Income and Wealth Distributions: An Application of Copulas

Thomas Hauner* Hui Liu

This draft: May 2015

1 Introduction

Copulas are a useful tool to measure the dependency between two or more random variables. While many economic models rely on the assumption of a marginal Gaussian distribution for each of the underlying random variables in estimation, this frequently fails empirically. The copula approach is attractive for its flexibility and precision with modeling the underlying marginal distributions. In an economic model it can improve, for example, the efficiency of a maximum-likelihood estimator by specifying a more accurate density function.

Aside from estimating models, copulas are also a useful tool for measuring the dependence between two non-elliptically distributed random variables as it changes over time–in particular the extreme values in their tails. Because copulas rely on concordance metrics, or measures of association between random variables independent of Pearson's linear correlation coefficient, copulas can more accurately describe the relationship between two non-elliptical variables.

We apply empirical copula estimation to measure the joint distribution of two random variables, a cross-section of US household income and wealth at a given point in time, and then compare predicted theoretical copulas as possible statistical models. Further, we aim to understand changes in the joint distribution as a result of the Global Financial Crisis and ensuing recession by paying particular attention to changes between 2007 and 2010, relying on the joint distribution of income and wealth in the United States based on the triennial Survey of Consumer Finances (SCF). Thus

*PhD Program in Economics, The Graduate Center, City University of New York (CUNY).
Contact: thauner@gradcenter.cuny.edu

our paper is a synthesis of both the income inequality literature as well as the applied statistical literature focusing on copulas.

Overall, we find that bivariate empirical copula density captures nuanced changes in the dependence between income and wealth. It also captures the lower quantile dependence better than the highest quantile dependence as the latter has too high a bound. The collapse of the housing market in 2007 caused a major drop in the households' net worth reflected in the lower tail dependence between income and wealth slightly drifting apart, while the upper tail dependence also increased through 2007 and then decreased. The Gumbel copula tail dependence most closely follows the empirical tail dependence estimates. Despite the empirical copula's resemblance to the shape and tail behavior of a Gumbel copula, we reject the null that Gumbel describes the income and wealth dependence–though it does so relative to the class of elliptical copulas.

2 Background

2.1 Measurements of Inequality Dependence

The study of income distribution has been active since Vilfredo Pareto's seminal research in 1896 on the apparent log-linear relationship between income and the percentage of the population with at least a certain level of income. It was another Italian statistician, Corrado Gini, who developed the popular Gini coefficient, a measure of income inequality across a population that is bounded between zero and one. Statisticians have developed myriad derivations of the Gini, however one intuitive metric is the ratio of the area between the 45 degree line and the Lorenz curve to the total area under the 45 degree line.

Such tools are applied to wealth and income distributions independently, as are other metrics such as "mean-median" ratios or the increasingly used income concentration of top quantiles as calculated from fiscal rather than household data. All mostly consider univariate distributions at a single point in time.

Of course we may also consider the actual marginal density functions by treating income or wealth as a random variable. While these provide a good understanding of how wealth or income are

distributed, they do not provide any information regarding the relationship between the two. The wealth-income ratio plotted over wealth/income can also be useful, but a better measure is the joint distribution. Kennickell (2009) describes the difficulty in estimating the joint distribution directly due to the wide interval of both wealth and incomes. Thus no true dependency measure exists between income and wealth suggesting a copula approach may be useful.

2.2 Copulas in Welfare Research

Application of copulas to questions of income and welfare is relatively fertile ground. However there do exist some novel contributions in recent literature upon which we can build.

While not the focus of this extensive summary of the 2007 SCF results, or even its main methodology, Kennickell (2009) utilizes empirical copulas to describe the dependence between wealth and income distributions between 1989 and 2007. He indirectly uses the copula method by simply dividing each marginal distributions into five-percentile-point groups and computing the cross product. He finds strikingly high dependence in the tails of the distributions (the top and bottom five-percentiles) and a dispersed relationship in between.

Of particular interest, given the Global Financial Crisis of 2007-2008, is the change in the dependence between the tails since 2007. Bricker et al. (2012) consider a panel subset of the SCF, collected between 2007 and 2009, however they do not employ copulas and instead focus on demographic and family unit characteristics to describe the changes in wealth. They also consider the most recent cross-sectional SCF in 2010, but do not use copulas and focus on changes in household wealth by demographic groups.

Decancq (2013) considers a multifaceted analysis of social well-being as a supplement to the Human Development Index, which is insensitive to dependence between parameters, arguing that well-being is dependent on many components and their interactions. He tests his multivariate copula approach to study welfare in the Russian Federation after Soviet disintegration, showing a richer analysis of improved well-being since 1995. Quinn (2007) introduces the use of copulas to measure dependency between health outcomes, typically an ordinal variable, and income, a continuous variable. These categorical incongruities traditionally made direct comparisons difficult. The author's results enable

a more accurate ranking of health inequality and overall health outcomes by country.

Somewhat more aligned with our paper's strategy, though with a labor focus, Chau (2010) estimates cross-sectional distributions of wages for the marginal distributions and uses copulas to describe their relative positions over time. That is, an individual's income y_t is some function of the previous periods, and thus the copula density function is of the form $f(y_t, y_{t-1})$. He applies this structure to examine wage mobility whereby the flexibility of the copula can model changing wage distributions over time.

Each of the above papers utilize the flexibility of copulas for a broad scope of welfare applications to gain a more nuanced understanding of the dependency between the random variables considered. Only Kennickell (2009) applies them to describe the joint distribution of income and wealth, however no attempt at defining a theoretical statistical model is made.

We aim to make a novel contribution to this seemingly intuitive application by considering such models as well as the effects of the Global Financial Crisis on the dependence between income and wealth.

3 Data

3.1 Survey of Consumer Finances

The Survey of Consumer Finances (SCF) is the most thorough household survey of wealth in the United States. A triennial cross-section organized by the Federal Reserve, it contains micro-level data on family balance sheets, pensions, income, and demographic characteristics among other variables. One important feature of the survey is its oversampling of higher incomes, correcting a known bias in other household surveys which under sample higher income households. Thus it provides a more representative sample of the total population's income and wealth distributions. For example, the 2010 SCF includes 6,492 interviews, 5,012 resulting from a standard probability design and 1,480 resulting from fiscal data in order to disproportionately select families that were likely to be relatively wealthy and less likely to participate. (The latter is referred to as the list sample.)

Our data sample begins with the 1989 SCF, the first implementation of the multiple-imputation methodology (discussed below), and continues through the most recent survey in 2010. While we will refer to the unit of measurement as a household, the actual observations are on primary economic units, i.e. the economically dominant occupant or couple in a household and all individuals financially dependent on them.

3.1.1 Imputation

Imputed values are critical to the SCF as many respondents omit answers or provide only a range of values to specific questions. Additionally imputation helps anonymize the identifiable observations from the list sample of higher incomes and accumulated wealth. Since 1989, the first SCF year included in our analysis, a finite value is imputed by the FRB Research & Statistics group five times using a multiple imputation model–the FRITZ (Federal Reserve Imputation Technique Zeta) system, a multiple imputation (MI) model developed specifically for the survey. (See Kennickell (1998) for further detail on the Fed's SCF imputation methodology.) The individual imputations are made by drawing repeatedly from a previous estimate of conditional distributions of the data variable. The five unique imputations are stored as successive observations for each data record, or survey respondent. Therefore in the 2010 SCF, for example, the number of observations in the full data set (32,460) is five times the actual number of respondents (6,492)

Kennickell (2000) cites two distinct advantages to multiple imputation. First, "it is more efficient in that one can expect to get a more efficient estimate from multiple estimates of a missing value than from a single estimate, at least if there is any randomization involved in the imputation process, as there is in the SCF model. Second, MI makes clearer the uncertainty induced by having to make estimates based on partial information."

3.1.2 Variables

Based on the publicly available SCF data extract, we focus on two summary variables computed by the Fed: total household income (inc), inclusive of wages, rent, capital gains, interest, business and farm, social security and retirement and other income sources; and net worth ($netw$), where

net worth is calculated as total financial and non-financial assets (including vehicles, homes, and business interests) minus total debt. Both variables are adjusted for inflation in the available data. Because the SCF is not an equal-probability survey, weights for each observation and imputed value are critical to interpreting results. We use SCF's public weight variable (wgt) to compute weighted means, medians, and standard deviations in tables (1) and (2). The weights, calculated by the Fed survey team, use original selection probabilities and frame information along with aggregate control totals estimated from the Current Population Survey.

Figure 8 from Kennickell (2009) (Appendix A) shows the overall cumulative distribution of net worth and income between 1989 and 2007. The marginal distribution functions (plotted on log scales) are clearly unique, changing over time, and non-Gaussian.

The inequality literature is primarily focussed on income inequality as its data are more often available both through fiscal sources and household surveys. Some authors, Kopczuk & Saez (2004), have utilized estate tax data from the IRS to impute wealth from bequests. Thus the SCF is unique in its information on household wealth formation.

