Factors hiding in the tails: Bias in cross-sectional tail estimates

Milian Bachem

Erasmus University Rotterdam

Lerby M. Ergun*

Bank of Canada
London School of Economics

Casper G. de Vries

Erasmus University Rotterdam

Monday 29th June, 2020

Draft

Abstract

Measurement of the scaling behavior is often extracted from cross-sectional realizations. We show that due to the non-location invariant nature of the estimator that common time-series variation, caused for instance by a linear factor structure, introduces a bias. The characterization of the bias for the Hill estimator shows that it moves in the opposite direction for the left and right tail index estimate and it diminishes as fewer tail observations are used in the estimation. We use two test cases, financial returns and county population data, to show this is the case in real world data.

*Corresponding author: lerbyergun@gmail.com, 234 Wellington Street West, Ottawa, ON, K1A 0G9, Canada. We would like to thank Bruno Feunou, Andreas Utheman, Jun Yang and Chen Zhou for the helpful comment. I also thank the seminar participants at the Bank of Canada.

1 Introduction

A wide variety of natural processes are best characterized as being heavy-tailed, i.e. the scaling behavior described by power laws. In economics, this type of scaling behavior is wide spread and found in wealth and income (Atkinson and Piketty, 2007), firm size (Axtell, 2001), executive compensation (Baker et al., 1988), productivity (Helpman et al., 2004), stock markets (Jansen and Vries, 1991), amongst others. these findings guide subsequent modeling choices (Gabaix and Landier, 2008; Gabaix et al., 2006; Reed, 2001).[1] The cross-sectional distribution is frequently used to estimate the tail index, which determines the specific scaling behavior. Recently, attention has been drawn towards using the cross-section and time-series to measure time variation in the shape of the tail (Kelly and Jiang, 2014; Quintos et al., 2001; Haan and Zhou, 2019). However, to date little is known about the statistical properties of these measurements over time and what drives the variation.

Assuming that the data generating process (DGP) is a linear factor model with heavy-tailed idiosyncratic shocks, Feller's (1971) convolution theorem implies that the dependent variable inherits the tail index of the heavy-tailed disturbance term.[2] We show that tail index estimates extracted from the cross-sectional observations, using Hill's estimator (1971), contain a bias. Time variation in the bias is driven by the (inconsequential) factors and coefficients of the factor model. Furthermore, we show that the bias goes in the opposite direction for the left and right tail-index estimates. They are ones mirror images. Additionally, the size of the bias is decreasing in the number of observations used to estimate the tail index. This leads to three predictions. First, variation in the estimated tail index is partially driven by the factors, as opposed to a change in the tail index of the idiosyncratic shock distribution. Second, the bias has the opposite sign for the left and right tail index estimates. Lastly, the effect of the factors diminishes as a smaller number of tail observations is used to estimate the tail index. While using fewer tail observations reduces the problem of the bias, it increases the variance of the estimator.

The source of the bias originates from the non-location invariant property of the Hill estimator, e.g. adding a constant to the data changes the estimate. Thus,

[1] An overview of theoretical and empirical power law research in economics, finance and a selection of work from other fields is given in Gabaix (2009).

[2] The underlying assumption here is that the factors are fixed from a cross sectional perspective. To relax this assumption, we need the assumption that the factors have a less heavy tail than the idiosyncratic shocks.

common time-series fluctuation, driven by the factors, causes fluctuations in the tail index estimates. A possible solution to attenuate this problem, is to run time-series regressions to estimate the unobserved disturbance terms. However, the bias may persist if disturbance term estimates are cross-sectionally correlated. Other tail index estimators, such as moment type estimators (Dekkers et al., 1989) and kernel type estimators (Csorgo et al., 1985), suffer from the same bias, due to their non-location invariant nature (Drees, 1998).[3]

To test the above empirical predictions we use monthly US stock returns and annual US Census county population data. Both datasets contain a wide cross-section and a long time-series dimension. The wide cross-section is vital to estimate the tail index accurately. The long time-series dimension helps to establish the effect of the bias caused by the linear factor structure.

There is a rich tradition in the asset pricing literature to find explanatory factors to explain stock returns (Lintner (1965); Sharpe (1964); Fama and French (1996); Carhart (1997); Stambaugh and Lubos (2003)). The combination of a strong factor structure, wide cross-section and long time-series dimension provides an ideal setting for a first test case. Assuming a linear five-factor model, we isolate the effect of variation in a single factor on the cross-sectional tail index estimate. We find that the individual contributions of the factors on the Hill estimate are highly correlated with the factors themselves. These isolated effects explain about 65% of the variation in the left tail index estimates and 56% of the right tail index estimates. Furthermore, we find that the bias causes a negative correlation between the left and right tail index estimates. Measuring the tail index deeper in the tail lowers these correlations. However, the tail indexes of the dependent variable on the one hand and the estimated idiosyncratic shocks on the other are more highly correlated. The change in threshold causes the R^2 of this relationship to increase from 18% to 34% in the right tail and 7% to 67% in the left tail. This implies that moving the threshold deeper into the tail, variation in the tail index of the idiosyncratic shocks becomes more dominant, i.e. the bias in the tail index diminishes. Thus as proposed, the bias caused by the factors can be reduced by measuring deeper into the tail.

The focus of the literature on the heavy tailed nature of geographical population (Gabaix (1999); Eeckhout (2004); Rozenfeld et al. (2011)) and our data requirements makes the county population growth an appropriate second test case to analyze the

[3]The Pickands III (1975) estimator is location invariant. However, due to cross sectional variation in the factor coefficients, cross-sectional dependencies persists to bias the tail index estimate.

impact of the bias. The convenient contrast with the financial returns data is that no consensus on the factor structure of population growth has been reached. Providing us with a test case with a weak factor structure. Chi and Ventura (2011) conduct a review of the existing literature and present a large number of factors that describe population growth to a varying degree. Given this lack of consensus and proliferation of variables, we summarize many of the proposed factors into five principal components.

We find that the bias persists in the population data. The isolated effect of the principal components on the Hill estimator induces the predicted direction in the correlations. Furthermore, we find that the left and right tail estimates are negatively correlated. From regression analysis we find that the principal components explain 10% and 9% of the variation in the Hill estimate on the left and right tail, respectively. Most of the variation is explained by the first principal component, which is the only variable which is significant. This shows that in data with a weak factor structure the bias is less prevalent.

Therefore, we advise caution when attributing variation in the measurement of cross-sectional tail-index estimate to variation in the tail index of the linear factor model. The variation may be driven by known or unknown factors which are inconsequential for the scaling behavior of the underlying model.

2 Theory

Feller's convolution theorem on additivity states that the scaling behavior of a sum of random variables exhibits the scaling behavior of the most heavy-tailed random variable(s) out of the sum. This implies that in a linear factor model where the idiosyncratic shock is the most heavy-tailed random variable, the dependent variable inherits the tail behavior of the idiosyncratic shock. Time variation in the tail behavior of the idiosyncratic shocks is hard to detect from time-series behavior of the dependent variable. Extracting the tail behavior from the cross-sectional realizations provide a potential solution. However, as we explore in this section, the factor structure induces a bias on these estimates.

Consider a linear-factor model with n factors g_i, $i = 1, ..., n$. At any point in time (omitting superfluous time indices) the dependent variable Y_j for cross-sectional en-

tity j is

$$Y_j = \sum_{i=1}^{n} \gamma_{ij} g_i + X_j,$$

where the X_j are the idiosyncratic shocks. At any point in time the factors g_i are *fixed*. We consider the $\gamma_{ij} g_i$ deterministic from a data point of view.[4] As a shortcut on notation, we write

$$\sum_{i=1}^{n} \gamma_{ij} g_i = h_j$$

so that at a specific point in time

$$Y_j = h_j + X_j.$$

For X_j we consider the class of distributions with regularly-varying tails, i.e.

$$\lim_{n \to \infty} \frac{G(-tx)}{G(-t)} = x^{-\alpha},$$

with $x > 0$ and $\alpha > 0$. In that case there exists a slowly varying function $L(x)$ such that we may write for large $x > 0$,

$$P(X \leq -x) \sim L(x) x^{-\alpha}.$$

By $x \sim y$ we mean that x is asymptotic to y. A function is slowly varying if,

$$\lim_{n \to \infty} \frac{L(-tx)}{L(-t)} = 1.$$

These distribution functions are characterized as being heavy-tailed with scaling behavior α, i.e. the tail index.

Initially we examine the implications for a single observation. We deal with a special case in which X_j is standard Pareto distributed and h_j is zero. Subsequently, we relax the assumption that X_j is standard Pareto distributed. Instead, we assume that it is heavy-tailed distributed. Then, we relax the assumption that h_j is zero. This introduces a bias in the estimator caused by a shift in location h_j. Next, we move to multiple observations in the cross section. In that section we allow the distribution of the X_j to differ over the cross-section. In such setting, we determine the marginal effect of a change in h_j for the left and right tail of the distribution.

[4] From the econometricians point of view, this may not be the case, they may not be known or contain an error when estimated. If the distribution of X_{jt} exhibits regular-varying tails and given the bounded distribution of γ_{ij}, we can use the conditioning argument by Breiman (1965) to get an asymptotic expression for the tail probabilities by replacing h_{jt} by the conditional expectation $E_{h_t}(h_{jt})$.

2.1 Hill estimator

The Hill (1975) estimator uses the K highest-order statistics above the threshold u to estimate the (inverse of the) tail index α. Let $Y_{(j)}$ denote the descending order statistics from a cross section with m observations, that is $Y_{(1)} \geq Y_{(2)} \geq ... \geq Y_{(K)} \geq u \geq ... \geq Y_{(m-1)} \geq Y_{(m)}$. Here u is the threshold, typically chosen as a rate depending on the sample size.[5] The Hill estimator only uses the higher-order statistics above the threshold u and uses the average of their logarithmic sum,

$$\frac{1}{\widehat{\alpha}} = \frac{1}{K} \sum_{i}^{K} \ln \left(\frac{Y_{(i)}}{u} \right).$$

This is the version of the Hill estimator considered in Goldie and Smith (1987). In case the sample comes from the standard Pareto distribution, the Hill estimator coincides with the maximum likelihood estimator. In this setting all observations can be used by setting $u = 1$. Given the estimator is unbiased in the pure Pareto case, $u = 1$ is optimal in the sense of lowest variance. In other cases, like the Student-t distribution, only the tail resembles the Pareto tail and u is chosen in the tail area to reduce bias.

