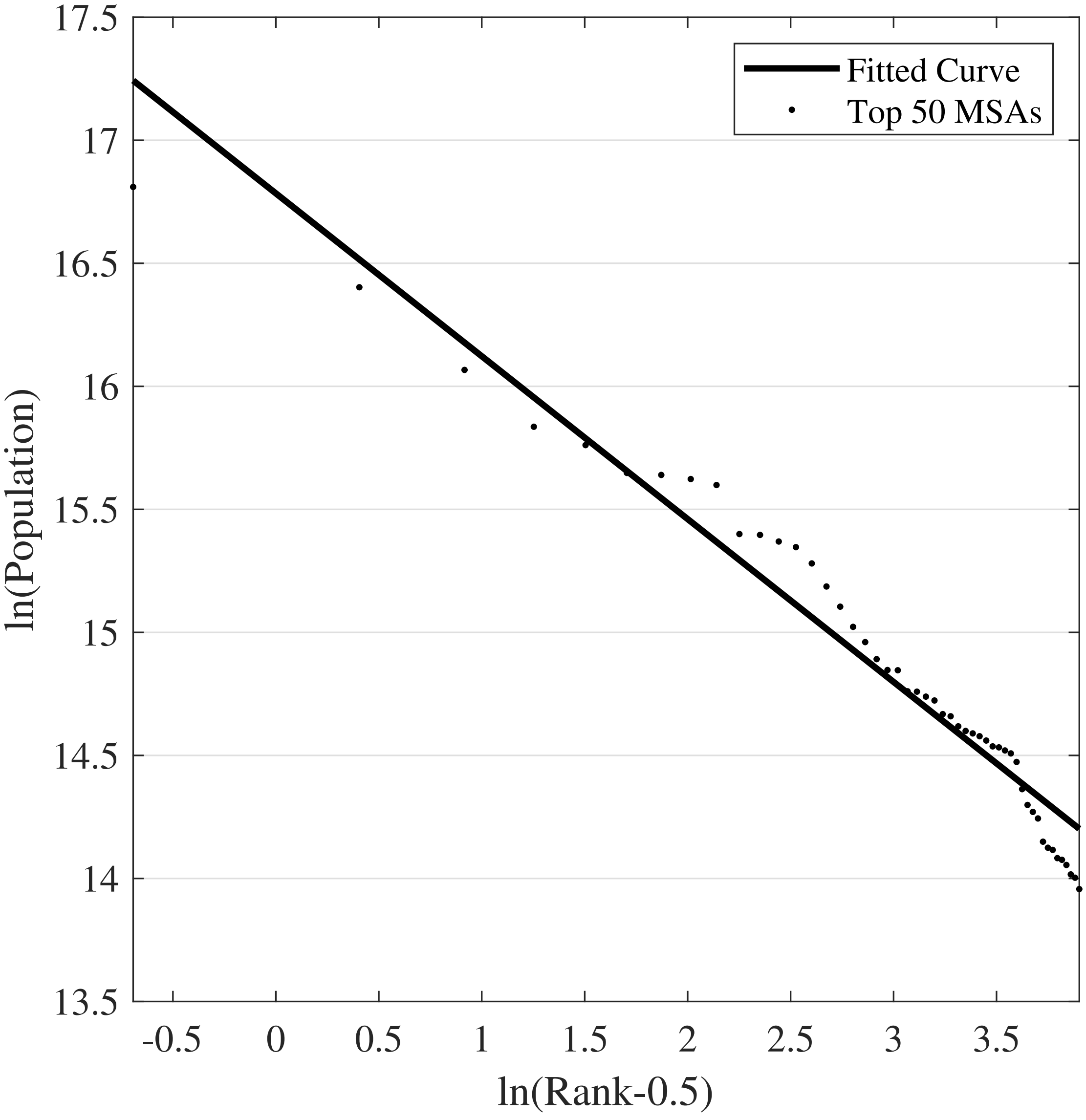

Figure 3.1. Linear Rank-Size Relation for the 50 Largest MSAs by Population

Note: Fitted linear rank-size relation for the populations across the 50 largest MSAs in the 2018 intercensal estimates from the U.S. Census. The estimated slope parameter suggests $\alpha = 0.662$ for the Zipf distribution in (45), and produces an adjusted R^2 of 0.953.

The estimation of the rank-size regression model has been studied for quite some time with a variety of approaches. For example, Nishiyama, Osada, and Sato (2008) recommend a trimmed OLS procedure, which removes an optimal number of the top ranked sites to reduce bias when testing for the traditional rank-size rule. However, rather than testing the hypothesis of $\alpha = 1$, this paper aims to simplify the concentration risk of mortgage debt and compare across different section, which would be incomplete without including the largest markets.

Another approach to the issue of autocorrelation and bias is to expand the functional form of the regression to allow for non-linearity in the fitted curve. The motivation for this is emphasized when expanding to the full set of 945 U.S. CBSAs, which are plotted along with their linear fit in Figure 3.2. From this figure, the non-linearity of this relationship is apparent with the top ranked metros being vastly overestimated by the linear approximation.

This result appears to contrast with the case of country-level populations, in which Laherrère and Sornette (1998) find that China and India appear as outliers while the remaining countries fit a straight line with an R^2 of 0.995. This phenomenon, which they term as a "king effect," also appears in cases such as the populations of French cities, where Paris is the "king" or underestimated outlier in the rank-size regression. Unlike these cases, the scenario with U.S. city populations shows gradual overestimation when moving from the well-fitting mid-section of the curve to the top ranked cities. This pattern suggests that the linear relationship can be improved upon by introducing an additional parameter to correct for the non-linearity apparent in the empirical data.

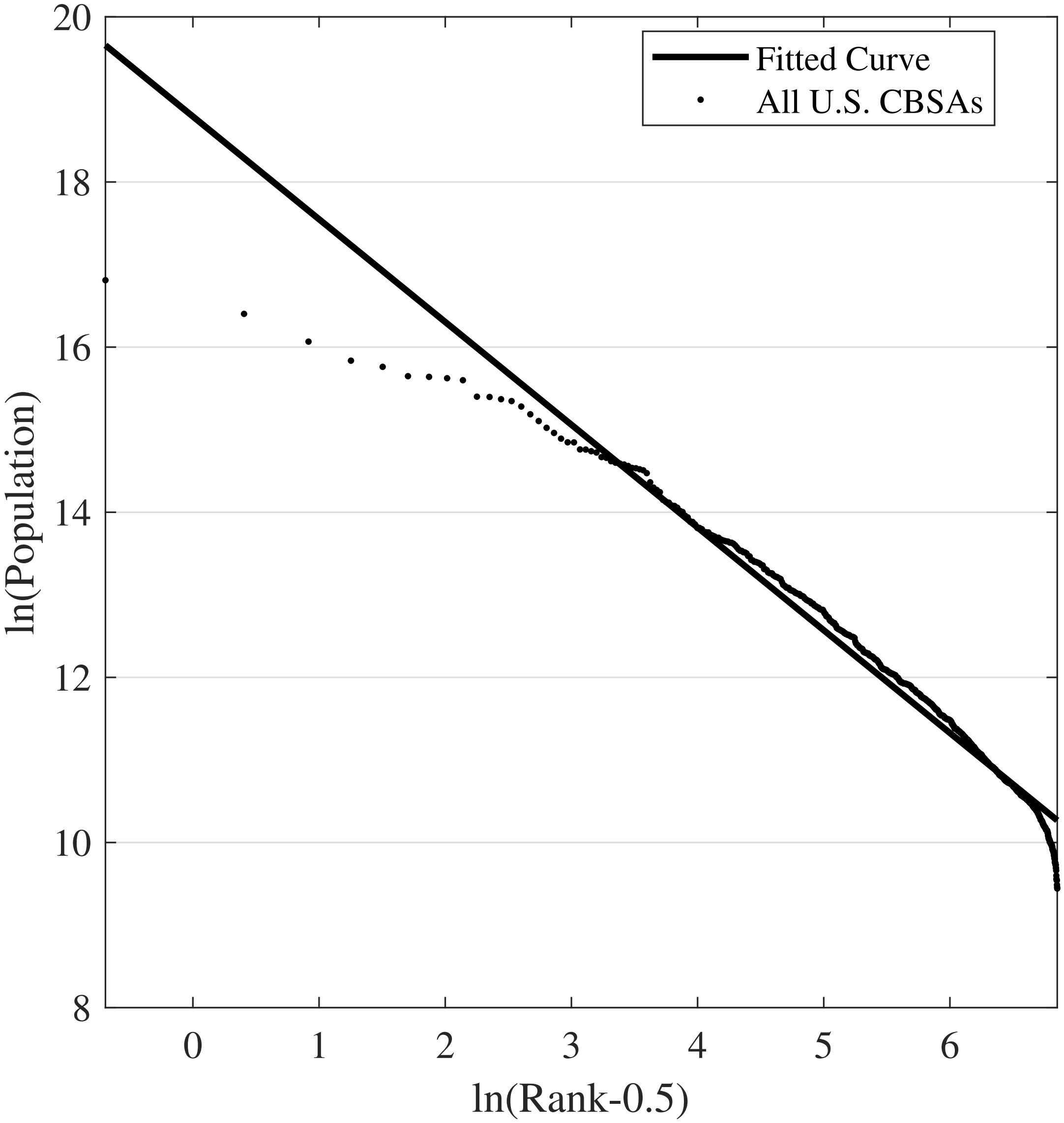

Figure 3.2. Linear Rank-Size Relation for All U.S. CBSA Populations

Note: Fitted linear rank-size relation for the populations across all 945 U.S. CBSAs in the 2018 intercensal estimates from the U.S. Census. The estimated slope parameter suggests $\alpha = 1.244$ for the Zipf distribution in (45), and produces an adjusted R^2 of 0.968.

3.1.2. Parabolic Fractal Distribution

One such non-linear model involves fitting the parabolic fractal distribution (Laherrère, 1996). This distribution expands the log-linear relationship from the Zipf distribution to include a quadratic term for ln(rank), as in (48).

$$\ln(x) = \beta_0 + \beta_1 \ln(r) + \beta_2 \ln^2(r) + \varepsilon \tag{48}$$

This resembles the translog extension of the Cobb-Douglas production function by fitting a second-order polynomial instead of a straight line to the log-log relationship. With this additional term, the model corrects for the non-linearity that arises when examining a more complete set of locations. As seen in the fitted curve for 2018 populations in Figure 3.3, this parabolic fit greatly improves upon the linear fit of the Zipf distribution, particularly around the top ranked sites.

With this additional term, the regression R^2 improves to 0.995 from the 0.968 of the linear Zipf fit from Figure 3.2. However, one drawback of this expanded functional form is the difficulty in comparing the estimates across variables. For example, when comparing the estimates of Zipf's α across the 2010 population counts and 2018 intercensal estimates from Table 3.1, the increase in magnitude of the slope estimate from 1.219 to 1.244 suggests that populations are becoming more concentrated in major metropolitan areas. On the other hand, the parabolic fractal estimates are less clear. The magnitude of the linear coefficient gets smaller (-0.174 to -0.155), and the quadratic coefficient becomes larger in magnitude (-0.107 to -0.112).

One way to address the interpretation issue is to orthogonalize the quadratic term in (48) with the linear term. This is accomplished by regressing $\ln^2(r)$ on $\ln(r)$, as in (49), and using the residuals (50) in place of the quadratic term. At this point, the estimated slope coefficient $\hat{\beta}_1$ will be equal to that of the linear Zipf regression and $\hat{\beta}_2$ will capture the impact of the curvature provided by the quadratic component. To make comparisons regarding the relative effects of each of these components, all of the variables are normalized to have unit variance

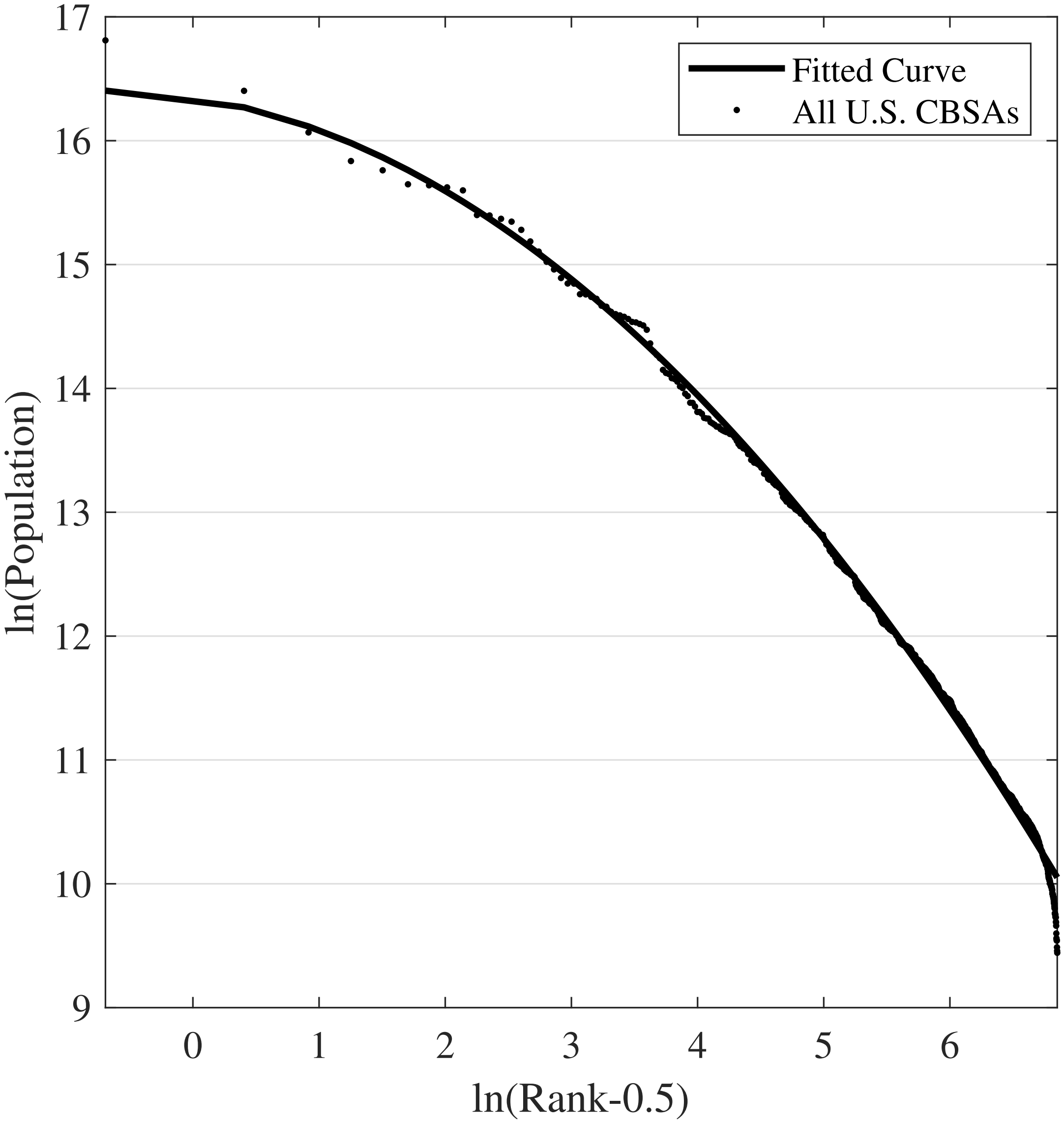

Figure 3.3. Fitted Quadratic Rank-Size Relation for All MSAs

Note: Fitted quadratic rank-size relation for the populations across all 945 U.S. CBSAs in the 2018 intercensal estimates from the U.S. Census. This functional form follows from the parabolic fractal distribution and produces an adjusted R^2 of 0.995.

Table 3.1. Rank-Size Estimates for U.S. CBSA Populations

	2010 Counts		2018 Estimates	
	ZIPF	PFO	ZIPF	PFO
$\hat{\beta}_0$	18.626	15.071	18.793	14.892
s.e.	0.155	0.028	0.162	0.027
$\hat{\beta}_1$	−1.219	−0.984	−1.244	−0.984
s.e.	0.026	0.005	0.027	0.005
$\hat{\beta}_2$	–	−0.162	–	−0.165
s.e.	–	0.005	–	0.004
$\bar{R}^2$	0.969	0.995	0.968	0.995
n	945	945	945	945
p_{50}	2.698	1.995	2.684	1.994
p_{75}	8.047	3.991	7.937	3.976
p_{90}	40.459	10.188	39.445	10.072

Rank-size regressions for the population estimates across all 945 U.S. CBSAs from the 2010 census counts and 2018 intercensal estimates. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 3.1. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, p_{50}, p_{75}, and p_{90} refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.2.

in the regression. The result, denoted by PFO, is given in (51) where σ_x, σ_r, and σ_u are the standard deviations of $\ln(x)$, $\ln(r)$, and $\hat{u}$, respectively.

$$\ln^2(r) = \gamma_0 + \gamma_1 \ln(r) + u \tag{49}$$

$$\hat{u} = \ln^2(r) - \hat{\gamma}_0 - \hat{\gamma}_1 \ln(r) \tag{50}$$

$$\frac{\ln(x)}{\sigma_x} = \beta_0 + \beta_1 \frac{\ln(r)}{\sigma_r} + \beta_2 \frac{\hat{u}}{\sigma_u} + \varepsilon \tag{51}$$

Since these transformations simply orthogonalize and scale the variables, this PFO regression results in the exact same set of predictions as the standard parabolic fractal regression. However, now the coefficients capture the relative impact of each of the independent components. For example, with the 2018 population estimates in Table 3.1, the linear coefficient of −0.984 is identical in effect to the Zipf slope coefficient of −1.244 after the scaling. However, now the $\hat{\beta}_2$ coefficient has some interpretive value.

Since each coefficient represents the standard deviation of the respective component in the fitted model, the coefficient of -0.165 relates to a variance of 0.027 (or 0.165^2). Meanwhile, the standard deviation of 0.984 from the linear component relates to a variance of 0.968, which is also the R^2 from the linear, Zipf model. Mathematically, since the dependent variable is also normalized to have unit variance, the sum of these variances, $0.968 + 0.027 = 0.995$, is equal to the PFO model R^2.

An additional consideration when fitting a parabola is the monotonicity implied by the ranking procedure. For the Zipf distribution, the slope parameter is guaranteed to be non-positive.[10] However, with the parabolic fit, it is possible that the estimates produce a parabola that is increasing for some portion of the domain. This is resolved by imposing two constraints, (53) and (54), which require that the first derivative, (52), be non-positive at both bounds of the domain, $r \in [0.5, n - 0.5]$. Since the derivative is linear in ln(rank), this forces the fitted curve to be non-increasing over the entire domain.

$$d(r) = \frac{d\ln(x)}{d\ln(r)} = \beta_1 + 2\beta_2 \ln(r) \tag{52}$$

$$d(0.5) \leq 0 \implies \beta_1 \leq -2\beta_2 \ln(0.5) \tag{53}$$

$$d(n-0.5) \leq 0 \implies \beta_2 \leq \frac{-\beta_1}{2\ln(n-0.5)} \tag{54}$$

For the orthogonalized variant of the parabolic fractal distribution, the two monotonicity constraints are similar; however, due to the orthogonalization and scaling, a few minor tweaks must be made. The details for the transformed constraints are included in Appendix D.

3.2. Implications for Portfolio Risk

The analysis of risk is an important consideration for both portfolio selection and optimization. In the traditional mean-variance setup of Markowitz (1952), risk is proxied with the variance of the portfolio returns as in (55). The components of this are the portfolio weights, w, and the covariance matrix for asset returns, Σ.

[10] A slope of zero is possible in the equal-weighted case; however, any variation in size will result in a negative slope due to the ranking procedure.

$$\sigma_p^2 = w' \cdot \Sigma \cdot w \tag{55}$$

In regards to geographical concentration, the focus is on modeling the distribution of the w component of this equation, given the return distribution characterized by Σ. The rank-size relations described in the previous section provide a way to simplify this concentration into just a few estimated parameters. To the degree that a portfolio follows similar weights, the fitted relations can be scaled to produce implied portfolio weights. Alternatively, these implied weights can be thought of an the proportional importance of each CBSA to the national market. This section demonstrates how to obtain the implied weights from the fitted rank-size regressions and the implications for the ability to diversify risk when the weights are distributed as such.

