A Multivariate Hill Estimator

Yves Dominicy, Pauliina Ilmonen & David Veredas

ECARES, Solvay Brussels School of Economics and Management, Université libre de Bruxelles

yves.dominicy@ulb.ac.be

Motivation

In practise, we often assume that the observations follow a Gaussian distribution. But in reality this assumption is often not the case. For example, in finance it is known that the tails of the returns are thicker than those from a Gaussian distribution. And hence we need other distributions like for example heavy-tailed distributions (univariate) and/or elliptical distributions (multivariate).

Some examples of domains where heavy-tailed or elliptical distributions can be applied: economy and finance (risk management, portfolio selection), physics and astrology (temperature, hydrodynamics, gravitational field of the stars), geology and environment (earthquakes, intensity of the perturbations, pollution), internet (World Wide Web traffic), ...

The parameters of those distributions are estimated via parametric methods, like MLE, GMM, ... But in the examples cited above often the interest lies on one feature, namely the *tail index*.

Introduction

Univariate Hill

As our focus lies on the tail index only, the most popular estimator for this purpose in the univariate setting is the Hill estimator proposed in the seminal paper of Bruce Hill in 1975. The Hill estimator is a semi-parametric estimate of the extreme value index τ and hence of the tail index α for heavy-tailed distributions.

We consider a sequence $X_1, \ldots, X_n$ of i.i.d. random variables from a common distribution with distribution function F and survival or tail function given by $\bar{F} = 1 - F$. We denote by $X_{n,n} \leq \ldots \leq X_{1,n}$ the corresponding order statistics. We say that X is a heavy-tailed distribution if

$$P(X > x) = \bar{F}(x) = 1 - F(x) = x^{-\alpha} L(x) \quad \text{for} \quad x > 0,$$

for some $0 < \alpha < \infty$ and some slowly varying function L at infinity. Another way to express the heavy-tailed distribution is to say that $\bar{F}$ is regularly varying with tail index $-\alpha$, $\alpha > 0$:

$$\lim_{t \to \infty} \frac{\bar{F}(tx)}{\bar{F}(t)} = x^{-\alpha}, \quad x > 0.$$

The Hill estimator can be defined as being the $k = k(n)$ exceedancees over a given order statistic $X_{k+1,n}$:

$$\hat{\tau}_{k,n} = \frac{1}{k} \sum_{j=1}^{k} \ln X_{j,n} - \ln X_{k+1,n}, \quad \text{for} \quad 1 \leq k \leq n-1.$$

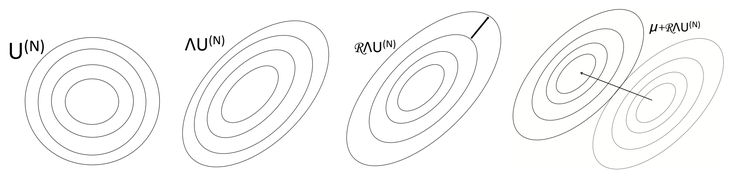
Assuming that k the number of upper order statistics is a sequence of positive integers satisfying $1 \leq k \leq n-1$, $k = k(n) \to \infty$ and $k/n \to 0$ as $n \to \infty$. Under these conditions, Hall (1982) proved the asymptotic normality of the Hill estimator, given by

$$\sqrt{k(n)}(\hat{\tau}_{k,n} - \tau) \to^d N(0, \tau^2),$$

provided a more stringent condition namely a second order refinement of regular variation.

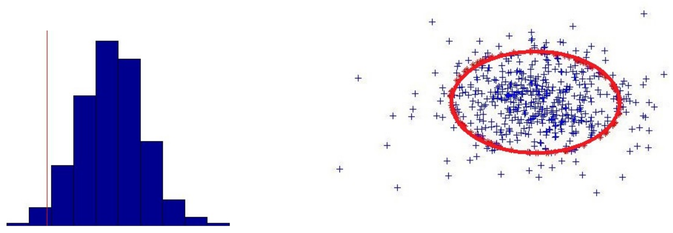
Elliptical distribution

A N–dimensional random vector $\mathbf{X}$ is elliptically distributed if $\mathbf{X} =_d \boldsymbol{\mu} + \mathcal{R}\boldsymbol{\Lambda}\mathbf{U}^{(N)}$.



