A simple modification of the Hill estimator with applications to robustness and bias reduction

Keith Knight
University of Toronto

Abstract: Suppose that $X_1, \cdots, X_n$ are i.i.d. random variables with $P(X_i > x) = x^{-\alpha} L(x)$ and define $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ to be the order statistics. The Hill estimator (Hill, 1975) of the tail index α is a pseudo-maximum likelihood estimator based on the exponential approximation of the normalized log-spacings $Y_j = j \ln(X_{(j)}/X_{(j+1)})$ for $j = 1, \cdots, k$. In practice, the Hill estimator can be extremely dependent on the choice of $k = k_n$ and is inherently non-robust to large values Y_j, which bias the Hill estimator downward. In this paper, we introduce a simple robustification of the Hill estimator that has a bounded influence curve and is Fisher consistent. The estimator is straightforward to compute and can be tuned to have a specified asymptotic efficiency (with respect to the Hill estimator) between 0 and 1. The resulting family of estimators can also be used to reduce the asymptotic bias of the Hill estimator. We also consider extensions to modifications of the Hill estimator based on exponential regression methods (Feuerverger and Hall, 1999; Beirlant *et al.*, 1999).

1 Introduction

Suppose that $X_1, \cdots, X_n$ be i.i.d. random variables whose distribution satisfies

$$P(X_i > x) = x^{-\alpha} L(x) \quad \text{for } x \geq 0 \tag{1}$$

where $L(x)$ is a slowly varying function. A commonly used estimator of α is the Hill estimator (Hill, 1975) defined for some k_n by

$$\widehat{\alpha}_n = \left\{ \frac{1}{k_n} \sum_{i=1}^{k_n} \ln(X_{(i)}/X_{(k_n+1)}) \right\}^{-1}$$

where $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(k_n+1)}$ are the $k_n + 1$ largest order statistics of $X_1, \cdots, X_n$.

The Hill estimator has been used in a variety of applications and its strengths and weaknesses are well known; see, for example, Drees *at al.* (2000). In finance, modeling the tails of the distribution of returns is important in the evaluation of risk (Embrechts *et al.*, 1997; Brooks *et al.*, 2005).

To motivate our proposed family of estimators, we begin by giving the simple derivation of the Hill estimator. The Hill estimator is justified by considering the point process of

$\{X_i/t\}$ above some high threshold t. In particular, if t is sufficiently large, we have for $x > 1$,

$$\begin{aligned} P\left(X_i/t > x | X_i > t\right) &= \frac{(tx)^{-\alpha} L(tx)}{t^{-\alpha} L(t)} \\ &\approx x^{-\alpha} \end{aligned}$$

If $X_{(k_n+1)} \geq t$ then the random variables

$$Y_j = j \ln(X_{(j)}/X_{(j+1)}) \quad (j = 1, \cdots, k_n) \tag{2}$$

will behave like independent exponential random variables with mean approximately $1/\alpha$. This leads to the pseudo-maximum likelihood estimator

$$\begin{aligned} \widehat{\alpha}_n &= \left\{ \frac{1}{k_n} \sum_{j=1}^{k_n} j \ln(X_{(j)}/X_{(j+1)}) \right\}^{-1} \\ &= \left\{ \frac{1}{k_n} \sum_{i=1}^{k_n} \ln(X_{(i)}/X_{(k_n+1)}) \right\}^{-1} \end{aligned}$$

which is simply the Hill estimator.

Using this formulation, it is clear that the Hill estimator is not sensitive so much to large values of X_i but rather to large values of the normalized log-spacings $\{Y_j\}$ defined in (2). Thus it makes sense to robustify the pseudo-maximum likelihood estimation based on the exponential distribution to limit the influence of large Y_j. At the same time, the exponential approximation of Y_j becomes less sound as j increases, which leads to bias in the Hill estimator.

As an example, we will consider data on calcium concentrations in soil samples from a particular city in the Condroz region of Belgium; see chapter 6 (section 6.1) of Beirlant *et al.* (2004) for more details. (A small amount of noise has been added to these data to avoid exact zeroes in $\{Y_j\}$.) A Pareto plot of the data is given in Figure 1; this plot seems to indicate that the largest six observations are somewhat anomalous. However, if one looks at the values of Y_j for $j = 1, \cdots, 100$ given in Figure 2, it seems that only two values are anomalous on this scale.

A number of robust estimators of the exponential parameter have been proposed; see, for example, Ahmed *et al.* (2005) as well as Gather and Schultze (1999). Gather and Schultze (1999) propose to estimate the exponential parameter based on scale equivariant statistical functionals having bounded influence functions while Ahmed *et al.* (2005) use weighted maximum likelihood estimation with extreme observations downweighted. In the context of

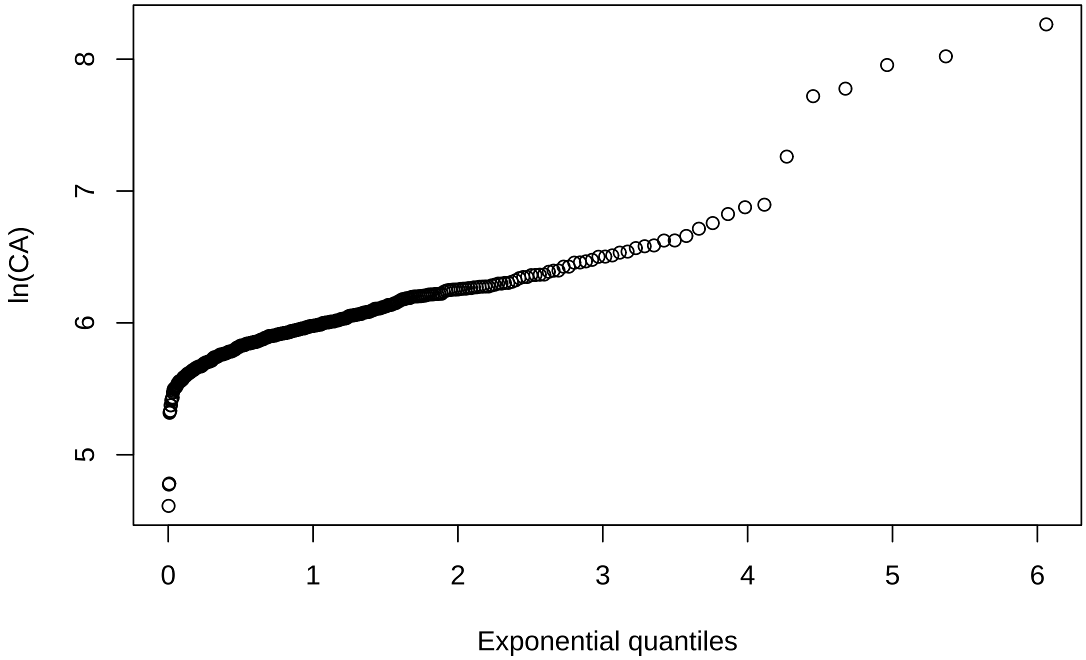

Figure 1: Pareto plot of Condroz calcium data

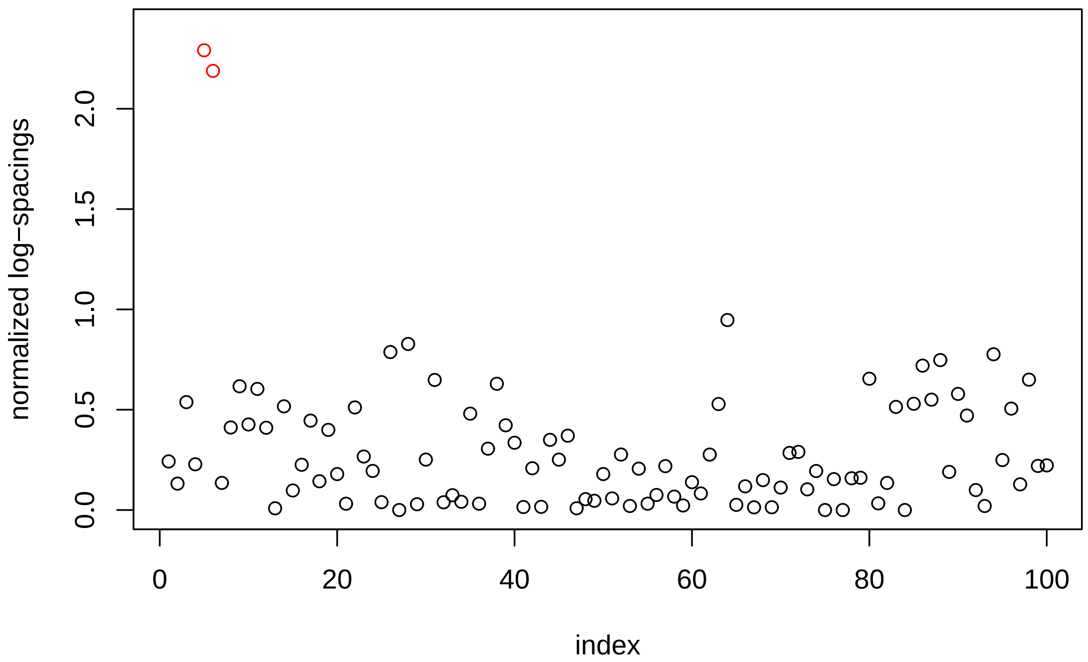

Figure 2: Plot of Y_j $(j = 1, \cdots, 100)$ for Condroz calcium data

tail index estimation, Vandewalle *et al.* (2004) and Vandewalle *et al.* (2007) have considered different approaches to robust estimation. Schluter and Trede (2008) consider methods for outlier identification in data assumed to have come from a heavy tailed distribution.

