Covariates hiding in the tails

Milian Bachem*

Erasmus University Rotterdam

Lerby M. Ergun†

Bank of Canada
London School of Economics

Casper G. de Vries‡

Erasmus University Rotterdam

Saturday 17th July, 2021

DRAFT

Abstract

Scaling behavior measured in cross-sectional studies through the tail index of a power law is prone to a bias. This hampers inference; in particular, time variation in estimated tail indices may be erroneous. In case of a linear factor model, the factor biases the tail indices in the left and right tail in opposite directions. This fact can be exploited to reduce the bias. We show how this bias arises from the factor, how to remedy for the bias and apply our methods to financial data and geographic location data.

*bachem@ese.eur.nl, Burgemeester Oudlaan 50, Rotterdam, 3062 PA, The Netherlands.

†Corresponding author: lerbyergun@gmail.com, 234 Wellington Street West, Ottawa, ON, K1A 0G9, Canada.

‡cdevries@ese.eur.nl, Burgemeester Oudlaan 50, Rotterdam, 3062 PA, The Netherlands.

We would like to thank Jason Allen, Maarten Bosker, Bruno Feunou, Sermin Gungor, Andreas Uthemann, Jun Yang and Chen Zhou for their helpful comments. We also thank the seminar participants at the Bank of Canada, ES World Congress 2020 and EEA meetings 2020.

1 Introduction

A wide variety of economic data and natural processes exhibit scaling behavior, in the sense that one variable varies as a power of another variable. For example, the number of households in higher income brackets varies as a power of the level of income (Pareto's law). This implies that the upper tail of the income density falls off by a power. If the tails of the density follow a power law, the shape is the same regardless the magnification of the quantiles, in other words the tails are self similar. For the extreme order statistics this implies that the logarithm of the probability of observing a realization above a certain threshold is linear in the logarithm of the threshold. The ratio of the two logarithms is equal to (minus) the power, which is generally known as the tail index. Distributions with such tail behavior only have bounded moments up to the value of the tail index and are therefore referred as heavy-tailed distributions.

In economics, scaling behavior is found in wealth and income (Atkinson and Piketty, 2007), firm size (Axtell, 2001), executive compensation (Baker et al., 1988), productivity (Helpman et al., 2004) and stock markets (Jansen and Vries, 1991). More generally, scaling behavior is found in a variety of natural processes, such as internet data traffic (Resnick, 1997), city size (Gabaix, 1999) and natural disasters (Pisarenko and Rodkin, 2010).

Considerable attention in economics has been paid to time series characterized by heavy-tailed innovations, like returns to financial investments. Since investors generally choose a portfolio from a large multitude of different assets, recent literature also investigates the cross-sectional scaling behavior. Kelly and Jiang (2014) estimate the tail index from monthly cross-sections of US stock returns. They find that the tail-index varies considerably across different months. More recently, Karagiannis and Tolikas (2019), Atilgan et al. (2020) and Agarwal et al. (2017) use measures related to the tail shape of cross-sections (such as Value-at-Risk) to expose risks of different assets not priced by the market.

To date little is known about the statistical properties of the estimates of scaling behavior in cross-sections. In the cited literature considerable time variation is observed in tail index estimates. The main issue that we investigate is the possible cause for this variation that is particular to the cross-sectional nature of the data. Suppose the data are generated by a linear factor model, that is, the dependent variable equals a weighted sum of the factors and some idiosyncratic noise. Furthermore, assume that the scaling behavior is identical for all idiosyncratic noises in the cross section, implying equality of the tail indices. In this setup, the tail index of the dependent variable equals the tail index of the idiosyncratic noise. We show that cross-sectional estimates of the tail index are biased and vary with the size of the factor realizations at different points in time.

Subsequently, we use the specific properties of linear factor models to remedy for the bias and variation in cross-sectional tail index estimates. Specifically, the factor induces bias in opposite directions for left and right tail index estimates. We show that this fact can be exploited to reduce the bias. If one does not correct for the bias, one may misinterpret observed variation in tail index estimates.

We suggest two simple-to-implement procedures to alleviate the bias due to the location shift originating from the factors. The first method takes advantage of the symmetry in the bias for the left and right tail estimate. By taking the average of the left and right tail estimates, the location shift is offset. Under tail symmetry, this not only cancels out the bias, but also reduces the variance of the estimator.

A second approach, is to subtract the average of the dependent variable in the cross-section from each observation before one applies the tail index estimator. This approach does not require the assumption that the left and right tail have the same tail shape. As opposed to the first approach, the left and right tail indices can be estimated separately. In this application we simply use the cross-sectional mean as the *tuning* parameter. In a simulation exercise we show that both methods alleviate bias caused by a cross-sectional location shift.

To test for the direction and size of the bias in real world data, we use monthly US stock returns and annual US Census county population data. Both data sets 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 elicit the effect of the bias caused by the linear factor structure.

Cochrane (2009) shows that under mild restrictions, asset returns naturally follow from a linear factor structure. The factor structure has been used to explain asset prices empirically (see e.g. Fama and French (2015); Stambaugh and Lubos (2003)). The combination of an innate 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 the variation that a single factor contributes to the bias in cross-sectional tail-index estimates. When considered in isolation, correlation between tail index estimates and factors explains a considerable amount of the variation over time. As predicted on the basis of our theory, we find that the bias indeed induces a negative correlation between the left and right tail index estimates.

A second test case is based on county population growth rate data motivated by literature on the heavy-tailed nature of geographical population clustering (Gabaix (1999); Eeckhout (2004); Rozenfeld et al. (2011)). Furthermore, the abundance of the data in both the time series and cross-section dimension implies that these data are amenable to our analysis. However, the difference with the financial data is that a clear factor structure is lacking. County data therefore provides us with an example in which there is only a weak factor structure. In the literature review by Chi and Ventura (2011) a large number of possible factors are identified that may explain population growth. We use five principal components to summarize a large subset of the proposed factors. The weak factor structure behind population growth induces less correlation between the bias and the principal components. Nevertheless, the correlation is still marginally significant and in the predicted direction. The contrast in results with the first test case underlines that a Data Generating Process (DGP) with a strong factor structure suffers more severely from the cross-sectional bias.

2 Theory

Consider a linear-factor model with n factors g_i, $i = 1, ..., n$. At any point in time the dependent variable Y_j for cross-sectional entity j is

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

where the X_j are idiosyncratic shocks (omitting superfluous time indices on Y_j, g_i and X_j). At any point in time the factors g_i are given.[1] Define, as a shortcut:

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

Thus at a specific point in time

$$Y_j = h_j + X_j. \tag{1}$$

[1] From the econometrician's point of view, the g_i may not be known or contain an error when estimated.

One can classify distributions according to their tail behavior. Either a distribution has a bounded support or it has unbounded support. In the former case, the tail behavior is captured by the Weibull distribution. In the latter case, the tail is either of exponential decay or resembles a power law. Hereafter we focus on innovations X_j, with tails that follow a power law. The tails of these distributions are always heavier than the distributions with exponential tail behavior, i.e., always have a higher probability of extreme observations. These distributions only have a finite number of bounded moments and are referred to as heavy-tailed distributions. Examples are the Pareto distribution, the Student-t and the F-distribution. The entire class of these distributions is closed under addition and characterized by the regular variation property. Let $G(.)$ denote a cumulative distribution function (cdf). For the left tail, regular variation entails:

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

and for the right tail:

$$\lim_{t\to\infty} \frac{1-G(tx)}{1-G(t)} = x^{-\alpha},$$

with $x > 0$ and $\alpha > 0$, (the left and right α's need not be equal). The power decline implies self-scaling behavior. Note that a decrease in the α gives a heavier tail as moments only exist up to α.

Below we first describe how α, commonly referred as the tail index, can be estimated for the Pareto distribution. Then we investigate how the estimator is influenced by adding a fixed factor such as h_j. We show that the introduction of h_j induces a bias in the cross-sectional estimates. Lastly, we consider how this bias can be remedied.

2.1 Hill estimator

The tail index α, can be estimated in a number of ways. Two of the most popular methods are the Hill (1975) estimator and the regression approach. While latter method is often applied in regional economics, the former method is often used in financial economics. One shows that the bias-variance trade-off differs for the two methods; both methods however attain the best possible rate in larger samples. We focus on the Hill estimator.

The Hill estimator uses the k highest-order statistics above 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:

$$Y_1 \geqslant Y_2 \geqslant ... \geqslant Y_k \geqslant u \geqslant ... \geqslant Y_{m-1} \geqslant Y_m.$$

For the lower tail, one takes the negative of the observations and reorders these from high to low. Here u is the threshold, typically chosen as a percentage of the sample size. The Hill estimator calculates the average logarithmic difference between the threshold and the higher-order statistics:

$$\frac{1}{\hat{\alpha}} = \frac{1}{K}\sum_{i=1}^{K}\ln(\frac{Y_i}{u}). \tag{2}$$

If the sample is drawn from a standard Pareto distribution, the Hill estimator coincides with the maximum likelihood estimator. In this case all observations can be used, i.e. $u = 1$. Given that 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 of the distribution resembles the Pareto tail and u must be chosen in the tail area to reduce bias. There are two versions of the Hill estimator, one is with a fixed threshold u as in (2), the other uses one of the upper order statistics as a threshold.[2] In the fixed threshold version, the number K of order statistics exceeding u is random.

2.2 Single observation

Consider a single observation X drawn from a standard Pareto distribution,

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

on $[1, \infty)$. Take $u > 1$ as one would do in the general case. In repeated samples, suppose one records a zero if $X < u$ and otherwise records $\ln(X/u)$. The expected value of the estimator (2) is the conditional expectation

$$E[\ln\frac{X}{u}|X > u] = \frac{\alpha}{u^{-\alpha}}\int_u^{\infty}(\ln\frac{x}{u})x^{-\alpha-1}dx = \frac{1}{\alpha}. \tag{3}$$

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

[2] Goldie and Smith (1987) argue that: "In practical terms, there is little to choose between these two points of view."

Next consider the case with a non-zero fixed factor, $h \neq 0$, added to the idiosyncratic noise as in (1). For large s, a first-order Taylor approximation around $hs^{-1} = 0$ yields an expression for the tail of the distribution of Y

$$\begin{aligned}\Pr\{Y \leqslant s\} = \Pr\{X + h \leqslant s\} &= 1 - (s-h)^{-\alpha} \\ &\simeq 1 - s^{-\alpha}[1 + \alpha h s^{-1}].\end{aligned} \tag{4}$$

Apply the expectation in (3) twice to get

$$E[\ln \frac{Y}{u} | Y > u] \simeq \frac{1}{\alpha} - \frac{1}{\alpha+1} h u^{-1}. \tag{5}$$

In comparison to (3), we now have an additional term signifying the bias due to the location shift. Thus, given a fixed α over time, variation in the cross-sectional estimates may stem from a variation in h.

The location shift has a different effect when considering the left tail. If the idiosyncratic noise term again follows a standard Pareto distribution, then

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

This again results in a bias dependent on h,

$$E[\ln \frac{Y}{u} | Y \leqslant -u] \simeq \frac{1}{\alpha} + \frac{1}{\alpha+1} h u^{-1}. \tag{6}$$

The bias, however, is of the opposite sign. This implies that a location shift h biases the left and right tail index estimates in different directions. The two biases are each others mirror image.

In general, heavy-tailed distributions, i.e. distributions that vary regularly at infinity, only resemble the Pareto distribution in the tail area. That is to say, these distributions have second-order terms not due to a shift. For example, the Student-t satisfies the following expansion

$$G(x) = 1 - Cx^{-\alpha}[1 + Dx^{-\theta} + o(x^{-\theta})]. \tag{7}$$

Here $\alpha > 0$, $C > 0$, $\theta > 0$ and D is a real number. In fact, most known heavy-tailed distributions satisfy this so-called Hall expansion (Hall and Welsh, 1985). The expansion also applies to the stationary distribution of an ARCH process, see e.g.

Sun and Vries (2018). But the second-order term is not necessarily a power function. In the remainder we assume that expansion (7) applies. For the general case in (7), we show in Appendix 7.1 that the expected value of the Hill statistic is

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

In the simplest case, that is, the shifted Pareto distribution in (4), $C = 1$, $D = \alpha h$ and $\theta = 1$. For a Student-t distribution $\theta = 2$ and expressions for C and D can be found in Sun and Vries (2018). One shows that adding a shift h to the Student-t distribution, changes the second-order term into the third-order term. The second-order term in expansion (7) in the shifted Student law then resembles the second-order term of the shifted Pareto distribution (with the sign of D depending on which tail is considered).