3.1.3 Descriptive Statistics

Though the SCF was first implemented in 1963, it has been updated triennially between 1989 - 2010. The following are the descriptive statistics of the full samples for each SCF year in that period–that is, data sets include each set of implicates as independent observations in each survey year, therefore increasing the total number of observations to a multiple of five from the original number of survey respondents.

Year	1989	1992	1995	1998	2001	2004	2007	2010
Median	79,374	75,349	81,864	95,640	106,288	107,125	126,769	77,000
Mean	319,397	283,610	301,853	378,344	487,155	516,812	583,351	494,916
SD	1,760,650	1,734,220	1,893,431	2,448,251	2,558,802	2,932,968	3,505,423	3,140,209
n	15,715	19,530	21,495	21,525	22,210	22,595	22,085	32,410

Table 1: US Household Net Worth Descriptive Statistics

Note: All figures inflation adjusted and weighted by survey design probability weights. Final values may differ from Bricker et al. (2012) because of slight discrepancies between the publicly available data and the Fed researchers' data. The number of observations includes multiple imputations.

Year	1989	1992	1995	1998	2001	2004	2007	2010
Median	43,985	40,417	43,513	44,633	49,131	49,633	49,561	45,743
Mean	69,181	60,734	63,264	70,784	84,712	81,313	88,162	78,332
SD	289,382	125,129	228,439	323,179	284,006	249,119	381,246	282,390
n	15,715	19,530	21,495	21,525	22,210	22,595	22,085	32,410

Table 2: US Household Income Descriptive Statistics

Note: All figures inflation adjusted and weighted by survey design probability weights. Final values may differ from Bricker et al. (2012) because of slight discrepancies between the publicly available data and the Fed researchers' data. The number of observations includes multiple imputations.

4 Methodology

4.1 Preliminaries

4.1.1 Rank Statistics

Copulas rely on a rank-based measure of variable dependence. That is, correlation of two random variables is not calculated as a ratio between covariance and variance per se but of relative rank of pairs of sorted observations observations. Therefore rank correlation is a non-parametric measure of dependence based on ranked data. In our estimates we rely on Kendall's tau, which compares all possible dyads, or pairs of observations, in the data:

$$\tau = \frac{N_C - N_D}{\frac{1}{2}n(n-1)}. \tag{1}$$

Pairs of observations $(\mathbf{x}, \mathbf{y})$ are said to be *concordant* if $(x_1 - x_2)$ has the same sign as $(y_1 - y_2)$. The number of such pairs is equal to N_C. (The number of discordant pairs is N_D.)

4.1.2 Copulas

At its simplest, a copula is a joint distribution function. Sklar's (1959) seminal theorem states that distinct copulas define distinct joint density functions, given a fixed set of continuous uniform marginal distributions. This implies that given a joint density of d-dimension $F(x_1, x_2, \ldots, x_d)$ with continuous and uniform marginals, one can "back out" a unique copula C such that the following holds:

$$F(x_1, x_2, \ldots, x_d) = C(F_1(x_1), F_2(x_d), \ldots, F_d(x_d))$$

One can also then find the associated copula density function:

$$c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) = \frac{\partial^d C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d))}{\partial F_1(x_1) \partial F_2(x_2) \ldots \partial F_d(x_d)}$$

Thus given the above copula density and the marginal densities, assuming they exist, it is possible to derive the joint density function of the original random variables:

$$f(x_1, x_2, \ldots, x_d) = f_1(x_1) f_2(x_2) \ldots f_d(x_d) c(F_1(x_1), F_2(x_2), \ldots, F_d(x_d))$$

4.2 Copula Estimation

4.2.1 Empirical Copula

From Alexander (2008), calculating the empirical copula distribution requires the sample order statistics of the random variables. That is, if $x^{(1)} = min\{x_1, \ldots, x_n\}$ then $x^{(2)}$ represents the second smallest observation in the sample of size n and $x^{(n)} = max\{x_1, \ldots, x_n\}$. Then the empirical copula distribution function for two random variables X and Y both of sample size n is

$$\hat{C}(\frac{i}{n}, \frac{j}{n}) = \frac{\text{Number of dyads } (x, y) \text{ s.t. } x \leqslant x^{(i)} \text{ and } y \leqslant y^{(j)}}{n}$$

where $x^{(i)}$ and $y^{(j)}$ represent the sample order statistics.

All of the above definitions require that the random variables be transformed such that they have uniform marginal distributions. And as mentioned previously, the flexibility of the copula is aided by its reliance on rank statistics rather than linear correlations. The goodness-of-fit test, described below, will similarly be rank-based. Thus the n observations for our two random variables, $X_1 = (x_1, y_1), \ldots, X_n = (x_n, y_n)$ can be transformed into pseudo-observations deduced from each observation's rank: $Z_{ij} = \frac{R_{ij}}{(n+1)} = \frac{n\hat{F}_j(X_{ij})}{(n+1)}$, where R_{ij} is the rank of X_{ij} amongst $(X_{1i}, \ldots, X_{ni})$ and Z_{ij} is the empirical marginal distribution of X_{ij}, or $n\hat{F}_j(X_{ij})$. In other words, each margin is transformed through their normalized ranks. Thus the empirical copula can be rewritten as

$$\hat{C}_n(\mathbf{u}) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{I}\{(Z_{i1} \leqslant u_1, Z_{i2} \leqslant u_2)\} \tag{2}$$

where $\mathbf{u} = (u_1, u_2) \in [0, 1]^2$ and $\mathbf{I}$ is an indicator function.

4.2.2 Gumbel Copula

The Gumbel copula, as a type of Archimedean copula, is derived from a generator function rather than an implicit multivariate distribution like the Gaussian or Student's-t copulas. More generally, for some generator function $\Psi(u)$ one can define the corresponding Archimedean copula by

$$C(u_1, u_2, \ldots u_d) = \Psi^{-1}(\Psi(u_1), \Psi(u_2), \ldots, \Psi(u_d))$$

and its corresponding copula density by

$$c(u_1, u_2, \ldots u_d) = \Psi^{-1}_{(d)}(\Psi(u_1) + \Psi(u_2) + \cdots + \Psi(u_d)) \prod_{i=1}^{d} \Psi(u_i)'$$

where $\Psi^{-1}_{(d)}$ is the dth derivative of the inverse generator function. (For the Gaussian copula, $\Psi(u) = \Phi(u)$.) The Gumbel copula has a generator function

$$\Psi(u) = -(\ln u)^{\alpha}, \ \alpha \geqslant 1. \tag{3}$$

and thus we can derive the bivariate Gumbel copula density as

$$c(u_1, u_2; \alpha) = (A + \alpha - 1)A^{1-2\alpha} \exp(-A)(u_1 u_2)^{-1}(-\ln u_1)^{\alpha-1}(-\ln u_2)^{\alpha-1} \tag{4}$$

where

$$A = [(-\ln u_1)^{\alpha} + (-\ln u_2)^{\alpha}]^{\frac{1}{\alpha}} . \tag{5}$$

When the parameter $\alpha = 1$ then the marginals are perfectly independent. While the copula literature extensively discusses efficient estimation of copula parameters using maximum likelihood and other methods, the general Archimedean parameter θ can be also estimated by inverting Kendall's tau such that $\hat{\theta}_n = \tau^{-1}(\tau_n)$, which yields the theoretical bivariate Gumbel copula density estimate $c_{\hat{\theta}_n}$.

We closely consider Gumbel copulas as a possible theoretical model for the bivariate distribution of income and wealth in the US. This is motivated by the fact that the income distribution in the US tends to follow a log-normal distribution, as shown for 2010 in figure (2) (Appendix B). Thus given the generator function in equation (3), the Gumbel copula is an obvious candidate. Furthermore, initial visual comparisons between estimates of the 2010 empirical copula density (panel (b), figure (3) in Appendix B) with the Gumbel copula reveals both display high density in the upper tail and lower density in the lower tail.

4.2.3 Tail Dependence

Given the high tail densities of each univariate empirical distribution we are interested in the concordance in the tails, or tail dependence. The lower tail dependence, for the ijth observation, is defined as

$$\lambda^l_{ij} = \lim_{q \to 0} \Pr\left(X_i < F_i^{-1}(q) | X_j < F_j^{-1}(q)\right) \Rightarrow \lambda^l = \lim_{q \to 0} \frac{C(q, q)}{q} \tag{6}$$

where q represents the quantile. It exists when the limit exists and $\lambda^l > 0$. Similarly, the upper tail dependence is defined as

$$\lambda^u_{ij} = \lim_{q \to 1} \Pr\left(X_i > F_i^{-1}(q) | X_j > F_j^{-1}(q)\right) \Rightarrow \lambda^u = \lim_{q \to 1} \left[\frac{\bar{C}(1 - q, 1 - q)}{(1 - q)}\right] \tag{7}$$

where $\bar{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2)$ and represents a survival copula.

4.2.4 Goodness of Fit

When we estimate the goodness-of-fit (GoF) of an empirical bivariate copula C we want to test the hypothesis

$$\mathcal{H}_0 : C \in \mathcal{C}_{\hat{\theta}} \text{ vs } \mathcal{H}_A : C \notin \mathcal{C}_{\hat{\theta}} \text{ where } \hat{\theta} \text{ is the parameter estimate.} \tag{8}$$

In order to determine the goodness of fit for any theoretical copula modeling the empirical copula, we must have a method of calculating a test-statistic and some sort of distribution to determine inference. Alexander (2008) suggests using root mean square errors to determine a parametric copula's fit to an empirical copula.

