Computational geometry for multivariate statistics

John P. Nolan

American University
Washington, DC, USA

CASD 2015
George Mason University
21 October 2015

Outline

ARO Multidisciplinary University Research Initiative (MURI)

5 year grant on Multivariate Heavy Tailed Phenomena

Richard Davis - Columbia University (Statistics)
Weibo Gong - UMass Amherst (ECE)
John Nolan - American University (Math)
Sidney Resnick - Cornell University (ORIE)
Gennady Samorodnitsky - Cornell University (ORIE)
Ness Shroff - Ohio State University
R. Srikant - University of Illinois at Urbana-Champaign (ECE)
Don Towsley - UMass Amherst (CS)
Zhi-Li Zhang - University of Minnesota

Some topics the MURI group has worked on

- Growth of social networks - preferential attachment model leads to joint heavy tailed model for (in-degree, out-degree)
- Random walks on large graphs
- Analysis of communication networks with heavy tailed traffic
- Resource allocation in cloud computing
- Reducing power consumption on mobile devices
- Multivariate extreme value distributions - calculations and estimation
- Threshold selection in heavy tailed inference
- Dimensionality reduction for heavy tailed data - robust PCA and ICA

Outline

There is a need for non-traditional models for multivariate data. Working in dimension $d > 2$ requires new tools.

- grids and meshes on non-rectangular shapes
- numerical integration over surfaces
- simulate from a shape

There is a need for non-traditional models for multivariate data. Working in dimension $d > 2$ requires new tools.

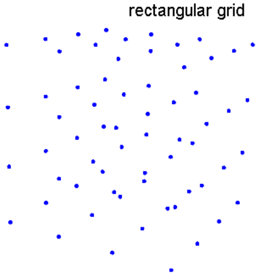
- grids and meshes on non-rectangular shapes
- numerical integration over surfaces
- simulate from a shape

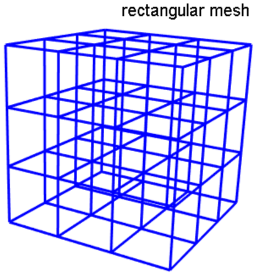
R software packages

- mvmesh - MultiVariate Meshes (CRAN)
- SphericalCubature (CRAN)
- Simplicial Cubature (CRAN)
- gensphere (manuscript submitted)
- mvevd - MultiVariate Extreme Value Distributions (in progress)

mvmesh

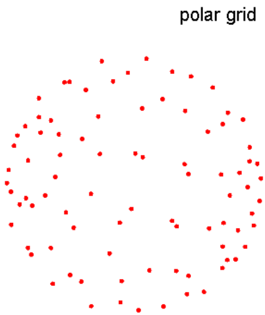
Rectangular grids are straightforward in any dimensions

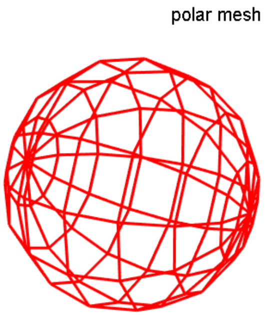




Grid points evenly spaced, easy to determine which cell a point is in.

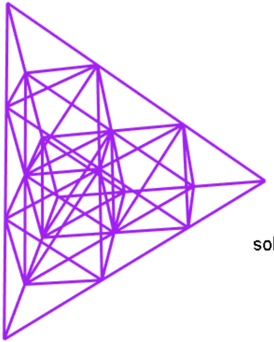
Other shapes are not well described by rectangular grids

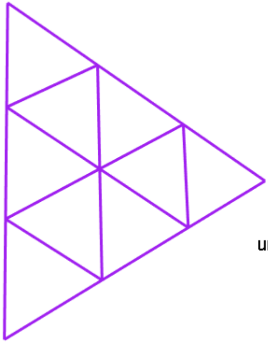




Points are not evenly spread, faces have different numbers of vertices.

