Contagion in Subprime Mortgage Defaults : a Composite Likelihood Approach

Andréas Heinen[*]
Jim Kau[†]
Donald C. Keenan[*]
Mi Lim Kim[*]
Carlos Slawson[‡§]

January 6, 2014

Abstract

Using a composite likelihood approach, we analyze the pairwise dependence of defaults within a set of securitized subprime mortgages originated in Los Angeles between 2000 and 2011. As the main factors affecting default dependence, we propose geographic proximity as well as the similarity of mortgages in terms of various other time-varing economic variables. Thus, in addition to geographic distance, we employ measures of non-geographic distance, in terms of both individual mortgages and their neighborhood characteristics. For physical distance, we use a squared exponential correlation function, a special case of the Matérn function. Our results show that physical distance has a strongly significant effect on default dependence. Furthermore, even after controlling for physical distance, a number of these non-geographic measures prove significant in explaining default dependence.

PRELIMINARY AND INCOMPLETE.
PLEASE DO NOT QUOTE WITHOUT PERMISSION

[*] THEMA, Université de Cergy-Pontoise

[†] Terry College of Business, University of Georgia

[‡] Department of Finance, Louisiana State University

[§] The authors gratefully acknowledge the support of the QUANTVALLEY/FdR: 'Quantitative Management Initiative. The usual disclaimers apply.

1 Introduction

The subprime credit crisis beginning in 2007 has had a profound effect, not just on the U.S. economy but on the world economy. An understanding of the nature of mortgage default, and in particular of how such defaults tend to occur together, with a resulting accumulated effect of great macroeconomic consequence, is an urgent task. At a more general, but technical level, dependence between defaults is an important dimension for credit risk of all kinds. Even with the individual probabilities of default taken as given, a pool of mortgages and derivative products written on it are riskier when the defaults become more dependent. Default dependence is central to the risk management of debt portfolios, and to the design, pricing, and risk management of securitized credit products, such as mortgage backed securities (MBS),[1] which can be viewed as derivatives on the default probabilities and correlations of their underlying assets (see Duffie & Singleton 2003, Lando 2004, Schoenbucher 2003).

In this paper, we analyze the default dependence within a large data set of individual U.S. nonconforming securitized mortgages originated in Los Angeles between 2000 and 2011. Our data both predate and include the subprime crisis. We propose, as the main factors affecting default dependence, geographic proximity, similarity of characteristics of mortgages and other time-varying economic variables. To determine the effect of these factors, we first use a multinomial logistic framework in order to estimate the default probability of each mortgage. Second, we rely on a copula model to provide a flexible dependence structure. To capture the effect of physical distance on dependence, we use a squared exponential correlation function, a special case of the Matérn function. Finally, given the difficulty of specifying a joint distribution of default, we rely on a bivariate composite likelihood approach to estimate the dependence parameters.

The main contribution of this paper is to explicitly model the default dependence of mortgages, where in addition to considering macroeconomic variables, we take into account

[1] Mortgage backed securities (MBSs) are assets whose pay-off depends on large pools of underlying mortgages.

the mortgages' geographic and non-geographic proximities. While default dependence of mortgages is an important part of the risk of pools of mortgages and mortgage backed securities (MBS), there have been very few studies on its determinants. This paper aims to fill this gap. Using a large portfolio of residential subprime loans from an anonymous subprime lender, Cowan & Cowan (2004) show that the correlation of defaults increases as the internal credit rating of mortgages declines. However, they refrain from modeling default probabilities, and rely instead on realized defaults.

Several papers in the real estate literature take account of spatial dependence. For instance, Case, Clapp, Dubin & Rodriguez (2004) and Bourassa, Cantoni & Hoesli (2007) study spatial dependence in house prices to aid in their prediction. Deng, Pavlov & Yang (2005) incorporate space-varying dependence in the residuals of a competing hazard model of mortgage termination with refinance, sale and default. They use a space-varying coefficient method to estimate the spatial dependence between the unobserved random effects of their competing hazards. Their assumption, that mortgages with similar characteristics tend to cluster in space, improves the performance of their model. Kau, Keenan & Li (2011) model termination risks of mortgages using a shared frailty model. They show that mortgages from the same MSA have higher correlations than mortgages originated from different MSAs, and that geographic effects are heterogeneous across MSAs. While these papers consider dependence in default probabilities, they attribute it exclusively to spatial proximity. In our approach, we explicitly consider both geographic and non-geographic distance of mortgages to measure their dependence. Moreover, we use copulas instead of considering the dependence as a nuisance parameter.

Conley & Topa (2002) use a similar approach in a different context. More specifically, they examine the spatial patterns of unemployment in Chicago, and use several distance metrics based on ethnicity, occupation and travel time, to estimate non-parametric estimates of autocorrelations in unemployment rates across census tracts. However, the method they use does not allow them to estimate the effect of more than two distance metrics at the same time. This paper more clearly shows the multidimensional interaction among distance metrics.

It is not a trivial task, however, to estimate dependence of default hazard using copulas, since building a multivariate dependence structure to incorporate the spatial correlation is not feasible when using a data set as large as ours. We use instead a composite likelihood approach, which allows us to estimate the dependence structure by modeling and then adding up the lower dimensional margins.Paik & Ying (2012) employ a Cox proportional hazard model with a Farlie Gumbel Morgenstern (FGM) copula with a composite likelihood in order to treat spatially correlated survival data using pairwise distributions. They assess whether geographic distance is an appropriate metric to describe dependence. This paper extends previous research through the use of copulas different than the FGM as well as different functional forms of distances.

The remainder of this paper is organized as follows. In section 2, we explain both the multinomial logit model we use the marginal estimation as well as the copula and the Matérn function we use to treat dependence. Section 3 explains our two-step estimation procedure based on the composite likelihood approach. Section 4 describes the data and the results. Section 5 concludes.

2 The Model

In this section, we first introduce the logistic regression model we use for the marginal estimation. We then introduce the copula models we apply to measure the dependence between pairs of mortgages across geographical distance. We also discuss the squared exponential correlation function, which describes the nonlinear effect of distance on the pairwise dependence between mortgages.

2.1 Multinomial logistic (MNL) regression model

Since our aim is to measure the correlation of default probabilities of mortgages across distance, we need to first estimate the default probability of each mortgage. At any given point in time, each mortgage faces three different risks: prepayment, default or censoring. There are several possible models for survival data with grouped durations, such as proportional hazard (PH), probit and logit models. Sueyoshi (1995) compares these three models in terms of the their baseline hazard and finds that duration (PH) models and binary response models (probit and logit) produce very similar results. Corrente, Chalita & Moreira (2003) also compare a grouped duration Cox proportional hazard model with a logit model, and find that both types of models give similar values for the Akaike Information criterion (AIC), an indicator of the quality of the fit, and also for the deviance of the model, which is akin to a residual in a linear regression. It is thus difficult to say that one model is better than the other. However, the logit models perform better in terms of standardized Pearson residuals. An additional complication that we need to deal with is the competing risk nature of our data. While we are mainly interested in default, households can also prepay or their behavior be censored, which is the case when by the end of our sample period, neither of the other two outcomes has occurred. Therefore we consider a right-censored multinomial logistic regression with two competing risks: prepayment and default (see for instance Liu (2012), Chapter 7). This model has been applied before in the context of default and prepayment in real estate (see e.g. Clapp, Goldberg, Harding & LaCour-Little 2001, Clapp, Deng & An 2006)

Let variable Y denote our multichotomous response variable with $J = 3$ possible outcomes (1 for *censoring*, 2 for *default*, 3 for *prepayment*) and x a vector of p explanatory variables. The probability that mortgage i at time t experiences event k, for $i = 1, \ldots, n$, and $k = 1, \ldots, J$, can be expressed as

$$\pi^{(k)}(x_{it}) = \frac{e^{x_{it}\beta_k}}{1 + \sum_{j=2}^{J} e^{x_{it}\beta_j}}, \tag{1}$$

where $k = 1$ is the reference event.[2] The loglikelihood function can thus be written as:

[2] To identify the model, we impose $\beta_1 = 0$ for the reference event. Thus all coefficients should be interpreted as effects relative to the reference case, which is censoring.

$$L_m(\beta) = \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=1}^{3} \delta_{itk} \log \left(\pi^{(k)}(x_{it})\right), \tag{2}$$

where $\delta_{itk} = \mathbb{1}_{\{Y_{it}=k\}}$ indicates that the i-th observation at time t experiences the k-th event, being zero otherwise.

2.2 Copulas

Copulas are a convenient tool to separate the dependence between variables from their marginal distributions, which is particularly convenient in non-Gaussian contexts. They have a long history, partly associated with modeling dependent durations (see Gumbel 1960, Gumbel 1961). Copulas have become a standard tool in finance, used to capture the dependence in returns as well as in the context of credit risk analysis. They have previously been used in real estate as well, to model the dependence between house prices in different metropolitan statistical areas (MSAs) (see e.g. Goorah 2007, Zimmer 2012). The use of copulas relies on the celebrated Sklar (1959) theorem, which states that a joint cumulative distribution function (CDF) F of a vector of n variables $(Y_1, \ldots, Y_n)$ can be written in terms of a copula function C with dependence parameter θ, whose arguments are the n marginal distribution functions F_i, $i = 1, \ldots, n$:

$$F(y_1, \ldots, y_n) = C(F_1(y_1), \ldots, F_n(y_n); \theta). \tag{3}$$

The joint probability density function (PDF), f obtains by differentiation, and can be written as a product of the marginals and of a copula density term, which captures all the dependence between the variables:

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_i(y_i) c(F_1(y_1), \ldots, F_n(y_n); \theta). \tag{4}$$

In the case of continuous marginals, copulas are uniquely identified, and they fully characterize the dependence between the variables, as well as rank correlation measures, such as

Kendall's tau, that are invariant to any strictly increasing transformation of the data. For instance, according to its analytical definition, Kendall's tau between a pair of variables (Y_i, Y_j) is given by

$$\tau(Y_i, Y_j) = P\{(Y_{i1} - Y_{i2})(Y_{j1} - Y_{j2}) > 0\} - P\{(Y_{i1} - Y_{i2})(Y_{j1} - Y_{j2}) < 0\}, \qquad (5)$$

where $(Y_{i1}, Y_{j1}), (Y_{i2}, Y_{j2})$ are two observations of random variables (Y_i, Y_j). The pairs are said to be *concordant* whenever $(Y_{i1} - Y_{i2})(Y_{j1} - Y_{j2}) > 0$, and *discordant* whenever $(Y_{i1} - Y_{i2})(Y_{j1} - Y_{j2}) < 0$. Alternatively, we can use the analytical definition of Kendall's tau, as a function of the copula CDF C:

$$\tau(C) = \int_0^1 \int_0^1 C(u_i, u_j) dC(u_i, u_j). \qquad (6)$$

With continuous marginals, the two definitions coincide.[3]

Copulas are uniquely identified only on the product range of the marginal CDFs. In the Bernoulli case, the CDF only takes on two values: $Ran\{F_i\} = \{\bar{\pi}_i, 1\}$, which corresponds to outcomes $\{0, 1\}$ of the variables, where $\bar{\pi}_i = 1 - \pi_i$, and π_i is the Bernoulli parameter. Consequently the copula is uniquely identified only at these points. Moreover, rank correlation coefficients, such as Kendall's tau now depend both on the copulas and on the marginals (see e.g. Denuit & Lambert 2005, Nešlehová 2007, Genest & Nešlehová 2007). This is due to the possibility of draws in discrete data, i.e. $P\{(Y_{i1} - Y_{i2})(Y_{j1} - Y_{j2}) = 0\} > 0$. With binary marginals, the probabilistic version of Kendall's tau can be written as

$$\tau_{ij} = 2\left(C(\bar{\pi}_i, \bar{\pi}_j; \theta) - \bar{\pi}_i \bar{\pi}_j\right), \qquad (7)$$

which illustrates the dependence of tau on the marginals $\bar{\pi}_i$ and $\bar{\pi}_j$ and on the copula C. Unlike in the case of continuous margins, when the margins are discrete, Kendall's tau is no longer in the (-1,1) interval.[4] However for binary variables, there exists a rescaled version

[3] For more on the probabilistic and analytical definitions of Kendall's tau, see Genest & Nešlehová (2007).