2.3 Cross-section

Following the bias based on a single observation, we examine how the Hill statistic fares for multiple observations in a cross-section with only idiosyncratic shocks X_j. Suppose that the X_j satisfy (7) and that the tail indices α and θ are equal. But the scale parameters C and D may differ.[3] Thus consider

$$G_j(x) = 1 - C_j x^{-\alpha}[1 + D_j x^{-\theta} + o(1)]. \tag{8}$$

Let $K = k \leqslant m$ be the number of elements in the cross-section that exceed u. The expected value of the Hill statistic in the cross-section is

$$E[\frac{1}{\hat{\alpha}}|K = k] \simeq \frac{1}{\alpha} - \frac{\theta}{\alpha(\alpha+\theta)} (\frac{1}{k} \sum_{j=1}^{k} D_j) u^{-\theta},$$

where $j = 1, ..., k$ are the elements of X_j that exceed u. The difference with the case of a single observation is that the bias term now contains the average of the second-order scale coefficients. Also note that the first-order scale coefficients C_j do not play a role.

[3]In Appendix 7.3, we relax the assumption that the powers are the same. In large samples, the X_j with the lowest α_j dominate. 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. Einmahl and He (2020), show that the tail-index estimates remain consistent if the scale coefficients differ.

Furthermore suppose, that each entity j is influenced by an individual factor h_j (non-stochastic from the point of view of the cross section). Hence, the linear model becomes $Y_j = h_j + X_j$. To estimate the tail index of Y_j we again use the Hill estimator from (2). The introduction of h_j now has a slightly more complicated effect on the value of the estimate. In particular, h_j affects the distribution of Y_j in the tail area as follows

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

A Taylor approximation yields

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

The question then 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 $\theta = 1$ the new second-order term is $(\alpha h_j + D_j)x^{-\alpha-1}$, while for $\theta > 1$, the new second-order term is $\alpha h_j x^{-\alpha-1}$. Denote the Hill estimate of Y_j by $1/\hat{\alpha}^Y$. We have the following intermediate result:

Lemma 1. The conditional expectation up to a second-order term with a shift factor h_j is as follows

$$E[\frac{1}{k}\sum_{j=1}^{k} \ln \frac{Y_j}{u} | Y_j > u] = E[\frac{1}{\hat{\alpha}^Y} | Y_j > u] \simeq \begin{cases} \frac{1}{\alpha} - \frac{\theta}{\alpha(\alpha+\theta)} \frac{1}{k} \sum_{j=1}^{k} D_j u^{-\theta}, & \text{if } \theta < 1 \\ \frac{1}{\alpha} - \frac{1}{\alpha(\alpha+1)} \frac{1}{k} \sum_{j=1}^{k} (D_j + \alpha h_j) u^{-1}, & \text{if } \theta = 1 \\ \frac{1}{\alpha} - \frac{1}{\alpha+1} \frac{1}{k} \sum_{j=1}^{k} h_j u^{-1}, & \text{if } \theta > 1 \end{cases}$$

Note that if $\theta > 1$ the tail-index of the second-order term changes. This affects the bias of the Hill estimator on Y_j.

Denote the average of the k shift factors of the Y_j 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 and if $\theta \geqslant 1$, the Hill statistic **declines** if $\overline{h}$ increases, since

$$\partial E[\frac{1}{k}\sum_{j=1}^{k} \ln \frac{Y_j}{u} | Y_j > u] / \partial \overline{h} \approx -\frac{1}{\alpha+1} u^{-1} < 0. \tag{9}$$

Thus if a single factor increases, this affects all Y_j with a positive coefficient γ_j 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

$$\partial E[\frac{1}{k}\sum_{j=1}^{k}\ln\frac{Y_j}{u}|Y_j > u]/\partial g \approx -\frac{\overline{\gamma}}{\alpha+1}u^{-1} \tag{10}$$

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 and if $\theta \geqslant 1$, the Hill statistic **increases** if $\overline{h}$ increases, since

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

Note that while the bias in the upper tail is negative, the bias in the lower tail is positive. This implies a negative (positive) correlation between the factor and lower (upper) tail index estimate (inverse of the Hill estimate). Moreover, the bias also generates a negative correlation between the left and the right tail-index estimates.[4]

Given the coefficients γ_{ij}, movement in the factors induce changes in the cross-sectional Hill estimates over time. Even if the coefficients are constant over time, the Y_j (and therefore the γ_{ij}) that are included in the Hill estimate at any time is random. This implies a variation in the included γ_{ij}. Thus the bias varies over time, due to changes in the g_i and due to the fact that different coefficients γ_{ij} enter the estimator at each point in time. Suppose that the γ_{ij} at some points in time have a sign opposite to the sign of factor g_i, this may even lead to a sign reversal of the correlation between Y_j and the factor g_i.

Suppose one wants to identify the contribution of the factors to the bias. This can be done, to some extent, by estimating the γ_{ij} and deducting the sum of the $\hat{\gamma}_{ij}g_i$ from the Y_j. Assume that the parameters γ_{ij} are constant over time and can be

[4] A third-order expansion provided in Appendix 7.2 reveals that the third-order term has the same sign for the left and right tail. This implies that, even though the second-order term dominates the correlation, the biases in the left and right tail are likely not perfectly negatively correlated.

recovered from a time-series regression. The estimate of idiosyncratic noise from the linear factor model reads:

$$\hat{X}_j = Y_j - \sum_{i=1}^{n} \hat{\gamma}_{ij} g_i.$$

Consider the (estimated) bias contribution of an individual factor, say g_f. To this end define the semi-residual with respect to g_f:

$$S_j^f = Y_j - \sum_{i \neq f} \hat{\gamma}_{ij} g_i. \tag{11}$$

Thus S_j^f contains the estimated contribution of the remaining factor $\gamma_f g_f$ and X_j.

Assume that the distribution of X_j satisfies (8) and $\theta > 1$, then:

$$\begin{aligned} E[\frac{1}{\hat{\alpha}^{\hat{X}_j}} | \hat{X}_t > u] &= E[\frac{1}{k} \sum_{j=1}^{k} \ln \frac{\hat{X}_j}{u} | \hat{X}_j > u] \\ &= \frac{1}{\alpha} - \frac{1}{\alpha + 1} \left(\frac{1}{k} \sum_{j=1}^{k} \left[\sum_{i=1}^{n} (\gamma_{ij} - \hat{\gamma}_{ij}) g_i \right] \right) u^{-1} + o(u^{-1}). \end{aligned}$$

A result with an opposite sign for the bias applies to the other tail. Note that in a correctly specified model some bias due to other factors nevertheless remains due to estimation error in the coefficients γ_{ij}.

If $\theta > 1$, the bias in the tail index-estimates of the semi-residuals is as follows:

$$\begin{aligned} E[\frac{1}{\hat{\alpha}^{S^f}} | S_j^f > u] &= E[\frac{1}{k} \sum_{j=1}^{k} \ln \frac{S_j^f}{u} | S_j^f > u] \\ &= \frac{1}{\alpha} - \frac{1}{\alpha + 1} \left(\frac{1}{k} \sum_{j=1}^{k} \left[\sum_{i \neq f} (\gamma_{ij} - \hat{\gamma}_{ij}) g_i + \gamma_{fj} g_f \right] \right) u^{-1} + o(u^{-1}). \end{aligned}$$

This partially isolates the influence of $\gamma_{fj} g_f$ on the variation in $\hat{\alpha}^Y$. Furthermore, the effect of the factors on the tail-index estimate is expected to be in a different direction for the left and right tail index estimates, implying a negative correlation between the two.

The contribution to the bias of g_f can be approximated by the difference between the tail-index estimate of the semi-residual and the tail index of the idiosyncratic noise:

$$\frac{1}{\hat{\alpha}^f} = \frac{1}{\hat{\alpha}^{S^f}} - \frac{1}{\hat{\alpha}^{\hat{X}_j}}. \tag{12}$$

If one ignores the estimation error in the $\hat{\gamma}_{ij}$ and $\theta > 1$, then the expected difference is approximately equal to

$$\begin{aligned} E[\frac{1}{\hat{\alpha}^f}|S_j^f, \hat{X}_j > u] &= E[\frac{1}{\hat{\alpha}^{S^f}} - \frac{1}{\hat{\alpha}^{\hat{X}_j}}|S_j^f, \hat{X}_j > u] \\ &\simeq -\frac{1}{1+\alpha}\left(\frac{1}{k}\sum_{j=1}^{k}\gamma_{fj}g_f\right)u^{-1} + o(u^{-1}). \end{aligned} \tag{13}$$

The RHS in (13) isolates the bias contributed by a single factor, with a different sign for the other tail. Given positive coefficients γ_{fj}, one can expect a negative correlation between $1/\hat{\alpha}^f$ and g_f in the right tail. Recall that α has the intuitive interpretation as the number of bounded moments. For this reason, we report in the empirical application how α relates to a factor. Note that this flips the sign predictions for the bias contribution of a factor in the left and the right tail.

From the above (9), (10) and (13) it can be noted that the size of the bias diminishes in the size of the threshold u. This also opens the possibility to reduce the bias by increasing the threshold u. For this reason we report correlation estimates at two different thresholds, at 5% (the conventionally used threshold) and at 0.5% of the sample fraction. There is a limit to how deep one can go into the tail area, i.e. how high one can take u, since with too few observations the tail-index estimates become highly variable. That is why we consider two alternative methods of bias correction based on the above sign predictions.

2.4 Bias correction in Hill estimates

The characterization of the bias in the Hill estimates specified in (5) and (6) suggests two potential methods of bias correction. First, under tail symmetry one can exploit the opposite sign of the bias in the left and the right tail. The average of the two cross-sectional Hill estimates ($\hat{\alpha}^{mirror}$) could reduce the bias due to the factor structure. If one is unsure about tail symmetry, one may also reduce the bias on a per-tail basis by removing the mean from the dependent variable, i.e. $Y_j - E[Y_j]$.

2.4.1 Exploiting the mirror image

To a first order the bias in the cross-sectional Hill estimates is caused by the contribution of the factor realizations (provided that $\theta > 1$). As we show in (5) and (6), the bias terms are their mirror image under tail symmetry. This gives an opportunity for bias reduction by taking the average of the left and right tail-index estimates.

Consider a single factor model. Without bias correction the Hill estimate for the upper tail has the following asymptotically normal distribution:

$$\sqrt{k}\Big(\frac{1}{\hat{\alpha}_+} - \frac{1}{\alpha}\Big) \sim \mathcal{N}\Big(- \frac{\overline{\gamma}_j g}{1+\alpha} u^{-1}; \frac{1}{\alpha^2}\Big)$$

and where k is the number of intermediate order statistics such that $k \to \infty$, while $k/m \to 0$ and $m \to \infty$. In the lower tail the bias has the opposite sign. If k increases at too high a rate, this diminishes the bias but raises the variance. Conversely, if k increases more slowly, the bias dominates asymptotically. The typical approach in tail index estimation tries to strike a balance between the two vices.

If the tails of the distribution of the X_j are symmetric, taking the average of the left tail estimate $1/\hat{\alpha}_-$ and the right tail estimate $1/\hat{\alpha}_+$ yields a new estimator with an asymptotic normal distribution:

$$\sqrt{k}\Big(\frac{1}{\hat{\alpha}^{mirror}} - \frac{1}{\alpha}\Big) = \sqrt{k}\Big(\frac{{}^1\!/\!\hat{\alpha}_- + {}^1\!/\!\hat{\alpha}_+}{2} - \frac{1}{\alpha}\Big) \sim \mathcal{N}\Big(\frac{1}{2}\frac{(\overline{\gamma}_{j-} - \overline{\gamma}_{j+})g}{1+\alpha} u^{-1}; \frac{1}{2\alpha^2}\Big), \tag{14}$$

where $\overline{\gamma}_{j-}$ and $\overline{\gamma}_{j+}$ are the average of the coefficients found in the left and right tail area, respectively. The mirror method exploits that in a large cross-section any γ_j has an equal probability of appearing in either tail if the idiosyncratic shock X_j dominates the tail behavior of Y_j, so that the average difference appearing in (14) is small. Due to the law of large numbers, as $k \to \infty$, asymptotically the difference goes to zero and hence the asymptotic bias will be of lower order.[5] Thus there is a double benefit of taking the average, the bias is reduced and the variance is halved.