2 Robust estimator

To define the robust estimator of α, we start by defining the function

$$\psi_c(x;\alpha) = \begin{cases} x - \phi(c)/\alpha & \text{if } x \le \{c + \phi(c)\}/\alpha \\ c/\alpha & \text{otherwise} \end{cases}$$

where $\phi(c)$ depends on the tuning constant $c > 0$ so that

$$\int_0^\infty \psi_c(x;\alpha) f_\alpha(x)\, dx = 0$$

where $f_\alpha(x) = \alpha \exp(-\alpha x)$ is the exponential density function; this guarantees Fisher consistency. The estimator $\widetilde{\alpha}_n(c)$ then satisfies the equation

$$\sum_{j=1}^{k_n} \psi_c(Y_j; \widetilde{\alpha}_n(c)) = 0. \tag{3}$$

The maximum likelihood estimator of α corresponds to $c = +\infty$. Computationally, it is somewhat easier to work with $\bar{\psi}_c(x;\alpha) = \alpha\psi_c(x;\alpha)$ since the solution of (3) is the same as the solution of

$$\sum_{j=1}^{k_n} \bar{\psi}_c(Y_j; \widetilde{\alpha}_n(c)) = 0. \tag{4}$$

For a given c, $\phi(c)$ can be determined by noting that

$$\begin{aligned}\int_0^\infty \psi_c(x;\alpha) f_\alpha(x)\,dx &= \frac{1}{\alpha}\int_0^{c+\phi(c)} \{t - \phi(c)\}\exp(-t)\,dt + \frac{1}{\alpha}\int_{c+\phi(c)}^\infty c\exp(-t)\,dt \\ &= \frac{1 - \phi(c) - \exp(-\{c+\phi(c)\})}{\alpha}\end{aligned}$$

Thus $\phi(c)$ is the solution of the equation

$$\phi(c) + \exp(-\{c + \phi(c)\}) = 1,$$

which gives

$$\phi(c) = \mathcal{W}(-1/\exp(c+1)) + 1$$

where $\mathcal{W}(x)$ is the principal branch of the Lambert W function (Corless *et al.*, 1996) defined by the equation

$$\mathcal{W}(x)\exp(\mathcal{W}(x)) = x$$

for $x \geq -1/e \approx -0.368$ with the constraint $\mathcal{W}(x) \geq -1$. $\mathcal{W}(x)$ is a real-valued function for $x \geq -1/e$ with $\mathcal{W}(-1/e) = -1$ and has the same sign as x. [1] Thus $\phi(c)$ is always well-defined and lies between 0 and 1 for $c > 0$; in fact, $\phi(c)$ provides an alternative parametrization for the tuning parameter with $\phi^{-1}(u) = -\ln(1-u) - u$.

Next, we will give the asymptotic properties of $k_n^{1/2}(\widetilde{\alpha}_n(c) - \alpha)$ as $k_n \to \infty$ at an appropriate rate, which will depend on the nature of the slowly varying function L in (1). Standard asymptotic theory suggests that

$$k_n^{1/2}(\widetilde{\alpha}_n(c) - \alpha) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\alpha, c))$$

[1] For $x < 0$, $\mathcal{W}(x)$ can be evaluated using the fixed point iteration

$$w_{k+1} = x\exp(-w_k)$$

with $w_0 = x$; more generally, it can be computed using an iterative algorithm described in Corless *et al.* (1996).

Efficiency	c	$\phi(c)$
0.99	4.25	0.995
0.95	2.57	0.971
0.90	1.84	0.938
0.75	0.91	0.823
0.50	0.30	0.590
0.25	0.06	0.314

Table 1: Values of c and $\phi(c)$ for various relative efficiencies.

as $k_n \to \infty$ where

$$\begin{aligned}\sigma^2(\alpha, c) &= \frac{\displaystyle\int_0^\infty \psi_c^2(x;\alpha) f_\alpha(x)\,dx}{\left\{\displaystyle\int_0^\infty \psi_c'(x;\alpha) f_\alpha(x)\,dx\right\}^2} \\ &= \alpha^2 \frac{h^2(c) - 2(c+1)h(c) + 1}{\{h^2(c) - (2+c)h(c) + 1\}^2} \end{aligned} \tag{5}$$

where $\psi_c'(x;\alpha)$ is the partial derivative of ψ_c with respect to α and

$$h(c) = -\mathcal{W}(-1/\exp(c+1)) = 1 - \phi(c). \tag{6}$$

Note that as $c \to \infty$, $ch(c) \to 0$ and so

$$\lim_{c\to\infty} \sigma^2(\alpha, c) = \alpha^2,$$

which is the limiting variance of the Hill estimator (Hall, 1982) under appropriate regularity conditions.

THEOREM 1. *Suppose that the distribution function F satisfies*

$$1 - F(x) = \lambda x^{-\alpha}\left\{1 + O(x^{-\beta})\right\}$$

for some $\beta > 0$ as $x \to \infty$. Define $\widetilde{\alpha}_n(c)$ as the solution of (3) for some c and k_n. If $\{k_n\}$ satisfies

$$\frac{k_n}{n^{2\beta/(2\beta+\alpha)}} \to 0 \quad \textit{as } n \to \infty$$

then

$$k_n^{1/2}\left\{\widetilde{\alpha}_n(c) - \alpha\right\} \xrightarrow{d} \mathcal{N}(0, \sigma^2(\alpha, c))$$

where $\sigma^2(\alpha, c)$ is defined by (5) and (6).

The proof of Theorem 1 is given in the appendix. Note that the proof implies that $\widetilde{\alpha}_n(c)$ is consistent provided that $k_n = o(n)$.

Table 1 gives the values of c, $\phi(c)$ for a given efficiency relative to the Hill estimator (corresponding to $c = \infty$). Note that $\phi(c)$ is the proportion of observations whose observed value is used in the computation of $\widetilde{\alpha}$ for a given value of c. Likewise, $h(c) = 1 - \phi(c)$ can be interpreted as a one-sided breakdown point; for a given c, $h(c)$ is the asympototic fraction of extremely large observations that can be observed without driving the estimator of α to 0. For purposes of robustness, it typically would not make sense to consider using an estimator with a breakdown point greater than 50%; $h(c) = 0.5$ for $c = 0.193$, which gives an asymptotic relative efficiency of 0.413.

The asymptotic covariance between $k_n^{1/2}\{\widetilde{\alpha}_n(c_1) - \alpha\}$ and $k_n^{1/2}\{\widetilde{\alpha}_n(c_2) - \alpha\}$ is $\alpha^2 K_0(c_1, c_2)$ where for $c_1 \leq c_2$,

$$K_0(c_1, c_2) = \frac{h^2(c_1) - (c_1 + 2)h(c_1) - c_1 h(c_2) + 1}{\{h^2(c_1) - (c_1 + 2)h(c_1) + 1\}\{h^2(c_2) - (c_2 + 2)h(c_2) + 1\}}. \tag{7}$$

Under the assumptions of Theorem 1, $\left\{k_n^{1/2}\{\widetilde{\alpha}_n(c) - \alpha\} : c \geq \epsilon\right\}$ converges weakly to a Gaussian process for any $\epsilon > 0$.

It is possible to obtain a refinement of Theorem 1 to the case where

$$\frac{k_n}{n^{2\beta/(2\beta+\alpha)}} \to r \geq 0 \quad \text{as } n \to \infty.$$

To do this, we need to make a slightly stronger assumption about the $O(x^{-\beta})$ term in $1 - F$ given in Theorem 1. Theorem 2, stated below, gives a representation of the bias of the robust estimator as a function of the tuning parameter c.