2.2 Single observation

Consider a single observation drawn from a standard Pareto distribution

$$G(x) = 1 - x^{-\alpha} \tag{1}$$

on $[1, \infty)$. Given a choice for $u > 1$, suppose we record a zero if $Y_1 < u$ and otherwise record $\ln(Y_1/u)$. Therefore, we need the conditional expectation

$$\begin{aligned} E\left[\ln \frac{Y}{u} | Y > u \right] &= \frac{\alpha}{u^{-\alpha}} \int_u^{\infty} \left(\ln \frac{x}{u} \right) x^{-\alpha-1} dx \\ &= \frac{1}{\alpha}. \end{aligned} \tag{2}$$

[5] Note that K is random given the deterministic threshold u. The alternative setup replaces u by the k-th order statistic $Y_{(k)}$, in which case k is deterministic, but it comes with a random threshold level u. Goldie and Smith (1987) argue that: "In practical terms, there is little to choose between these two points of view". They show that if u is chosen in such a way that $u \sim \lambda n^{1/(\alpha+2\theta)}$, then as the sample size increases, u increases in such a way that the mse is minimized asymptotically, the number of observations exceeding u increases, however, the percentage of these observations declines towards zero.

This shows that the conditional expectation of the Hill estimator from a standard Pareto sample of just one observation is unbiased, even if we choose $u > 1$.

Goldie and Smith (1987) relax the Pareto assumption and provide an explicit expression for the bias based on the following so called Hall expansion (Hall and Welsh, 1985)

$$G(x) = 1 - Cx^{-\alpha}\left[1 + Dx^{-\theta} + o(x^{-\theta})\right]. \tag{3}$$

Here $\alpha > 0$, $C > 0$, $\theta > 0$ and D is a real number. Here C and D are the first and second-order scale parameters, where α and θ are the first and second-order shape parameters. The first-order term is equivalent to the scaled Pareto distribution and more generally a first-order approximation to the tail probability. However, the second-order term is also a power. The expansion applies to many well-known distribution functions such as the Student-t distributions and the stationary distribution of an ARCH process, see e.g. Sun and Vries (2018). In case the distribution adheres to this expansion, the bias in the Hill statistic is (see Appendix 6.1 for a short derivation)

$$E\left[\ln\frac{Y}{u}|Y > u\right] = \frac{1}{\alpha} - \frac{\theta}{\alpha(\alpha+\theta)}Du^{-\theta} + o(u^{-\theta}). \tag{4}$$

Now we ask how the estimator fares if we allow for a factor $h \neq 0$ contribution to the model. Suppose idiosyncratic noise term X follows the standard Pareto distribution, then

$$\begin{aligned}\Pr\{Y \leq s\} = \Pr\{X + h \leq s\} &= 1 - (s-h)^{-\alpha} \\ &\simeq 1 - s^{-\alpha}\left[1 + \alpha h s^{-1}\right].\end{aligned}$$

The last step uses a first-order Taylor approximation of $(1 - hs^{-1})^{-\alpha}$ around $hs^{-1} = 0$ for s sufficiently large (larger than h). The corresponding coefficients in (3) are $C = 1$, $D = \alpha h$ and $\theta = 1$. Thus we get[6]

$$E\left[\ln\frac{Y}{u}|Y > u\right] \simeq \frac{1}{\alpha} - \frac{1}{\alpha+1}hu^{-1}.$$

The additional term, compared to (2), is the bias caused by a location shift. Given a fixed α over time, variation in the estimates come from variation in h. In this setting it is clear that variation in the estimates is caused by variation in the factors and coefficients as opposed to the scaling behavior of the linear factor model.

[6] Alternatively, one could directly use the density of the shifted Pareto distribution in the conditional expectation. The resulting integral is difficult to solve, but, differentiating with respect to h by using Leibniz's rule, the marginal effect of h is equivalent to this result.

2.3 Cross-section

Following the bias based on a single observation, we ask how the Hill statistic fares if we have multiple observations in the cross section. Suppose that the tail indices α and θ are equal, but the scale parameters C and D differ in the cross section.[7] Take

$$G_j(x) = 1 - C_j x^{-\alpha}\left[1 + D_j x^{-\theta} + o(1)\right]. \tag{5}$$

Calculating the conditional expectation per element gives

$$E\left[\frac{1}{\widehat{\alpha}}\middle| K = k\right] \simeq \frac{1}{\alpha} - \frac{\theta}{\alpha(\alpha+\theta)}\left(\frac{1}{k}\sum_{j=1}^{k} D_j\right)u^{-\theta}. \tag{6}$$

The only difference is that the bias term now contains the average of the second-order scale coefficients. In particular, the first-order scale coefficients C_j do not play a role.

Now suppose, however, that each entity j is influenced by an individual (non-stochastic from the point of view of the cross section) factor h_j. Assuming the distribution of the X_j conform to the first-order approximation, as in (1), the bias is as in (6) with $D_j = \alpha h_j$ and $\theta = 1$. Under a second-order expansion for the distribution of X_j, we arrive at

$$G_j(x) = 1 - C_j(x - h_j)^{-\alpha} - C_j D_j(x - h_j)^{-\alpha-\theta} + o(1).$$

By means of the Taylor approximation, we get

$$\begin{aligned} G_j(x) &= 1 - C_j x^{-\alpha}\left(1 - h_j x^{-1}\right)^{-\alpha} - C_j D_j(x - h_j)^{-\alpha-\theta} + o(1) \\ &\simeq 1 - C_j x^{-\alpha} - \alpha C_j h_j x^{-\alpha-1} - C_j D_j x^{-\alpha-\theta} - (\alpha+\theta) C_j D_j h_j x^{-\alpha-\theta-1}. \end{aligned}$$

The question now is which term is the second-order term? This depends on the value of θ. If $\theta < 1$, then the same first and second-order terms figure as before. But if $\theta = 1$ and $\theta > 1$, then the new second-order term is $(\alpha h_j + D_j) x^{-\alpha-1}$ and $\alpha h_j x^{-\alpha-1}$, respectively. We have the following intermediate result:

[7] In Appendix 6.2, we also relax the assumption that the powers are the same. For large samples, X_j with the lowest α_j dominates. For smaller samples, we show that the idiosyncratic shocks with less heavy tails bias the estimates. The cross-sectional tail estimate is then a weighted average of the tail indexes of the cross-section.

Lemma 2. If $\theta < 1$, the shift factor h_j exerts no influence on the bias of the Hill estimator, but if $\theta \geq 1$, the conditional expectation

$$E\left[\ln \frac{Y_j}{u} | Y_j > u\right]$$

changes from

$$\frac{1}{\alpha} - \frac{\theta}{\alpha\left(\alpha+\theta\right)} D_j u^{-\theta}$$

to

$$\begin{cases} \frac{1}{\alpha} - \frac{1}{\alpha(\alpha+1)}\left(D_j + \alpha h_j\right) u^{-1} & \text{if } \theta = 1 \\ \frac{1}{\alpha} - \frac{1}{\alpha+1} h_j u^{-1} & \text{if } \theta > 1. \end{cases}$$

Denoting the average of the shift factors by

$$\overline{h} = \frac{1}{k}\sum_{j=1}^{k} h_j,$$

we have the following proposition:

Proposition 1. For the upper tail of the cross-sectional distribution, if $\theta \geq 1$, the Hill statistic **declines** if $\overline{h}$ increases, since

$$\partial E\left[\frac{1}{k}\sum_{j=1}^{k} \ln \frac{Y_j}{u} \middle| Y_j > u\right] / \partial \overline{h} \approx -\frac{1}{\alpha+1} u^{-1} < 0.$$

Thus if a single factor increases, this affects all Y_j with a positive coefficient in such a way that the downward bias in the Hill statistic becomes more severe. Note that one has to sum over all h_j. So in a one factor model setup $h_j = \gamma_j g$, we get

$$\partial E\left[\frac{1}{k}\sum_{j=1}^{k} \ln \frac{Y_j}{u} \middle| Y_j > u\right] / \partial g \approx -\frac{\overline{\gamma}}{\alpha+1} u^{-1}$$

and where $\overline{\gamma}$ is the average of the k number of γ_j coefficients.

A similar proposition applies for the lower tail:

Proposition 2. For the lower tail of the cross-sectional distribution, if $\theta \geq 1$, the Hill statistic **increases** if $\overline{h}$ increases, since

$$\partial E\left[\frac{1}{k}\sum_{j=1}^{k} \ln \frac{Y_j}{-u} \middle| Y_j \leq -u\right] / \partial \overline{h} \approx \frac{1}{\alpha+1} u^{-1} > 0.$$

While the bias in the upper tail is negative, the bias in the lower tail is positive. This implies that the correlation between the factor and lower tail index estimate is negative.

Note that the two derivatives in Proposition 1 and 2 are mirror images. This implies that the bias generates a negative correlation between the left and right tail-index estimates. At first this result may seem counter-intuitive. However, the location shift, in combination with the transformation of the data for the left tail-index estimate gives the result intuitive support. A downward location shift lowers the order statistics in the right tail. While the same is true for the left tail observations, the Hill estimator necessitates multiplying the observations by -1. This transformation alters the initial downwards shift in the left tail data into an upward shift in the transformed lower tail data. This produces the sign difference in the bias for the left and right tail estimate.

Given the coefficients, movement in the factors induces changes in the cross-sectional Hill estimate, as discussed in Proposition 1 and 2. Even though the coefficients are constant over time, the inclusion of the k entities in the Hill estimate is random. The variation in the k number of included γ_{ij} causes additional bias in excess of the movement in the factors. One can estimate γ_{ij} and remove the effect of factor i in the Hill estimate. Take

$$Y_{jt} = \sum_{i=1}^{n} \gamma_{ij} g_{it} + X_{jt}.$$

One can define;

$$S_{jt}^{(f)} = Y_{jt} - \sum_{i \neq f} \hat{\gamma}_{ij} g_{it} \tag{7}$$

Here the factor g_f is the remaining factor. Henceforth, we refer to $S_{jt}^{(f)}$ as the semi-residuals. Working from a first-order approximation for the distribution of X_j this implies,

$$E\left[\frac{1}{\widehat{\alpha}_t^{S^{(f)}}} \middle| S_{(jt)}^{(f)} > u\right] = E\left[\frac{1}{k}\sum_{j=1}^{k} \ln \frac{S_{(jt)}^{(f)}}{u} \middle| S_{(jt)}^{(f)} > u\right] \approx \frac{1}{\alpha} - \frac{1}{(1+\alpha)}\left[\frac{1}{k}\sum_{j=1}^{k}\left(h_{jt}^{*}\right) u^{-1}\right],$$

where $h_{jt}^{*} = \gamma_{fj} g_{ft}$. This isolates the influence of $\gamma_{fj} g_{ft}$ on the variation in $\widehat{\alpha}_t^{Y}$, especially for intermediate levels of u.