2.4.2 Exploiting the cross-sectional mean

If the tail indices on the left and the right side of the distribution differ, the averaging procedure described above is less meaningful. In this case, assume again that second-order tail index $\theta > 1$. If the bias is caused by a factor shift, one can try to correct for

[5] Due to the contribution of higher-order terms in the tail expansion of the idiosyncratic noise terms, X_j, some bias will remain; see Appendix 7.2.

the bias by shifting the data back in the other direction (we refer to this procedure as the shift method). This is similar to the procedure outlined in Ivette Gomes and Oliveira (2003). Consider the following revised estimator:

$$\frac{1}{\hat{\alpha}^{Y-\overline{Y}}} = \frac{1}{k}\sum_{j=1}^{k}\ln(\frac{Y_j - \overline{Y}}{u}).$$

The idea is that by subtracting the average of the cross-section of Y_j, factor contributions that generate the bias are more or less netted out. But the variance remains as $1/\alpha^2$. For ease of presentation, we derive the bias for this estimator under the assumption of a one factor model, i.e. $i = 1$. The average in the cross-section is

$$\overline{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j = \frac{1}{m}\sum_{j=1}^{m}(\gamma_j g + X_j).$$

Note that this average is approximately equal to $(1/m)(\sum_{j=1}^{m}\gamma_j)g$, as the idiosyncratic noise is assumed to have mean zero. By shifting the observations in the cross-section by the average factor realization in the cross-section we get

$$Y_j - \overline{Y} = (\gamma_j - \frac{1}{m}\sum_{j=1}^{m}\gamma_j)g + X_j - \frac{1}{m}\sum_{j=1}^{m}X_j.$$

The first term containing the factor is non-stochastic in the cross-section.

The second term contains the random elements. Our assumption is that the X_j are i.i.d. and at large s

$$Pr\{X_j \leqslant -s\} \sim C_j s^{-\alpha}.$$

If the X_j exhibit the same tail behavior, it follows from Feller's convolution theorem (1971, Section VIII, 8) that the linear combination exhibits the same tail behavior. Thus the sum also declines by a power of α. But the scale parameter as in (7) changes, resulting in

$$Pr\Big\{\frac{1}{m}\sum_{j=1}^{m}X_j \leqslant -s\Big\} \sim m^{-\alpha}\Big(\sum_{j=1}^{m}C_j s^{-\alpha}\Big).$$

This specification allows us to use (6) to derive the bias in the left tail, by substituting (where $\overline{\gamma} = (1/m)\sum_{j=1}^{m}\gamma_j$):

$$h_j = (\gamma_j - \overline{\gamma})g.$$

This yields a bias in the left tail as

$$E[\ln \frac{Y - \overline{Y}}{u} | Y - \overline{Y} \leqslant -u] \simeq \frac{1}{\alpha} + \frac{1}{\alpha + 1} \Big(\frac{1}{k} \sum_{j=1}^{k} (\gamma_j - \overline{\gamma}) g \Big) u^{-1}. \tag{15}$$

Recall that the bias in a one factor model in the left tail before bias correction equals

$$\frac{1}{\alpha + 1} \Big(\frac{1}{k} \sum_{j=1}^{k} \gamma_j g \Big) u^{-1}. \tag{16}$$

Note that the elements γ_j that enter into (16) may not necessarily enter into (15), as due to the fixed threshold, the shift changes the number of k elements that enter into the estimator (15). The efficacy of correcting the bias by the cross-sectional mean therefore depends on how well the average of all m factor coefficients approximate the average of the k factor coefficients that are part of the tail observations.

3 Simulations

To investigate the efficacy of the above methods of bias reduction, we conduct simulations. In the first model we simulate a cross-section by adding a deterministic constant (positive or negative) to a heavy-tailed innovations term. The second approach simulates from a linear factor model, where factor g is multiplied with a coefficient specific to entity j.

Thus, in the first model data is simulated from

$$Y_j = h + X_j, \tag{17}$$

where h is constant and $j = 1, ..., 20000$ signifies the cross-section. The X_j are heavy-tailed innovations drawn from a Student-t distribution with 3 degrees of freedom.[6]

In the second model, data is simulated from

$$Y_j = \gamma_j g + X_j, \tag{18}$$

[6] Note that in finite samples, bias will remain due to the fact that we simulate from a Student-t distribution; the Hill estimator is only unbiased when data is simulated from the Pareto distribution.

where $X_j \sim$ Student-t and where $\gamma_j \sim \mathcal{N}(1, 0.5^2)$.[7] We again take $j = 1, ..., 20000$.

The above can be used to study the efficacy of the mirror method of bias reduction. To examine the efficacy of the shift method in the case of tail asymmetry, we simulate from a Student-t (2.5) and a Student-t (3.5) for the left and right side of the distribution, respectively. In the calculation of Hill estimates, we set the threshold u at 4.54, which corresponds to the 99% quantile for a Student-t (3) distribution.

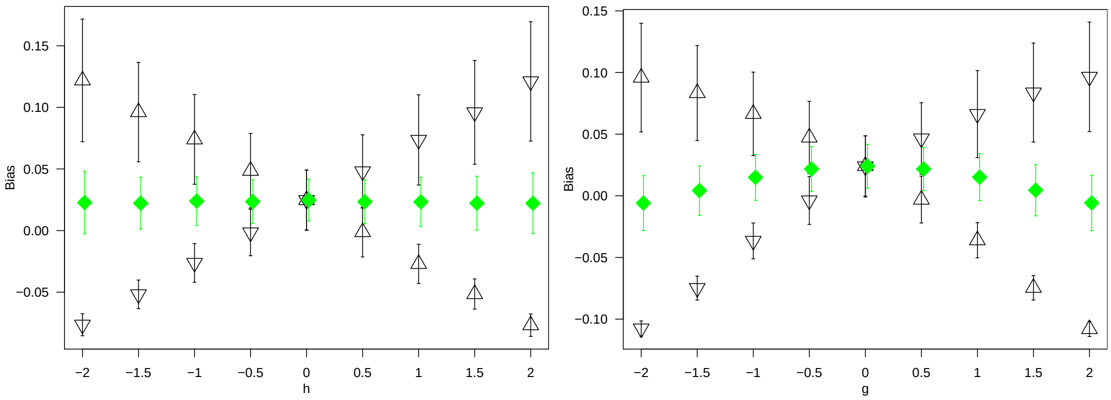
3.1 Mirror method

Figure 1 presents cross-sectional Hill estimates on the simulated data for the left and the right tail with the inclusion of bias-corrected estimates using the mirror method. Panel (a) shows the Hill estimates of the model with deterministic shift described in (17). The figure confirms the linear relationship between the value of h and the bias in Hill estimates as derived in (5) and (6). The value of the bias in left (right) tail-index estimates is signified by the downward (upward) pointing triangles. In the left tail an increase in h produces an increased bias, as shown in (6). This opposite relationship can be observed for the right tail, as in (5). Since the Hill estimates change at the same (absolute) linear rate in the left and the right tail, bias correction using the mirror method produces estimates (green diamonds) that are close to being unbiased. Furthermore, the bars indicate that the variances of $\hat{\alpha}_{-}^{Y}$ and $\hat{\alpha}_{+}^{Y}$ differ. This is due to the fact that the bias shifts the estimates away from α. As a result, the mirror method reduces the average of these two variances. Panel (b) presents the Hill estimates of the linear model described in (18). For the model with random shift, the relationship between the value of the bias and g is no longer linear. Instead, the bias also depends on the values of the coefficients γ_j that enter through the observations in the tail. However, their influence is moderate to small on the efficiency of the mirror method.

Using the relationship between (5) and (6) to remove the cross-sectional location shift works under tail symmetry only. Under tail asymmetry there are three problems that surface. The obvious one is that the estimate neither reflects the left nor the right tail index. The more subtle effect is that $1/(1 + \alpha)$ in (5) and (6) are different in the left and the right tail, i.e. α_{-} and α_{+} differ. This induces a different effect of h in

[7] To bring the simulations closer to our empirical application we also draw γ_j from a subsample of estimated CAPM betas ($n = 13,535$) for US stocks obtained from the CRSP database. The distribution of the CAPM betas has mean 1.14 and standard deviation 0.72. The unreported results are very similar and are available on request.

Figure 1: Bias correction for $1/\hat{\alpha}$ using the mirror method



(a) Constant model (b) Linear model

This figure presents the results of bias correction using the mirror method discussed in Section 2.4.1. In panel (a) and (b), data is simulated from the models described in (17) and (18), respectively, where the X_j are drawn from a Student-t (3). The x-axis gives different values of the deterministic constant h in panel (a) and different values of factor g in panel (b). The y-axis indicates the bias by subtracting 1/3 from the Hill estimates with threshold $u = 4.54$ (99% quantile of the Student-t (3)). The upward pointing black triangles ($\triangle$) show the bias of uncorrected Hill estimates in the right tail, the downward pointing black triangles ($\triangledown$) show the bias of uncorrected Hill estimates in the left tail and the green diamonds (♦) show the bias of corrected Hill estimates using the mirror method. The bars surrounding the mean estimates present the standard deviation of the estimates.

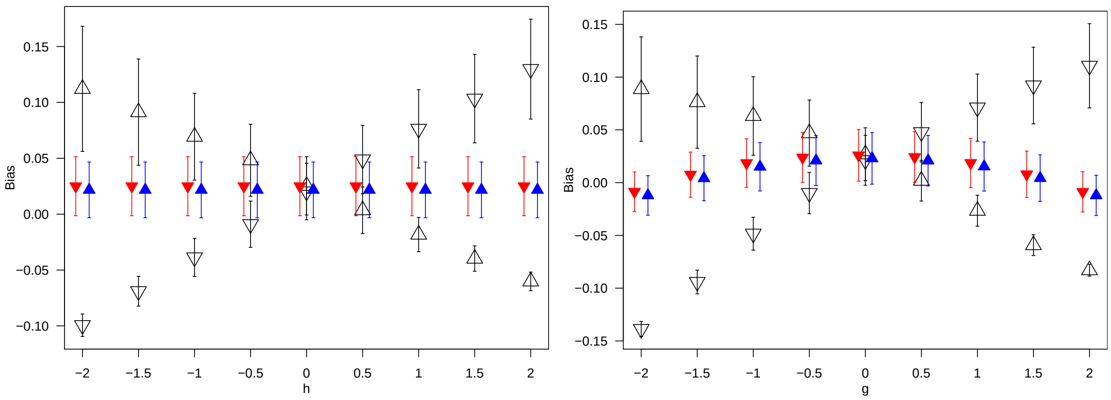
the left and right tail. Additionally, the optimal threshold u may differ for the left and right tail. The optimal threshold u varies inversely with the tail index α, thus it may not be optimal to use the same threshold for both tails.

3.2 Shift method

The cross-sectional Hill estimates for the case with asymmetric tails is presented in Figure 2. Panel (a) shows the Hill estimates for the model with deterministic shift described in (17). The bias-corrected estimates using the shift method for the left (right) tail are indicated by the red downward (blue upward) pointing arrows. Bias correction using the shift method also produces estimates with near zero bias, despite the asymmetry between the bias in Hill estimates on the left and the right tail.

Panel (b) presents the Hill estimates of the linear model described in (18). Due to the inclusion of a random shift term, the value of the bias has again become dependent on the values of the coefficients in the tail. By moving from a deterministic shift to a random shift the efficacy of bias correction decreases somewhat, especially for the estimate in the left tail. However, bias correction remains highly effective, producing tail estimates with far smaller bias than the uncorrected estimates.