THEOREM 2. *Suppose that the distribution function F satisfies*

$$1 - F(x) = \lambda x^{-\alpha}\left\{1 + \theta x^{-\beta} + o(x^{-\beta})\right\}$$

for some $\beta > 0$ as $x \to \infty$ where $-\infty < \theta < \infty$. Define $\widetilde{\alpha}_n(c)$ as the solution of (3) for some c and $k = k_n$. If $\{k_n\}$ satisfies

$$\frac{k_n}{n^{2\beta/(2\beta+\alpha)}} \to r \geq 0 \quad \textit{as } n \to \infty$$

then

$$k_n^{1/2}\{\widetilde{\alpha}_n(c) - \alpha\} \xrightarrow{d} \mathcal{N}(\mu(\alpha, \beta, c), \sigma^2(\alpha, c))$$

Efficiency	c	$\rho(c)$
0.99	4.25	1.029
0.95	2.57	1.118
0.90	1.84	1.226
0.75	0.91	1.592
0.50	0.30	2.635
0.25	0.06	5.795

Table 2: Values of $\rho(c)$ for various relative efficiencies.

where $\sigma^2(\alpha, c)$ is defined by (5) and (6) and

$$\mu(\alpha, \beta, c) = \frac{\phi(c)}{h^2(c) - (2+c)h(c) + 1} \left(\frac{\theta \lambda^{-\beta/\alpha} \alpha \beta}{\alpha + \beta} \right) r^{\beta/\alpha + 1/2}. \tag{8}$$

The asymptotic bias in Theorem 2 can be written as $\mu(\alpha, \beta, c) = \rho(c)\mu_0(\alpha, \beta)$ where $\mu_0(\alpha, \beta)$ is the asymptotic bias of the Hill estimator (under the same conditions) and

$$\rho(c) = \frac{\phi(c)}{h^2(c) - (2+c)h(c) + 1}. \tag{9}$$

It can be shown that $\rho(c)$ is a decreasing function of c with $\rho(c) \to 1$ as $c \to \infty$ and $\rho(c) \to \infty$ as $c \to 0$. Table 2 contains values of $\rho(c)$ for the asymptotic efficiencies considered in Table 1.

To illustrate, we use the Condroz calcium data introduced in section 1. The Hill estimates of α are influenced considerably by two large Y_j values with j small. Figure 3 gives “robust” Hill plots for $c = 0.06$ (25% efficiency), $c = 0.3$ (50% efficiency), and $c = 0.91$ (75% efficiency). Note that there is very little difference between the two higher efficiency estimates and the Hill estimate for larger values of k but that these estimates are larger than the Hill estimate for smaller values of k. The lower efficiency estimate is significantly larger than the Hill estimate except for very small values of k. The fact that there is a greater difference between the estimates for larger k may be an indication of greater bias (that is, $|\mu(\alpha, \beta)|$ large) in the both $\widehat{\alpha}$ and $\widetilde{\alpha}(c)$ for such values of k. At the same time, one might be tempted to interpret a region of values of k for which $\widehat{\alpha}$ and $\widetilde{\alpha}(c)$ are approximately equal as one where $|\mu(\alpha, \beta)|$ is perhaps closer to 0 and the resulting estimates more reliable.

In the next section, we will explore bias reduction.

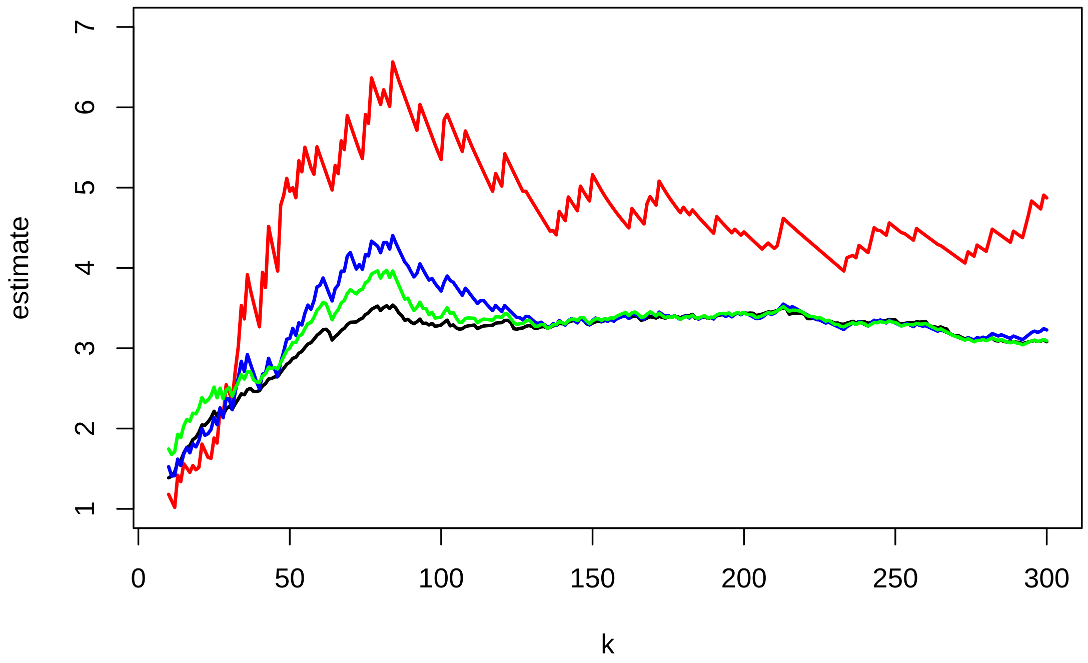

Figure 3: Hill plot (black) with robust alternatives: red – $c = 0.06$ (25% efficiency); blue – $c = 0.3$ (50% efficiency); green – $c = 0.91$ (75% efficiency).

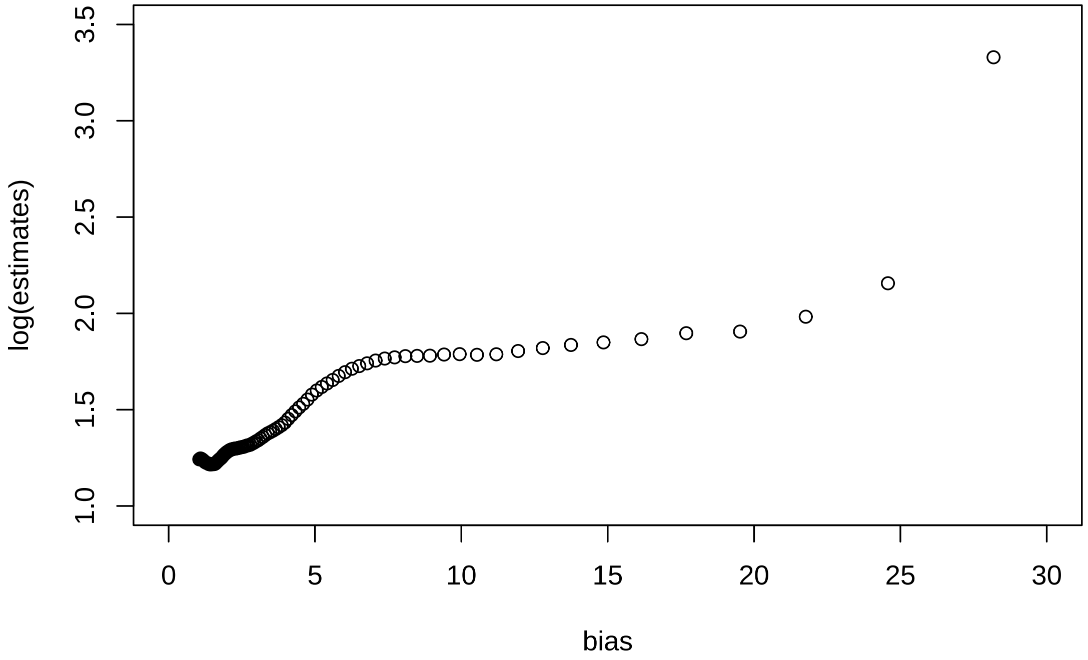

Figure 4: Plot of estimates (logarithmic scale) for the Condroz calcium data (with $k = 100$) for various values of c versus $\rho(c)$.