The above propositions predict several empirical regularities. First, we look at the

correlation between the factors, g_{it}, and $\widehat{\alpha}_t^Y$ for the left and right tail of the distribution. Given positive coefficients γ_{ij}, the most dominant factor should have a positive correlation with the right tail estimate and a negative correlation with the left tail estimate. Second, in order to isolate the effect of the individual factors on the tail index we take the difference,

$$E\left[\frac{1}{\widehat{\alpha}_t^{S^{(f)}}} - \frac{1}{\widehat{\alpha}_t^X}\middle| S_{jt}^{(f)} > u\right] = E\left[\frac{1}{\widehat{\alpha}_t^{(f)}}\middle| S_{jt}^{(f)} > u\right]$$
$$\approx -\frac{1}{(1+\alpha)}\left[\frac{1}{k}\sum_{j=1}^{k} h_{jt}^* u^{-1}\right]. \tag{8}$$

Here $\widehat{\alpha}_t^X$ is the estimated cross-sectional tail index on $\widehat{X}_{jt}$. Third, the relationship depends on the $\gamma_{ij} > 0$. The size of the bias depends on the value of the γ_{ij}'s included in the estimation of $\widehat{\alpha}_t^{S^{(f)}}$. Therefore, we also multiply g_{ft} by the average of the coefficient estimate with $S_{(jt)} \geq u$, i.e.

$$W_t^{(f)} = {}^1\!/\!\left(\textstyle\sum_{j=1}^m I\{S_{jt}^{(f)} > u\}\right) \displaystyle\sum_{j=1}^{m} \widehat{\gamma}_{fj} I\{S_{jt}^{(f)} > u\}. \tag{9}$$

Fourth, the effect of the factors on the tail index should be in the opposite direction for the left and right tail index estimate. This predicts a negative correlation between $\widehat{\alpha}_t^{S^{(f)}} - \widehat{\alpha}_t^{X_t}$ for the left and right tail-index estimates. Finally, *ceteris paribus* the threshold u should diminish the effect of the factors on $\widehat{\alpha}_t^Y$. Therefore, we estimate correlations at two different thresholds, one at 5% and the other at 0.5% of the sample fraction.

3 Data

To test the above empirical predictions, we use financial return and geographical population concentration data. The data requirements for our analysis and the ample empirical evidence of their heavy tailed nature allows us analyze the effect of the bias in real world data.

3.1 Firm stock returns

The Center for Research in Security Prices (CRSP) provides a wide cross-section of firm return data for the US equity market with 13,535 individual US traded firms.

These daily data are collected from the NYSE, AMEX, NASDAQ and NYSE Arca exchanges since 1925. In accordance with the financial literature on asset pricing, we use monthly stock returns from 1963 to 2019.[8]

There is a large body of literature which uses factors to explain the cross-sectional variation in expected excess stock returns, excess over the risk-free rate. The combination of the rich dimensions of the data and the theoretical and empirical backing for a factor structure in stock returns should provide a good test case to verify factor bias in tail index estimates. In line with existing literature, we use the Fama and French (1996) three-factor model augmented by the Momentum (MOM) factor from Carhart (1997) and the Liquidity factor from Stambaugh and Lubos (2003). In Table 10, the analysis is repeated using the Fama and French (2015) five-factor model. In their model the momentum and liquidity factor are substituted by the Robust-Minus-Weak (RMW) and Conservative-Minus-Aggressive (CMA) factor.[9]

3.2 County population data

The other heavily researched field in power laws is the geographical distribution of population. The US Census Bureau has collected since 1970 county population statistics. Analysis on the county level grants the most consistent cross-sectional classification over time. The annual county population data provided from the Census is from 1970 to 2017. In contrast to the 648 time-series dimension for the US stock data, this data contains 46 time-series observations. Moreover, the US Census is only conducted every 10 years. Annual data is estimated using births, deaths and net migration, including net immigration from abroad. In every census after 2000, the county populations for each year of the census are updated yearly, leading to inconsistent comparison between the last year of the previous census and the first of the current census. Consequently, we omit the years 2000 and 2010 from our data. We conduct our analysis on the growth rate of the population in line with existing literature. For the creation of population change, we use the Federal Information Processing Standards (FIPS) code which uniquely identifies counties and county equivalents in the United States.

[8]Due to stale prices and liquidity issues, it is common practice to excluded stocks with a price below 5 dollars. Including stocks with an average price below 5 dollars yields almost identical results. Stocks with exchange code -2, -1 or 0 are not included in the analysis. In addition, only common stocks with share code 10 and 11 are included in the analysis.

[9]We obtain the five Fama and French (1996) factors and the momentum factor from the data library of Kenneth R. French and the Liquidity factor from the website of Lubos Pastor.

The documentation on a clear factor structure is notably weaker than our first empirical test case. Chi and Ventura (2011) conduct a review of the existing literature and propose variables that can broadly be placed in one of five categories: demographic characteristics, socio-economic conditions, transportation accessibility, natural amenities, and land development. So far, the models used to explain population growth have varying degrees of success and significance. As there is no consensus in the literature on which exact model is the most valid, we conduct a Principal Component Analysis (PCA) and extract the first five principal components. The factors used as inputs for the PCA are suggested in Chi and Ventura (2011), which we describe in Table 13 in the Appendix. Using principal components (PCs), we avoid multicollinearity, which is likely to arise in a model with many explanatory variables. Furthermore, using five PCs avoids issues with overfitting.

3.3 Empirical implementation

To obtain $S_{jt}^{(f)}$ in (7) for different factors, we run linear regressions to estimate γ_{ij}. Take $Y_{jt} = R_{jt} - R_t^f$, where R_{jt} is the return of stock j at time t. Here R_t^f is the risk-free rate at time t. We run regression:

$$Y_{jt} = \sum_{i \neq f} \gamma_{ij} g_{it} + S_{jt}^{(f)} \quad \text{for} \quad t = 1, 2, ..., T.$$

We repeat this for all m entities. Stacking all the $S_j^{(f)}$'s, we obtain matrix $S^{(f)}$. Given matrix $S^{(f)}$, we estimate

$$\frac{1}{\hat{\alpha}_t^{S^{(f)}}} = \frac{1}{k} \sum_{j=1}^{k} \ln \left(\frac{S_{(jt)}^{(f)}}{S_{(kt)}^{(f)}} \right)$$

at time t. The $\hat{\alpha}_t^{S^{(f)}}$ isolate the influence of $\gamma_{ij} g_{ft}$ and X_{jt} on the Hill estimate.

The Hill estimator needs a choice of threshold, $u(k)$. Ordinarily, u is set to the k^{th} order-statistic, i.e. $S_{(kt)}^{(f)}$. We take various choices of k to study the influence of k on the empirical results. We choose $S_{(kt)}^{(f)}$ at the 5% and 0.5% empirical quantile.

4 Results

4.1 US financial returns

Feller's convolution theorem on additivity suggests that one could extract the variation in scaling behavior from the cross-sectional realization of the dependent variable Y_{it}. In turn, Section 2 demonstrates that these measurement contain a bias caused by underlying factors and coefficients. For the stock return data, Table 1 shows the correlation between the asset pricing factors and α^Y_{t-}. The "+" ("-") subscript for the Hill estimate indicates that the estimate is on the right (left) tail of the empirical distribution. The correlation for the market factor is particularly strong, showing a first sign of the strong role it plays in the cross-section of stock returns and consequently the tail index estimates. Although smaller, the correlation for the SMB factor is still relatively strong. The correlation of the other factors is smaller still but more importantly point in the wrong direction. This is partly caused by the simultaneous effect that the different factors have on α^Y_{t-}. Furthermore, for a given stock the coefficients for the different factors may vary in size and have different signs. These issues may dilute the effect of the bias caused by a single factor. To

Table 1: Cross-sectional tail index

	Market	SMB	HML	MOM	Liq
$\rho(\alpha^Y_{t-}, g_t)$	-0.76	-0.50	0.20	0.11	0.04
$\rho(\alpha^Y_{t+}, g_t)$	0.70	0.45	-0.10	-0.10	-0.03

This table presents the correlation between cross-sectional Hill estimates and the asset pricing factors. Here α^Y_{t-} and α^Y_{t+} are the cross-sectional Hill (1975) estimates for the cross-section of stock returns for the left and right tail, respectively. Here G_t represent the five factors are the market, small-minus-big, high-minus-low, momentum and the liquidity factor reported in the columns. The threshold u is set to 5% of the sample fraction.

isolate the bias that a single factor induces in $\hat{\alpha}^Y_t$, we use expression $\hat{\alpha}^{(f)}_{t+}$ in (8). Panel (a) in Figure 1 displays this effect graphically. The time-series of the market factor (solid line) and $\hat{\alpha}^{(M)}_{t+}$ (dotted line) present a clear positive relationship. Given the already strong correlation between the market factor and $\hat{\alpha}^Y_t$, the use of expression (8) further confirms Proposition 1. In panel (b), we contrast $\alpha^{(f)}_{t+}$ and $\alpha^{(f)}_{t-}$, i.e. the right and left tail estimates, for the market factor. The $\hat{\alpha}^{(M)}_{t-}$ series (dashed line) shows a clear negative relationship with the right tail estimates. This is in line with the prediction of Proposition 2. Figure 2, in the Appendix, shows the same figures for the SMB, HML, momentum and the liquidity factors. The relationship for the SMB and HML factors is not as clear as that of the market factor. The momentum and liquidity factors show the weakest relationship. The relationship of the

tail index and the factors hinges on the validity of the factor structure and the correct specification of the factors. A number of these constructed factors are possibly poor proxies for the factors in the DGP, leading to some of the weaker relationships.[10]

The first and second row in Table 2 show that, by isolating the effect of the factors on $\widehat{\alpha}_{t-}^{Y}$, the correlations improve for most factors. The correlations for the market and SMB factor have increased. Isolating the effect of the HML factor changes the correlation in the predicted direction for both the left and right tail of the distribution. The correlations for the momentum factor, in the first and second row, are not in the expected direction. The observations included in the tail measurement could have negative coefficients, leading to the unexpected signs. In rows three and four, we control for the effect of γ_{ij} in the bias of $\hat{\alpha}_{t-}^{Y}$ with $W_t^{(f)}$ in (9). Applying this correction changes the direction of the correlation for the momentum factor. The correlations in the first two rows for the liquidity factor are in the expected direction, but very small. Correcting for the coefficients, turns the sign of the correlation for the right tail in the wrong direction. However, this correlation is weak.