Figure 2: Bias correction using shift method



(a) Constant model

(b) Linear model

This figure presents the results of bias correction using the shift method discussed in Section 2.4.2. In panel (a) and (b), data is simulated from the models described in (17) and (18), respectively. The X_j innovations are drawn from a Student-t (2.5) and (3.5) for the left and right tail respectively. The x-axis gives different values of the deterministic constant h in panel (a) and different values of factor g in panel (b). The y-axis indicates the bias by subtracting 1/2.5 and 1/3.5 from the Hill estimates in the left and right tail respectively. The threshold is set at $u = 4.54$ (99% quantile of the Student-t (3)). The upward pointing black triangles (△) show the bias of uncorrected Hill estimates in the right tail, the downward pointing black triangles (▽) show the bias of uncorrected Hill estimates in the left tail. The blue upward pointing triangles (▲) show the bias of the corrected Hill estimates using the cross-sectional mean in the right tail. The red downward pointing triangles (▼) show the bias of the corrected Hill estimates using the cross-sectional mean in the left tail. The bars surrounding the mean estimates present the standard deviation of the estimates.

4 Data

To test for the presence, direction and size of the bias, in real world data, we use monthly US stock returns and annual US Census county population data. We also test the two above outlined methods for bias correction on US stock returns. Both datasets are known to exhibit power law behavior. Moreover, the data are sufficiently rich in both the time series and cross-sectional dimension to investigate the efficacy of our methods.

4.1 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.[8]

[8] We exclude stocks with a price below 5 dollars, as is the standard in the asset pricing literature, noting that their inclusion leads to almost identical results. Stocks with exchange codes -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.

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 (log) returns from 1963 to 2019.

There is a large body of literature which uses co-movement between excess returns and factors to explain the cross-sectional variation in expected excess stock returns. The combination of the rich dimensions of the data and the theoretical and empirical backing for a factor structure in stock returns provides an exemplary 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 7 in Appendix 7.4, 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]

4.2 County population data

Another heavily researched field in power laws is the geographical distribution of population. The US Census Bureau has collected county population statistics since 1970. Analysis on the county level offers 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 monthly US stock data, this data has a length of 46. 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 (log) growth rate of the population in line with existing literature.[10] For the creation of population change, we use the Federal Information Processing Standards (FIPS) code which uniquely identifies counties and

[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. Table 8 presents the pairwise correlations between the different factors.

[10] Most studies focus on the upper tail of geographical density of population. Due to data limitations, only recently have studies (Devadoss and Luckstead, 2016; Ioannides and Skouras, 2013) shown that the left tail also adheres to power law behavior.

county equivalents in the United States. The documentation on a clear factor structure is notably weaker than for stock returns. 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 what constitutes the best combination of factors, we conduct a 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 9 in Appendix 7.4. By using five principal components (PCs), we avoid multicollinearity and over-fitting, which is likely to arise in a model with many explanatory variables.

4.3 Empirical implementation

To distinguish between factor values at different instances in time, we now introduce a time index t. To obtain estimates of X_{jt} and semi-residuals S^f_{jt} in (11) for factor f, we run linear time-series regressions to estimate the factor coefficients γ_{ij}. In a linear factor model for stock returns; $Y_{jt} = R_{jt} - r_t$, where R_{jt} is the log return of stock j at time t. Here r_t is the one-month Treasury bill rate which is used as the risk-free rate at time t. In case of county population growth, Y_{jt} is the percentage change in county j's population at time t. Thus we run regressions:

$$Y_{jt} = \sum_{i=1}^{n} \gamma_{ij} g_{it} + X_{jt} \quad \text{for} \quad t = 1, 2, ..., T.$$

We repeat this for all $j = 1, ..., m$ entities. We also use these regressions to construct estimates of the X_{jt}. To estimate the tail index for the different quantities like $\hat{X}_{jt}$ and S^f_{jt} we use the Hill estimator as defined in (2).

The Hill estimator requires a choice of threshold, $u(k)$. We follow common practice of selecting the threshold on order statistic k at a fixed percentage of the sample size. Specifically, we choose k at 5% and 0.5% of the empirical quantile to study the influence of factors in a linear model on estimates of the tail index. Recall that the thresholds selected in this way are MSE consistent estimates of the quantiles at 5% and 0.5% of the cross-section, respectively.

5 Results

5.1 US financial returns

Section 2 demonstrates that the Hill estimator in (2) applied to Y_{jt}, that is $\hat{\alpha}_t^Y$, contains a specific bias caused by underlying factors and coefficients. We investigate this relationship between commonly known asset pricing factors and estimates of the tail index in the cross-section. Table 1 shows the partial correlations between the asset pricing factors and $\hat{\alpha}_t^Y$. The "+" ("-") subscript for the Hill estimate indicates that the estimate is on the right (left) tail of the empirical distribution.

From Table 1, we note that the correlation between tail index estimates and the market factor is particularly strong. This is a first indication of the influence a factor can have in cross-sectional tail estimation. Although somewhat smaller, the correlation for the SMB factor is still pronounced. The correlation between the tail index and the other factors is smaller and the signs are the opposite of what one would initially expect. This may be partly caused by the simultaneous effect that the different factors have on $\hat{\alpha}_t^Y$, see Table 8 in Appendix 7.4 which records the partial correlations between all factors. Furthermore, for a given stock the coefficients for the different factors can vary in size and sign. These issues may dilute the effect of the bias caused by a single factor.

Table 1: Cross-sectional tail index

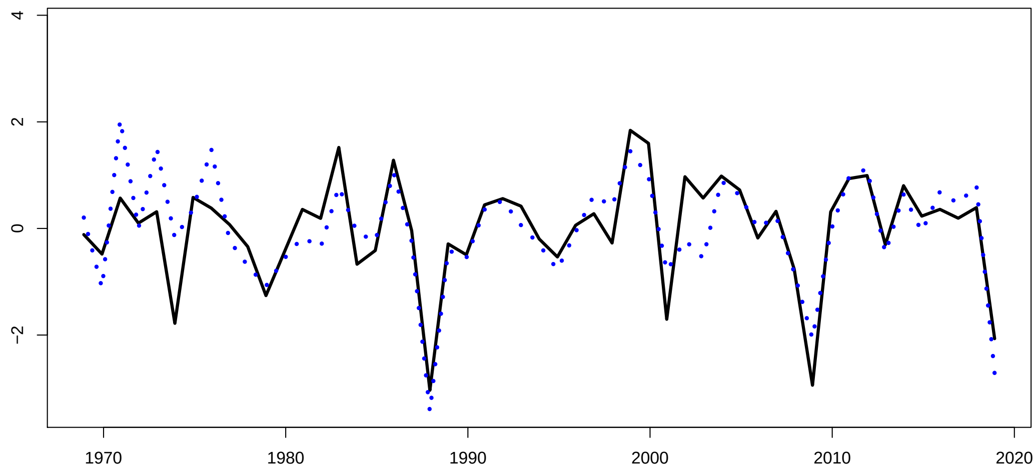
	Market	SMB	HML	MOM	Liq
$\hat{\alpha}_{t-}^Y$	-0.69	-0.45	0.19	0.09	0.05
$\hat{\alpha}_{t+}^Y$	0.75	0.52	-0.12	-0.15	-0.02

This table presents the correlation between cross-sectional Hill estimates on excess returns and the asset pricing factors. The $\hat{\alpha}_{t-}^Y$ and $\hat{\alpha}_{t+}^Y$ are the inverses of the cross-sectional Hill estimates for the cross-section of stock returns for the left and right tail, respectively. The threshold u is set to 5% of the sample fraction. The factor with which the correlation is calculated is reported in the columns.

To isolate the bias that a single factor induces in $\hat{\alpha}_t^Y$, we use $\hat{\alpha}_{t+}^f$ as defined in (12). Figure 3 plots the (normalized) time-series of the market factor and $\hat{\alpha}_{t+}^M$ graphically. The time series illustrates the clear positive relationship between $\hat{\alpha}_{t+}^M$ and the market factor, as predicted by the inverse of the expectation in (13).

In Section 2, we derived that a shift in heavy-tailed variables induces a bias in left and the right tail-index estimates of opposite sign. To investigate whether this is indeed the case for the most dominant factor, we plot the $\hat{\alpha}_{t+}^M$ and $\hat{\alpha}_{t-}^M$ in Figure 4. The two estimates indeed appear to be each others mirror image. This confirms the predictions from Proposition 1 and Proposition 2. Figure 7 in the Appendix

Figure 3: M_t vs $\hat{\alpha}^M_{t+}$



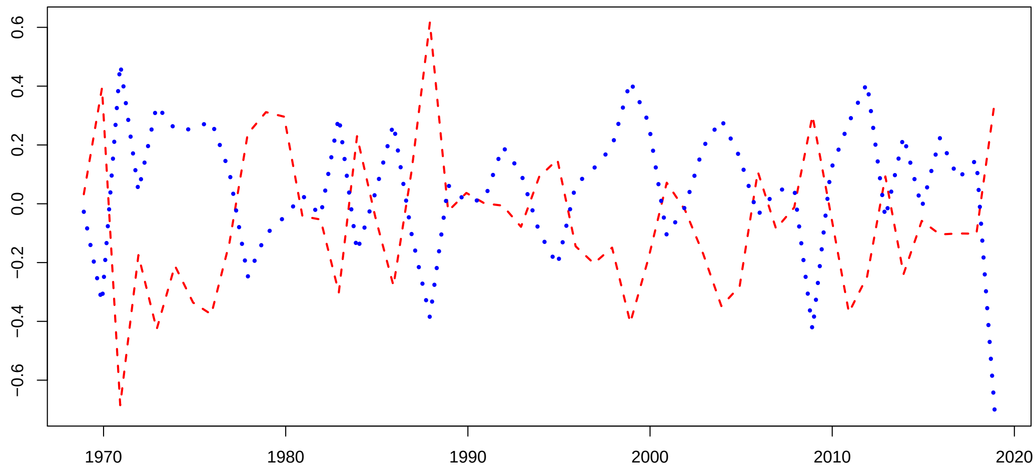
This figure presents the time series of the market factor (solid black line) and its isolated effect on $\hat{\alpha}^M_{t+}$ (blue dotted line). The time-series of the market factor and $\hat{\alpha}^M_{t+}$ are normalized. Furthermore, the data has been annualized by averaging the monthly estimates for a given year.

presents the same figures for the SMB, HML, momentum and the liquidity factors. While weaker, the relationship for the SMB and HML factors still shows substantial negative co-movement between the estimate of the tail index in the right and the left tail during some periods. The momentum and liquidity factors show a somewhat weaker pattern. The relationship between the tail index and the factors hinges on the validity of the factor structure, i.e. the relative importance of factors 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.[11]

We summarize the patterns observed in Figure 3 and 4 for all factors by means of correlations in Table 2. In Table 2 the first two rows show that isolating the contribution of a specific factor leads to a higher correlation between the tail index estimates and the factor. This implies that the interaction between the factors obfuscates the relationships as shown in Table 1. The correlations for the market and SMB factor are substantial. Isolating the effect of the HML factor changes the correlation in the predicted direction for both the left and right tail of the distribution, cf. Table 1. The correlations in Table 2 for the momentum factor have a sign that is somewhat counter intuitive. One possibility that could generate the counter-intuitive sign is that the observations included in the tail measurement have negative coefficients. In

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

Figure 4: $\hat{\alpha}^M_{t+}$ vs $\hat{\alpha}^M_{t-}$



This figure presents the time series of the isolated effect of the market factor on the cross-sectional Hill estimator, $\hat{\alpha}^M_{t-(+)}$ defined in (13), for the left (right) tail. The right tail $\hat{\alpha}^M_{t+}$ (blue dotted line) is contrasted with left tail estimate $\hat{\alpha}^M_{t-}$ (red dashed line). The presented data is annualized by averaging the monthly estimates for a given year. In contrast to Figure 3, these series are not normalized.

unreported results we multiply the factor by the average coefficients found for the tail observations each month, which alters the signs in the predicted direction.