3.2.1. Implied Weights

Similar to the derivation of the Zipf distribution in Section 3.1, the fitted sizes from the estimated rank-size regressions must be scaled by their sum to produce the implied weights. From the fitted values of each regression model, the logarithmic functional form typically requires an adjustment to account for the log-normality of the transformed residuals. Using the Zipf distribution as an example, the natural exponential of (47) produces (56), which simplifies to (57) following from the properties of exponentials.

$$x = \exp(\beta_0 + \beta_1 \ln(r) + \varepsilon) \tag{56}$$

$$= \exp(\beta_0) \cdot r^{\beta_1} \cdot \exp(\varepsilon) \tag{57}$$

Thus, when taking the expectation of x given β, the first two terms are fixed and factor out, as in (58). However, if ε is normally distributed, then $\mathbb{E}[\exp(\varepsilon)] = \exp(\sigma^2/2)$, and scales the expected size from the regression.

$$\mathbb{E}[x|\beta] = \exp(\beta_0) \cdot r^{\beta_1} \cdot \mathbb{E}[\exp(\varepsilon)] \tag{58}$$

Although this is important for calculating the expected values for the size of each CBSA, this scaling is equal for each site. As a result, its effect is offset when dividing by the sum to obtain implied weights, as in (59)–(61).

$$
\begin{aligned}
w &= \frac{\mathbb{E}[x|\beta]}{\sum \mathbb{E}[x|\beta]} && (59)\\
&= \frac{\exp(\beta_0) \cdot r^{\beta_1} \cdot \mathbb{E}[\exp(\varepsilon)]}{\sum_{r=1}^{n} \exp(\beta_0) \cdot r^{\beta_1} \cdot \mathbb{E}[\exp(\varepsilon)]} && (60)\\
&= \frac{r^{\beta_1}}{\sum_{r=1}^{n} r^{\beta_1}} && (61)
\end{aligned}
$$

These implied weights in (61) are identical to the probability mass function in (45) defining the Zipf distribution. This same process can be repeated for the implied weights from the parabolic fractal distribution; however, the solution does not simplify as neatly and is left for Appendix E.

3.2.2. Portfolio Variance

With these implied weights from the rank-size regressions, these simplifications can be substituted into the portfolio variance equation in (55). In this framework, the implications of the asset concentration can be examined and compared across the two rank-size models. As a way to compare the fit in terms of portfolio risk, these can also be compared with the empirical weights given by the original data. As will be demonstrated, the parabolic fractal distribution outperforms the linear Zipf regression.

To begin, consider the simple case of independent and identically distributed returns. In this case, the covariance matrix is $\sigma^2 I_n$ where I_n is an n by n identity matrix with ones on the diagonal and zeros for the off-diagonal elements. This leads to a portfolio variance equal to the individual asset variance multiplied by the sum of squared weights, as in (64).

$$
\begin{aligned}
\sigma^2_{iid} &= w'\sigma^2 I w && (62)\\
&= \sigma^2 w'w && (63)\\
&= \sigma^2(w_1^2 + w_2^2 \ldots w_n^2) && (64)
\end{aligned}
$$

For the case of Zipf-distributed weights, the sum of squared weights can be simplified using the harmonic number notation from (46). For each asset i, the implied weight can be written as (65) and its squared value in (66).

$$w(n,\alpha)_i = \frac{i^{-\alpha}}{H_{n,\alpha}} \tag{65}$$

$$w(n,\alpha)_i^2 = \frac{i^{-2\alpha}}{H_{n,\alpha}^2} \tag{66}$$

Substituting this into the i.i.d. portfolio variance from (64), the numerator in (67) is simply another harmonic number, which yields the solution in (68).

$$\sigma_{iid}^2 = \sigma^2 \frac{\sum_{i=1}^n i^{-2\alpha}}{H_{n,\alpha}^2} \tag{67}$$

$$= \sigma^2 \frac{H_{n,2\alpha}}{H_{n,\alpha}^2} \tag{68}$$

In this scenario, the asymptotic properties of harmonic sums demonstrates a potential limiting effect in regards to diversification. As noted back in Section 3.1, when the α parameter is greater than one, the harmonic sum converges as $n \to \infty$. This implies that the denominator in (68) approaches a finite value. Similarly, the numerator also converges and the i.i.d. portfolio variance has an asymptotic bound greater than zero. This suggests a limit to the amount of diversification that one can obtain.

On the other hand, if the concentration is low ($\alpha \leq 1$), the harmonic sum diverges and increases to infinity along with n. Since the denominator continues to increase, the i.i.d. portfolio variance will approach zero for sufficiently large n. This rate of decline is faster for less concentrated asset classes.

For example, Figure 3.4 demonstrates the decline in the i.i.d. portfolio variance as n increases with different concentrations. The equal-weighted portfolio ($\alpha = 0$) diversifies at a rate of $1/n$. The traditional rank-size rule of $\alpha = 1$ declines at a slower rate, but still approaches zero as n increases. The convergent case of $\alpha = 2$ exhibits a much slower rate of decline as well as an asymptotic lower bound of 0.4, which demonstrates the limit on diversification that can be imposed by a highly concentrated market.

Building off of this idea of an asymptotic limit to the level of diversification that one can obtain, another way to compare the implied weights from the rank-size regressions is to examine the number of sites required to reach a certain percentage of the potential diversi-

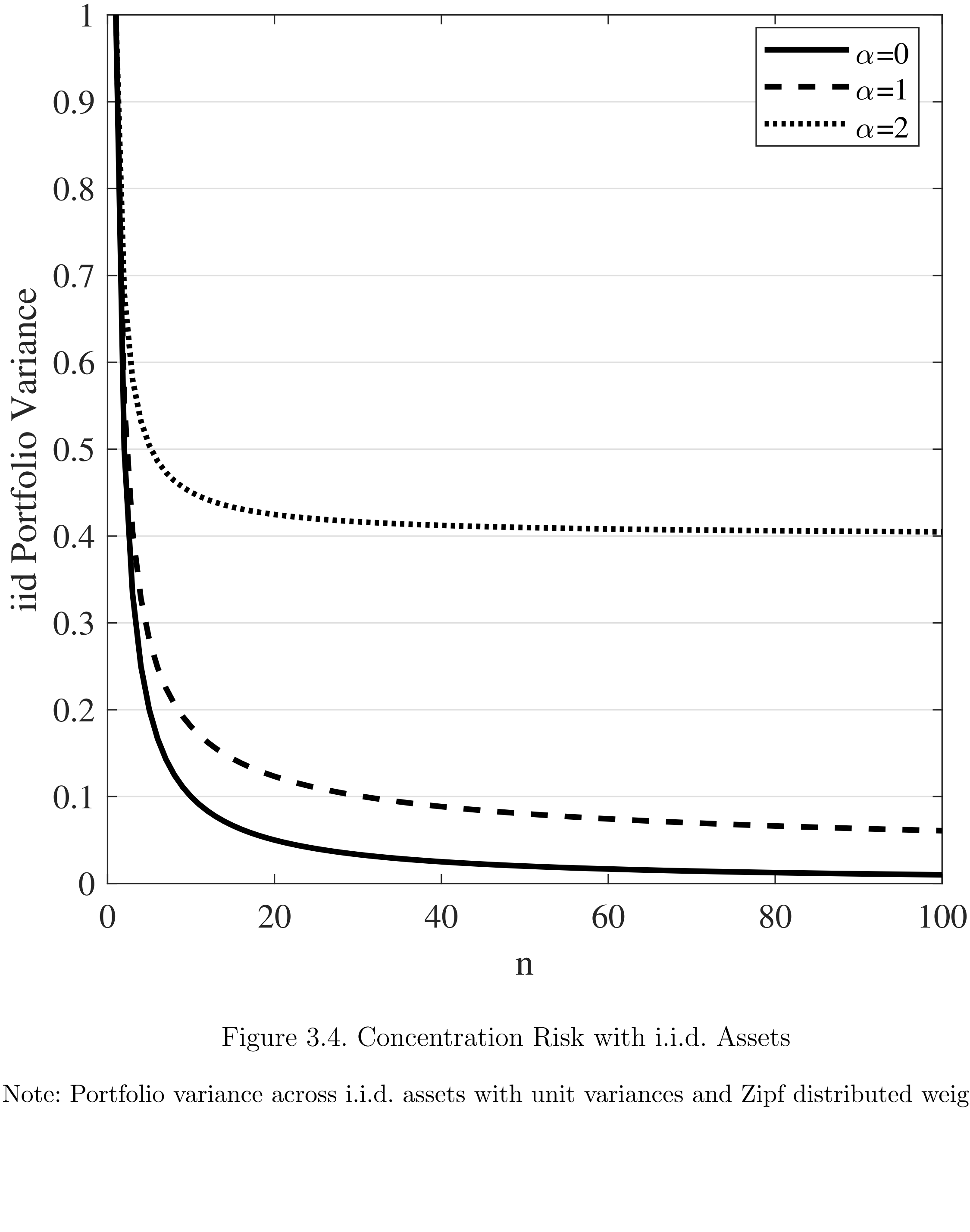

Figure 3.4. Concentration Risk with i.i.d. Assets

Note: Portfolio variance across i.i.d. assets with unit variances and Zipf distributed weights.

fication. For example, since the $\alpha = 2$ case has a lower bound of 0.4 for the i.i.d. variance, 50% diversification would be obtained at the rank where the i.i.d. variance is equal to 0.7.

This involves re-weighting each subset of weights such that they sum one. Using the earlier example of $n = 400$ and $\alpha = 2$, a portfolio with only the top ranked site will re-weight the 0.609 to 1, which produces a variance of 1. Similarly, for a portfolio of the top two sites, the weights of 0.609 and 0.152 re-weight to 0.8 and 0.2. This produces a variance of $0.8^2 + 0.2^2 = 0.68$. Thus, more than 50% of the potential risk reduction is obtained just by including the second ranked site in the portfolio. This rank for this 50% reduction in risk in denoted as p_{50} and is presented along with p_{75} and p_{90} in the bottom three rows of each table underneath the regression estimates.

As a visual aid to understanding these percentile ranks, Figure 3.5 presents a comparison between the linear Zipf weights, the parabolic PFO weights, and the empirical weights for mortgage debt across the 929 CBSAs present in the BKFS dataset.[11] For each curve, the p_{50} rank indicates the number of cities required to lower the i.i.d. portfolio variance 50% of the way to its fully diversified variance. Unlike the theoretical case where $n \to \infty$, this fully diversified variance is simply based off of the full weight vector. As can be seen in the figure, the Zipf distribution's overestimation of the top ranked sites tends to overestimate the limiting factor; however, the PFO model greatly improves on this and traces the empirical weights quite well.

The correlation across asset returns is another critical component for portfolio variance and the ability to diversify risk. For mortgages, this component can be represented as a function of the correlation across the underlying property returns and the default rate of the loans (Dombrowski, Pace, and Narayanan, 2020). Real estate assets demonstrate high levels of both spatial (weak) and macroeconomic (strong) cross-sectional dependence. In regards to

[11] Although there are 945 CBSAs in the U.S. as per the Census, the 16 unmatched CBSAs all correspond with micropolitan areas with less than 50,000 residents. For more details on the BKFS dataset and its coverage, see Section 3.3 and Appendix F.

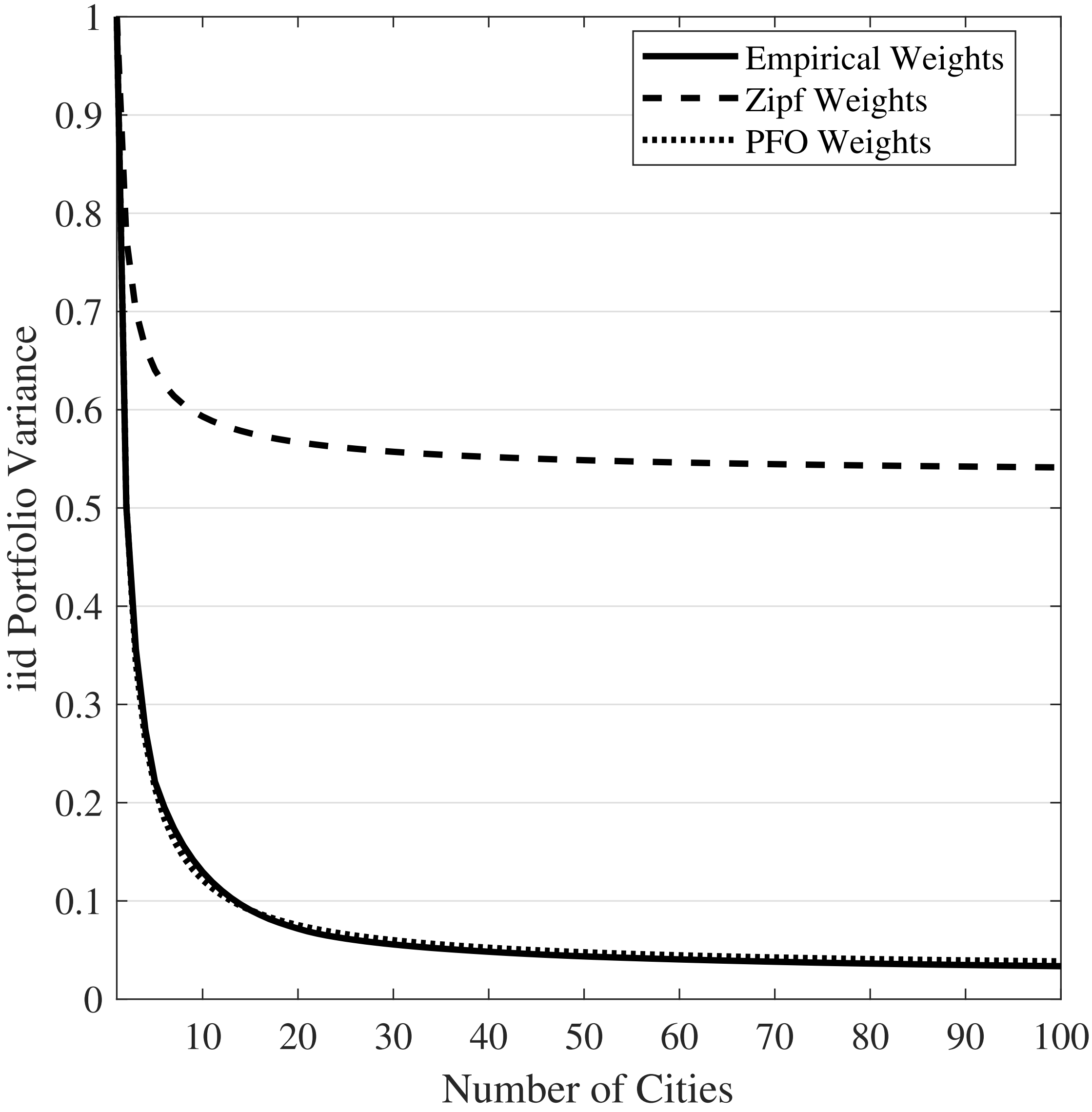

Figure 3.5. Comparison of Rank-Size Estimates with Empirical Weights

Note: Comparison between linear Zipf weights, PFO weights, and empirical weights for i.i.d. portfolio variances as portfolio size increases.

geographical diversification, risk can be reduced by diversifying away the weak dependence; however, the strong dependence remains regardless of portfolio size.

To examine the effect of strong cross-sectional dependence in this context, a correlation of ρ is introduced to the i.i.d. returns and a unit variance is assumed, as in (69). This can be rewritten into matrix form as in (70), where I_n is an n by n identity matrix and ι_n is a vector of ones.

$$V_{ij} = \begin{cases} 1, \text{ if } i = j \\ \rho, \text{ if } i \neq j \end{cases} \tag{69}$$

$$V = (1-\rho)I_n + \rho\iota_n\iota_n' \tag{70}$$

Substituting this correlation matrix into (55), (71) simplifies to (72).

$$\sigma_p^2 = w'Vw \tag{71}$$

$$= (1-\rho)w'w + \rho w'\iota_n\iota_n'w \tag{72}$$

Then (73) follows since both $w'\iota_n$ and $\iota_n'w$ are equal to one since the weight vectors sum to one. Thus, (74) provides the portfolio variance as a function of the macroeconomic risk (ρ) and i.i.d. portfolio variance described previously.

$$= (1-\rho)w'w + \rho \tag{73}$$

$$= (1-\rho)\sigma_{iid}^2 + \rho \tag{74}$$

The effect of this correlation on portfolio variance is a reduction in the potential for diversification. This subsequently lowers the impact of any concentration risk coming from the portfolio weights. For example, if $\rho = 0.3$, this portion of the variance is not diversifiable. The sum of squared weights (or σ_{iid}^2) can still limit the ability to achieve the potential diversification; however, this term is scaled by $(1-\rho)$ or 0.7. More strong dependence (ex. specialized portfolios) diminishes the diversification potential in general and lowers the impact of high concentrations. However, as will be demonstrated in the next section, some sectors of the mortgage market (such as jumbo mortgages) exhibit substantially larger degrees of concentration, which may offset the attenuation from large correlations.

3.3. Empirical Concentration

Now that the previous examples for the rank-size relations have demonstrated the concentration of population across U.S. CBSAs, the focus shifts to the mortgage market. Using a large set of loan-level mortgage data from BKFS, this section models the geographical concentration of various sectors by partitioning the data across a number of dimensions, including GSE-eligibility, lien priority, interest rate type, loan purpose, occupancy status, and documentation level.

The raw BKFS dataset includes a loan table, which provides origination characteristics for more than 173 million loans. Additionally, monthly remittance tables begin in January 1989 and provide updates on loan statuses, balances, and interest rates. After a relatively mild cleaning process,[12] we are left with just over 150 million loans originating between January 1990 and November 2016, which is the most recent month of observation.

For financial variables such as mortgage debt, the concentration of population only reflects part of the equation. In addition to large population centers originating more loans, there is also a pricing differential between major cities and smaller, less urban locations. This suggests that while measurements of population concentration may act as a reasonable baseline for the quantity of loans, the value of the debt is likely to be even more highly concentrated.

This distinction between concentration in quantity vs. concentration of balances is demonstrated in Table 3.2. This table presents the rank-size estimates for both the total number of loans originated over the sample period along with the estimates for the concentration of the aggregate balances. From the Zipf slope estimates, the increase from $\hat{\alpha} = 1.56$ for quantities to $\hat{\alpha} = 1.73$ for balances shows a materially larger degree of concentration for the aggregate debt compared to simply the number of loans. When compared to the 2018 population estimate of $\hat{\alpha} = 1.24$, both mortgage quantities and balances appear to be far

[12] See Appendix F for a detailed account of the data cleaning process.

more heavily concentrated than is suggested by measures of population concentration. Thus, the geographical concentration of mortgage debt is larger than that of the number of loans, which are both larger than the concentration present in populations across the country.

In regards to the orthogonalized parabolic fractal estimates, the additional fit provided by the curvature explains a similar proportion of the loan quantity concentration ($0.966/0.163 = 5.93$) as with the aggregate debt balances ($0.968/0.163 = 5.94$). This suggests that the comparison of the linear components compares similar degrees of variation in the rank-size relationship. The fitted relations for the aggregate mortgage debt are presented in Figures 3.6 and 3.7, respectively for the linear Zipf and quadratic PFO models.

For the diversification percentile ranks, the Zipf estimates suggest that 90% of the potential diversification is obtained by diversifying across the top 21 ranked cities. However, since the overestimation of top ranked sites leads to an overestimated lower bound for the diversification potential (as in Figure 3.5), these ranks are less reliable than those for the PFO model, which more closely tracks the true empirical weights.

