Computational tools for multivariate heavy tailed data

John Nolan

American University
Washington, DC

MURI Cornell Meeting
1 October 2013

1. Past
   - Graphical diagnostics for heavy tails
   - Meshes on spheres, simplices and manifolds
   - Spherical Cubature
2. Ongoing
   - Generalized Spherical Distributions (Natick)
   - Ratios of stable variables
   - Simplicial Cubature
3. Future
   - Measure of dependence for multivariate stable laws
   - Computational methods for multivariate stable distributions

Outline

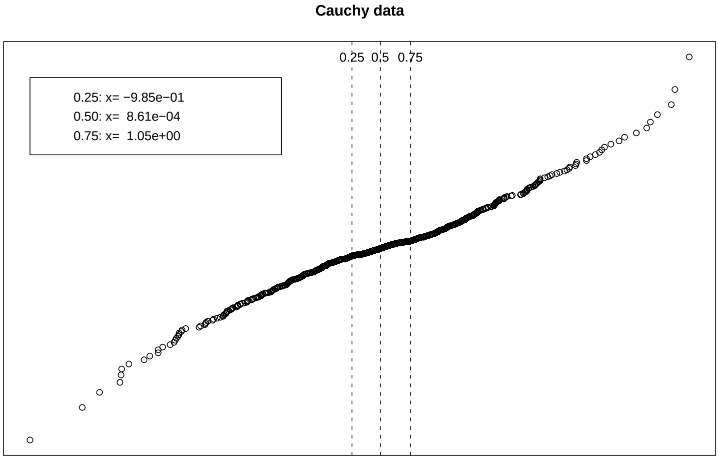
1. Past
   - Graphical diagnostics for heavy tails
   - Meshes on spheres, simplices and manifolds
   - Spherical Cubature

2. Ongoing
   - Generalized Spherical Distributions (Natick)
   - Ratios of stable variables
   - Simplicial Cubature

3. Future
   - Measure of dependence for multivariate stable laws
   - Computational methods for multivariate stable distributions

Transformed plot to show heavy tailed data

Univariate case: log-log empirical CDF plot



Pick three values $0 \le q_1 \le q_2 \le q_3 \le 1$ (default $q_1 = 1/4$, $q_2 = 1/2$, $q_3 = 3/4$). Then define the corresponding data quantiles: $t_i = \widehat{F}^{-1}(q_i)$. Use these values to define the two functions

$$h(x) = h(x|t_1, t_2, t_3) = \begin{cases} -1 - \log(-z) & x < t_1 \\ z & t_1 \le x \le t_3 \\ 1 + \log(z) & x > t_3 \end{cases} \qquad z = 2\frac{x - t_2}{t_3 - t_1}$$

$$g(p) = g(p|q_1, q_2, q_3) = \begin{cases} q_1(1 + \log \frac{p}{q_1}) & p < q_1 \\ p & q_1 \le p \le q_3 \\ q_3 - (1 - q_3) \log \frac{1-p}{1-q_3} & p > q_3 \end{cases}$$

Plot $(h(x_i), g(p_i))$, where x_i are sorted data values and $p_i = (i - 1/2)/n$.

log-log empirical CDF plot (continued)

Transformations are continuous, linear in the middle interval and logarithmically scaled on the outer intervals. So if there is power law decay, get linear behavior on tails.

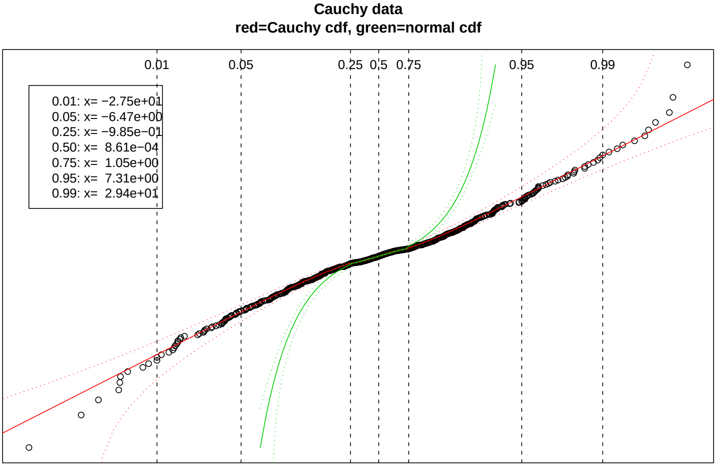
Compresses extremes to give better picture of the whole range data.

Model free: makes no assumption about the data

Can adjust q_i's: for one-sided data on the right use $0 = q_1 = q_2 < q_3$, $q_3 = 1/2$.

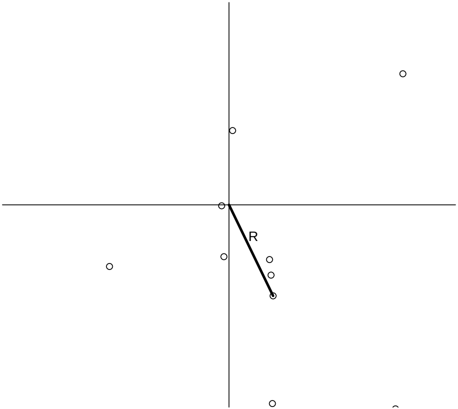
Can compare to one or more models, add more annotations, ...