Genest et al. (2009) propose an intuitive method analogous to univariate nonparametric testing whereby a distance measure compares the empirical copula $\hat{C}_n$ to the hypothetical copula $\mathcal{C}_{\hat{\theta_n}}$ imposed under the null. A rank-based version of the familiar Cramér-von Mises statistic follows:

$$\hat{T}_n = n \int_{[0,1]^2} \{\hat{C}(\mathbf{z}) - C_{\hat{\theta}}(\mathbf{z})\}^2 d\hat{C}(\mathbf{z}) = \sum_{i=1}^{n} \{\hat{C}(z_i) - C_{\hat{\theta}}(z_i)\}^2. \tag{9}$$

In each of the GoF tests we consider $C_{\hat{\theta}}(\mathbf{z})$ takes an analytical form (Gaussian, Student's-t, Clayton and Gumbel copulas). The test statistic must be bootstrapped but is consistent, though computationally intensive. The p-value is estimated as

$$\hat{p} = \frac{1}{R+1} \sum_{r=1}^{R} \mathbf{I}\{\hat{T}_r \geqslant \hat{T}\} \tag{10}$$

where $\hat{T}_r$ is the test statistic for each bootstrap sub-sample $r \in \{1, \ldots, R\}$.

The method outlined above is the default used in the `gofCopula` command from the `Copula` library. Its consistency is proven in Genest & Rémillard (2008). Berg (2009) determines that the above is the best approach overall out of nine different GoF methods compared, based on extensive Monte Carlo trials. He finds the test statistic demonstrates increasing power with dimension, sample size and dependence.

5 Results

In estimating the above copulas and test statistics we alternately rely on the `fCopulae` (Wuertz et al. (2007)) and `Copula` (Yan (2007)) libraries in R. The former is used mainly to calculate rank statistics and estimate bivariate copulas whereas the latter is used for estimating GoF test statistics.

5.1 Empirical Copula Densities

Graphical results of the bivariate empirical copula densities and contour plots are presented in Appendix C for each SCF year.

Two stylized facts are observed regarding the bivariate income and net worth empirical copula densities. First, there exists extremely high upper tail densities for the top wealth and income quantiles. The sharp spike in the far corner of the plot is several times higher than any other density and is relatively static. This makes intuitive sense since high incomes enable saving and therefore asset accumulation which appears as net worth, or wealth, in the SCF. This is consistent with Piketty & Saez (2006), who show that the US income distribution is highly skewed at the top quantiles, even fractal in a way. Our copula densities capture the overall distributions and hence the static spike at top quantiles.

Second, there is relatively high density in the lowest income and wealth quantiles. This smaller spike is more dynamic and representative of broader economic fluctuations. It also captures the fact that incomes are bounded below by zero while wealth is not. All years of the SCF have observations with negative net worth.

From 1995 to 2001 (figures (6) - (8) in Appendix C) there is a noticeable increase in the density in the middle of the copula for both income and net worth quantiles. The contour plots make seemingly nuanced changes clearer. This could be attributed to the overall growth of the economy in those periods, aided by the stock-market boom, particularly the Nasdaq. However, there is a distinct shift in 2001. Whereas somewhat of a ridge along the diagonal existed in previous years, in 2001 it shifts decidedly towards higher wealth quantiles. Also a ridge forms along the net worth

quantile at zero income, punctuated by a large spike at the lowest quantiles. The latter contradicts the assertion by Sufi & Mian (2014) that the poorest households were not substantially effected by the tech bubble bursting.

From 2001 to 2007 one again sees a noticeable ridge forming along the diagonal of the density plots. This parallels the inflation of the mortgage bubble. As households across the income distribution were sold over-leveraged mortgages, their net worth increased as well. This net worth was mostly composed of housing. In 2010 the body of the density plot smooths out for all quantiles, except the top and bottom, as the Global Financial Crisis eroded enormous amounts of wealth. Notably, the spike in the lowest quantiles shifts for the first time. In all other survey years the spike occurred for both the lowest income and net worth quantiles. In 2010 the spike in density remained in the lowest wealth quantile, however shifted to the second-lowest income quantile. This suggests that households who lost the most wealth, and are "underwater" on their mortgages, had low, but not the lowest incomes. Perhaps those with the lowest incomes successfully declared bankruptcy, expunging their negative net worths, while those with some income faced increased indebtedness.

From the empirical marginal transformation of the income and net worth variables, the scatterplots in Appendix C.1 make clear the dispersion of negative wealth after the crisis in 2010 as well as the broad general decreases in income and wealth.

5.1.1 Tail Dependence

We use tail dependence to measure how likely income and net worth (wealth) concentrate in the extreme values. We consider both upper tail and lower tail dependence as they represent two different subsets of the population, the very poor and the very rich. The graphs created (see Appendix D) therefore include two parts: for percentiles lower than 0.5 we look at the lower tail dependence, while for percentiles higher than 0.5 we examine upper tail dependence.

The black lines plot empirical tail dependence, which is the probability of wealth being higher (lower) than a certain percentile given the income is higher (lower) than this percentile. For each year the lower tail appears to get very close to zero in limits. On the other hand, the upper tail does not go to zero for most of the years. This is consistent with the quantile scatter plots that

the observations are highly concentrated on the upper right corner (the richest and the wealthiest) while the observations at the lower left corner is relatively scattered. Intuitively there is a higher chance that an observation is among the richest and the wealthiest, whereas the observation does not have to be among the least wealthy ones when the income is among the lowest.

An interesting pattern within upper tail dependence is the slight increase when the percentile goes from roughly 0.5 to 0.8. This supports the idea that a larger portion of wealth can be accumulated when income is higher, particularly greater than median income and wealth.

For each SCF dataset, the empirical tail dependence is contrasted with the ones that should be obtained with the corresponding Kendall's tau for Gaussian, Gumbel, Clayton and Student-t copulas. Among the parametric results, Gumbel copula appears to have the best fit, especially for the lower tail. None of the parametric copulas capture the upper tail dependence very well. Further research could look into other potential copulas to obtain a better fit.

Too see how the tail dependence changes over time, figure (28) (Appendix D) compares the empirical results across SCF years. The lower tails are fairly similar for the years 2007 and before. The 2010 graph is significantly lower than the other years. This may be due to the financial crisis in 2008. Specifically, the collapse of the housing market caused a major drop in the households' net worth, therefore the dependence between income and wealth slightly drifted apart.

As for the upper tail, a steady upward movement can be observed for years 1992 to 2007, showing a higher chance to be among the more wealthy households when the income is higher. Again, the dependence dropped in 2010 and this phenomenon can be explained similarly as above. Also notice that this dependence is not affected much for the households with the highest income (above 0.8). The probability of their net worth being higher than 0.8 or above is still high and the graph follows the upward movement trend since 1992.

5.2 Goodness-of-Fit

We test hypothetical bivariate copulas $\mathcal{C}_{\hat{\theta}}$ under the null in equation (8) for Gaussian, Student's-t, Clayton and Gumbel copulas, respectively. Results for test-statistics, parameters, and approximated p-values are presented in Appendix E.

As expected, the bootstrapped p-value in equation (10) is decreasing in the sample size and number of bootstrap sub-samples. Thus when we use $R = 100$ iterations we find p-values of 0.00495 across the surveys for the Gumbel copula fit, but a p-value of 0.0455 for $R = 10$ iterations. For both iteration sizes we reject the null of a bivariate Gumbel copula for income and net worth.

Because Genest & Rémillard (2008) show that the GoF test statistics based on T_n are consistent, i.e., if $C \notin \mathcal{C}_{\hat{\theta}}$, $\mathcal{H}_0$ is rejected with probability 1 as $n \to \infty$. We have a relatively large sample size for each SCF year, thus rejection is consistent even at lower iteration numbers–an advantage given the computational intensity of bootstrapping.

We also reject hypothetical Gaussian, Student's-t and Clayton copulas to describe the data for any given survey year, though test statistics are an order of magnitude larger than the Gumbel copula test statistics. Thus while we reject the hypothetical Gumbel copula as a good fit, it is the best fit of all four.

6 Conclusion

In this paper we examined the changing dependency between income and net worth in the US and their relative distributions over the last 20 years by applying copula estimation. Using a newer technique to examine an old idea, we found that the bivariate empirical copula density provides a nuanced understanding of the overall distribution changes, particularly the changes in lower tail densities. The higher tail densities are so high and concentrated that it's difficult to parse any significance to its changes. From this simple exploration the top quantile looks relatively static, which is intuitive given the rank-based statistical reliance. Both upper and lower empirical tail dependence is most closely modeled by the Gumbel copula tail dependence behavior.

From these observations we posited to test the fit of a hypothetical Gumbel copula, however rejected the null. This seems driven by the extreme density in the top quantile tile relative to the rest of the copula.