For the PFO estimates, the percentile ranks suggest that approximately 50% of the potential diversification is obtained just by including the two largest cities in the portfolio. To attain 75% of the diversification potential, p_{75} suggests this is accomplished with the top four cities. Then for 90% of the diversification potential, the top ten cities reduce the i.i.d. portfolio variance 90% of the way to its fully diversified lower bound.

To narrow the focus and provide some comparisons across different segments of the mortgage market, Tables 3.3–3.9 present the rank-size estimates for various subsets of loans. The first partition that is examined is the conforming vs. jumbo loan markets. One characteristic of the mortgage market that distinguishes it from other debt markets is the prominence of the GSEs, which stimulate demand in the secondary mortgage market by purchasing and securitizing any loans that meet their criteria. One such criterion is the loan amount, which is set by the Federal Housing Finance Agency. Mortgages that fall below this threshold are classified as conforming loans and are eligible for purchase by the GSEs. Loans with bal-

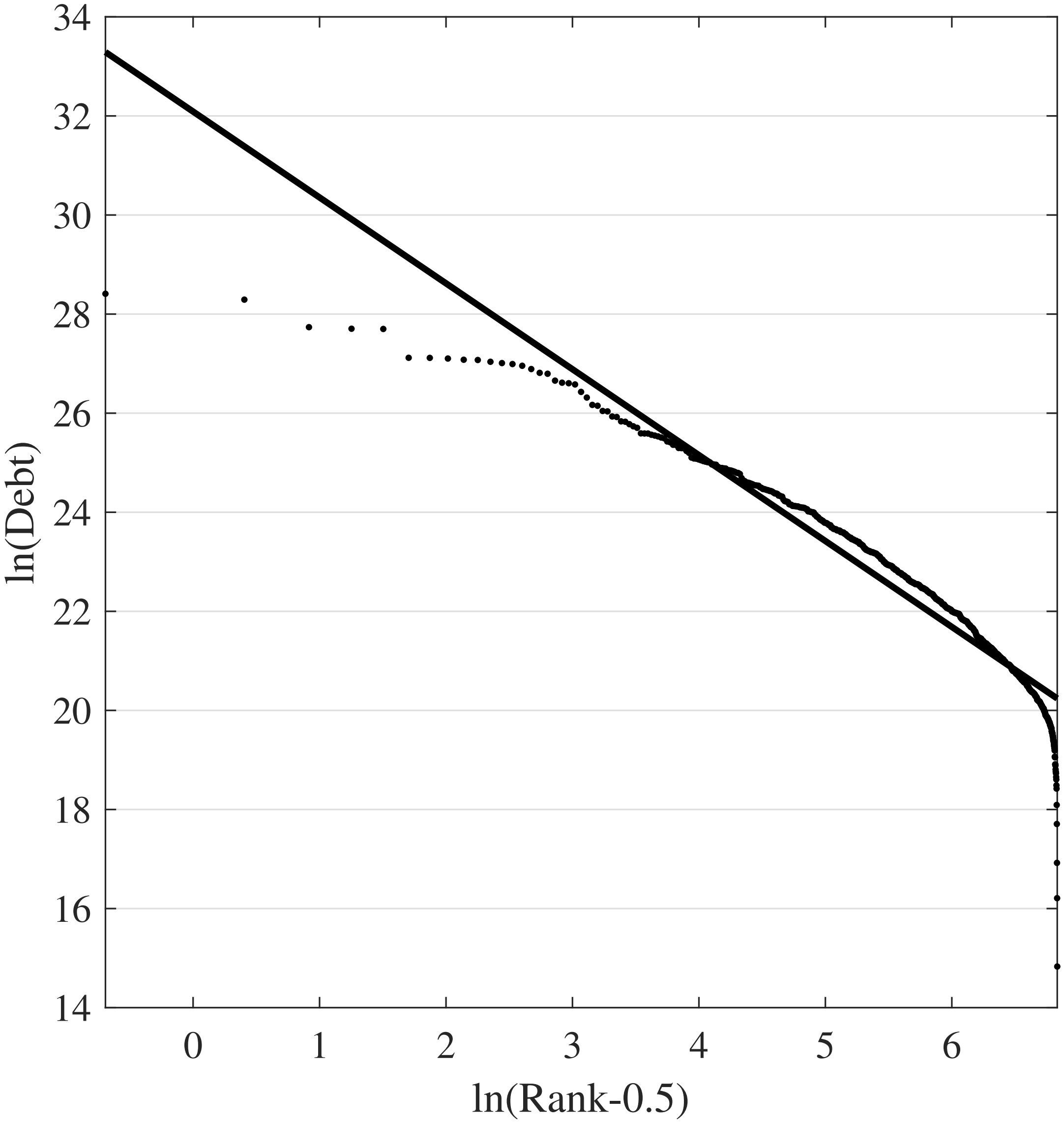

Figure 3.6. Fitted Linear Rank-Size Relation for All Mortgages

Note: Fitted linear rank-size relation for the total mortgage debt originated across 929 U.S. CBSAs from the BKFS dataset. The estimated slope parameter suggests $\alpha = 1.734$ for the Zipf distribution in (45), and produces an adjusted R^2 of 0.914.

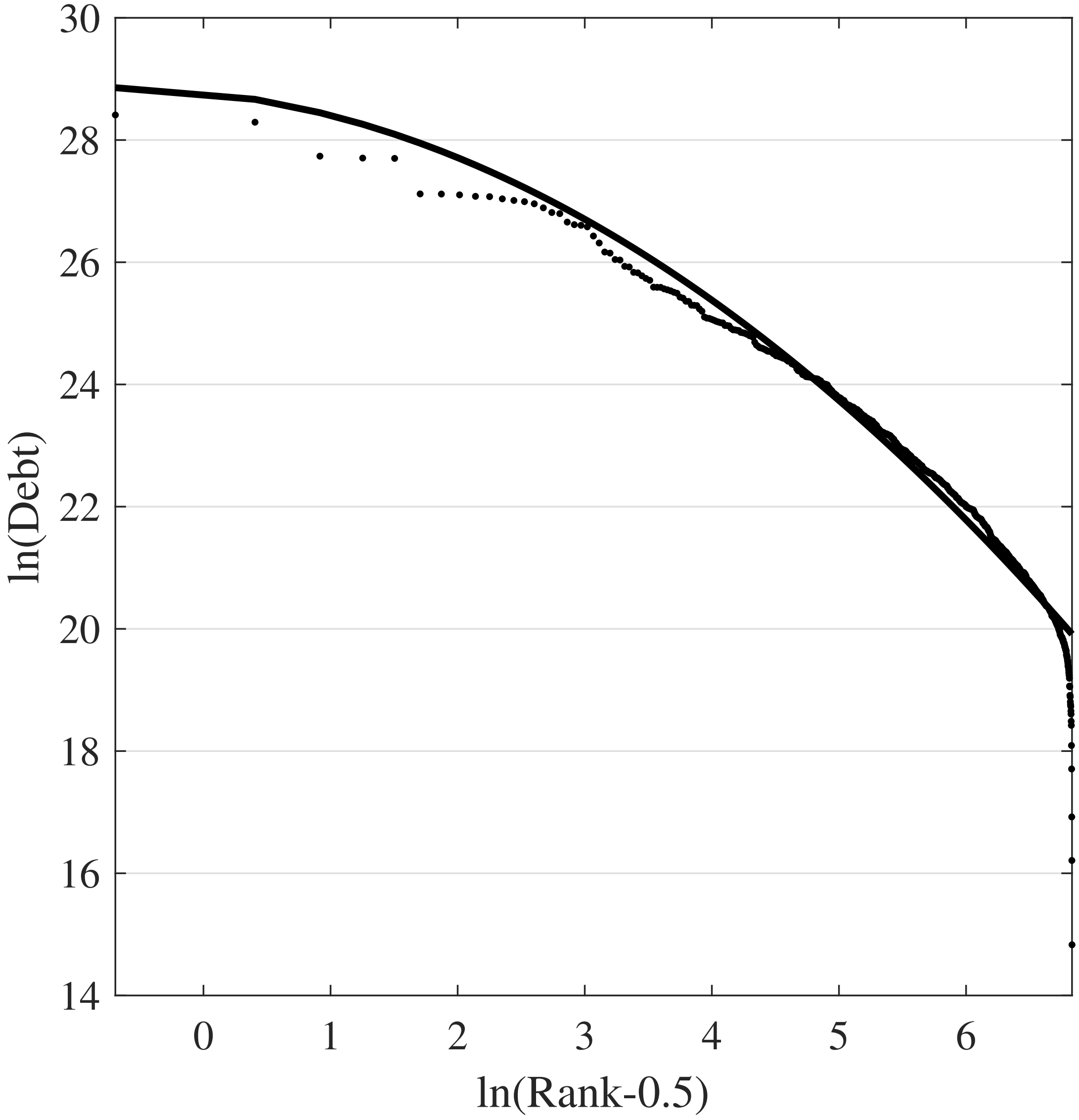

Figure 3.7. Fitted Quadratic Rank-Size Relation for All Mortgages

Note: Fitted quadratic rank-size relation for the total mortgage debt originated across 929 U.S. CBSAs from the BKFS dataset. This expanded functional form follows from the parabolic fractal distribution in Section 3.1.2 and produces an adjusted R^2 of 0.963.

Table 3.2. Rank-Size Estimates for All Mortgages

	Loan Quantities		Balances	
	ZIPF	PFO	ZIPF	PFO
$\hat{\beta}_0$	19.375	11.992	32.088	17.808
s.e.	0.270	0.091	0.285	0.101
$\hat{\beta}_1$	-1.555	-0.966	-1.734	-0.968
s.e.	0.045	0.016	0.048	0.018
$\hat{\beta}_2$	–	-0.163	–	-0.163
s.e.	–	0.024	–	0.026
$\bar{R}^2$	0.912	0.960	0.914	0.963
n	929	929	929	929
p_{50}	2.353	1.985	2.110	1.978
p_{75}	5.690	3.916	4.604	3.869
p_{90}	20.825	9.705	13.727	9.398

Rank-size regressions for the total quantities and balances of mortgage originations across 929 U.S. CBSAs over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 3.1. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, p_{50}, p_{75}, and p_{90} refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.2.

ances above this limit are classified jumbo loans and are often held as portfolio loans on bank balance sheets or packaged into private-label mortgage-backed securities (MBS).

Table 3.3 examines the non-conforming, jumbo loan market, which suggests drastically larger degrees of concentration ($\hat{\alpha} = 2.53$ for loan quantities and 2.58 for aggregate debt). On the other hand, conforming loans (Table 3.4) are far less concentrated with $\hat{\alpha} = 1.54$ and 1.69, respectively for quantities and balances. Since the less concentrated, conforming loans are purchased and securitized by the GSEs, this would suggest that the loans held in bank portfolios or in private-label MBS tend to be more highly concentrated than those that are in the GSE securities. In regards to the diversification percentile ranks, the conforming loan market estimates appear fairly similar to those for the full mortgage market with p_{50}, p_{75}, and p_{90} respectively equal to 2, 4, and 10. However, for the jumbo market, p_{75} reduces to 3.5 and p_{90} falls to approximately 7.7.

Table 3.3. Rank-Size Estimates for Jumbo Mortgages

	Loan Quantities		Balances	
	ZIPF	PFO	ZIPF	PFO
$\hat{\beta}_0$	19.954	7.598	33.380	12.362
s.e.	0.461	0.150	0.481	0.163
$\hat{\beta}_1$	−2.534	−0.967	−2.579	−0.963
s.e.	0.077	0.026	0.081	0.028
$\hat{\beta}_2$	–	−0.163	–	−0.162
s.e.	–	0.040	–	0.041
$\bar{R}^2$	0.901	0.962	0.891	0.954
n	913	913	913	913
p_{50}	1.770	1.939	1.761	1.936
p_{75}	2.621	3.564	2.574	3.544
p_{90}	4.595	7.724	4.437	7.636

Rank-size regressions for the total quantities and balances of jumbo mortgage originations across U.S. CBSAs over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 3.1. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, p_{50}, p_{75}, and p_{90} refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.2.

Another separation for the mortgage market is the lien priority. In event of foreclosure, the priority of the mortgage lien indicates the riskiness for the debtholder to recover some of the losses. After foreclosure sales, which tend to be at a substantial discount (Clauretie and Daneshvary, 2009), first-lien debtholders are paid down prior to any recovery for junior liens.

In Table 3.5, the left columns model the concentration of first-lien mortgages. These are the vast majority of the loans with approximately 96% of all loans and 99% of the aggregate debt. The first-lien debt demonstrates similar results to the full set of mortgages in Table 3.2. This is contrasted with junior-lien mortgages (right columns of Table 3.5) that have only a residual claim to recovery in event of a default. The concentration estimate of 1.83 suggests a mildly larger degree of concentration for this smaller market; however, this does not appear to be statistically significant given the standard errors.

Table 3.4. Rank-Size Estimates for Conforming Mortgages

	Loan Quantities		Balances	
	ZIPF	PFO	ZIPF	PFO
$\hat{\beta}_0$	19.276	12.021	31.774	18.027
s.e.	0.272	0.090	0.296	0.098
$\hat{\beta}_1$	-1.541	-0.965	-1.691	-0.966
s.e.	0.046	0.016	0.050	0.017
$\hat{\beta}_2$	–	-0.162	–	-0.163
s.e.	–	0.024	–	0.026
$\bar{R}^2$	0.909	0.958	0.908	0.959
n	929	929	929	929
p_{50}	2.372	1.985	2.169	1.980
p_{75}	5.783	3.919	4.818	3.880
p_{90}	21.544	9.724	15.095	9.474

Rank-size regressions for the total quantities and balances of conforming mortgage originations across U.S. CBSAs over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 3.1. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, p_{50}, p_{75}, and p_{90} refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.2.

As with the previous results for the PFO model, both the first-lien and junior-lien partitions exhibit similar relative effects between the linear portion of the fit and the correction for the non-linearity. For the first-lien subset, the relative effect is nearly identical to the full market ($0.968/0.163 = 5.94$). Then for the smaller junior-lien market, the explanatory power is slightly lower; however, the relative effect of the two components ($0.958/0.161 = 5.95$) is still fairly constant.

In regards to the percentile ranks, the results are both quantitatively similar to those for the full mortgage market. The number of cities needed to obtain 50%, 75%, and 90% shares of the diversification potential remain in a similar range around 2, 4, and 9, respectively in the PFO model.

The type of interest rate for a mortgage is another loan characteristic that appears to suggest some variation around the degree of geographical concentration. In Table 3.6, the

Table 3.5. Rank-Size Estimates by Lien Priority

	First-Lien		Junior-Lien	
	ZIPF	PFO	ZIPF	PFO
$\hat{\beta}_0$	32.075	17.805	27.999	14.609
s.e.	0.288	0.101	0.313	0.117
$\hat{\beta}_1$	-1.733	-0.968	-1.825	-0.958
s.e.	0.048	0.018	0.053	0.021
$\hat{\beta}_2$	–	-0.163	–	-0.161
s.e.	–	0.027	–	0.028
$\bar{R}^2$	0.914	0.963	0.893	0.943
n	929	929	929	929
p_{50}	2.172	1.978	1.997	1.974
p_{75}	4.830	3.869	4.170	3.840
p_{90}	15.186	9.399	11.466	9.211

Rank-size regressions for the balances of mortgage originations partitioned by lien priority over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 3.1. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, p_{50}, p_{75}, and p_{90} refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.2.

estimated concentration for adjustable rate mortgages ($\hat{\alpha} = 2.07$) is quite larger than that of the fixed rate mortgage market ($\hat{\alpha} = 1.69$). Similar to previous results, the PFO model suggests consistent improvements to the linear fit of the Zipf distribution and the estimated percentile ranks are also unchanged.

When comparing mortgages that are originated for new purchases and refinancing activity (Table 3.7), both subsets suggest similar degrees of geographical concentration, which are also similar to the full mortgage market.

In Table 3.8, the concentration of investment properties ($\hat{\alpha} = 2.08$) appears to be larger than for owner-occupied properties ($\hat{\alpha} = 1.71$). The estimates for the investment property subset do tend to carry slightly less explanatory power in the rank-size regressions; however, as with the junior-lien market, if the relative effect of the two components are compared (0.955/0.161), then the proportional importance of each component is shown to be constant.

Lastly, the geographical concentration between loans with full documentation are compared to those with less than full documentation. In Table 3.9, the estimates for the fully documented mortgage market suggest only slightly less concentration than the non-fully documented loans. On the other dimensions of comparison, such as the relative effects in the PFO model and the percentile ranks, these estimates also produce fairly similar results.

Table 3.6. Rank-Size Estimates by Interest Rate Type

	Fixed Rate		Adjustable Rate	
	ZIPF	PFO	ZIPF	PFO
$\hat{\beta}_0$	31.699	18.037	31.672	14.783
s.e.	0.284	0.099	0.350	0.117
$\hat{\beta}_1$	-1.688	-0.967	-2.072	-0.972
s.e.	0.048	0.018	0.059	0.020
$\hat{\beta}_2$	–	-0.163	–	-0.164
s.e.	–	0.026	–	0.032
$\bar{R}^2$	0.911	0.961	0.922	0.972
n	929	929	927	927
p_{50}	2.172	1.980	1.899	1.963
p_{75}	4.830	3.881	3.369	3.756
p_{90}	15.186	9.481	7.512	8.713

Rank-size regressions for the balances of mortgage originations partitioned by interest rate type over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 3.1. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, p_{50}, p_{75}, and p_{90} refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.2.

3.4. Discussion

This paper characterizes the geographical concentration of the mortgage market using the empirical regularity from regional science known as rank-size rule. This allows for the simplification of a set of portfolio weights into just one or two parameters that measure the degree of concentration. Rank-size distributions such as the Zipf or parabolic fractal distribution demonstrate how high levels of geographical concentration can impose limits on the ability to achieve diversify a portfolio of mortgages.

Table 3.7. Rank-Size Estimates by Loan Purpose

	New Purchases		Refinances	
	ZIPF	PFO	ZIPF	PFO
$\hat{\beta}_0$	31.074	17.372	31.214	17.042
s.e.	0.297	0.098	0.287	0.110
$\hat{\beta}_1$	-1.724	-0.969	-1.755	-0.964
s.e.	0.050	0.017	0.048	0.020
$\hat{\beta}_2$	–	-0.163	–	-0.162
s.e.	–	0.027	–	0.027
$\bar{R}^2$	0.916	0.966	0.906	0.955
n	929	929	929	929
p_{50}	2.124	1.979	2.081	1.977
p_{75}	4.654	3.871	4.497	3.862
p_{90}	14.022	9.416	13.102	9.356

Rank-size regressions for the balances of mortgage originations partitioned by loan purpose over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 3.1. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, p_{50}, p_{75}, and p_{90} refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.2.

The linear relationship between ln(size) and ln(rank) suggested by the Zipf distribution is extended to allow for non-linearity by fitting the parabolic fractal distribution along with an orthogonalized variant, which isolates the effect of the curvature. This extension helps correct for the overestimation of the top ranked cities evident from the linear fit of the Zipf distribution and provides more accurate predictions for the degree of concentration in mortgage debt.

The application of these rank-size relations to data from BKFS suggests considerable degrees of concentration in the mortgage market. When compared to the estimate of Zipf's α for populations (1.24 in 2018), the quantity of mortgage originations produces a materially larger estimate of 1.56. Taking into account the higher property values of these large metros, this concentration estimate increases to 1.73 for the aggregate balances of the debt originations.

Table 3.8. Rank-Size Estimates by Occupancy Status

	Owner-Occupied		Investment	
	ZIPF	PFO	ZIPF	PFO
$\hat{\beta}_0$	31.577	17.798	28.505	12.915
s.e.	0.276	0.101	0.414	0.150
$\hat{\beta}_1$	−1.713	−0.970	−2.084	−0.955
s.e.	0.046	0.018	0.069	0.026
$\hat{\beta}_2$	–	−0.163	–	−0.161
s.e.	–	0.027	–	0.033
$\bar{R}^2$	0.921	0.967	0.868	0.938
n	929	929	928	928
p_{50}	2.139	1.979	1.895	1.962
p_{75}	4.708	3.875	3.337	3.743
p_{90}	14.377	9.443	7.388	8.641

Rank-size regressions for the balances of mortgage originations partitioned by occupancy status over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 3.1. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, p_{50}, p_{75}, and p_{90} refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.2.

