Estimating extreme bivariate quantile regions

John H.J. Einmahl

Tilburg University

Joint work with Laurens de Haan

EVA/Graybill VIII, Fort Collins,

June 26, 2009

INTRODUCTION

Bivariate probability distribution.

Quantile?

When the probability density has some monotonicity property, a natural definition seems possible: for elliptic distributions we choose an ellipse on which the density is constant as the quantile curve.

When the density lives on $(0, \infty)^2$ and is monotone in both variables separately, a similar quantile region can be defined.

Let the pair (X, Y) have df F with density f on $(0, \infty)^2$. Denote the corresponding probability measure with P. Suppose that f is decreasing in each variable.

We define quantile regions determined by the levels of f:

$$Q = \{(x, y) \in (0, \infty)^2 : f(x, y) \leq \varepsilon\}.$$

So, for a (very small) $p \in (0, 1)$ we try to find a Q of this form such that $PQ = p$.

The region

$$Q^c = \{(x, y) \in (0, \infty)^2 : f(x, y) > \varepsilon\}$$

has the property that everywhere on Q^c, f is larger than everywhere on Q, i.e. the quantile region Q is the set of *less likely points*. As a consequence, Q^c is the *smallest* region such that $PQ^c = 1 - p$.

In this setup, we shall consider estimation of quantiles in the far tail.

Suppose we simultaneously monitor two possibly dependent risks $X, Y > 0$ and we have a random sample of size n, i.e. n i.i.d. copies of (X, Y).

Let $p = p_n$ be very small. In particular when we want to protect ourselves against a calamity that has not yet occurred, we consider the case where $p < 1/n$. For the asymptotics we consider $np_n \to c \in [0, \infty)$, so $c = 0$ is possible.

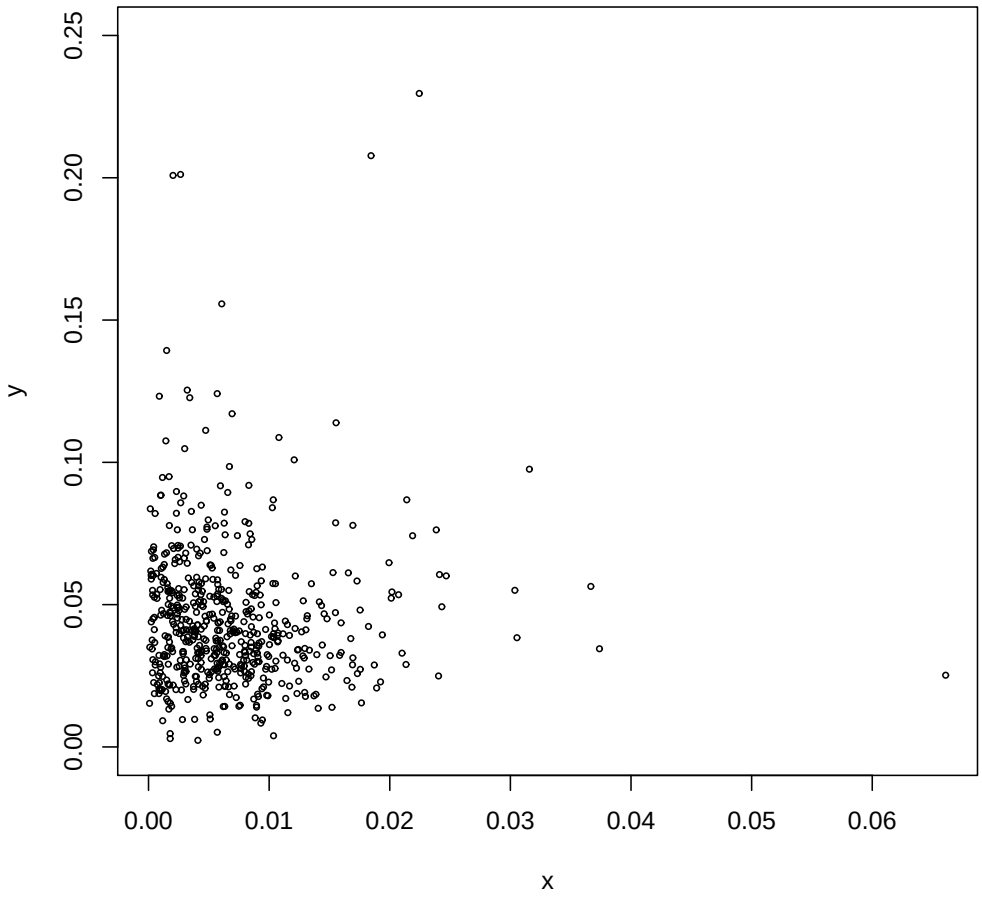
How to estimate $Q = Q_n$ (if we do not assume a parametric model)?