Multivariate Hill

Intuition



Methodology

Let $\mathbf{X} = \{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$ denote a sample from a continuous p-variate elliptical distribution with parameter set $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)$. Assume that $\boldsymbol{\Sigma}$ is full rank. Let $e_{\boldsymbol{\mu},\boldsymbol{\Sigma},k,n}$ denote the smallest probability ellipsoid (defined using $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$) containing $n-k$ data points (ν the probability mass), and let $O_{\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Sigma},k,n} \subset \mathbf{X}$ denote the corresponding set of the outer points (i.e. set of the points that do not lie on or in $e_{\boldsymbol{\mu},\boldsymbol{\Sigma},k,n}$). For each data point $\mathbf{X_i}$ we define a corresponding point $\mathbf{X_i}^* \in \mathbb{R}^p$ such that $\mathbf{X_i} - \boldsymbol{\mu} = c(\mathbf{X_i}^* - \boldsymbol{\mu})$, where $c \in \mathbb{R}_+$, and $\mathbf{X_i}^*$ lies on the border of the separating ellipsoid $e_{\boldsymbol{\mu},\boldsymbol{\Sigma},k,n}$. Let

$$D_p(\mathbf{X_i}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k, n) = \frac{L_1(\mathbf{X_i}, \boldsymbol{\mu})}{L_1(\mathbf{X_i}^*, \boldsymbol{\mu})},$$

where $L_1(a, b)$ denotes the L_1 distance between the points a and b.

Definition 1. *The multivariate Hill is defined as*

$$H_p(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k, n) = \frac{1}{k} \sum_{\mathbf{X_i} \in O_{\mathbf{X},\boldsymbol{\mu},\boldsymbol{\Sigma},k,n}} \ln(D_p(\mathbf{X_i}, \mu, \Sigma, k, n)).$$

Now, as long as the univariate marginal distributions are regularly varying, the inverse of $H_p(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k, n)$ can be used to estimate the tail parameter α.

Lemma 1. *Let $\mathbf{X} = \{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$ denote a sample from a continuous p-variate elliptical distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)$, and let $Z = \{Z_1, \ldots, Z_n\}$ denote a sample from continuous univariate elliptical distribution with parameters $(0, 1, \alpha)$. Assume that $\boldsymbol{\Sigma}$ is full rank. We have*

$$D_p(\mathbf{X_i}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k, n) \sim D_1(Z_i, 0, 1, k, n).$$

Theorem 1. *Let $\mathbf{X} = \{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$ denote a sample from a continuous p-variate elliptical distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)$. Assume that $\boldsymbol{\Sigma}$ is full rank, $0 < \alpha < \infty$, and assume second order regular variation of the corresponding generating variate $\mathcal{R}_\alpha$. Assume that the additional requirements for the sequence $k(n)$ depending on the second order condition are fulfilled. Let $\tau = \alpha^{-1}$. Now, for $k \to \infty$ and $\frac{k}{n} \to 0$ as $n \to \infty$, we have that*

$$\sqrt{k}((H_p(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k, n)) - \tau) \xrightarrow{d} N(0, \tau^2).$$

MCD

In practise, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have to be estimated. We use the Minimum Covariance Determinant (MCD) method introduced by Rousseeuw (1985). MCD method is based on considering all subsets containing $n-k$ sample points of the original observations, and estimating the covariance matrix, and the mean vector, on the data of the subset associated with the smallest covariance matrix determinant. Moreover, the estimates $\hat{\boldsymbol{\Sigma}}_{MCD}$ and $\hat{\boldsymbol{\mu}}_{MCD}$ are highly robust against outlying observations. Thus, we replace $D_p(\mathbf{X_i}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k, n)$ with $D_p(\mathbf{X_i}, \hat{\boldsymbol{\mu}}_{MCD}, \hat{\boldsymbol{\Sigma}}_{MCD}, k, n)$, and $H_p(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k, n)$ with $H_p(\mathbf{X}, \hat{\boldsymbol{\mu}}_{MCD}, \hat{\boldsymbol{\Sigma}}_{MCD}, k, n)$.