log-log cdf plot with comparisons

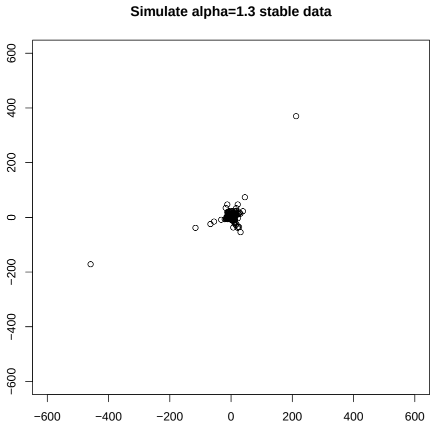


Bivariate diagnostics

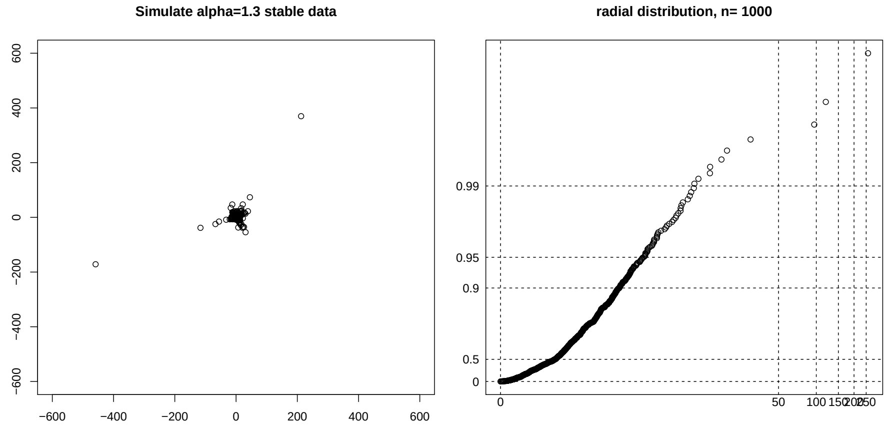
Heavy tailed data has same problems as in univariate case: large values visually dominate. Look at amplitude/radii: $R = \sqrt{X^2 + Y^2}$



Bivariate diagnostics based on amplitude



Bivariate diagnostics based on amplitude

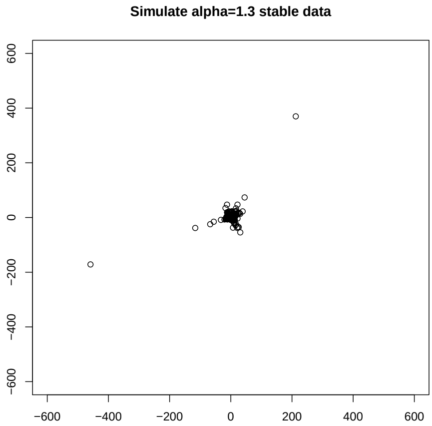


If directional behavior is very different, can select sector/cone to examine.

Bivariate diagnostics scaling x and y by amplitude

Use log transform based for $R > R_0$, preserving direction:

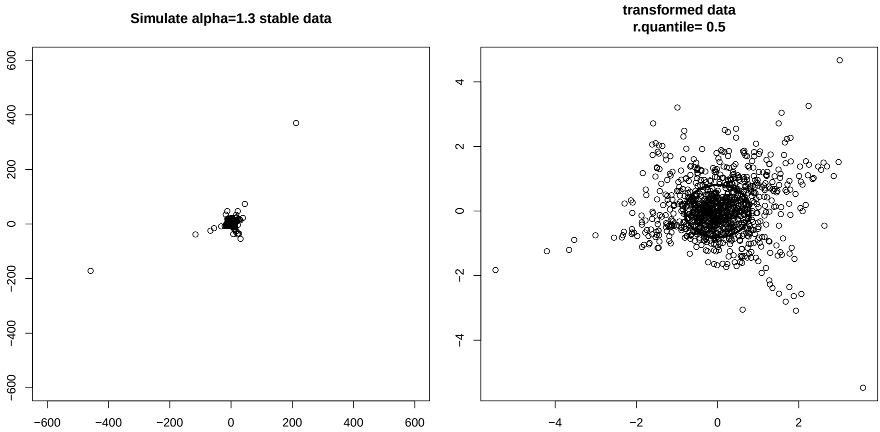
$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} \frac{h(R_i|0,0,R_0)}{R_i}.$$



Bivariate diagnostics scaling x and y by amplitude

Use log transform based for $R > R_0$, preserving direction:

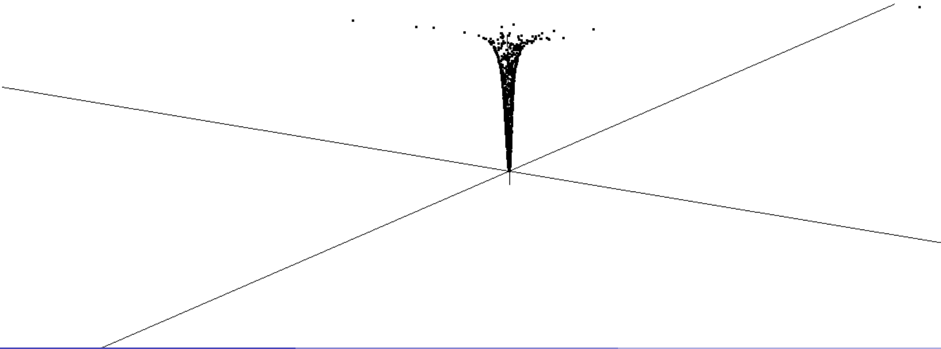
$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} \frac{h(R_i|0,0,R_0)}{R_i}.$$