We will propose an estimator based on multivatiate extreme value theory.

Our results are useful in, e.g., aviation safety.

The Federal Aviation Administration (FAA) needs a system that provides instant assessments of airline performances and that in particular signals those that appear to be extreme. The task now is to identify such an extreme risk region. See Einmahl, Li and Liu (2009).

Available is a data set for two key airline performance measures (Incident and Operational). The bivariate data are positive and higher values correspond to a worse performance.



Our estimator of Q is a very natural extreme risk region and hence could be used for flagging events of extreme aviation risk.

RESULTS

We assume that $F \in D(G)$, with extreme-value indices $\gamma_1, \gamma_2 > 0$, i.e. when $t \to \infty$

$$(1) \qquad t\,(1 - F(U_1(t)x^{\gamma_1}, U_2(t)y^{\gamma_2})) \to \iint_{u>x \text{ or } v>y} g(u,v)dudv,$$

on $(0,\infty]^2 \backslash \{(\infty,\infty)\}$, with

$$U_j(t) = F_j^{-1}(1 - 1/t)$$

and $F_j,\ j = 1, 2$, the marginals of F.

Here g is the density corresponding to $-\log G_0$; G_0 is obtained from G after standardization to standard Fréchet marginals: $G_0(x,y) = G(x^{\gamma_1}, y^{\gamma_2})$.

We also need a density version of (1): when $t \to \infty$,

$$(2) \qquad tU_1(t)U_2(t)f(U_1(t)x^{\gamma_1}, U_2(t)y^{\gamma_2}) \to \frac{1}{\gamma_1\gamma_2}x^{1-\gamma_1}y^{1-\gamma_2}g(x,y),$$

on $(0,\infty)^2$.

It follows that $g(ax,ay) = a^{-3}g(x,y),\ a > 0$.

We assume that f is decreasing in each coordinate, outside $(0,M]^2$ (for some $M > 0$) and on $(0,M]^2$, f is bounded away from zero.

Recall

$$Q_n = \{(x, y) \in (0, \infty)^2 : f(x, y) \leq \varepsilon\}$$

with ε such that $PQ_n = p$. Note that ε is only defined implicitly.

Set

$$S = \{(x, y) : x^{1-\gamma_1} y^{1-\gamma_2} g(x, y) \leq \gamma_1 \gamma_2\},$$

see (2). S is a fixed (not depending on n) basis for our estimator of Q_n. We will estimate it later and then transform that estimator, using in particular p.

Throughout let k be a sequence of positive numbers such that $k \to \infty$ and $k/n \to 0$.

A first step is to replace ε by

$$\widetilde{\varepsilon} = \left(\frac{np}{k\nu(S)}\right)^{\gamma_1+\gamma_2+1} \frac{1}{(n/k)U_1(n/k)U_2(n/k)},$$

where ν is the exponent measure, the measure corresponding to $-\log G_0$.

Let $z = (x, y)$ and define, in vector notation, the map T_n by

$$T_n(z) = U(n/k)z^{\gamma}.$$

Write

$$\widetilde{Q}_n = T_n\left(\frac{k\nu(S)}{np}S\right) = U\left(\frac{n}{k}\right)\left(\frac{k\nu(S)}{np}\right)^{\gamma} S^{\gamma}.$$

$\widetilde{Q}_n$ is obtained from

$$\{(x, y) \in (0, \infty)^2 : f(x, y) \leq \widetilde{\varepsilon}\}$$

by using the limiting relation between f and g, i.e. the domain of attraction condition for densities (2).