[4] It can be shown that when the dependence goes from the Fréchet-Höffding lower to upper bound (from perfect negative to perfect positive dependence), tau goes through the range $(-\frac{1}{2}, \frac{1}{2})$.

of Kendall's tau, which is in the familiar (-1,1) range. This is called Goodman's gamma (see Goodman & Kruskal 1954), and it is defined as:

$$\gamma_{ij} = \frac{\tau_{ij}}{2\xi_{ij}(\theta)}, \tag{8}$$

where $\xi_{ij}(\theta) = 2C(\bar{\pi}_i, \bar{\pi}_j; \theta)^2 + \bar{\pi}_i \bar{\pi}_j + C(\bar{\pi}_i, \bar{\pi}_j; \theta)(-3 + 2\pi_i + 2\pi_j)$.

As discussed by Genest & Nešlehová (2007), the differences in interpretation of rank correlation and the the lack of unicity of the copula when the marginals are discrete do not invalidate the use of copulas in this context, since even with discrete marginals, the copula parameter retains its interpretation as a measure of association. There is indeed a tradition in the statistics literature of using copula with binary data. For instance, Meester & MacKay (1994) propose a model for binary data based on Archimedean copulas, such as the Frank copula for exchangeable dependence, while Gauvreau & Pagano (1997) use the Farlie Gumbel Morgenstern (FGM) copula, whereas Molenberghs & Lesaffre (1994) advocate the use of a multivariate extension of the Plackett copula for multivariate ordinal data. Finally Song (2000) builds multivariate distributions from a Gaussian copula whose marginals are dispersion models, a class that includes discrete distributions, such as the Poisson and the Bernoulli.

In the discrete case, the probability mass function (PMF), which is the discrete counterpart of the PDF, obtains by finite differencing of Equation (3):

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \sum_{\nu \in \mathcal{S}} \text{sign}(\nu) \quad C(F_1(\nu_1), \ldots, F_n(\nu_n, \theta)), \tag{9}$$

where the sum is over all $\nu = (\nu_1, \ldots, \nu_n) \in \mathcal{S} = \prod_{i=1}^{n} \{y_i, y_i - 1\}$ and $\text{sign}(\nu) \in \{-1, 1\}$ equals 1 if and only if $\#\{k : \nu_k = y_k - 1\}$ is even. The copula parameter θ can still be meaningfully estimated by maximum likelihood (ML). With a pair of Bernoulli variables (Y_i, Y_j), the contributions to the PMF are given as:

$$
\begin{aligned}
P(Y_i = 1, Y_j = 1) = &\quad 1 - \bar{\pi}_i - \bar{\pi}_j + C(\bar{\pi}_i, \bar{\pi}_j, \theta), \\
P(Y_i = 1, Y_j = 0) = &\quad \bar{\pi}_i - C(\bar{\pi}_i, \bar{\pi}_j, \theta), \\
P(Y_i = 0, Y_j = 1) = &\quad \bar{\pi}_j - C(\bar{\pi}_i, \bar{\pi}_j, \theta), \\
P(Y_i = 0, Y_j = 0) = &\quad C(\bar{\pi}_i, \bar{\pi}_j, \theta).
\end{aligned}
\tag{10}
$$

With the previous result, we can then define the loglikelihood contribution of a single observation:

$$
\begin{aligned}
L(Y_i, Y_j) \quad = \quad & \delta_{i1}\delta_{j1} log\left(1 - \bar{\pi}_i - \bar{\pi}_j + C(\bar{\pi}_i, \bar{\pi}_j, \theta)\right) + (1 - \delta_{i1})(1 - \delta_{j1}) log\left(C(\bar{\pi}_i, \bar{\pi}_j, \theta)\right) \\
& + \delta_{i1}(1 - \delta_{j1}) log\left(\bar{\pi}_i - C(\bar{\pi}_i, \bar{\pi}_j, \theta)\right) + (1 - \delta_{i1})\delta_{j1} log\left(\bar{\pi}_j - C(\bar{\pi}_i, \bar{\pi}_j, \theta)\right).
\end{aligned}
\tag{11}
$$

To close our model, we still need to specify how the dependence parameter of the copula varies with regressors. The problem, when estimating different copulas, is that their parameters are not directly comparable, since they are defined on different ranges. Therefore we map each copula parameter into its analytical Kendall's tau, as defined in Genest & Nešlehová (2007), and we model the effect of regressors and distance on the analytical Kendall's tau.[5] This has the advantage that the effect of the regressors will then all be on the same dependence measure that is in the $(-1, 1)$ range. A similar approach is followed by Nikoloulopoulos & Karlis (2008), who study a multivariate logit model, where the dependence is handled by a number of parametric copulas.

2.3 Squared exponential correlation function

In order to measure spatial dependence, we use a *squared exponential correlation* function, which is a special case of the Matérn correlation function. Matérn functions are used in geostatistical modeling (see Gneiting, Kleiber & Schlather 2010, Bai 2011).[6] Dubin (1998) introduces the Matérn function into the real estate literature to account for the

[5] Appendix A contains the analytical Kendall's tau as a function of the copula parameter for the copulas we use in this paper.

[6] For a more detailed coverage of geospatial covariance functions, see Rasmussen & Williams (2006), Part 4.

correlation between house prices in a prediction exercise. The Matérn function explains the decrease of correlation with distance. It expresses the dependence between any two mortgages separated by distance d as

$$k_{matern}(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} d}{\alpha} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} d}{\alpha} \right), \tag{12}$$

with positive parameters ν and α, where K_ν is a modified Bessel function of the second kind of order ν. When $\nu \to \infty$, we obtain the squared exponential correlation function:

$$k_{SE}(d) = \exp\left(-\frac{d^2}{2\alpha^2} \right). \tag{13}$$

Whenever we estimate the Matérn function with an upper bound on the parameters, the optimal parameter value for ν is always at the upper bound, which indicates that the best model is indeed the squared exponential function (see for instance Kazianka 2013). We follow Bai (2011) and use the center-to-center distance between our clusters. We assume that mortgages within the same cluster (zip code) are equidependent, with dependence parameter τ_1, and that the dependence of mortgage pairs across different clusters depends on their distance according to the quadratic correlation function introduced above. Bai (2011) assumes that the pairwise dependence of observation pairs across clusters is smaller than the dependence of observation pairs within a cluster. However, since we do not know the exact locations of the houses in our sample, we cannot exclude that a mortgage pair across different clusters could be more correlated than a pair within the same cluster. For example, even if two arbitrary mortgages belong to different zip codes, if they share a common border, it is quite possible that the distance between the mortgages will be smaller than the distance between mortgages within a zip code. This is why we allow for an additional parameter, τ_2, that multiplies the distance function, and that can be different from the within cluster dependence, τ_1. The Mat'ern function was originally designed to capture correlation, but since we are modeling dependence with copulas, we use it to parameterize the analytical Kendall's tau of each on of our copulas. Thus the analytical

Kendall's tau between mortgages i in cluster c and j in cluster d is given as:

$$\tau_{ic,jd} = \begin{cases} \tau_2 \exp\left(-\frac{d_{cd}^2}{2\alpha^2}\right) & \text{if } c \neq d \text{ and } i \neq j, \\ \tau_1 & \text{if } c = d \text{ and } i \neq j, \end{cases} \tag{14}$$

where d_{cd} is center-to-center distance between clusters c and d, $0 < \tau_1, \tau_2 < 1$ are dependence parameters, and $\alpha > 0$ is a spatial scaling parameter, which controls the speed with which correlation decreases with distance. The larger α, the faster the dependence decreases with distance.

2.4 Dependence across non-geographic distances

So far, our focus was on geographic distance. But geographic distance could also capture the effect of other sources of dependence. For instance, neighbors are likely to share similar characteristics in terms of income, jobs or credit history (see Deng et al. 2005) and are thus more likely to be correlated than mortgagors who live far away. Alternatively, geographic distance could also capture a contagion effect due to the occurrence of local shocks, such as foreclosures, leading to a collapse in house prices. Harding, Rosenblatt & Yao (2009) find that foreclosures reduce the prices of neighbors' non-distressed sales through such a contagion effect. More specifically, foreclosures of an immediate neighbor (located within 300 feet) induce a discount in house prices.

In order to incorporate the effect of other variables on the analytical Kendall's tau, we modify Equation (14) as follows:

$$\tau_{ic,jd} = \begin{cases} \tau_2 \nu_{ij} \exp\left(-\frac{d_{cd}^2}{2\alpha^2}\right) & \text{if } c \neq d \text{ and } i \neq j, \\ \nu_{ij} & \text{if } c = d \text{ and } i \neq j, \end{cases} \tag{15}$$

where

$$\nu_{ij} = \frac{\exp(\psi^T X_{ij})}{1 + \exp(\psi^T X_{ij})}, \tag{16}$$

and d_{cd}, α, τ_2 are as defined in Section 2.3, and X_{ij} is a matrix of explanatory variables. Equation (16) corresponds to a logit link function that maps regressors in the real line to positive dependence, in the $(0,1)$ interval. For each non-geographic measure, X_i, listed in Table 1, we define both the distance $D_{ic,jd}(X) = |X_{ic} - X_{jd}|/2$, and the average $A_{ic,jd}(X) = (X_{ic} + X_{jd})/2$. When $c = d$, X_{ij} consists of a vector of dimension $(n \times 1)$, whose elements are all 1s. If distance has a significant negative (positive) impact on dependence, this means that the dependence is higher (lower) for mortgages with similar characteristics. If the average across pairs has a positive impact, this implies that dependence is higher for higher values of the characteristic. Moreover, if the coefficients of the difference and average are both equal to α, then this is equivalent to $\alpha\ max(X_{ic}, X_{jd})$, which means that only the mortgage in the pair, with the largest value of characteristic X matters for dependence, whereas it is only the smallest value of X that matters if the coefficients on distance and average are equal but of opposite sign.[7] Including both the distance and the average can capture a wide range of possible effects. For example, Figure 1 shows the pattern of dependence that emerges when distance has a negative, and average a positive effect on dependence. Besides non-geographic distance, we also consider the effect of the macroeconomic variables listed in Panel C of Table 1.

3 Estimation

3.1 Two-step estimation

Our interest in this paper focuses on the estimation of the determinants of the dependence between default risk of individual subprime mortgages. Thus, in principle, we should specify and estimate a full joint distribution for the default probabilities of all the mortgages in our sample. Given the number of mortgages we have in our data, this would be a daunting task. Instead, we rely on a composite likelihood approach. Composite likelihood (CL) consists in adding up lower-dimensional margins in situations where full multivariate modeling is infeasible, see Varin (2008) and Varin, Reid & Firth (2011) for reviews. A

[7] This follows from the fact that $\frac{1}{2}(X_{ic} - X_{jd}) + \frac{1}{2}|X_{ic} - X_{jd}| = \max(X_{ic}, X_{jd})$, while $\frac{1}{2}(X_{ic} - X_{jd}) - \frac{1}{2}|X_{ic} - X_{jd}| = \min(X_{ic}, X_{jd})$.

popular implementation of CL is bivariate composite likelihood (BCL), also called pairwise likelihood (PL), a sum of bivariate marginal models. Le Cessie & van Houwelingen (1994) introduce CL for bivariate and more generally clustered correlated binary data. Kuk & Nott (2000) analyzes BCL for clustered and longitudinal binary data. de Leon (2005) studies a BCL estimation like the one of Kuk & Nott (2000), but for the grouped continuous model, a model for multivariate ordinal data with a normally distributed latent variable.

It is common in copula modeling to separate the estimation of the margins from the copula, mostly for reasons of computational feasibility. Joe (2005) calls this inference for the margins (IFM). He studies the asymptotic relative efficiency (ARE)[8] of the two-step estimation procedure compared to the full maximum likelihood approach, and concludes that for discrete margins with few categories (we have three categories), the two-step estimator is highly efficient, and that efficiency deteriorates only slowly with the number of categories. The IFM method can also be applied to the composite likelihood framework. This is the approach we follow in this paper.