Table 2: Correlations cross-sectional tail index

	Market	SMB	HML	MOM	Liq
$\hat{\alpha}^f_{t-}$	-0.81	-0.81	-0.24	0.47	-0.05
$\hat{\alpha}^f_{t+}$	0.85	0.85	0.39	-0.42	-0.04
$\rho(\hat{\alpha}^f_{t-}, \hat{\alpha}^f_{t+})$	-0.91	-0.83	-0.07	-0.28	-0.02

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

	Market	SMB	HML	MOM	Liq
$\hat{\alpha}^f_{t-}$	-0.51	-0.49	0.04	0.25	0.01
$\hat{\alpha}^f_{t+}$	0.54	0.39	0.07	0.13	-0.02
$\rho(\hat{\alpha}^f_{t-}, \hat{\alpha}^f_{t+})$	-0.39	-0.24	-0.11	0.15	0.03

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

This table reports the correlations between the isolated effect of factors on the cross-sectional Hill estimates and an individual factor. In the first and second row of each panel, $\hat{\alpha}^f_t$ is the cross-sectional tail index estimate where factor f's effect is isolated, as defined in (12). The sign "+" ("-"), indicates that the estimate is made on the right (left) tail of the distribution. The factor with which the correlation is calculated is reported in the column. The **last** row shows the correlation between the left and right tail estimates of $\hat{\alpha}^f_t$ for the respective factors.

The implication of Propositions 1 and 2 is that $\hat{\alpha}^f_{t-}$ and $\hat{\alpha}^f_{t+}$ should be each others mirror image and their correlation is therefore expected to be negative. The negative sign in the last row of panel (a) 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, as is apparent from Figure 4 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 to the quality of these factors as proxies for factors in the underlying DGP.

Equation (13) implies that the bias originating from the factors in $\hat{\alpha}_{t-}^{Y}$ diminishes as threshold u increases. In panel (b) of Table 2, we lower the percentage of the sample fraction used in the tail estimation to 0.5%. This indeed leads to a sharp decrease in the correlations between the factors and the tail-index estimates for most factors.

In Table 3, we report on regressions that investigate the degree to which variation in the isolated bias $\hat{\alpha}_{t+}^{f}$ explains variation in the cross-sectional Hill estimate of Y_t, i.e. $\hat{\alpha}_{t+}^{Y}$. Panel (a) shows results of these regressions for the right tail-index estimate. The coefficients for the Market and SMB factors are significantly different from zero, resonating the strong correlations in Table 2. The R^2 of the first regression shows that about 42% of the variation in the cross-sectional tail index is driven by the market factor in the right tail.[12] The second most important factor is the SMB factor which contributes about 12% to the variation in $\hat{\alpha}_{t+}^{Y}$. The HML, momentum and liquidity factors have a marginal role in explaining the variation in $\hat{\alpha}_{t+}^{Y}$. Similar regressions results for the left tail are reported in panel (b). The contribution of the individual factors are quantitatively similar in the left tail.

The explanatory power of the idiosyncratic part of the linear factor model explains only about 7% of the variation in the right tail. This suggests that indeed most variation in cross-sectional tail-index estimates stems from variation in the factor realizations. This is somewhat different for the left tail. The R^2 of the regression with the idiosyncratic factor is 24%. Aside from correlated measurement errors and variation in α, we may not have isolated all the factors that influence the left tail realisations.

In the last column, we report the results of a multiple regression to investigate the contribution of all individual factors together on $\hat{\alpha}_{t+}^{Y}$. The contribution of all factors is significant and produces a high R^2 of 64%. This suggests that each factor contributes significantly to the bias, even when considering the correlation amongst the factors. The results for the left tail in panel (b) are quite similar to the results for the right tail.

[12] See Table 10 in the Appendix for regression results for the factors (instead of for the $\hat{\alpha}_{t}^{f}$).

Table 3: Regression cross-sectional tail index

$\hat{\alpha}^{M}_{t+}$	0.95***						1.09***
	(0.05)						(0.04)
$\hat{\alpha}^{SMB}_{t+}$		0.90***					1.22***
		(0.10)					(0.07)
$\hat{\alpha}^{HML}_{t+}$			0.31				−0.69***
			(0.27)				(0.17)
$\hat{\alpha}^{MOM}_{t+}$				−0.40			0.41**
				(0.31)			(0.20)
$\hat{\alpha}^{Liq}_{t+}$					−0.25		−0.53**
					(0.40)		(0.25)
$\hat{\alpha}^{X}_{t+}$						0.52***	
						(0.08)	
Constant	2.57***	2.58***	2.57***	2.56***	2.56***	1.20***	2.59***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.21)	(0.02)
R^2	0.42	0.12	0.002	0.003	0.001	0.07	0.64

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

$\hat{\alpha}^{M}_{t-}$	0.96***						1.16***
	(0.05)						(0.05)
$\hat{\alpha}^{SMB}_{t-}$		0.82***					1.23***
		(0.10)					(0.07)
$\hat{\alpha}^{HML}_{t-}$			0.001				−1.26***
			(0.30)				(0.21)
$\hat{\alpha}^{MOM}_{t-}$				−0.46			−0.50**
				(0.32)			(0.23)
$\hat{\alpha}^{Liq}_{t-}$					0.79**		0.36
					(0.38)		(0.27)
$\hat{\alpha}^{X}_{t-}$						0.78***	
						(0.06)	
Constant	2.76***	2.77***	2.76***	2.75***	2.76***	0.58***	2.76***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.16)	(0.02)
R^2	0.33	0.10	0.00	0.003	0.01	0.24	0.55

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

This table presents the regression results for the effect of the factors on the cross-sectional Hill estimate. The dependent variable in the upper panel is the Hill estimate for the left tail of the raw cross-sectional excess returns ($\hat{\alpha}^{Y}_{t-}$). The independent variable, is the cross-sectional tail index where the factor f's effect is isolated ($\hat{\alpha}^{f}_{t}$). In the sixth row $\hat{\alpha}^{X}_{t-}$ is the tail index estimated on the estimated idiosyncratic noise terms of the five-factor asset pricing model. Panel (b) illustrates the results for the right tail of the distribution. The threshold u used to estimate the Hill estimate is set to 5% of the sample fraction. The stars in the table indicate: * p<0.1; ** p<0.05; *** p<0.01.

Table 11 in Appendix 7.4 presents the regression results for a threshold based on the 0.5% sample fraction. We find that only a small share of the variation in $\hat{\alpha}^{Y}_{t}$ is explained by the individual factors. The role of the bias caused by the factor has diminished by looking deeper into the tail. In this case, variation in $\hat{\alpha}_t^X$ explains about 58% for the left and 51% for the right tail of the variation in $\hat{\alpha}_t^Y$. Abstracting away from correlated measurement errors, the increase in R^2 strongly suggests that the role of known and unknown factors in the bias has diminished. Therefore, $\hat{\alpha}_t^X$ captures variation in $\hat{\alpha}_t^Y$ more strongly.

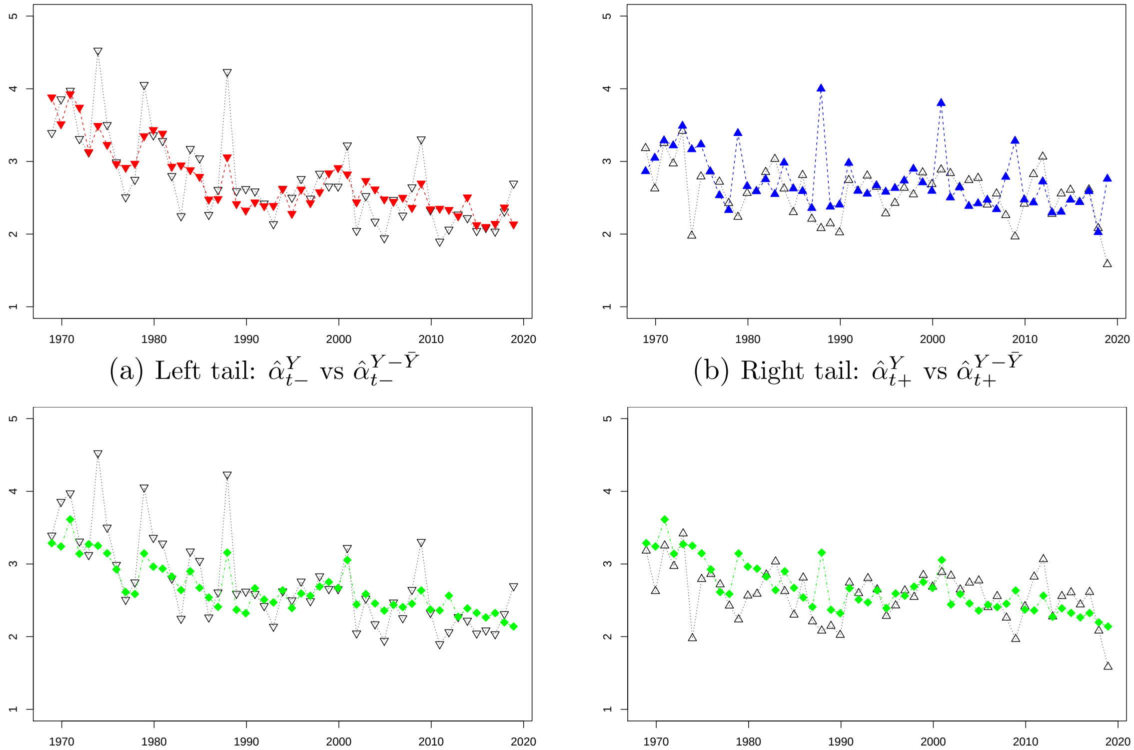
5.1.1 Bias correction for US stock returns

The analysis above provides ample evidence for the type of bias arising from factors. Therefore, we subject the US stock return data to the two bias reduction methods. In Figure 5 we show the results of bias correction, by presenting the time-series of uncorrected tail-index estimates in tandem with bias-corrected estimates.

Panels (a) and (b) in Figure 5 present the results of bias correction using the cross-sectional mean for the left and the right tail, respectively. In panel (a), for the left tail, one notices a number of outliers in the uncorrected tail estimates $\hat{\alpha}_{t-}^Y$. These outliers can be dated at respectively the first and second oil crisis, the Black Monday (1987), the Dot-com bubble burst (2001) and the Credit crises (2008). The bias-corrected estimates $\hat{\alpha}_{t-}^{Y-\bar{Y}}$ for these crisis periods show that the bias correction is substantial. The corrected estimates are more inline with the preceding and succeeding estimates. This can be understood from (6). Due to the negative value of the shift parameter h, capturing the large declines in the market factor during crises periods, the estimated $1/\alpha$ is lowered and hence the estimated α is larger. This also explains why we see the opposite pattern arise around crisis periods on the right tail $\hat{\alpha}_{t+}^Y$ presented in panel (b).

Panel (c) and (d) present the mirror method for bias correction with inclusion of the (inverse) Hill estimate on US stock returns, $\hat{\alpha}_t^Y$, on the left and right tail, respectively. Note that the green diamonds (mirror method estimates) also give a substantial correction during periods of market turmoil. When comparing panel (a) and panel (c), very similar behavior is observed for the mirror estimate $\hat{\alpha}_t^{mirror}$ and the shift method in the left tail $\hat{\alpha}_{t-}^{Y-\bar{Y}}$. For both correction methods, the estimators successfully dampen the effects of large factor realizations. This correspondence is not as clearly observed when comparing $\hat{\alpha}_t^{mirror}$ and the shift estimate in the right tail $\hat{\alpha}_{t+}^{Y-\bar{Y}}$ in panel (d) and (b), respectively. The mirror estimate does not increase to the same degree as the shift estimate in the right tail during economic crises. Figure 6 in the Appendix presents the three bias-corrected estimates in one figure. The figure confirms the close correspondence between $\hat{\alpha}_{t-}^{Y-\bar{Y}}$ and $\hat{\alpha}_t^{mirror}$.

Figure 5: Bias corrected $\hat{\alpha}$



(a) Left tail: $\hat{\alpha}^{Y}_{t-}$ vs $\hat{\alpha}^{Y-\bar{Y}}_{t-}$

(b) Right tail: $\hat{\alpha}^{Y}_{t+}$ vs $\hat{\alpha}^{Y-\bar{Y}}_{t+}$

(c) Left tail: $\hat{\alpha}^{Y}_{t-}$ vs $\hat{\alpha}^{mirror}_{t}$

(d) Right tail: $\hat{\alpha}^{Y}_{t+}$ vs $\hat{\alpha}^{mirror}_{t}$

This figure presents the time series of tail estimates for US stock returns after applying bias reduction methods. The y-axis shows the value of the estimates in terms of α. In panel (a) the red triangles (▼) show the Hill estimates for the left tail of $Y_{jt} - \overline{Y}_t$. The black open downward triangles (▽) are tail estimates extracted from Y_{jt} on the left side of the distribution. In panel (b) the blue triangles (▲) show the Hill estimates for the right tail of $Y_{jt} - \overline{Y}_t$. The black open upward pointing triangles (△) are tail estimates extracted from Y_{jt} on the right side of the distribution. The green diamonds (♦) in panel (c) and (d) are the estimates of the tail index after correcting for bias using the mirror method. The estimates are extracted from the cross-section each month and subsequently averaged within a year for ease of presentation. The threshold is set at the 5% sample fraction.