$\widetilde{Q}_n$ is a good approximation to Q_n: we have

$$\frac{P(Q_n \triangle \widetilde{Q}_n)}{p} \to 0.$$

Here $\triangle$ denotes 'symmetric difference': $A \triangle B = A \backslash B \cup B \backslash A$.

The obvious step to obtain an estimator of Q_n is now to estimate $\widetilde{Q}_n$, which can be done by estimating $T_n, \nu(S)$ and in particular S.

We write S in polar coordinates ($r = \sqrt{x^2 + y^2}$, $\theta = \arctan(y/x)$):

$$\left\{ (x, y) : r \geq \left(\frac{1}{\gamma_1 \gamma_2} \psi(\theta) \cos^{1-\gamma_1} \theta \sin^{1-\gamma_2} \theta \right)^{\frac{1}{\gamma_1 + \gamma_2 + 1}} \right\},$$

where

$$\psi(\theta) = g(\cos\theta, \sin\theta).$$

The function ψ is the density of the spectral measure Ψ. Just like the exponent measure, Ψ describes the dependence structure of the limit distribution G.

In order to estimate these quantities, we have to estimate $U_1(n/k), U_2(n/k), \gamma_1, \gamma_2$ and the spectral density ψ.

Estimation of the first four is well-known. We estimate the $U_j(n/k)$ with the corresponding marginal order statistics and the two extreme-value indices with the moment estimator.

The estimator $\widehat{\psi}$ for ψ will be obtained by smoothing the empirical likelihood estimator of the spectral measure Ψ in Einmahl and Segers (2010).

Hence we finally obtain the 'non-trivial' estimator

$$\widehat{Q}_n = \widehat{U}\left(\frac{n}{k}\right)\left(\frac{k\widehat{\nu(S)}}{np}\right)^{\widehat{\gamma}}\widehat{S}^{\widehat{\gamma}}.$$

All these building block estimators have good asymptotic properties. In particular we show that for every $\eta \in (0, \pi/4)$:

$$\sup_{\theta \in [\eta, \frac{\pi}{2} - \eta]} |\widehat{\psi}(\theta) - \psi(\theta)| \xrightarrow{\mathbb{P}} 0.$$

Hence $\widehat{Q}_n$ is close to $\widetilde{Q}_n$ and hence to Q_n:

THEOREM. Under the above and some other conditions, we have that, as $n \to \infty$,

$$\frac{P(\widehat{Q}_n \triangle Q_n)}{p} \xrightarrow{\mathbb{P}} 0.$$

Remarks.

1) The consistency formulation in a ratio setting is appropriate here. Since $p = O(1/n)$, the statement $P(\widehat{Q}_n \triangle Q_n) \xrightarrow{\mathbb{P}} 0$ is pointless: it even holds when taking $\widehat{Q}_n$ the empty set. Our result states that the estimation error is much smaller than the already extremely small p. Observe that it follows from the Theorem that $P(\widehat{Q}_n)/p \xrightarrow{\mathbb{P}} 1$.

2) In practice it is important that the tuning parameters k used in the estimation of the marginal quantities (γ_j and U_j, $j = 1, 2$) and in the estimation of ψ can be chosen to be different, i.e. we take k_1, k_2 and k_ψ. In this case the theorem remains true.

3) Note that the estimated quantile region $\widehat{Q}_n$ depends on p in a monotone way: if $p < p'$ then $\widehat{Q}_n(p) \subset \widehat{Q}_n(p')$. It is also a continuous function of p. Hence, starting from a very small $\widehat{Q}_n$ we can enlarge it until it first hits an observation. This observation can then be considered the largest one and it has a 'p-value' attached to it.

In this way we can introduce a ranking of the larger observations and give them p-values. This could be helpful in deciding whether some two-dimensional observation is the most atypical (or: an outlier).

ILLUSTRATION

(Thanks to Andrea Krajina.)