Because this was a preliminary exploration of a new application, there exist many avenues for future research. Given certain assumptions about the primary economic units' survey responses

one could decompose net worth into financial and non-financial assets and income into wage income and capital income. Thus for the US we may indirectly consider the fiscal evidence presented by **?**, that the share of capital income in the US is growing over the last 30 years.

Examining the role of real estate assets over the past decade would also be instructive given the enormous role distributional assumptions imbedded in mortgage financing and portfolio management played in the crisis.

Another consideration could be mixed-formation copulas of the myriad other components captured by the SCF, including various asset classes. Finally, there also exists potential to expand the analysis to other countries and use copulas as a tool to better understand the nuanced dependency amongst income and wealth variables. While the inaccessibility to individual observations and reliance on STATA programming make the Luxembourg Income Study and Luxembourg Wealth Study data less ideal, the recent Household Finance and Consumption Study implemented by the ECB, and modeled after the SCF, would be a natural comparison.

References

Alexander, C. (2008). *Market Risk Analysis, Practical Financial Econometrics*. Market Risk Analysis. John Wiley & Sons.

Berg, D. (2009). Copula Goodness-of-Fit Testing: An Overview and Power Comparison. *The European Journal of Finance*, *15*(7-8), 675–701.

Bricker, J., Kennickell, A. B., Moore, K. B., & Sabelhaus, J. (2012). Changes in US Family Finances from 2007 to 2010: Evidence from the Survey of Consumer Finances. *Federal Reserve Bulletin*, (June), 1–80.

Chau, T. W. (2010). *Essays on Earnings Mobility Within and Across Generations Using Copula*. Ph.D. thesis, University of Rochester.

Decancq, K. (2013). Copula-Based Measurement of Dependence Between Dimensions of Well-Being. *Oxford Economic Papers*, (p. gpt038).

Genest, C., & Rémillard, B. (2008). Validity of the Parametric Bootstrap for Goodness-of-Fit Testing in Semiparametric Models. In *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques*, vol. 44, (pp. 1096–1127).

Genest, C., Rémillard, B., & Beaudoin, D. (2009). Goodness-of-Fit Tests for Copulas: A Review and a Power Study. *Insurance: Mathematics and Economics*, *44*(2), 199–213.

Kennickell, A. B. (1998). Multiple Imputation in the Survey of Consumer Finances. In *Proceedings of the Section on Business and Economic Statistics*, Joint Statistical Meetings, (pp. 63–74). Dallas, TX.

Kennickell, A. B. (2000). Wealth Measurement in the Survey of Consumer Finances: Methodology and Directions for Future Research. Tech. rep., Federal Reserve Board.

Kennickell, A. B. (2009). *Ponds and Streams: Wealth and Income in the US, 1989 to 2007*. Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board.

Kopczuk, W., & Saez, E. (2004). Top Wealth Shares in the United States: 1916-2000: Evidence from Estate Tax Returns. NBER Working Paper 10399, National Bureau of Economic Research.

Piketty, T., & Saez, E. (2006). The Evolution of Top Incomes: A Historical and International Perspective. *American Economic Review*, *96*(2), 200–205.

Quinn, C. (2007). Using Copulas to Measure Association Between Ordinal Measures of Health and Income. Tech. Rep. Working Paper 07/24, HEDG, c/o Department of Economics, University of York.

Sufi, A., & Mian, A. (2014). Why the Housing Bubble Tanked the Economy And the Tech Bubble Didn't.
URL http://fivethirtyeight.com/features/why-the-housing-bubble-tanked-the-economy-and-the-tech-bubble-didnt

Wuertz, D., Wuertz, M. D., & Team, R. C. (2007). The fCopulae Package.

Yan, J. (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, *21*(4), 1–21.

A Appendix

A Cumulative Distributions

Figure 8: Cumulative distributions of 1988 and 2006 income and 1989 and 2007 net worth.

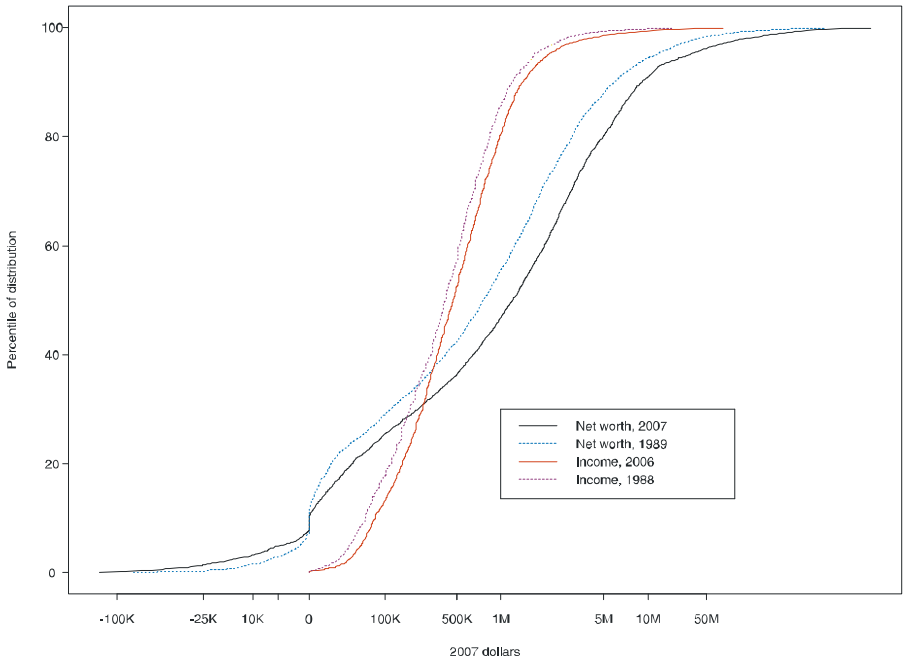

Figure 1: Kennickell (2009)

B Gumbel Copula Motivation

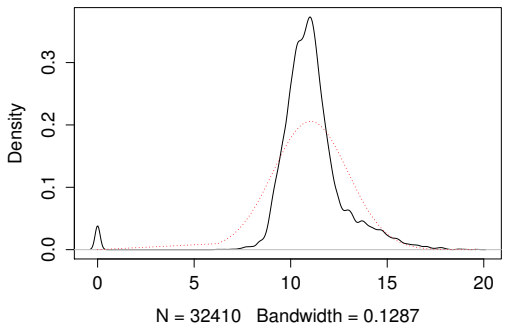

Figure 2: Estimated Density of $log(inc)$, from 2010 SCF, with Imposed log-Normal Distribution

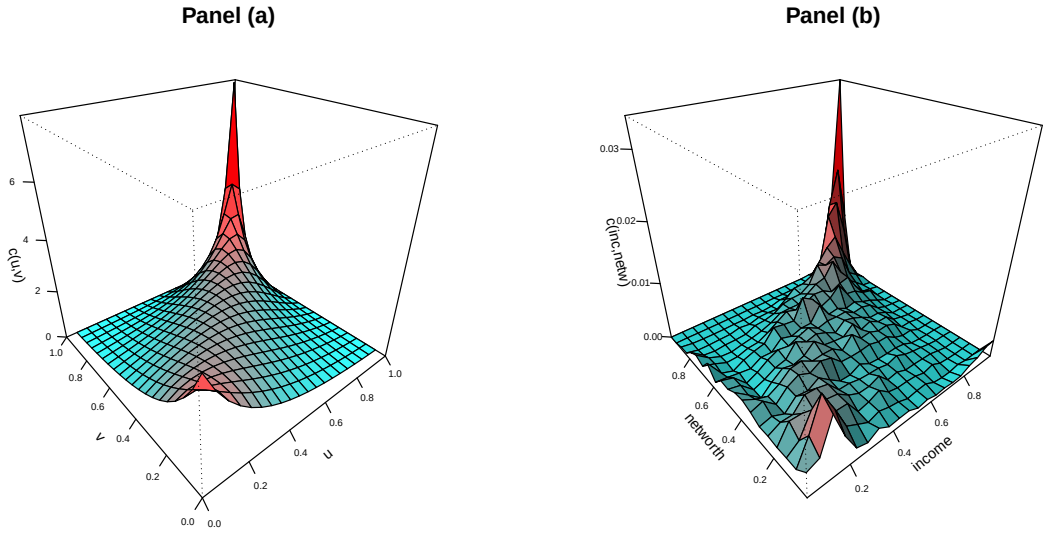

Figure 3: Theoretical and Empirical Copula Densities
Panel (a) displays the bivariate Gumbel copula density with a parameter $\alpha = 2.0639$, as calibrated with Kendall's tau from 2010 SCF income and net worth variables. Panel (b) displays the bivariate empirical copula density for income and net worth in the 2010 SCF.