Zhao & Joe (2005) compare two-step and one-step estimation in BCL for clustered probit data with a normal copula, when there is correlation between individual units in the same cluster, but no cross-cluster correlation. Their two-step procedure is akin to the IFM method of Joe (2005), where the marginals are estimated first, and then the dependence parameter, conditional on the marginal parameters. They estimate common regression parameters for the means of the different units, and common dependence parameters for all bivariate margins. They find that for the multivariate probit model, the higher the within-cluster dependence, the lower the efficiency of the two-step method for the mean compared to a one-step method. However, while the one-step method performs well for marginal regression parameters, it is less efficient for the dependence parameters under weak dependence. Since the dependence between the mortgages in our data is not very large, even compared to the smallest dependence used by Zhao & Joe (2005), the loss of

[8] $ARE(\hat{\theta}_{CL}) = Avar(\hat{\theta}_{MLE})/Avar(\hat{\theta}_{CL})$, where $Avar(\hat{\theta}_{MLE})$ is obtained from the diagonals of the inverses of the Fisher information matrix and $Avar(\hat{\theta}_{CL})$ is obtained from the Godambe information matrix, and $Avar$ refers to the asymptotic variance.

efficiency stemming from our use of a two-step estimation method should be extremely reduced.

We are now ready to set up the likelihood functions for a two-step estimation. We decompose the likelihood into two parts. The loglikelihood of the marginal is as defined in Equation (1) and L_c is the composite likelihood method. For $\mathbf{Y} = (y_1, \ldots, y_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$, we can express the marginal log likelihood and the composite likelihood function as:

$$\begin{array}{rcll} L_m(\mathbf{Y}|\mathbf{x};\beta) & = & \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{k=1}^{3} & \delta_{ik}\log(\pi^{(k)}(x_i)), \\ L_c(\mathbf{Y}|\mathbf{x};\beta,\theta_c) & = & \sum_{t=1}^{T}\sum_{\substack{(i,j)\in D \\ i<j}} & \{\delta_{i1}\delta_{j1}log(1-\bar{\pi}_i-\bar{\pi}_j+C(\bar{\pi}_i,\bar{\pi}_j,\theta)) \\ & & & +(1-\delta_{i1})(1-\delta_{j1})\log(C(\bar{\pi}_i,\bar{\pi}_j,\theta)) \\ & & & +\delta_{i1}(1-\delta_{j1})\log(\bar{\pi}_i-C(\bar{\pi}_i,\bar{\pi}_j,\theta)) \\ & & & +(1-\delta_{i1})\delta_{j1}\log(\bar{\pi}_j-C(\bar{\pi}_i,\bar{\pi}_j,\theta))\}, \end{array} \tag{17}$$

where $\beta = (\beta_2, \beta_3)$ collects the parameters of the multinomial logit model for the competing risks of prepayment and default, respectively, $\theta_c = (\tau_1, \tau_2, \alpha)$ are the parameters of the dependence model, $\delta_{it} = \mathbf{1}_{\{Y_{it}=1\}}$, and D is a set of mortgage pairs. Estimation proceeds in two stages. In a first step, we maximize the likelihood of the margins and we derive $\hat{\beta}$. In a second step, we derive $\hat{\theta}_c$ by maximizing the composite likelihood, taking $\hat{\beta}$ as given. The composite likelihood estimator of the dependence parameters is asymptotically normal:

$$\sqrt{N}(\hat{\theta}_c - \theta_c) \sim N(0, \Gamma(\theta_c)), \tag{18}$$

where $\Gamma(\theta_c) = H(\theta_c)^{-1}G(\theta_c)H(\theta_c)^{-1}$ is the Godambe matrix, $H(\theta_c) = \mathbb{E}\left[-\frac{\partial^2 L_c(\theta_c)}{\partial\theta_c\partial\theta_c'}\right]$ is the Hessian, and $G(\theta_c) = \mathbb{E}\left[\frac{\partial L_c(\theta_c)}{\partial\theta_c}\frac{\partial L_c(\theta_c)}{\partial\theta_c'}\right]$ is the score.

3.2 Estimation of dependence across distances

3.2.1 Random mortgage pairs

According to the pairwise marginal likelihood function in Equation (17), there are $\sum_s n_s n_t$ mortgage pairs across different clusters s and t, and $\sum_{s \neq t} \frac{n_s(n_s-1)}{2}$ mortgage pairs from the same cluster s. The number of mortgages in LA is 31,962. If we consider the time history of each loan, the number of observations is 239,486. Thus, using all mortgage pairs is infeasible. Instead, in order to estimate the correlation of mortgages across geographic distance, we need to resort to a sampling strategy. Varin & Vidoni (2005) propose to exclude the pairs which are located too far away from each other in order to improve numerical and statistical efficiency of the estimation procedure. Varin et al. (2011) and Bai, Song & Raghunathan (2012) further show that excluding pairs that are too far apart increases efficiency. We can specify the set D of mortgages in Equation (17) as $D = D_{within} \bigcup D_{across}$, with

$$\begin{aligned} D_{within} &= \{(i,j) : i \neq j, s = t, i \in \Im_s, j \in \Im_s\}, \\ D_{across} &= \{(i,j) : 0 < ||s-t|| \leq d, i \in \Im_s, j \in \Im_t\}, \end{aligned} \tag{19}$$

where $\Im_s$ and $\Im_t$ are two sets of mortgages of cluster s and t. The set of mortgage pairs across different clusters, D_{across} depends on the cutoff distance d, which is the threshold for inclusion into the sample. The measure of efficiency that is commonly used in the literature to determine this optimal cutoff distance, is based on the trace of the Godambe matrix. The Godambe matrix is the variance covariance matrix of the estimate, and its trace sums the variances of all the parameters of the dependence model. In that context, the optimal maximum distance d is the one which minimizes $tr(\hat{\Gamma}(d))$.[9]

We test cutoffs of 30 km, 35 km and 40 km. For each cutoff we randomly select 20 million mortgage pairs from the whole data set, whose distances are less than the cutoff. We then select the common history of each mortgage pair, and we exclude those mortgage pairs, that lack a common history.[10] Our results in Table 5 show that the optimal threshold

[9] In other words, it is the cutoff that yields the sample with the most precisely estimated coefficients.

[10] Two mortgages share a common history if they are both active during at least one quarter.

is around 35 km.

3.2.2 Estimation with copula functions

First, we estimate the dependence across physical distance. We assume that all the mortgage pairs in the same cluster (zip code) have the same dependence τ_1,[11] while for the mortgage pairs which are from different clusters, the correlation is determined by their distance as well as the parameters $\hat{\tau}_2$ and $\hat{\alpha}$ via Equation (14). For ease of estimation, we split the likelihood function into one for mortgage pairs within a cluster, D_{within}, and one for mortgage pairs across clusters, D_{across}. This is possible since there is no common parameter between both groups.

With the estimated parameters, $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\alpha}$, we calculate a unique Kendall's tau for each mortgage pair. For every copula and every mortgage pair, we calculate out the value of the dependence parameter, and we combine this with the default probabilities of the marginal estimation in order to obtain pairwise Goodman's gammas, via Equation (8).[12] In order to see how our Goodman's gamma changes across distance, we average gammas for mortgage pairs falling in intervals proceeding by one-half kilometer to the maximum distance. We estimate Gumbel, Frank, FGM, rotated Gumbel and Clayton copulas, all of which describe different shpaes of dependence. The properties of these copulas are explained in Appendix A.

Second, we estimate dependence across non-geographic distance. We apply the squared correlation function for continuous variables, such as $Amount_{ij}$, $FICO_{ij}$, $Current\ LTV_{ij}$ and $Current\ contract\ rate_{ij}$. In addition, we try to assess the effect of various non-geographic variables one by one in Equation (16), in order to compare the results on the distance effect with the squared correlation function. We fix physical distance and estimate one after another the dependence of other distances so as to figure out the interaction between physical distance and these other distances. Finally, we put all distances into Equation

[11] This is the best we can do, given that we do know the location of the mortgages only up to their zip code.

[12] As explained in Section 2.2, Goodman's gamma replaces Kendall's tau for the purpose of comparing dependence across copulas because of the discreteness of the margins.

(16), to see whether the parameters of the various distance measures are jointly significant.

4 Data and Results

In this section, we first explain the data we use for the estimation. We then present the results for the marginal model. Finally, we introduce the dependence results for the mortgages in LA.

4.1 Data

We use data provided by *Black Box Logic, LLC.* The data set contains a large number of privately-securitized nonconventional mortgages originated from 1997 to 2008, covering about 80% of the value of all securitized subprime mortgages in the US. We focus on mortgages in the Metropolitan Statistical Area of Los Angeles, observed between the third quarter of 2000 and the second quarter of 2011. Using the same selection criteria as Kau, Keenan, Lyubimov & Slawson (2011) we narrow our focus on the subprime mortgages which proved problematic during financial crisis. More specifically, we retain 30-year single-family, residential adjustable rate mortgages (ARMs), located in the Los Angeles MSA, whose FICO credit scores were below 720 with loan-to-value ratios (LTVs) at or above 80%. We also use data on zip code-level income and the unemployment rate taken from the *2007-2011 American Community Survey from U.S. census Bureau.* We employ the monthly Case-Shiller house price index (HPI) at the MSA level to compute the quarterly HPI return for LA. This results in 31,962 loans, and 239,486 quarter-loan observations when each loan's time history is included. We focus on Los Angeles, since that MSA contains the largest number of loans satisfying the conditions mentioned above. Moreover, one of our interests lies in the spatial dependence between defaults, and since we expect this to be a local phenomenon, we do not expect that focusing on a single MSA leads to any significant loss of information. The data set includes the zip code in which the mortgages are located and we use this information to compute distances between zip codes.

4.2 Marginal Model

We estimate a multinomial logistic (MNL) regression model for the default probability of each loan in Los Angeles. Table 1 shows the explanatory variables we use. We estimate common regression parameters of the covariates for the marginal distribution of each mortgage. Our choice of covariates is in line with previous literatures on mortgage defaults (see e.g. Kau, Keenan, Lyubimov & Slawson 2011, Ashcraft, Goldsmith-Pinkham & Vickery 2010, Deng et al. 2005, Voicu, Jacob, Rengert & Fang 2012). We assume that mortgages are in default if they are either in foreclosure, real estate ownership (REO) or bankruptcy. As shown in Table 2, we classify covariates as either static variables, which usually describe the characteristics of the mortgages at origination, or as dynamic variables, which reflect time-varying characteristics of mortgages or local economic outcomes, such as house price index returns and past defaults, which are meant to capture a contagion effect. In addition, we also use geographical variables to capture local economic conditions, such as zip code level income and unemployment. Bourassa et al. (2007) discuss the importance, in terms of prediction, of including geographic information in a spatial model of house prices. We capture the equivalent of a nonparametric hazard function of the Cox proportional hazard model by including a set of mortgage time dummy variables, built from *Loan Age*. This reflects the systematic variation of default and prepayment risk over the life time of a mortgage.[13]

Table 3 shows the default and prepayment rate of the mortgage cohorts by origination year. It is remarkable how different the default rates are from one cohort to another. What is especially striking is the increased default rates of mortgages that were originated after the first quarter of 2005, compared to the earlier mortgages. Because this suggests that the quality of mortgages can be different according to their origination year, we follow Kau, Keenan, Lyubimov & Slawson (2011) and estimate the MNL model on a sample, split in two according to origination date. This produces two separate sets of parameter estimates,

[13] Sueyoshi (1995) shows that the natural analogues of the PH model involve estimating pooled logit or porbit models with period specific constant terms.

and this added flexibility enhances the quality of the marginal model.[14] Table 4 shows the estimation results, which mostly agree with our expectations. The effect of contract rate is captured by both a linear and a quadratic term, but in the range of our data, the effect is increasing, in accordance with intuition.

4.3 Dependence model

4.3.1 Geographic distance

The results of Table 5 suggest that the optimal cutoff is around 40 km for Frank, FGM and Clayton copulas, since the minimum value of $tr(\hat{\Gamma}(d))$ for these copula is found to occur at this cutoff of 40 km. On the other hand, Gumbel and rotated Gumbel copula has optimal cutoffs at 35 km. Moreover, while the $tr(\hat{\Gamma}(d))$ drops significantly from 30 km to 35 km, there is hardly a change from 35 km to 40 km except Gumbel copula. This means that even though there is an efficiency gain (loss) from enlarging the data set from 35 km to 40 km, this gain (loss) is small. Figure 2 shows that spatial dependence, measured with Goodman's gamma, decays monotonically with distance. This effect of distance on dependence is significant regardless of the copula used, but the best results obtain with the rotated Gumbel copula. The results confirm the importance of geographic proximity for the dependence of mortgage defaults. Column (3) of Table 6 shows that the effect of geographic distance is not eliminated, even after taking account of other measures of non-geographic distances. In Table 6, column (2) shows the effect of geographic distance when we use a Matérn function, while in column (3) we add geographic distance to the list of regressors in the logit link of Equation (16).