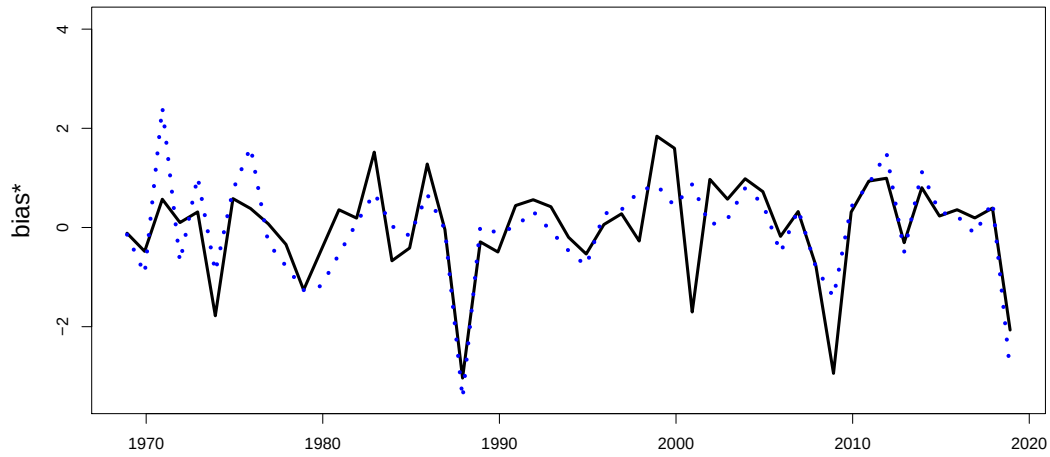
The implication of Proposition 1 and 2 being a mirror image of one another is that the correlation between $\widehat{\alpha}_{t-}^{(f)}$ and $\widehat{\alpha}_{t+}^{(f)}$ should be negative. The negative sign in the last row of Table 2 illustrates that the effect of variation in the factors on the left and right tail-index estimates is in the opposite direction. This is apparent in Figure 1 for the market factor. The market and SMB factors have the strongest effect on the cross-sectional estimate and also the strongest negative correlation between their respective left and right tail estimates. This might be attributed the quality of these factors as proxies for the underlying factors in the DGP.

Equation (8) shows that the bias in $\widehat{\alpha}_{t-}^{Y}$ originating from the factors should diminish as threshold u increases. In panel (b) of Table 2, we lower the threshold to 0.5% of the sample fraction. This lowers the correlations between the factors and the tail-index estimates for most factors. Thus, at this lower threshold, less of the variation in the tail index is driven by the factors and coefficients.

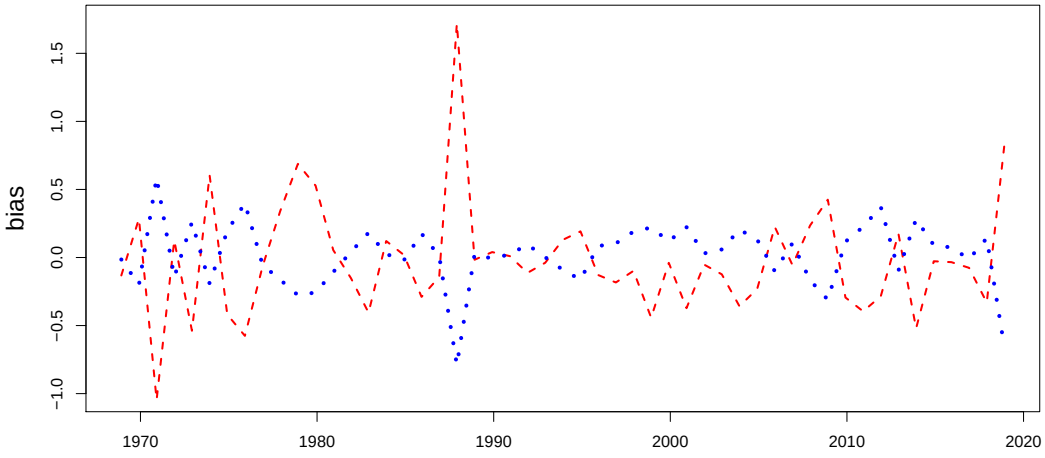
Panel (a) of Table 3 shows how much of the variation in $\widehat{\alpha}_{t-}^{Y}$ can be explained by variation in the factors and coefficients. The R^2 of the first regression shows that about 44% of the variation in the cross-sectional tail index is driven by the market factor.[11]

[10] A possible cause of the weaker relationship is time varying explanatory power of asset pricing factors. This is further discussed in Hwang and Rubesam (2015).

[11] See Table 9 in the Appendix for regression results for factors (instead of the $\hat{\alpha}_t^{(f)}$).



(a) M_t vs $\alpha_+^{(M)}$



(b) $\alpha_+^{(M)}$ vs $\alpha_-^{(M)}$

Figure 1: Market factor effect on cross-sectional Hill estimator

This figure displays the time series of the market factor and its isolated effect on the cross-sectional Hill estimator, $\alpha^{(M)}$ defined in (8). In panel (a) the market factor (solid black line) is contrasted with $\alpha_+^{(M)}$ (blue dotted line). In panel (a) the time series of annual observations are standardized by their respective mean and standard deviation. In panel (b) the right tail estimate $\alpha_+^{(M)}$ is contrasted with left tail estimate $\alpha_-^{(M)}$ (red dashed line).

The second most important factor is the SMB factor which contributes about 12% to the variation in $\widehat{\alpha}_{t-}^{Y}$. The HML, momentum and liquidity factors have a marginal role in explaining the variation in $\widehat{\alpha}_{t-}^{Y}$. The explanatory power of the idiosyncratic part of the linear factor model explains about 19% of the variation. This explanatory power can originate from three sources. First, the model we use could be misspec-

Table 2: Correlations cross-sectional tail index

	Market	SMB	HML	MOM	Liq
$\rho(\alpha^{(f)}_{t-}, g_t)$	-0.85	-0.82	-0.12	0.38	-0.03
$\rho(\alpha^{(f)}_{t+}, g_t)$	0.84	0.83	0.37	-0.51	0.02
$\rho(\alpha^{(f)}_{t-}, W^{(f)}_{t-} * g_t)$	-0.84	-0.84	-0.30	-0.37	-0.18
$\rho(\alpha^{(f)}_{t+}, W^{(f)}_{t+} * g_t)$	0.81	0.74	0.11	0.22	-0.14
$\rho(\alpha^{(f)}_{t-}, \alpha^{(f)}_{t+})$	-0.92	-0.80	-0.09	-0.21	-0.06

(a) Threshold u at **5%** of sample fraction

	Market	SMB	HML	MOM	Liq
$\rho(\alpha^{(f)}_{t-}, g_t)$	-0.62	-0.41	0.01	0.09	-0.04
$\rho(\alpha^{(f)}_{t+}, g_t)$	0.52	0.34	-0.04	0.01	0.02
$\rho(\alpha^{(f)}_{t-}, W^{(f)}_{t-} * g_t)$	-0.62	-0.40	-0.17	-0.16	0.06
$\rho(\alpha^{(f)}_{t+}, W^{(f)}_{t+} * g_t)$	0.47	0.25	0.05	-0.08	-0.05
$\rho(\alpha^{(f)}_{t-}, \alpha^{(f)}_{t+})$	-0.47	-0.24	-0.01	-0.03	-0.05

(b) Threshold u at **0.5%** of sample fraction

This table presents the correlation between various specifications of the cross-sectional Hill estimates. Here $\alpha^{(f)}_{t-}$, stated in the **first** and **second** rows of each panel, is the cross sectional tail index where the factor f's effect is isolated, as defined in (8). Here g_t represents the five factors, the market, small-minus-big, high-minus-low, momentum and the liquidity factor reported in the columns. In the **third** and **fourth** row, the factor realization at time t is multiplied by the average factor loading, as measured by γ_{ij}, of the k stocks included in the estimation of $\alpha^{S^{(f)}}$, i.e $W^{(f)}_t$ in (9). The **last** row shows the correlation between the left and right tail estimates of $\alpha^{(f)}_t$. The upper panel presents the correlations where the threshold u is set to 5% of the sample fraction and for the lower panel this threshold is set to 0.5% of the sample fraction.

ified and therefore the variation partially comes from missing factors. Second, the measurement error for $\hat{\alpha}^Y_t$ and $\hat{\alpha}^X_t$ are correlated. Lastly, the variation comes from the variation in scaling behavior of the disturbance term.

Table 7 in the Appendix presents the regression results for a threshold based on the 0.5% sample fraction. For the lower threshold, we find that only a small share of the variation in $\widehat{\alpha}^Y_t$ is explained by the factors. Now variation in $\widehat{\alpha}^X_t$ explains about 35% for the left and 69% for the right tail of the variation in $\widehat{\alpha}^Y_t$. As previously discussion, variation in $\widehat{\alpha}^X_t$ can come from measurement error, common time-series variation caused by unknown factors and/or from the true α_t of the DGP. Abstracting away from correlated measurement errors, the increase in R^2 strongly suggests that the role of known and unknown factors in the bias diminishes. Therefore, capturing variation α_t more strongly.