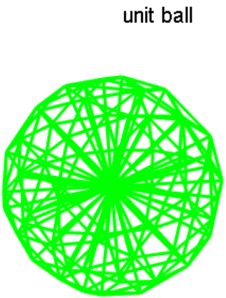
Simplices - equal area subdivisions

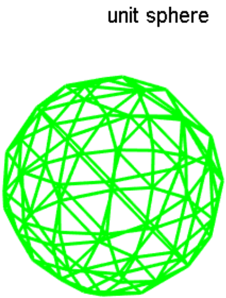




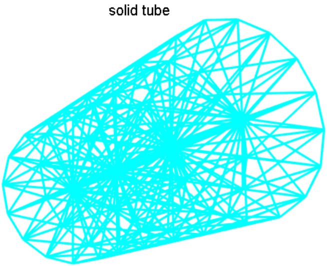
This and following shapes can be generated in any dimension $d \geq 2$.

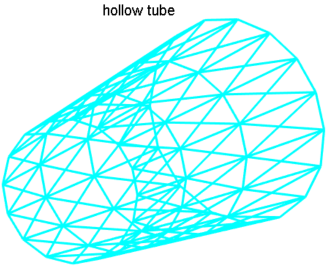
Balls/spheres - approximate equal area subdivisions



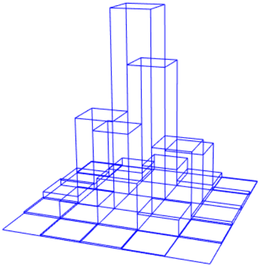


Tubes- approximate equal area subdivisions

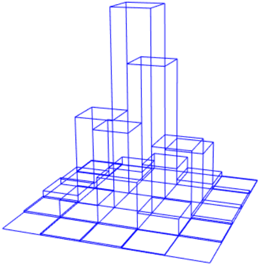


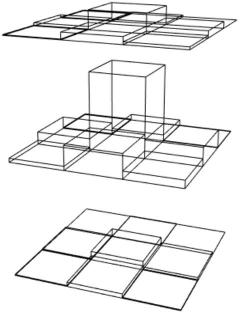


Rectangular histograms

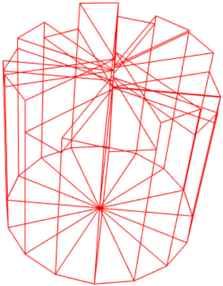


Rectangular histograms

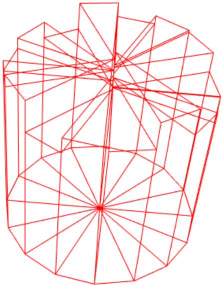


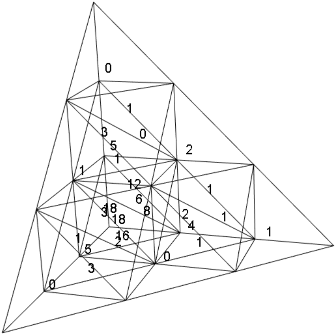


Histograms of non-rectangular regions

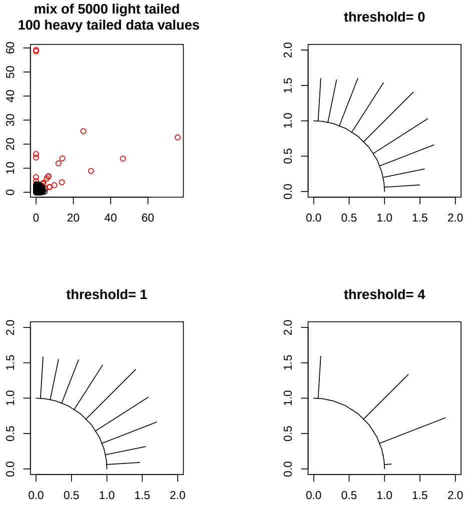


Histograms of non-rectangular regions



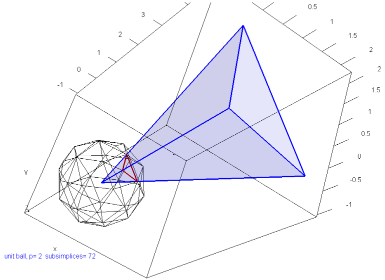


Directional histogram $d = 2$ - count how many in each "direction"

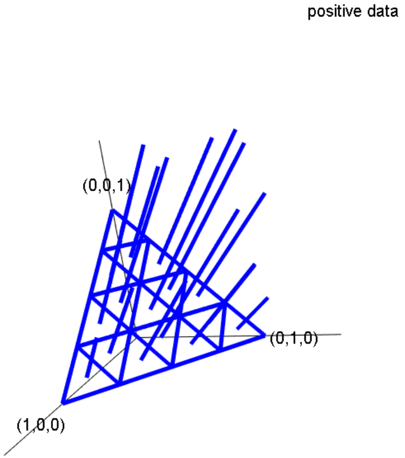


Generalize to $d \geq 3$?