This dataset also allows for examination of specific sectors of the mortgage market to examine the differential degrees of geographical concentration. One such sector that exhibits a substantial degree of concentration is the jumbo loan market. With α estimates of approximately 2.5, this suggests that these loans tend to be heavily clustered in just a few large markets. Since these loans are not eligible for purchase by the GSEs, these jumbo loans tend to either be held as portfolio loans or consolidated into private-label mortgage-backed securities. Thus, the risk of the private market is exacerbated by the geographically disperse nature of the conforming loan market.

In addition to the jumbo mortgage market, several other sectors of the mortgage market exhibit relatively large degrees of geographical concentration, albeit to a lesser degree. For example, mortgages secured by investment properties are substantially more concentrated when compared to owner-occupied properties. Similarly, junior-lien mortgages are slightly

Table 3.9. Rank-Size Estimates by Documentation Type

	Full Doc		Non-Full Doc	
	ZIPF	PFO	ZIPF	PFO
$\hat{\beta}_0$	30.736	17.258	31.821	17.487
s.e.	0.283	0.108	0.294	0.101
$\hat{\beta}_1$	-1.702	-0.961	-1.753	-0.969
s.e.	0.048	0.020	0.049	0.018
$\hat{\beta}_2$	–	-0.162	–	-0.163
s.e.	–	0.026	–	0.027
$\bar{R}^2$	0.901	0.950	0.916	0.966
n	928	928	929	929
p_{50}	2.154	1.979	2.083	1.977
p_{75}	4.763	3.878	4.506	3.863
p_{90}	14.731	9.457	13.150	9.359

Rank-size regressions for the balances of mortgage originations partitioned by documentation status over the period from 1990–2016. ZIPF and PFO refer to the linear Zipf and orthogonalized parabolic fractal models described in Section 3.1. Standard errors for the parameter estimates are computed across 10,000 bootstrap iterations. The percentile ranks, p_{50}, p_{75}, and p_{90} refer to the number of CBSAs needed to produce the respective percentages of diversification, as described in Section 3.2.2.

more concentrated than their first-lien counterparts, and adjustable rate mortgages show slightly more concentration than fixed rate mortgages.

With these large degrees of geographical concentration in mortgage debt, portfolios constructed from these assets are likely to experience limits to the amount of diversification that can be obtained. The Zipf distribution and the convergence of the generalized harmonic sum allow for a rigorous demonstration for how such limits to geographical diversification can lead to a lower bound on portfolio risk. As a result, the relative weights for various cities demonstrates how local economic shocks for some cities remain local, but others can propagate through to global shocks for highly concentrated markets. For example, a local shock to California is effectively a global shock to the jumbo mortgage market.

Chapter 4. Imputing Dynamics in Mortgage Default Models

In traditional analysis of mortgage default, an unaddressed issue is that many predictors of default are only observable at the time of loan origination. Some predictors, such as interest rate type and whether a borrower provides full documentation, are characteristics of a mortgage and have no time-varying attributes. However, some other important variables do change over time. Examples of such dynamic variables include property values, borrower income, wealth, and credit scores.

For property values, a common solution is to use house price indices (HPIs) to capture dynamics at a macro-level. However, factors like borrower income, wealth, and credit scores are likely to be driven more by borrower-specific factors, so this approach may not be as effective. Although some studies simply use the estimated house price dynamics as independent variables (Archer, Elmer, Harrison, and Ling, 2002; Ambrose, Conklin, and Yoshida, 2016), many others construct a measure of equity in the form of a current loan-to-value ratio (CLTV).[13]

Although some of these models (Deng et al., 2000; Archer et al., 2002; Foote et al., 2008; Bajari et al., 2008; Ambrose et al., 2016; Bhutta et al., 2017) include aggregated dynamic variables such as divorce and unemployment rates, for many of these models, the only borrower-specific variable that varies over time is the CLTV ratio. However, other variables (such as borrowers' credit scores) also vary over the life of the loan, but standard practice is to use the credit score at origination (Bajari et al., 2008; Mayer et al., 2009; McCollum et al., 2015; Ambrose et al., 2016; Bhutta et al., 2017). If credit scores or any other omitted dynamic predictors vary systematically with house prices or any other included regressors, then the traditional approach could create a bias by omitting these dynamics.

Another potential omission in traditional mortgage default models is the inclusion of borrower-specific fixed effects. The assumption of exogeneity requires that any unobserved

[13] See Deng, Quigley, and Order (2000); Foote, Gerardi, and Willen (2008); Bajari, Chu, and Park (2008); Mayer, Pence, and Sherlund (2009); Campbell and Cocco (2015); McCollum, Lee, and Pace (2015); Bhutta, Dokko, and Shan (2017).

heterogeneity be uncorrelated with the included regressors. If the cross-sectional variation of this heterogeneity is related with any predictor variables, this will also lead to biased coefficients.

For example, if a borrowers' income is not observed (as in our dataset), its variation will be captured by the error term. Thus, if borrower income is systematically related with any other predictors, the resulting estimates will be biased. Although some studies, such as Bajari et al. (2008), use debt-to-income ratios to infer borrower income at origination, and other studies use regional income measures (Ambrose et al., 2016; Bhutta et al., 2017) the potential issue of omitting income dynamics arises.

Another component of this unobserved heterogeneity is the variation in borrowers' attitudes towards default. A survey by Guiso, Sapienza, and Zingales (2013) finds that views regarding the morality of default are relevant when determining the willingness of individuals to strategically default. Such attitudes regarding the morality of default may also be critical omissions if certain classes of borrowers are more likely to have specific loan characteristics, such as full documentation or adjustable rate mortgages.

If all other factors driving the default decision are controlled for in the model, then including borrower-specific intercepts will capture the variation around these attitudes surrounding default. Deng et al. (2000) find evidence of statistically significant heterogeneity among borrowers and suggests that its omission leads to errors in estimation of prepayment behavior. This supports the idea that there are distinctions between borrowers relating to innate attitudes towards default.

The goal of this paper is to (1) impute unobserved heterogeneity and predictor dynamics, (2) determine the effectiveness of this methodology using cross-validation, and (3) evaluate the resulting influence on the importance of the traditional variables, such as CLTV, in the default decision. Using a ridge regression framework, we are able to estimate an econometric model to capture these effects through a specification of the model fit, which is tuned to maximize out-of-sample performance.

We use residential mortgage data provided by Blackbox Logic LLC along with HPIs from the S&P/Case-Shiller 20-City Composite Home Price Index. This allows us to construct a sample of nearly 90 million borrower-month observations spanning over 2 million residential loans for properties in the 20 CS-MSAs originating in 2003–2014.

Based on this sample, we find the largest difference between the analysis proposed here and the conventional approach is concentrated in the California MSAs of our sample. The estimates of the full documentation and CLTV parameters increase in magnitude when imputing static heterogeneity compared to the estimates from a traditional OLS model. When imputing unobserved dynamics, the opposite effect results and those variables become less important in the default decision. To the degree that latent fixed effects and omitted dynamics predict differential default behavior of fully documented versus undocumented borrowers, this technique may have value in better understanding the mortgage behavior.

In the context of mortgage default literature, our study and its findings make a contribution by proposing a solution to two econometric issues: consideration of borrower heterogeneity beyond observed loan characteristics and imputation of unobserved dynamics for time-varying predictors. In predictive models, using a measure of house price appreciation to reflect changes in borrower equity may be sufficient maximize predictive power and out-of-sample performance. However, from an explanatory perspective, the correlation between housing returns and important liquidity factors, such as borrower income or wealth, may bias the relative importance of negative equity in the default decision. Thus, when explaining why defaults on residential mortgages skyrocketed at the onset of the 2008 recession, understanding the nature of these potential biases should help untangle the joint effects of the falling house prices and deteriorating economic conditions affecting borrower liquidity.

The rest of this chapter is structured as follows: In Section 4.1, we describe the source and nature of the empirical data and provide summary statistics. In Section 4.2, we discuss the theoretical foundations of mortgage default analysis, quantify the proposed biases, and describe the relevant econometric considerations for our analysis. In Section 4.3, we propose several model specifications to impute borrower heterogeneity and dynamics. Section 4.4 goes into detail regarding the parameter selection method. In Section 4.5, we discuss some insights regarding the results.

4.1. Data

The primary source of data for this study comes from Blackbox Logic LLC (BBX). This dataset includes residential, privately securitized mortgages over all credit categories. We examine loans originating between 2003–2014 for properties within one of the 20 CS-MSAs of the S&P CoreLogic Case-Shiller Home Price Indices.[14] We further restrict the sample to loans with standard 15- or 30-year terms and those with estimated CLTV ratios within the range (0,3).[15] Table 4.1 presents some summary statistics for the variables partitioned by MSA to show geographical differences in leverage, loan characteristics, and performance. In the appendices, we provide the names and descriptions for the variables from the BBX dataset (Table G.1) and for the variables used in our statistical models (Table H.1).

4.1.1. Estimating CLTV Ratios

In order to estimate CLTV ratios, we use the origination appraisal value and scale it by the accumulated housing appreciation (or depreciation) at the MSA-level using the S&P CoreLogic Case-Shiller Home Price Indices. This yields monthly estimates for the underlying collateral (75). Combined with the monthly balance updates from the BBX remittances, we obtain estimates of the CLTV ratios using (76). We use the loan origination date to approximate the appraisal date ($t = 0$), then (76) produces the estimated CLTV for borrower i in month t, where the CS terms refer to the index values of MSA j. Thus, $i \in j$.

$$CSValue_{it} = OrigAppraisalValueCalc_i \cdot \left(\frac{CS_{jt}}{CS_{j0}}\right) \tag{75}$$

$$CLTV_{it} = ActualBalance_{it} / CSValue_{it} \tag{76}$$

[14] These restrictions are to limit the sample size and for ease of presentation.

[15] A CLTV equal to zero indicates full repayment and a CLTV exceeding three is unlikely to occur in reality and likely results from an error with either the mortgage balance or appraisal value.

Table 4.1. Summary Statistics for BBX Variables

	Dynamic Variables			Static Variables				
MSA	Default	CLTV	N_{obs}	FICO	Fulldoc	ARM	Term	N_{loans}
ATL	0.102	0.776	3,527,390	686.87	0.336	0.611	0.123	78,373
BOS	0.179	0.772	1,594,709	663.72	0.288	0.628	0.100	43,065
CHA	0.109	0.701	2,047,125	670.09	0.417	0.524	0.142	43,176
CHI	0.180	0.840	5,617,409	655.65	0.313	0.681	0.105	153,044
CLE	0.203	0.845	1,799,405	634.64	0.488	0.512	0.115	35,239
DAL	0.097	0.665	3,892,547	661.65	0.419	0.429	0.175	72,259
DEN	0.085	0.681	3,001,515	686.29	0.337	0.614	0.134	66,843
DET	0.232	0.952	2,180,196	621.84	0.499	0.713	0.090	55,823
LV	0.172	1.057	5,545,130	680.77	0.227	0.643	0.137	145,399
LA	0.133	0.827	11,946,578	682.38	0.192	0.660	0.113	316,390
MIA	0.223	0.974	7,944,618	664.46	0.252	0.628	0.105	177,860
MIN	0.138	0.825	2,968,995	671.55	0.355	0.645	0.115	71,232
NY	0.155	0.767	3,926,385	676.16	0.300	0.587	0.093	87,842
PHX	0.109	0.923	6,157,528	676.78	0.293	0.662	0.129	173,029
POR	0.097	0.668	2,593,714	685.24	0.325	0.532	0.156	62,465
SD	0.093	0.863	4,100,043	706.33	0.197	0.680	0.094	92,923
SF	0.083	0.805	9,410,340	713.37	0.206	0.673	0.087	229,142
SEA	0.100	0.693	3,343,664	689.63	0.327	0.602	0.127	81,343
TPA	0.188	0.944	3,154,583	661.36	0.304	0.608	0.110	70,318
DC	0.115	0.786	4,137,220	677.09	0.331	0.562	0.115	98,535
ALL	0.137	0.837	88,889,094	678.21	0.283	0.629	0.115	2,154,300

Summary statistics for variables partitioned by MSA. Dynamic variables are default indicators and CLTV ratios. Static variables include origination FICO scores and indicators for full documentation, adjustable rate, and 180-month term mortgages.

4.2. Theory

Much of the existing mortgage default literature focuses on two main rationales for default: liquidity default and strategic default. The latter has been extensively studied since the Great Recession when housing appreciation halted and declining property values coincided with drastic increases in defaults. However, liquidity defaults can be more challenging to study due to the difficulty in obtaining critical information regarding individual borrower liquidity.

For example, incentives to strategically default can be captured by the CLTV ratio, which measures the equity position of a borrower. However, liquidity defaults are often driven by individual factors such as the borrower's income or wealth, which are often not available to academic researchers. In the context of this paper, if any time-varying liquidity factors are correlated with housing appreciation, then their effects will be partially captured by through the CLTV variable. This may lead to a biased estimate of the true coefficient for CLTV. This potential bias holds for all predictor variables included in the model.

An intuitive way to model the default decision is for a borrower to default on their mortgage when either their wealth can be increased by defaulting (strategic default) or if a liquidity constraint is binding (liquidity default). For analysis of liquidity defaults, it is important to consider any additional liabilities of the borrowers and how they prioritize their debt payments. Andersson, Chomsisengphet, Glennon, and Li (2013) study this issue and find that pre-crisis borrowers were eight times more likely to prioritize mortgage payments over credit card payments. However, once house prices began falling and strategic incentives began to take hold, borrowers prioritized mortgage payments about the same as credit card payments.

This suggests that in the context of liquidity defaults, borrowers are likely to prioritize their mortgage payment over other types of debt. Thus, it is likely that factors such as income or credit score begin to decline prior to a borrower's first missed mortgage payment. This motivates the idea that such unobserved dynamics are important in differentiating strategic factors from liquidity factors.

In the remainder of this section, we evaluate a simple linear probability model, solve for the biases introduced from key omitted variables, propose a novel solution to impute the omitted variables, and discuss the relevant econometric issues that arise.

4.2.1. Ordinary Least Squares (Naive Model)

Before delving into unobserved heterogeneity and dynamics, we evaluate a traditional linear probability model with a single, static intercept for all borrowers. The resulting regression coefficients from this model will be used as benchmark values for comparison in the subsequent sections. In this naive model (78), our matrix of regressors, X_0, is regressed on $Default$ (represented as y), where $YEARS$ is a matrix of year fixed effects to control for macroeconomic factors over time.

$$X_0 = \begin{bmatrix} 1 & FICO & CLTV & fulldoc & ARM & term & YEARS \end{bmatrix} \tag{77}$$

$$y_{it} = X_{0,it}\beta_0 + \varepsilon_{it} \tag{78}$$

For this model, the optimal solution from ordinary least squares is:

$$\hat{\beta}_0 = (X_0^T X_0)^{-1} X_0^T y \tag{79}$$

In Table 4.2, we present the naive OLS coefficients for each MSA and the weighted average for the full sample. Since our samples are sufficiently large, all predictors are statistically significant at the 99% level. In future sections, we restrict our focus to the magnitudes of important coefficients as the standard errors are sufficiently small to assume significance.

4.2.2. Quantifying Bias from Omission of Borrower Heterogeneity

Let us first rationalize the inclusion of borrower-specific intercepts in the mortgage default setting. The first consideration for a fixed effects model is the nature of the omitted variables. Borrower fixed effects effectively control for static omitted variables. In our model, allowing for borrower-specific intercepts captures any loan or borrower characteristics not already controlled for explicitly in the model. For example, since we do not have data on borrower occupation or marital status, these will be omitted variables that can partially be captured by the intercepts. If any of the included regressors is systematically related with these omitted variables, then the estimated coefficients will be biased.

Even in a more complete model that includes liquidity factors, this unobserved heterogeneity is likely to remain an important consideration for strategic default due to views regarding the morality of default. Guiso et al. (2013) find that while negative equity is a necessary condition for strategic default, the most predictive factors are moral and social considerations.

Table 4.2. Naive OLS Default Model

MSA	Intercept	FICO	CLTV	Fulldoc	ARM	Term	R^2
ATL	0.0684	−0.0674	0.0485	−0.0329	0.0113	0.0309	0.0969
BOS	0.0123	−0.0752	0.0721	−0.0287	0.0470	0.0493	0.1578
CHA	0.0366	−0.0594	0.0434	−0.0181	0.0325	0.0284	0.0908
CHI	0.0756	−0.0607	0.0577	−0.0414	0.0548	0.0250	0.1247
CLE	0.1322	−0.0750	0.0460	−0.0226	0.0715	−0.0070	0.0806
DAL	0.0365	−0.0508	0.0353	−0.0157	0.0372	0.0200	0.0722
DEN	0.0216	−0.0531	0.0409	−0.0148	0.0143	0.0352	0.0754
DET	0.1169	−0.0635	0.0571	−0.0300	0.0766	0.0168	0.0996
LV	0.0616	−0.0570	0.0649	−0.0361	0.0286	0.0451	0.1597
LA	0.0349	−0.0583	0.0720	−0.0280	0.0330	0.0697	0.1519
MIA	0.0655	−0.0503	0.0991	−0.0257	0.0504	0.0664	0.1965
MIN	0.0128	−0.0716	0.0567	−0.0317	0.0418	0.0304	0.1201
NY	0.0600	−0.0697	0.0809	−0.0375	0.0299	0.0450	0.1786
PHX	0.0429	−0.0503	0.0559	−0.0285	0.0212	0.0469	0.1414
POR	0.0370	−0.0522	0.0564	−0.0188	0.0212	0.0559	0.1274
SD	0.0314	−0.0565	0.0593	−0.0269	0.0112	0.0648	0.1205
SF	0.0381	−0.0550	0.0642	−0.0232	0.0094	0.0653	0.1330
SEA	0.0304	−0.0575	0.0612	−0.0183	0.0188	0.0618	0.1455
TPA	0.0811	−0.0642	0.0783	−0.0470	0.0364	0.0333	0.1641
DC	0.0694	−0.0677	0.0626	−0.0323	0.0220	0.0459	0.1406
ALL	0.0507	−0.0587	0.0642	−0.0283	0.0310	0.0482	0.1373

OLS regression coefficients for each MSA. All coefficients statistically significant at the 99% level.

If these unobserved attitudes about default are related with borrowers' selection of mortgage characteristics, the estimated coefficients for those characteristics will be biased.

To quantify the bias resulting from a common-intercept model, we will assume the data generating process (DGP) in (80). Let y_{it} be the true probability of default for individual i at time t, M_i control for time-constant loan characteristics (such as indicators for full documentation or mortgage term), and N_{it} control for dynamic predictors which vary throughout the life of the loan (such as CLTV).

$$y_{it} = \kappa_i + M_i\gamma_1 + N_{it}\gamma_2 + \varepsilon_{it} \tag{80}$$

$$= \kappa_0 + M_i\gamma_1 + N_{it}\gamma_2 + (\kappa_i - \kappa_0) + \varepsilon_{it} \tag{81}$$

$$= X_{it}\beta + \Delta\kappa_i + \varepsilon_{it} \tag{82}$$

If we estimate a linear probability model by OLS as in (77)–(79), then the estimated model differs from (82) by replacing $\Delta\kappa_i + \varepsilon_{it}$ with u_{it}, which is composed of the idiosyncratic error term and borrower heterogeneity. If we substitute (82) into (79), we can solve for the omitted variable bias as such:

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta + \Delta\kappa + \varepsilon) \tag{83}$$

$$= \beta + (X^T X)^{-1} X^T \Delta\kappa + (X^T X)^{-1} X^T \varepsilon \tag{84}$$

$$\mathbb{E}[\hat{\beta}|X] = \beta + (X^T X)^{-1} X^T \Delta\kappa \tag{85}$$

From inspection of (85), it is clear that this proposed bias will be non-zero whenever borrower heterogeneity ($\Delta\kappa$) is systematically related with the included predictor variables in X. The bias will primarily be a function of the relationship between the each regressor and the true heterogeneity; however, any non-zero covariances between the included regressors in the off-diagonal elements of $(X^T X)^{-1}$ may also have minor effects on the bias.