C Empirical Copula Densities

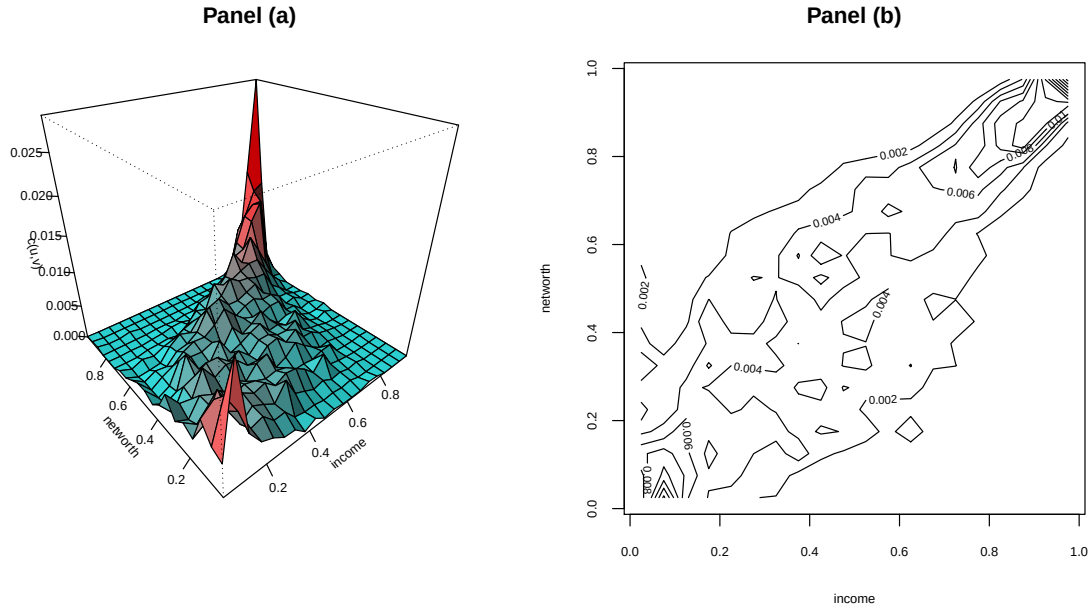

Figure 4: 1989 SCF

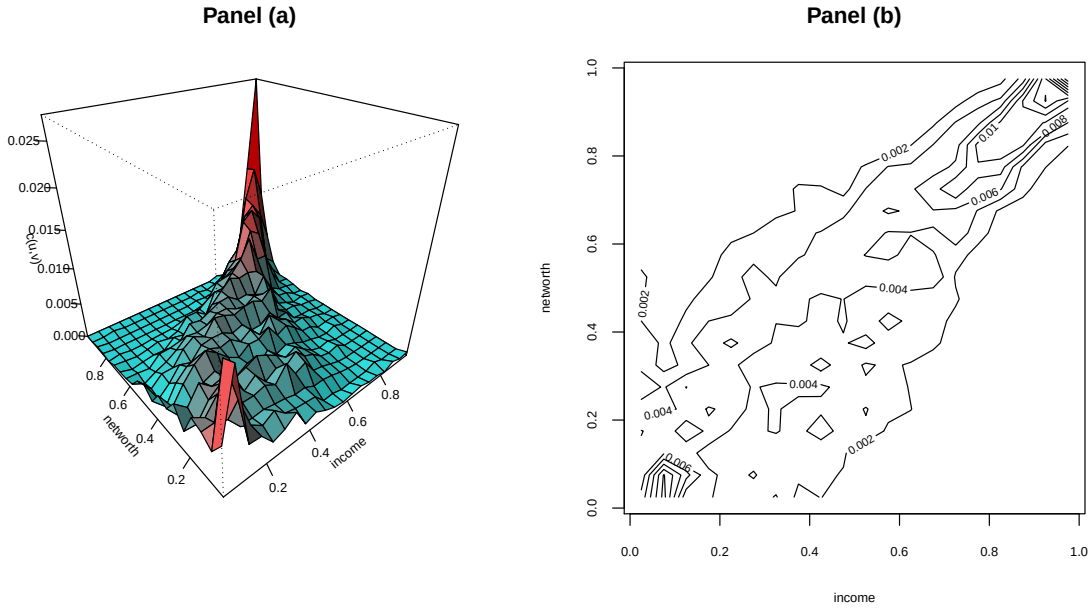

Figure 5: 1992 SCF

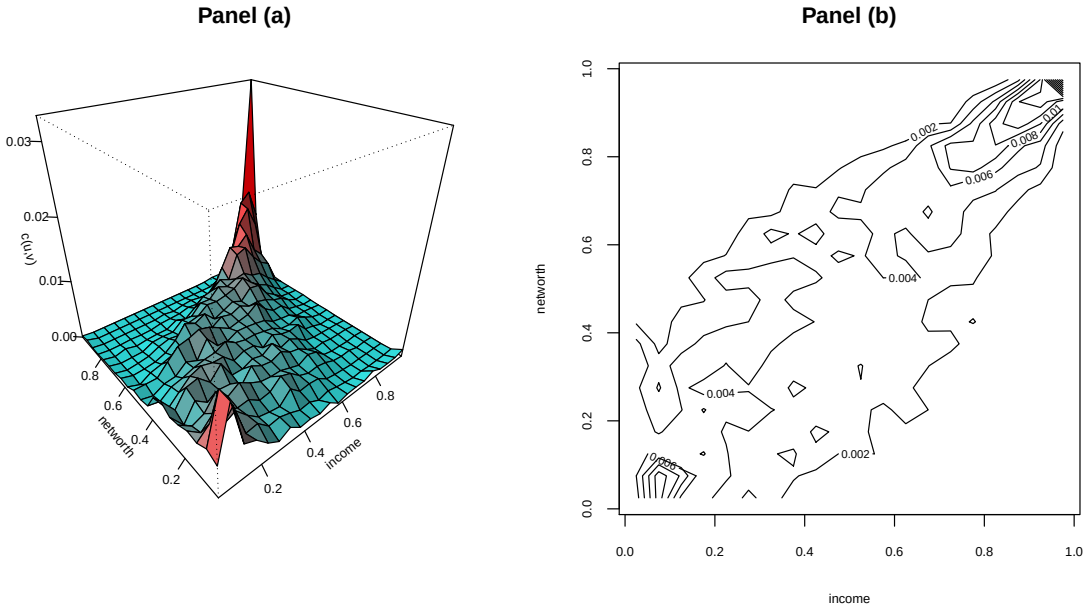

Figure 6: 1995 SCF

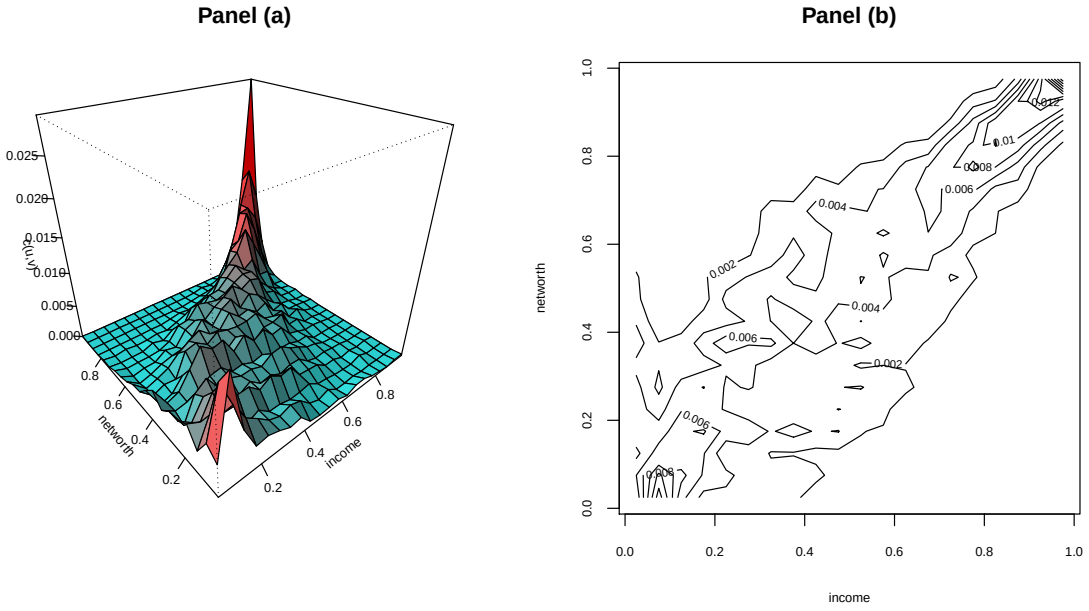

Figure 7: 1998 SCF

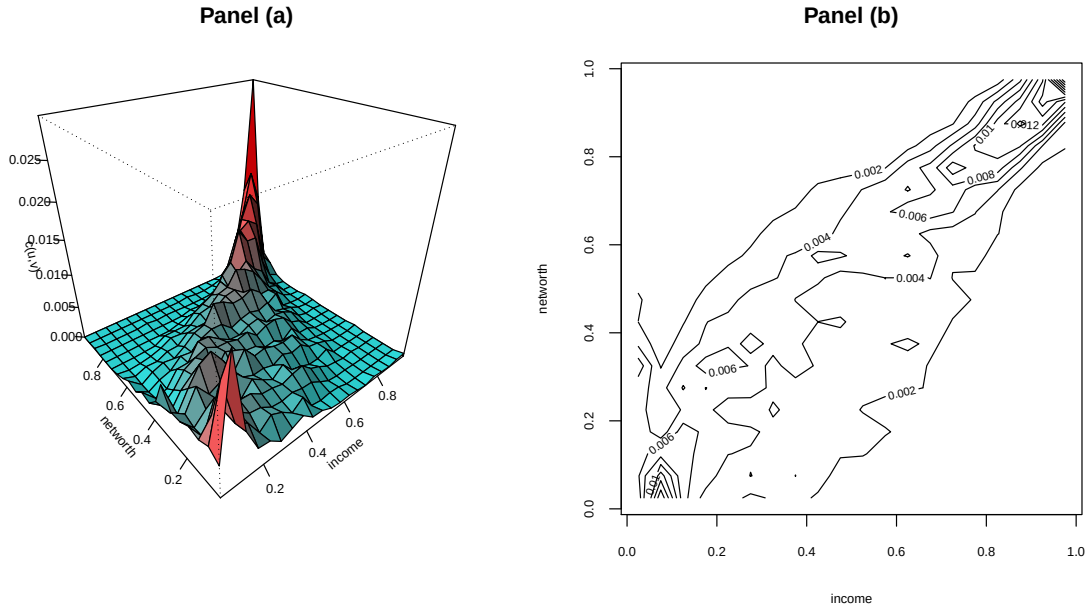

Figure 8: 2001 SCF

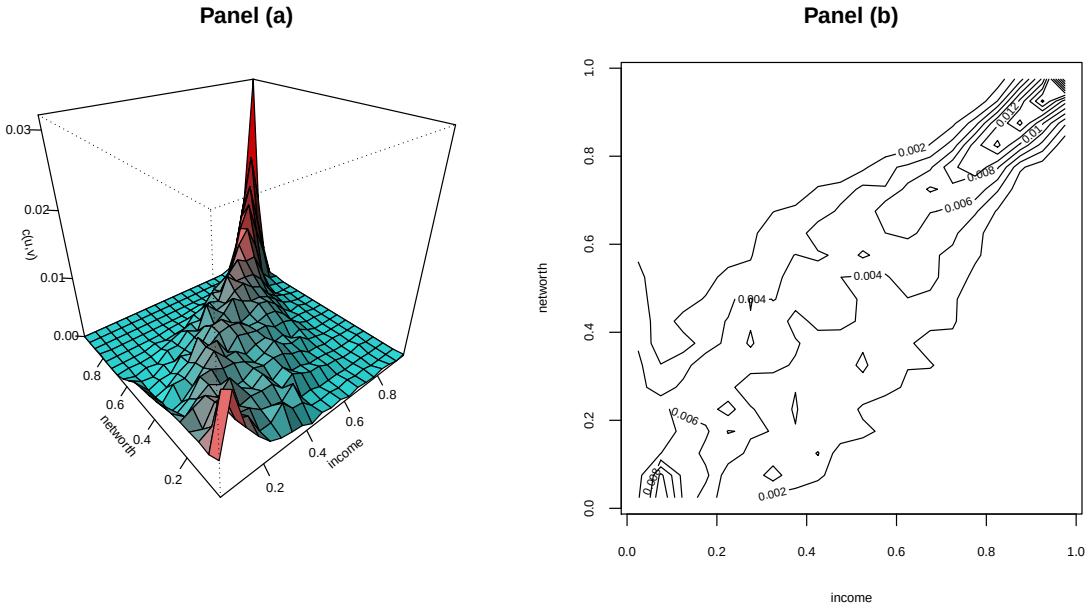

Figure 9: 2004 SCF

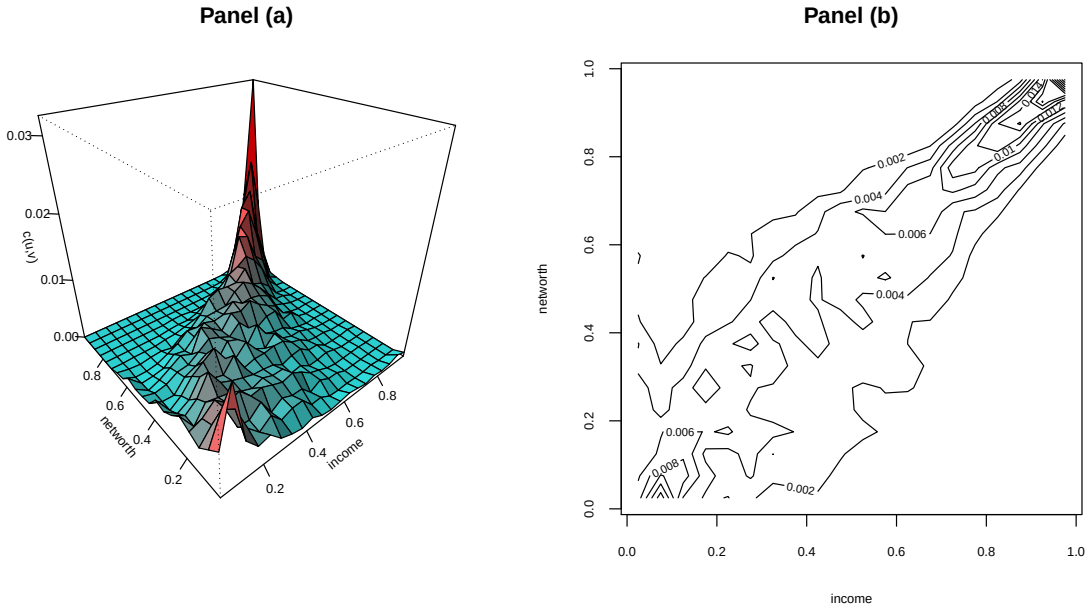

Figure 10: 2007 SCF

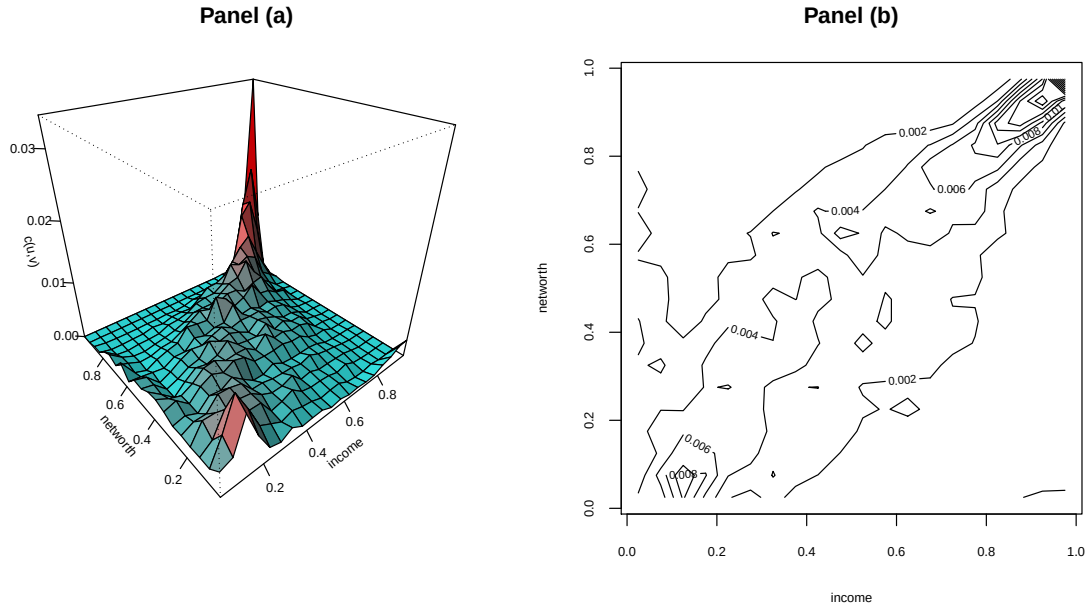

Figure 11: 2010 SCF

C.1 Quantile Scatterplots

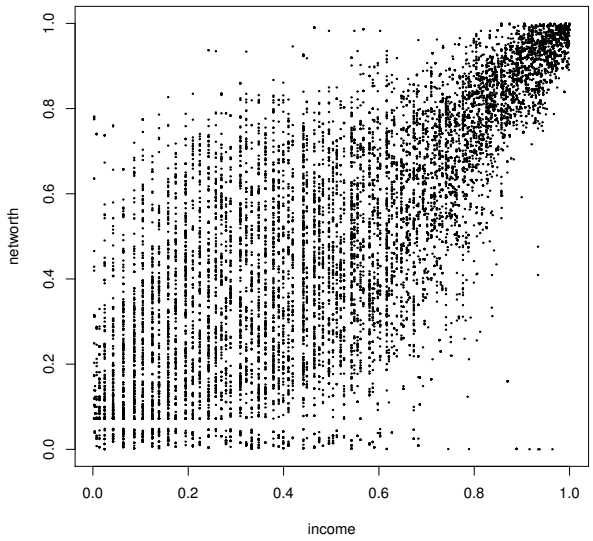

Figure 12: 1989 SCF

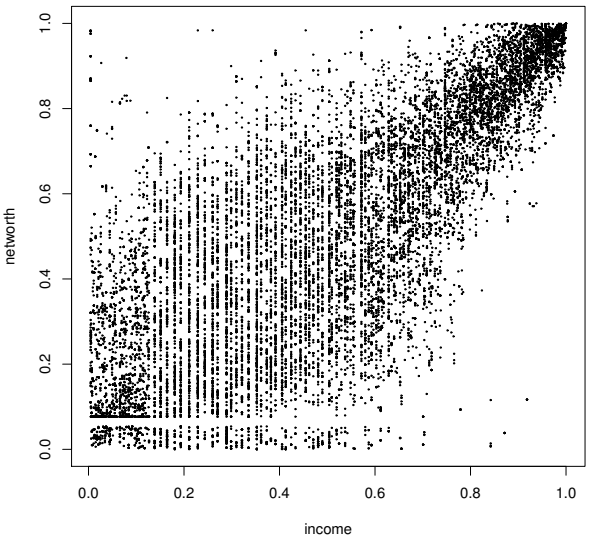

Figure 14: 1992 SCF

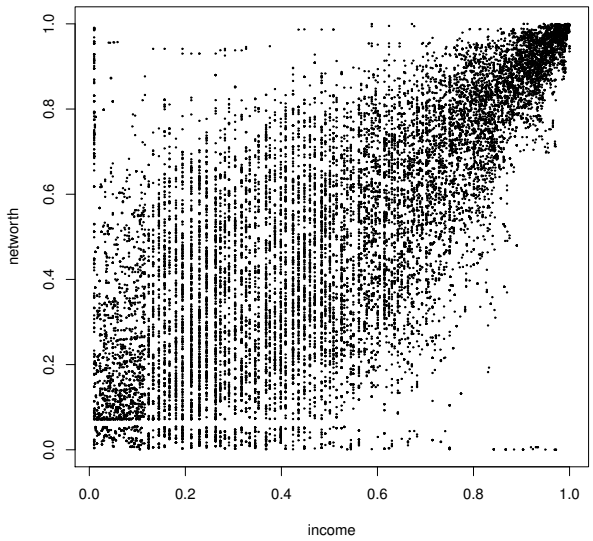

Figure 13: 1995 SCF

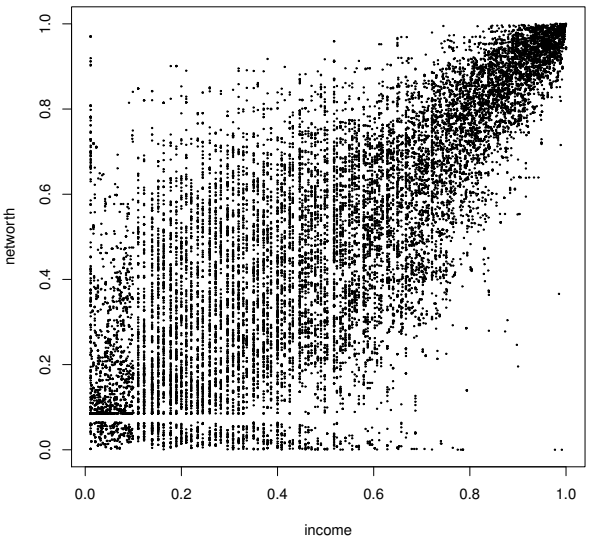

Figure 15: 1998 SCF

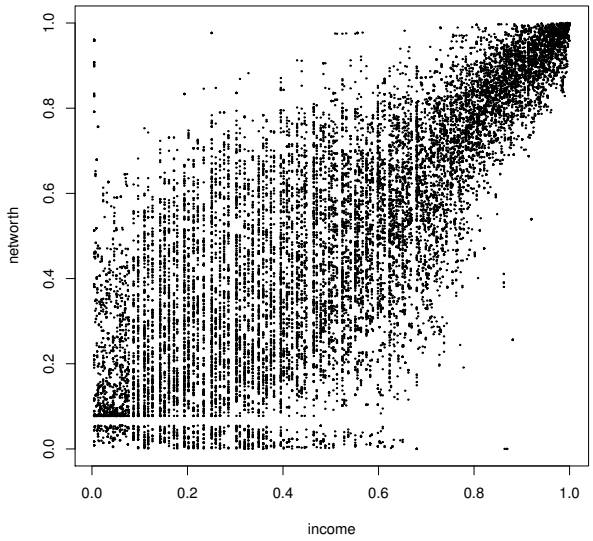

Figure 16: 2001 SCF