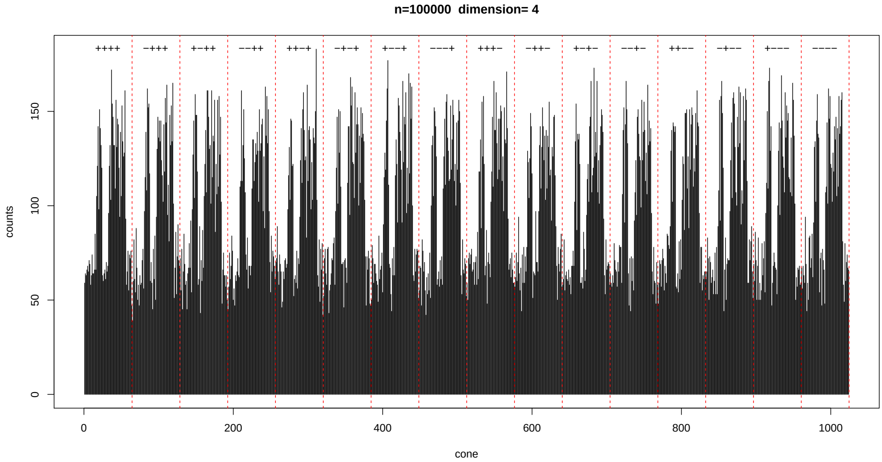
- triangulate/tessellate the sphere
- each simplex on sphere determines a cone starting at the origin
- loop through data points, seeing which cone each falls in
- plot histogram



Directional histogram $d = 3$, positive data

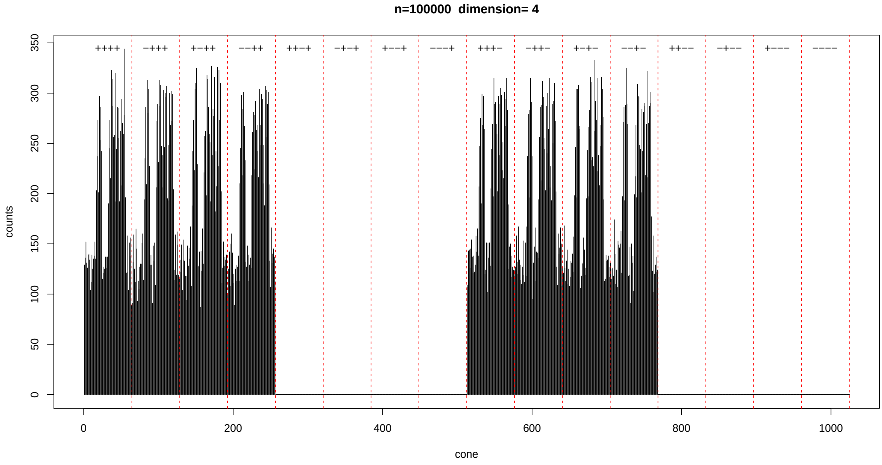


Directional histogram $d = 4$



Radially symmetric data in $\mathbb{R}^4$

Directional histogram $d = 4$



All octants where 3^{rd} component is negative are empty.

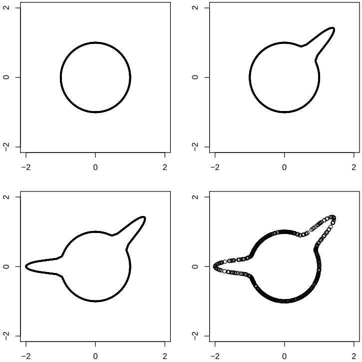
Outline

Generalized spherical distributions

Distributions with level sets that are all scaled versions of a star shaped region. Flexible scheme for building up nonstandard star shaped contours.

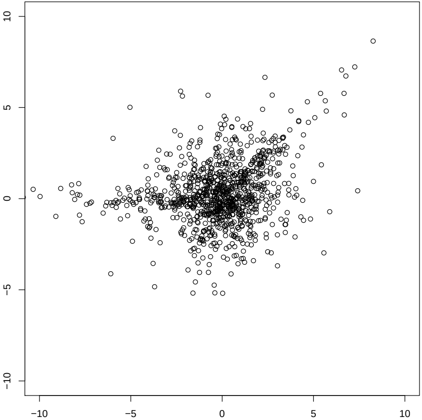
Generalized spherical distributions

Distributions with level sets that are all scaled versions of a star shaped region. Flexible scheme for building up nonstandard star shaped contours.