4.3.2 Non-geographic distance

Tables 7 shows the estimation results with non-geographic distances for mortgage pairs within and across clusters. Even after considering individual mortgage and cluster level

[14] Using data for the 20 largest MSAs, Kau, Keenan, Lyubimov & Slawson (2011) detect a structural break by origination year in the last quarter of 2004, using a sup Wald test statistic.

characteristics in the marginal estimation, we still find a significant effect of these characteristics on default dependence. We now report the results of each estimation.[15]

First, we investigate the effect of non-geographic measures of distance on default dependence, taken one at a time. The effects are quite different for mortgages between within and across. While most distance measures, except for *Low Doc*$_{ij}$, *Refinance*$_{ij}$ and *Investment*$_{ij}$, have a significant effect on default dependence within clusters, the same measures have no effect for pairs across clusters, except for *Current LTV*$_{ij}$, *Current Contract Rate*$_{ij}$ and *Penalty*$_{ij}$. In particular, while the mortgage characteristics at origination, such as *FICO*, *Loan Amount* and *Investment*, affect the default dependence within cluster, those factors lose explanatory power across clusters. The distance measures in terms of income and unemployment do not have a significant effect for mortgage pairs from different zip codes.[16] The pattern of dependence implied by the coefficients on distances and averages are shown across pairs of measures in Figures 3 and 4. The different economic characteristics of clusters determine the dependence of a mortgage pair from the same cluster differently. The estimation result of Table 7, Panel B, indicate that the higher unemployment rate in the cluster is, the higher is the dependence among mortgages in that cluster, and that this effect is statistically significant. The difference of average income, however, does not have a significant effect. The results indicate that mortgage pairs with similar characteristics within a cluster tend to have positive dependence. As can be seen in Figure 5, which reports the results of a Matérn correlation function with non-geographic distances, these effects are less important for mortgages across clusters, whose dependence is affected more by geographic distance.

Second, in an attempt to verify how robust the effect of the mortgage and cluster characteristics on dependence are, we take account of geographic distance and additionally introduce each non-geographic distance in turn. Table 6 shows the results of this estimation.

[15] Since the gain from enlarging the data set from 35 km to 40 km is small compared to the amount of computing resources required to estimtate, we use 35 km subsample instead of 40 km subsample for the estimation with non-geographic distances.

[16] Since these measures are only available at cluster level, they can only be used for mortgage pairs in different clusters.

To the sake of comparison, estimation (1) of Table 6, repeats the results of the last columns of Table 7, about the effect of the characteristics, when geographic distance is not included. τ_2 and α are the parameters of the Matérn function. The results suggest that geographic distance dominates all the distances in terms of all other variables. Once we condition on geographic distance, only $Current\ LTV_{ij}$, $Current\ Contract\ Rate_{ij}$ and $Penalty_{ij}$ continue to significantly affect dependence, as they do when their effect is accounted for, one at a time, without geographic distance.

Third, we treat geographic distance like all the other variables, and we introduce it into the logit link function of Equation (16), along with non-geographic distances and macroeconomic variables. Table 8 shows that the distance, based on each non-geographic characteristic, has a similar effect on default dependence for mortgage pairs within and across clusters. $FICO_{ij}$ is the only variable that is significant only for mortgages from the same cluster. The effects of $Refinance_{ij}$ and $Penalty_{ij}$ change compared to when we consider these variables separately or with geographic distance. Thus there is some merit in considering the effect of all these variables jointly, since some that are not individually significant become so in the larger model. This shows that the various measures interact in complex ways. The results show that the effects of non-geographic distance based on LTV_{ij} and $Investment_{ij}$ are stronger for closely located mortgage pairs. However, mortgage pairs far apart can still face relatively high dependence if they share similar mortgage contracts or characteristics.

Finally, we examine the effect of the time variation in the economic situation on the dependenc of mortgage defaults. Table 8 shows that the house price index (HPI) return has a significant negative impact on the dependence of mortgage pairs. This means that mortgage pairs become more dependent when the house price index is low. This captures the sort of contagion in mortgage defaults that occurred after a collapse in house prices during the last financial crisis. Moreover, this shows that on top of the direct effect of house price returns on default probabilities, there is also an additional effect via increased default dependence. This dependence channel is new in the literature, to the best of our knowledge. Since *Current* LTV_{ij} and *Current Contract* $Rate_{ij}$ both reflect current economic situation, they might compete for statistical significance with other time-varying variables. This could explain why *Past Default*, which is a variable that captures contagion, is not significant.

5 Conclusion

This paper models the dependence of mortgage defaults using alternative measures of distances, at different levels of aggregation. In particular, we use geographic distance, as well as non-geographic measures of the distance between pairs of mortgages in terms of mortgage and cluster level characteristics. We use a multinomial logistic regression (MNL) model in order to estimate the default probability of each individual mortgage and a copula for the dependence. The estimation relies on a pairwise composite likelihood. In order to measure spatial dependence, we introduce a squared exponential correlation function, which is a special case of the Matérn correlation function.

Our results show that the effect of geographic distance on default correlation is strongly significant, even if we control for other measures of distance. This might be due to the fact that distances captures the effect of unobserved factors, other than the ones we explicitly include in our estimations. Our results concerning other non-geographic distance measures suggest that in order to measure the risk stemming from the dependence of defaults, one needs to consider both geographic and non-geographic distance among mortgages. Finally we show that, in addition to the well-known effect on default probabilities, there is also a dependence channel, whereby low house price index returns have an impact on default dependence. These two effect concur to making MBSs riskier during economically difficult periods.

Further research is needed to determine whether these results are in line with prime mortgages or mortgage pairs from different MSAs. In addition, more disaggregated house prices, geographic distance data or other cluster level demographic data would be necessary to tease out the unobserved variables, whose effect is captured by geographic distance.

6 Tables

Table 1: Variable definition

Variable name	Definition
	Panel A. Mortgage level variables
Current LTV	Loan to value ratio (LTV) at the time of observation (in percent).
Original LTV	Loan to value (LTV) ratio at origination.
LTV80	Indicator variable that equals 1 if the loan to value ratio at origination is exactly 80%.[a]
Amount	Loan amount at origination in constant (2000) dollars ($ million).
FICO	Fair, Isaac, and Company (FICO) credit score of the borrower at origination (divided by 100).
Low Doc	Indicator variable that equals 1 for a loan with partial or no verification of borrower income or assets.
Penalty	Indicator variable that equals 1 for a loan with a prepayment penalty.
Investment	Indicator variable that equals 1 for a loan secured by a property other than the borrower's primary residence
Refinance	Indicator variable that equals 1 if the loan is taken out for the purpose of refinancing.[17]
Current Contract Rate	Contract rate at the time of observation (in percent).
Current Contract Rate2	Square of contract rate.
Loan Age	Age of loan from origination date (in months).
	Panel B. Cluster level variables
Income	Zip code level median household income ($ thousands).
Unemployment	Zip code level unemployment rate.
	Panel C. MSA level variables
HPI return	Quarterly rate of change in the house price index (HPI) of Los Angeles MSA.
Past Default	Past quarter's default rate of Los Angeles MSA, calculated as the ratio of the number of defaults that occurred in the last quarter to the number of mortgages that survived until the previous quarter.
Season	Sinusoidal seasonal trend with minimum in June and maximum in December.

This table defines the variables employed in the marginal as well as in the dependence model.

[a] There is a concern that mortgages with origination loan to value of exactly 80% (the maximum rate for eligibility for federal mortgage insurance), might have a hidden second-tier loan, which would make those loans more risky.

Table 2: Descriptive Statistics of model variables

Variable	Mean	SD	Min	Max
Panel A. Static variables ($n = 31962$)				
Original LTV (%)	83.68	5.36	80	107
LTV80	0.61	0.49	0	1
Amount ($ million)	0.36	0.14	0.04	3.06
FICO	6.43	0.53	4.01	7.2
Low Doc	0.48	0.50	0	1
Penalty	0.72	0.45	0	1
Investment	0.10	0.29	0	1
Refinance	0.15	0.35	0	1
Income ($ thousands)	84.52	22.13	44.41	210.34
Unemployment (%)	10.43	2.69	0	20.3
Panel B. Dynamic variables ($n = 239486$)				
Current LTV (%)	100.14	23.51	0.01	207.58
Current Contract Rate (%)	6.85	1.50	0.5	37.5
Season	0.0019	0.71	-1	1
Loan Age (month)	23.03	15.87	2	151
HPI return (%)	-0.74	0.64	-10.11	8.38
Past Default (%)	4.78	4.79	0.46	17.25

This table provides key statistics of the explanatory variables used in the marginal multinomial logistic model.

Table 3: Number of defaults and prepayments in LA by origination year and quarter

Year	Quarter	# Loans	Prepaid	Default	% Prepaid	% Default
1997	1	2	.	.	.	.
	2	2	2	.	100.0%	.
	3	3	2	.	66.67%	.
	4	3	1	.	33.33%	.
1998	1	1	1	.	100.0%	.
	2	3	1	.	33.33%	.
	3	6	3	.	50.00%	.
	4	9	6	2	66.67%	22.22%
1999	1	4	4	.	100.0%	.
	2	14	10	2	71.43%	14.29%
	3	29	17	3	58.62%	10.34%
	4	66	13	10	19.70%	15.15%
2000	1	53	9	11	16.98%	20.75%
	2	87	10	14	11.49%	16.09%
	3	82	22	10	26.83%	12.20%
	4	61	18	10	29.51%	16.39%
2001	1	43	13	7	30.23%	16.28%
	2	82	12	8	14.63%	9.76%
	3	90	23	18	25.56%	20.00%
	4	126	53	15	42.06%	11.90%
2002	1	130	28	10	21.54%	7.69%
	2	214	94	16	43.93%	7.48%
	3	197	83	18	42.13%	9.14%
	4	309	140	16	45.31%	5.18%
2003	1	289	175	10	60.55%	3.46%
	2	541	323	31	59.70%	5.73%
	3	969	479	112	49.43%	11.56%
	4	860	578	51	67.21%	5.93%
2004	1	1079	787	61	72.94%	5.65%
	2	1508	1187	158	78.71%	10.48%
	3	1531	1191	193	77.79%	12.61%
	4	1729	1228	339	71.02%	19.61%
2005	1	2277	1594	507	70.00%	22.27%
	2	2903	1704	921	58.70%	31.73%
	3	2906	1306	1307	44.94%	44.98%
	4	2480	857	1376	34.56%	55.48%
2006	1	1968	494	1310	25.10%	66.57%
	2	2399	506	1683	21.09%	70.15%
	3	2040	254	1582	12.45%	77.55%
	4	1996	148	1655	7.41%	82.92%
2007	1	1649	107	1373	6.49%	83.26%
	2	941	61	735	6.48%	78.11%
	3	250	17	193	6.80%	77.20%
	4	16	2	8	12.50%	50.00%
2008	1	15	.	8	.	53.33%

Table 4: Estimates of the multinomial logistic model

	1997-2004 Origination				2005-2008 Origination			
Parameter	Estimate	Std Err	Pr > χ^2	Odds ratio	Estimate	Std Err	Pr> χ^2	Odds ratio
LTV80	0.3753	0.0815	<.0001	1.455	0.0368	0.0259	0.1551	1.037
Amount	0.0325	0.3127	0.9208	1.033	0.0280	0.0954	0.7691	1.028
Low Doc	0.1927	0.0665	0.0037	1.213	0.0586	0.0214	0.0062	1.060
Current Contract Rate	0.5354	0.1082	<.0001	1.708	−0.0657	0.0364	0.0871	0.936
Current Contract Rate2	−0.0155	0.0069	0.0247	0.985	0.0209	0.0025	<.0001	1.021
Investment	0.0172	0.1128	0.8789	1.017	0.0154	0.0327	0.6375	1.016
FICO	−0.7209	0.0629	<.0001	0.486	−0.5196	0.0186	<.0001	0.595
Refinance	−0.5075	0.0946	<.0001	0.602	−0.1281	0.0296	<.0001	0.878
Penalty	−0.2135	0.0799	0.0075	0.808	0.0534	0.0244	0.0289	1.055
Current LTV	0.0373	0.0040	<.0001	1.038	0.0099	0.0017	<.0001	1.010
Season	−0.3254	0.0477	<.0001	0.722	−0.1751	0.0141	<.0001	0.839
HPI return	−0.0349	0.0134	0.0093	0.966	−0.0256	0.0032	<.0001	0.975
Past Default	0.0900	0.0208	<.0001	1.094	0.1327	0.0049	<.0001	1.142
Income	−0.0466	0.0228	0.0414	0.954	−0.0755	0.0075	<.0001	0.927
Unemployment	−0.0075	0.0163	0.6479	0.993	0.0016	0.0050	0.7506	1.002
Intercept	Yes				Yes			
Loan Age	Yes				Yes			

This table provides results of the marginal multinomial logistic model estimated on a subsample of 10,122 and 21,840 loans for the first period and second period, respectively.

