Covariates hiding in the tails*

Milian Bachem¹, Lerby M. Ergun², and Casper G. de Vries¹

¹Erasmus University Rotterdam ²Bank of Canada

Sunday 3rd November, 2024

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, observed time variation in estimated tail indices may not originate from the data generating process. In the case of a linear factor model, the factors bias 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 factors, how to remedy for the bias and how to apply our methods to financial data and geographic location data.

JEL Classification: C01, C13, C21

Keywords: Tail index estimation, Power laws, Scaling behavior, Cross-sectional bias

*Ergun (corresponding author): Financial Markets Department, Bank of Canada, 234 Wellington St W, Ottawa, ON K1A 0G9 (lerbyergun@gmail.com).Bachem: bachem@ese.eur.nl. De Vries: cdevries@ese.eur.nl. We would like to thank Jason Allen, Maarten Bosker, John Einmahl, Bruno Feunou, Jean-Sebastian Fontaine, 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, EEA meetings 2020, CFE meetings 2021 and the IAAE conference 2023.

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. Distributions with such tail behavior only have bounded moments up to the value of the tail index and are therefore referred to 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 multitude of different assets, recent literature also investigates the cross-sectional scaling behavior. Kelly and Jiang (2014), Almeida et al. (2022), Andersen et al. (2024) and Faias (2023) estimate the tail index from cross sections of US stock returns. They find that the tail index varies considerably through time. Additionally, Karagiannis and Tolikas (2019), Atilgan et al. (2020) and Agarwal et al. (2017) use measures (such as Value-at-Risk) related to the tail shape in cross sections to expose risks of different assets not priced by previously known factor models.

The statistical properties of scaling behavior estimates in repeated cross-sectional data remain under explored. Some of the authors are aware of a possible bias and first clean the returns for e.g., the FF-3 factors. But this requires knowing the right factors and necessitates time series estimation of their coefficients, possibly introducing other biases. Our correction methods, however, do not require knowledge or identification of such factors, since we exploit how these factors bias the tail estimates in opposite directions depending on the upper or lower tail. So we use information as to how factors impinge on the tail estimates without having to know the factors. This bias is due to the Hill estimator's non-location invariant property, i.e., adding a constant to the data changes the tail exponent estimate. As factors vary over time, it causes the cross-sectional distribution to shift and change the tail index estimate. This links variation in tail index estimates to the non-heavy tailed part of the data generating process (DGP), risking false inference.

Our theoretical framework, assumes the data are generated by a linear factor model with heavytailed idiosyncratic shocks. Furthermore, assume that the scaling behavior is identical for all idiosyncratic shocks 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 derive the asymptotic distribution of the estimator for the dependent variable and show that cross-sectional estimates of the tail index are biased. Furthermore, our analysis shows that this bias varies with the size of the factor realizations at different points in time. Interestingly, the factor induces bias to go in opposite directions for left and right tail index estimates. If one does not correct for the bias this may lead one to misinterpret bias fluctuations for time varying tail indices. Through simulations and real data we demonstrate that this feature is empirically important.

Typically bias reduction in tail estimation is not straightforward as it hinges on second-order parameters that are difficult to estimate. Fortunately, the current case is easier to handle due to the fact that the biases in the two tails run in opposite directions. We suggest two simple-to-implement procedures to alleviate the bias due to the shift in location originating from the factors. The first method takes advantage of the symmetry in the bias for the left and right tail estimates. 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. And, this method has the added benefit that it allows the left and right tail indices to be estimated separately. In a simulation exercise we show that both asymptotically justified methods alleviate bias caused by a cross-sectional location shift for intermediate sample sizes. 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 datasets contain a wide cross section and a long time-series. The wide cross section is vital for accurate tail index estimation. The long time-series dimension helps to elicit the effect of the bias caused by the linear factor structure.

The linear factor structure is the workhorse model for explaining asset prices empirically (see e.g., Fama and French (2015); Stambaugh and Lubos (2003)). Cochrane (2009) shows that under mild restrictions, asset returns naturally follow from a linear factor structure. 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 in the estimated tail index 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. Moreover, correcting the bias causes the time series of left tail estimates to be smoother. The high (less heavy tail) estimates during financial market turmoil become more inline with the surrounding low (heavy) tail exponent estimates. For example, the biased left tail exponent estimate for the 1973 crisis is adjusted downwards from 4.3 to 3.4, implying the fourth moment of the cross-sectional distribution does not exist. For the right tail the opposite occurs. Very heavy-tailed (biased) crisis estimates adjust to become lighter than the average right tail estimate, i.e., extreme positive returns are now modelled to occur less frequently during crisis times.

A second test case is based on county-population 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 data in both time-series and cross-section dimensions makes this dataset amenable to our analysis. However, the difference with the financial data is that a clear factor structure is lacking. 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 lack of explanatory power of the factors behind population growth induces less correlation between the bias and the principal components. Nevertheless, the correlation is still significant and in the predicted direction. The contrast in results with the first test case underlines that a DGP with factors that have strong explanatory power suffers more severely from the cross-sectional bias.

There are three main contributions that we want to highlight. First, on the theory side we derive the asymptotic distribution of the cross-sectional estimator under a linear factor structure. This differs from literature that studies bias correction for a single cross-section or time series under the i.i.d. assumption, see for instance Beirlant et al. (2012) for an overview of this literature. Our article is related to a small but growing literature on the effects of heteroskedasticity in extreme value analysis. In particular, Einmahl et al. (2016) allow for variation in the first-order scale parameter, but do not focus on how this affects the bias. In the case at hand it turns out that the bias is the main issue and stems from variation in the second-order scale parameter.

The second contribution of this work, to the best of our knowledge, is highlighting that the timevarying nature of the bias, and the drivers of this variation, have largely gone unnoticed. It is well known that the Hill estimator is non-location invariant, which induces bias. Moreover, most crosssectional studies focus on a single cross section, making it difficult to detect the source of this bias. Only recently have repeated cross sections been employed to estimate time variation in scaling behavior, underscoring the need to investigate the impact of location shifts in the cross-sectional distribution over time.

Third, we propose three solutions to alleviate the bias. The first method is standard in extreme value theory (EVT). With a sufficient number of data, one can increase the tail threshold to decrease the bias, however, this comes at the cost of increasing the variance of the estimator. But in cross-sections, data limitations constrain the use of this approach. The second method is inspired by Ivette Gomes and Oliveira (2003), who suggest to deduct a tuning parameter from the data. For the case at hand we show that one can use the cross-sectional average as the tuning parameter.

The third approach, which is specific to the cross-sectional structure of the data, is our own mirror estimator.

The paper is structured as follows. Section 2 derives the asymptotic distribution of the crosssectional Hill estimator and introduces the two proposed estimators to alleviate the bias. Section 3 describes the data used in the two empirical illustrations. Subsequently, Section 4 documents the bias and evaluates the newly proposed estimators followed by the conclusion.

2 Theory

Consider a linear factor model with *n* factors G_i , i = 1, ..., n and idiosyncratic shocks X_j . At any point in time the dependent variable Y_j (omitting superfluous time indices on Y_j , G_i and X_j) for cross-sectional entity *j* is

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

To be specific, in the finance application the Y_j 's are logarithmic excess returns and in our other application the data are logarithmic population growth rates. The factors G_i are stochastic over time. Further, the coefficients γ_{ij} are assumed to be fixed. Define, as a shortcut:

$$H_j = \sum_{i=1}^n \gamma_{ij} G_i$$

Thus at a specific point in time

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

The H_j are allowed to vary stochastically over time. However, as factors G_i do not vary over the cross-section observations and the γ_{ij} are fixed, H_j is functionally a constant for tail behaviour in the cross-section. We can therefore speak of the conditional distribution $Pr(H_j + X_j \le s | H_j = h_j)$.