3-dim visualization in terms of cdf of amplitude

First, take original data and lift based on distance from origin:

Z_i =empirical cdf of amplitude at (X_{1i}, X_{2i})

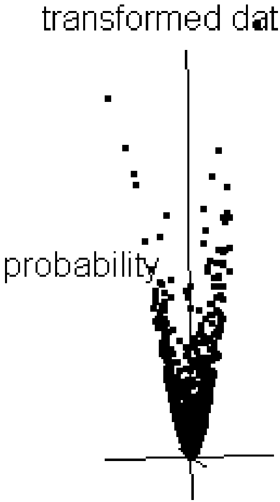


3-dim visualization in terms of cdf of amplitude

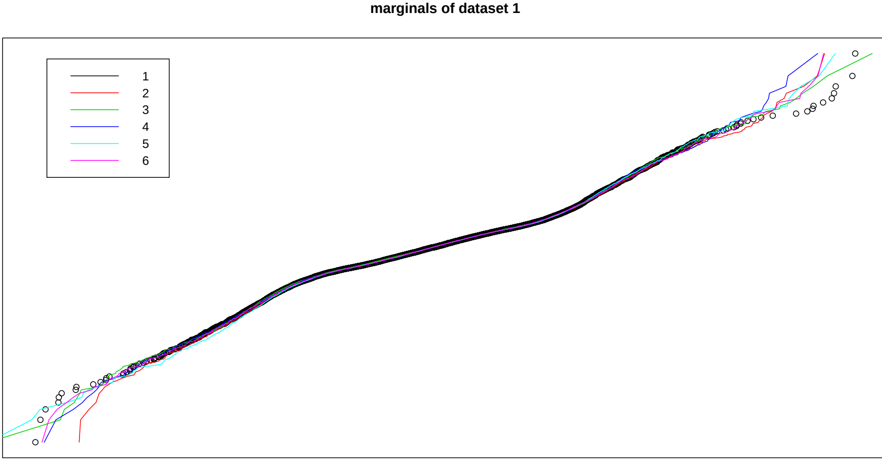
Now combine previous 2 slides: transform in xy plane based on amplitude, transform in z direction based on $g(p)$ from univariate case.

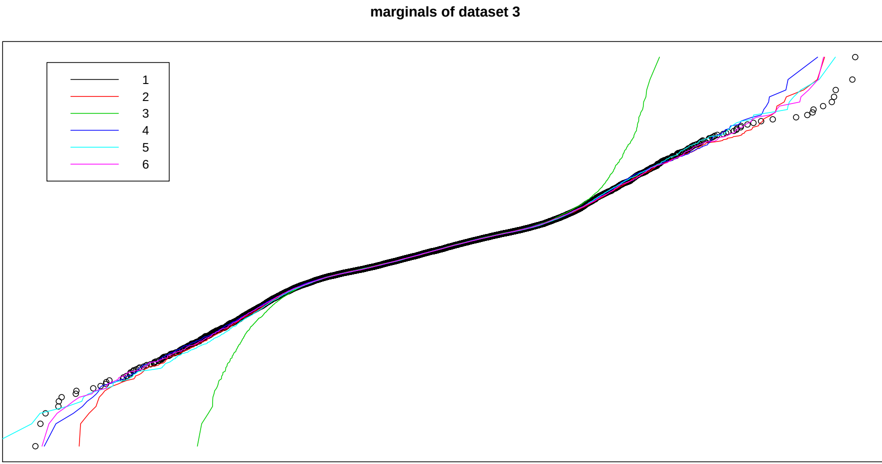
3-dim visualization in terms of cdf of amplitude

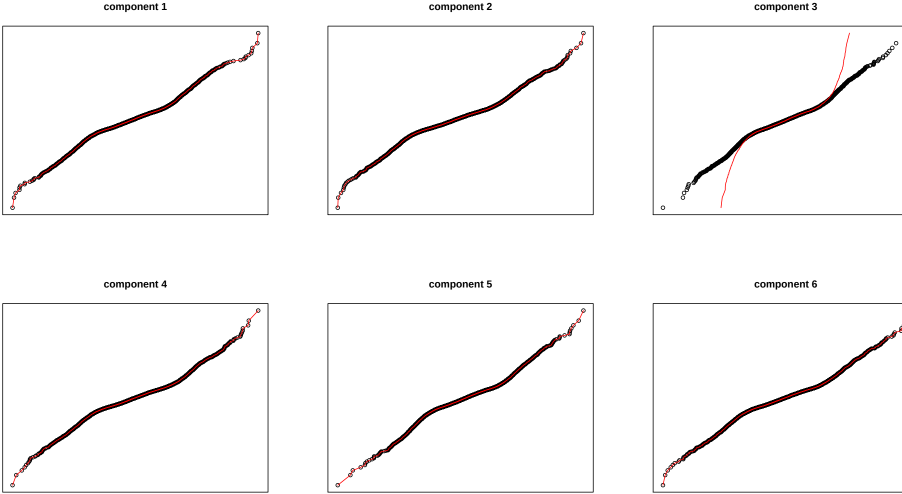
Now combine previous 2 slides: transform in xy plane based on amplitude, transform in z direction based on $g(p)$ from univariate case.