Table 5: The effect of geographic distance on dependence, using a Matérn function (40 km cutoff distance)

	Gumbel		Frank		FGM		Rotated Gumbel			Clayton	
	Within	Across	Within	Across	Within	Across	Within	Across		Within	Across
τ_1	0.0040		0.0171		0.0173		0.0200		0.0308		
t-stat	(17.46)		(20.54)		(20.29)		(21.11)		(20.20)		
α		39.2841		30.7106		30.6153		32.3190		30.9216	
t-stat		(6.46)		(9.39)		(8.57)		(9.18)		(9.39)	
τ_2		0.0024		0.0112		0.0113		0.0130		0.0210	
t-stat		(8.34)		(9.36)		(8.90)		(9.63)		(9.45)	
$tr(\hat{\Gamma}(d))$		36.9894		10.6925		12.7581		12.4006		10.8509	
$tr(\hat{\Gamma}(35))$		23.5189		10.9187		14.5876		12.0478		11.2424	
$tr(\hat{\Gamma}(30))$		43.5037		224.5859		219.23		49.9097		281.72	
LogL= (-5730000+)	-395.76		-349.05		-349.88		-341.27			-363.78	
LogL= (-15670000+)		-6655.78		-6655.43		-6655.44		-6649.81			-6655.31
Sum= (-21400000+)	-7051.54		-7004.48		-7005.32		-6991.08			-7019.09	

This table provides estimates of the dependence model with the Matérn function and various copulas using a subsample of mortgage pairs whose maximum distance is 40 km. It also provides $tr(\hat{\Gamma}(d))$ of data samples with cutoffs of 30 km and 35 km. The number of mortgage pairs is 41,904,285 with a cutoff of 40 km, which represents 3.87% of the exhaustive number of mortgage pairs in Los Angeles (1,149,871,859). The number of mortgage pairs for the subsamples with 30 km and 35 km cutoffs is 28,634,993 and 35,349,314, respectively. The exhaustive number of possible mortgage pairs within the same cluster is 13,744,974.

Table 6: The effect of non-geographic distances on dependence, with and without accounting for geographic distance, FGM copula.

Geographic Distance	Not Included		In Matérn		In logit	
	(1)		(2)		(3)	
Variables	Estimates	t-stat	Estimates	t-stat	Estimates	t-stat
Panel A. Mortgage characteristics						
ψ_0	−4.7765	−7.40			−4.3631	−9.16
Amount (average)	3.9765	2.55	3.0429	6.11	−1.4780	−1.30
Amount$_{ij}$	−5.4258	−1.39	−2.8150	−0.94	9.6740	3.79
d_{ij}					−0.0480	−4.87
τ_2			0.0158	9.18		
α			29.3416	9.00		
ψ_0	−2.7129	−1.39			−6.5503	−2.46
FICO (average)	0.4400	1.04	0.0801	0.60	0.6290	1.54
FICO$_{ij}$	−0.9318	−0.76	−0.8344	−0.86	−0.5434	−0.59
d_{ij}					−0.0420	−4.62
τ_2			0.0204	2.11		
α			29.6774	8.87		
ψ_0	−2.3889	−7.42			−1.9582	−5.97
Current LTV (average)	0.0092	3.66	0.8919	2.91	0.9305	3.76
Current LTV$_{ij}$	−0.8634	−9.33	−126.5806	−7.74	−84.1000	−9.50
d_{ij}					−0.0228	−4.20
τ_2			0.0714	6.57		
α			43.1272	8.12		
ψ_0	0.1878	0.52			0.6847	1.78
Current Contract Rate (average)	−0.4224	−7.90	−0.3444	−9.21	−0.4093	−7.56
Current Contract Rate$_{ij}$	−1.0176	−3.65	−0.8612	−3.24	−0.9470	−3.43
d_{ij}					−0.0302	−3.90
τ_2			0.2222	4.37		
α			36.0778	7.31		
ψ_0	−3.0455	−16.66			2.3305	−10.86
Low Doc (average)	−0.3924	−1.71	−0.5297	−1.57	−0.3255	−1.45
Low Doc$_{ij}$	−0.6996	−1.81	−1.0178	−2.01	−0.6754	−1.83
d_{ij}					−0.0386	−4.52
τ_2			0.0316	5.57		
α			29.9658	8.78		
ψ_0	−0.7519	−13.20			−1.9127	−8.04
Penalty (average)	−0.6650	−2.79	−1.0593	−3.90	−0.6731	−3.00
Penalty$_{ij}$	−1.1606	−2.61	−1.9430	−3.75	−1.3842	−3.05
d_{ij}					−0.0413	−4.95
τ_2			0.0481	5.24		
α			28.6644	9.65		
ψ_0	−3.4461	−33.30			−2.6830	−16.74
Refinance (average)	0.2503	0.41	0.0218	0.02	0.0756	0.11
Refinance$_{ij}$	−0.2878	−0.40	0.0830	0.05	0.0042	0.01
d_{ij}					−0.0393	−4.55
τ_2			0.0229	8.65		
α			29.6105	8.86		

ψ_0	-3.3841	-34.70			-2.6330	-16.79
Investment (average)	0.3436	0.63	1.3796	0.65	0.4688	1.02
Investment$_{ij}$	-1.0976	-1.44	-2.2596	-1.03	-1.0430	-1.59
d_{ij}					-0.0390	-4.54
τ_2			0.0238	9.04		
α			29.7851	8.81		
ψ_0	3.2637	-16.63			-2.6231	-11.72
LTV80 (average)	-0.2459	-0.98	-0.0158	-0.03	-0.0731	-0.28
LTV80$_{ij}$	-0.1496	-0.41	-0.0340	-0.05	-0.0550	-0.15
d_{ij}					-0.0389	-4.44
τ_2			0.0232	4.94		
α			29.6615	8.57		
Panel B. Economic distance						
ψ_0	-5.1509	-7.42			-4.4664	-8.55
Income (average)	0.2201	3.15	0.1445	3.79	0.2213	3.82
Income$_{ij}$	-0.1350	-0.71	-0.0096	-0.04	0.0502	0.37
d_{ij}					-0.0460	-4.64
τ_2			0.0266	8.30		
α			29.3051	9.00		
ψ_0	-3.0226	-4.86			-2.0260	-3.55
unemployment (average)	-0.0396	-0.64	-0.0823	-1.36	-0.0649	-1.23
unemployment$_{ij}$	-0.0048	-0.06	0.0870	0.73	0.0550	0.71
d_{ij}					-0.0419	-4.71
τ_2			0.0365	2.21		
α			28.8693	9.15		
Panel C. Economic situation						
ψ_0	-3.3119	-20.17			-2.6595	-13.61
HPI return	2.5133	0.89	0.1264	0.03	0.3334	0.11
d_{ij}					-0.0390	-4.31
τ_2			0.0231	6.60		
α			29.6562	8.59		
ψ_0	-2.8196	-13.48			-2.3210	-11.36
Past Default (t-1)	-6.5392	-2.76	-6.0592	-2.25	-4.3482	-1.87
d_{ij}					-0.0345	-3.77
τ_2			0.0314	6.20		
α			31.4667	7.76		
ψ_0	-3.3883	-30.24			-2.6182	-15.52
ΔPast Default	-0.3012	-0.80	-0.6195	-0.83	-0.2941	-0.81
d_{ij}					-0.0393	-4.61
τ_2			0.0245	7.73		
α			29.6371	9.00		

This table provides FGM copula estimation results for the dependence parameters with Equation (16) in columns (1) and (3), and with the Matérn function in column (2). Various non-geographic variables are placed into Equation (16) and Equation (15), one at a time along with physical distance. Column (1) shows mortgage pairs within cluster, while columns (2) and (3) show results with mortgage pairs across clusters, with a cutoff of 35 km. Past Default (t-1) is the previous quarter's default rate of LA. d_{ij} is physical distance between different clusters. τ_2 and α are the parameters of the Matérn function. ψ_0 is the constant.

Table 7: The effect of non-geographic distances on dependence, within and across cluster

	Within cluster (1)				Across clusters (2)	
Variables	Estimates	t-stat	Estimates	t-stat	Estimates	t-stat
Panel A. Mortgage characteristics						
ψ_0	-2.6371	-9.54			-4.7765	-7.40
Amount (average)	0.9706	1.19			3.9765	2.55
Amount$_{ij}$	-4.7832	-2.10			-5.4258	-1.39
ψ_0	-3.3889	-3.35			-6.0497	-2.23
FICO (average)	0.2013	1.30			0.4400	1.04
FICO$_{ij}$	-1.2942	-3.90			-0.4659	-0.76
ψ_0	0.2264	0.70			-2.3889	-7.42
Current LTV (average)	-1.3773	-4.59			0.9167	3.66
Current LTV$_{ij}$	-34.6568	-8.88			-86.3388	-9.33
ψ_0	-1.8285	-4.63			0.1878	0.52
Current Contract Rate (average)	-0.0194	-0.38			-0.4224	-7.90
Current Contract Rate$_{ij}$	-0.7384	-6.45			-1.0176	-3.65
ψ_0	-3.0240	-13.16			-3.0455	-16.66
Low Doc$_i$=Low Doc$_j$=1	0.7985	3.33			-0.3924	-1.71
Low Doc$_i$ $\neq$ Low Doc$_j$	0.4420	2.01			-0.5460	-2.35
ψ_0	-1.8491	-12.00			-0.7519	-13.20
Penalty$_i$=Penalty$_j$=1	-0.7937	-4.59			-0.6650	-2.79
Penalty$_i$ $\neq$ Penalty$_j$	-0.4553	-2.83			-0.9128	-3.42
ψ_0	-2.4278	-41.85			-3.4461	-33.30
Refinance$_i$=Refinance$_j$=1	-0.8231	-0.90			0.2503	0.41
Refinance$_i$ $\neq$ Refinance$_j$	-0.2025	-0.42			-0.0187	-0.05
ψ_0	-2.4085	-41.73			-3.3841	-34.70
Investment$_i$= Investment$_j$=1	0.2973	0.91			0.3436	0.63
Investment$_i$ $\neq$ Investment$_j$	-0.3748	-1.66			-0.3770	-0.97
ψ_0	-2.9794	-15.18			3.2637	-16.63
LTV80$_i$=LTV80$_j$=1	1.0192	4.96			-0.2459	-0.98
LTV80$_i$ $\neq$ LTV80$_j$	0.1586	0.79			-0.1978	-0.85
Panel B. Economic difference						
ψ_0	-2.8289	-8.64				
Income (level)	0.0484	1.12				
ψ_0	-3.1672	-9.53				
Unemployment (level)	0.0570	2.21				
Panel C. Economic distance between clusters						
ψ_0					-5.1509	-7.42
Income (average)					0.2201	3.15
Income$_{ij}$					-0.0675	-0.71
ψ_0					-3.0226	-4.86
Unemployment (average)					-0.0396	-0.64
Unemployment$_{ij}$					-0.0024	-0.06
Panel D. Economic situation						
ψ_0	-2.8182	-5.43			-3.3119	-20.17
HPI return	-5.5791	-16.19			2.5133	0.89
ψ_0	-1.8081	-18.57	-1.9193	-17.31	-2.8196	-13.47
Past Default (t-1)	-7.3600	-6.75	3.2859	0.74	-6.5392	-2.76
Past Default (t-2)			-11.4984	-2.40		
ψ_0	-2.6131	-35.88			-3.3883	-30.24
ΔPast Default	0.5992	3.11			-0.3012	-0.80

This table provides estimation results of dependence parameters with the logit link function in Equation (16) and the FGM copula. Various non-geographic variables are placed into Equation (16), one at a time. Column (1) shows the estimation results with mortgage pairs within cluster. Estimation with mortgage pairs across clusters in column (2) is based on the subsample with a cutoff of 35 km. Past Default (t-1) is the previous quarter's default rate of LA, and Past default (t-2) is the default rate two quarters ago.