The X_i are non-normally heavy-tailed distributed, like in the case of a Student-t distribution. That

is, the tail of the distribution follows a power law in the sense of regular variation. To define the concept of regular variation, 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 α is known as the tail index. The class of regularly varying distributions is closed under addition (convolutions).

The power decline implies self-scaling behavior. This follows most clearly from Feller's convolution theorem. If two independent X_i 's vary regularly at infinity and follow the same distribution, then $\lim_{s\to\infty} Pr(X_1 + X_2 > s)/Pr(X_1 > s) = 2$, thus the scale changes but the tail index does not. Furthermore, as moments only exist up to α , i.e., $E \mid X_i \mid^p < \infty$ for $p < \alpha$, a decrease in α gives a heavier tail. For both applications the heavy-tail assumption is backed by the empirical literature.

The X_j are i.i.d. random variables both from a cross-section and a time-series perspective. This assumption is build into most asset pricing models. For instance, the most relied upon asset pricing model for empirical applications, Arbitrage Pricing Theory, explicitly assumes that the idiosyncratic shocks are cross-sectionally uncorrelated. This prevents the DGP from describing any arbitrary set of returns. However, we could accommodate cross-sectional dependence in X_j . In this case, the Hill estimator is still consistent, but comes with wider confidence intervals.

Below we first describe how α for Y_j can be estimated, assuming the X_j are exactly Pareto distributed. Then we investigate how the estimator is influenced by a fixed factor h_j . We show that h_j induces a bias in the cross-sectional estimates. Interestingly, the bias in the two tails are of the opposite sign. Lastly, we consider how the bias can be remedied, exploiting the sign difference.

2.1 Hill estimator

We consider Hill (1975)'s estimator to estimate α . Let $Y_{(j)}$ denote the descending order statistics from a cross section of *m* observations Y_j :

$$Y_{(1)} \ge Y_{(2)} \ge \ldots \ge Y_{(K)} \ge u \ge Y_{(K+1)} \ge \ldots \ge Y_{(m-1)} \ge Y_{(m)}.$$

The Hill estimator uses the K < m highest-order statistics above threshold u to estimate the (inverse of the) tail index α . For the lower tail estimate, one takes the negative of the observations and reorders these from high to low. The threshold u is typically chosen as a percentage of the sample size m (and deep into one of the tails); we suppress the dependency of u on m whenever possible, otherwise we write u_m . The Hill estimator calculates the average logarithmic difference between the threshold and the higher-order statistics:

$$\frac{1}{\hat{\alpha}} = \frac{1}{K} \sum_{i=1}^{K} \ln\left(\frac{Y_{(i)}}{u}\right).$$
(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), while the other uses one of the upper-order statistics, $Y_{(k+1)}$, as a threshold. In the fixed threshold version, the K number of order statistics exceeding u is random, otherwise the threshold is random.¹

¹Goldie and Smith (1987) argue that "In practical terms, there is little to choose between these two points of view."

2.2 Single observation

To explain most clearly the effect and, consequently, the difference in sign due to the bias caused by the h_i , we first 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 estimator (2) is the conditional expectation

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

Next consider the case with a non-zero fixed location shift, $h \neq 0$, 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*:

$$\Pr\{Y \le s\} = \Pr\{X + h \le s\} = 1 - (s - h)^{-\alpha}$$
$$= 1 - s^{-\alpha} [1 + \alpha h s^{-1} + o(s^{-1})].$$
(4)

Apply the expectation in (3) twice to get

$$E\left[\ln\frac{Y}{u}|Y>u\right] = \frac{1}{\alpha} - \frac{1}{\alpha+1}hu^{-1} + o\left(u^{-1}\right).$$
(5)

In comparison to (3), we now have an additional term signifying the bias due to the location shift. (See the online Appendix for a more detailed derivation.) The location shift has a different effect when considering the left tail. If the idiosyncratic noise term again follows a standard Pareto distribution, then

$$\Pr\{Y \le -s\} = \Pr\{-X > s+h\} = (s+h)^{-\alpha}$$
$$= s^{-\alpha} [1 - \alpha h s^{-1} + o(s^{-1})].$$
(6)

This again results in a bias dependent on *h*:

$$E\left[\ln\frac{Y}{u}|Y\leqslant -u\right] = \frac{1}{\alpha} + \frac{1}{\alpha+1}hu^{-1} + o\left(u^{-1}\right).$$
(7)

The bias, however, is of the opposite sign. This implies that h biases the left and right tail index estimates in opposite directions. Furthermore, the two biases are each other's 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:

$$G(x) = 1 - Ax^{-\alpha} [1 + Bx^{-\theta} + o(x^{-\theta})].$$
(8)

Here $\alpha > 0$, A > 0, $\theta > 0$ and *B* 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 (G)ARCH process, see e.g., Sun and Vries (2018). In the remainder of the paper we assume that expansion (8) applies with $\theta > 1$, as is the case for the Student-t distribution and the (G)ARCH processes. One shows that if $\theta > 1$, the first-order bias terms are as in (5) and (7), i.e., the second-order term is only of order $o(u^{-1})$.

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 location shift h_j and idiosyncratic shocks X_j . Suppose that the X_j satisfy (8) and that the tail indices α and $\theta > 1$ do not vary cross-sectionally. In that case (4) and (6) adhere to the following general expansion:

$$G_{i}(x) = 1 - Cx^{-\alpha} [1 + D_{i}x^{-1} + o(1)].$$
(9)

Here D_j captures the specific second-order scales as in (4) and (6) for the *j*-th observation, i.e., αh_j and $-\alpha h_j$ respectively for the right and left tail data. We also allow the first-order scale parameters *C* to be different from 1.² In the online Appendix, 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 also contribute to the bias. The cross-sectional tail estimate is then a weighted average of the tail indices of the cross section.

In the cross section let $\bar{H} = \frac{1}{m} \sum h_j$ be the cross-sectional average shift. Furthermore, suppose that $\lim_{m \to \infty} \bar{H} = \bar{H}$ exists. Similarly, for the left tail data we then have $-\bar{H}$, c.f. to (6). Given this setting, we have the following general result:

Proposition 1 Suppose the X_j are independently distributed with distribution function that conforms with the expansion as in (8), have the same α , same scale A and $\theta > 1$. Then the distribution of $Y_j = h_j + X_j$ satisfies the upper tail expansion in (9) as $x \to \infty$. Consider the distribution of the

²The first-order scale coefficient *C* may also vary in size. This affects the size of the bias and the variance of the limiting distribution. The effect is quantitatively very minor in nature. For the sake of simplicity we only present the proof with fixed *C*, but a varying *C* can be easily accommodated. Einmahl and He (2023) show that the tail index estimates remain consistent if the first-order scale coefficients differ.