We are now able to investigate whether the bias, likely caused by some of the factors on $\hat{\alpha}^{Y}_{t}$, is still present after bias correction. To enable a direct comparison between Table 3 and Table 4, the independent variable $\hat{\alpha}^{f}_{t+}$ is left unchanged, while in Table 4 the dependent variable $\hat{\alpha}^{Y}_{t}$ is bias-corrected. Panel (a) in Table 4 gives the effect of the shift method and panel (b) gives the effect of the mirror method.[13]

If the bias correction methods work (i.e. the bias arising due to the factors has reduced), one should observe less significance for the coefficients for the isolated bias of

[13]Table 10 in Appendix 7.4 presents the results on the factors themselves.

Table 4: Bias Correction for US Stock returns

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\hat{\alpha}_{t-}^{M}$	0.02						0.04
	(0.05)						(0.05)
$\hat{\alpha}_{t-}^{SMB}$		−0.13*					−0.12
		(0.08)					(0.08)
$\hat{\alpha}_{t-}^{HML}$			−0.42**				−0.57***
			(0.20)				(0.21)
$\hat{\alpha}_{t-}^{MOM}$				−0.34			−0.45*
				(0.22)			(0.23)
$\hat{\alpha}_{t-}^{Liq}$					0.67***		1.03***
					(0.26)		(0.27)
$\hat{\alpha}_{t-}^{X}$						0.99***	
						(0.02)	
Constant	2.80***	2.80***	2.80***	2.80***	2.80***	0.04	2.80***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.06)	(0.02)
R^2	0.0002	0.01	0.01	0.004	0.01	0.76	0.04

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

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\hat{\alpha}_{t-}^{M}$	0.03						0.07**
	(0.03)						(0.03)
$\hat{\alpha}_{t-}^{SMB}$		0.09*					0.11**
		(0.05)					(0.05)
$\hat{\alpha}_{t-}^{HML}$			−0.22*				−0.35***
			(0.13)				(0.13)
$\hat{\alpha}_{t-}^{MOM}$				−0.06			−0.13
				(0.14)			(0.15)
$\hat{\alpha}_{t-}^{Liq}$					0.25		0.39**
					(0.16)		(0.17)
$\hat{\alpha}_{t-}^{X}$						0.58***	
						(0.02)	
Constant	2.66***	2.66***	2.66***	2.66***	2.66***	1.04***	2.66***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.05)	(0.01)
R^2	0.002	0.01	0.01	0.0003	0.004	0.66	0.03

(b) Mean left and right tail index estimate cross-sectional tail index, i.e. $\hat{\alpha}_{t}^{mirror}$

This table presents the regression results for the two bias reduction methods and the asset pricing factors. In panel (a) the dependent variable is the Hill estimate for the (left) tail of $Y_i - \overline{Y}$. In panel (b) the dependent variable is the average of the left and right tail estimate on the Y_i, i.e. $\hat{\alpha}_t^{mirror}$.The independent variable, is the cross-sectional tail index where the factor's effect is isolated (α_{t+}^{f}). The lower panel shows the results for the right tail of the distribution. The threshold u to estimate the Hill estimate is set to 5% of the sample fraction. The stars in the table indicate: * p<0.1; ** p<0.05; *** p<0.01.

the different individual factors. Moreover, the R^2 of the regressions should decrease. Panel (a) shows that the isolated bias with respect to the market ($\hat{\alpha}_{t-}^{M}$) and SMB factor ($\hat{\alpha}_{t-}^{SMB}$), lose almost all of their explanatory power. While the estimated coefficients for both factors were close to 1 in Table 3, now the estimates have almost become indistinguishable from 0. The R^2 drops from 33% and 10% to 0% and 1% for the isolated bias of the market and SMB factor, respectively. This breaks the link between the bias found in the Hill estimate and the factors. Although the coefficients for α_{t-}^{HML} and α_{t-}^{Liq} have become significant, the explanatory power remains low. The R^2 for the regression of $\hat{\alpha}_t^{Y-\bar{Y}}$ on $\hat{\alpha}_{t-}^{X}$ has increased to 0.76 (and a similar result applies for the mirror method). This indicates a far more intimate relationship between the bias corrected estimate and the true value of the tail index.

The results for $\hat{\alpha}_t^{mirror}$ in panel (b) show that the relationship between the bias and the estimates again decreases significantly. For the original isolated bias of the market and SMB factors, both the coefficients become almost zero. Additionally, the R^2 values decrease to close to zero for all factors. The relationship for the HML, momentum and liquidity factors are largely unchanged.

5.2 County population

Due to the lack of a clear emergent set of factors in the population size literature, we use PCA to extract five 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 summary statistics of the PCs.

In the same vein as for the US stock data, Table 5 presents the correlations for the county population growth data and is comparable to the combination of Tables 1 and 2 combined. We first consider panel (a) where tail index estimates are measured at 5% of the sample fraction. The correlations between $\hat{\alpha}_t^{Y}$ and the PCs is weaker than for the US stock return data. However, as is the case for the data on US stock returns, these correlations are stronger when the effect of the PC's are isolated, depicted in rows three and four. The correlation with the first PC increases in magnitude from -0.45 to -0.83 for the left tail and from -0.07 to 0.82 for the right tail. Since the sign of a PC is indeterminate, one should not interpret the direction of the bias. Finally, the last row indicates that the isolated bias in tail index estimates on the left and right side of the distribution are negatively correlated for all but the second PC.

Panel (b) of Table 5 presents the correlations when the threshold is lowered to 0.5%.

Table 5: Correlations of cross-sectional tail indices (County data)

	PC1	PC2	PC3	PC4	PC5
$\hat{\alpha}^{Y}_{t-}$	-0.45	-0.09	-0.34	0.18	0.28
$\hat{\alpha}^{Y}_{t+}$	-0.07	0.02	0.06	0.01	-0.05
$\hat{\alpha}^{f}_{t-}$	-0.83	-0.22	-0.74	0.75	0.59
$\hat{\alpha}^{f}_{t+}$	0.82	0.57	0.75	-0.68	-0.58
$\rho(\hat{\alpha}^{f}_{t-}, \hat{\alpha}^{f}_{t+})$	-0.71	0.23	-0.64	-0.41	-0.45

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

	PC1	PC2	PC3	PC4	PC5
$\hat{\alpha}^{Y}_{t-}$	-0.18	-0.17	-0.44	0.08	-0.05
$\hat{\alpha}^{Y}_{t+}$	0.07	-0.08	-0.31	-0.08	0.06
$\hat{\alpha}^{f}_{t-}$	-0.66	0.01	-0.39	-0.09	-0.16
$\hat{\alpha}^{f}_{t+}$	0.42	0.05	0.40	0.17	0.20
$\rho(\hat{\alpha}^{f}_{t-}, \hat{\alpha}^{f}_{t+})$	-0.31	-0.24	-0.15	-0.10	-0.10

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

This table reports the correlations between the isolated effects of PCs on the cross-sectional Hill estimates and the PCs themselves. Here $\hat{\alpha}^{Y}_{t-}$ and $\hat{\alpha}^{Y}_{t+}$ are the cross-sectional Hill estimates for the cross-section of county population growth for the left and right tail, respectively. The $\hat{\alpha}^{f}_{t}$, stated in the third and fourth rows of each panel, is the cross-sectional tail index where the effect of the PC's is isolated, as defined in (13). The five factors are the first five principal components from an assortment of variables suggested by the literature. The last row shows the correlation between the left and right tail estimates of $\hat{\alpha}^{f}_{t}$. 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.

As with the data on US stock returns, correlations decrease in magnitude substantially and frequently have signs opposite from those found in panel (a). In the final row, negative correlations are still observed between the isolated bias on the left and the right tail-index estimates but the magnitude has decreased substantially. Thus again, lowering the threshold limits the influence of the factor structure in cross-sectional tail-index estimates.

Table 6 presents the results of the regressions between the isolated bias of the different individual PCs $\hat{\alpha}^{f}_{t}$ and the cross-sectional Hill estimate on Y_t, i.e. $\hat{\alpha}^{Y}_{t}$. Panel (a) illustrates the results when using the estimate of the left cross-sectional tail-index and α^{f}_{t-}, while panel (b) uses the estimate of the right cross-sectional tail-index. We observe that only the isolated bias with respect to PC1 in the left tail ($\hat{\alpha}^{PC1}_{t-}$) can significantly account for variation in the cross-sectional tail-index estimate of county population change ($\hat{\alpha}^{Y}_{t-}$). For the right tail, $\hat{\alpha}^{PC1}_{t+}$ is not significant but attains the highest R^2 of the five principal components in panel (b). This implies that the previously presented correlations for the PCs most likely come with large standard errors. The high R^2 attained by $\hat{\alpha}^{X}_{t-}$ and $\hat{\alpha}^{X}_{t+}$ in the penultimate column, further illustrates the marginal influence of the factor structure on cross-sectional tail index estimates, in other words the idiosyncratic shock X_{jt} is dominant and the factors contribute little towards explaining the dependent variable. The contrast with the strong results for financial return data highlights the role of the strength of the factors in the DGP that drives the bias.[14] Due to the orthogonality of the PCs, the penultimate columns in panel (a) and (b) do not change much relative to the respective univariate regressions.

The foregoing shows, that while county population change may not be perfectly described by a linear factor model, factor variation does bias tail index estimates in the cross-section. Furthermore, the bias tail-index estimates in the right and left tail are negatively correlated. Thus, even when investigating factors with a weak or unclear underlying factor structure, inference on the basis of cross-sectional tail-index may lead to incorrect conclusions.

[14] Table 13 in Appendix 7.4 presents the regression results for a 0.5% threshold. The coefficient for $\hat{\alpha}_{t+}^{PC2}$ becomes significant t this lower threshold. Given that this does not occur for the right tail estimates in panel (b), it is possible this is caused by correlated measurement errors in $\hat{\alpha}_{t+}^{PC2}$ and $\hat{\alpha}_{t+}^{Y}$.

Table 6: Regression cross-sectional tail index (County population growth)

$\hat{\alpha}_{t-}^{PC1}$	0.41*						0.49*
	(0.21)						(0.26)
$\hat{\alpha}_{t-}^{PC2}$		0.13					-0.77
		(0.51)					(0.61)
$\hat{\alpha}_{t-}^{PC3}$			0.32				0.30
			(0.33)				(0.35)
$\hat{\alpha}_{t-}^{PC4}$				0.29			-0.02
				(0.40)			(0.41)
$\hat{\alpha}_{t-}^{PC5}$					0.44		0.53
					(0.38)		(0.40)
$\hat{\alpha}_{t-}^{X}$						0.63***	
						(0.12)	
Constant	2.46***	2.50***	2.50***	2.50***	2.49***	0.90**	2.43***
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.32)	(0.07)
R^2	0.08	0.001	0.02	0.01	0.03	0.38	0.15

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

$\hat{\alpha}_{t+}^{PC1}$	-0.18						-0.21
	(0.18)						(0.19)
$\hat{\alpha}_{t+}^{PC2}$		0.006					0.02
		(0.29)					(0.35)
$\hat{\alpha}_{t+}^{PC3}$			0.08				0.04
			(0.30)				(0.32)
$\hat{\alpha}_{t+}^{PC4}$				0.16			0.30
				(0.30)			(0.35)
$\hat{\alpha}_{t+}^{PC5}$					-0.15		-0.26
					(0.23)		(0.29)
$\hat{\alpha}_{t+}^{X}$						0.66***	
						(0.11)	
Constant	2.99***	3.00***	3.00***	3.00***	3.00***	1.45***	2.99***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.27)	(0.05)
R^2	0.02	$8.48{\cdot}10^{-6}$	0.002	0.007	0.01	0.44	0.06

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

This table presents the regression results for the cross-sectional Hill estimate extracted from US county level population growth. The dependent variable in the upper panel is $\hat{\alpha}_{t-}^{Y}$, i.e. is the Hill estimate for the left tail of the cross-sectional county level population growth. The independent variables $\hat{\alpha}_{t}^{f}$ given in the first column, is the cross-sectional tail index where the effect is isolated with respect to the given PC, as defined in (13). The tail index estimated on the disturbance terms ($\hat{\alpha}_{t-}^{X}$) is given in the penultimate row. The five PCs are the first five principal components from an assortment of variables suggested by the literature discussed in Table 9. Panel (b) shows the results for the right tail of the distribution. The Hill estimate is calculated by setting the threshold u at 5% of the sample fraction. The stars in the table indicate the following: * p<0.1; ** p<0.05; *** p<0.01.