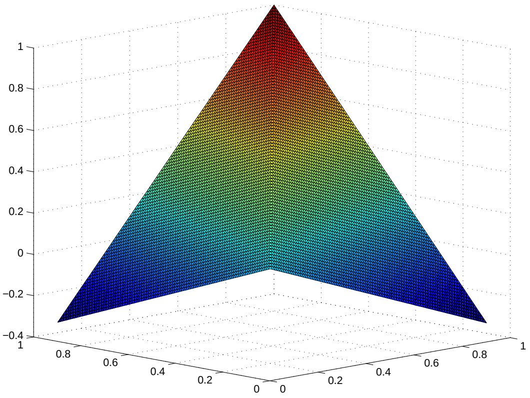
Table 8: The effect of all distances

	Across clusters				Within cluster	
	In logit		In Matérn			
Variables	Estimates	t-stat	Estimates	t-stat	Estimates	t-stat
ψ_0	0.0733	0.21			−0.2106	−0.99
Panel A. Physical distance						
d_{ij}	−0.0173	−2.93				
τ_2			0.2120	7.10		
α			60.4096	5.13		
Panel B. Mortgage characteristics						
Amount$_{ij}$	0.0429	0.08	−0.0368	−0.07	−1.7561	−1.56
FICO$_{ij}$	−0.2068	−1.29	−0.2518	−1.58	−0.3609	−3.42
Current LTV$_{ij}$	−53.4107	−13.92	−55.4654	−12.71	−27.6117	−5.37
Current Contract Rate$_{ij}$	−0.3507	−4.06	−0.4241	−5.68	−0.2958	−6.49
Low Doc$_{ij}$	−0.1502	−1.38	−0.1734	−1.60	0.0066	0.07
Penalty$_{ij}$	−0.0394	−0.34	−0.1610	−1.48	0.0715	0.72
Refinance$_{ij}$	−0.5569	−2.99	−0.5789	−2.89	−0.6328	−3.11
Investment$_{ij}$	−0.5880	−3.21	−0.5868	−3.19	−0.5288	−3.48
LTV80$_{ij}$	−1.9666	−4.87	−2.0585	−4.94	−0.8455	−2.98
Panel C. Economic distance						
Income$_{ij}$	−0.0135	−0.37	−0.0239	−0.65		
Unemployment$_{ij}$	−0.0039	−0.17	−0.0102	−0.44		
Panel D. Economic situation						
HPI return	−6.2521	−2.66	−6.2996	−2.82	−12.0806	−5.98
Past Default (t-1)	3.1034	1.94	6.0319	3.96	−3.1014	−1.80
Δ Past Default	−2.0239	−4.65	−1.7482	−4.08	−1.0351	−3.22

This table provides estimation results of the dependence parameters in Equation (16) with the FGM copula and estimation results of the dependence parameters with the Matérn function using rotated Gumbel copula. All the non-geographic variables including physical distance are placed together into the logit link function of Equation (16) and the Matérn function respectively for the mortgage pairs across clusters. We use the logit link function with the non-geographic variables including physial distance for the mortgage pairs within cluster. For the estimation with mortgage pairs across clusters, we use the subsample of mortgage pairs with a cutoff of 35 km. We use all possible mortgage pairs within cluster (exhaustive mortgage pairs). Past Default (t-1) is the previous quarter's default rate of LA. d_{ij} is physical distance between different clusters. ψ_0 is a constant term.

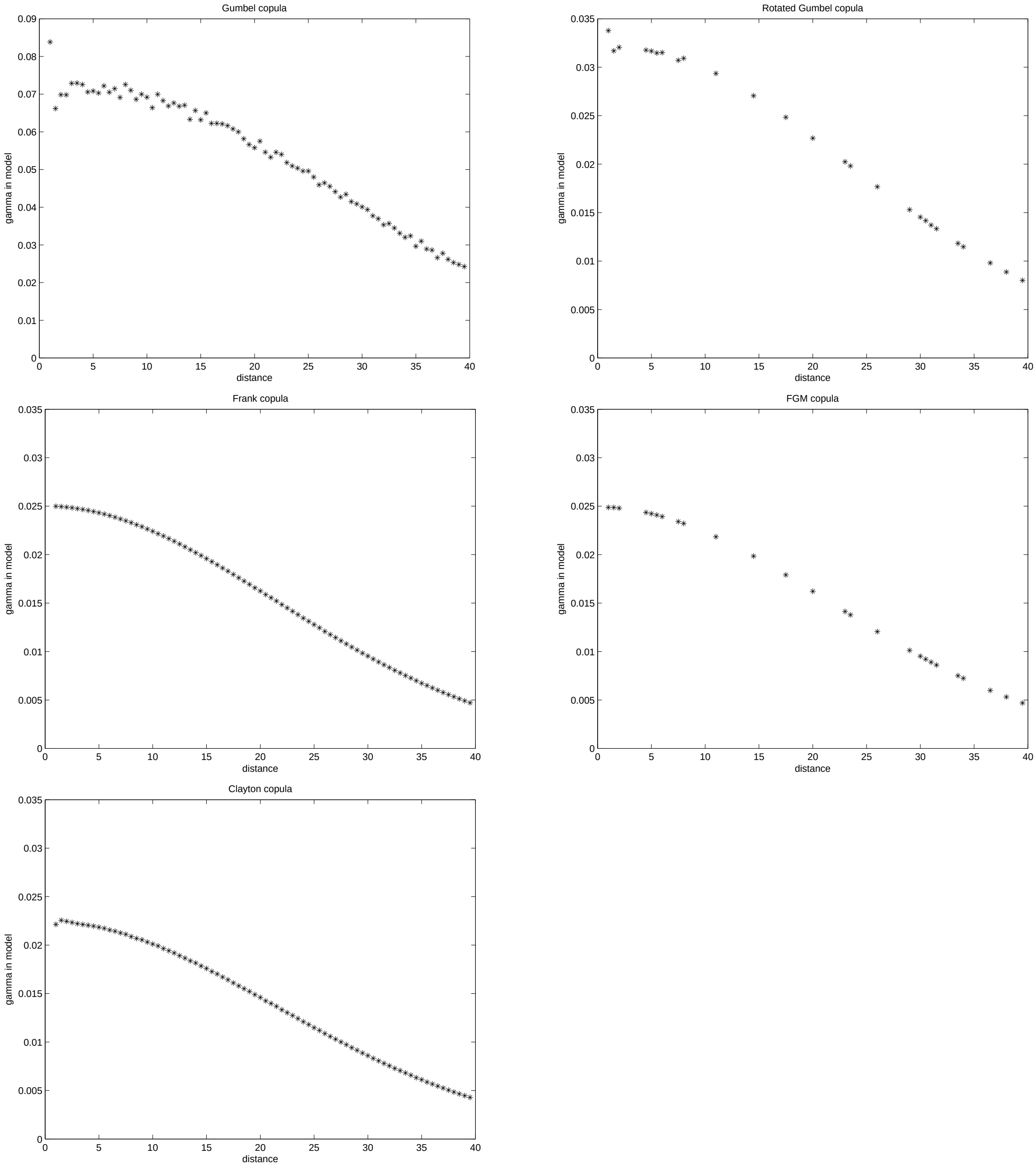
7 Figures

Figure 1: Functional form of dependence with distance and average for a non-geographic variable X_i.



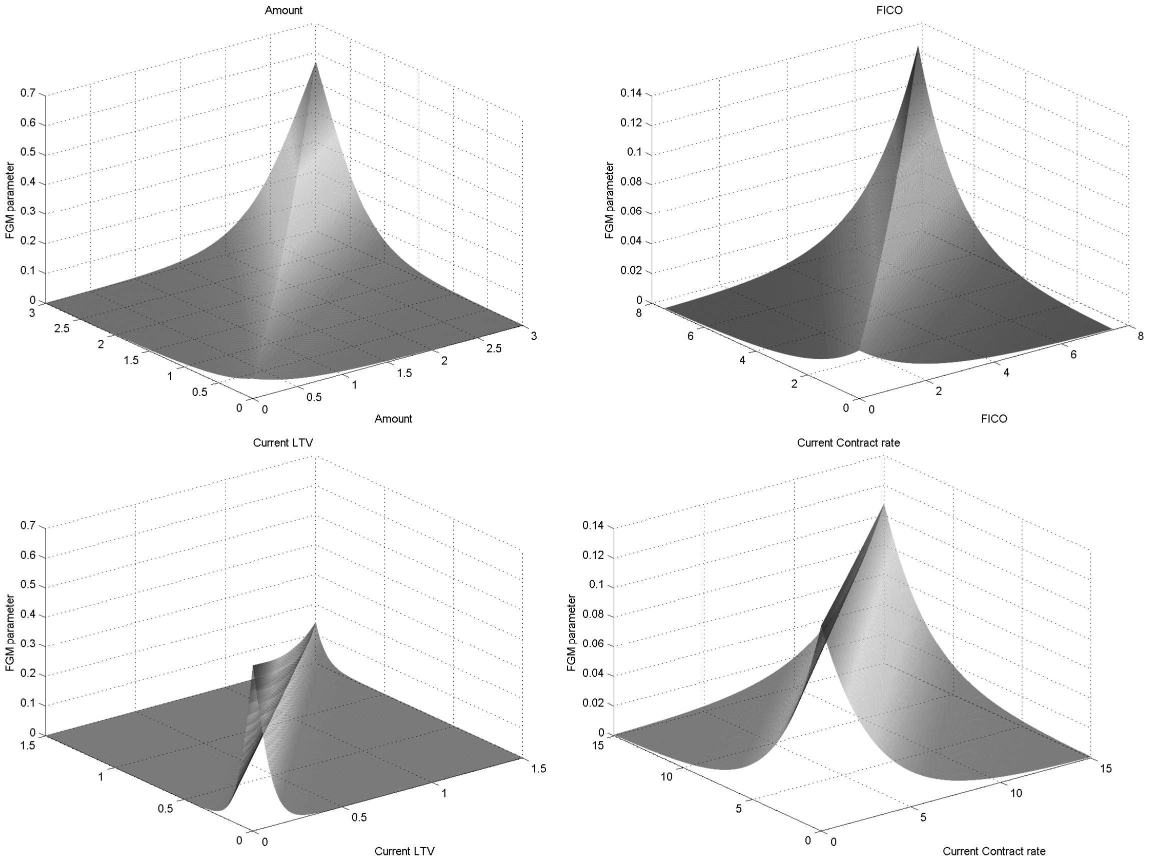
This figure shows the pattern of dependence that emerges when distance has a negative, and average a positive effect on dependence.