Table 3: Regression cross-sectional tail index

$\alpha_{t-}^{(M)}$	1.16***						1.31***
	(0.05)						(0.05)
$\alpha_{t-}^{(SMB)}$		0.98***					1.27***
		(0.11)					(0.07)
$\alpha_{t-}^{(HML)}$			1.05***				−0.48***
			(0.28)				(0.18)
$\alpha_{t-}^{(MOM)}$				0.19			0.28
				(0.33)			(0.20)
$\alpha_{t-}^{(Liq)}$					0.67*		−0.40
					(0.40)		(0.25)
α_{t-}^{X}						0.96***	
						(0.08)	
R^2	0.44	0.12	0.02	0.001	0.005	0.19	0.64

(a) Left cross-sectional tail index, i.e. $\hat{\alpha}_{t-}^{Y}$

$\alpha_{t+}^{(M)}$	0.93***						1.02***
	(0.04)						(0.04)
$\alpha_{t+}^{(SMB)}$		0.77***					0.95***
		(0.10)					(0.07)
$\alpha_{t+}^{(HML)}$			0.30				−0.66***
			(0.29)				(0.20)
$\alpha_{t+}^{(MOM)}$				−0.40			0.42*
				(0.31)			(0.22)
$\alpha_{t+}^{(Liq)}$					0.23		−0.48*
					(0.37)		(0.27)
α_{t+}^{X}						0.41***	
						(0.06)	
R^2	0.41	0.09	0.002	0.003	0.001	0.06	0.56

(b) Right cross-sectional tail index, i.e. $\hat{\alpha}_{t+}^{Y}$

This table displays the regression results for the cross-sectional Hill estimate. Here the dependent variable in the upper panel is α_{t-}^{Y}, i.e. is the Hill (1975) estimate for the left tail of the raw cross-sectional returns. The independent variables $\widehat{\alpha}_{t}^{(f)}$, stated in the first column, is the cross sectional tail index where the factor f's effect is isolated, as defined in (8). Furthermore, α_{t-}^{X} is the tail index estimated on the estimated disturbance terms of the five-factor asset pricing model. The five factors are the market, small-minus-big, high-minus-low, momentum and the liquidity factor. The lower panel shows the results right tail of the distribution. The threshold u to estimate the Hill estimate is set to 5% of the sample fraction. The constant included in the regression is excluded from the presented results. * p<0.1; ** p<0.05; *** p<0.01

To show that possibly also location invariant estimators might suffer from a similar bias, we use the Pickands estimator. The Pickands III (1975) tail index estimators is an example of a location invariant estimator. The estimator is proportional to the ratio of the difference between two order statistics, namely

$$1/\hat{\alpha}_k = \frac{1}{\log 2} \log \frac{X_{(k)} - X_{(2k)}}{X_{(2k)} - X_{(4k)}}.$$

Due to variation of the factor coefficients in the cross-section $[X_{(j)} + h_{(j)}] - [X_{(2j)} + h_{(2j)}] \neq X_{(j)} - X_{(2j)}$. Even though this bias is smaller than for the non-location invariant estimators, it causes variation in the Pickands estimate due to the factor structure. Table 11 show that only the market factor significantly influences the estimates of tail index, $\hat{\alpha}_t^Y$. The previous analysis suggests that the market factor is the most dominant out of all the factors. Therefore, its effect through the cross sectional variation in coefficients is the most likely to still affect the Pickands estimates. Furthermore, there is no strong relationship between $\hat{\alpha}_t^Y$ and $\hat{\alpha}_t^X$. This is most likely because the Pickands estimator is weakly consistent. This causes estimates to be much more volatile than the Hill estimates.

4.2 County population data

Due to the lack of a clear emergent set of factors in the population size literature, we us PCA to extract 5 PCs from 39 suggested factors. The first five principal components explain about 60 % of the variation in our original variables. Table 12 in the Appendix presents the summary statistics of the PCA.

In a similar vein to the US stock data, Table 4 presents the correlations for the county population growth data. The correlations between $\hat{\alpha}^Y$ and the PCs is relative weak and occasionally in the wrong direction. However, the increase in correlations by applying $\hat{\alpha}^{(f)}$ in the third and fourth row, is illuminating. As the data on which the Hill estimate is calculated, is made independent of the respective principal components, correlations become far more clear. The correlation in the first PC increases from -0.43 to -0.83 for the left tail and from -0.16 to 0.84 for the right tail. Caution is advised, as the length of the time series do not permit a narrow delineation of the confidence interval around the estimates. Furthermore, since the factors are PC's, there is little interpretation that we can draw from the direction of the correlations when we ignore the effect of γ_{ij}. Multiplying the PCs by $W_t^{(f)}$ specified in (9), we observe that all correlations point in the direction suggested by Propositions 1 and 2, in rows five and six. Finally, the last row indicates that the Hill estimate on the left

and right side of the distribution are negatively correlated for all but the second PC.[12]

Panel (b) of Table 4 presents the correlations when the threshold is lowered to 0.5% of the data. Here, with the exception of the first principal component, correlations decrease in magnitude and now point in the wrong direction. In the final row we still observe negative correlations, but the magnitude decreases substantially for all but one principal component. Thus, as is observed in the US stock data, lowering the threshold limits the influence of the factor structure in the cross-sectional tail-index estimate.

The regression analysis in Table 5 shows that only the first PC can significantly account for variation in the left tail-index estimate. For the right tail, $\alpha_{t+}^{(PC1)}$ is not significant but attains the highest R^2 of the five principal components. This implies that the previously presented correlations for the PCs most likely come with large standard errors. The high R^2 attained by α_{t-}^{X} and α_{t+}^{X} further illustrates the marginal influence of the factor structure in the cross-sectional tail index estimate. Abstracting away from the weak inference due to the limiting time-series dimension in the county data, the contrast in results with the financial return data showcases the effect of the factors in the cross-sectional tail index estimates.

[12]Given that PC2 to PC5 are insignificant in the regression analysis, it is likely the case that the estimated correlations in Table 4 have large confidence intervals.

Table 4: Correlations cross-sectional tail index (County Data)

	PC1	PC2	PC3	PC4	PC5
$\rho(\alpha^Y_{t-}, g_t)$	-0.43	-0.07	-0.40	-0.09	0.14
$\rho(\alpha^Y_{t+}, g_t)$	-0.16	0.02	0.06	0.09	0.06
$\rho(\alpha^{(f)}_{t-}, g_t)$	-0.83	-0.20	-0.60	0.59	0.67
$\rho(\alpha^{(f)}_{t+}, g_t)$	0.84	0.66	0.65	-0.64	-0.67
$\rho(\alpha^{(f)}_{t-}, W^{(f)}_{t-} * g_t)$	-0.86	-0.39	-0.63	-0.60	-0.70
$\rho(\alpha^{(f)}_{t+}, W^{(f)}_{t+} * g_t)$	0.61	0.66	0.38	0.31	0.39
$\rho(\alpha^{(f)}_{t-}, \alpha^{(f)}_{t+})$	-0.78	0.23	-0.72	-0.57	-0.49

(a) Threshold u at **5%** of sample fraction

	PC1	PC2	PC3	PC4	PC5
$\rho(\alpha^Y_{t-}, g_t)$	-0.22	-0.05	-0.41	-0.06	-0.17
$\rho(\alpha^Y_{t+}, g_t)$	-0.04	-0.06	-0.23	-0.19	-0.17
$\rho(\alpha^{(f)}_{t-}, g_t)$	-0.69	0.07	-0.34	0.01	-0.37
$\rho(\alpha^{(f)}_{t+}, g_t)$	0.45	-0.22	0.19	0.18	-0.16
$\rho(\alpha^{(f)}_{t-}, W^{(f)}_{t-} * g_t)$	-0.71	-0.44	-0.42	0.03	0.07
$\rho(\alpha^{(f)}_{t+}, W^{(f)}_{t+} * g_t)$	0.22	-0.31	-0.10	-0.35	0.14
$\rho(\alpha^{(f)}_{t-}, \alpha^{(f)}_{t+})$	-0.44	-0.57	-0.11	-0.10	-0.25

(b) Threshold u at **0.5%** of sample fraction

This table presents the correlation between various specifications of the cross-sectional Hill estimates. Here α^Y_{t-} and α^Y_{t+} are the cross-sectional Hill (1975) estimates for the cross-section of county population growth for the left and right tail, respectively. The $\alpha^{(f)}_{t-}$, stated in the third and fourth rows of each panel, is the cross sectional tail index where the factor f's effect is isolated, as defined in (8). The five factors are the first five principal components from an assortment of variables suggested by the literature. In the **fifth** and **sixth** row, the factor realization at time t is multiplied by the average factor loading, as measured by γ_{ji}, of the k counties included in the estimation of $\alpha^{S^{(f)}}$, i.e $W^{(f)}_t = {}^1/\left(\sum_{j=1}^m I\{S^{(f)}_{jt} > u\}\right) \cdot \sum_{j=1}^m \hat{\gamma}^{(f)}_{jt} I\{S^{(f)}_{jt} > u\}$. The **last** row shows the correlation between the left and right tail estimates of $\alpha^{(f)}_t$. The upper panel presents the correlations where the threshold u is set to 5% of the sample fraction and for the lower panel this threshold is set to 0.5% of the sample fraction.

Table 5: Regression cross-sectional tail index (County population growth)

$\alpha_{t-}^{(PC1)}$	0.44*						0.56*
	(0.25)						(0.32)
$\alpha_{t-}^{(PC2)}$		−0.03					−0.73
		(0.64)					(0.76)
$\alpha_{t-}^{(PC3)}$			0.19				0.06
			(0.33)				(0.35)
$\alpha_{t-}^{(PC4)}$				0.36			0.11
				(0.44)			(0.48)
$\alpha_{t-}^{(PC5)}$					0.17		0.12
					(0.39)		(0.41)
α_{t-}^{X}						0.66***	
						(0.12)	
R^2	0.07	0.0001	0.01	0.01	0.004	0.40	0.10

(a) Left cross-sectional tail index, i.e. $\hat{\alpha}_{t-}^{Y}$

$\alpha_{t+}^{(PC1)}$	−0.29						−0.33*
	(0.18)						(0.20)
$\alpha_{t+}^{(PC2)}$		0.06					0.12
		(0.34)					(0.37)
$\alpha_{t+}^{(PC3)}$			0.09				−0.001
			(0.30)				(0.32)
$\alpha_{t+}^{(PC4)}$				−0.03			0.09
				(0.33)			(0.36)
$\alpha_{t+}^{(PC5)}$					−0.19		−0.32
					(0.23)		(0.28)
α_{t+}^{X}						0.73***	
						(0.11)	
R^2	0.05	0.001	0.002	0.0002	0.02	0.50	0.09

(b) Right cross-sectional tail index, i.e. $\hat{\alpha}_{t+}^{Y}$

This table displays the regression results for the cross-sectional Hill estimate extracted from US county level population growth. Here the dependent variable in the upper panel is α_{t-}^{Y}, i.e. is the Hill (1975) estimate for the left tail of the raw cross-sectional returns. The independent variables $\widehat{\alpha}_{t}^{(f)}$, stated in the first column, is the cross sectional tail index where the factor f's effect is isolated, as defined in (8). Furthermore, α_{t-}^{X} is the tail index estimated on the estimated disturbance terms of the five principal components model. The five factors are the first five principal components from an assortment of variables suggested by the literature. The lower panel shows the results right tail of the distribution. The threshold u to estimate the Hill estimate is set to 5% of the sample fraction. The constant included in the regression is excluded from the presented results. * p<0.1; ** p<0.05; *** p<0.01

5 Conclusion

In this paper we show that tail-index estimates extracted from the cross-section contain a bias. Under the assumption that the underlying process is a linear factor model with heavy-tailed idiosyncratic shocks, the tail behavior of the linear factor model is inherited from the idiosyncratic shocks. In an intermediate sample size, the bias is driven by the inconsequential factors and coefficients. Assuming that the coefficients of the factors are positive, an increase in the factors increases the tail index estimate in the right tail. For the left tail this effect is in the opposite direction and decreases the left tail index estimates. Furthermore, the effect of the factors should diminish with the threshold that defines the tail region. We test these predictions using data on US stock returns and US county population change. For US stock return data, we find that variation in the cross-sectional tail index can be largely explained by variation in US stock factors. Since US county population data lacks a clear factor structure, the ability of county population factors to explain the variation in the cross-sectional tail index is diminished. Several conclusions however, hold for both datasets. We find that the isolated effect of the factors on the Hill estimate is highly correlated with the realizations of the factor. This effect appears in opposite directions for the left and the right tail. Furthermore, we find that their influence becomes smaller as the tail threshold is moved deeper into the tail of the distribution.