Monte Carlo Simulations

We carry out a Monte Carlo experiment for the elliptical Student-t distribution. We consider dimensions 1, 2, 10 and 50. For each dimension we simulate 1000 draws of 500, 1000 and 5000 observations. As tail indexes of the Student-t distribution we consider $\alpha = 2$, 3, 5 and 10. We set the location vector to the null vector and the scatter matrix to the identity matrix. For the threshold ν in the MCD approach we take 0.95, 0.975, 0.99 and 0.999.

The main findings of our simulation study are:

- The more observations the better.
- The higher the threshold the better the estimates.
- The thicker the tails the better.
- An increase in the dimension p is a blessing and a curse at the same time.

Empirical Application

We illustrate our estimator with 13 years of daily stock log returns of 21 major world-wide equity market indexes that represent three geographical regions, namely America (S&P500, NASDAQ, TSX, Merval, Bovespa, and IPC), Europe (AEX, ATX, FTSE, DAX, CAC40, SMI, and MIB), and East Asia and Oceania (HgSg, Nikkei, StrTim, SSEC, BSE, KLSE, KOSPI, and AllOrd). The sample spans from January 4, 2000 to April 26, 2013, giving rise to 3474 observations for each series. We adjust each return series with an AR(2)-GARCH(1,1) model such that the remaining tail dependencies are not conditional.

The tail dependence coefficient between, say, X_i and X_j describes the limiting probability that X_i exceeds a certain threshold given that X_j has already exceeded that threshold:

$$\lambda_{ij} = \frac{\int_0^{f(\rho_{ij})} \frac{s^\alpha}{\sqrt{s^2-1}} ds}{\int_0^1 \frac{s^\alpha}{\sqrt{s^2-1}} ds},$$

where $f(\rho_{ij}) = \sqrt{(1+\rho_{ij})/2}$.

We do a rolling window analysis, by using sub-samples of five years and rolling them one year.

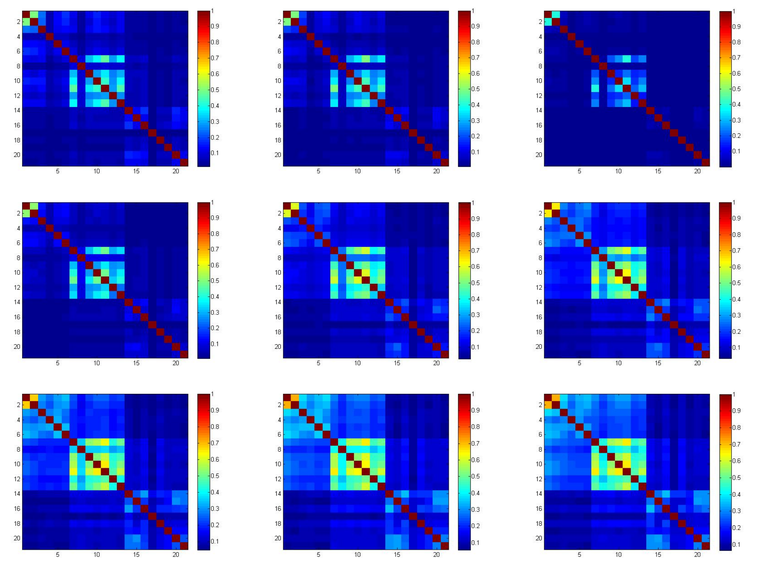


Figure 1: Starting in 2000-2005 and ending in 2008-2013

Window	$H_{21}(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, k, n)$
2000-2005	7.325
2001-2006	9.266
2002-2007	16.21
2003-2008	10.20
2004-2009	6.918
2005-2010	6.353
2006-2011	5.996
2007-2012	5.480
2008-2013	5.083

Table 1: Multivariate Hill

Conclusion

We propose a multivariate elliptical extension of the well-known semi-parametric Hill estimator. We show that it possesses the same characteristics as the univariate one and we prove the asymptotic theory. For univariate random variables, our estimator boils down to the popular Hill estimator. The presented estimator is easy to implement and the simulation results show that the estimator behaves well.

The paper is available at www.ssrn.com