We apply the method to two simulated data sets of size $n = 5000$.

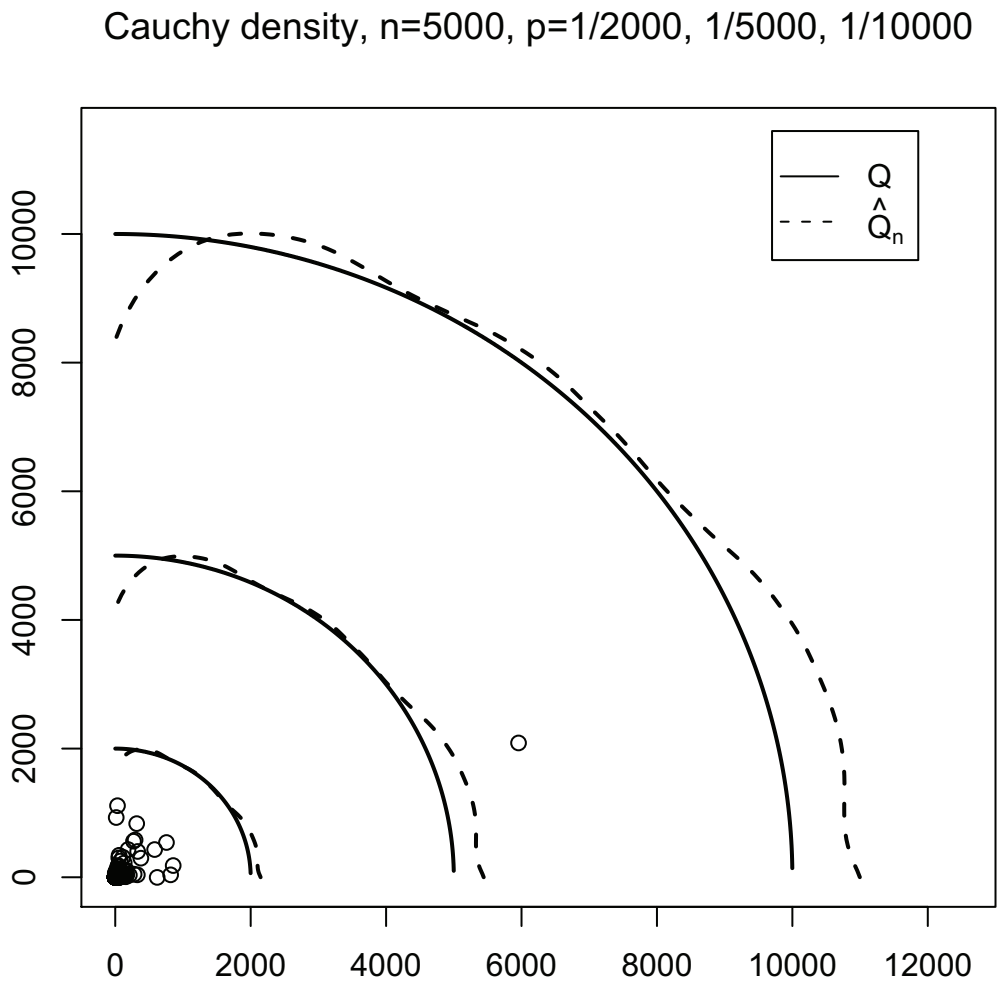
$p = 1/2000, 1/5000$, and $1/10{,}000$. Note that for the latter value we have $np = 0.5$.

Bivariate Cauchy distribution on first quadrant:

$$f(x, y) = \frac{2}{\pi(1 + x^2 + y^2)^{3/2}}, \quad x, y > 0.$$

This is a heavy-tailed density, symmetric in the coordinates x and y and a function of the radius r.

$\gamma_1 = \gamma_2 = 1$ and $\psi(\theta) = 1$, for $\theta \in (0, \pi/2)$



We see excellent behavior of the procedure. Observe that the regions ar far away from almost all the data. We also calculated $P(\widehat{Q}_n)$. These three values only deviate a few percent from p, a very small error.