Hill estimator

$$\frac{1}{\widehat{\alpha}(m)} = \sum_{j=1}^m \left(\ln \frac{Y_{(j)}}{u_m} \right) \mathbb{1}_{Y_{(j)} > u_m}$$

for $m \to \infty$ such that

$$u_m = m^{1/(\alpha+2)}$$

it follows that $1/\hat{\alpha}(m)$, appropriately scaled, has the following limit normal distribution:

$$m^{1/(\alpha+2)}\left(\frac{1}{\widehat{\alpha}(m)}-\frac{1}{\alpha}\right) \xrightarrow{d} N\left(-\frac{1}{\alpha+1}\overline{\overline{H}},\frac{1}{\alpha^2}\frac{1}{C}\right)$$

Proof See Appendix A.

Corollary 1 For the lower tail data the complementary result is

$$m^{1/(\alpha+2)}\left(\frac{1}{\widehat{\alpha}(m)}-\frac{1}{\alpha}\right) \xrightarrow{d} N\left(\frac{1}{\alpha+1}\overline{\overline{H}},\frac{1}{\alpha^2}\frac{1}{C}\right).$$

The typical Hill estimator as presented in e.g., Haan and Ferreira (2006), assumes identical first and second-order scales and tail indexes. The difference here is that we allow for variation in the second-order scales. Therefore, the bias term now contains the expectation of the second-order scale coefficients. A priori, we do not know the sign of $\sum h_j$ at a certain point in time, but its effect on the bias of the left and right tail estimates of α is such that these are of opposite sign.³

From hereon for the purpose of clarity we introduce a superscript on α to indicate the data the estimate refers to, e.g., in case of the raw data we use *Y*. The subscript notes to which tail the estimate refers ("+" for the right tail and "-" for the left tail). The above proposition implies:

³Furthermore, since the estimation is based on data from the tail only, the typical rate is lower than the square root of the sample size. The asymptotic distribution with the rate that minimizes the asymptotic mean squared error is presented in Appendix A.

Corollary 2 For the tail index estimated on the upper tail data

$$\partial Bias(1/\hat{\alpha}_{+}^{Y})/\partial \bar{H} = -\frac{1}{\alpha+1}.$$

and

Corollary 3 For the tail index estimated on the lower tail data

$$\partial Bias(1/\hat{\alpha}_{-}^{Y})/\partial \bar{H} = rac{1}{lpha+1}.$$

Specifically, consider, e.g., the one factor model where $h_j = \gamma_j g$. Take for example the Capital Asset Pricing Model, where g is the market excess return. In that case

$$rac{\partial Bias(1/\hat{lpha}^Y_+)}{\partial g} = -rac{ar{ar{\gamma}}}{lpha+1},$$

where $\overline{\overline{\gamma}} = \lim_{m \to \infty} \overline{\gamma}$ and $\overline{\gamma}$ is the average of the *m* number of different γ_j coefficients. Thus, if the market factor is positive and 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. For the left tail data the opposite result applies. As the factor *g* varies over time the above result implies a positive correlation between the factor and the upper tail index estimate, $\hat{\alpha}_+^Y$. Moreover, the bias also generates a negative correlation between the left and the right tail index estimates.⁴

Suppose one wants to identify the contribution of the individual factors to the bias. Assuming one knows the factors, 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 recovered

⁴A third-order expansion, provided in the online Appendix, reveals that some higher-order terms have 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.

from a time-series regression. Consequently, 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 :

$$S_j^f = Y_j - \sum_{i \neq f} \hat{\gamma}_{ij} g_i.$$
⁽¹⁰⁾

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

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

$$\frac{1}{\Delta^f} = \frac{1}{\hat{\alpha}^{S^f}} - \frac{1}{\hat{\alpha}^{\hat{X}}}.$$
(11)

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

$$Bias(1/\Delta_{+}^{f}) = -\frac{1}{1+\alpha} \lim_{m \to \infty} \left(\frac{1}{m} \sum_{j=1}^{m} \gamma_{fj} g_f \right).$$
(12)

The RHS isolates the bias contributed by factor f, with a different sign for the left tail. Given positive coefficients γ_{fj} , one expects a negative correlation between $1/\Delta^f_+$ and g_f in the right tail.

Recall that α has the intuitive interpretation as the number of bounded moments. For this reason, in the empirical application we report how α , instead of $1/\alpha$, relates to a factor. This implies that the signs of the biases, like in Proposition 1, change direction.

From Proposition 1 it can be noted that the size of the bias diminishes with the size of the threshold u_m . This opens up the possibility to reduce the bias by increasing the threshold u_m . For this reason we report correlation estimates at two different thresholds: 5% (the conventionally used threshold)

and 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_m . If the threshold increases at too high a rate, this diminishes the bias but raises the variance. Conversely, if u_m increases at a slower rate, the bias dominates asymptotically. The typical approach in tail index estimation tries to strike a balance between the two vices. That is why we consider two alternative methods of bias correction.

2.4 Bias correction in Hill estimates

The characterization of the bias in the Hill estimates specified in (5) and (7) 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 $(1/\hat{\alpha}^{mirror})$ 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 cross-sectional mean from the dependent variable, i.e., $Y_j - E[Y_j]$. Furthermore, for both methods no prior knowledge of the factors is required. We consider both methods below.

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. As we show in (5) and (7), the bias terms are their mirror images under tail symmetry. This gives an opportunity for bias reduction by taking the average of the left and right tail index estimates.

If the tails of the distribution of the X_j are symmetric, taking the average of the left tail estimate $1/\hat{\alpha}_{+}^{Y}$ and the right tail estimate $1/\hat{\alpha}_{+}^{Y}$ yields a new estimator. Noting Proposition 1 and Corollary 1, the new estimator has asymptotic normal distribution:

$$m^{1/(\alpha+2)}\left(\frac{1}{\hat{\alpha}^{mirror}}-\frac{1}{\alpha}\right)=m^{1/(\alpha+2)}\left(\frac{1/\hat{\alpha}_{-}^{Y}+1/\hat{\alpha}_{+}^{Y}}{2}-\frac{1}{\alpha}\right)\stackrel{d}{\to} N\left(0;\frac{1}{2\alpha^{2}C}\right).$$

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 . Due to the law of large numbers, as $m \to \infty$ and $k \to \infty$, asymptotically the left and right tail are each other mirror images and hence the asymptotic bias evaporates at a higher rate.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

The procedure described above is less meaningful, if the tail indices on the left and the right side of the distribution differ. If the bias is caused by a factor shift, one correct the bias by shifting the data back in the other direction (henceforth referred to as the shift method). This is similar to the procedure outlined in Ivette Gomes and Oliveira (2003). Consider the following revised estimator:

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

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 $1/(\alpha^2 C)$. For ease of interpretation, we state the bias for this estimator under the assumption of a one-factor model, i.e., i = 1. By demeaning the observations 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. Note that the third term on the RHS is zero as one sums over the *m* observations.

The second and third term contain the random elements. Our assumption is that the X_i are i.i.d.