6 Conclusion

We show that tail-index estimates representing scaling behavior extracted from the cross-section contain a bias. This bias is caused by common time-series fluctuations, which for instance, can originate from an underlying factor structure. The bias fluctuates with the factors. For the left and right tail of the distribution the sign of the bias differs. This offers an opportunity to correct for the bias induced by the factors. We propose two methods to alleviate this bias. Moreover, the bias can also be diminished by looking deeper into the tail.

We find that data with a strong underlying factor structure, as is the case for US stock return data, there is considerable cross-sectional bias which dominates fluctuations in tail index estimates. In data with a weak factor structure (US county population growth) this bias is present, but weak.

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 fluctuations in known factors, unknown factors, measurement error or the tail index. Therefore, we advise caution when attributing variation in the tail index estimates to the scaling behavior in the DGP. Be aware of the factor and choose your threshold wisely.

References

Agarwal, V., S. Ruenzi, and F. Weigert (2017). "Tail risk in hedge funds: A unique view from portfolio holdings". In: *Journal of Financial Economics* 125.3, pp. 610–636.

Atilgan, Y., T. G. Bali, K. O. Demirtas, and A. D. Gunaydin (2020). "Left-tail momentum: Underreaction to bad news, costly arbitrage and equity returns". In: *Journal of Financial Economics* 135.3, pp. 725–753.

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.

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.

Cochrane, J. H. (2009). *Asset pricing: Revised edition*. Princeton university press, 560 p.

Devadoss, S. and J. Luckstead (2016). "Size distribution of US lower tail cities". In: *Physica A: Statistical Mechanics and its Applications* 444, pp. 158–162.

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

Einmahl, J. H. and Y. He (2020). "Unified extreme value estimation for heterogeneous data". In:

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. (1999). "Zipf's law for cities: An explanation". In: *The Quarterly Journal of Economics* 114.3, pp. 739–767.

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.

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.

Ioannides, Y. and S. Skouras (2013). “US city size distribution: Robustly Pareto, but only in the tail”. In: *Journal of Urban Economics* 73.1, pp. 18–29.

Ivette Gomes, M. and O. Oliveira (2003). “How can non-invariant statistics work in our benefit in the semi-parametric estimation of parameters of rare events”. In: *Communications in Statistics-Simulation and Computation* 32.4, pp. 1005–1028.

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.

Karagiannis, N. and K. Tolikas (2019). “Tail Risk and the Cross-Section of Mutual Fund Expected Returns”. In: *Journal of Financial and Quantitative Analysis* 54.1, pp. 425–447.

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

Pisarenko, V. and M. Rodkin (2010). *Heavy-tailed distributions in disaster analysis.* Vol. 30. Springer Science & Business Media.

Resnick, S. I. (1997). “Heavy tail modeling and teletraffic data: special invited paper”. In: *The Annals of Statistics* 25.5, pp. 1805–1869.

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.

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.

7 Appendix

7.1 Bias under second-order Hall expansion

Recall the so-called Hall expansion (Hall and Welsh, 1985) in (7)

$$F(x) = 1 - Cx^{-\alpha}[1 + Dx^{-\theta} + o(x^{-\theta})] \tag{19}$$

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

$$f(x) = \alpha Cx^{-\alpha-1} + (\alpha + \theta)CDx^{-\alpha-\theta-1} + o(1)$$

The conditional expectation can now be found as follows

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

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 the calculus result

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

to the two parts separately, we obtain

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

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

7.2 Higher order bias due to the factor model

Using the second order expansion in (7)

$$F(x) = 1 - Cx^{-\alpha}[1 + Dx^{-\theta} + o(x^{-\theta})],$$

we can analyse the bias in Hill estimates induced by factors in a linear model. Consider the single linear factor model

$$Y_j = \gamma_j g + X_j,$$

where the tails of the distribution of the idiosyncratic shocks X_j are symmetric and satisfy the Hall expansion to the second order:

$$\Pr\{X_j > s\} = \Pr\{-X_j > s\} \simeq C_j s^{-\alpha} + C_j D_j s^{-\alpha-\theta} + o(x^{-\theta})$$

Here $C_j > 0$, but D_j can be of either sign and we assume that $\theta = 2$. This expansion holds for e.g. Student-t distributions and the stationary distribution of stochastic processes like ARCH and GARCH. The right side of the distribution function of the Y_j can be written as:

$$\begin{aligned}
Pr\{Y_j > s\} &= Pr\{X_j > s - \gamma_j g\} \\
&\simeq C_j(s-\gamma_j g)^{-\alpha} + C_j D_j (s-\gamma_j g)^{-\alpha-\theta} + o(x^{-\theta}) \\
&\simeq C_j s^{-\alpha}(1-\gamma_j g/s)^{-\alpha} + C_j D_j s^{-\alpha-\theta}(1-\gamma_j g/s)^{-\alpha-\theta} + o(x^{-\theta}) \\
&\simeq C_j s^{-\alpha} + \alpha C_j \gamma_j g s^{-\alpha-1} + \frac{\alpha(\alpha+1)}{2} C_j (\gamma_j g)^2 s^{-\alpha-2} + C_j D_j s^{-\alpha-\theta} + o(x^{-\theta})
\end{aligned}$$

where the last line follows from taking a second order Taylor approximation of $(1 - \gamma_j g/s)^{-\alpha}$ and a first order Taylor approximation of $(1-\gamma_j g/s)^{-\alpha-\theta}$. Under the monotone density assumption this results in a density function with right tail:

$$\begin{aligned}
f(s) &\simeq \alpha C_j s^{-\alpha-1} + \alpha(\alpha+1) C_j \gamma_j g s^{-\alpha-2} + \frac{\alpha(\alpha+1)(\alpha+2)}{2} C_j (\gamma_j g)^2 s^{-\alpha-3} \\
&\quad + (\alpha+\theta) C_j D_j s^{-\alpha-\theta-1}
\end{aligned}$$

A similar expression applies for the left tail. Using the calculus result from the previous subsection one can derive the following conditional expectation:

$$E[\ln \frac{Y_j}{u} | Y_j > u] - \frac{1}{\alpha} \simeq -\frac{\frac{1}{\alpha+1}\gamma_j g u^{-1} + \frac{\alpha+1}{\alpha+2}(\gamma_j g)^2 u^{-2} + \frac{\theta/\alpha}{\alpha+\theta} D_j u^{-\theta}}{1 + \alpha \gamma_j g u^{-1} + \frac{\alpha(\alpha+1)}{2}(\gamma_j g)^2 u^{-2} + D_j u^{-\theta}}$$

By tail symmetry the conditional expectation we find on the left tail:

$$E[\ln \frac{-Y_j}{u} | -Y_j > u] - \frac{1}{\alpha} \simeq -\frac{-\frac{1}{\alpha+1}\gamma_j g u^{-1} + \frac{\alpha+1}{\alpha+2}(\gamma_j g)^2 u^{-2} + \frac{\theta/\alpha}{\alpha+\theta} D_j u^{-\theta}}{1 - \alpha\gamma_j g u^{-1} + \frac{\alpha(\alpha+1)}{2}(\gamma_j g)^2 u^{-2} + D_j u^{-\theta}}$$

Note that the two conditional expectations are almost symmetric, except for the sign on the first term in the numerators and the sign on the second term in the denominators.

7.3 Differences in tail indices

To analyze the bias in case of cross-sectional heterogeneity in tail indexes, we return to the analysis of the pure Pareto case. However, 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 in case of the Pareto distribution

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

and

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

Conditional on being above threshold u, we get from the above to a first order

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

and

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

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,

$$E[\frac{1}{K}\sum_{m=1}^{K}\ln\frac{Y_m}{u}|Y_m > u] = \frac{\lambda u^{-\alpha}n}{\lambda u^{-\alpha}n + \lambda u^{-\alpha-\varepsilon}n}\frac{1}{\alpha} + \frac{\lambda u^{-\alpha-\varepsilon}n}{\lambda u^{-\alpha}n + \lambda u^{-\alpha-\varepsilon}n}\frac{1}{\alpha+\varepsilon}$$
$$= \frac{1}{1+u^{-\varepsilon}}\frac{1}{\alpha} + \frac{u^{-\varepsilon}}{1+u^{-\varepsilon}}\frac{1}{\alpha+\varepsilon}$$

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 $\alpha + \varepsilon$.

7.4 Tables and Figures

Table 7: Correlations cross-sectional tail index **RMW and CMA Factors**

	Market	SMB	HML	RMW	CMA
$\hat{\alpha}_{t-}^{Y}$	-0.69	-0.45	0.19	0.16	0.28
$\hat{\alpha}_{t+}^{Y}$	0.75	0.52	-0.12	-0.19	-0.22
$\hat{\alpha}_{t-}^{f}$	-0.76	-0.78	-0.17	-0.06	0.08
$\hat{\alpha}_{t+}^{f}$	0.82	0.81	0.21	0.04	0.01
$\rho(\hat{\alpha}_{t-}^{f}, \hat{\alpha}_{t+}^{f})$	-0.91	-0.82	-0.04	-0.04	-0.01

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

	Market	SMB	HML	RMW	CMA
$\hat{\alpha}_{t-}^{Y}$	-0.31	-0.25	0.04	0.09	0.13
$\hat{\alpha}_{t+}^{Y}$	0.33	0.33	-0.02	-0.19	-0.07
$\hat{\alpha}_{t-}^{f}$	-0.49	-0.44	-0.08	0.10	-0.02
$\hat{\alpha}_{t+}^{f}$	0.48	0.41	0.06	-0.03	0.01
$\rho(\hat{\alpha}_{t-}^{f}, \hat{\alpha}_{t+}^{f})$	-0.40	-0.16	-0.01	0.04	-0.06

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

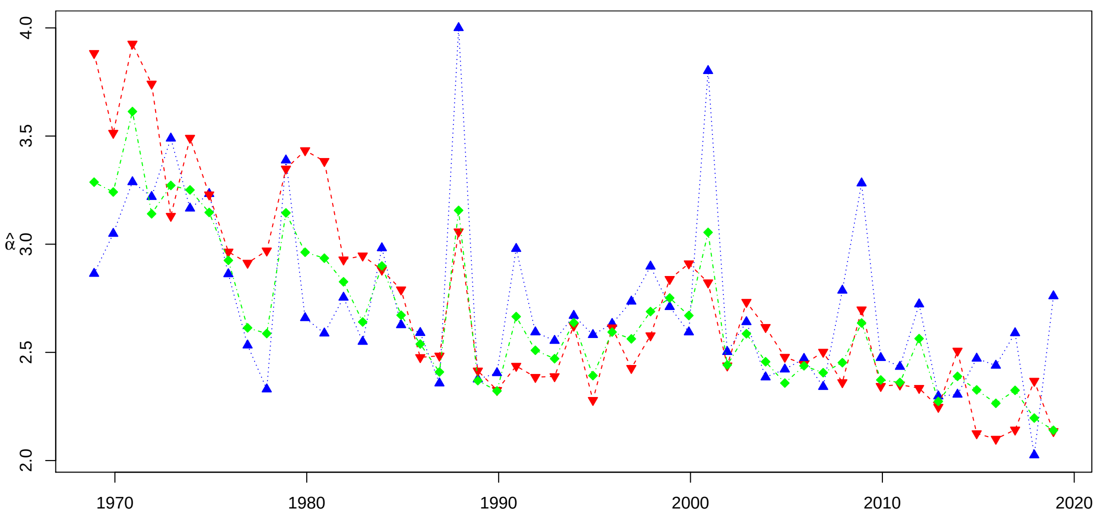
This table presents the correlation between various specifications of the cross-sectional Hill estimates. In the first two rows, $\hat{\alpha}_{t-}^{Y}$ and $\hat{\alpha}_{t+}^{Y}$ are the cross-sectional Hill estimates for excess stock returns for the left and right tail, respectively. The $\hat{\alpha}_{t-(+)}^{f}$, 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 (13) for the left (right) tail. The five factors are the market, small-minus-big (SMB), high-minus-low (HML), robust-minus-weak (RMW) and conservative-minus-aggressive (CMA). The last row shows the correlation between the left and right tail estimates of $\hat{\alpha}_{t}^{f}$. 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 8: 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

This table reports the correlation between the factors for the financial returns data. The seven factors are the market, small-minus-big (SMB), high-minus-low (HML), momentum (SMB), liquidity factor (Liq), robust-minus-weak (RMW) and the conservative-minus-aggressive (CMA) factor.