Figure 2: Goodman's gamma across distance for different copulas



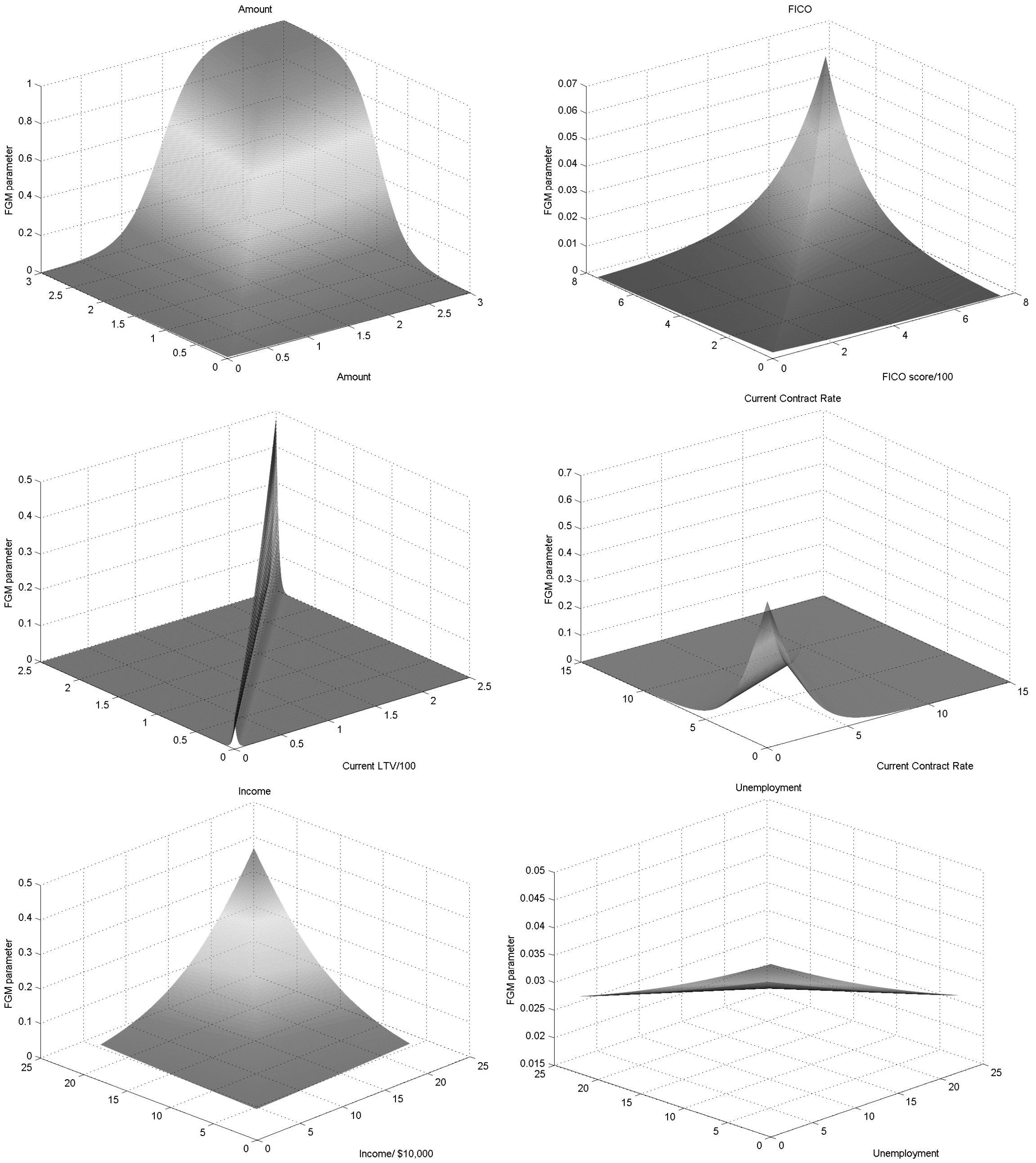
This figure shows that spatial dependence, measured by Goodman's gamma, decays monotonically with physical distance. The figure is based on the estimates in Table 5. The effect of geographic distance is taken into account via a Matérn function (squared exponential correlation function) with various copulas. We combine the probability of default from the marginal estimation with the copula estimates in order to compute Goodmans's gamma for every mortgage pair, using Equation (8). We then average Goodman's gamma for mortgage pairs in 0.5 km bins. We use the subsample of mortgage pairs with a 40 km cutoff.

Figure 3: Dependence across parameter space (mortgage pairs within cluster)



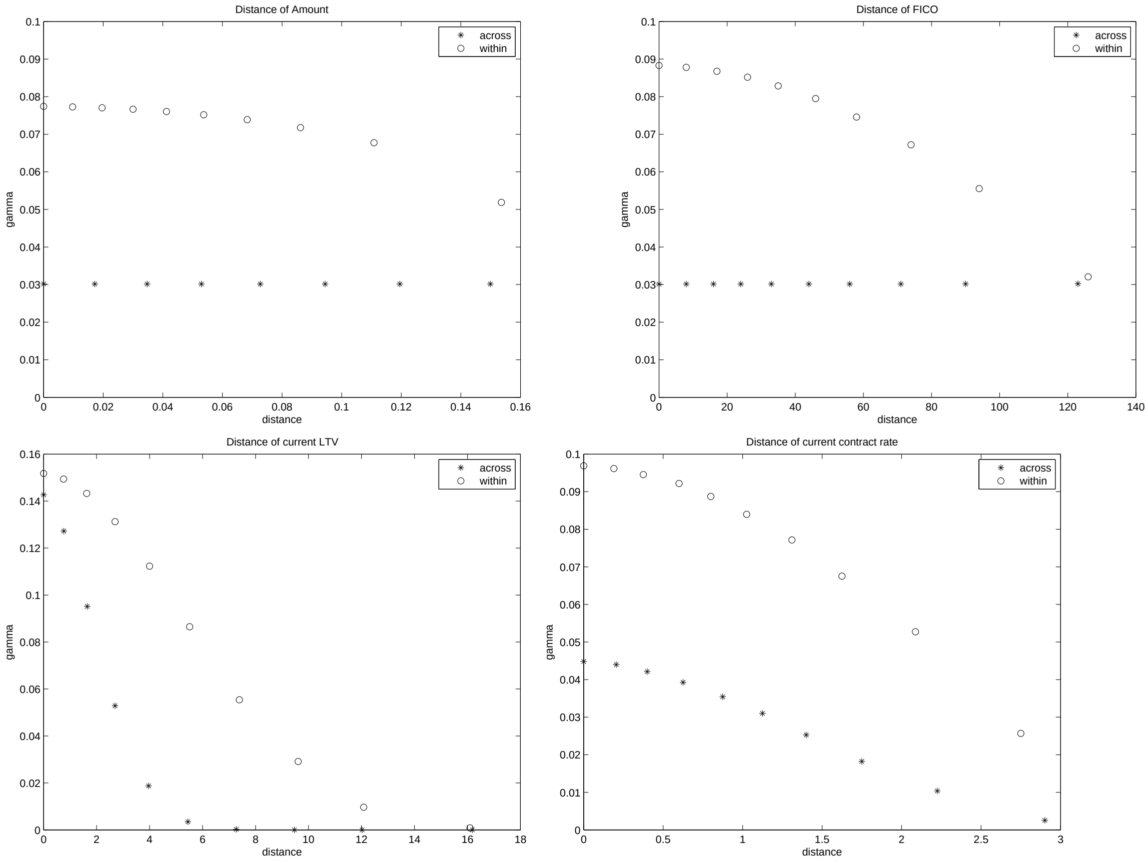
This figure shows the pattern of dependence, measured by the FGM parameter, implied by the coefficients on non-geographic distance and average. We use the non-geographic variables in Panel A, column (1) of Table 7, except for the dummy variables. We use only mortgage pairs within the same cluster.

Figure 4: Dependence across parameter space (mortgage pairs across clusters)



This figure shows the pattern of dependence, measured by the FGM parameter, implied by the coefficients on non-geographic distance and average. We use the non-geographic variables in Panel A, column (2) of Table 7, except for the dummy variables. We use only mortgage pairs within the same cluster.

Figure 5: The gammas across non-geographic distance



This figure shows how dependence, measured by Goodman's gamma, varies with non-geographic distances for mortgages pairs within and across clusters. We take into account the effect of non-geographic distance via a Matérn function (squared exponential correlation function) with a FGM copula. We use the subsample of mortgage pairs across clusters with a cutoff of 35 km and pairs within a cluster.

References

Ashcraft, A., Goldsmith-Pinkham, P. & Vickery, J. (2010), MBS ratings and the mortgage credit boom, Technical report, Federal Reserve Bank of New York.

Bai, Y. (2011), Joint Composite Estimating Functions in Spatial and Spatio-Temporal Models, PhD thesis, University of Michigan.

Bai, Y., Song, P. X.-K. & Raghunathan, T. E. (2012), 'Joint composite estimating functions in spatiotemporal models', *Journal of the Royal Statistical Society B* **75**(5), 799–824.

Bourassa, S. C., Cantoni, E. & Hoesli, M. (2007), 'Spatial dependence, housing submarkets and house price prediction', *Journal of Real Estate Finance and Economics* **35**(2), 143–160.

Case, B., Clapp, J., Dubin, R. & Rodriguez, M. (2004), 'Modelling spatial and temporal house price patterns : A comparison of four models', *Journal of Real Estate Finance and Economics* **29**(2), 167–191.

Clapp, J. M., Deng, Y. & An, X. (2006), 'Unobserved heterogeneity in models of competing mortgage termination risks', *Real Estate Economics* **34**(2), 243–273.

Clapp, J. M., Goldberg, G. M., Harding, J. P. & LaCour-Little, M. (2001), 'Movers and shuckers: Interdependent prepayment decisions', *Real Estate Economics* **29**(3), 411–450.

Conley, T. G. & Topa, G. (2002), 'Socio-economic distance and spatial patterns in unemployment', *Journal of Applied Eonometrics* **17**, 303–327.

Corrente, J. E., Chalita, L. V. A. S. & Moreira, J. A. (2003), 'Choosing between Cox proportional hazards and logistic models for interval-censored data via bootstrap', *Journal of Applied Statistics* **30**(1), 37–47.

Cowan, A. M. & Cowan, C. D. (2004), 'Default correlation: An empirical investigation of a subprime lender', *Journal of Banking and Finance* **28**, 753–771.

de Leon, A. R. (2005), 'Pairwise likelihood approach to grouped continuous model and its extension', *Statistics and Probability Letters* **75**, 49–57.

Deng, Y., Pavlov, A. D. & Yang, L. (2005), 'Spatial heterogeneity in mortgage terminations by refinance, sale and default', *Real Estate Economics* **33**(4), 739–764.

Denuit, M. & Lambert, P. (2005), 'Constraints on concordance measures in bivariate discrete data', *Journal of Multivariate Analysis* **93**(1), 40–57.

Dubin, R. (1998), ‘Predicting house prices using multiple listings data’, *Journal of Real Estate Finance and Economics* **17**(1), 35–59.

Duffie, D. & Singleton, K. J. (2003), *Credit Risk*, Princeton Series in Finance.

Gauvreau, K. & Pagano, M. (1997), ‘The analysis of correlated binary outcomes using multivariate logistic regression’, *Biometrical Journal* **39**(3), 309–325.

Genest, C. & Nešlehová, J. (2007), ‘A primary on copulas for count data’, *Astin Bulletin* **37**(2), 475–515.

Gneiting, T., Kleiber, W. & Schlather, M. (2010), ‘Matérn cross-covariance functions for multivariate random fields’, *Journal of the American Statistical Association* **105**(491), 1167–1177.

Goodman, L. A. & Kruskal, W. H. (1954), ‘Measures of association for cross classifications’, *Journal of the American Statistical Association* **49**(268), 732764.

Goorah, A. (2007), ‘Real estate risk management with copulas’, *Journal of Property Research* **24**(4), 289–311.

Gumbel, E. J. (1960), ‘Bivariate exponential distributions’, *Journal of the American statistical Association* **55**, 698–707.

Gumbel, E. J. (1961), ‘Bivariate logistic distributions’, *Journal of the American statistical Association* **56**, 335–349.

Harding, J. P., Rosenblatt, E. & Yao, V. W. (2009), ‘The contagion effect of foreclosed properties’, *Journal of Urban Economics* **66**, 164–178.

Joe, H. (2005), ‘Asymptotic efficiency of the two-stage estimation method for copula-based models’, *Journal of Multivariate Analysis* **94**(2), 401–419.

Kau, J. B., Keenan, D. C. & Li, X. (2011), ‘An analysis of mortgage termination risks: A shared frailty approach with msa-level random effects’, *Journal of Real Estate Finance and Economics* **42**, 51–67.

Kau, J. B., Keenan, D. C., Lyubimov, C. & Slawson, C. V. (2011), ‘Subprime mortgage default’, *Journal of Urban Economics* **70**, 75–87.

Kazianka, H. (2013), ‘spatialCopula: A Matlab toolbox for copula-based spatial analysis’, *Stochastic Environmental Research and Risk Assessment* **27**, 121–135.

Kuk, A. Y. & Nott, D. J. (2000), ‘A pairwise likelihood approach to analyzing correlated binary data’, *Statistics and Probability Letters* **47**, 329–335.

Lando, D. (2004), *Credit Risk Modeling, Theory and Applications*, Princeton Series in Finance.

Le Cessie, S. & van Houwelingen, J. (1994), 'Logistic regression for correlated binary data', *Applied Statistics* **43**(1), 95–108.

Liu, X. (2012), *Survival Analysis: Models and Applications*, Wiley, New York.