3 Bias reduction

The fact that the asymptotic bias $\mu(\alpha, \beta, c)$ defined in (8) increases (in absolute terms) as c decreases (under the conditions of Theorem 2) raises the possibility that we can use estimators of α obtained with different values of c to decrease the bias in the Hill estimator. To see this, write

$$\begin{aligned} \widetilde{\alpha}_n(c) &= \alpha + \rho(c)\left\{\frac{\theta\lambda^{-\beta/\alpha}\alpha\beta}{\alpha+\beta}k_n^{-1/2}r^{\beta/\alpha+1/2}\right\} + \nu_n(c) \\ &= \alpha + \kappa(\alpha,\beta,r)\rho(c) + \nu_n(c) \qquad (10) \end{aligned}$$

where $\{\nu_n(c) : c > 0\}$ is approximately a Gaussian process with a covariance function proportional to that given in (7). The parameter $\kappa(\alpha, \beta, r)$ in (10) depends on unknowns α, β, and r (as well as θ) but *not* c; thus if we plot $\widetilde{\alpha}_n(c)$ versus $\rho(c)$ and fit a straight line to the points, the intercept gives an estimate of $\ln(\alpha)$. (An analogous model to (10) will hold if we replace $\widetilde{\alpha}_n(c)$ and α by $g(\widetilde{\alpha}_n(c))$ and $g(\alpha)$, respectively, for some monotone differentiable function g with $g'(\alpha) \neq 0$.) Although such an estimator will have smaller bias (assuming that the approximate Gaussian model is sufficiently good), its variance will increase. For example, suppose we estimate α using generalized least squares using $\widetilde{\alpha}_n(c_1), \cdots, \widetilde{\alpha}_n(c_m)$ where $c_1 < c_2 < \cdots < c_m$ where c_m may equal $+\infty$. Defining $D = D(c_1, \cdots, c_m)$ to be the symmetric matrix with elements

$$D_{ij} = \frac{h^2(c_i) - (c_i+2)h(c_i) - c_i h(c_j) + 1}{\{h^2(c_i) - (c_i+2)h(c_i) + 1\}\{h^2(c_j) - (c_j+2)h(c_j) + 1\}} \quad \text{for } i \leq j,$$

then it is straightforward to verify that if $\bar{\alpha}_n$ is the generalized least squares estimator of the intercept

$$k_n^{1/2}(\bar{\alpha}_n - \alpha) \xrightarrow{d} \mathcal{N}(0, \alpha^2\sigma^2(c_1, \cdots, c_m))$$

where

$$\sigma^2(c_1, \cdots, c_m) = \boldsymbol{e}^T(P^T D^{-1} P)^{-1}\boldsymbol{e} \qquad (11)$$

where $\boldsymbol{e}^T = (1, 0)$ and

$$P = \begin{pmatrix} 1 & \rho(c_1) \\ \vdots & \vdots \\ 1 & \rho(c_m) \end{pmatrix}.$$

As an illustration, consider using equally spaced values of $\phi = \phi(c)$ with the Hill estimator; for various values of r, we take $\phi_i = \phi(c_i) = i/(r+1)$ for $i = 1, \cdots, r$, noting that $\phi = 1 = \phi(\infty)$ corresponds to the Hill estimator. The limiting variances of the generalized least squares estimators are given in Table 3.

r	$\sigma^2(\boldsymbol{c})$
2	1.1385
5	1.0564
10	1.0285
20	1.0143
50	1.0058
100	1.0029

Table 3: Values of the generalized least squares varaince $\sigma^2(c_1, \cdots, c_{r+1})$ defined in (11) for different values of r where $c_i = \phi^{-1}(i/(r+1))$ for $i = 1, \cdots, r$ and $c_{r+1} = \infty$.

Perhaps surprisingly, the price paid for asymptotic unbiasedness is quite small. In fact, it is possible to reduce the asymptotic mean square error of the Hill estimator arbitrarily close to α^2 by combining the Hill estimator $\widehat{\alpha}_n = \widetilde{\alpha}_n(\infty)$ with $\{\widetilde{\alpha}_n(c) : c \in \mathcal{I}\}$ for some set $\mathcal{I}$. In particular, define

$$\begin{aligned}\widetilde{\alpha}_n(\nu) &= \widehat{\alpha}_n + \int_{\mathcal{I}} \{\widetilde{\alpha}_n(c) - \widehat{\alpha}_n\} \, \nu(dc) \\ &= \left(1 - \int_{\mathcal{I}} \nu(dc)\right) \widehat{\alpha}_n + \int_{\mathcal{I}} \widetilde{\alpha}_n(c) \, \nu(dc)\end{aligned}$$

where ν is some signed measure on $\mathcal{I}$. (Alternatively, for some smooth function g, could write

$$g(\widetilde{\alpha}_n(\nu)) = g(\widehat{\alpha}_n) + \int_{\mathcal{I}} \{g(\widetilde{\alpha}_n(c)) - g(\widehat{\alpha}_n)\} \, \nu(dc)$$

and transform back.) It is easy to show that the asymptotic mean square error of such an estimator cannot be less than α^2. However, it is straightforward to find a $\widetilde{\alpha}_n(\nu)$ whose asymptotic mean square error is arbitrarily close to α^2. To do so, take $\mathcal{I} = \{c_0\}$; putting mass w_0 gives asymptotic mean square error

$$\text{AMSE}(w_0) = \alpha^2 \left[1 + \kappa_0^2 - 2w_0\kappa_0^2 \{1 - \rho(c_0))\} + w_0^2 \left\{K_0(c_0, c_0) - 1 + \kappa_0^2(1 - \rho(c_0))^2\right\}\right] \quad (12)$$

where $\kappa_0 = \alpha^{-1}\mu_0(\alpha, \beta)$. If κ_0 is known then the AMSE(w_0) in (12) is minimized at

$$w_0^* = \frac{\kappa_0^2}{\{1 - \rho(c_0)\}} K_0(c_0, c_0) - 1 + \kappa_0^2 \{1 - \rho(c_0)\}^2,$$

which gives

$$\text{AMSE}(w_0^*) = \alpha^2 \left[1 + \kappa_0^2 - \frac{\kappa_0^4 \{1 - \rho(c_0)\}^2}{K_0(c_0, c_0) - 1 + \kappa_0^2 \{1 - \rho(c_0)\}^2}\right].$$

Taking c_0 sufficiently small makes $\text{AMSE}(w_0^*)$ arbitrary close to α^2. Of course, κ_0 is typically unknown; in this case, we can take

$$w_0^\dagger = \{1 - \rho(c_0)\}^{-1}$$

so that the asymptotic bias of $\widehat{\alpha}_n + w_0^\dagger \{\widetilde{\alpha}_n(c_0) - \widehat{\alpha}_n\}$ is 0. Then we have

$$\text{AMSE}(w_0^\dagger) = \alpha^2 \left[1 + \frac{K_0(c_0, c_0) - 1}{\{1 - \rho(c_0)\}^2}\right],$$

which again can be made arbitrarily close to α^2 by taking c_0 sufficiently close to 0.

Of course, the problem with the simple two point estimators considered above is the fact that we are relying on the accuracy of the normal approximation (and the resulting bias and variance approximations) for values of c close to zero. When c is close to zero, the value of $\widetilde{\alpha}_n(c)$ is determined by the smallest values of $\{Y_j\}$; in data that are rounded or discretized in some way, exact zeroes are possible, which greatly complicates the estimation. (For small values of c, the asymptotics for $\widetilde{\alpha}_n(c)$ are better approximated by a functional of a Poisson process.) Given this, it seems more desirable to consider estimators putting positive mass on a wider range of tuning parameter values, thereby given less weight to small values of c. Given no knowledge of κ_0, a sensible approach is to consider asymptotically unbiased estimators; that is, a given $\mathcal{I}$, we assume that the signed measure ν satisfies

$$\int_{\mathcal{I}} \{1 - \rho(c)\}\ \nu(dc) = 1.$$

If $\mathcal{I}$ is a finite set then the asymptotically unbiased estimator of α with the minimum asymptotic variance is simply the generalized least squares estimator defined above; this estimator will put most of its weight on values of c close to 0, which, as suggested above, is less than desirable. Another possibility is to find a measure ν satisfying the asymptotic unbiasedness property whose "distance" from the zero measure (which corresponds to the Hill estimator) is minimum; we can also add an upper bound on the asymptotic variance as a constraint. An example of such a distance or discrepancy function for a measure ν on a finite set $\mathcal{I}$ is the L_2 discrepancy

$$\sum_{c \in \mathcal{I}} w^2(c)$$

where $w(c) = \nu(\{c\})$.