Density on $(0,\infty)^2$:

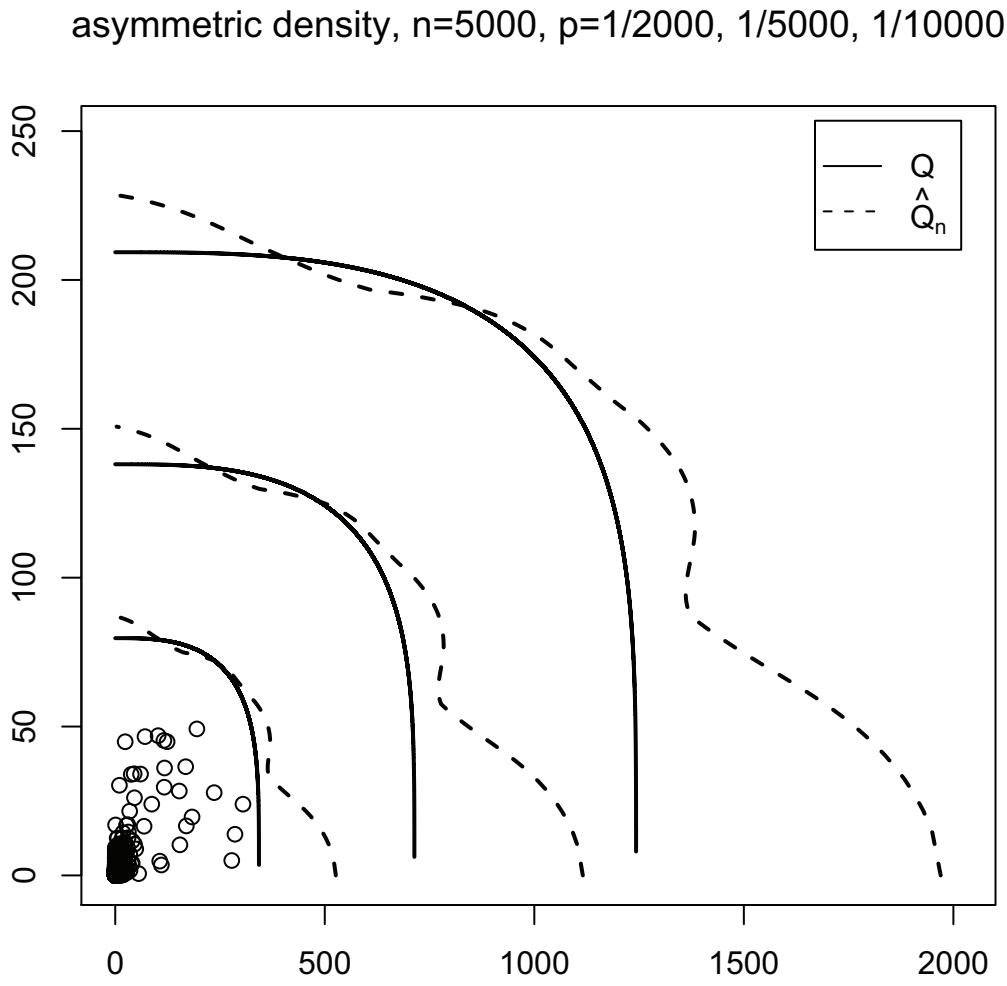
$$f(x,y) = \frac{c}{x^3 + y^4 + 1}, \tag{3}$$

with $c \approx 0.581$.

This density is less heavy tailed: $\gamma_1 = 4/5$ and $\gamma_2 = 3/5$.

$$\psi(\theta) = \frac{12cc_1c_2 \cos^{-1/5}\theta \sin^{-2/5}\theta}{25\left(c_1^3 \cos^{12/5}\theta + c_2^4 \sin^{12/5}\theta\right)},$$

$\theta \in (0, \pi/2)$, with $c_1 \approx 0.589,\ c_2 \approx 0.593$.



For these data the procedure shows the same excellent behavior. The three values of $P(\widehat{Q}_n)$ are now 10–15% too low, a small error given the statistical difficulty of the estimation problem.