Multivariate diagnostics I - plot all marginals




marginals of dataset 3
1
2
3
4
5
6

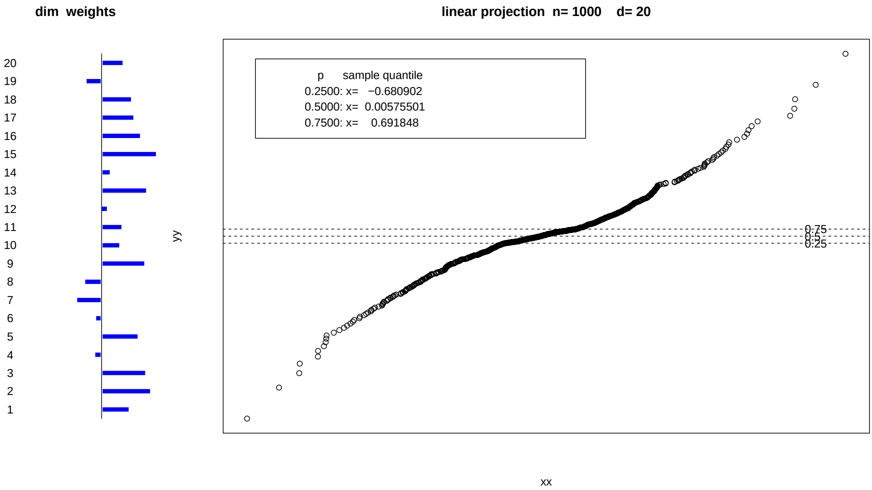

component 1
component 2
component 3
component 4
component 5
component 6

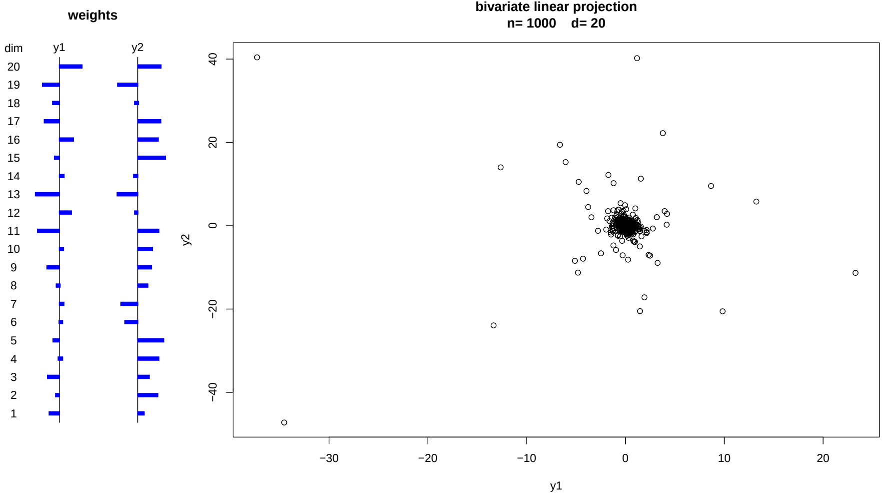
Multivariate diagnostics II

View projections of $\mathbf{X} = (X_1, \ldots, X_d)$:

- can show the amplitude of arbitrary dimensional data
- linear projections $\sum a_j X_j$ or max projections $\bigvee a_j X_j$
- 1, 2 or 3d visualization
- possible interactive way to choose weights and animation to cycle through components (pairs/triples) or "grand tour" through sequence of directions

Examples use 20-dimensional data set (elliptical stable with $\alpha = 1.3$) with $n = 1000$ observations.


dim weights
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
linear projection n= 1000 d= 20
p sample quantile
0.2500: x= −0.680902
0.5000: x= 0.00575501
0.7500: x= 0.691848
0.75
0.5
0.25
yy
xx


weights
dim
y1
y2
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
bivariate linear projection
n= 1000 d= 20
40
20
0
−20
−40
y2
−30
−20
−10
0
10
20
y1

Integration on spheres: multivariate sum stable distributions

Lévy and Feldheim (1930s): **X** sum stable, with index α and centered, then there is a finite measure Λ on the unit sphere $\mathbb{S}$ with

$$E\exp(i\langle \mathbf{u},\mathbf{X}\rangle) = \exp\left(-\int_{\mathbb{S}} \omega_\alpha(\langle \mathbf{u},\mathbf{s}\rangle)\,\Lambda(d\mathbf{s})\right),$$

where

$$\omega_\alpha(u) = \begin{cases} |u|^\alpha \left[1 - i(\operatorname{sign} u)\tan\frac{\pi\alpha}{2}\right] & \alpha \neq 1 \\ |u|\left[1 + i(\operatorname{sign} u)\frac{2}{\pi}\log|u|\right] & \alpha = 1. \end{cases}$$

Here the spread of mass by Λ determines the joint structure.

Integration on simplices: multivariate max stable distributions

de Haan and Resnick (1977): **X** max stable, centered with index ξ, then there is a finite measure Λ on the unit simplex $\mathbb{W}_+$ with

$$P(\mathbf{X} \leq \mathbf{x}) = \exp\left(-\int_{\mathbb{W}_+} \left(\bigvee_{i=1}^{d} \frac{w_i}{x_i^{\xi}} \right) \Lambda(d\mathbf{w}) \right)$$

Again the spread of mass by Λ determines the joint structure.

How to work with spectral measures in higher dimensions?

In both cases, the specification of the dependence structure for max stable and sum stable laws is done in terms of a spectral measure.

Can work with discrete spectral measures, but currently hard to handle much else when dimension is bigger than 2.

Idea (joint work with A.-L. Fougeres and C. Mercadier): use piecewise polynomial spectral densities. Allow you to put mass in different regions. Simplest case is piecewise linear spectral densities.