As an example, we take $\mathcal{I} = \{c_i = \phi^{-1}(i/(r+1)) : i = 1, \cdots, r\}$. Setting $\phi_i = \phi(c_i) = i/(r+1)$, we consider estimating $\{w(\phi_i)\}$ to minimize

$$\sum_{i=1}^{r} w^2(\phi_i) \tag{13}$$

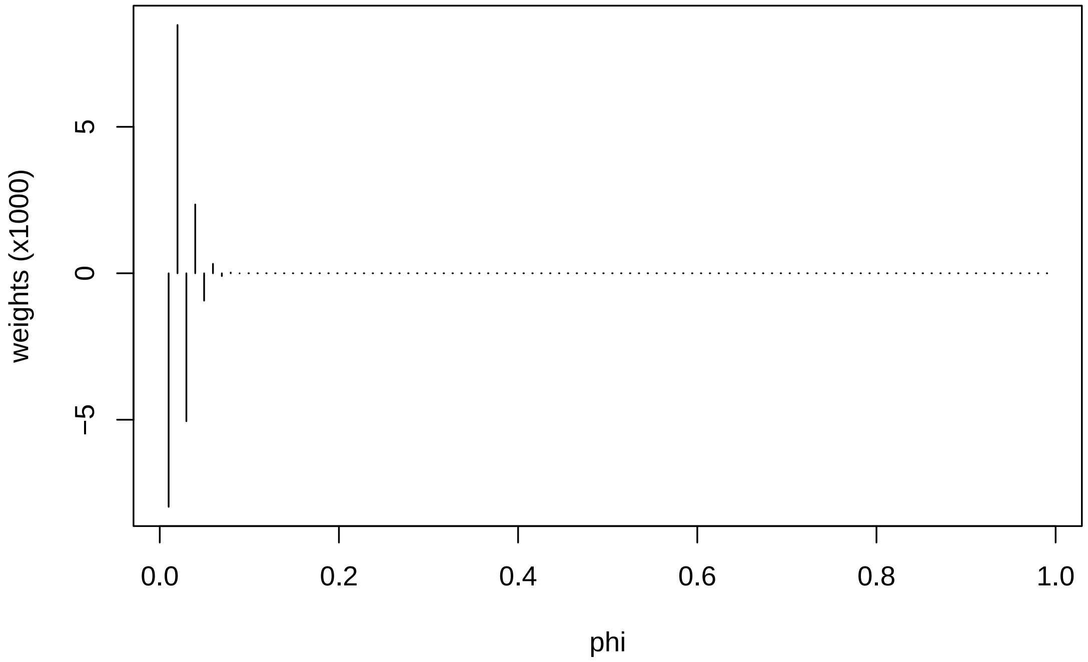

Figure 5: Weights for the generalized least squares estimator using $c_i = \phi^{-1}(i/101)$ for $i = 1, \cdots, 100$; the asymptotic variance of the resulting estimator is $1.0029\alpha^2$.

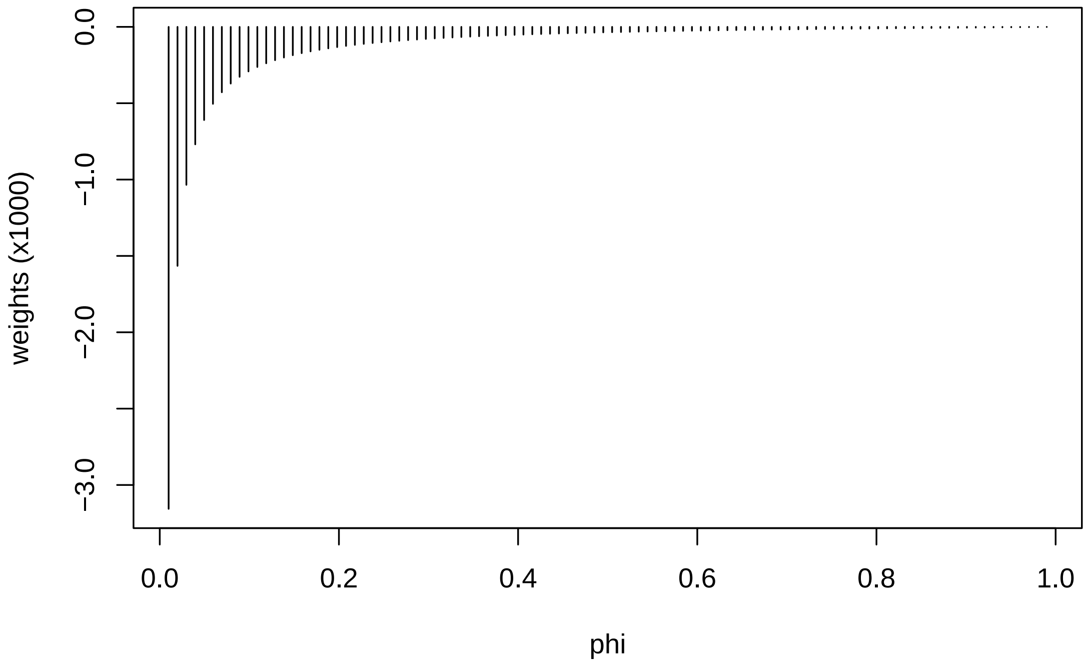

Figure 6: Weights for the asymptotically unbiased estimator minimizing (13) using $c_i = \phi^{-1}(i/101)$ for $i = 1, \cdots, 100$; the asymptotic variance of the resulting estimator is $1.0049\alpha^2$.

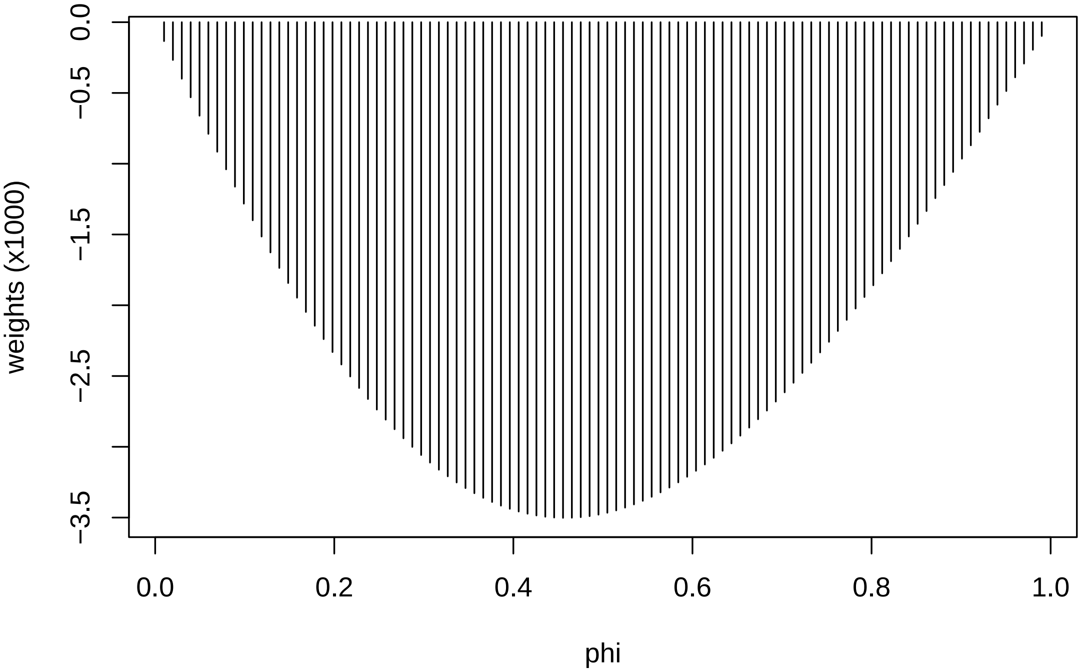

Figure 7: Weights for the asymptotically unbiased estimator minimizing (14) using $c_i = \phi^{-1}(i/101)$ for $i = 1, \cdots, 100$; the asymptotic variance of the resulting estimator is $1.0811\alpha^2$.

and

$$\sum_{i=1}^{r} \{w(\phi_{i-1}) - 2w(\phi_i) + w(\phi_{i+1})\}^2 \tag{14}$$

(where $w(0) = w(1) = 0$) subject to the asymptotic unbiasedness constraint

$$\sum_{i=1}^{r} w(\phi_i)\left\{1 - \rho(h^{-1}(\phi_i))\right\} = 1 \tag{15}$$

To illustrate, we will set $r = 100$. Figure 5 gives the weight function for the generalized least squares estimator (which will have the minimum asymptotic variance) while Figures 6 and 7 give the weight functions for the discrepancy functions (13) and (14), respectively.