⁵Due to the contribution of higher-order terms in the tail expansion of X_i , some bias remains; see online Appendix.

and at large s

$$Pr\{X_j \ge s\} = Cs^{-\alpha} [1 + o(s^{-\alpha})]$$

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 *C* changes, resulting in

$$Pr\left\{\frac{1}{m}\sum_{j=1}^{m}X_{j} \ge s\right\} = m^{-\alpha+1}Cs^{-\alpha}\left(1+o(1)\right).$$

This specification allows us to use (5) to derive the bias in the right tail by substituting:

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

(where $\overline{\gamma} = (1/m) \sum_{j=1}^{m} \gamma_j$). This yields a bias in the right tail as

$$Bias\left(1/\alpha_{+}^{Y-\overline{Y}}\right) = -\frac{1}{\alpha+1}\left(\lim_{m\to\infty}\frac{1}{m}\sum_{j=1}^{m}(\gamma_{j}-\overline{\gamma})g\right) = 0.$$

2.4.3 Simulations

To investigate the efficacy of the above methods of bias reduction in intermediate sample sizes, we conduct simulations that are presented and discussed in the online Appendix. We simulate data from a linear model with Student-t distributed idiosyncratic shocks, where the tails are (a)symmetric. The non-idiosyncratic part is either with a constant or a single factor with normally distributed coefficients. The simulations confirm the theoretical result of the linear relationship between the value of *h* and the bias in Hill estimates as derived in (5) and (7). Therefore, the mirror method produces estimates that are close to being unbiased. As the mirror method averages over two estimates, the standard deviation of estimates is smaller, by a factor close to $1/\sqrt{(2)}$. In case the tails of the idiosyncratic shocks are asymmetric the mirror method is not applicable. Using

the shift method is as effective in reducing the bias, forgoing the variance reduction in the mirror method. One thing to note from the simulations with normally distributed coefficients, is that the effect of this thin-tailed component biases the Hill estimates downwards in finite samples.

3 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. 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.

3.1 Firm stock returns

The Center for Research in Security Prices (CRSP) provides a wide cross section of firm return data for the US equity market with 13,535 individual US traded firms. These daily data are collected from the NYSE, AMEX, NASDAQ and NYSE Arca exchanges since 1925. In accordance with the financial literature on asset pricing, we use monthly stock (log) returns from 1963 to 2019.⁶

There is a large body of literature that 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 unreported results, the analysis is repeated with a model where the momentum

⁶Following common practice in the asset pricing literature, we only include common stocks (share code 10 and 11) with a price above 5 dollars, noting that relaxing these filters leads to almost identical results.

and liquidity factors are substituted for the Robust-Minus-Weak (RMW) and Conservative-Minus-Aggressive (CMA) factors (Fama and French, 2015). The results are very similar.⁷

3.2 County population data

Another heavily researched field in power laws is the geographical distribution of population sizes. 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 runs from 1970 to 2017. In contrast to the 648 time-series dimension for the monthly US stock data, these data have a length of only 46. Moreover, the US Census is only conducted every 10 years. Annual data are 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 comparisons 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.⁸ For the creation of population change, we use the Federal Information Processing Standards (FIPS) codes, which uniquely identify counties and county equivalents in the US. 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 as to

⁷We obtain the five Fama and French (2015) factors and the momentum factor from Kenneth R. French's website and the liquidity factor from Lubos Pastor's website. Table B.1 in the Appendix, presents their pairwise correlations.

⁸Due 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. Our analysis considers both tails of the distribution.

what constitutes the best combination of factors, we conduct a principal component analysis (PCA) and extract the first five PCs. The variables used for the PCA are suggested in Chi and Ventura (2011), which we describe in the online Appendix. By using five PCs, we avoid multicollinearity and over-fitting, which is likely to arise in a model with many explanatory variables.

3.3 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_{jt}^{f} in (10) for factor f, we run linear time-series regressions to estimate the factor coefficients γ_{ij} . In case of the financial application we use excess stock returns, $Y_{jt} = R_{jt} - r_t$, where R_{jt} is the log return of stock j at time t and r_t is the one-month Treasury bill rate (risk-free rate). In case of county population growth, Y_{jt} is the percentage change in county j's population at time t.

Thus we run regression (including a constant):

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

We repeat this for all j = 1, ..., m entities. We use these regressions to construct estimates of the X_{jt} . To estimate the tail index for Y_{jt} , \hat{X}_{jt} and S_{jt}^{f} , 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 + 1 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.

4 Results

4.1 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. The first two rows in Table 1 show, as a crude initial attempt at capturing the relation, the partial correlations between the asset pricing factors and $\hat{\alpha}_t^Y$. Note that, to interpret the signs in the table correctly we report the correlations with the tail index, i.e., the inverses of the Hill estimates.

The correlation between tail index estimates and the market factor is particularly strong and is of the conjectured sign. 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 B.1 in Appendix B 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 dilute the effect of the bias caused by a single factor.

	Market	SMB	HML	Mom	Liq			Market	SMB	HML	Mom	Liq
$\hat{\alpha}_{t-}^{Y}$	-0.69	-0.45	0.19	0.09	0.05	â	.Y 't —	-0.31	-0.25	0.04	0.10	0.00
\hat{lpha}_{t+}^Y	0.75	0.52	-0.12	-0.15	-0.02	â	Y + t +	0.33	0.33	-0.02	-0.13	0.01
Δ_{t-}^{f}	-0.81	-0.81	-0.24	0.47	-0.05	Δ	f_{t-}	-0.51	-0.49	0.04	0.25	0.01
Δ_{t+}^{f}	0.85	0.85	0.39	-0.42	-0.04	Δ	f_{t+}	0.54	0.39	0.07	0.13	-0.02
$\rho(\Delta_{t-}^f, \Delta_{t+}^f)$	-0.91	-0.83	-0.07	-0.28	-0.02	$\underline{\rho(\Delta_{t-}^f,\Delta_{t-}^f)}$	+)	-0.39	-0.24	-0.11	0.15	0.03

Table 1: Correlations cross-sectional tail index and factors

(a) Threshold u at 5% of sample fraction

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

To isolate the bias a single factor induces in $\hat{\alpha}_{t+}^{Y}$, we use Δ_{t+}^{f} as defined in (11). Panel (a) of Figure

This table reports the correlations between the cross-sectional Hill estimates and an individual factor. In the first and second row of each panel $\hat{\alpha}_{t-}^{Y}$ 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 Δ_{t}^{f} is the cross-sectional tail index estimate where factor f's effect is isolated, as defined in (11). The sign "-" ("+"), indicates that the estimate is made on the left (right) tail of the distribution. The factor with which the correlation is calculated is reported in the first row. The **last** row reports the correlation between the left and right tail estimates of Δ_{t}^{f} for the respective factors. The left panel presents the correlations where the threshold u is set to 5% of the sample fraction, and for the right panel this threshold is set to 0.5% of the sample fraction.

1 plots the (normalized) time series of the market factor and Δ_{t+}^M . The time series illustrate the clear positive relationship between Δ_{t+}^M and M_t , as predicted by the inverse of the bias in (12).



Figure 1: M_t and Δ_t^M

These figures present the time series of the market factor (solid black line), its isolated effect on right tail Δ_{t+}^{M} (blue dash-dotted line) and left tail (red dashed line). Panel (a) shows the time series of the market factor and Δ_{t+}^{M} , which are normalized due to a difference in units. Panel (b) contrasts the right tail and left tail estimate of (not normalized) Δ_{t}^{M} . The data have been annualized by averaging the monthly estimates for a given year.

In Section 2, we derived that a shift in the factor induces a bias in left and the right tail index estimates of opposite sign. To this end, we plot the Δ_{t+}^{M} and Δ_{t-}^{M} in panel (b). The two estimates indeed appear to be each other's mirror image. This confirms the predictions from Proposition 1 and Corollary 1. In the online Appendix similar figures are presented for the SMB, HML, momentum and liquidity factors. While weaker, the relationship for the SMB and HML factors shows substantial negative co-movement between the estimate of the tail index in the right and the left tail. The momentum and liquidity factors show a 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 their correct specification. A number of these constructed factors are possibly poor proxies for the factors in the DGP, leading to the weaker relationships. Another possible explanation is the time varying explanatory power of asset pricing factors (see Hwang and Rubesam (2015)).