Figure 6: Bias corrected estimates for US stock returns (mirror and shift method)



This figure presents the time series of tail estimates for US stock returns after applying the two bias reduction methods. The y-axis shows the value of the estimates in terms of α. The blue triangles (▲) and red triangles (▼) show the Hill estimates for the right and left tail of $Y_{jt} - \overline{Y}_t$, respectively. The green diamonds (♦) are the estimates of the tail index after correcting for bias using the mirror method. The estimates are extracted from the cross-section each month and subsequently averaged within a year for presentational purposes. The threshold for the Hill estimates is set at 5% of the sample fraction.

Table 9: 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.

Table 10: 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) $\hat{\alpha}^Y_{t-}$

	(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) $\hat{\alpha}^Y_{t+}$

This table presents the regression results for the cross-sectional Hill estimate and the five asset pricing factors. Here the dependent variable in the upper panel is $\hat{\alpha}^Y_t$, i.e. is the Hill estimate for the left tail of the raw cross-sectional excess returns. The independent variables are the five asset pricing factors; the market, small-minus-big (SMB), high-minus-low (HML), momentum (MOM) and the liquidity (Liq) factor. The lower panel shows the results for the right tail of the distribution. The threshold u to estimate the Hill estimate is set to **5%** of the sample fraction. The stars in the table indicate: * p<0.1; ** p<0.05; *** p<0.01.

Table 11: Regression cross-sectional tail index **0.5% Threshold**

$\hat{\alpha}_{t-}^{M}$	0.38***						0.46***
	(0.10)						(0.11)
$\hat{\alpha}_{t-}^{SMB}$		0.41***					0.46***
		(0.10)					(0.11)
$\hat{\alpha}_{t-}^{HML}$			−0.04				−0.32**
			(0.13)				(0.14)
$\hat{\alpha}_{t-}^{MOM}$				−0.06			−0.31*
				(0.15)			(0.16)
$\hat{\alpha}_{t-}^{Liq}$					0.17		0.10
					(0.14)		(0.16)
$\hat{\alpha}_{t-}^{X}$						0.81***	
						(0.03)	
R^2	0.02	0.02	0.0002	0.0003	0.002	0.58	0.06

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

$\hat{\alpha}_{t+}^{M}$	0.32***						0.35***
	(0.08)						(0.09)
$\hat{\alpha}_{t+}^{SMB}$		0.47***					0.53***
		(0.09)					(0.09)
$\hat{\alpha}_{t+}^{HML}$			0.09				−0.07
			(0.12)				(0.13)
$\hat{\alpha}_{t+}^{MOM}$				−0.06			−0.52***
				(0.14)			(0.15)
$\hat{\alpha}_{t+}^{Liq}$					0.27*		0.16
					(0.14)		(0.15)
$\hat{\alpha}_{t+}^{X}$						0.80***	
						(0.03)	
R^2	0.02	0.04	0.001	0.0003	0.01	0.51	0.08

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

This table presents the regression results for the effect of the factors on the cross-sectional Hill estimate for a lower threshold u. The threshold u to estimate the Hill estimate is set to **0.5%** of the sample fraction. Here the dependent variable in the upper panel is $\hat{\alpha}_{t-}^{Y}$, i.e. is the Hill estimate for the left tail of the raw cross-sectional excess returns. The independent variables $\hat{\alpha}_{t}^{f}$, stated in the first column, is the cross-sectional tail index where the factor f's effect is isolated, as defined in (13). Furthermore, $\hat{\alpha}_{t-}^{X}$ is the tail index estimated on the estimated disturbance terms of the five-factor asset pricing model; the market, small-minus-big (SMB), high-minus-low (HML), momentum (MOM) and the liquidity (Liq) factor. The lower panel shows the results for the right tail of the distribution. The constant included in the regression is excluded from the presented results. The stars in the table indicate: * p<0.1; ** p<0.05; *** p<0.01.

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 extracted from the percentage change of county population. 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: Regression cross-sectional tail index **0.5% Threshold** (County Data)

$\hat{\alpha}_{t-}^{PC1}$	0.04						0.20
	(0.21)						(0.25)
$\hat{\alpha}_{t-}^{PC2}$		−0.19					-0.50
		(0.24)					(0.31)
$\hat{\alpha}_{t-}^{PC3}$			0.42				0.68*
			(0.33)				(0.37)
$\hat{\alpha}_{t-}^{PC4}$				−0.21			-0.13
				(0.37)			(0.48)
$\hat{\alpha}_{t-}^{PC5}$					−0.14		-0.07
					(0.33)		(0.46)
$\hat{\alpha}_{t-}^{X}$						0.83***	
						(0.11)	
Constant	3.30***	3.32***	3.33***	3.28***	3.30***	0.42	3.32***
	(0.19)	(0.18)	(0.18)	(0.19)	(0.19)	(0.40)	(0.200)
R^2	0.0009	0.02	0.04	0.008	0.004	0.57	0.11

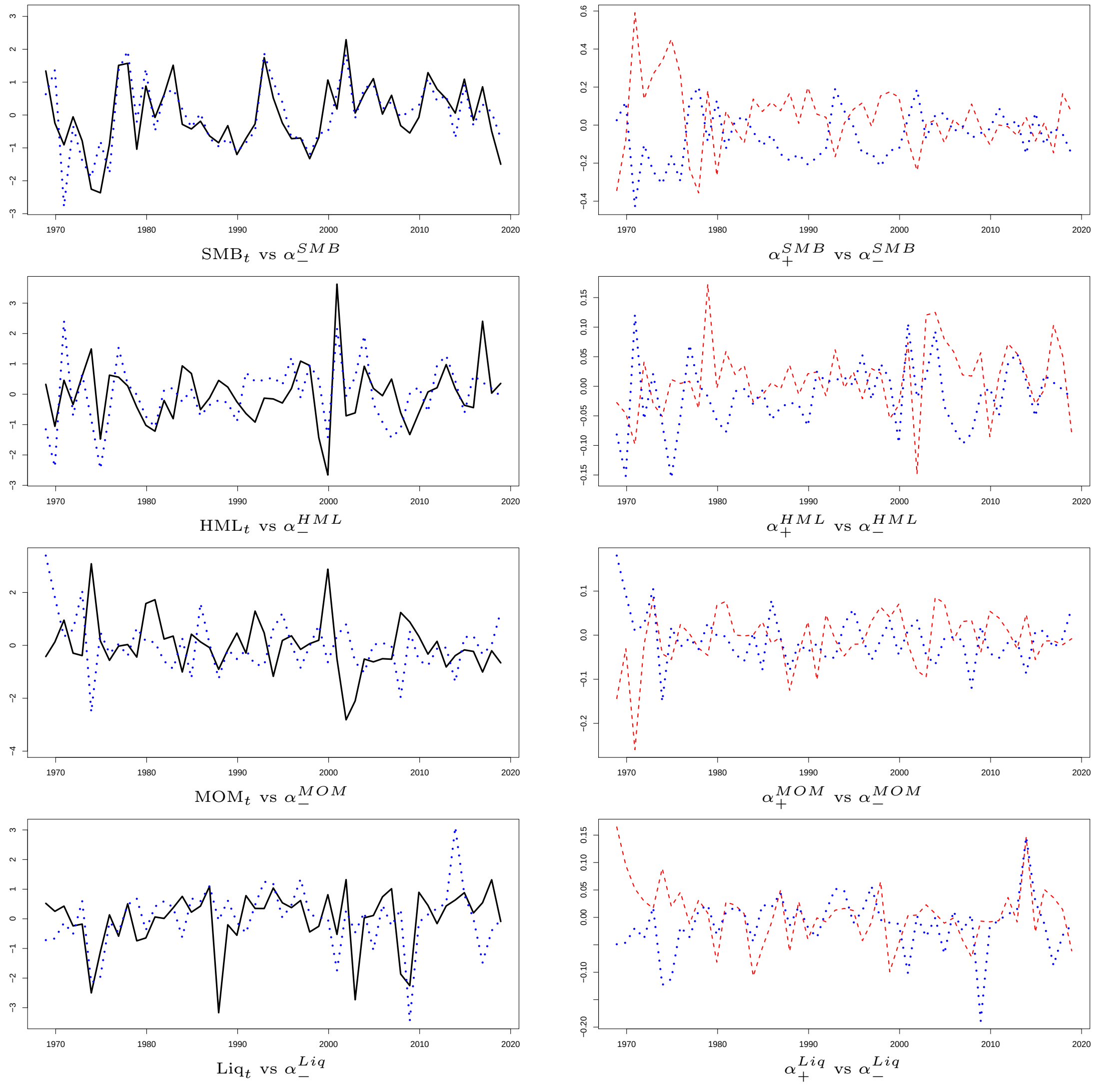
(a) Left cross-sectional tail index, i.e. $\hat{\alpha}_{t-}^{Y}$

$\hat{\alpha}_{t+}^{PC1}$	0.49						0.68
	(0.43)						(0.47)
$\hat{\alpha}_{t+}^{PC2}$		−1.43*					-1.65**
		(0.67)					(0.78)
$\hat{\alpha}_{t+}^{PC3}$			−0.31				-0.08
			(0.38)				(0.43)
$\hat{\alpha}_{t+}^{PC4}$				−0.52			-0.29
				(0.64)			(0.66)
$\hat{\alpha}_{t+}^{PC5}$					0.34		0.19
					(0.72)		(0.78)
$\hat{\alpha}_{t+}^{X}$						1.04***	
						(0.25)	
Constant	4.39***	4.37***	4.41***	4.40***	4.40***	1.10	4.33***
	(0.22)	(0.22)	(0.22)	(0.22)	(0.23)	(0.83)	(0.22)
R^2	0.03	0.10	0.02	0.02	0.01	0.28	0.17

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

This table presents the regression results for the effect of the PCs on the cross-sectional Hill estimate extracted from US county level population growth for a lower threshold u. The threshold u to estimate the Hill estimate is set to **0.5%** of the sample fraction. Here the dependent variable in the upper panel is $\hat{\alpha}_{t-}^{Y}$, i.e. is the Hill estimate for the left tail of the cross-sectional county level population growth. The independent variables $\hat{\alpha}_{t}^{f}$, stated in the first column, is the cross-sectional tail index where the PC f's effect is isolated, as defined in (13). Furthermore, $\hat{\alpha}_{t-}^{X}$ is the tail index estimated on the estimated disturbance terms of the five principal components model. The five PCs are the first five principal components from an assortment of variables suggested by the literature. The lower panel shows the results for the right tail of the distribution. The constant included in the regression is excluded from the presented results. The stars in the table indicate the following: * p<0.1; ** p<0.05; *** p<0.01.

Figure 7: Effects of asset pricing factor on cross-sectional Hill estimator



These figures display the time series of the asset pricing factors and the isolated effect of these factors on the cross-sectional Hill estimates $\hat{\alpha}_t^f$, as defined in (13). In the left column of figures $\hat{\alpha}_{t+}^f$ for the right tail (blue dotted line) and the respective normalized factors (solid black line) are plotted. In the right figures $\hat{\alpha}_{t-}^f$ (red dashed line) and $\hat{\alpha}_{t-}^f$ (blue dotted line) are plotted (not normalized). These annual observation are created by averaging the monthly observations within a year. The plots are for the small-minus-big (SMB), high-minus-low (HML), momentum (MOM) and the liquidity factor (Liq). The threshold for the Hill estimates is set at 5% of the sample fraction.