As an example, for the full documentation coefficient, the primary driver of the bias in (85) will be the relationship between fully documented borrowers and the unobserved heterogeneity. Ambrose et al. (2016) find evidence of borrower income misrepresentation concentrated among borrowers who originated low-doc loans but could have originated full-doc loans instead. It is intuitive to associate the willingness to misrepresent income with moral views that impact the propensity to strategically default. One might also expect this relationship to be true regarding liquidity defaults since borrowers with larger and more stable income likely qualify for loans with good terms and have no incentive to misrepresent items on a mortgage application.

Such relations suggest an inverse relationship between full documentation and unobserved borrower heterogeneity. This would suggest that our estimate for $\beta_{fulldoc}$ is biased downwards or that the signal of providing full documentation is smaller if we could observe borrowers' views regarding morality of default. Since the findings of Ambrose et al. (2016) relate to income misrepresentation, this relation would hold true even in models that include borrower income as an independent variable.

An alternative perspective regarding this bias would be to consider weak verification standards of loans classified as fully documented. For example, if income and employment are appropriately verified for loans classified as fully documented, then this classification signals a significant reduction in asymmetric information between borrowers and underwriters. However, if verification is lacking or fully documented borrowers are not properly screened, then the information asymmetry will persist and the classification will be a weaker signal regarding the default risk of the borrower. At the extreme of this case, classification of full documentation is as simple as a borrower checking a box on the mortgage application. In the context of our model, this would suggest that full documentation has a smaller magnitude than in a scenario with more stringent verification standards.

Another consequence of potentially weak verification standards failing to reduce information asymmetries between borrowers and underwriters is an increased sensitivity to house price dynamics. With weak verification standards, it may be possible for weak or fraudulent borrowers to misrepresent income or employment to obtain the loan or get more favorable terms. Such borrowers are likely to have less stable incomes and face binding budget constraints. They may also be less likely to have reservations about strategically defaulting on their mortgages.

Thus, an average borrower's decision to default may appear more sensitive to house price fluctuations when documentation standards are lacking. This would suggest that by imputing these unobserved factors (as in our models), we should expect an increased signal from full documentation and that the sensitivity to the CLTV ratio should decrease in magnitude reflecting a lower sensitivity of default to house price dynamics.

4.2.3. Quantifying Bias from Omission of Variable Dynamics

When it comes to omitted dynamic variables, we often have one of two scenarios: the variable is entirely omitted (borrower income) or the variable is observed at origination and only the dynamics are omitted (credit scores). For a completely omitted variable, the bias would be similar to that of the omitted heterogeneity. However, for variables that are recorded at origination, there is a subtle difference where use of the origination value implicitly assumes that it remains constant over time.

Let's consider credit scores as an example of a variable with omitted dynamics. The inclusion of the origination credit score in the default model assumes that credit scores do not change over time. If credit score dynamics are systematically related with any of the other independent variables, this will result in a bias to their estimated coefficients.

Recall the DGP in (80) and consider a credit score variable, s_{it}, that is only observed at loan origination (we only have s_{i0}). If we estimate a model with borrower fixed effects using individuals' origination credit scores, then the estimated model is (86) where the error term, v_{it} includes $(s_{it} - s_{i0})\gamma_0$ in addition to the idiosyncratic error.

$$y_{it} = \kappa_0 + \Delta\kappa_i + s_{i0}\gamma_0 + M_i\gamma_1 + N_{it}\gamma_2 + (s_{it} - s_{i0})\gamma_0 + \varepsilon_{it} = X_{it}\beta + v_{it} \tag{86}$$

Similar to the prior section, if we substitute (86) into the solution of our estimated model (79) we can solve for the bias.

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta + (s_{it} - s_{i0})\gamma_0 + \varepsilon_{it}) \tag{87}$$

$$= \beta + (X^T X)^{-1} X^T (s_{it} - s_{i0})\gamma_0 + (X^T X)^{-1} X^T \varepsilon \tag{88}$$

$$\mathbb{E}[\hat{\beta}|X] = \beta + (X^T X)^{-1} X^T (s_{it} - s_{i0})\gamma_0 \tag{89}$$

This bias is primarily driven by the relationship between the omitted dynamics and the included predictor variables, as well as the true marginal effect of credit scores on one's propensity to default, γ_0. Intuitively, and from our naive model results, we know that larger credit scores are associated with a lower probability of default. Thus, $\gamma_0 < 0$. If any of the included regressors is systematically related with future credit score dynamics (or any other omitted dynamics), then the regression coefficients will be biased.

If we consider the CLTV ratio, it is likely that house prices have a non-zero correlation with the incomes of some borrowers. As documented by Mayer et al. (2009), increases in unemployment (decreases to income) result in lower demand for housing, and thus house prices. This suggests a positive relationship between house prices and borrower incomes. This would imply that β_{CLTV} is biased upwards and that the actual effect of house price fluctuations is smaller than is suggested by the OLS model.

Intuitively, this bias can be explained as the CLTV dynamics partially capturing the effects of changes to borrower income or credit scores. Since contingent claims approaches to mortgage default often associate borrower equity with strategic incentives to default, this bias would suggest that the impact of falling house prices on strategic defaults may be overstated due to the correlated changes in unobserved liquidity factors.

4.2.4. Econometric Issues

We have now provided a rationale for the use of borrower-specific fixed effects and expressed the bias introduced from omitting them; however, it is crucial to note the issue of perfect multicollinearity in modeling (80). If we allow for a distinct intercept for each borrower, by estimating $\Delta\kappa_i$, any non-time-varying predictor (often everything but CLTV) will be a linear combination of the fixed effects intercepts. Thus, without a remedy to this perfect multicollinearity issue, these loan characteristic variables will be unidentifiable in the regression model. Our solution to this issue is to adapt a ridge regression model, which is the primary focus of Section 4.3.

To resolve the omitted variable bias induced by the omission of predictor dynamics, we refine our model to capture these dynamics through the estimated model parameters. Our first proposed solution is to estimate a heterogeneous trend model, which allows for borrowers' distinct intercepts to change linearly over time. Alternatively, we employ a variable parameter design as in (90) to impute non-linear dynamics.

$$\kappa_{it} = \kappa_0 + \Delta\kappa_i + \Delta\kappa_{it} \tag{90}$$

In this model specification, a model overfitting problem is apparent from allowing the $\Delta\kappa_{it}$ parameter to vary for each observation. As with the perfect multicollinearity issue arising from fixed effects intercepts, we are able to resolve this issue through the use of a ridge regression model.

4.3. Proposed Methodology

Once we allow for the model to have a distinct intercept for each individual borrower, the issue of perfect multicollinearity arises. Similarly, the additional parameters for imputed dynamics introduce additional estimation issues in the form of model overfitting. These are remedied by adapting a penalized regression framework, such as ridge regression introduced by Hoerl (1959). This section provides a brief overview of ridge analysis and subsequently constructs structured penalties to allow for the estimation of the additional parameters.

4.3.1. Ridge Regression

The standard approach for ridge regression adds a penalty term, $\rho\Gamma$ in (91), to the usual OLS solution. The ridge matrix Γ is traditionally set to an identity matrix,[16] which penalizes the magnitudes of estimated coefficients with strength ρ. This approach leads to biased estimates with smaller magnitudes and smaller variances than OLS estimates.

$$\hat{\beta}_{ridge}(\rho) = (X^T X + \rho\Gamma)^{-1} X^T y \tag{91}$$

Alternatively, the penalty structure can be adjusted to impose specific penalties on the additional parameters introduced by borrower-specific intercepts and imputed dynamics. We propose a two-ridge model with separate ridge parameters for each penalty (ρ_1, ρ_2). The first ridge effectively tunes the amount of variation allowed among the borrower-specific intercepts. For the dynamics ridge, we consider two alternatives that trade-off on simplicity versus flexibility. The first proposal is effectively a heterogeneous trend model where each borrower has a distinct time trend, whose slopes are penalized towards 0. Alternatively, a smoothness penalty is considered for the variable parameter design described by (90).

A key attribute of ridge regression coefficients is that they are a function of the ridge parameter(s). For each specific ridge, we can examine changes in coefficient estimates over

[16] See Hoerl and Kennard (1970a,b); Marquardt and Snee (1975); Kasarda and Shih (1977); Vinod (1978).

a range of ridge parameter values. The concept of a ridge trace refers to plotting relevant coefficients against the ridge parameters to evaluate the impact of the ridge on specific parameters from the model.

For uniformity, we will refer to the matrix of the regressors as X, the dependent variable ($Default$) as y, and all of the estimated coefficients as β, which includes the additional parameters for the various models (labeled by κ, δ, and α) in addition to the naive model regressors from (77). Additionally, we will refer to the number of observations (rows of X) as N and the number of loans as K.

4.3.2. Borrower Heterogeneity

In this section, we formulate a penalty function to resolve the perfect multicollinearity that arises from the inclusion of borrower-specific fixed effects. To alleviate the issue, we penalize the squared deviations from the common intercept parameter. The penalty function (and its corresponding gradient and Hessian) are defined in (92), (93), and (94).

$$P_1(\hat{\beta}) = \frac{1}{2K}\Delta\kappa^T\Delta\kappa \tag{92}$$

$$\frac{\partial P_1}{\partial \beta_i} = \begin{cases} \frac{1}{K}\Delta\kappa & \text{if } \beta_i \in \Delta\kappa \\ 0 & \text{otherwise} \end{cases} \tag{93}$$

$$\frac{\partial^2 P_1}{\partial \beta_i \partial \beta_j^T} = \begin{cases} \frac{1}{K} & \text{if } i = j \text{ and } \beta_i \in \Delta\kappa \\ 0 & \text{otherwise} \end{cases} \tag{94}$$

From this penalty function, we obtain its ridge matrix from its Hessian. Thus, for the fixed effects ridge, $\Gamma_1 = I_\kappa \cdot \frac{1}{K}$, where I_κ refers to a sparse matrix with the K rows and columns corresponding to the coefficients, $\Delta\kappa$, forming an identity matrix. Thus, the closed form solution for the fixed effects model is of the following form:

$$\hat{\beta}_{FE}(\rho) = (X^TX + \rho_1\Gamma_1)^{-1}X^Ty \tag{95}$$

With this penalty, we allow for the model to accept various degrees of variation in the borrower-specific intercepts. We refer to the ridge parameter as ρ_1 as we will include a second ridge in the subsequent sections, which uses a separate tuning parameter. For values of ρ_1 close to 0, the model suffers from multicollinearity, time-constant regressors are not identified, and loans with few observations may have extreme intercepts. For very large values of ρ_1, the penalty will force the fixed effects intercepts to all be equal to κ_0 and the results are equivalent to the naive model. This relation can be seen in the ridge trace on the left plot of Figure 4.1.

4.3.3. Heterogeneous Trend Model

Once we have allowed for borrower fixed effects in the model, our next contribution is to resolve the issue of omitting time-varying predictor dynamics. For our first potential solution, we allow for the imputed borrower heterogeneity to change linearly over time. This can be achieved by including borrower-specific time trends. With these K additional parameters, which we call δ_i, we allow for individuals' intercepts to change linearly within the model through the regression coefficients. It is possible to extend this methodology to allow for quadratic or higher order paths over time; however, for each additional degree polynomial we allow for, we must estimate K additional parameters.

Once again, the issue of perfect multicollinearity arises from the inclusion of the borrower-specific slope parameters. Similar to the fixed effects penalty, we introduce a ridge to penalize the magnitude of these slope coefficients towards 0. This ridge matrix, $\Gamma_\delta = I_\delta \cdot \frac{1}{K}$, is similar to that of the fixed effect ridge where I_δ is similar to I_κ; however, the identity matrix is located in the block that corresponds to the parameters in δ. The derivation of this ridge mimics that of the fixed effects ridge. The resulting model yields the following solution:

$$\hat{\beta}_{lin}(\rho) = (X^T X + \rho_1 \Gamma_1 + \rho_2 \Gamma_\delta)^{-1} X^T y \tag{96}$$

Intuitively, this penalty forces the slopes of the time trends to approach 0 as the penalty strengthens (as ρ_2 increases). As ρ_2 approaches 0, the δ's vary widely and the estimation suffers from perfect multicollinearity. This relation is depicted in the right plot of Figure 4.1, where ρ_1 is fixed to be large (effectively implementing a single-intercept model) to maintain the two-dimensional nature of the ridge trace.

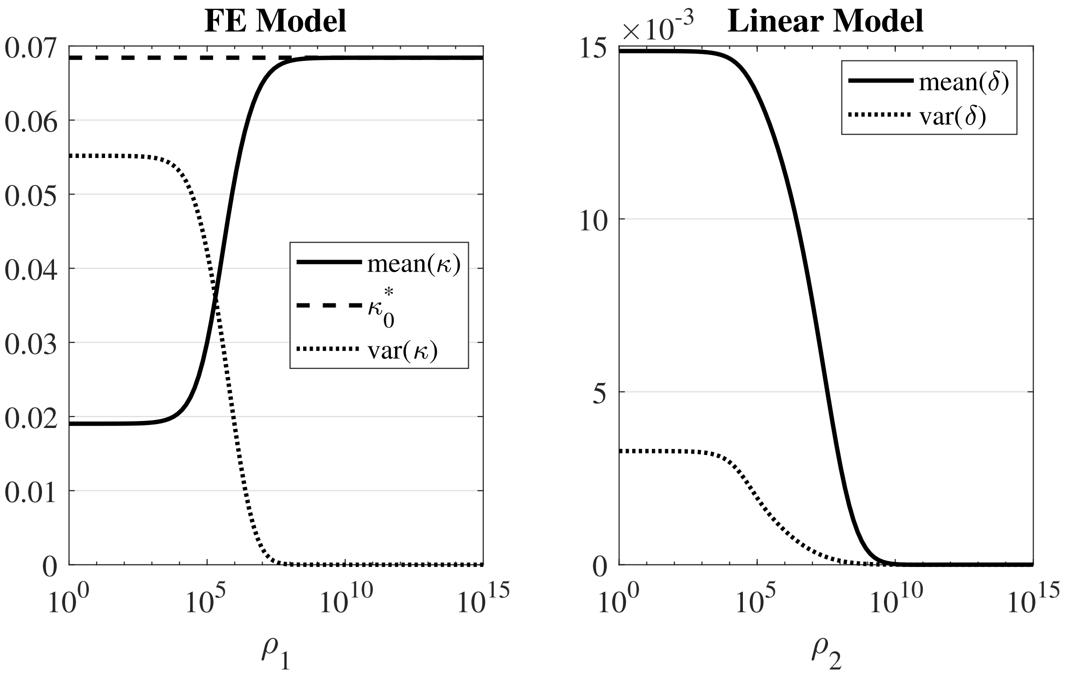

Figure 4.1. Ridge Traces of Penalty Functions

Note: Imputed distribution of heterogeneity parameters and their relation with ridge parameters (Atlanta sample)

4.3.4. Nonlinear Dynamics

The assumption of linear trends in heterogeneity can be fairly restrictive. In this section, we propose an alternative model to allow for non-linear dynamics over time. Rather than expand the polynomial form as described in the previous section, we do so by replacing the trend parameters with N time-varying parameters (represented as $\Delta\kappa_{it}$ in (90)) and imposing a smoothness penalty for each individual. To distinguish these parameters from the time-constant heterogeneity parameters, we label them as α_{it}.

To resolve the model overfitting issue that arises from these additional parameters, we penalize the squared successive deviations in the α parameters within each loan. This smoothness penalty can be computed using the differencing matrix, A, which is a sparse matrix, where $A_\alpha \equiv blkdiag(A_i)$ for $i = 1, ..., K$ is in the block corresponding to the α parameters.

$$A_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix}$$

This matrix allows for us to compute $A_\alpha \alpha$, which computes the successive deviations in α. From this, we formulate the penalty function in (97).

$$P_2(\hat{\beta}) = \frac{1}{2K}\alpha^T A_\alpha^T A_\alpha \alpha \tag{97}$$

$$\frac{\partial P_2}{\partial \beta_i} = \begin{cases} \frac{1}{K} A_\alpha^T A_\alpha \alpha & \forall \beta_i \in \alpha \\ 0 & \text{otherwise} \end{cases} \tag{98}$$

$$\frac{\partial^2 P_2}{\partial \beta_i \partial \beta_j^T} = \begin{cases} \frac{1}{K} A_\alpha^T A_\alpha & \forall \beta_i, \beta_j \in \alpha \\ 0 & \text{otherwise} \end{cases} \tag{99}$$

With this penalty, we construct the ridge matrix $\Gamma_\alpha = \frac{1}{K}A^T A$ from (99). Since this formulation replaces the linear ridge from the prior subsection, we replace Γ_δ with Γ_α, yielding the solution:

$$\hat{\beta}_{smooth}(\rho) = (X^T X + \rho_1 \Gamma_1 + \rho_2 \Gamma_\alpha)^{-1} X^T y \tag{100}$$

The intuition surrounding this penalty at its ridge parameter bounds are that for ρ_2 close to 0, the model is severely overfit and each α parameter is being estimated from only one observation. As ρ_2 increases towards infinity, the α parameters will all converge to 0. We fix ρ_1 to be large and depict the trace of the mean α coefficient on the left and penalty function values on the right of Figure 4.2.

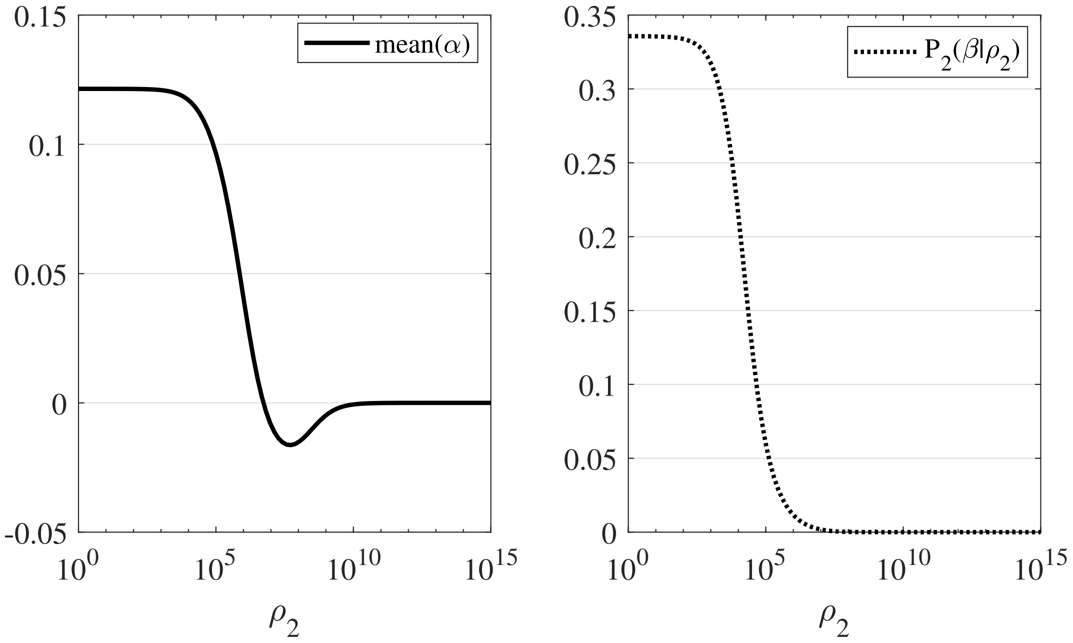

Figure 4.2. Nonlinear Dynamics Ridge Trace

Note: Nonlinear penalty ridge traces for $\bar{\alpha}_{it}$ and $P_2(\beta)$ (Atlanta sample)

4.4. Parameter Selection and Interpretation

To obtain a single set of parameter estimates for each model, we must select specific values for the ridge parameters. This parameter selection issue has been a topic of interest in ridge regression for quite some time having been discussed by Hoerl and Kennard (1970a,b); Marquardt and Snee (1975); Kasarda and Shih (1977); Vinod (1978); and Golub et al. (1979). In this section, we adapt a variation of 10-fold cross-validation to select parameter values that maximize out-of-sample predictive performance. The relationship between the ridge parameters and the R^2 of the ridge regression model provides an intuitive transformation of the parameter selection problem into an optimization problem where the domain of interest is the degree of increased model fit allowed by each additional ridge parameter.