Piecewise polynomial (PWP) spectral densities

We need tools to:

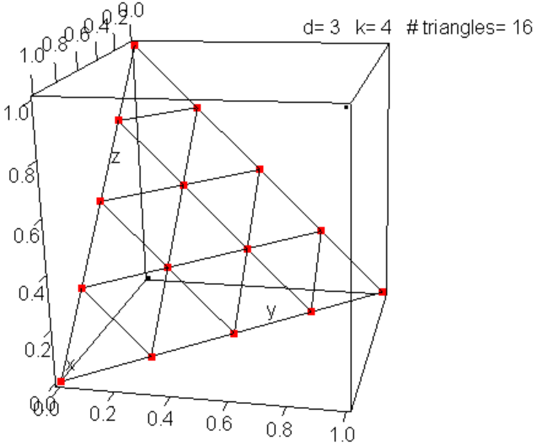
- define meshes simplex and spheres in higher dimensions
- define PWP densities
- methods to evaluate the integrals in the exponent of the max-stable and sum-stable formulas above
- routines to estimate PWP spectral densities from data

Initial work on PWP algorithms

R packages

- mesh.R - define uniform partitions on the simplex, approx. unif. on the sphere
- rlrs.R - R interface to the lrs program of David Avis, CS @ McGill

edgewise 4-subdivision in $\mathbb{R}^3$



Integrating over spheres and balls

Developed an R package SphericalCubature to evaluate integrals of the form:

$$\int_{\mathbb{S}} f(\mathbf{s}) d\mathbf{s} \quad \text{and} \quad \int_{\mathbb{B}} f(\mathbf{x}) d\mathbf{x},$$

where $\mathbb{S}$ is the d-dimensional sphere and $\mathbb{B}$ is the d-dimensional ball.

Publicly available package from the CRAN repository.

- Exact integration of polynomials f in any dimension.
- Numerical quadrature of smooth functions in moderate dimensions.
- 'Directed' numerical quadrature of non-smooth functions in moderate dimensions.

Outline

1. Past
   - Graphical diagnostics for heavy tails
   - Meshes on spheres, simplices and manifolds
   - Spherical Cubature
2. Ongoing
   - Generalized Spherical Distributions (Natick)
   - Ratios of stable variables
   - Simplicial Cubature
3. Future
   - Measure of dependence for multivariate stable laws
   - Computational methods for multivariate stable distributions

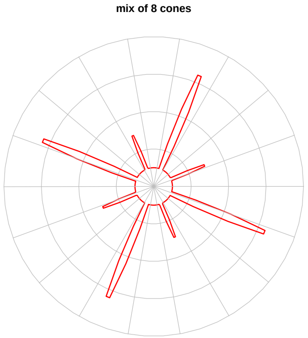
Generalized spherical distributions (Natick)

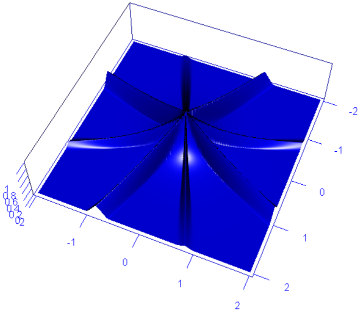
Let $\mathbf{X}$ have a density $f(x)$ on $\mathbb{R}^d$ and let $\mathbb{S}$ be the unit sphere $\{\mathbf{x} : |\mathbf{x}| = 1\}$, $\mathbb{B}$ be the unit ball $\{\mathbf{x} : |\mathbf{x}| \leq 1\}$. A distribution is spherically distributed if $f(\cdot)$ is constant on each sphere $r\mathbb{S}$, $r > 0$.

A distribution is generalized spherical if there is a curve/surface $\mathbb{S}_*$ with $f(\cdot)$ being constant on all multiples $r\mathbb{S}_*$, $r > 0$.

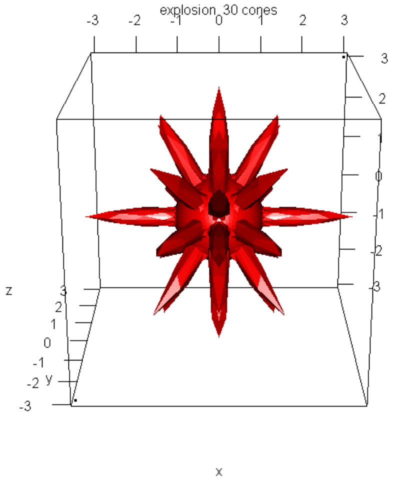
Goal: to have flexible program to work with large classes of such distributions in $d-$dimensions.

Star shaped distributions in 2D





Start shaped contour in 3D

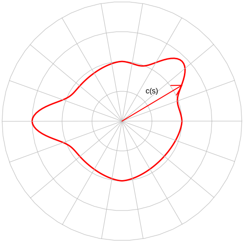


Previous work

Fernandez, Osiewalski and Steel (1995) gave idea, Arnold, Castillo and Sarabia (2008) extended some and advocated modeling data with these.

We will start with a contour/surface given by a polar representation:

$$\mathbb{S}_* = \{c(\mathbf{s})\mathbf{s} : \mathbf{s} \in \mathbb{S}\}$$

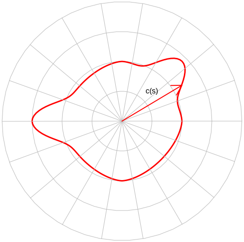


Previous work

Fernandez, Osiewalski and Steel (1995) gave idea, Arnold, Castillo and Sarabia (2008) extended some and advocated modeling data with these.