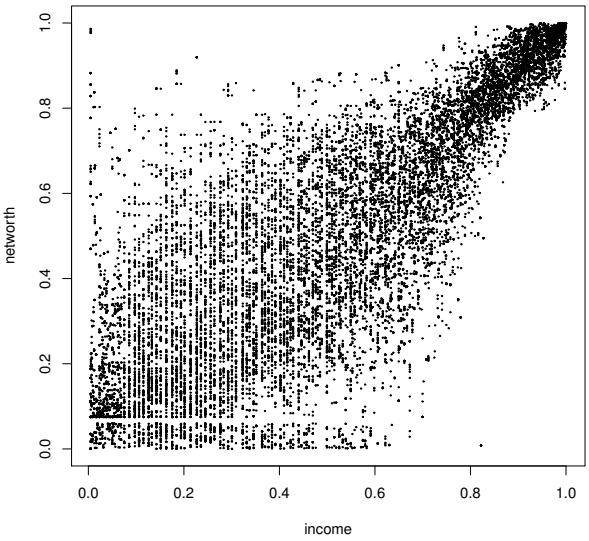

Figure 18: 2004 SCF

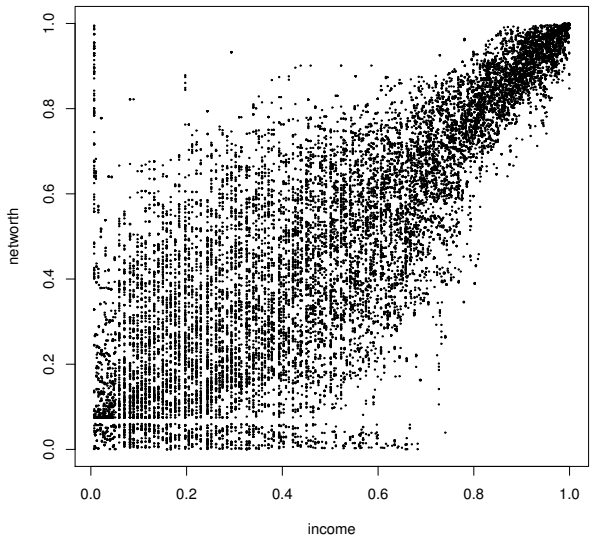

Figure 17: 2007 SCF

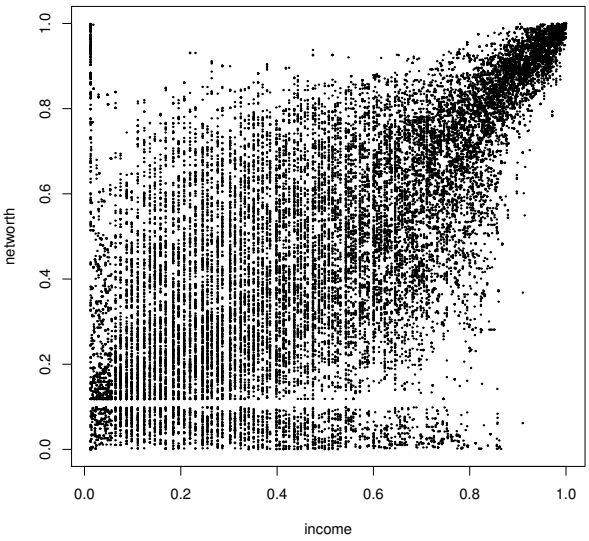

Figure 19: 2010 SCF

D Tail Dependence Estimation

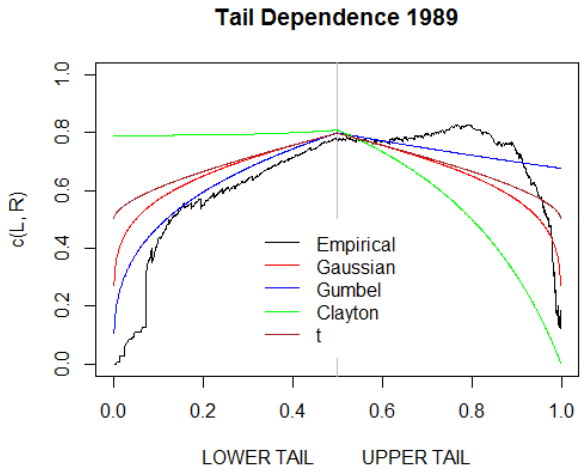

Figure 20: 1989 SCF

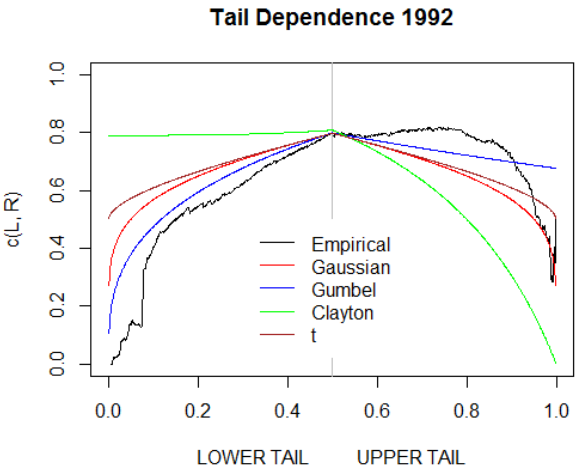

Figure 22: 1992 SCF

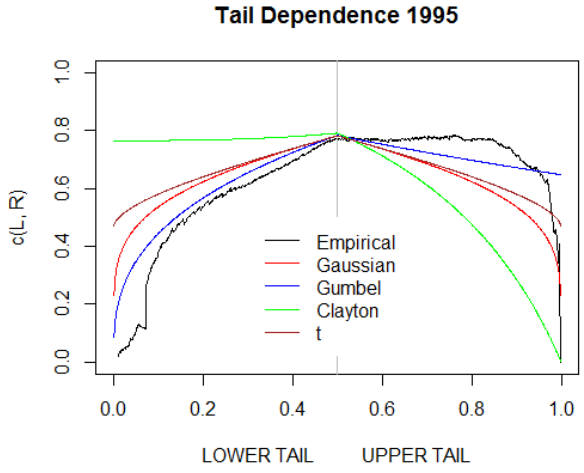

Figure 21: 1995 SCF

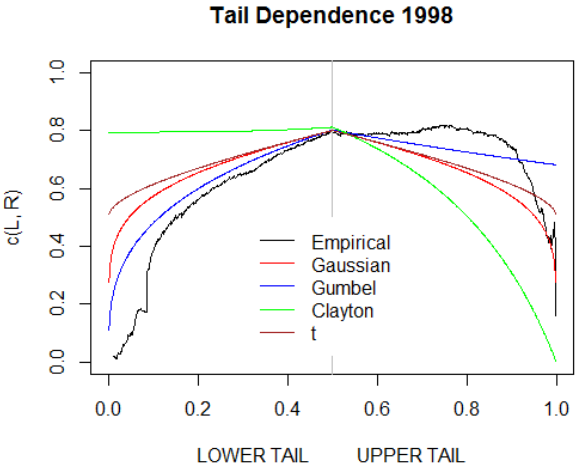

Figure 23: 1998 SCF

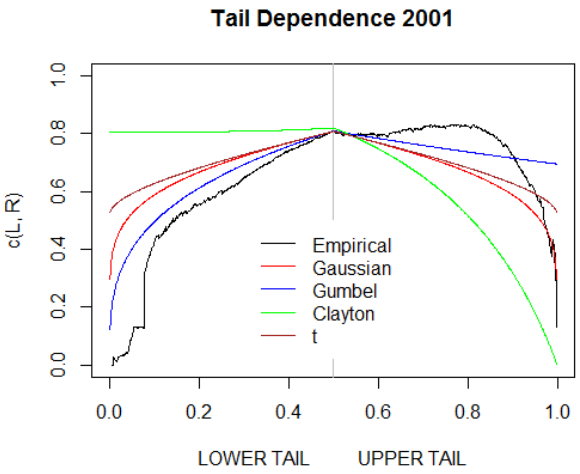

Figure 24: 2001 SCF

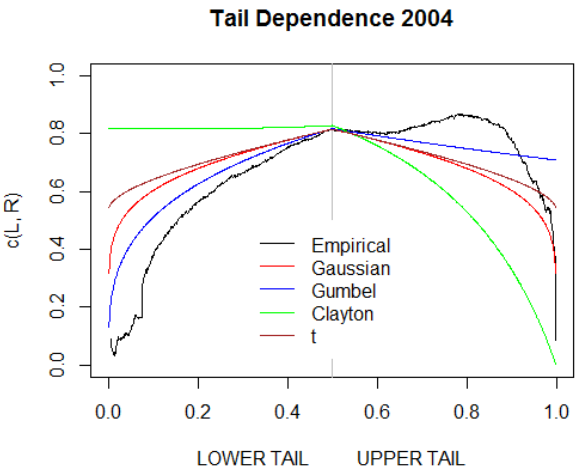

Figure 26: 2004 SCF

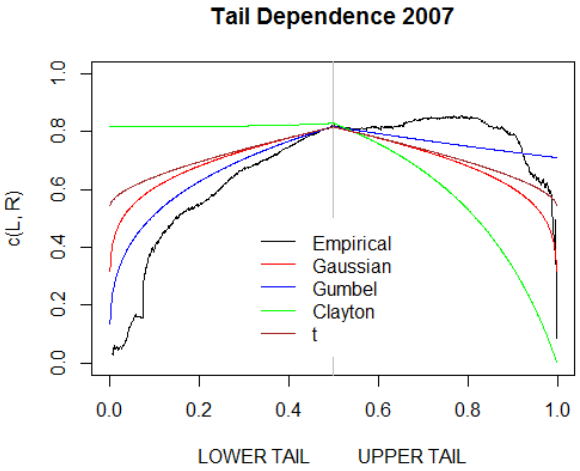

Figure 25: 2007 SCF

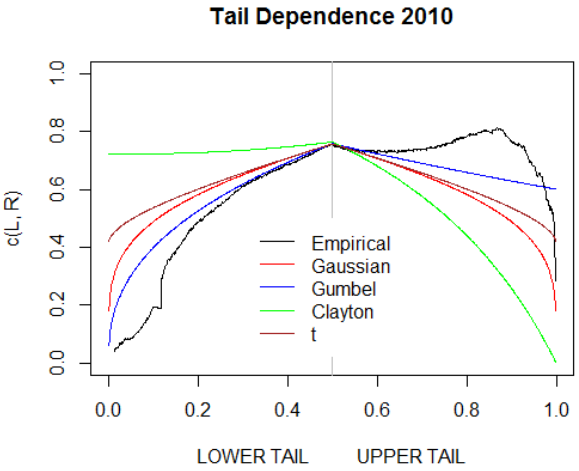

Figure 27: 2010 SCF

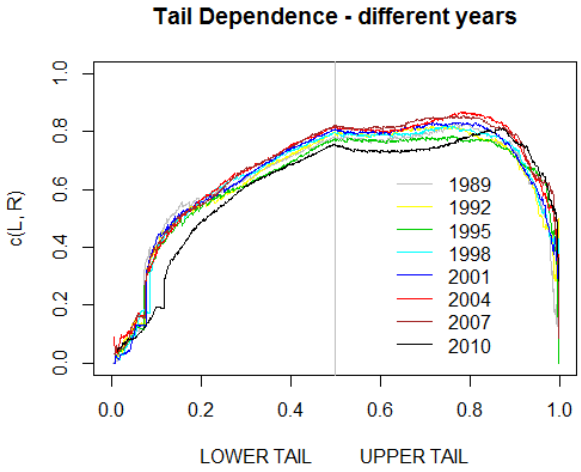

Figure 28: All years

E Goodness-of-Fit Test Statistics

	1989	1992	1995	1998	2001	2004	2007	2010