4.4.1. Using Model Fit to Transform Ridge Parameters

One of the benefits of allowing for borrower heterogeneity and dynamics is a better fit model. Consequently, we turn to the relationship between ρ and R^2. If we start with a large penalty, the penalized parameters are trivial; however, as we relax the penalty strength (ρ decreases), the model R^2 monotonically increases as depicted in Figure 4.3.

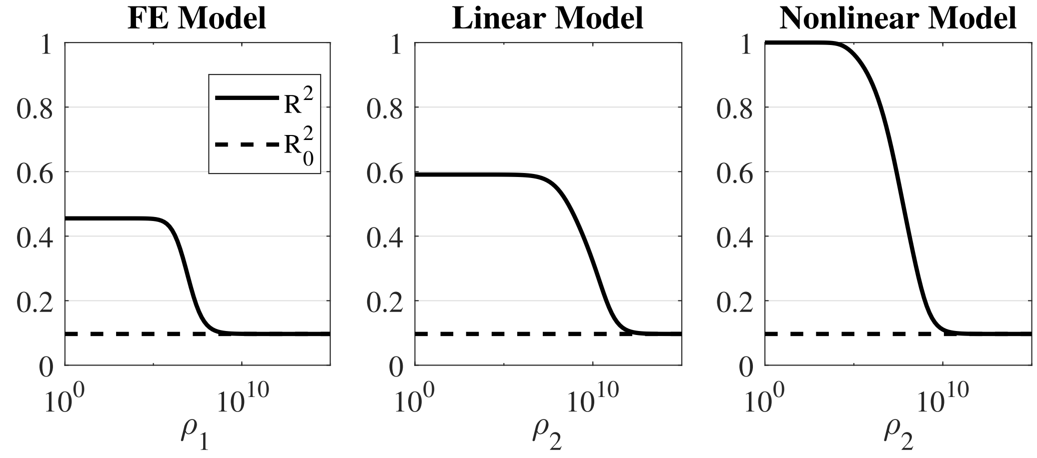

Figure 4.3. Ridge Trace of Model Fit

Note: Ridge traces of model R^2 for each penalty function (Atlanta sample)

This injective relation between the ridge parameters and R^2 provides a more interpretable specification of the ridge parameters, where Δ_1 is the marginal increase of the fixed effects model R^2 over the naive model fit. Thus, if we specify how much variation we wish to be captured by time-constant borrower heterogeneity, we obtain a unique ρ_1, and thus, a unique $\hat{\beta}$.

$$\Delta_1 = R^2_{FE}(\rho_1) - R^2_{naive} \tag{101}$$

Similarly, we can transform the selection of ρ_2 with a specification of Δ_2, which represents the additional explanatory power captured by imputed dynamics. This design, as shown in (102), transforms the parameter selection problem into a specification of model fit.

$$\Delta_2 = R^2 - R^2_{FE} \tag{102}$$

Although this transformation still has two parameters for specification, selecting Δ_1 and Δ_2 is less arbitrary than other methods of parameter selection. This specification gives the ridge traces for the ridge regression parameters a degree of interpretability.

4.4.2. Maximizing Out-of-Sample Prediction

After transforming the ridge parameter selection issue into a specification of model fit, we incorporate cross-validation methods to tune the transformed parameter so as to maximize the out-of-sample prediction accuracy. For each MSA, we partition the set of loans into 10 disjoint testing sets. For each testing set, the remaining loans are used as training sets in the ridge regression models. This 10-fold procedure is independently applied on two random partitions of loans, which yields 20 distinct training and test sets.

For each training set, we obtain a unique ρ_1 that achieves the marginal increase in model fit specified by Δ_1. This yields updated coefficients for the observed independent variables (the fixed effects coefficients are disregarded as they relate to unobserved variables). Using these updated coefficients, we apply the testing set and evaluate the predicted values.

Since the outcome variable is dichotomous, we measure the testing set predictions using the receiver operating characteristic (ROC) curve. An ROC curve plots the true positive prediction rate against the false positive rate over the entire domain for the cutoff to predict default. This effectively varies the required probability of default for the model to predict that a borrower will default. In machine learning, models of various complexity can be compared by evaluating the integral (AUC) of this curve (with larger values indicating better performance).

The testing AUCs are calculated and averaged for each of the 20 cross-validation iterations. The resulting curve is traced out across the Δ_1 domain for the Atlanta subsample in Figure 4.4. The optimal parameter is selected from maximizing this curve ($\Delta_1^* = 0.036$ for ATL). This transforms the parameter selection issue into the optimization problem in (103).

$$\Delta_1^* = \max_{\Delta_1} A\bar{U}C_{test}(\Delta_1) \tag{103}$$

However, this marginal increase in out-of-sample performance is not guaranteed for all MSA-penalty combinations. As detailed in Table 4.3, which provides optimal parameters values and the impact on the explanatory variable coefficient magnitudes, 7/20 MSAs do not obtain any improvements beyond the naive model estimates.

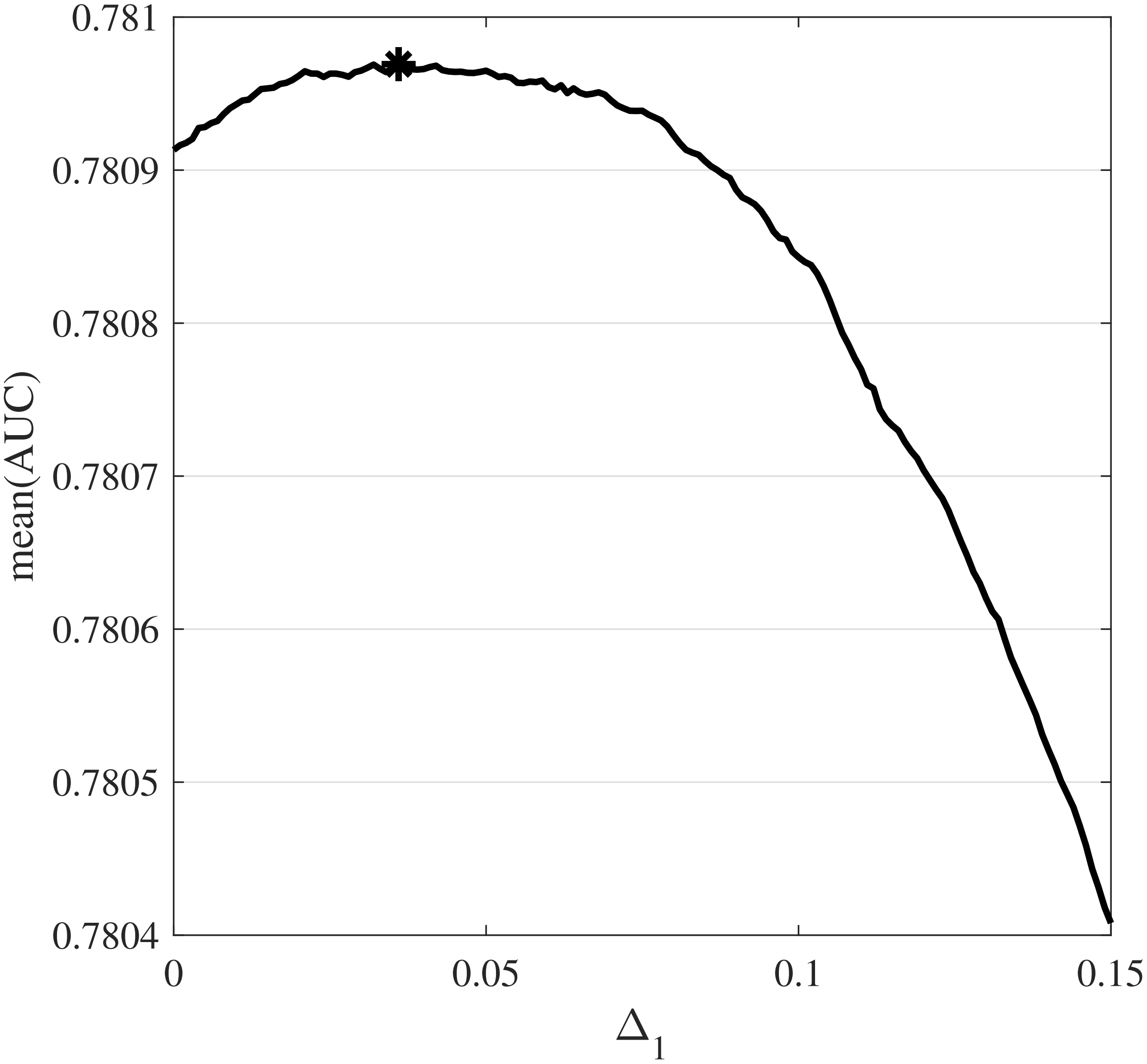

Figure 4.4. AUC Maximization Example

Note: Ridge trace for the mean AUC from two partitions of 10-fold cross-validation for the ATL subsample. The star marks the optimal value for Δ_1 that maximizes the out-of-sample predictive performance.

This procedure is repeated in (104) for Δ_2 in the linear and non-linear models with Δ_1^* fixed. These results are presented in Tables 4.4 and 4.5.

$$\Delta_2^* = \max_{\Delta_2} A\bar{U}C_{test}(\Delta_2|\Delta_1) \tag{104}$$

Table 4.3. Optimal Fixed Effects Ridge Strengths

MSA	R_0^2	Δ_1^*	Δ_{FICO}	Δ_{CLTV}	$\Delta_{Fulldoc}$	Δ_{ARM}	Δ_{Term}
ATL	9.69%	3.60%	0.63%	0.53%	1.86%	0.74%	1.53%
BOS	15.78%	0.70%	−0.04%	0.02%	0.22%	−0.15%	0.34%
CHA	9.08%	3.20%	0.25%	0.04%	0.15%	−0.72%	1.94%
CHI	12.47%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CLE	8.06%	0.90%	0.26%	0.06%	−0.44%	0.15%	−0.74%
DAL	7.22%	4.10%	0.46%	0.11%	1.54%	0.53%	1.25%
DEN	7.54%	9.00%	3.75%	3.14%	7.09%	9.31%	9.21%
DET	9.96%	1.00%	0.20%	0.23%	0.20%	0.43%	1.38%
LV	15.97%	4.00%	−0.06%	−0.62%	1.20%	0.36%	1.86%
LA	15.19%	14.30%	−0.33%	4.45%	1.66%	−8.62%	12.65%
MIA	19.65%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
MIN	12.01%	5.80%	0.95%	1.44%	3.39%	3.14%	5.37%
NY	17.86%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PHX	14.14%	8.80%	0.28%	3.18%	5.62%	−3.74%	7.88%
POR	12.74%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
SD	12.05%	14.10%	1.56%	6.06%	3.49%	−10.10%	12.22%
SF	13.30%	14.50%	1.06%	4.54%	2.93%	−17.30%	10.18%
SEA	14.55%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TPA	16.41%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
DC	14.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ALL	13.73%	5.92%	0.37%	1.73%	1.66%	−3.23%	4.68%

Optimal fixed effects ridge penalties for each MSA and the percent change in coefficient magnitudes resulting from including omitted variable imputations. Estimates for ALL are weighted averages across the 20 CS-MSAs.

Table 4.4. Optimal Linear Ridge Strengths

MSA	$R_0^2 + \Delta_1^*$	Δ_2^*	Δ_{FICO}	Δ_{CLTV}	$\Delta_{Fulldoc}$	Δ_{ARM}	Δ_{Term}
ATL	13.29%	1.90%	0.16%	−0.62%	0.50%	3.74%	−0.25%
BOS	16.48%	4.20%	−1.03%	−2.02%	1.04%	−0.51%	−0.23%
CHA	12.28%	0.50%	−0.01%	−0.25%	−0.21%	−0.06%	0.18%
CHI	12.47%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CLE	8.96%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
DAL	11.32%	1.00%	−0.01%	−0.50%	0.18%	0.20%	−0.21%
DEN	16.54%	1.30%	0.39%	−0.32%	0.22%	3.80%	1.29%
DET	10.96%	0.60%	0.22%	−0.16%	−0.04%	0.58%	0.27%
LV	19.97%	3.10%	−0.97%	−3.80%	−0.24%	3.78%	−1.95%
LA	29.49%	3.90%	−1.29%	−3.29%	−2.53%	−1.42%	1.35%
MIA	19.65%	5.40%	−1.34%	−3.62%	−4.35%	−1.32%	−0.52%
MIN	17.81%	1.40%	−0.02%	−0.65%	0.29%	2.16%	0.32%
NY	17.86%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PHX	22.94%	2.50%	−0.78%	−2.52%	0.58%	3.03%	−0.33%
POR	12.74%	7.30%	−1.78%	−2.96%	−0.72%	−2.35%	−0.22%
SD	26.15%	4.20%	−1.32%	−2.69%	−2.41%	6.30%	0.44%
SF	27.80%	6.20%	−2.18%	−4.40%	−2.87%	−1.58%	0.33%
SEA	14.55%	0.70%	−0.15%	−0.27%	−0.14%	−0.08%	−0.04%
TPA	16.41%	0.10%	−0.02%	−0.06%	−0.03%	0.01%	0.02%
DC	14.06%	4.70%	−0.89%	−1.64%	0.11%	0.62%	0.87%
ALL	19.64%	3.00%	−0.79%	−2.07%	−1.08%	0.58%	0.12%

Optimal heterogeneous trend ridge penalties for each MSA and the percent change in coefficient magnitudes resulting from including omitted variable imputations. Estimates for ALL are weighted averages across the 20 CS-MSAs.

Table 4.5. Optimal Nonlinear Ridge Strengths

MSA	$R_0^2 + \Delta_1^*$	Δ_3^*	Δ_{FICO}	Δ_{CLTV}	$\Delta_{Fulldoc}$	Δ_{ARM}	Δ_{Term}
ATL	13.29%	2.20%	0.04%	−0.51%	0.54%	2.73%	−0.06%
BOS	16.48%	5.50%	−1.27%	−2.04%	0.99%	−1.00%	0.24%
CHA	12.28%	0.40%	−0.02%	−0.16%	−0.12%	−0.08%	0.10%
CHI	12.47%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CLE	8.96%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
DAL	11.32%	1.20%	−0.04%	−0.46%	0.18%	0.11%	−0.18%
DEN	16.54%	1.50%	0.22%	−0.29%	0.42%	2.80%	1.05%
DET	10.96%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
LV	19.97%	2.50%	−0.70%	−2.20%	−0.19%	1.79%	−0.91%
LA	29.49%	4.20%	−1.58%	−2.35%	−2.30%	−2.38%	0.93%
MIA	19.65%	6.80%	−1.70%	−3.52%	−4.45%	−1.76%	−0.07%
MIN	17.81%	1.50%	−0.08%	−0.42%	0.29%	1.42%	0.53%
NY	17.86%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
PHX	22.94%	3.40%	−1.24%	−2.05%	0.30%	0.98%	−0.11%
POR	12.74%	8.50%	−1.95%	−2.63%	−0.60%	−3.34%	0.10%
SD	26.15%	5.90%	−2.10%	−2.53%	−2.68%	2.53%	0.41%
SF	27.80%	7.90%	−2.86%	−3.80%	−3.27%	−6.13%	0.16%
SEA	14.55%	1.20%	−0.25%	−0.35%	−0.19%	−0.28%	−0.02%
TPA	16.41%	0.10%	−0.02%	−0.04%	−0.02%	0.00%	0.02%
DC	14.06%	6.40%	−1.20%	−1.60%	0.23%	−0.39%	1.15%
ALL	19.64%	3.62%	−1.04%	−1.70%	−1.12%	−0.72%	0.20%

Optimal non-linear dynamics ridge penalties for each MSA and the percent change in coefficient magnitudes resulting from including omitted variable imputations. Estimates for ALL are weighted averages across the 20 CS-MSAs.

4.5. Discussion

In mortgage default analysis, many important predictor variables pertain to loan or borrower characteristics that are only recorded at origination of the loan. For example, since a borrower's credit score is often used to model default risk, the omission of its dynamics potentially results in biases in coefficients correlated with those credit score dynamics. Such omitted dynamics may arise from other dynamic variables such as borrower income or wealth. Additionally, the issue of perfect multicollinearity arises when including borrower fixed effects to capture unobserved heterogeneity.

We resolve these issues by introducing a two-dimensional ridge regression model to impute predictor dynamics and borrower heterogeneity. Although these models require a subjective specification of ridge parameters, ρ, we transform this problem into an optimization problem where parameters are selected such that out-of-sample performance is maximized.

From analysis of relative changes in parameter magnitudes, we can infer some details regarding the bias introduced by these omissions. The most apparent evidence of the bias is in the CLTV coefficient where the fixed effects ridge increases the magnitude of its coefficient and the dynamic ridges result in a decrease. This is consistent with CLTV ratio capturing some of the variation of dynamic omitted factors such as income or wealth. In the context of our models, this implies that the CLTV ratio is actually a lesser signal of default if we could observe information such as borrowers' views about the morality of default or omitted dynamics.

The insights from these findings can aid in understanding the importance of many factors relating to both strategic behavior and default risk. In a period with lax documentation standards, such as the pre-crisis period, weak or even potentially fraudulent borrowers may have qualified for loans. If some of these weak but fully documented borrowers were more strategic, then when prices fell they may have found default optimal. Therefore, the price variable may have absorbed some of the effects of the documentation variable.

To make this clearer, imagine a world where there are two groups of borrowers. One group consists of strategic, fraudulent borrowers and the other of non-strategic, legitimate borrowers. Consider two scenarios. The first scenario allows borrowers to simply check a box to declare full documentation status. Whether someone checked the box would not provide any information on their default probability. On the other hand, falling house prices would lead the strategic fraudulent borrowers to default and therefore the price variable would explain the variation in borrower performance. The second scenario involves perfect documentation. In this case no fraudulent borrower would pass the documentation requirements and all fully documented loans would be held by non-strategic borrowers. In this scenario, full documentation would explain much of loan performance and price would explain much less of the loan performance.

To further support our results, extensions of this project include lifting the penalized regression framework into a logit or probit setting and implementing bounded estimation to ensure plausible results. In a non-linear estimation model, such as logit and probit, computation of penalized coefficients is far more time consuming due to the lack of closed form solutions. In regards to bounded estimation, due to the bounded nature of the probability of default, it may be possible to further restrict parameter estimation to more closely reflect reality.

Chapter 5. Conclusion

The measurement of risk in mortgage portfolios and subsequent pricing of mortgage-backed securities involves modeling the multivariate return distribution of a large number of loans with highly correlated collateral values. The risk implications of these correlations are usually masked whenever economic conditions are strong and most loans simply yield their constant, scheduled coupon rate. However, when the economy faces a downturn and a default shock occurs, the VaR framework in Section 2.4 shows how the shock has compounding effects from the increased risk to the individual loans, but also from the larger correlations across the asset returns.

Even under the assumption of normally distributed housing returns, the relationship between the degree of censoring (default rate) and theoretical correlation across the loan returns is highly non-linear. For example, when the underlying properties have a correlation of ρ, the model predicts a correlation of $0.297 \cdot \rho$ when default rates are as low (1.41%) and $0.599 \cdot \rho$ at just 11.54% default. This suggests that the observed asset correlations for these portfolios in the years preceding the crisis would have been roughly half of the levels experienced at the peak of the crisis. These findings from Chapter 2 demonstrate an important structural relationship between the underlying distribution of housing returns and the risk of a portfolio of loans secured by those properties.