We will start with a contour/surface given by a polar representation:

$$\mathbb{S}_* = \{c(\mathbf{s})\mathbf{s} : \mathbf{s} \in \mathbb{S}\}$$



IDEA: define a distribution with all level curves a scale of this contour. Need two parts: (a) a flexible way of describing multivariate contours and (b) a (univariate) radial function to specify decay.

Radial function and density

Let $g(r) \geq 0$ be an integrable function on $[0, \infty)$, it will determine the radial decay of the density. Using the contour function $c(\mathbf{s})$ and the radial function, define

$$f(\mathbf{x}) = \begin{cases} g\left(\frac{|\mathbf{x}|}{c(\mathbf{x}/|\mathbf{x}|)}\right) & |\mathbf{x}| > 0 \\ g(0) & |\mathbf{x}| = 0. \end{cases} \tag{1}$$

With suitable integrability conditions (see below), this gives a density on $\mathbb{R}^d$.

Fernandez, Osiewalski and Steel (1995) started with a homogeneous function $v(\mathbf{x})$ on $\mathbb{R}^d$ ($v(r\mathbf{x}) = rv(\mathbf{x})$) and defined $\mathbb{B}_* = \{\mathbf{x} \in \mathbb{R}^d : v(\mathbf{x}) \leq 1\}$. If $\mathbb{B}_*$ is convex and symmetric, then $v(\cdot)$ is a norm on $\mathbb{R}^d$ with unit ball $\mathbb{B}_*$ and unit sphere given by it's boundary $\mathbb{S}_* = \{\mathbf{x} \in \mathbb{R}^d : v(\mathbf{x}) = 1\}$. In general, $v(\cdot)$ is not a norm, but we may still call $\mathbb{S}_*$ a "unit ball". In their approach, the density

$$f(\mathbf{x}) = g(v(\mathbf{x}))$$

is called a v-spherical density. A.K.A. homothetic distributions.

In terms of the contour function, $v(\mathbf{x}) = |\mathbf{x}|/c(\mathbf{x}/|\mathbf{x}|)$. We find using the contour function as the starting point a more intuitive approach: it defines the unit ball, which defines the level curves.

In terms of the contour function, $v(\mathbf{x}) = |\mathbf{x}|/c(\mathbf{x}/|\mathbf{x}|)$. We find using the contour function as the starting point a more intuitive approach: it defines the unit ball, which defines the level curves.

For (1) to be a proper density, it is required that

$$k_*^{-1} := \int_{\mathbb{S}} c^d(\mathbf{s}) d\mathbf{s} \in (0, \infty) \quad (2)$$

and

$$\int_0^\infty r^{d-1} g(r) dr = k_*.$$

We will assume $c(\mathbf{s})$ is continuous and bounded away from 0 on compact $\mathbb{S}$, so the k_* is finite. An easy way to guarantee the second condition is to start with a univariate density $h(r)$ on $[0, \infty)$ and define $g(r) = k_* r^{1-d} h(r)$.

Specifying the contour

We wanted flexible, parametric families of generalized spherical distributions that worked in arbitrary dimensions that included most of the cases described in the earlier work. We allow contour functions of the form

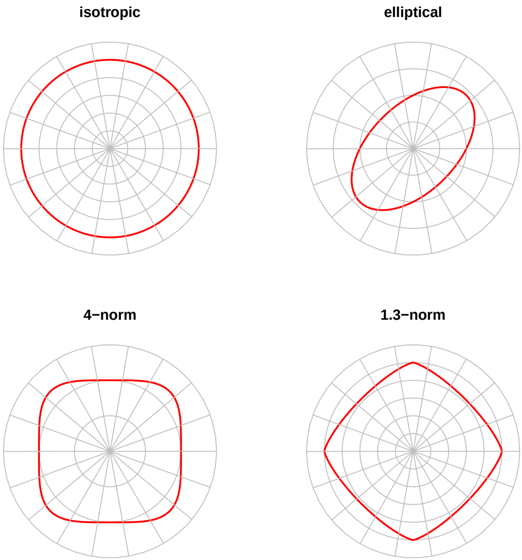
$$c(\mathbf{s}) = \sum_{j=1}^{N_1} a_j c_j(\mathbf{s}) + \frac{1}{\sum_{j=1}^{N_2} a_j^* c_j^*(\mathbf{s})},$$

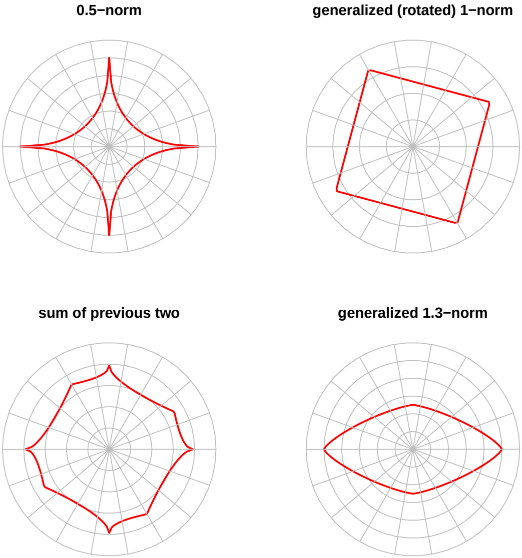
where $c_j > 0$, $c_j^* > 0$, and $c_j(\cdot)$ and/or $c^*(\cdot)$ are one of the cases discussed below.