We summarize the patterns observed in Figure 1 for all factors by means of correlations in rows three to five of Table 1. The third and fourth 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 shown in the first two rows. 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. The correlations for the momentum factor have a sign that remains somewhat counter-intuitive. One possibility is that the observations included in the tail measurement have negative coefficients. In unreported results we multiply the factor realization by the average of the coefficients found in that month's cross section, which alters the signs in the predicted direction.

The last rows of Table 1 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 1. 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.

Proposition 1 implies that the bias originating from the factors in $\hat{\alpha}_{t-}^{Y}$ diminishes as threshold *u* increases. In panel (b) of Table 1, 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. But using 0.5% is about how deep one can go into the tail area. Fortunately, the two novel bias correction methods are less data demanding.

In Table 2, we report on regressions that investigate the degree to which variation in the isolated bias Δ_t^f explains variation in the cross-sectional Hill estimate of Y_t , i.e., $\hat{\alpha}_t^Y$. Panel (a) shows results for the right tail index estimates. The coefficients for the Market and SMB factors are significantly different from zero, resonating the strong correlations in Table 1. The R^2 of the first regression for the right tail shows that about 42% of the variation in the cross-sectional tail index is driven by the market factor. The second most important factor is the SMB factors have a marginal role in explaining variation in $\hat{\alpha}_{t+}^Y$. The HML, momentum and liquidity factors have a marginal role in explaining variation in $\hat{\alpha}_{t+}^Y$. Similar regression results for the left tail are reported in the right side of the table. The contribution of the individual factors are quantitatively similar in the left tail.⁹

⁹Regression results (unreported) for the factors (instead of Δ_t^f) are very similar.

				left tail								
	Δ^M_{t+}	Δ_{t+}^{SMB}	Δ_{t+}^{HML}	Δ_{t+}^{Mom}	Δ_{t+}^{Liq}	$\hat{lpha}_{t+}^{\hat{X}}$	Δ^M_{t-}	Δ_{t-}^{SMB}	Δ_{t-}^{HML}	Δ_{t-}^{Mom}	Δ_{t-}^{Liq}	$\hat{lpha}_{t-}^{\hat{X}}$
coef.	0.95***	0.90***	0.31	-0.40	-0.25	0.52***	0.96***	0.82***	0.00	-0.46	0.79**	0.78***
s.e.	(0.05)	(0.10)	(0.27)	(0.31)	(0.40)	(0.08)	(0.05)	(0.10)	(0.30)	(0.32)	(0.38)	(0.06)
\mathbb{R}^2	0.42	0.12	0.00	0.00	0.00	0.07	0.33	0.10	0.00	0.00	0.01	0.24

Table 2: Regression cross-sectional tail index and factors

This table presents the regression results for the effect of the factors on the cross-sectional Hill estimate. The dependent variable in the left panel is the Hill estimate for the right tail of the raw cross-sectional excess returns $(\hat{\alpha}_{t+}^{Y})$. The independent variable is the cross-sectional tail index where the factor *f*'s effect is isolated (Δ_{t+}^{f}) . In the last column of the left panel, $\hat{\alpha}_{t+}^{\hat{X}}$ is the tail index estimated on the fitted idiosyncratic noise terms of the five-factor asset pricing model. We include an unreported constant in the regressions, which are significant and positive for all regressions. The right side illustrates the results for the left tail of the distribution. The threshold *u* used to estimate the Hill estimate is set to 5% of the sample fraction. The asterisks in the table indicate: * p<0.1; ** p<0.05; *** p<0.01.

The explanatory power of the idiosyncratic part of the linear factor model explains only about 7% of the variation in the right tail index estimate. 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 tail index of the idiosyncratic shocks is 24%. Aside from correlated measurement errors and variation in α , we may not have isolated all the factors that influence stock returns.

Unreported results of a multi-variable regression to investigate the contribution of all individual factors together on $\hat{\alpha}_{t+}^{Y}$, shows that the contributions of all factors are significant and produce 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 are comparable.

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

4.1.1 Bias correction for US stock returns

The analysis above provides ample evidence for bias arising from factors. Therefore, we subject the US stock return data to the two bias reduction methods. In Figure 2, we show the results of bias correction, by presenting the time series of uncorrected tail-index estimates in tandem with bias-corrected estimates.



This figure presents the time series of tail index estimates for US stock returns before and after applying the proposed bias reduction methods. The y-axis shows the value of the estimates in terms of α . In panel (a) the red triangles (\checkmark) show the Hill estimates for the left tail of $Y_{jt} - \overline{Y}_t$. The black open downward triangles (\bigtriangledown) are tail index estimates extracted from Y_{jt} on the left side of the distribution. The bars surrounding the black triangular estimates are sized to be two times the standard errors of these estimates. In panel (b) the blue triangles (\bigstar) show the Hill estimates for the right tail of $Y_{jt} - \overline{Y}_t$. The black open upward pointing triangles (\bigtriangleup) are tail index estimates extracted from Y_{jt} on the right side of the distribution. The green diamonds (\diamondsuit) in panel (c) are the estimates of the tail index after correcting for bias using the mirror method. Panel (d) shows all three bias reduction estimates in a single plot. All 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.

Panels (a) and (b) in Figure 2 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 (1973) and second oil crisis (1979), Black Monday (1987), the dot-com bubble burst (2001) and the credit crisis (2008). The bias-corrected estimates $\hat{\alpha}_{t-}^{Y-\bar{Y}}$ for these crisis periods show that the bias correction is substantial and significant. The corrected estimates are more in line with the preceding and succeeding estimates. This can be understood from (7). Due to the negative value of the shift parameter *h*, capturing the large declines in the market factor during crisis periods, the estimated $1/\alpha_{t-}^{Y}$ 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). The bias corrected estimates imply that there are only few firms with large positive returns in times of crisis, causing significant deviations from biased estimates.

Panels (c) present the mirror method for bias correction with inclusion of the (inverse) Hill estimate on US stock returns, $\hat{\alpha}_{t-}^{Y}$. Note that the green diamonds (mirror method estimates) also give a substantial and significant 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 $\hat{\alpha}_{t-}^{Y-\bar{Y}}$. Both correction methods, successfully dampen the effects of large factor realizations. This correspondence is less clearly observed when comparing $\hat{\alpha}_{t-}^{mirror}$ and the shift estimate in the right tail $\hat{\alpha}_{t+}^{Y-\bar{Y}}$ in panel (c) and (b), respectively. The mirror estimate does not increase to the same degree as the shift estimate in the right tail during economic crises. Correcting for the large negative shift, reveals that during crisis times $\hat{\alpha}_{t+}^{Y-\bar{Y}}$ indicates that large positive stock returns are even more rare than the uncorrected estimates would indicate. Panel (d) presents the three bias-corrected estimates and the close correspondence between $\hat{\alpha}_{t-}^{Y-\bar{Y}}$ and $\hat{\alpha}_{t-}^{mirror}$.

We are now able to investigate whether the bias, likely caused by some of the factors, is still present after bias correction. To enable a direct comparison between Table 2 and Table 3, the independent variable Δ_{t+}^{f} is left unchanged, while in Table 3 the dependent variable $\hat{\alpha}_{t}^{Y}$ is bias-corrected. The left side of Table 3 gives the effect of the shift method and the right side for the mirror method.