Test Statistic	4.515	7.232	7.208	7.133	7.889	9.193	10.159	16.465
Parameter	0.764	0.751	0.718	0.759	0.773	0.786	0.777	0.648
p-value	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
n	15,715	19,530	21,495	21,525	22,210	22,595	22,085	32,410

Table 3: Goodness-of-Fit test using Genest & Rémillard (2008) method with null of a theoretical Gaussian copula. Number of bootstrap iterations, $R = 100$.

	1989	1992	1995	1998	2001	2004	2007	2010
Test Statistic	4.050	5.635	4.930	5.497	6.467	7.014	6.747	9.824
Parameter	0.772	0.777	0.760	0.784	0.792	0.813	0.820	0.734
p-value	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046
n	15,715	19,530	21,495	21,525	22,210	22,595	22,085	32,410

Table 4: Goodness-of-Fit test using Genest & Rémillard (2008) method with null of a theoretical Student's-t copula. Number of bootstrap iterations, $R = 10$.

	1989	1992	1995	1998	2001	2004	2007	2010
Test Statistic	33.890	46.230	44.516	50.321	55.434	61.467	60.666	72.581
Parameter	1.212	1.176	1.092	1.192	1.228	1.272	1.285	0.862
p-value	0.046	0.046	0.046	0.046	0.046	0.046	0.046	0.046
n	15,715	19,530	21,495	21,525	22,210	22,595	22,085	32,410

Table 5: Goodness-of-Fit test using Genest & Rémillard (2008) method with null of a theoretical Clayton copula. Number of bootstrap iterations, $R = 10$.

	1989	1992	1995	1998	2001	2004	2007	2010
Test Statistic	1.419	2.029	1.443	1.770	2.227	2.407	2.392	3.595
Parameter	2.361	2.372	2.274	2.423	2.485	2.647	2.674	2.143
p-value	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
n	15,715	19,530	21,495	21,525	22,210	22,595	22,085	32,410

Table 6: Goodness-of-Fit test using Genest & Rémillard (2008) method with null of a theoretical Gumbel copula. Number of bootstrap iterations, $R = 100$.