- $c(\mathbf{s}) = 1$, which makes $\mathbb{S}_*$ the Euclidean ball.
- $c(\mathbf{s}) = c(\mathbf{s}|\boldsymbol{\mu}, \theta)$ is a linear cone with peak 1 at center $\boldsymbol{\mu} \in \mathbb{S}$ and height 0 at the base given by the circle $\{\mathbf{x} \in \mathbb{S} : \boldsymbol{\mu} \cdot \mathbf{x} = \cos\theta\}$. It is assumed that $|\theta| \leq \pi/2$.
- $c(\mathbf{s}) = c(\mathbf{s}|\boldsymbol{\mu}, \sigma) = \exp(-t(\mathbf{s})^2/(2\sigma^2))$ is a Gaussian bump centered at location $\boldsymbol{\mu} \in \mathbb{S}$ and "standard deviation" $\sigma > 0$. Here $t(\mathbf{s})$ is the distance between $\boldsymbol{\mu}$ and the projection of $\mathbf{s} \in \mathbb{S}$ linearly onto the plane tangent to $\mathbb{S}$ at $\boldsymbol{\mu}$.

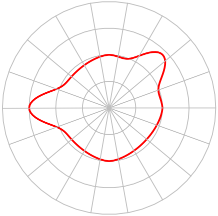
- $c^*(\mathbf{s}) = ||\mathbf{s}||_{\ell^p(\mathbb{R}^d)}$, $p > 0$.
- $c^*(\mathbf{s}) = ||A\mathbf{s}||_{\ell^p(\mathbb{R}^m)}$, $p > 0$, A an $(m \times d)$ matrix. This allows a generalized p-norm. If A is $d \times d$ and orthogonal, then the resulting contour will be a rotation of the standard unit ball in ℓ^p. If A is $d \times d$ and not orthogonal, then the contour will be sheared. If $m > d$, it will give the ℓ^p norm on $\mathbb{R}^m$ of $A\mathbf{s}$.
- $c^*(\mathbf{s}) = (\mathbf{s}^T A\mathbf{s})^{1/2}$, where A is a positive definite $(d \times d)$ matrix. Then the level curves of the distribution are ellipses.

Sums of the last three types allow us to consider contours that are familiar unit balls, or generalized unit balls, or sums of such shapes. Sums of the first three types allow us to describe star shaped contours. Combinations allow more general shapes, see the following plots.

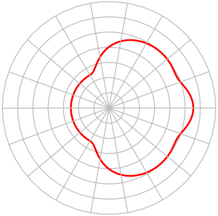

isotropic
elliptical
4−norm
1.3−norm


0.5−norm
generalized (rotated) 1−norm
sum of previous two
generalized 1.3−norm

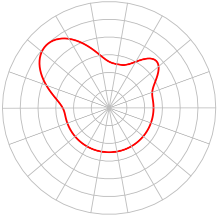
blob #1



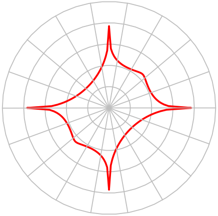
blob #2

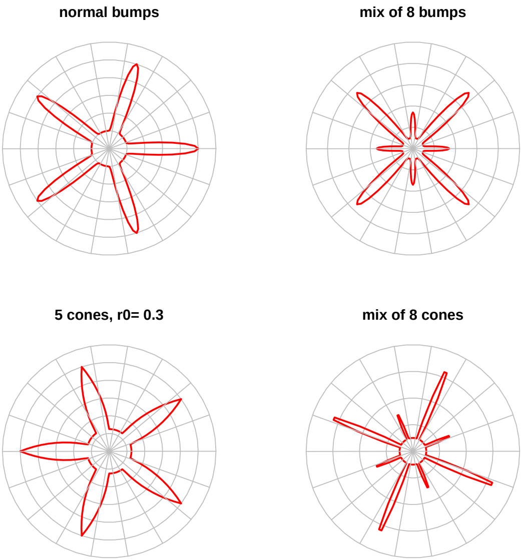


blob #3

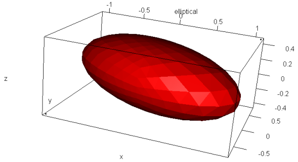


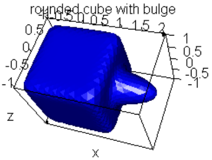
blob #4

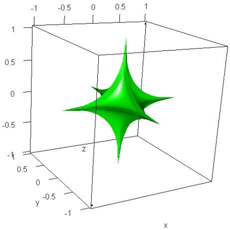


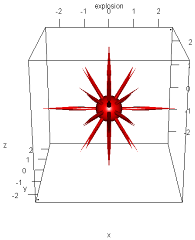

normal bumps
mix of 8 bumps
5 cones, r0= 0.3
mix of 8 cones

Some 3D contours









The radial function

Generally want $g(r)$ to be decreasing so that $f(\mathbf{x})$ is unimodal on $\mathbb{R}^d$. Here are two accessible classes, defined in terms of univariate r.v. $R \geq 0$, which has pdf $h(r)$.

The radial function

Generally want $g(r)$ to be decreasing so that $f(\mathbf{x})$ is unimodal on $\mathbb{R}^d$. Here are two accessible classes, defined in terms of univariate r.v. $R \geq 0$, which has pdf $h(r)$.

- $R\sim$Gamma$(d+a+1)$, then $g(r) = k_* r^{1-d} h(r)$ is a constant times a Gamma(a). If $a \in (0,1]$, then $g(r)$ is decreasing. Is finite at 0 if and only if $a = 1$, always has a light tail.

The radial function

Generally want $g(r)$ to be decreasing so that $f(\mathbf{x})$ is unimodal on $\mathbb{R}^d$. Here are two accessible classes, defined in terms of univariate r.v. $R \geq 0$, which has pdf $h(r)$.