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

				Table	3: Bias-	-corrected	α and	factors				
			$\hat{\alpha}_{t}^{Y}$	\bar{Y}		$\hat{\alpha}_{t}^{mirror}$						
	Δ_{t-}^{M}	Δ_{t-}^{SMB}	Δ_{t-}^{HML}	Δ_{t-}^{Mom}	Δ_{t-}^{Liq}	$\hat{\pmb{lpha}}_{t-}^{\hat{X}}$	Δ_{t-}^{M}	Δ_{t-}^{SMB}	Δ_{t-}^{HML}	Δ_{t-}^{Mom}	Δ_{t-}^{Liq}	$\hat{lpha}_{t-}^{\hat{X}}$
coef.	0.02	-0.13*	-0.42**	-0.34	0.67***	0.99***	0.03	0.09*	-0.22^{*}	-0.06	0.25	0.58***
s.e.	(0.05)	(0.08)	(0.20)	(0.22)	(0.26)	(0.02)	(0.03)	(0.05)	(0.13)	(0.14)	(0.16)	(0.02)
\mathbb{R}^2	0.00	0.01	0.01	0.00	0.01	0.76	0.00	0.01	0.01	0.00	0.00	0.66

T 1 1 2 D' . 1 . 1 . .

This table presents the regression results for the two bias reduction methods. In left side of the table the dependent variable is the Hill estimate for the left tail of $Y_j - \overline{Y}$. In the right side of the table the dependent variable is the average of the left and right tail estimate on the Y_j , i.e., $\hat{\alpha}_t^{mirror}$. The independent variable is the cross-sectional tail index where the factor's effect is isolated $(\Delta_{I_{-}}^{f})$. We include an unreported constant in the regressions. With the exception of the sixth column $(\hat{\alpha}_{t-}^{\hat{\chi}})$, the constants are significant and positive. The threshold *u* to estimate the Hill estimate is set to 5% of the sample fraction. The asterisks in the table indicate: * p<0.1; ** p<0.05; *** p<0.01.

factors. Moreover, the R^2 of the regressions should decrease. The results for $\hat{\alpha}_{t-}^{Y-\bar{Y}}$ show that the isolated bias with respect to the market (Δ_{t-}^M) and SMB factor (Δ_{t-}^{SMB}) lose almost all of their explanatory power. While the estimated coefficients for both factors were close to 1 in Table 2, 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. Bias reduction thus breaks the link between the bias found in the Hill estimate and the factors. Although the coefficients for Δ_{t-}^{HML} and Δ_{t-}^{Liq} have become significant, the explanatory power remains low. The R^2 for the regression of $\hat{\alpha}_{t-}^{Y-\bar{Y}}$ on $\hat{\alpha}_{t-}^{\hat{X}}$ has increased to 0.76 (and a similar result is found 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 the right side of the table show that the relationship between the bias and the estimates decreases significantly. For the original isolated bias of the market and SMB factors, both coefficients become almost zero. Additionally, the R^2 values decrease close to zero for all factors. The relationship for the HML, momentum and liquidity factors are largely unchanged.

4.2 **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 PCs explain about 60% of the variation in our original variables. Table B.3 in the Appendix presents summary statistics of the PCs.

	PC1	PC2	PC3	PC4	PC5			PC1	PC2	PC3	PC4	PC5
\hat{lpha}_{t-}^Y	-0.45	-0.09	-0.34	0.18	0.28	â	ℓ_{t-}^{Y}	-0.18	-0.17	-0.44	0.08	-0.05
$\hat{\alpha}_{t+}^{Y}$	-0.07	0.02	0.06	0.01	-0.05	â	ℓ_{t+}^{Y}	0.07	-0.08	-0.31	-0.08	0.06
Δ_{t-}^{f}	-0.83	-0.22	-0.74	0.75	0.59	Δ	f_{t-}	-0.66	0.01	-0.39	-0.09	-0.16
Δ_{t+}^f	0.82	0.57	0.75	-0.68	-0.58	Δ	f_{t+}	0.42	0.05	0.40	0.17	0.20
$\underline{\rho(\Delta_{t-}^f, \Delta_{t+}^f)}$	-0.71	0.23	-0.64	-0.41	-0.45	$\underline{\rho(\Delta_{t-}^f, \Delta_{t-}^f)}$	r +)	-0.31	-0.24	-0.15	-0.10	-0.10

Table 4: Correlations of cross-sectional tail indices (county population growth)

(a) Threshold u at 5% of sample fraction

(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 $\hat{\alpha}_{t}^{Y}$ and $\hat{\alpha}_{t+}^{Y}$ are the cross-sectional Hill estimates for the cross section of county population growth for the left and right tail, respectively. The Δ_{t}^{f} , stated in the third and fourth rows of each panel, is the cross-sectional tail index where the effect of the PCs is isolated, as defined in (11). The five factors are the first five principal components from an assortment of variables suggested by the literature. The last row reports the correlation between the left and right tail estimates of Δ_{t}^{f} . The left (right) panel presents the correlations where the threshold *u* is set to 5% (0.5%) of the sample fraction.

In the same vein as for the US stock data, Table 4 presents the correlations for the county population growth data. 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 PCs are isolated, as depicted in rows three and four. The correlation with the first PC increases from -0.45 to -0.83 for the left tail and from -0.07 to positive 0.82 for the right tail. Since the sign of a PC is indeterminate, one should not interpret the direction of the bias. But, as the last row indicates, in accordance with our theory, 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 4 presents the correlations when the threshold is lowered to 0.5%. As with the data on US stock returns, correlations decrease substantially in magnitude 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 5 presents the results of the regressions between the isolated bias of the different individual PCs (Δ_t^f) and the cross-sectional Hill estimate on Y_t , i.e., $\hat{\alpha}_t^Y$. The left side of the table illustrates the results when using the estimate of the right cross-sectional tail index and Δ_{t+}^f , while the right side uses the estimate of the left cross-sectional tail index. We observe that only the isolated bias with

respect to PC1 in the left tail (Δ_{t-}^{PC1}) can significantly account for variation in the cross-sectional tail index estimate of county population change $(\hat{\alpha}_{t-}^Y)$. For the right tail, Δ_{t+}^{PC1} is not significant but attains the highest R^2 of the five principal components. This implies that the previously presented correlations for the PCs most likely come with large standard errors. The high R^2 attained by $\hat{\alpha}_{t-}^{\hat{\chi}}$ and $\hat{\alpha}_{t+}^{\hat{\chi}}$ in the last 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, only leading to small location shifts in the cross-sectional observations. This contrasts with the strong results for financial return data that highlight the role of the strength of the factors in the DGP that drives the bias.¹⁰

									<u>r r r</u>	0		
			rig	ht tail			left tail					
	Δ_{t+}^{PC1}	Δ_{t+}^{PC2}	Δ_{t+}^{PC3}	Δ_{t+}^{PC4}	Δ_{t+}^{PC5}	$\hat{lpha}_{t+}^{\hat{X}}$	Δ_{t-}^{PC1}	Δ_{t-}^{PC2}	Δ_{t-}^{PC3}	Δ_{t-}^{PC4}	Δ_{t-}^{PC5}	$\hat{lpha}_{t-}^{\hat{X}}$
coef.	-0.18	0.006	0.08	0.16	-0.15	0.66***	0.41*	0.13	0.32	0.29	0.44	0.63***
s.e.	(0.18)	(0.29)	(0.30)	(0.30)	(0.23)	(0.11)	(0.21)	(0.51)	(0.33)	(0.40)	(0.38)	(0.12)
R ²	0.02	0.00	0.00	0.007	0.01	0.44	0.08	0.00	0.02	0.01	0.03	0.38

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

This table presents the regression results for the cross-sectional Hill estimate extracted from US county level population growth. The dependent variable in the left (right) side of the table is $\hat{\alpha}_{t+(-)}^{f}$, i.e., the Hill estimate for the right (left) tail of the cross-sectional county level population growth. The independent variable Δ_{t+}^{f} , given in the first row, is the cross-sectional tail index where the effect is isolated with respect to the given PC, as defined in (11). We include an unreported constant in the regressions, which are significant and positive for all regressions. The five PCs are the first five principal components from an assortment of variables suggested by the literature. The tail index estimated on the disturbance terms $(\hat{\alpha}_{t+}^{A})$ is given in the last column of each side. The Hill estimate is calculated by setting the threshold *u* at 5% of the sample fraction. The asterisks in the table indicate the following: * p<0.1; ** p<0.05; *** p<0.01.