It is also possible to define a weighted estimator that minimizes a discrepancy function such as (13) or (14) subject to asymptotic unbiasedness and an upper bound on the asymptotic variance. Writing $\boldsymbol{w} = (w(\phi_1), \cdots, w(\phi_r))^T$, both the discrepancies (13) and (14) can be written in the general form

$$\boldsymbol{w}^T \Upsilon \boldsymbol{w} \tag{16}$$

for some matrix Υ and the minimizer of (16) subject to asymptotic unbiasedness (15) is

$$\boldsymbol{w}_0 = \frac{\Upsilon^{-1}\boldsymbol{\gamma}}{\boldsymbol{\gamma}^T \Upsilon^{-1}\boldsymbol{\gamma}}$$

where $\boldsymbol{\gamma} = (1 - \rho(\phi^{-1}(\phi_1)), \cdots, 1 - \rho(\phi^{-1}(\phi_r)))^T$. Defining the symmetric matrix C by

$$C_{ij} = \frac{h^2(c_i) - (c_i + 2)h(c_i) - c_i h(c_j) + 1}{\{h^2(c_i) - (c_i + 2)h(c_i) + 1\}\{h^2(c_j) - (c_j + 2)h(c_j) + 1\}} - 1 \quad \text{for } 1 \le i \le j \le r,$$

we can define for $\lambda \ge 0$

$$\boldsymbol{w}_0(\lambda) = \frac{(\Upsilon + \lambda C)^{-1}\boldsymbol{\gamma}}{\boldsymbol{\gamma}^T(\Upsilon + \lambda C)^{-1}\boldsymbol{\gamma}}, \tag{17}$$

which will minimize (16) subject to asymptotic unbiasedness and a upper bound (dependent on λ) of the asymptotic variance; as $\lambda \to \infty$, $\boldsymbol{w}_0(\lambda)$ in (17) will converge to the generalized least squares estimator. A plot of these estimates (using both untransformed estimates and log-transformed estimates) as a function of λ computed for the Condroz data (with $r = 100$ and $k = 100$) using the second difference discrepancy function (14) is given in Figure 8. The Hill estimator in this case is 3.289; the bias adjustment (in this case downwards) for the untransformed estimates is greater than that for the log-transformed estimates (although these differences are not much greater than one standard error).

The biased-reduced estimates can also be used to construct alternative Hill plots. To illustrate this, we consider daily returns of Intel from 15 November, 1999 to 11 November, 2008, focusing on estimating the tail index of the lower tail (that is, negative returns). For

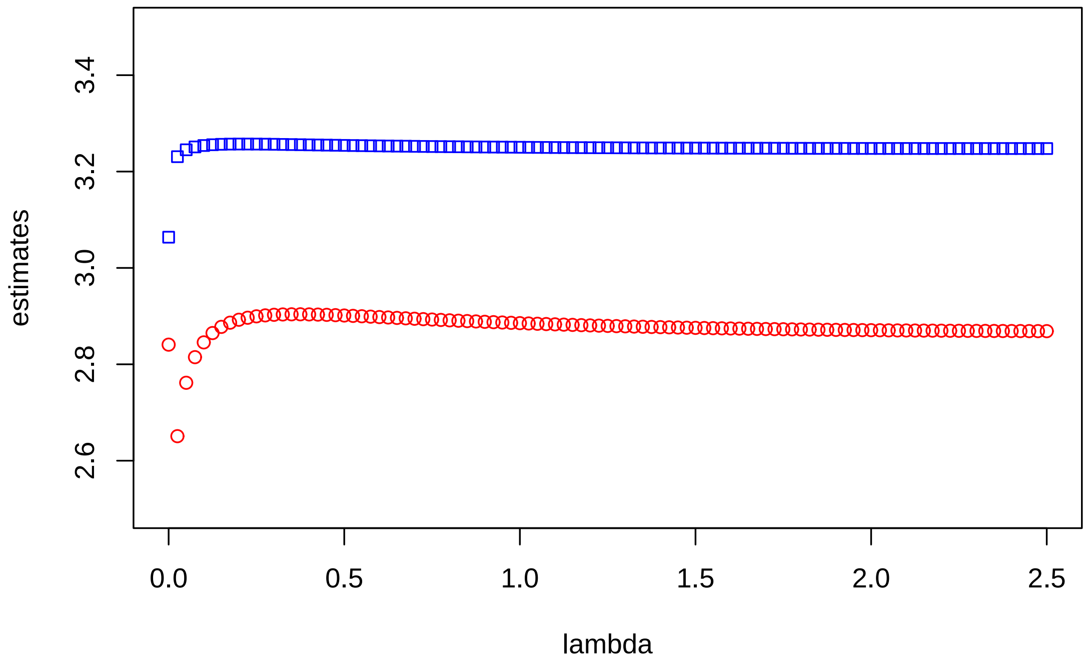

Figure 8: Estimates of α as a function of λ; the red circles indicate the estimates computed using the untransformed $\{\widehat{\alpha}_n(c)\}$ while the blue squares indicate estimates computed using the logarithms.

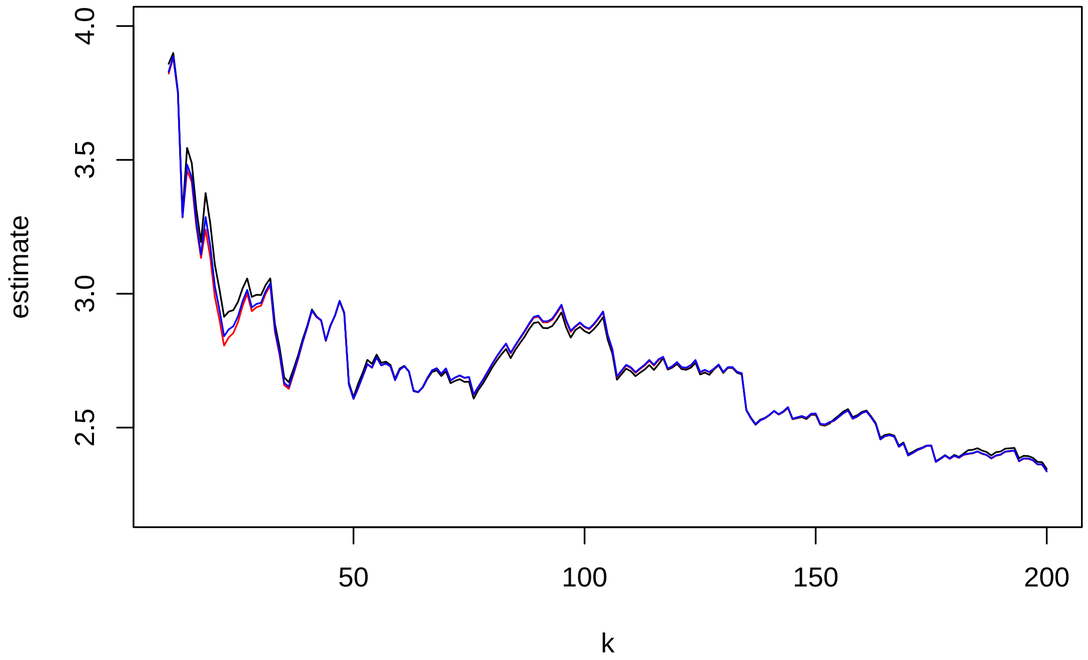

Figure 9: Hill plot for Intel data with asymptotically unbiased estimates minimizing (13) with $r = 20$; the black line are Hill estimates, the red line uses untransformed estimates, and the blue line log-transformed estimates.

values of k from 10 to 200, we compute both the Hill estimator and asymptotically unbiased estimators with weights minimizing (13) and (14) with $r = 20$ using both untransformed estimates and log-transformations. These are given in Figures 9 and 10. In both cases (particularly when the weights minimize (13)), the differences in the three graphs are minimal, perhaps indicating that the bias in the Hill estimates for these data is small; in particular, an absence of significant bias would imply that we could use a larger value of k_n in the computation of $\widehat{\alpha}_n$.