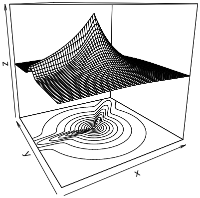


A tessellation based on the added 'bumps' is automatically generated and used in simulating from the contour. Process requires the arclength/ surface area of the contour.

Add a radial component to get a distribution: $\mathbf{X} = R\mathbf{Z}$, where $\mathbf{Z}$ is uniform w.r.t. $(d-1)$-dimensional surface area on contour. Here $R \sim \Gamma(2,1)$

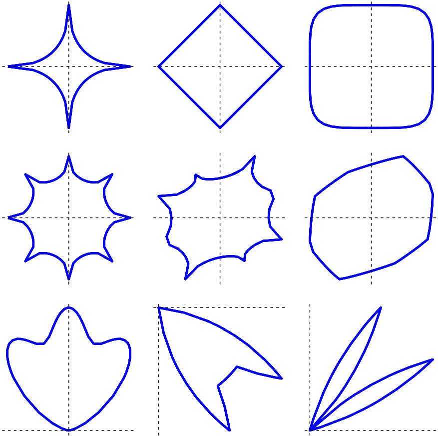


Sample of $\mathbf{X} = R\mathbf{Z}$

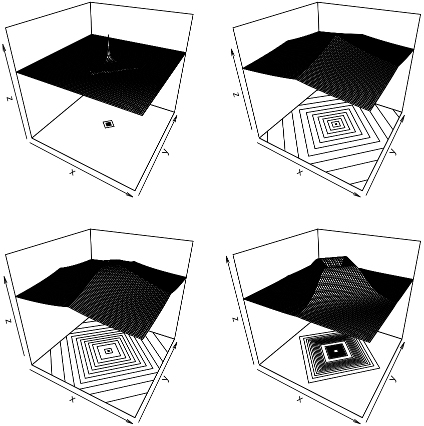


density surface

Many contour shapes possible



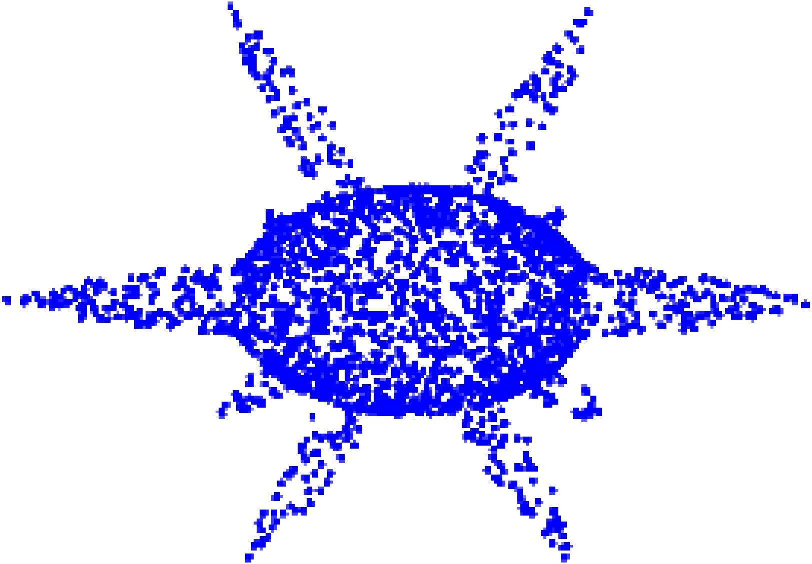
Choice of R determines radial behavior



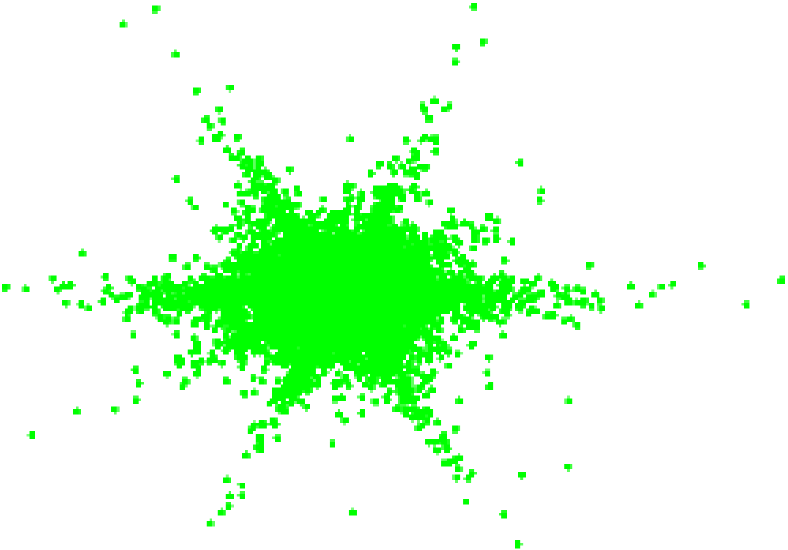
In all cases, contour is a diamond. (a) $R \sim$Uniform(0,1) (b) $R \sim \Gamma(2,1)$
(c) $R = |\mathbf{Y}|$ where $\mathbf{Y}$ is 2D isotropic stable (d) $R \sim \Gamma(5,1)$

3D example - contour

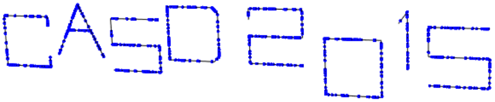
uniform sample from contour



sample from distribution **X** with $R \sim \Gamma(2,1)$

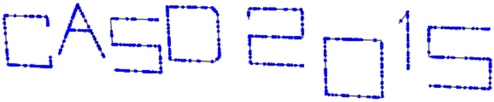


Flexible shapes

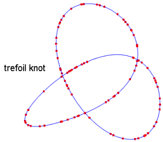


Specified some letters in 3D, can sample from this word proportional to arclength

Flexible shapes



Specified some letters in 3D, can sample from this word proportional to arclength



Outline

Multivariate Fréchet Distributions

de Haan and Resnick (1977): **X** max stable, centered with shape index ξ, is characterized by the angular measure H on the unit simplex $\mathbb{W}_+$. The spread of mass by H determines the joint structure. Define the scale function

$$\sigma^{\xi}(\mathbf{u}) = \int_{\mathbb{W}_+} \left(\bigvee_{i=1}^{d} u^{\xi} w_i \right) H(d\mathbf{w}).$$