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 in the tail index estimates in the right and left tail are negatively correlated. Thus, even when investigating factors with marginal explanatory power, inference on the basis of cross-sectional tail index estimates may lead to incorrect conclusions.

¹⁰Table B.4 in Appendix B presents the regression results for a 0.5% threshold. The coefficient for Δ_{t+}^{PC2} becomes significant. As this is contained to the left tail only, it is possibly caused by correlated measurement errors in Δ_{t+}^{PC2} and $\hat{\alpha}_{t+}^{Y}$.

5 Conclusion

We show that tail index estimates representing scaling behavior extracted from a 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 in a systematic fashion. For the left and right tail of the distribution the sign of the bias moves in exactly the opposite direction. This data feature has, as of today, gone undetected. It offers an opportunity to correct for the bias induced by the factors. We propose two methods to alleviate this bias. Moreover, the bias also diminishes by looking deeper into the tail, but this may run against data limitations of the cross section.

We find that data from a DGP with factors that have strong explanatory power, as is the case for US stock return data, show considerable cross-sectional bias, which dominates fluctuations in tail index estimates. In data with factors that have little explanatory power (US county population growth) this bias is present, but weaker.

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. In future cross-sectional studies regarding tail estimates, gauging the influence of factors is crucial for unbiased interpretation of the results. In the paper, several ways are proposed to correct for the possible bias that factors may cause.

A Asymptotic distribution of the Hill Estimator

Below we provide a concise version of the steps to derive the asymptotic distribution of the Hill estimator for cross-sectional observations, with varying second-order scale parameters, as in (9). Furthermore, for consistency with the main text, we set $D_j = \alpha h_j$ and $\bar{H} = 1/m \sum h_j$. A detailed exposition of the following derivation is available on request. See Haan and Ferreira (2006, p.76) for the standard case. Consider the Hill estimator

$$\frac{1}{\widehat{\alpha}}(u_m) = \frac{\frac{1}{m}\sum_{j=1}^m \left(\ln\frac{Y_j}{u_m}\right) \mathbb{1}_{Y_j > u_m}}{\frac{1}{m}\sum_{j=1}^m \mathbb{1}_{Y_j > u_m}} = \frac{I_1(u_m)}{I_2(u_m)},$$
(A.1)

say, where $\mathbb{1}_{Y_j > u_m}$ is an indicator function for order statistics larger than u_m . Assume the idiosyncratic shocks are drawn from a heavy-tailed distribution that satisfies (8), implying that the tail expansion of the Y_j distribution follows (9). By application of the Lindeberg-Feller theorem and its Corollary stated in Serfling (1980), we obtain the asymptotic normality (AN) of I_1 and I_2 in (A.1) as:

$$\begin{pmatrix} I_1 \\ I_2 \end{pmatrix} \operatorname{are} \begin{pmatrix} AN\left(\frac{1}{\alpha}Cu_m^{-\alpha} + \frac{\alpha}{\alpha+1}C\bar{H}u_m^{-\alpha-1}(1+o(1)), \frac{2}{m}\frac{1}{\alpha^2}Cu_m^{-\alpha}(1+o(1))\right) \\ AN\left(Cu_m^{-\alpha} + C\bar{H}u_m^{-\alpha-1}(1+o(1)), \frac{1}{m}Cu_m^{-\alpha}(1+o(1))\right) \end{pmatrix}.$$

By using Minkowski's inequality one shows that the condition of the Corollary can be invoked, to guarantee that the Lindenberg-Feller theorem holds.

To show that the ratio I_1/I_2 is asymptotically normally distributed, we first show that I_1 and I_2 are jointly normally distributed. Note that the covariance between I_1 and I_2 is:

$$cov[I_1, I_2] = \frac{1}{m} \frac{1}{\alpha} C u_m^{-\alpha} (1 + o(1)).$$

Notice that the covariance is of the same order as the two variances.

Using the results above, it follows that every linear combination of I_1 and I_2 converges to a normal distribution, showing that I_1 and I_2 are jointly normally distributed. Furthermore, for any vector $(z,t) \in \{[0,1] \times [0,1]\} \setminus \{(0,0)\}:^{11}$

$$(zI_1 + tI_2) \text{ is } AN\left(\left[\left(z\frac{1}{\alpha} + t\right) + \left(z\frac{\alpha}{\alpha+1} + t\alpha\right)\overline{H}u_m^{-1}\left(1 + o\left(1\right)\right)\right]Cu_m^{-\alpha}, \\ \frac{1}{m}\left[z^2\frac{2}{\alpha^2} + 2tz\frac{1}{\alpha} + t^2\right]Cu_m^{-\alpha}\left(1 + o\left(1\right)\right)\right).$$

Since we have established the joint asymptotic normality of the components of (A.1), one can obtain the asymptotic normality for the ratio by using Cramér (1974)'s delta argument. A first-order Taylor approximation of h(x, y) = x/y around the point $(I_1(m), I_2(m))$ gives

$$\frac{\mu_{I_1}(m)}{\mu_{I_2}(m)} + \frac{1}{\mu_{I_2}(m)}x - \frac{\mu_{I_1}(m)}{\mu_{I_2}(m)^2}y,$$

where μ_{I_1} and μ_{I_2} are the means of I_1 and I_2 , respectively. This results in the following asymptotic distribution for $1/\hat{\alpha}(u_m)$:

$$\frac{1}{\widehat{\alpha}}(u_m) \text{ is } \frac{1}{\alpha} - \frac{1}{\alpha+1}\overline{H}u_m^{-1}(1+o(1)) + AN\left(0, 2\frac{1}{m}\frac{1}{\alpha^2}\frac{u_m^{\alpha}}{C}\right) - AN\left(0, \frac{1}{m}\frac{1}{\alpha^2}\frac{u_m^{\alpha}}{C}\right).$$

Then using $cov[I_1, I_2]$ to derive the variance of the combined distribution gives

$$\frac{1}{\hat{\alpha}(u_m)} - \frac{1}{\alpha} \text{ is } AN\left(-\frac{1}{\alpha+1}\bar{H}u_m^{-1}\left(1+o\left(1\right)\right), \frac{1}{m}\frac{1}{\alpha^2}\frac{u_m^{\alpha}}{C}\right).$$
(A.2)

If the threshold u_m increases at the rate $m^{1/(\alpha+2)}$ and define $\lim_{m\to\infty} \bar{H} = \bar{H}$, then we have the following asymptotic result. For $m \to \infty$

$$m^{1/(\alpha+2)}\left(\frac{1}{\widehat{\alpha}}(m)-\frac{1}{\alpha}\right) \xrightarrow{d} N\left(-\frac{1}{\alpha+1}\overline{\overline{H}},\frac{1}{\alpha^2}\frac{1}{C}\right).$$

¹¹The analysis for the other three quadrants follows analogously.