The conclusions drawn from studying tail index estimates extracted from the cross section could therefore be misleading. The time variation in these estimates, can be caused by known factors, unknown factors, measurement error or tail index. Therefore, we advise caution when attributing measured variation to the scaling behavior of the linear factor model and choose the threshold appropriately.

References

Atkinson, A. B. and T. Piketty (2007). *Top incomes over the twentieth century: a contrast between continental european and english-speaking countries*. Oxford University Press.

Axtell, R. L. (2001). "Zipf distribution of US firm sizes". In: *Science* 293.5536, pp. 1818–1820.

Baker, G. P., M. C. Jensen, and K. J. Murphy (1988). "Compensation and incentives: Practice vs. theory". In: *The Journal of Finance* 43.3, pp. 593–616.

Bingham, N. H., C. M. Goldie, and J. L. Teugels (1987). *Regular variation*. Vol. 27. Cambridge University Press.

Breiman, L. (1965). "On some limit theorems similar to the arc-sin law". In: *Theory of Probability & Its Applications* 10.2, pp. 323–331.

Carhart, M. M. (1997). "On persistence in mutual fund performance". In: *The Journal of Finance* 52.1, pp. 57–82.

Chi, G. and S. J. Ventura (2011). "An Integrated Framework of Population Change: Influential Factors, Spatial Dynamics, and Temporal Variation". In: *Growth and Change* 42.4, pp. 549–570.

Csorgo, S., P. Deheuvels, and D. Mason (1985). "Kernel estimates of the tail index of a distribution". In: *The Annals of Statistics* 13, pp. 1050–1077.

Dekkers, A. L., J. H. Einmahl, and L. De Haan (1989). "A moment estimator for the index of an extreme-value distribution". In: *The Annals of Statistics* 17, pp. 1833–1855.

Drees, H. (1998). "A general class of estimators of the extreme value index". In: *Journal of Statistical Planning and Inference* 66.1, pp. 95–112.

Eeckhout, J. (Dec. 2004). "Gibrat's Law for (All) Cities". In: *American Economic Review* 94.5, pp. 1429–1451.

Fama, E. F. and K. R. French (1996). "Multifactor explanations of asset pricing anomalies". In: *Journal of Finance* 51.1, pp. 55–84.

Fama, E. F. and K. R. French (2015). "Dissecting anomalies with a five-factor model". In: *Review of Financial Studies* 29.1, pp. 69–103.

Feller, W. (1971). *An Introduction to probability theory and its applications*. 2nd ed. Vol. 2. Michigan: Wiley, p. 626.

Gabaix, X. (2009). "Power laws in economics and finance". In: *Annual Review of Economics* 1.1, pp. 255–294.

Gabaix, X., P. Gopikrishnan, V. Plerou, and H. E. Stanley (2006). "Institutional investors and stock market volatility". In: *Quarterly Journal of Economics* 121.2, pp. 461–504.

Gabaix, X. (1999). "Zipf's Law for Cities: An Explanation". In: *The Quarterly Journal of Economics* 114.3, pp. 739–767.

Gabaix, X. and A. Landier (2008). "Why has CEO pay increased so much?" In: *The Quarterly Journal of Economics* 123.1, pp. 49–100.

Goldie, C. M. and R. L. Smith (1987). "Slow variation with remainder: Theory and applications". In: *The Quarterly Journal of Mathematics* 38.1, pp. 45–71.

Haan de, L. and C. Zhou (2019). "Trends in extreme value indices". In: *Journal of the American Statistical Association* just-accepted.

Hall, P. and A. Welsh (1985). "Adaptive estimates of parameters of regular variation". In: *The Annals of Statistics* 13.1, pp. 331–341.

Helpman, E., M. J. Melitz, and S. R. Yeaple (2004). "Export versus FDI with heterogeneous firms". In: *American Economic Review* 94.1, pp. 300–316.

Hill, B. M. (1975). "A simple general approach to the inference about the tail of a distribution". In: *The Annals of Statistics* 3.5, pp. 1163–1174.

Hwang, S. and A. Rubesam (2015). "The disappearance of momentum". In: *The European Journal of Finance* 21.7, pp. 584–607.

Jansen, D. W. and C. G. de Vries (1991). "On the frequency of large stock returns: Putting booms and busts into perspective". In: *The Review of Economics and Statistics* 73.1, pp. 18–24.

Kelly, B. and H. Jiang (2014). "Tail risk and asset prices". In: *Review of Financial Studies* 27.10, pp. 2841–2871.

Lintner, J. (1965). "The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets". In: *The Review of Economics and Statistics* 47.1, pp. 13–37.

Pickands III, J. (1975). "Statistical inference using extreme-order statistics". In: *The Annals of Statistics* 3.1, pp. 119–131.

Quintos, C., Z. Fan, and P. C. B. Philips (2001). "Structural change tests in tail behaviour and the Asian crisis". In: *Review of Economic Studies* 68.3, pp. 633–663.

Reed, W. J. (2001). "The Pareto, Zipf and other power laws". In: *Economics Letters* 74.1, pp. 15–19.

Rozenfeld, H. D., D. Rybski, X. Gabaix, and H. A. Makse (Aug. 2011). "The Area and Population of Cities: New Insights from a Different Perspective on Cities". In: *American Economic Review* 101.5, pp. 2205–25.

Sharpe, W. F. (1964). "Capital asset prices: A theory of market equilibrium under conditions of risk". In: *The Journal of Finance* 19.3, pp. 425–442.

Stambaugh, R. F. and P. Lubos (2003). "Liquidity risk and expected stock returns". In: *Journal of Political Economy* 111.3, pp. 642–685.

Sun, P. and C. G. de Vries (2018). "Exploiting tail shape biases to discriminate between stable and Student-t alternatives". In: *Journal of Applied Econometrics* 33.5, pp. 708–726.

6 Appendix

6.1 Bias under second-order Hall expansion

Goldie and Smith (1987) provide an explicit expression for the bias based on the following so called (Hall and Welsh, 1985) expansion

$$F(x) = 1 - Cx^{-\alpha}\left[1 + Dx^{-\theta} + o(1)\right]. \tag{10}$$

Here $\alpha > 0$, $C > 0$, $\theta > 0$ and D is a real number. Here C and D are the first and second-order scale parameters, where α and θ are the first and second-order shape parameters.We give a short derivation of the (conditional) bias in case the distribution function adheres to the above expansion. If the distribution function satisfies the monotone density theorem (see Bingham et al., 1987), it is sufficiently smooth so that the derivative gives its density with tail expansion

$$f(x) = \alpha C x^{-\alpha-1} + (\alpha+\theta) CD x^{-\alpha-\theta-1} + o(1).$$

We therefore need

$$\begin{aligned} &E\left[\ln \frac{R}{u} | R > u\right] \\ &\simeq \frac{1}{Cu^{-\alpha}\left[1 + Du^{-\theta}\right]} \int_u^\infty \left(\ln \frac{x}{u}\right) \left\{\alpha C x^{-\alpha-1} + (\alpha+\theta) CD x^{-\alpha-\theta-1}\right\} dx, \end{aligned}$$

if we omit the terms that are of order small. Note that we can cancel the C factor from the numerator and denominator. If we then apply

$$\alpha \int_u^\infty \left(\ln \frac{s}{u}\right) s^{-\alpha-1} ds = -\left(\ln \frac{s}{u}\right) s^{-\alpha}|_u^\infty + \int_u^\infty s^{-\alpha-1} ds = \frac{1}{\alpha} u^{-\alpha}$$

to the two parts separately, we get

$$\begin{aligned} E\left[\ln \frac{R}{u} | R > u\right] &\simeq \frac{u^\alpha}{1 + Dx^{-\theta}} \left\{\frac{1}{\alpha} u^{-\alpha} + \frac{1}{\alpha+\theta} D u^{-\alpha-\theta}\right\} \\ &= \frac{1}{\alpha} + \left(\frac{1}{\alpha+\theta} - \frac{1}{\alpha}\right) \frac{Du^{-\theta}}{1 + Du^{-\theta}} \\ &\simeq \frac{1}{\alpha} - \frac{1}{\alpha(\alpha+\theta)} D u^{-\theta} \end{aligned}$$

as $1 + Du^{-\theta} \to 1$ for u large.

6.2 Differences in tail indices

Return to the analysis of the pure Pareto case, but now assume that one proportion of the sample comes with a tail index α and the other proportion with a larger index $\alpha + \varepsilon$, $\varepsilon > 0$. Denote the observation with index α by Y_i and the observation with index $\alpha + \varepsilon$ by Y_j. Thus

$$\Pr\{Y_i > s\} = s^{-\alpha}$$

and

$$\Pr\{Y_j > s\} = s^{-\alpha+\varepsilon}.$$

Conditional on being above threshold u, we get from the above

$$E\left[\ln \frac{Y_i}{u} | Y_i > u\right] = \frac{1}{\alpha}$$

and

$$E\left[\ln \frac{Y_j}{u} | Y_j > u\right] = \frac{1}{\alpha+\varepsilon}.$$

Suppose that $u > 1$ as a proportion λ of the sample size n. Of course in the Pareto case it is optimal to use all data, but the presumption is that the researcher does not have this detailed information. Thus

$$u \sim \lambda n.$$

Then in larger samples, the proportion of observations on each type that are above the threshold u are respectively

$$\lambda u^{-\alpha} n$$

and

$$\lambda u^{-\alpha-\varepsilon} n$$

in probability. The expected value of the Hill estimator is a mixture of the two conditional expectation weighted by the proportion by which the two types of observations appear,

$$\begin{aligned} E\left[\frac{1}{K}\sum_{m=1}^{K} \ln \frac{Y_m}{u} \middle| Y_m > u\right] &= \frac{\lambda u^{-\alpha} n}{\lambda u^{-\alpha} n + \lambda u^{-\alpha-\varepsilon} n} \frac{1}{\alpha} + \frac{\lambda u^{-\alpha-\varepsilon} n}{\lambda u^{-\alpha} n + \lambda u^{-\alpha-\varepsilon} n} \frac{1}{\alpha+\varepsilon} \\ &= \frac{1}{1+u^{-\varepsilon}} \frac{1}{\alpha} + \frac{u^{-\varepsilon}}{1+u^{-\varepsilon}} \frac{1}{\alpha+\varepsilon} \end{aligned}$$

for $m = i, j$. Thus in large samples as u increases one recovers $1/\alpha$. In smaller samples there is a bias. One can extend this analysis with the effects of a shift factor h. If $\varepsilon > 1$, then the second order term is due to the shift factor, otherwise the second-order term is due to $\alpha + \varepsilon$.