4 Exponential regression

Exponential regression models have been used by a number of authors to adjust for the presence of the slowly varying component in (1) and hence hopefully reduce the bias of the Hill estimator. In particular, Beirlant *et al.* (1999) as well as Feuerverger and Hall (1999)

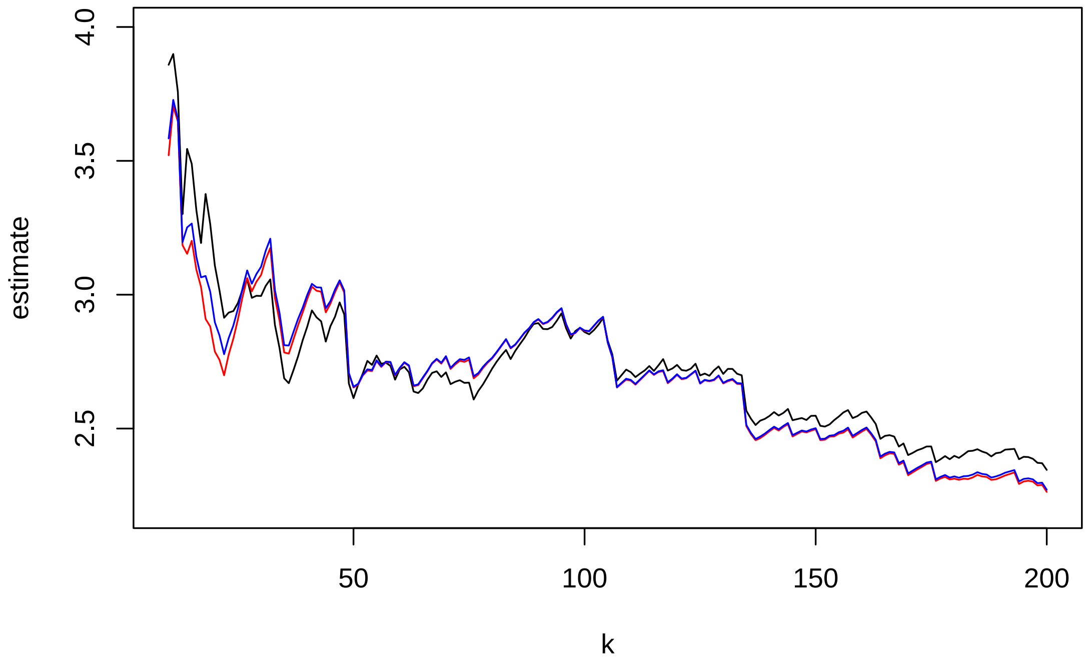

Figure 10: Hill plot for Intel data with asymptotically unbiased estimates minimizing (14) with $r = 20$; the black line are Hill estimates, the red line uses untransformed estimates, and the blue line log-transformed estimates.

assume a parametric form for $L(x)$, similar to that assumed in Theorem 2.

We assume that $Y_1, \cdots, Y_{k_n}$ are approximately independent exponential random variables with means $\mu_1, \cdots, \mu_{k_n}$ where

$$\mu_j^{-1} = \tau(j/n)$$

where τ is a smooth function with the tail index $\alpha = \tau(0)$. If τ is linear in some parameter $\boldsymbol{\beta}$ then this is a generalized linear model with inverse link (McCullagh and Nelder, 1989) and the maximum likelihood estimates of α and $\boldsymbol{\beta}$ can be computed quite easily. Here, we will estimate the function τ non-parametrically using an adaptation of the local scoring algorithm as described in Hastie and Tibshirani (1990).

The algorithm for estimating τ non-parametrically is most easily motivated from the algorithm for a linear parametric model. Suppose that $\mu_j^{-1} = \boldsymbol{\tau}(j/n)^T\boldsymbol{\beta}$ where $\boldsymbol{\tau}(t) = (1, \tau_1(t), \cdots, \tau_p(t))^T$ for functions $\tau_1, \cdots, \tau_p$. Then we can define an estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ as the solution of

$$\sum_{j=1}^{k_n} \bar{\psi}_c(Y_j; \boldsymbol{\tau}(j/n)^T\widehat{\boldsymbol{\beta}})\boldsymbol{\tau}(j/n) = \mathbf{0} \tag{18}$$

where $\bar{\psi}_c(Y_j;\alpha) = \alpha\psi_c(Y_j;\alpha)$ as in (4). We can compute $\widehat{\boldsymbol{\beta}}$ satisfying (9) using a Newton-Raphson algorithm, which can be expressed as an iteratively reweighted least squares (IRLS) algorithm; to be more precise, if $\widehat{\boldsymbol{\beta}}_{(\ell)}$ is the estimate at the ℓ-th step then

$$\widehat{\boldsymbol{\beta}}_{(\ell+1)} = \left(X^T W(\widehat{\boldsymbol{\beta}}_{(\ell)})X\right)^{-1} X^T W(\widehat{\boldsymbol{\beta}}_{(\ell)})\boldsymbol{\xi}(\widehat{\boldsymbol{\beta}}_{(\ell)})$$

where X is a matrix with rows $\boldsymbol{\tau}(j/n)$ for $j = 1, \cdots, k_n$, $W(\widehat{\boldsymbol{\beta}}_{(\ell)})$ is a diagonal matrix with elements $\bar{\psi}'_c(Y_j; \boldsymbol{\tau}(j/n)^T\widehat{\boldsymbol{\beta}}_{(\ell)})$ for $j = 1, \cdots, k_n$, and the so-called adjusted dependent variable $\boldsymbol{\xi}(\widehat{\boldsymbol{\beta}}_{(\ell)})$ is a vector with elements

$$\boldsymbol{\tau}(j/n)^T\widehat{\boldsymbol{\beta}}_{(\ell)} - \frac{\bar{\psi}_c(Y_j; \boldsymbol{\tau}(j/n)^T\widehat{\boldsymbol{\beta}}_{(\ell)})}{\bar{\psi}'_c(Y_j; \boldsymbol{\tau}(j/n)^T\widehat{\boldsymbol{\beta}}_{(\ell)})} \quad \text{for } j = 1, \cdots, k_n.$$

Note that $\bar{\psi}'_c(Y_j; \boldsymbol{\tau}(j/n)^T\widehat{\boldsymbol{\beta}}_{(\ell)})$ can equal 0, which is problematic in the IRLS algorithm; however, this is easily resolved by simply replacing the zeroes by some small positive number δ and defining $\widehat{\boldsymbol{\beta}}_{(\ell)}$ as the limit as δ tends to 0. (As $\bar{\psi}'_c(Y_j; \boldsymbol{\tau}(j/n)^T\widehat{\boldsymbol{\beta}}_{(\ell)})$ tends to 0 then the j-th element of $W(\widehat{\boldsymbol{\beta}}_{(\ell)})\boldsymbol{\xi}(\widehat{\boldsymbol{\beta}}_{(\ell)})$ tends to $-\bar{\psi}_c(Y_j; \boldsymbol{\tau}(j/n)^T\widehat{\boldsymbol{\beta}}_{(\ell)})$.) The IRLS algorithm can be extended to given non-parametric estimates of τ by iteratively smoothing the adjusted dependent variable (with $\widehat{\tau}_{(\ell)}(j/n)$ replacing $\boldsymbol{\tau}(j/n)^T\widehat{\boldsymbol{\beta}}_{(\ell)}$) using weights $\bar{\psi}'_c(Y_j; \widehat{\tau}_{(\ell)}(j/n))$; a simple initial estimate of τ is simply $\widehat{\tau}_{(0)}(j/n) = \widetilde{\alpha}$.

Estimates of τ for the Condroz data are given in Figure 11 using $c = 0.3$, $c = 0.91$, and $c = \infty$. These are computed using spline estimation with 3 effective degrees of freedom;