Finally, one can show that the asymptotic MSE minimizing optimal rate at which u_m increases is

$$w(m) = C^{\frac{2}{\alpha+2}} \left\{ 2\alpha \left(\frac{\bar{H}}{\alpha+1} \right)^2 \right\}^{-\frac{\alpha}{\alpha+2}} m^{\frac{2}{\alpha+2}}.$$

This rate accounts for tail observations coming in more slowly than at the usual speed \sqrt{m} . Using w(m) we can restate (A.2) as

$$\sqrt{w(m)} \left(\frac{1}{\widehat{\alpha}} (u_m) - \frac{1}{\alpha} \right) \xrightarrow{d} N \left(-\frac{1}{\sqrt{2\alpha}} \cdot \operatorname{sgn}(\bar{\bar{H}}), \frac{1}{\alpha^2} \right).$$
(A.3)

From which the influence of \overline{H} of the sign of the bias is immediate.

B Tables

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

Table B.1: Correlation asset pricing factors.

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

Table B.2: Regres	sion cross-secti	onal tail index	x 0.5%	threshold
-------------------	------------------	-----------------	--------	-----------

			right	tail			left tail						
	Δ^M_{t+}	Δ^{SMB}_{t+}	Δ_{t+}^{HML}	Δ_{t+}^{Mom}	Δ_{t+}^{Liq}	$\hat{lpha}_{t+}^{\hat{X}}$	Δ^M_{t-}	Δ_{t-}^{SMB}	Δ_{t-}^{HML}	Δ_{t-}^{Mom}	Δ_{t-}^{Liq}	$\hat{lpha}_{t-}^{\hat{X}}$	
coef.	0.32***	0.47***	0.09	-0.06	0.27*	0.80***	0.38***	0.41***	-0.04	-0.06	0.17	0.81***	
s.e.	(0.08)	(0.09)	(0.12)	(0.14)	(0.14)	(0.03)	(0.10)	(0.10)	(0.13)	(0.15)	(0.14)	(0.03)	
R ²	0.02	0.04	0.00	0.00	0.01	0.51	0.02	0.02	0.00	0.00	0.00	0.58	

This table presents the regression results for the effect of the factors on the cross-sectional Hill estimate for a lower threshold *u*. Here *u* is set to **0.5%** of the sample fraction. The dependent variable in the left (right) side of the table is $\hat{\alpha}_{t+(-)}^{Y}$, i.e., the Hill estimate for the right (left) tail of the raw cross-sectional excess returns. The independent variable Δ_{t+}^{f} , stated in the first row, is the cross-sectional tail index where the factor *f*'s effect is isolated, as defined in (11). Furthermore, $\hat{\alpha}_{t+}^{\hat{X}}$ is the tail index estimated on the fitted disturbance terms of the five-factor asset pricing model: the market, SMB, HML, momentum (Mom) and the liquidity (Liq) factor. We include an unreported constant in the regressions, which are significant and positive for all regressions. The asterisks in the table indicate: * p<0.1; ** p<0.05; *** p<0.01.

Table B.3: 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

This table presents a summary of the PCs extracted from variables to explain changes in county population, as discussed in the data section. The first two rows give the standard deviation and the proportion of variance explained by each principal component.

			righ	t tail				left tail						
	Δ_{t+}^{PC1}	Δ_{t+}^{PC2}	Δ_{t+}^{PC3}	Δ_{t+}^{PC4}	Δ_{t+}^{PC5}	$\hat{lpha}_{t+}^{\hat{X}}$	Δ_{t-}^{PC1}	Δ_{t-}^{PC2}	Δ_{t-}^{PC3}	Δ_{t-}^{PC4}	Δ_{t-}^{PC5}	$\hat{lpha}_{t-}^{\hat{X}}$		
coef.	0.49	-1.43*	-0.31	-0.52	0.34	1.04***	0.04	-0.19	0.42	-0.21	-0.14	0.83***		
s.e.	(0.43)	(0.67)	(0.38)	(0.64)	(0.72)	(0.25)	(0.21)	(0.24)	(0.33)	(0.37)	(0.33)	(0.11)		
R ²	0.03	0.10	0.02	0.02	0.01	0.28	0.00	0.02	0.04	0.01	0.00	0.57		

This table presents regression results for the effect of the PCs on the cross-sectional Hill estimates extracted from US county level population growth for a lower threshold *u*, set at **0.5**% of the sample fraction. Here the dependent variable in the left (right) side of the table is $\hat{\alpha}_{t+(-)}^{Y}$, i.e., the Hill estimate for the right (left) tail of the cross-sectional county level population growth. The independent variable Δ_{t+}^{f} , stated in the first row, is the cross-sectional tail index where the PC *f*'s effect is isolated, as defined in (11). Furthermore, $\hat{\alpha}_{t+}^{\hat{X}}$ is the tail index estimated on the fitted disturbance terms of a model with the first five PC's. These PCs are extracted from an assortment of variables stated in the online Appendix. We include an unreported constant in the regressions. With the exception of the sixth $(\hat{\alpha}_{t+}^{\hat{X}})$ and twelfth column $(\hat{\alpha}_{t-}^{\hat{X}})$, the constants are significant and positive. The asterisks in the table indicate the following: * p<0.1; ** p<0.05; *** p<0.01.

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.
- Almeida, C., G. Freire, R. Garcia, and R. Hizmeri (2022). "Tail risk and asset prices in the short-term". In: *SSRN Working Paper*.
- Andersen, G. T., Y. Ding, and V. Todorov (2024). "The granular origins of tail dispersion risk". In: *Working Paper*, pp. 1–36.
- 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.
- Beirlant, J., F. Caeiro, and M. I. Gomes (2012). "An overview and open research topics in statistics of univariate extremes". In: *REVSTAT-Statistical Journal* 10.1, pp. 1–31.
- 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.
- Cramér, H. (1974). *Mathematical methods of statistics (PMS-9), Volume 9*. Princeton university press.
- 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. (2004). "Gibrat's law for (all) cities". In: *American Economic Review* 94.5, pp. 1429–1451.
- Einmahl, J. H. J., L. Haan, and C. Zhou (2016). "Statistics of heteroscedastic extremes". In: *Journal of the Royal Statistical Society Series B* 78.1, pp. 31–51.
- Einmahl, J. H. J. and Y. He (2023). "Extreme value estimation for heterogeneous data". In: *Journal* of Business & Economic Statistics 41.1, pp. 255–269.
- Faias, J. A. (2023). "Predicting the equity risk premium using the smooth cross-sectional tail risk: The importance of correlation". In: *Journal of Financial Markets* 63, p. 100769.
- 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.
- Haan, L. de and A. Ferreira (2006). Extreme value theory: an introduction. Springer.
- 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 (2011). "The area and population of cities: New insights from a different perspective on cities". In: *American Economic Review* 101.5, pp. 2205–25.
- Serfling, R. J. (1980). Approximation theorems of mathematical statistics. John Wiley & Sons.
- 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.