- $R \sim$Gamma$(d + a + 1)$, then $g(r) = k_* r^{1-d} h(r)$ is a constant times a Gamma(a). If $a \in (0, 1]$, then $g(r)$ is decreasing. Is finite at 0 if and only if $a = 1$, always has a light tail.
- R is the amplitude of an isotropic α-stable distribution $(0 < \alpha < 2)$ on $\mathbb{R}^d$: $R = |\mathbf{Z}|$. Using result of Wolfe (1975) on unimodality of isotropic stable laws, it can be shown that $r^{1-d} h(r)$ is decreasing, bounded at the origin, and has a heavy tail: $r^{1-d} h(r) \sim c r^{-(d+\alpha)}$.

The radial function

Generally want $g(r)$ to be decreasing so that $f(\mathbf{x})$ is unimodal on $\mathbb{R}^d$. Here are two accessible classes, defined in terms of univariate r.v. $R \geq 0$, which has pdf $h(r)$.

- $R \sim$Gamma$(d+a+1)$, then $g(r) = k_* r^{1-d} h(r)$ is a constant times a Gamma(a). If $a \in (0,1]$, then $g(r)$ is decreasing. Is finite at 0 if and only if $a = 1$, always has a light tail.
- R is the amplitude of an isotropic α-stable distribution $(0 < \alpha < 2)$ on $\mathbb{R}^d$: $R = |\mathbf{Z}|$. Using result of Wolfe (1975) on unimodality of isotropic stable laws, it can be shown that $r^{1-d}h(r)$ is decreasing, bounded at the origin, and has a heavy tail: $r^{1-d}h(r) \sim cr^{-(d+\alpha)}$.

To do: methods to simulate from such distributions.

Ratios of stable r.v. (Davis, N., & Resnick)

In several applications, the ratio of independent stable r.v. arise:

$$X = Z_1/Z_0, \tag{3}$$

where $Z_i \sim S(\alpha_i, \beta_i, \gamma_i, \delta_i; k)$.

Ratios of stable r.v. (Davis, N., & Resnick)

In several applications, the ratio of independent stable r.v. arise:

$$X = Z_1/Z_0, \tag{3}$$

where $Z_i \sim S(\alpha_i, \beta_i, \gamma_i, \delta_i; k)$.

For example, a centered heavy tailed time series with terms having tail $RV_{-\alpha}$, $1 < \alpha < 2$, the lag 1 ACF

$$\frac{\sum_{i=1}^{n-1} X_i X_{i+1}}{\sum_{i=1}^{n} X_i^2}$$

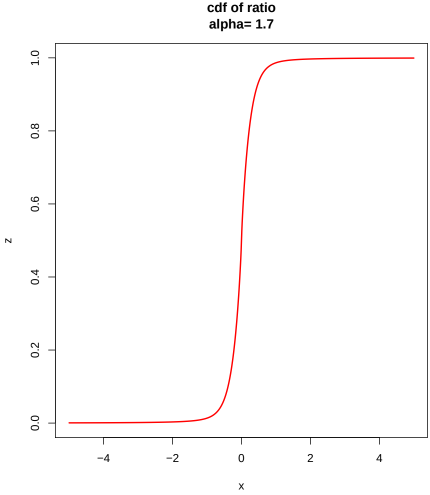
converges when $n \to \infty$ to the ratio (3) with $\alpha_1 = \alpha$, $\alpha_0 = \alpha/2$.

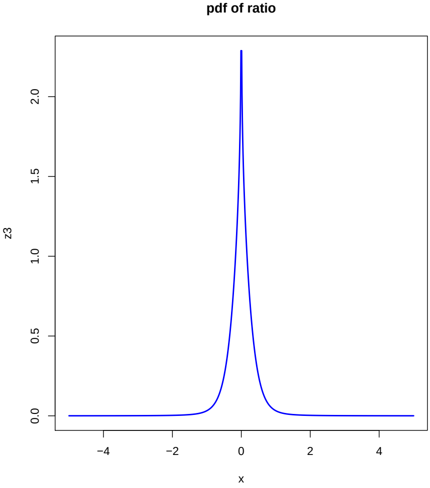
Program to compute density, d.f. and quantiles of ratio

Standard formulas for density and d.f. of a ratio of independent r.v. are combined with numerical calculations of the densities/d.f. of the stable terms to get

- $F(t) = P(Z_1/Z_0 \leq t)$
- $f(t) = F'(t)$.
- quantile function $F^{-1}(u)$.

Distribution function and density of Z_1/Z_0, where $Z_1 \sim \mathbf{S}(\alpha, 0; 1)$ and $Z_0 \sim \mathbf{S}(\alpha/2, 1; 1)$ with $\alpha = 1.70$.





Stable ratios and products - in progress

Tabulate critical values $F^{-1}(1-p)$ for the ratio of symmetric α-stable and independent positive $(\alpha/2)$-stable.

Compare to the proposed method of L. James (private communication).

Integration over simplices

polyintegrate.R - exact integration of polynomials over simplices

More general functions. May be possible to integrate $|\langle \mathbf{s}, \mathbf{u} \rangle|^{\alpha}$ exactly over a simplex?

Outline

Future

Measure of dependence for stable laws - joint work with T. Alparslan, using directional scale function to measure ‘distance from independence’.

Future

Measure of dependence for stable laws - joint work with T. Alparslan, using directional scale function to measure 'distance from independence'.

Computational methods for multivariate stable laws.