6.3 Tables and Figures

Table 6: Correlation asset pricing factors.

	Market	SMB	HML	MOM	Liq	RMW	CMA
Market	1.00	0.27	-0.27	-0.14	-0.01	-0.24	-0.40
SMB	0.27	1.00	-0.08	-0.05	0.00	-0.37	-0.08
HML	-0.27	-0.08	1.00	-0.19	0.04	0.08	0.70
MOM	-0.14	-0.05	-0.19	1.00	-0.01	0.11	-0.01
Liq	-0.01	0.00	0.04	-0.01	1.00	-0.01	0.02
RMW	-0.24	-0.37	0.08	0.11	-0.01	1.00	-0.01
CMA	-0.40	-0.08	0.70	-0.01	0.02	-0.01	1.00

Table 7: Regression cross-sectional tail index **(0.5% Threshold)**

$\alpha_{t-}^{(M)}$	0.61^{***}						0.55^{***}
	(0.08)						(0.08)
$\alpha_{t-}^{(SMB)}$		0.41^{***}					0.31^{***}
		(0.08)					(0.08)
$\alpha_{t-}^{(HML)}$			0.38^{***}				0.22^{**}
			(0.11)				(0.11)
$\alpha_{t-}^{(MOM)}$				0.10			-0.19
				(0.12)			(0.12)
$\alpha_{t-}^{(Liq)}$					0.21^{*}		-0.11
					(0.12)		(0.12)
α_{t-}^{X}						0.78^{***}	
						(0.04)	
R^2	0.10	0.05	0.02	0.001	0.01	0.35	0.13

(a) Left cross-sectional tail index, i.e. $\widehat{\alpha}_{t-}^{Y}$

$\alpha_{t+}^{(M)}$	0.45^{***}						0.46^{***}
	(0.10)						(0.10)
$\alpha_{t+}^{(SMB)}$		0.34^{***}					0.35^{***}
		(0.11)					(0.12)
$\alpha_{t+}^{(HML)}$			-0.15				-0.56^{***}
			(0.15)				(0.16)
$\alpha_{t+}^{(MOM)}$				0.18			-0.10
				(0.20)			(0.21)
$\alpha_{t+}^{(Liq)}$					0.69^{***}		0.65^{***}
					(0.19)		(0.20)
α_{t+}^{X}						0.89^{***}	
						(0.02)	
R^2	0.03	0.01	0.002	0.001	0.02	0.69	0.07

(b) Right cross-sectional tail index, i.e. $\widehat{\alpha}_{t+}^{Y}$

This table displays the regression results for the cross-sectional Hill estimate. Here the dependent variable in the upper panel is $\widehat{\alpha}_t^Y$, i.e. is the Hill (1975) estimate for the left tail of the raw cross-sectional returns. The independent variables $\widehat{\alpha}_t^{(f)}$, stated in the first column, is the cross sectional tail index where the factor f's effect is isolated, as defined in (8). Furthermore, $\widehat{\alpha}_t^X$ is the tail index estimated on the estimated disturbance terms of the five-factor asset pricing model. The five factors are the market, small-minus-big, high-minus-low, momentum and the liquidity factor. The lower panel shows the results right tail of the distribution. The threshold u to estimate the Hill estimate is set to **0.5%** of the sample fraction. The constant included in the regression is excluded from the presented results. * p<0.1; ** p<0.05; *** p<0.01

Table 8: Regression cross-sectional tail index **(0.5% Threshold)** (County Data)

$\alpha_{t-}^{(PC1)}$	0.61**						0.58**
	(0.24)						(0.24)
$\alpha_{t-}^{(PC2)}$		0.06					−0.31
		(0.26)					(0.25)
$\alpha_{t-}^{(PC3)}$			0.81***				0.74***
			(0.27)				(0.26)
$\alpha_{t-}^{(PC4)}$				0.93*			0.74
				(0.49)			(0.47)
$\alpha_{t-}^{(PC5)}$					0.19		−0.09
					(0.37)		(0.36)
α_{t-}^{X}						0.85***	
						(0.14)	
R^2	0.13	0.001	0.18	0.08	0.01	0.45	0.33

(a) Left cross-sectional tail index, i.e. $\widehat{\alpha}_{t-}^{Y}$

$\alpha_{t+}^{(PC1)}$	0.51						0.97*
	(0.45)						(0.53)
$\alpha_{t+}^{(PC2)}$		−0.22					0.12
		(0.58)					(0.78)
$\alpha_{t+}^{(PC3)}$			−0.37				−0.76
			(0.34)				(0.56)
$\alpha_{t+}^{(PC4)}$				−0.34			−0.03
				(0.56)			(0.60)
$\alpha_{t+}^{(PC5)}$					−0.41		−0.05
					(0.50)		(0.61)
α_{t+}^{X}						1.04***	
						(0.21)	
R^2	0.03	0.003	0.03	0.01	0.02	0.36	0.11

(b) Right cross-sectional tail index, i.e. $\widehat{\alpha}_{t+}^{Y}$

This table displays the regression results for the cross-sectional Hill estimate. Here the dependent variable in the upper panel is $\widehat{\alpha}_t^Y$, i.e. is the Hill (1975) estimate for the left tail of the raw cross-sectional returns. The independent variables $\widehat{\alpha}_t^{(f)}$, stated in the first column, is the cross sectional tail index where the factor f's effect is isolated, as defined in (8). Furthermore, $\widehat{\alpha}_t^X$ is the tail index estimated on the estimated disturbance terms of the five-factor asset pricing model. The five factors are the market, small-minus-big, high-minus-low, momentum and the liquidity factor. The lower panel shows the results right tail of the distribution. The threshold u to estimate the Hill estimate is set to **0.5%** of the sample fraction. The constant included in the regression is excluded from the presented results. * p<0.1; ** p<0.05; *** p<0.01

Table 9: Regression cross-sectional tail index and factors

	(1)	(2)	(3)	(4)	(5)	(6)
Market	-0.18^{***}					-0.16^{***}
	(0.01)					(0.01)
SMB		-0.17^{***}				-0.11^{***}
		(0.01)				(0.01)
HML			0.07^{***}			-0.003
			(0.01)			(0.01)
MOM				0.03^{***}		-0.001
				(0.01)		(0.01)
Liq					1.23	1.04
					(1.25)	(0.71)
Constant	3.33^{***}	3.27^{***}	3.21^{***}	3.22^{***}	3.23^{***}	3.33^{***}
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)
R^2	0.58	0.25	0.04	0.01	0.002	0.68

(a) $\widehat{\alpha}^Y_{t-}$

	(1)	(2)	(3)	(4)	(5)	(6)
Market	0.08^{***}					0.08^{***}
	(0.003)					(0.003)
SMB		0.08^{***}				0.05^{***}
		(0.01)				(0.005)
HML			-0.02^{**}			0.02^{***}
			(0.01)			(0.01)
MOM				-0.01^{**}		0.003
				(0.005)		(0.003)
Liq					-0.42	-0.39
					(0.62)	(0.41)
Constant	2.17^{***}	2.20^{***}	2.22^{***}	2.22^{***}	2.21^{***}	2.16^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)
R^2	0.49	0.20	0.01	0.01	0.001	0.57

(b) $\widehat{\alpha}^Y_{t+}$

This table displays the regression results to explain the time variation in the cross-sectional Hill estimate with asset pricing factors. Here α^Y_{t-}, i.e. is the Hill (1975) estimate of the raw cross-sectional excess returns, is the dependent variable. The upper panel shows the results for the left tail and the lower panel shows the results for the right tail of the cross-sectional distribution. The threshold u in the Hill estimator is set to 5% of the sample fraction. The five factors are the market, small-minus-big, high-minus-low, momentum and the liquidity factor. * p<0.1; ** p<0.05; *** p<0.01

Table 10: Correlations cross-sectional tail index **(RMW and CMA Factors)**

	Market	SMB	HML	RMW	CMA
$\rho(\alpha^Y_{t-}, g_t)$	-0.76	-0.50	0.20	0.23	0.31
$\rho(\alpha^Y_{t+}, g_t)$	0.70	0.45	-0.10	-0.12	-0.20
$\rho(\alpha^{(f)}_{t-}, g_t)$	-0.80	-0.78	-0.16	-0.06	0.04
$\rho(\alpha^{(f)}_{t+}, g_t)$	0.81	0.79	0.26	-0.02	0.08
$\rho(\alpha^{(f)}_{t-}, W^{(f)}_{t-} * g_t)$	-0.80	-0.79	-0.13	-0.15	-0.13
$\rho(\alpha^{(f)}_{t+}, W^{(f)}_{t+} * g_t)$	0.78	0.71	-0.14	0.02	-0.12
$\rho(\alpha^{(f)}_{t-}, \alpha^{(f)}_{t+})$	-0.92	-0.82	-0.14	-0.00	-0.02

(a) Threshold u at **5%** of sample fraction

	Market	SMB	HML	RMW	CMA
$\rho(\alpha^Y_{t-}, g_t)$	-0.47	-0.37	0.06	0.22	0.20
$\rho(\alpha^Y_{t+}, g_t)$	0.21	0.20	0.02	-0.09	-0.02
$\rho(\alpha^{(f)}_{t-}, g_t)$	-0.50	-0.38	-0.03	0.06	-0.04
$\rho(\alpha^{(f)}_{t+}, g_t)$	0.49	0.41	0.01	-0.05	-0.07
$\rho(\alpha^{(f)}_{t-}, W^{(f)}_{t-} * g_t)$	-0.49	-0.40	-0.07	0.04	-0.09
$\rho(\alpha^{(f)}_{t+}, W^{(f)}_{t+} * g_t)$	0.42	0.30	0.01	0.02	-0.04
$\rho(\alpha^{(f)}_{t-}, \alpha^{(f)}_{t+})$	-0.31	-0.30	0.13	0.03	0.01

(b) Threshold u at **0.5%** of sample fraction

This table presents the correlation between various specifications of the cross-sectional Hill estimates. Here α^Y_{t-} and α^Y_{t+} are the cross-sectional Hill (1975) estimates for the cross-section of stock returns for the left and right tail, respectively. The $\alpha^{(f)}_{t-}$, stated in the third and fourth rows of each panel, is the cross sectional tail index where the factor f's effect is isolated, as defined in (8). The five factors are the market, small-minus-big, high-minus-low, robust-minus-weak and the conservative-minus-aggressive factor. In the **fifth** and **sixth** row, the factor realization at time t is multiplied by the average factor loading, as measured by γ_{ij}, of the k stocks included in the estimation of $\alpha^{S^{(f)}}$, i.e $W^{(f)}_t = {}^1/\left(\sum_{j=1}^m I\{S^{(f)}_{jt} > u\}\right) \cdot \sum_{j=1}^m \hat{\gamma}^{(f)}_{jt} I\{S^{(f)}_{jt} > u\}$. The **last** row shows the correlation between the left and right tail estimates of $\alpha^{(f)}_t$. The upper panel presents the correlations where the threshold u is set to 5% of the sample fraction and for the lower panel this threshold is set to 0.5% of the sample fraction.