Given the spatially dependent distribution of mortgage returns, which effectively defines the diversification potential for the mortgage market, the geographical concentration of the outstanding mortgage market can lead to additional limits to diversification. The examination of the geographical concentration of mortgage debt in Chapter 3 produces some interesting results when contrasting between conforming loans eligible for purchase by the GSEs and the jumbo mortgage market, which tend to be held as portfolio loans on the balance sheets of large financial institutions. Unlike the geographically disperse governmental programs that currently capture a large share of new mortgage originations, private market loans tend to be heavily concentrated in major metropolitan areas where high property values often require large loans beyond the thresholds for GSE eligibility.

Regardless of the potential for investors to create a geographically diversified portfolio of mortgages, the decision of borrowers to default on these loans lies at the center of the measurement of portfolio risk. The censored variable framework in Chapter 2 demonstrates how the correlations across the underlying housing are revealed in a portfolio through the default rate, or the ex-ante probability of default. This mortgage default decision can be challenging to statistically model due to data limitations. The econometric methodology outlined in Chapter 4 aims to impute unobserved borrower heterogeneity and dynamics to examine if these potential omissions lead to biases in the estimation of the effects of various observed default factors.

As a whole, this dissertation contributes to the existing literature on the modeling of risk in the mortgage market and the literature examining geographical diversification, particularly for assets that are spatially dependent. The findings of these three essays demonstrate the challenges that investors may face when constructing a geographically diversified portfolio of mortgages and how the benefits of this diversification are substantially reduced in periods with poor economic conditions and increased default rates.

Appendix A. The CS-20 MSAs

Table A.1. List of CS-20 MSAs

MSA	Metropolitan Statistical Area
ATL	Atlanta-Sandy Springs-Marietta, GA
BOS	Boston-Cambridge-Quincy, MA-NH
CHA	Charlotte-Gastonia-Rock Hill, NC-SC
CHI	Chicago-Joliet-Naperville, IL-IN-WI
CLE	Cleveland-Elyria-Mentor, OH
DAL	Dallas-Fort Worth-Arlington, TX
DEN	Denver-Aurora-Broomfield, CO
DET	Detroit-Warren-Livonia, MI
LV	Las Vegas-Paradise, NV
LA	Los Angeles-Long Beach-Santa Ana, CA
MIA	Miami-Fort Lauderdale-Pompano Beach, FL
MIN	Minneapolis-St. Paul-Bloomington, MN-WI
NY	New York-Northern New Jersey-Long Island, NY-NJ-PA
PHX	Phoenix-Mesa-Glendale, AZ
POR	Portland-Vancouver-Hillsboro, OR-WA
SD	San Diego-Carlsbad-San Marcos, CA
SF	San Francisco-Oakland-Fremont, CA
SEA	Seattle-Tacoma-Bellevue, WA
TPA	Tampa-St. Petersburg-Clearwater, FL
DC	Washington-Arlington-Alexandria, DC-VA-MD-WV
ALL	Totals or averages across all CS-MSAs
USA	Totals or averages of U.S. Census data

List of the 20 Metropolitan Statistical Areas in the S&P CoreLogic Case-Shiller 20-City Composite Home Price NSA Index.

Appendix B. HPI Correlation Matrix

Table B.1. Average House Price Return Correlations Across ZIP Codes

	ATL	BOS	CHA	CHI	CLE	DAL	DEN	DET	LV	LA	MIA	MIN	NY	PHX	POR	SD	SF	SEA	TPA	DC
ATL	0.68																			
BOS	0.52	0.80																		
CHA	0.57	0.34	0.59																	
CHI	0.58	0.43	0.51	0.69																
CLE	0.49	0.32	0.42	0.53	0.53															
DAL	0.46	0.32	0.37	0.25	0.24	0.55														
DEN	0.52	0.38	0.37	0.34	0.41	0.52	0.70													
DET	0.55	0.45	0.43	0.54	0.55	0.31	0.50	0.75												
LV	0.58	0.51	0.42	0.66	0.49	0.38	0.46	0.65	0.91											
LA	0.45	0.47	0.33	0.54	0.34	0.20	0.21	0.46	0.74	0.83										
MIA	0.52	0.46	0.38	0.58	0.38	0.32	0.39	0.52	0.82	0.67	0.80									
MIN	0.65	0.63	0.50	0.64	0.55	0.42	0.57	0.64	0.69	0.55	0.63	0.77								
NY	0.48	0.69	0.36	0.53	0.31	0.22	0.23	0.37	0.57	0.55	0.55	0.59	0.74							
PHX	0.53	0.48	0.43	0.59	0.41	0.37	0.45	0.58	0.81	0.63	0.78	0.64	0.55	0.86						
POR	0.54	0.17	0.54	0.60	0.47	0.37	0.42	0.48	0.67	0.41	0.59	0.54	0.31	0.67	0.81					
SD	0.47	0.55	0.30	0.55	0.41	0.24	0.33	0.54	0.73	0.78	0.66	0.63	0.55	0.61	0.38	0.81				
SF	0.52	0.52	0.38	0.53	0.40	0.29	0.37	0.55	0.67	0.73	0.61	0.60	0.51	0.62	0.40	0.74	0.78			
SEA	0.58	0.25	0.55	0.62	0.46	0.38	0.39	0.49	0.70	0.57	0.61	0.59	0.37	0.66	0.73	0.55	0.57	0.82		
TPA	0.59	0.52	0.46	0.63	0.44	0.41	0.45	0.54	0.80	0.67	0.79	0.68	0.60	0.79	0.65	0.66	0.63	0.68	0.83	
DC	0.43	0.52	0.32	0.61	0.39	0.17	0.22	0.44	0.70	0.74	0.69	0.59	0.65	0.65	0.46	0.71	0.64	0.55	0.69	0.77

Average correlations across housing returns across ZIP codes within the 20 CS-MSAs. The annual HPI data at the five-digit ZIP code level is obtained from the FHFA and spans various starting points through 2017. Returns are calculated by differencing the natural logarithm of HPI levels. Presented correlations are the means after allocating the 8,280,415 ZIP code level correlations across 4,070 ZIP codes to the 210 CS-MSA combinations.

Appendix C. Truncated and Censored Normal Distributions

The statistical foundations for deriving the closed form solution of the correlation across censored loan returns in Section 2.3 comes from the work on the bivariate truncated normal distribution in Chapter 46 Section 9 of Kotz, Balakrishnan, and Johnson (2000). The relationship between truncated and censored data allows us to adapt these concepts to the return structure of mortgage loans.

In this appendix, we derive the necessary moments to obtain the correlation for the cases of right truncation and right censoring of the bivariate standard normal distribution. In Table C.1, we consolidate the relevant moments for each distribution, which are procured in more detail throughout the remainder of this appendix.

As in Kotz et al. (2000), we start with (105), a standard bivariate normal distribution with a latent correlation of ρ. For the truncated case, (106), only values of $X_1 < h$ are used.[17] Censoring, on the other hand, observes the values that were truncated, but replaces the true latent value with a constant, censoring value, c, as in (107).

$$X_1, X_2 \sim N(0,0,1,1,\rho) \tag{105}$$

$$T_1, T_2 = X_1, X_2 \mid X_1 < h \tag{106}$$

$$C_1, C_2 = X_1^*, X_2 \text{ where } X_1^* = \begin{cases} X_1 & \text{if } X_1 < h \\ c & \text{if } X_1 \geq h \end{cases} \tag{107}$$

In many cases of censoring, the censoring value, c, is equal to h maintaining a continuous relationship with the latent variable. However, the coupon rate (censoring return for mortgages) is unlikely to align perfectly with the factors influencing the default decision (determining h), so we use a separate constant, c. To match with the example in Section 2.3, we also present the simplified results where $c = 0$.

[17] This deviates slightly from the example in Kotz et al. (2000) with a simple adjustment to right truncation as opposed to left truncation as in their example.

By definition, the expectation of a standard normal variable, (108) where $\phi(\cdot)$ refers to the standard normal pdf, is equal to 0.

$$\mathbb{E}[X_1] = \int_{-\infty}^{\infty} x\phi(x)dx = 0 \tag{108}$$

To obtain the expectation of a truncated variable, the normal density function is scaled by the cumulative density at the point of truncation, $\Phi(h)$. This rescales the density function to (109) to ensure that the area under the curve remains equal to 1. This cumulative density factors out of the integral in (110) yielding the result in (111).

$$\phi_T(x) = \begin{cases} \dfrac{\phi(x)}{\Phi(h)} & \text{if } x < h \\ 0 & \text{if } x \geq h \end{cases} \tag{109}$$

$$\mathbb{E}[T_1] = \int_{-\infty}^{h} x\phi_T(x)dx \tag{110}$$

$$= \frac{1}{\Phi(h)} \int_{-\infty}^{h} x\phi(x)dx = -\frac{\phi(h)}{\Phi(h)} \tag{111}$$

For the censored case, the limits of integration are partitioned into two intervals as in (112). The first integral yields the expectation conditional on $x < h$ (since $\lim_{x\to-\infty} \phi(x) = 0$). This is also the truncated expectation multiplied by $\Phi(h)$, which is equal to $1 -$ the probability of censoring. The latter integral spans the censored range and yields the censoring value, c, multiplied by the probability of censoring. The assumption of $c = 0$ eliminates the latter term in (114), yielding (115).

$$\mathbb{E}[C_1] = \int_{-\infty}^{h} x\phi(x)dx + \int_{h}^{\infty} c\phi(x)dx \tag{112}$$

$$= (-\phi(x))\,\Big|_{-\infty}^{h} + c \cdot (\Phi(x))\,\Big|_{h}^{\infty} \tag{113}$$

$$= -\phi(h) + c \cdot (1 - \Phi(h)) \tag{114}$$

$$c = 0 \implies = -\phi(h) \tag{115}$$

The second raw moments are solved by similar methods, (116)–(119) for truncated case and (120)–(123) for censored case, and presented in (141) of Table C.1. The steps from

(118) to (119) and from (121) to (122) require the use of L'Hôpital's Rule to resolve the indeterminate solution (of the form $-\infty \cdot 0$) for $\lim_{x\to-\infty}(x\phi(x))$, which approaches 0.

$$\begin{aligned}
\mathbb{E}\left[T_1^2\right] &= \int_{-\infty}^{h} x^2 \phi_T(x)dx & (116)\\
&= \frac{1}{\Phi(h)} \int_{-\infty}^{h} x^2 \phi(x)dx & (117)\\
&= \frac{1}{\Phi(h)} \cdot \left(\Phi(x) - x\phi(x)\right)\Big|_{-\infty}^{h} & (118)\\
&= 1 - \frac{h\phi(h)}{\Phi(h)} & (119)
\end{aligned}$$

Just as with the first raw moment, the censored second raw moment when $c = 0$ is equal to that of the truncated case multiplied by $\Phi(h)$.

$$\begin{aligned}
\mathbb{E}\left[C_1^2\right] &= \int_{-\infty}^{h} x^2 \phi(x)dx + \int_{h}^{\infty} c^2 \phi(x)dx & (120)\\
&= \left(\Phi(x) - x\phi(x)\right)\Big|_{-\infty}^{h} + c^2 \cdot \left(\Phi(x)\right)\Big|_{h}^{\infty} & (121)\\
&= \Phi(h) - h\phi(h) + c^2 \cdot (1 - \Phi(h)) & (122)\\
c = 0 \implies &= \Phi(h) - h\phi(h) & (123)
\end{aligned}$$

These raw moments are combined to produce the variances in (125) and (127) along with the simplified censored case in (128) for $c = 0$.

$$\begin{aligned}
\mathrm{var}(T_1) &= \mathbb{E}\left[T_1^2\right] - \mathbb{E}[C_1]^2 & (124)\\
&= 1 - \frac{h\phi(h)}{\Phi(h)} - \left(\frac{\phi(h)}{\Phi(h)}\right)^2 & (125)\\
\mathrm{var}(C_1) &= \mathbb{E}\left[C_1^2\right] - \mathbb{E}[C_1]^2 & (126)\\
&= \Phi(h) - h\phi(h) - \phi(h)^2 + c(1 - \Phi(h))(c\Phi(h) + 2\phi(h)) & (127)\\
c = 0 \implies &= \Phi(h) - h\phi(h) - \phi(h)^2 & (128)
\end{aligned}$$

For the truncated distribution, the truncation changes the unconditional distribution of T_2. Kotz et al. (2000) show this, and the remaining moments for the truncated case in (144)–(149) of Table C.1, come from Equations (46.155)–(46.160) in Kotz et al. (2000).[18]

However, for the censored case, the unconditional distribution of the second variable, C_2, is unaffected and remains a standard normal variable; thus, the covariance, (129), simplifies to (130) since $\mathbb{E}[C_2] = 0$.

$$\begin{aligned} \operatorname{cov}(C_1, C_2) &= \mathbb{E}[C_1C_2] - \mathbb{E}[C_1]\mathbb{E}[C_2] && (129) \\ &= \mathbb{E}[C_1C_2] && (130) \end{aligned}$$

The expectation of the product of C_1 and C_2 can be written as the sum of the two double integrals in (131) where $f(x, y)$ is the joint density function for the bivariate standard normal distribution.

$$\operatorname{cov}(C_1, C_2) = \int_{-\infty}^{h} \int_{-\infty}^{\infty} xyf(x, y)dydx + \int_{h}^{\infty} \int_{-\infty}^{\infty} cyf(x, y)dydx \tag{131}$$

In the first term, x can be factored out of the inner integral, and since c is a constant, it can be factored out of both integrals in the latter term.

$$\operatorname{cov}(C_1, C_2) = \int_{-\infty}^{h} x \int_{-\infty}^{\infty} yf(x, y)dydx + c\int_{h}^{\infty} \int_{-\infty}^{\infty} yf(x, y)dydx \tag{132}$$

Since $\int yf(x, y)dy = \rho x\phi(x)$, we can substitute this into both terms, which yields (133).

$$\operatorname{cov}(C_1, C_2) = \rho \int_{-\infty}^{h} x^2\phi(x)dx + \rho c \int_{h}^{\infty} x\phi(x)dx \tag{133}$$

The first integral in (133) is identical to the second raw moment when $c = 0$ and the latter integral yields the probability density at h. This yields the solution in (134), which simplifies

[18] The variance in (142) deviates from (46.161) in Kotz et al. (2000) by using the variance of a right truncated variable rather than a left truncated variable as in their example. Additionally, the conditional expectation in (143) is from the text preluding (46.155).

to (135), and further to (136) when $c = 0$.

$$\operatorname{cov}(C_1, C_2) = \rho(\Phi(h) - h\phi(h)) + \rho c(-\phi(h)) \tag{134}$$

$$\operatorname{cov}(C_1, C_2) = \rho(\Phi(h) + (c - h)\phi(h)) \tag{135}$$

$$c = 0 \implies = \rho(\Phi(h) - h\phi(h)) \tag{136}$$

Substituting (127) and (134) respectively into the definition of correlation, (137), we obtain the solution for the correlation in (138).

$$\operatorname{corr}(C_1, C_2) = \frac{\operatorname{cov}(C_1, C_2)}{\sqrt{\operatorname{var}(C_1)}\sqrt{\operatorname{var}(C_2)}} \tag{137}$$

$$= \rho \cdot \frac{\Phi(h) + (c - h)\phi(h)}{\sqrt{\Phi(h) - h\phi(h) - \phi(h)^2 + c(1 - \Phi(h))(c\Phi(h) + 2\phi(h))}} \tag{138}$$

If we substitute (128) and (135) into (137) for the simplified case of censoring, this greatly reduces the simplicity of the closed form solution and it thus applied to the example in Section 2.3.

$$= \rho \cdot \frac{\Phi(h) - h\phi(h)}{\sqrt{\Phi(h) - h\phi(h) - \phi(h)^2}} \tag{139}$$

Table C.1. Relevant Moments of Standard, Truncated, and Censored Normal Distributions

Distribution:	Standard Normal	Truncated Normal	Censored Normal	
Variables:	X_1, X_2	T_1, T_2	$C_1, C_2 \mid c = 0$	
$\mathbb{E}[X_1]$	0	$-\dfrac{\phi(h)}{\Phi(h)}$	$-\phi(h)$	(140)
$\mathbb{E}[X_1^2]$	1	$1 - \dfrac{h\phi(h)}{\Phi(h)}$	$\Phi(h) - h\phi(h)$	(141)
$\text{var}(X_1)$	1	$1 - \dfrac{h\phi(h)}{\Phi(h)} - \left(\dfrac{\phi(h)}{\Phi(h)}\right)^2$	$\Phi(h) - h\phi(h) - (\phi(h))^2$	(142)
$\mathbb{E}[X_2 \vert X_1]$	$\rho \cdot X_1$	$\rho \cdot T_1$	$\rho \cdot C_1$	(143)
$\mathbb{E}[X_2]$	0	$\rho \cdot \mathbb{E}[T_1]$	0	(144)
$\mathbb{E}[X_2^2]$	1	$\rho^2 \cdot \mathbb{E}[T_1^2] + 1 - \rho^2$	1	(145)
$\text{var}(X_2)$	1	$\rho^2 \cdot \text{var}(T_1) + 1 - \rho^2$	1	(146)
$\mathbb{E}[X_1 X_2]$	ρ	$\rho \cdot \mathbb{E}[T_1^2]$	$\rho \cdot \mathbb{E}[C_1^2]$	(147)
$\text{cov}(X_1, X_2)$	ρ	$\rho \cdot \text{var}(T_1)$	$\rho \cdot \mathbb{E}[C_1^2]$	(148)
$\text{corr}(X_1, X_2)$	ρ	$\rho \cdot \left[\rho^2 + \dfrac{1 - \rho^2}{\text{var}(T_1)}\right]^{-1/2}$	$\rho \cdot \dfrac{\mathbb{E}[C_1^2]}{\sqrt{\text{var}(C_1)}}$	(149)

Appendix D. PFO Monotonicity Constraints

Similar to the two monotonicity constraints for the standard parabolic fractal given by (53) and (54) in Section 3.1.2, the quadratic functional form requires two constraints to force the fitted curve to be non-increasing. To start, substitute the definition of orthogonalized term from (50) into the regression equation from (51).

$$\frac{\ln(x)}{\sigma_x} = \beta_0 + \beta_1 \frac{\ln(r)}{\sigma_r} + \beta_2 \frac{\ln^2(r) - \hat{\gamma}_0 - \hat{\gamma}_1 \ln(r)}{\sigma_e} + \varepsilon \tag{150}$$

As with the standard parabolic fractal regression, the constraints require the first derivative (151) to be non-positive over the relevant domain.

$$d(r) = \frac{d\ln(x)/\sigma_x}{d\ln(r)} = \frac{\beta_1}{\sigma_r} + \beta_2 \frac{2\ln(r) - \hat{\gamma}_1}{\sigma_e} \tag{151}$$

Another result from the normalization of $\ln(r)$ is that the domain shifts as well. Thus, rather than being defined over $r \in [0.5, n - 0.5]$, the domain is now from $\ln(0.5)/\sigma_r$ to $\ln(n - 0.5)/\sigma_r$. This leads to the two following monotonicity constraints:

$$\frac{\beta_1}{\sigma_r} + \beta_2 \frac{2\ln(0.5)/\sigma_r - \hat{\gamma}_1}{\sigma_e} \leq 0 \tag{152}$$

$$\frac{\beta_1}{\sigma_r} + \beta_2 \frac{2\ln(n - 0.5)/\sigma_r - \hat{\gamma}_1}{\sigma_e} \leq 0 \tag{153}$$

Appendix E. Implied Weights from PFO Regression

Following the scaling of the fitted, linear curve for the Zipf distribution to obtain implied portfolio weights in Section 3.2.1, this appendix repeats the process to derive the implied weights from the orthogonalized parabolic fractal regression.