(If the components of **X** are normalized and $\xi = 1$, then this is the tail dependence function $\ell(\mathbf{u})$.)

Multivariate Fréchet Distributions

de Haan and Resnick (1977): **X** max stable, centered with shape index ξ, is characterized by the angular measure H on the unit simplex $\mathbb{W}_+$. The spread of mass by H determines the joint structure. Define the scale function

$$\sigma^\xi(\mathbf{u}) = \int_{\mathbb{W}_+} \left(\bigvee_{i=1}^{d} u^\xi w_i \right) H(d\mathbf{w}).$$

(If the components of **X** are normalized and $\xi = 1$, then this is the tail dependence function $\ell(\mathbf{u})$.) The scale function determines the joint distribution:

$$G(\mathbf{x}) = P(\mathbf{X} \le \mathbf{x}) = \exp\left(-\sigma^\xi(\mathbf{x}^{-1}) \right).$$

Observation: need to (a) describe different types of measures and (b) integrate over a surface

R package `mvevd`, $d \geq 2$

- Define classes of mvevds: discrete H, generalized logistic, Dirichlet mixture, piecewise constant and linear angular measures (computational geometry)
- Compute scale functions $\sigma(\mathbf{u})$ for above classes (integrate over simplices, computational geometry)
- Fitting mvevd data with any of the above classes (max projections)
- Exact simulation from these classes (Dirichlet mix - Dombry, Engelke & Oesting (EVA 2015), Dieker and Mikosch (2015))
- Compute cdf $G(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) = \exp(-\sigma^{\xi}(\mathbf{x}^{-1}))$, $(\boldsymbol{\mu} = 0, \mathbf{x} \geq 0)$.
- Computation of density $g(\mathbf{x})$ when known (partitions)
- Computation of $H(S)$ for a simplex S to estimate tail probabilities in the direction S. (computational geometry & integrate over simplices)