Table 11: Regression cross-sectional tail index (**Pickands estimator**)

$\alpha_{t-}^{(M)}$	0.05*						0.08
	(0.03)						(0.05)
$\alpha_{t-}^{(SMB)}$		0.04					0.03
		(0.02)					(0.03)
$\alpha_{t-}^{(HML)}$			0.01				−0.06
			(0.03)				(0.04)
$\alpha_{t-}^{(MOM)}$				0.03			−0.02
				(0.03)			(0.05)
$\alpha_{t-}^{(Liq)}$					0.03		0.02
					(0.03)		(0.04)
α_{t-}^{X}						−0.04	
						(0.03)	
R^2	0.01	0.004	0.0002	0.001	0.002	0.003	0.01

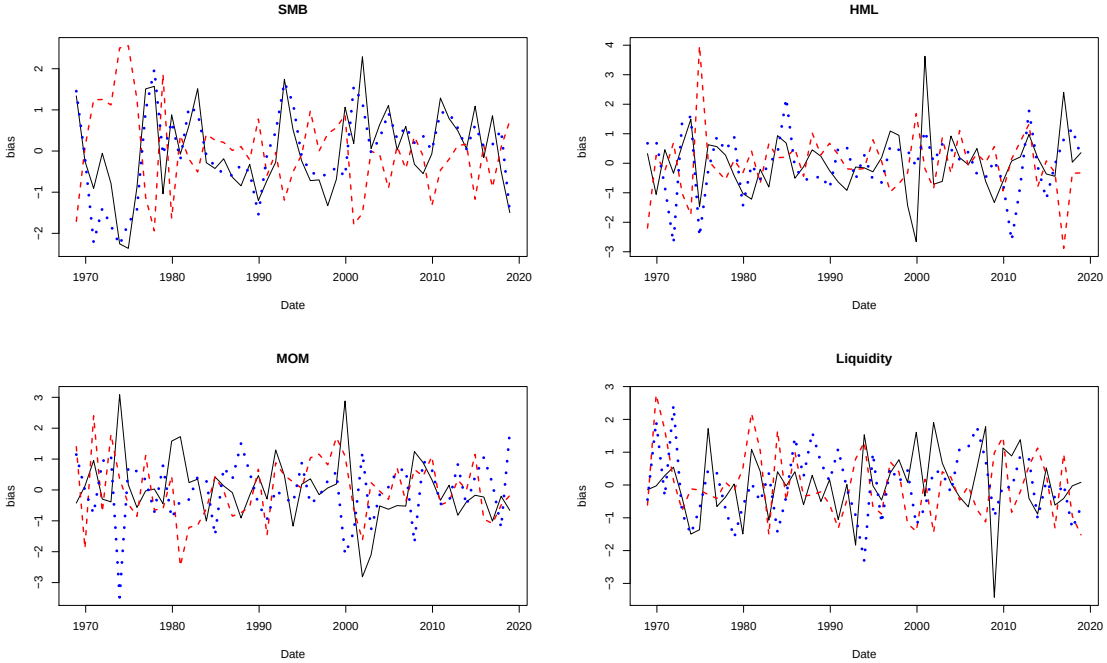
(a) Left cross-sectional tail index, i.e. $\hat{\alpha}_{t-}^{Y}$

$\alpha_{t+}^{(M)}$	−0.03						−0.07**
	(0.02)						(0.03)
$\alpha_{t+}^{(SMB)}$		0.01					0.02
		(0.03)					(0.04)
$\alpha_{t+}^{(HML)}$			−0.01				−0.04
			(0.02)				(0.03)
$\alpha_{t+}^{(MOM)}$				0.02			0.03
				(0.02)			(0.03)
$\alpha_{t+}^{(Liq)}$					0.03		0.05*
					(0.02)		(0.03)
α_{t+}^{X}						−0.003	
						(0.03)	
R^2	0.004	0.0003	0.0003	0.002	0.004	0.0000	0.02

(b) Right cross-sectional tail index, i.e. $\hat{\alpha}_{t+}^{Y}$

This table displays the regression results for the cross-sectional **Pickands** estimate. Here the dependent variable in the upper panel is α_{t-}^{Y}, i.e. is the Pickands estimate for the left tail of the raw cross-sectional returns. The independent variables $\widehat{\alpha}_{t}^{(f)}$, stated in the first column, is the cross sectional tail index where the factor f's effect is isolated, as defined in (8). Furthermore, α_{t-}^{X} is the tail index estimated on the estimated disturbance terms of the five-factor asset pricing model. The five factors are the market, small-minus-big, high-minus-low, momentum and the liquidity factor. The lower panel shows the results right tail of the distribution. The threshold u to estimate the Hill estimate is set to 5% of the sample fraction. The constant included in the regression is excluded from the presented results. * p<0.1; ** p<0.05; *** p<0.01

Figure 2: Effect asset pricing factor on cross-sectional Hill estimator



This figure displays the time series of the asset pricing factors and the isolated effect of these factors on the cross-sectional Hill estimator, as defined in (8). The time series of annual observations are standardized by their respective mean and standard deviation. The annual observation are created by averaging the monthly observations within a year. The two top plots are for the two Fama-French factors, small-minus-big and high-minus-low. The bottom left plot is for the momentum factors by Carhart (1997) and the bottom right is for the liquidity factor by Stambaugh and Lubos (2003). All series are standardized by their respective mean and standard deviation and are the annual average from the monthly observations. The solid black line depicts the time series of the factor. The red dashed line is the isolated effect of the factor in the left tail index estimate. The blue dotted line is the isolated effect of the factor in the right tail index estimate.

Table 12: Summary of principal components for county data

	PC1	PC2	PC3	PC4	PC5
Standard deviation	2.89	2.66	1.87	1.65	1.59
Proportion of Variance	0.21	0.18	0.09	0.07	0.06
Cumulative Proportion	0.21	0.39	0.47	0.54	0.60

This table presents a summary of the principal components used in the analysis for county data. The first two rows give the standard deviation and the proportion of variance explained by each principal component respectively. The proportion of variance being explained is that of the variables discussed in the data section for county data. The last row gives the cumulative proportion of variance that is explained by the corresponding principal component and all those previous to it.

Table 13: PCA input variables used for county data

Variable name	Description	Source
B279RA3A086NBEA	Real private investment: Trucks, buses, and truck trailers	FRED
B280RA3A086NBEA	Real private investment: Autos	FRED
B281RA3A086NBEA	Real private investment: Aircraft	FRED
B282RA3A086NBEA	Real private investment: Ships & Boats	FRED
DMUSRC1A027NBEA	Personal consumption: Museums and libraries	FRED
FCTAX	Tax Receipts on Corporate Income	FRED
G160291A027NBEA	Government expenditures: Education	FRED
I3GTOTL1SN000	Government Fixed Assets Investment: Transportation	FRED
SPDYNTFRTINUSA	Fertility Rate	FRED
STTMINWGFG	Federal Minimum Wage Rate	FRED
W188RC1A027NBEA	Government Fixed Assets: Transportation structures	FRED
W691RC1A027NBEA	Government expenditures: Libraries	FRED
AHETPI	Average Earnings of Production and Nonsupervisory Employees	FRED
CPITRNSL	Consumer Price Index: Transportation	FRED
CUSR0000SETB01	Consumer Price Index: Gasoline	FRED
CUUR0000SAS4	Consumer Price Index: Transportation Sevices	FRED
CWSR0000SA0	Consumer Price Index: All Items in U.S. City Average	FRED
FEDMINFRMWG	Minimum Hourly Wage for Farm Workers	FRED
FEDMINNFRWG	Minimum Hourly Wage for Non-Farm Workers	FRED
LNS12300002	Employment-Population Ratio: Women	FRED
MSACSR	Monthly Supply of Houses	FRED
UNRATE	Unemployment Rate	FRED
USEHS	Employees: Education & Health Services	FRED
A939RC0Q052SBEA	Gross domestic product/capita	FRED
A939RX0Q048SBEA	Real gross domestic product/capita	FRED
ASPUS	Average Sales Price of Houses Sold	FRED
FGEXPND	Federal Government: Current Expenditures	FRED
GCEC1	Real Government Consumption and Gross Investment	FRED
MSPUS	Median Sales Price of Houses Sold for the United States	FRED
W006RC1Q027SBEA	Federal government current tax receipts	FRED
W068RCQ027SBEA	Government total expenditures	FRED
W369RG3Q066SBEA	Terms of trade index	FRED
HCCSDODNS	Consumer Credit; Liability	FRED
Citizens 25+ College degree (%)		US Census
NE.TRD.GNFS.ZS	Trade (% of GDP)	World Bank
SP.DYN.CDRT.IN	Death rate, crude (per 1,000 people)	World Bank
SI.POV.GINI	GINI index	World Bank
BN.CAB.XOKA.CD	Current account balance	World Bank
SM.POP.NETM	Net migration	World Bank

This table presents the variables used as input for the PCA, to describe the county data. The variables are obtained from the websites FRED, the World Bank and US Census.