After orthogonalizing and scaling the variables to estimate the fit for the parabolic fractal distribution, the resulting regression from (51) is repeated in (154).

$$\frac{\ln(x)}{\sigma_x} = \beta_0 + \beta_1 \frac{\ln(r)}{\sigma_r} + \beta_2 \frac{\hat{u}}{\sigma_u} + \varepsilon \tag{154}$$

The first step involves multiplying both sides by σ_x.

$$\ln(x) = \sigma_x \beta_0 + \frac{\sigma_x \beta_1}{\sigma_r} \ln(r) + \frac{\sigma_x \beta_2}{\sigma_u} \hat{u} + \sigma_x \varepsilon \tag{155}$$

After taking the exponential of both sides, we obtain (156), which simplifies to (157) using the properties of exponentials.

$$\begin{align} x &= \exp\left(\sigma_x \beta_0 + \frac{\sigma_x \beta_1}{\sigma_r} \ln(r) + \frac{\sigma_x \beta_2}{\sigma_u} \hat{u} + \sigma_x \varepsilon\right) \tag{156} \\ &= \exp(\sigma_x \beta_0) \cdot r^{\sigma_x \beta_1 / \sigma_r} \cdot \exp\left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right) \cdot \exp(\sigma_x \varepsilon) \tag{157} \end{align}$$

As with the Zipf distribution, the fitted values are scaled by their sum to obtain the implied probability mass function (or weights). The first and last terms are both constants, and thus cancel out, which simplifies (158) to (159).

$$\begin{align} w &= \frac{\exp(\sigma_x \beta_0) \cdot r^{\sigma_x \beta_1 / \sigma_r} \cdot \exp\left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right) \cdot \exp(\sigma_x \varepsilon)}{\sum_{r=1}^{n} \exp(\sigma_x \beta_0) \cdot r^{\sigma_x \beta_1 / \sigma_r} \cdot \exp\left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right) \cdot \exp(\sigma_x \varepsilon)} \tag{158} \\ &= \frac{r^{\sigma_x \beta_1 / \sigma_r} \cdot \exp\left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right)}{\sum_{r=1}^{n} r^{\sigma_x \beta_1 / \sigma_r} \cdot \exp\left(\frac{\sigma_x \beta_2}{\sigma_u} \hat{u}\right)} \tag{159} \end{align}$$

Appendix F. BKFS Data Cleaning

The full loan table from the Black Knight Financial Services (BKFS) dataset contains loan-level information for 173,310,331 loans. Although the dataset includes loans from as early as October 1949, we remove observations with $ClosingMonth < 121$, which indicate loans originating prior to 1990. The BKFS data format is a monthly sequence where $ClosingMonth = 0$ for December 1979, $ClosingMonth = 1$ for January 1980, and so on. Additionally, 35 more observations with $ClosingMonth > 443$ are removed as these suggest loans originating after the November 2016 cutoff of the dataset.

Although the data includes a variable for the ZIP code, there are two primary reasons for conducting the analyses at the Core-Based Statistical Area (CBSA) level. First, the ZIP code variable is only reliable at the three-digit level. Despite the BKFS documentation indicating only the first three digits, there are some observations that contain all five digits. However, many simply have the first three, followed by 00. Also, ZIP codes are constructed such that their population sizes are fairly consistent, and thus, larger cities simply have more ZIP codes. Thus, for the approach in this paper, the CBSA geography level is most appropriate.

The BKFS variable, $CBSA\_MetroDivId$, provides the identifier for the CBSA of each property securing the respective loans. However, for the 11 CBSAs that are broken into Metropolitan Divisions, the variable provides the division code instead of the CBSA code. These divisions are aggregated into their respective CBSAs using the July 2015 delineation file from the U.S. Census. Although there have been more recent updates to these delineations, the July 2015 update is the most recent prior to our obtainment of the dataset. Thus, any changes do not impact the assignment. For example, the Chicago-Naperville-Evanston, IL, Metropolitan Division code was changed from 16974 to 16984 in the Sept. 2018 update.

Additionally, observations are removed if there are missing or non-positive original loan balances or terms (or terms longer than 480 months). This leaves a final set of 150,468,530 loans spanning 929 CBSAs.

Appendix G. BBX Variable Descriptions

Table G.1. Variable Descriptions from BBX Dataset

Variable	Description
LoanID	This is the unique asset identifier that is generated by concatenating the Deal Id and the data provider's supplied Loan Number. In the general case the loan number is the identifier supplied by the servicer and carried by the trustee.
ActualBalance	Amount of loan outstanding at the end of the remit period from the perspective of the borrower. This amount can differ from the ending scheduled balance if the servicer has advanced principal payments on the loan.
OrigAppraisalValueCalc	Cleansed or derived estimate of the property value at the time of loan origination, as supplied by the data provider.
FicoScoreOriginationCalc	Cleansed Fair Isaacs borrower credit score at the time of loan closing.
DocTypeSummary	A normalized code across providers that indicates the amount of income documentation provided by the borrower.
IntRtTypeSummary	Specifies whether the coupon on the loan is fixed or adjustable.
OriginalTermCalc	The cleansed or derived number of months between the first payment date and the date the principal is due from the borrower.
PropertyCityCalc	Cleansed or derived municipality that the property is located in.
PropertyStateCalc	Cleansed or derived two character state code that indicates the state that the property is located in.
DelinqStatus	Cleansed or derived amount of time between when the borrower last made scheduled payments and the current remittance period as measured in days using the MBA delinquency calculation.

Appendix H. Variable Definitions

Table H.1. Mortgage Default Model Variable Definitions

Variable	Definition
FICO	Borrowers' origination FICO credit score
CLTV	Current loan-to-value ratio as estimated in Section 4.1.1
Fulldoc	1 if borrower provided full documentation, 0 otherwise
ARM	1 if adjustable rate mortgage, 0 otherwise (fixed rate)
Term	1 if loan term is 15-years, 0 if loan-term is 30-years
Default	1 if borrower is 90+ days delinquent, 0 otherwise
04	1 if observation occurs in 2004, 0 otherwise
05	1 if observation occurs in 2005, 0 otherwise
06	1 if observation occurs in 2006, 0 otherwise
07	1 if observation occurs in 2007, 0 otherwise
08	1 if observation occurs in 2008, 0 otherwise
09	1 if observation occurs in 2009, 0 otherwise
10	1 if observation occurs in 2010, 0 otherwise
11	1 if observation occurs in 2011, 0 otherwise
12	1 if observation occurs in 2012, 0 otherwise
13	1 if observation occurs in 2013, 0 otherwise
14	1 if observation occurs in 2014, 0 otherwise

Definitions for the variables used in the mortgage default models throughout Chapter 4. Summary statistics for these variables can be found in Table 4.1.

References

Ackerman, A. and K. Davidson (2019a, Sept. 30). Fannie, Freddie to Retain Earnings. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/fannie-freddie-to-retain-earnings-11569851912.

Ackerman, A. and K. Davidson (2019b, Sept. 5). Trump Administration Aims to Privatize Fannie Mae and Freddie Mac. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/trump-administration-aims-to-privatize-fannie-mae-and-freddie-mac-11567717213.

Aitkin, M. A. (1964). Correlation in a singly truncated bivariate normal distribution. *Psychometrika 29*(3), 263–270.

Ambrose, B. W., J. Conklin, and J. Yoshida (2016). Credit rationing, income exaggeration, and adverse selection in the mortgage market. *The Journal of Finance 71*(6), 2637–2686.

Andersson, F., S. Chomsisengphet, D. Glennon, and F. Li (2013). The changing pecking order of consumer defaults. *Journal of Money, Credit and Banking 45*(2-3), 251–275.

Anselin, L. (2016). Variogram. University of Chicago Lecture Notes. [online]. https://spatial.uchicago.edu/sites/spatial.uchicago.edu/files/5_variogram_r.pdf.

Archer, W. R., P. J. Elmer, D. M. Harrison, and D. C. Ling (2002). Determinants of multifamily mortgage default. *Real Estate Economics 30*(3), 445–473.

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen 59*, 74–76.

Axtell, R. L. (2001). Zipf distribution of us firm sizes. *Science 293*(5536), 1818–1820.

Bailey, M. J., R. F. Muth, and H. O. Nourse (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association 58*(304), 933–942.

Bailey, N., S. Holly, and M. H. Pesaran (2016). A two-stage approach to spatio-temporal analysis with strong and weak cross-sectional dependence. *Journal of Applied Econometrics 31*(1), 249–280.

Bailey, N., G. Kapetanios, and M. H. Pesaran (2016). Exponent of cross-sectional dependence: Estimation and inference. *Journal of Applied Econometrics 31*(6), 929–960.

Bajari, P., C. S. Chu, and M. Park (2008, December). An empirical model of subprime mortgage default from 2000 to 2007. Working Paper 14625, National Bureau of Economic Research.

Bhutta, N., J. Dokko, and H. Shan (2017). Consumer ruthlessness and mortgage default during the 2007 to 2009 housing bust. *Journal of Finance 72*(6), 2433–2466. Forthcoming.

Birnbaum, Z. W. (1950). Effect of linear truncation on a multinormal population. *The Annals of Mathematical Statistics 21*(2), 272–279.

Birnbaum, Z. W., E. Paulson, and F. C. Andrews (1950). On the effect of selection performed on some coordinates of a multi-dimensional population. *Psychometrika 15*(2), 191–204.

Board of Governors of the Federal Reserve System (US) (2019). Delinquency Rate on Single-Family Residential Mortgages, Booked in Domestic Offices, All Commercial Banks [DRSFRMACBS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DRSFRMACBS, June 14, 2019.

Bogin, A., W. Doerner, and W. Larson (2018). Local house price dynamics: New indices and stylized facts. *Real Estate Economics*.

Brakman, S., H. Garretsen, C. Van Marrewijk, and M. Van Den Berg (1999). The return of Zipf: towards a further understanding of the rank-size distribution. *Journal of Regional Science 39*(1), 183–213.

Bureau of Labor Statistics (2019). Percent unemployed in labor force in the United States [cpsaat01]. Current Population Survey. https://www.bls.gov/cps/cpsaat01.htm.

Campbell, J. Y. and J. F. Cocco (2015). A model of mortgage default. *The Journal of Finance 70*(4), 1495–1554.

Case, K. E. and R. J. Shiller (1987). Prices of single-family homes since 1970: New indexes for four cities. *New England Economic Review* (Sep), 45–56.

Case, K. E. and R. J. Shiller (1989). The efficiency of the market for single-family homes. *The American Economic Review 79*(1), 125–137.

Chaney, T. (2018). The gravity equation in international trade: An explanation. *Journal of Political Economy 126*(1), 150–177.

Cheng, P. and S. E. Roulac (2007). Measuring the effectiveness of geographical diversification. *Journal of Real Estate Portfolio Management 13*(1), 29–44.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011). Weak and strong cross-section dependence and estimation of large panels. *The Econometrics Journal 14*(1), C45–C90.

Clauretie, T. M. and N. Daneshvary (2009). Estimating the house foreclosure discount corrected for spatial price interdependence and endogeneity of marketing time. *Real Estate Economics 37*(1), 43–67.

Corgel, J. B. and G. D. Gay (1987). Local economic base, geographic diversification, and risk management of mortgage portfolios. *Real Estate Economics 15*(3), 256–267.

Cotter, J., S. Gabriel, and R. Roll (2014). Can housing risk be diversified? a cautionary tale from the housing boom and bust. *The Review of Financial Studies 28*(3), 913–936.

D'Acunto, F. and A. G. Rossi (2019, July). Regressive mortgage credit redistribution in the post-crisis era. Robert H. Smith School Research Paper No. RHS 2833961.

Daneshvary, N., T. Clauretie, and A. Kader (2011). Short-term own-price and spillover effects of distressed residential properties: The case of a housing crash. *Journal of Real Estate Research 33*(2), 179–207.

Deng, Y., J. M. Quigley, and R. Order (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica 68*(2), 275–307.

Dombrowski, T. P., R. K. Pace, and R. P. Narayanan (2020). Mortgage portfolio diversification in the presence of cross-sectional and spatial dependence. *Advances in Econometrics 41*, 383–411.

Dubin, R. A. (1998). Spatial autocorrelation: A primer. *Journal of Housing Economics 7*(4), 304–327.

Dubin, R. A., R. K. Pace, and T. Thibodeau (1999). Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature 7*(1), 79–95.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review 94*(5), 1429–1451.

Eeckhout, J. (2009). Gibrat's law for (all) cities: Reply. *American Economic Review 99*(4), 1676–83.

Fischer, M., R. Füss, and S. Stehle (2019). Local house price comovements. University of St.Gallen, School of Finance Research Paper No. 2019/06 . Available at SSRN: https://ssrn.com/abstract=3403551 or http://dx.doi.org/10.2139/ssrn.3403551.

Foote, C. L., K. Gerardi, and P. S. Willen (2008). Negative equity and foreclosure: Theory and evidence. *Journal of Urban Economics 64*(2), 234–245.

Freddie Mac (2019). 30-Year Fixed Rate Mortgage Average in the United States [MORTGAGE30US], retrieved from FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/MORTGAGE30US.

Gabaix, X. (1999). Zipf's law for cities: an explanation. *The Quarterly Journal of Economics 114*(3), 739–767.

Gabaix, X. (2009). Power laws in economics and finance. *Annual Review of Economics 1*(1), 255–294.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica 79*(3), 733–772.

Gabaix, X. and R. Ibragimov (2011). Rank- 1/2: a simple way to improve the OLS estimation of tail exponents. *Journal of Business & Economic Statistics 29*(1), 24–39.

Gibrat, R. (1931). *Les inégalités économiques; applications: aux inégalités des richesses, á la concentration des entreprises, aux populations des villes, aux statistiques des families, etc., d'une loi nouvelle, la loi de l'effet proportionnel.* Paris: Librairie du Recueil Sirey.

Golub, G. H., M. Heath, and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics 21*(2), 215–223.

Gourieroux, C. and J. Jasiak (2002). Nonlinear autocorrelograms: An application to inter-trade durations. *Journal of Time Series Analysis 23*(2), 127–154.

Guiso, L., P. Sapienza, and L. Zingales (2013). The determinants of attitudes toward strategic default on mortgages. *The Journal of Finance 68*(4), 1473–1515.

Hartley, D. (2014). The effect of foreclosures on nearby housing prices: Supply or dis-amenity? *Regional Science and Urban Economics 49*, 108 – 117.

Heaney, R. and S. Sriananthakumar (2012). Time-varying correlation between stock market returns and real estate returns. *Journal of Empirical Finance 19*(4), 583–594.

Hoerl, A. E. (1959). Optimum solution of many variables equations. *Chemical Engineering Progress 55*(11), 69–78.

Hoerl, A. E. and R. W. Kennard (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics 12*(1), 69–82.

Hoerl, A. E. and R. W. Kennard (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*(2), 65–70.

Hoogstra, G. J., J. van Dijk, and R. J. G. M. Florax (2017). Do jobs follow people or people follow jobs? a meta-analysis of carlino-mills studies. *Spatial Economic Analysis 12*(4), 357–378.

Ioannides, Y. M. and H. G. Overman (2003). Zipf's law for cities: an empirical examination. *Regional Science and Urban Economics 33*(2), 127–137.

Kasarda, J. D. and W.-F. P. Shih (1977). Optimal bias in ridge regression approaches to multicollinearity. *Sociological Methods & Research 5*(4), 461–470.

Klass, O. S., O. Biham, M. Levy, O. Malcai, and S. Solomon (2006). The Forbes 400 and the Pareto wealth distribution. *Economics Letters 90*(2), 290–295.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2000). *Continuous Multivariate Distributions* (2nd ed.), Volume 1 of *Wiley Series in Probability and Statistics.* Wiley-Interscience.

Kuethe, T. H., K. A. Foster, and R. J. G. M. Florax (2008). A spatial hedonic model with time-varying parameters: A new method using flexible least squares. *Selected paper in the American Agricultural Economics Association 2008 Annual Meeting, Orlando, FL. http://ageconsearch.umn.edu/handle/6306*.

Laherrère, J. (1996). Distributions de type fractal parabolique dans la nature. *Comptes rendus de l'Académie des sciences. Série 2. Sciences de la terre et des planètes 322*(7), 535–541.

Laherrère, J. and D. Sornette (1998, Apr). Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B - Condensed Matter and Complex Systems 2*(4), 525–539.

LeSage, J. P. and R. K. Pace (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC Press.

Levy, M. (2009). Gibrat's law for (all) cities: Comment. *American Economic Review 99*(4), 1672–75.

Lin, Z., E. Rosenblatt, and V. W. Yao (2009). Spillover effects of foreclosures on neighborhood property values. *The Journal of Real Estate Finance and Economics 38*(4), 387–407.

Malevergne, Y., V. Pisarenko, and D. Sornette (2011). Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Physical Review E 83*(3), 036111.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance 7*(1), 77–91.

Marquardt, D. W. and R. D. Snee (1975). Ridge regression in practice. *The American Statistician 29*(1), 3–20.

Mayer, C., K. Pence, and S. M. Sherlund (2009). The rise in mortgage defaults. *The Journal of Economic Perspectives 23*(1), 27–50.

McCollum, M. N., H. Lee, and R. K. Pace (2015). Deleveraging and mortgage curtailment. *Journal of Banking & Finance 60*, 60–75.

Muthén, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology 43*(1), 131–143.

Nishiyama, Y., S. Osada, and Y. Sato (2008). OLS estimation and the t test revisited in rank-size rule regression. *Journal of Regional Science 48*(4), 691–716.

Pace, R. K. (1997). Performing large spatial regressions and autoregressions. *Economics Letters 54*(3), 283–291.

Paletta, D. (2019, Oct. 2). Federal government has dramatically expanded exposure to risky mortgages. *The Washington Post*. https://www.washingtonpost.com/business/economy/federal-government-has-dramatically-expanded-exposure-to-risky-mortgages/2019/10/02/d862ab40-ce79-11e9-87fa-8501a456c003_story.html.

Pesaran, M. H. (2004). General diagnostic tests for cross section dependence in panels. *CESifo Working Paper Series 1229. Available from: https://ssrn.com/abstract=572504*.

Pesaran, M. H. (2007). A simple panel unit root test in the presence of cross-section dependence. *Journal of Applied Econometrics 22*(2), 265–312.

Pesaran, M. H. (2015). Testing weak cross-sectional dependence in large panels. *Econometric Reviews 34*(6-10), 1089–1117.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review 21*(5), 1112–1130.

Reed, W. J. (2002). On the rank-size distribution for human settlements. *Journal of Regional Science 42*(1), 1–17.

Sinnott, R. W. (1984). Virtues of the haversine. *Sky and Telescope 68*(2), 159.

Smith, T. E. (2020). Notebook on Spatial Data Analysis. University of Pennsylvania Lecture Notes. [online]. https://www.seas.upenn.edu/∼ese502/#notebook.

Urban Institute Housing Finance Policy Center (2019, Dec.). Housing Finance at a Glance: A Monthly Chartbook. [online]. https://www.urban.org/policy-centers/housing-finance-policy-center/projects/housing-finance-glance-monthly-chartbooks.

U.S. Department of Housing and Urban Development (2019, Q3). FHA-Insured Single-Family Mortgage Market Share Report. [online]. https://www.hud.gov/program_offices/housing/rmra/oe/rpts/fhamktsh/fhamktqtrly.

Vinod, H. D. (1978). A survey of ridge regression and related techniques for improvements over ordinary least squares. *The Review of Economics and Statistics 60*(1), 121–131.

Zipf, G. K. (1949). Human behavior and the principle of least effort. *Addison Wesley, Cambridge, Mass.*.

Vita

Timothy Dombrowski earned an Associate in Arts from Pasco-Hernando State College in 2012 as a dual enrollment student at Pasco High School in Dade City, FL. Subsequently, he attended Saint Leo University and completed a Bachelor in Arts double-major in accounting and mathematics in 2015. After entering the Finance Department at the E.J. Ourso College of Business in the fall of 2015 as a candidate for the degree of Doctor of Philosophy in finance, this dissertation is being submitted during the spring 2020 semester of Louisiana State University for consideration to be awarded in May 2020. The major topics of interest for this research are the examination of various spatial considerations when modeling the risk of a pool of mortgage loans. Upon completion of this degree, Mr. Dombrowski has accepted a position as Assistant Professor of Finance at the University of Missouri – St. Louis.