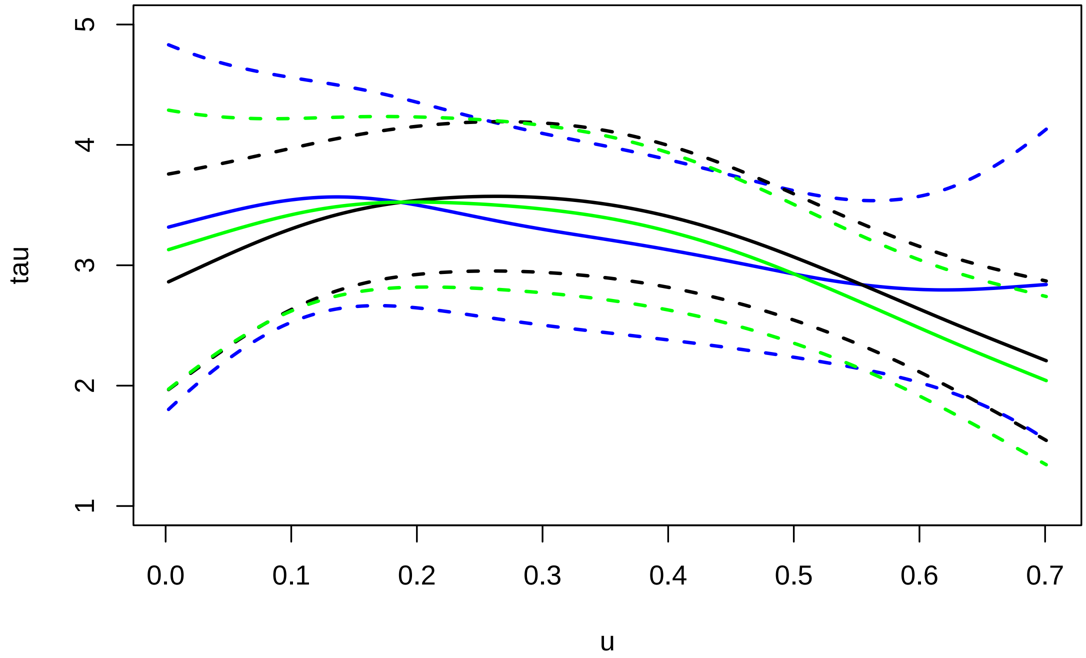

Figure 11: Estimates of τ with pointwise (approximate) 95% confidence intervals (dashed lines) for the Condroz data: blue – $c = 0.3$; green – $c = 0.91$; black – $c = \infty$.

approximate 95% pointwise confidence bands (Hastie and Tibshirani, 1990) for spline estimates are indicated with dashed lines. The tail index α is estimated by $\widehat{\tau}(0)$. These three estimates are in the range from 2.9 to 3.1, which agrees with the bias reduced estimates of the previous section.

5 Discussion

In this paper, we have presented a family of estimators for the tail index parameter α with a view towards robustness and bias reduction. A number of bias reduction methods have been proposed elsewhere. For example, Huisman *et al.* (2001) propose a generalized least squares procedure for reducing bias using a representation of the bias of the Hill estimator as a function of the number of order statistics k_n; in particular, under the assumptions of Theorem 2, the asymptotic bias of the Hill estimator (given by (8) with $c = \infty$) is (approximately) a linear function of $k_n^{\beta/\alpha}$ and so (given an estimator of β/α), a generalized least squares estimator can be used to produce an asymptotically unbiased estimator of α. The advantage of the procedures proposed in section 3 is that we do not need to estimate any auxiliary parameters (allow it is sometimes reasonable to assume that $\beta/\alpha = 1$). Other approaches involve estimating second order parameters, for example, by fitted a parametric or semi-parametric exponential regression model as in Beirlant *et al.* (1999) and Feuerverger and Hall (1999). The methods introduced in section 3 for reducing bias are attractive as they do not involve direct estimation of any nuisance parameters in the slowly varying function $L(x)$. However, these latter methods seem to be quite dependent on the form of $L(x)$ assumed in Theorem 2 and would be less appropriate for a more general form of $L(x)$.

Appendix: Proofs of Theorems 1 and 2

If $\{Y_j : j = 1, \cdots, k_n\}$ were exactly i.i.d. exponential random variables with mean $1/\alpha$ then the results would hold trivially if $k_n \to \infty$ (setting $\theta = 0$ in Theorem 2). In general, we can use the representation of Beirlant *et al.* (2002) to approximate $\{Y_j : j = 1, \cdots, k_n\}$ uniformly by independent exponential random variables if $k_n \to \infty$ sufficiently slowly. In particular, under the conditions of Theorem 2, we have

$$Y_j = \frac{1}{\alpha}\left\{1 - \frac{\beta\theta}{\alpha}\lambda^{-\beta/\alpha}\left(\frac{j}{n}\right)^{\beta/\alpha}\right\} E_j + R_j$$

where $\{E_j\}$ are i.i.d. exponential random variables and

$$R_j = R_j^{(1)} + R_j^{(2)}$$

with

$$\begin{aligned}\sum_{j=1}^{k_n} \frac{R_j^{(1)}}{j} &= o_p\left((k_n/n)^{\beta/\alpha}\ln(k_n)\right) \\ \sup_{j \le k_n} |R_j^{(2)}| &= o_p\left((k_n/n)^{\beta/\alpha}\right).\end{aligned}$$

Thus, if $k_n = O(n^{2\beta/(2\beta+\alpha)})$, we have

$$\begin{aligned}Z_n(u) &= k_n^{1/2}\sum_{j=1}^{k_n}\bar{\psi}(Y_j; \alpha + k_n^{-1/2}u) \\ &= k_n^{1/2}\sum_{j=1}^{k_n}\bar{\psi}(\mu_{nj}E_j; \alpha + k_n^{-1/2}u) + o_p(1)\end{aligned}$$

uniformly over u in compact sets where

$$\mu_{nj} = \frac{1}{\alpha}\left\{1 - \frac{\beta\theta}{\alpha}\lambda^{-\beta/\alpha}\left(\frac{j}{n}\right)^{\beta/\alpha}\right\}.$$

Applying (for example) the Lyapunov central limit theorem, a direct calculation gives

$$Z_n(u) \xrightarrow{d} -\phi(c)r^{\beta/\alpha+1/2}\frac{\beta\theta}{\alpha+\beta}\lambda^{-\beta/\alpha} - W + \frac{1}{\alpha}\left\{h^2(c) - (2+c)h(c) + 1\right\}u$$

where $W \sim \mathcal{N}(0, h^2(c) - 2(c+1)h(c) + 1)$. The proof of Theorem 1 follows likewise noting that taking $k_n = o(n^{2\beta/(2\beta+\alpha)})$ implies that

$$Z_n(u) \xrightarrow{d} -W + \frac{1}{\alpha}\left\{h^2(c) - (2+c)h(c) + 1\right\}u.$$

References

Ahmed, E.S., Volodin, A.I. and Hussein, A.A. (2005) Robust weighted likelihood estimation of exponential parameters. *IEEE Transactions on Reliability.* **54**, 389-395.

Beirlant, J., Dierckx, G., Goegebeur, Y. and Matthys, G. (1999) Tail index estimation and an exponential regression model. *Extremes.* **2**, 177-200.

Beirlant, J., Dierckx, G., Guillou, A. and Starica, C. (2002) On exponential representation of log-spacings of extreme order statistics. *Extremes.* **5**, 157-180.

Beilant, J., Goegebeur, Y., Teugels, J., Segers, J., De Waal, D. and Ferro, C. (2004) *Statistics of Extremes: Theory and Applications.* New York: Wiley.

Brooks, C., Clare, A.D., Dalle Molle, J.W. and Persand, G. (2005) A comparison of extreme value theory approaches for determining value at risk. *Journal of Empirical Finance.* **12**, 339-352.

Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J. and Knuth, D.E. (1996) On the Lambert W function. *Advances in Computational Mathematics.* **5**, 329-359.

Drees, H., de Haan, L. and Resnick, S. (2000) How to make a Hill plot. *Annals of Statistics.* **28**, 254-274.

Dupuis, D. and Morgenthaler, S. (2002) Robust weighted likelihood estimators with an application to bivariate extreme value problems. *Canadian Journal of Statistics.* **30**, 17-36.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extreme Events for Insurance and Finance.* Berlin: Springer.

Feuerverger, A. and Hall, P. (1999) Estimating a tail exponent by modelling departure from a Pareto distribution. *Annals of Statistics.* **27**, 760-781.

Gather, U. and Schultze, V. (1999) Robust estimation of scale of an exponential distribution. *Statistica Neerlandica.* **53**, 327-341.

Hall, P. (1982) On simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society, Series B.* **44**, 37-42.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models.* London: Chapman and Hall.

Huisman, R., Koedijk, K.G., Kool, C.J.M. and Palm, F. (2001) Tail index estimates in small samples. *Journal of Business and Economic Statistics.* **19**, 208-216.

Hill, B.M. (1975) A simple general approach to inference about the tail of a distribution. *Annals of Statistics.* **13**, 331-341.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models (2nd edition).* London: Chapman and Hall.

Schluter, C. and Trede, M. (2008) Identifying multiple outliers in heavy-tailed distributions with an application to market crashes. *Journal of Empirical Finance.* **15**, 700-713.

Vandewalle, B., Beirlant, J., Christmann, A. and Hubert, M. (2007) A robust estimator for the tail index of Pareto-type distributions. (unpublished manuscript)

Vandewalle, B., Beirlant, J. and Hubert, M. (2004) A robust estimator of the tail index based on an exponential regression model. In *Theory and Applications of Recent Robust Methods,* editors M. Hubert, G. Pison, A. Struyf, and S. van Aelst. 367-376. Basel: Birkhauser.