Meester, S. K. & MacKay, R. J. (1994), 'A parametric model for cluster correlated categorical data', *Biometrics* **50**, 954–963.

Molenberghs, G. & Lesaffre, E. (1994), 'Marginal modeling of correlated ordinal data using a multivariate Plackett distribution', *Journal of the American Statistical Association* **89**(426), 633–644.

Nešlehová, J. (2007), 'On rank correlation measures for non-continuous random variables', *Journal of Multivariate Analysis* **98**, 544–567.

Nikoloulopoulos, A. K. & Karlis, D. (2008), 'Multivariate logit copula model with an application to dental data', *Statistics in Medicine* **27**, 6393–6406.

Paik, J. & Ying, Z. (2012), 'A composite likelihood approach for spatially correlated survival data', *Computational Statistics and Data Analysis* **56**, 209–216.

Rasmussen, C. E. & Williams, C. K. (2006), *Gaussian Processes for Machine Learning*, The MIT Press.

Schoenbucher, P. J. (2003), *Credit Derivatives Pricing Models: Model, Pricing and Implementation*, John Wiley & Sons.

Sklar, A. (1959), 'Fonctions de répartition à n dimensions et leurs marges', *Pub. Inst. Statist. Univ. Paris* **8**, 229–231.

Song, P. X.-K. (2000), 'Multivariate dispersion models generated from Gaussian copulas', *Scandinavian Journal of Statistics* **27**, 305–320.

Sueyoshi, G. T. (1995), 'A class of binary response models for grouped duration data', *Journal of Applied Econometrics* **10**(4), 411–431.

Varin, C. (2008), 'On composite marginal likelihoods', *Advances in Statistical Analysis* **92**, 1–28.

Varin, C., Reid, N. & Firth, D. (2011), 'An overview of composite likelihood methods', *Statistica Sinica* **21**, 5–42.

Varin, C. & Vidoni, P. (2005), 'A note on composite likelihood inference and model selection', *Computational Statistics and Data Analysis* **92**, 519–528.

Voicu, I., Jacob, M., Rengert, K. & Fang, I. (2012), ‘Subprime loan default resolutions: Do they vary across mortgage products and borrower demographic groups?’, *Journal of Real Estate Finance and Economics* **45**, 939–964.

Zhao, Y. & Joe, H. (2005), ‘Composite likelihood estimation in multivariate data analysis’, *Canadian Journal of Statistics* **33**(3), 265–284.

Zimmer, D. M. (2012), ‘The role of copulas in the housing crisis’, *Review of Economics and Statistics* **94**(2), 607–620.

Appendices

A Copulas

A.1 functional forms

In this appendix, we introduce the bivariate copulas we work with. In the sequel, we use the notation $u_i = F_i(y_i)$.

a. Farlie Gumbel Morgenstern (FGM) The FGM copula has the form

$$C^{FGM}(u_i, u_j; \theta) = u_i u_j (1 + \theta(1 - u_i)(1 - u_j)), \tag{20}$$

where the parameter θ is restricted to the interval $[-1, 1]$, and $\theta = 0$ corresponds to independence. The FGM copula is symmetric and its Kendall's tau is $\tau = \frac{2\theta}{9}$, which can only capture moderate dependence, since it is in the [-$\frac{2}{9}$,$\frac{2}{9}$] range.

b. Gumbel The Gumbel copula has the form

$$C^G(u_i, u_j; \theta) = \exp -((-\log u_i)^\theta + (-\log u_j)^\theta)^{\frac{1}{\theta}}). \tag{21}$$

The Gumbel does not allow for negative dependence and it goes from independence to the Fréchet upper bound of perfect positive dependence, as its parameter θ moves in the range $[1, \infty)$. The Gumbel copula is asymmetric since it has upper, but no lower tail dependence. The Gumbel is often used in rotated form, which obtains by interverting upper and lower tail, as follows: $C^{RG}(u_i, u_j; \theta) = u_i + u_j - 1 + C^G(1 - u_i, 1 - u_j; \theta)$. The Kendall's tau of the Gumbel or rotated Gumbel is $\frac{\theta - 1}{\theta}$.

c. Frank The Frank copula has the form

$$C(u_i, u_j; \theta) = -\frac{1}{\theta} \log \left(1 + \frac{(\exp(-\theta u_i) - 1)(\exp(-\theta u_j) - 1)}{\exp(-\theta) - 1}\right), \tag{22}$$

where $\theta \in (-\infty, \infty) \setminus \{0\}$, and the dependence covers the full possible range, including both the Fréchet upper and lower bound. The Frank copula has neither upper nor lower tail dependence. Its Kendall's tau is $1 - \frac{4(D_1(\alpha) - 1)}{\alpha}$, where $D_1(\alpha) = \frac{1}{\alpha} \int_0^\alpha \frac{t}{\exp t - 1} dt$ is the Debye function.

d. Clayton The Clayton copula has the form

$$C(u_i, u_j; \theta) = (u_i{}^{-\theta} + u_j{}^{-\theta} - 1)^{-1/\theta}, \tag{23}$$

where $\theta \in (0, \infty)$, and the dependence covers only postive dependence, including the Fréchet upper bound. The clayton copula is asymmetric since the dependence is concentrated in the lower tail.

B Loglikelihood, Gradient and Hessian

This appendix develops the loglikelihood function for composite likelihood estimation of the probability mass function (PMF) for bivariate default probability with copula. Define Y_i, $i = 1, 2$ to be Bernoulli variables with probability $\bar{\pi}_i$ of default for mortgage i and $C_\theta(\bar{\pi}_1, \bar{\pi}_2, \theta)$, the copula for joint default. There are four possible outcomes for this bivariate variable, whose probabilities are as follows:

$$\begin{array}{rcl} P(Y_1 = 1, Y_2 = 1) = p_{11} & = & 1 - \bar{\pi}_1 - \bar{\pi}_2 + C(\bar{\pi}_1, \bar{\pi}_2, \theta) \\ P(Y_1 = 1, Y_2 = 0) = p_{10} & = & \bar{\pi}_1 - C(\bar{\pi}_1, \bar{\pi}_2, \theta) \\ P(Y_1 = 0, Y_2 = 1) = p_{01} & = & \bar{\pi}_2 - C(\bar{\pi}_1, \bar{\pi}_2, \theta) \\ P(Y_1 = 0, Y_2 = 0) = p_{00} & = & C_\theta(\bar{\pi}_1, \bar{\pi}_2, \theta) \end{array} \tag{24}$$

In the sequel we leave out the arguments $(\bar{\pi}_1, \bar{\pi}_2, \theta)$ and write $C \equiv C(\bar{\pi}_1, \bar{\pi}_2, \theta)$. The loglikelihood is the sum of the contributions of all four possible outcomes:

$$L = Y_1 Y_2 \log(p_{11}) + (1 - Y_1) Y_2 \log(p_{01}) + Y_1 (1 - Y_2) \log(p_{10}) + (1 - Y_1)(1 - Y_2) \log(p_{00}). \tag{25}$$

The score can be written as

$$\frac{\partial L}{\partial \theta} = C_\theta \Big(Y_1 Y_2 \frac{1}{p_{11}} - (1 - Y_1) Y_2 \frac{1}{p_{01}} - Y_1 (1 - Y_2) \frac{1}{p_{10}} + (1 - Y_1)(1 - Y_2) \frac{1}{p_{00}} \Big), \tag{26}$$

where $C_\theta = \frac{\partial C}{\partial \theta}$.

B.1 Standard Errors

In order to compute standard errors, we first need to evaluate the Hessian. Using chain rule, the Hessian can be written as

$$\begin{aligned}\tfrac{\partial^2 L}{\partial \theta^2} &= C_{\theta\theta}\Big(Y_1 Y_2 \tfrac{1}{p_{11}} - (1-Y_1)Y_2\tfrac{1}{p_{01}} - Y_1(1-Y_2)\tfrac{1}{p_{10}} + (1-Y_1)(1-Y_2)\tfrac{1}{p_{00}}\Big) \\ &- (C_\theta)^2\left(Y_1 Y_2 \tfrac{1}{p_{11}^2} + (1-Y_1)Y_2\tfrac{1}{p_{01}^2} + Y_1(1-Y_2)\tfrac{1}{p_{10}^2} + (1-Y_1)(1-Y_2)\tfrac{1}{p_{00}^2}\right),\end{aligned} \tag{27}$$

where $C_{\theta\theta} = \frac{\partial^2 C}{\partial \theta^2}$. In the next section, we give expressions for C_θ and $C_{\theta\theta}$ for each copula.

a. Farlie Gumbel Morgenstern (FGM)

$$\begin{aligned} C_\theta &= \bar{\pi}_1\bar{\pi}_2(1-\bar{\pi}_1)(1-\bar{\pi}_2) \\ C_{\theta\theta} &= 0 \end{aligned} \tag{28}$$

For all Archimedean copulas, such as the Gumbel, Frank and Clayton, the score and Hessian depend on their generator function:[18]

$$\begin{aligned} C_\theta &= \tfrac{1}{\phi_C(C,\theta)}\left[\phi_\theta(\bar{\pi}_1,\theta) + \phi_\theta(\bar{\pi}_2,\theta) + \phi_\theta(C,\theta)\right] \\ C_{\theta\theta} &= \tfrac{1}{\phi_t(C,\theta)}\left[\phi_{\theta\theta}(\bar{\pi}_1,\theta) + \phi_{\theta\theta}(\bar{\pi}_2,\theta) - \phi_{\theta\theta}(C,\theta) - \left[\phi_{CC}(C,\theta)C_\theta + 2\phi_{C\theta}(C,\theta)\right]C_\theta\right] \end{aligned} \tag{29}$$

We now provide expressions for the partial derivatives of the generator ϕ that are needed for computation of score and Hessian for each of our remaining copulas.

b. Gumbel

$$\begin{aligned} \phi_C(t,\theta) &= -\tfrac{\theta(-\log(t))^{\theta-1}}{t} \\ \phi_\theta(t,\theta) &= \log(-\log(t))(-\log(t))^\theta \\ \phi_{CC}(t,\theta) &= -\tfrac{\theta(-\log(t))^\theta(\log(t)-\theta+1)}{t^2\log(t)^2} \\ \phi_{C\theta}(t,\theta) &= -((-\log(t))^{\theta-t}(\theta\log(-\log(t))+1))/t \\ \phi_{\theta\theta}(t,\theta) &= \log(-\log(t))^2(-\log(t))^\theta \end{aligned} \tag{30}$$

c. Frank

[18] This obtains by total differentiation of $\phi\left(C(\bar{\pi}_1,\bar{\pi}_2,\theta)\right) = \phi(\bar{\pi}_1,\theta) + \phi(\bar{\pi}_2,\theta)$ with respect to θ.

$$\begin{aligned}
\phi_C(t,\theta) &= -\frac{\theta}{e^{\theta t}-1}\\
\phi_\theta(t,\theta) &= \frac{e^{\theta t}-e^{\theta}}{(e^{\theta}-1)(e^{\theta t}-1)}\\
\phi_{CC}(t,\theta) &= \frac{\theta^2}{4\sinh^2(\theta t/2)}\\
\phi_{C\theta}(t,\theta) &= \frac{e^{\theta t}(\theta t-1)+1}{(e^{\theta t}-1)^2}\\
\phi_{\theta\theta}(t,\theta) &= -\frac{e^{\theta}(e^{2\theta t}+1)-e^{\theta t}(t^2e^{2\theta}-e^{\theta}(2t^2-2)+t^2)}{(e^{\theta t}-1)^2(e^{\theta}-1)^2}
\end{aligned}\tag{31}$$

d. Clayton

$$\begin{aligned}
\phi_C(t,\theta) &= -t^{-(\theta+1)}\\
\phi_\theta(t,\theta) &= -\frac{t^{-\theta}-1}{\theta^2}-\frac{1}{\theta}t^{-\theta}\log(t)\\
\phi_{CC}(t,\theta) &= (\theta+1)t^{-(\theta+2)}\\
\phi_{C\theta}(t,\theta) &= \log(t)t^{-(\theta+1)}\\
\phi_{\theta\theta}(t,\theta) &= \frac{\theta^2\log(t)^2-2t^{\theta}+2\theta\log(t)+2}{t^{\theta}\theta^3}
\end{aligned}\tag{32}$$