

Algorithms for working with multivariate distributions and PCA/ICA for heavy tails

John Nolan

American University
Washington, DC

MURI Columbia Meeting
7 October 2014

1 Spheres, simplices, meshes and integration

2 PCA and ICA for multivariate heavy tailed data

Outline

- 1 Spheres, simplices, meshes and integration
- 2 PCA and ICA for multivariate heavy tailed data

Integration on spheres

Lévy and Feldheim (1930s): \mathbf{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

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

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

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

In both cases, the spread of mass by Λ determines the joint structure.

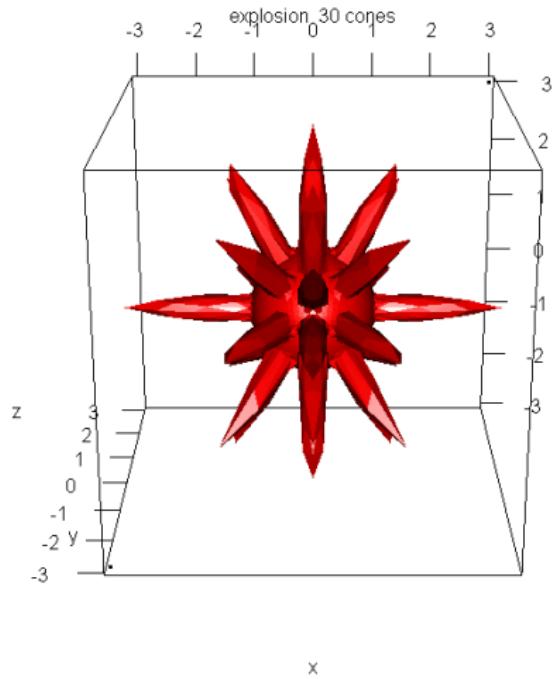
Star-shaped/generalized spherical distributions (Natick)

A distribution is generalized spherical if there is a curve/surface \mathbb{S}_* such that the density $f(\cdot)$ 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.

Need to be able to compute surface integration to get norming constant.
Working on a way to simulate from such distributions using a triangulation of the contour.

Star shaped contour in 3D



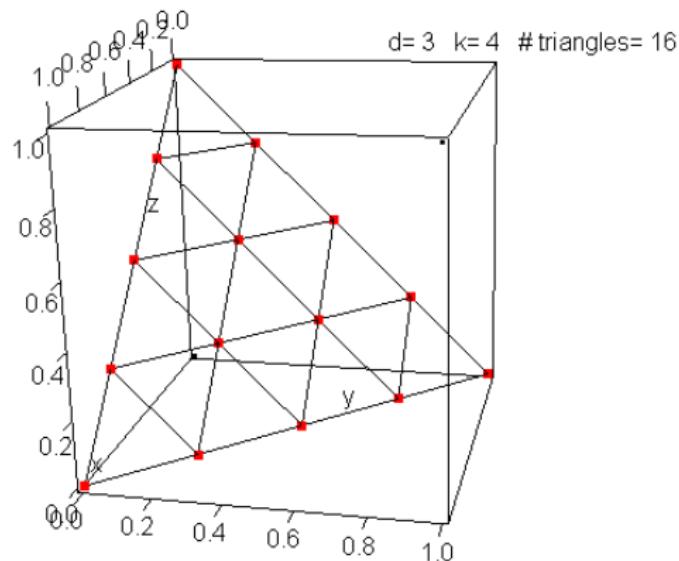
R packages

How to work with spectral measures in higher dimensions? Currently hard to handle much when dimension is bigger than 2.

- SphericalCubature* - integrate over spheres and balls in \mathbb{R}^d
- SimplicialCubature* - integrate over m -dim. simplices in \mathbb{R}^d
- Mesh - define uniform partitions on the simplex, approx. unif. on the sphere
- GeneralizedSpherical - models determined by a contour

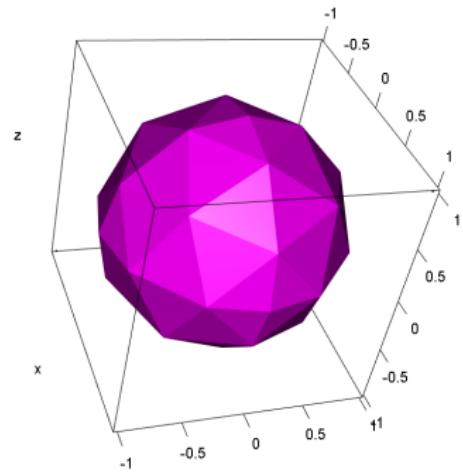
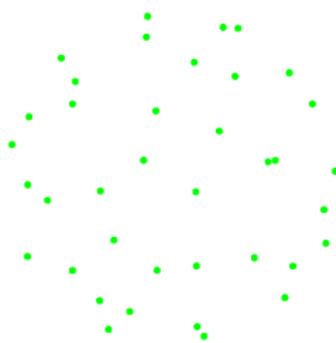
* = on public open source repository CRAN

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



Can project onto a sphere or other surface to get triangulations on them.

Grid vs Mesh



Outline

- 1 Spheres, simplices, meshes and integration
- 2 PCA and ICA for multivariate heavy tailed data

Robust PCA and ICA, joint with T. Alparslan, R. Davis, N., S. Resnick

Principal components analysis (PCA) is a popular technique for analyzing data, tries to extract the directions with maximum dispersion. Traditional PCA can behave poorly when the data is heavy tailed; we propose a robust PCA.

When there is no elliptical structure, we propose using a modified version of Independent Component Analysis (ICA).

We will start with the assumption that the data is multivariate stable.

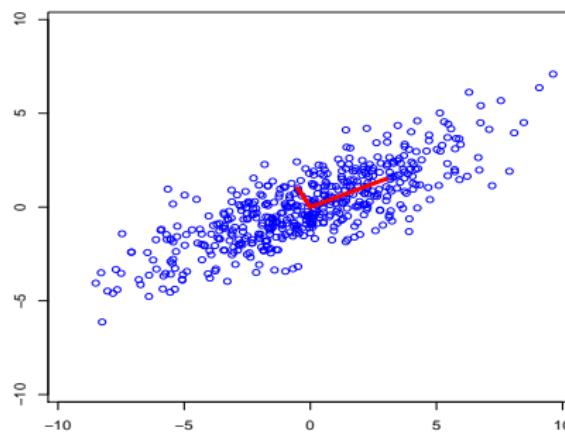
Traditional PCA

Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be a d -dim. sample

Compute the sample covariance matrix S .

Perform an eigenvalue decomposition of S : eigenvalue λ_j with associated eigenvector \mathbf{v}_j , $j = 1, \dots, d$. Assume eigenvalues are ranked:

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$. Then \mathbf{v}_1 is the first principal component, \mathbf{v}_2 is the second, etc.



What happens when data is heavy tailed?

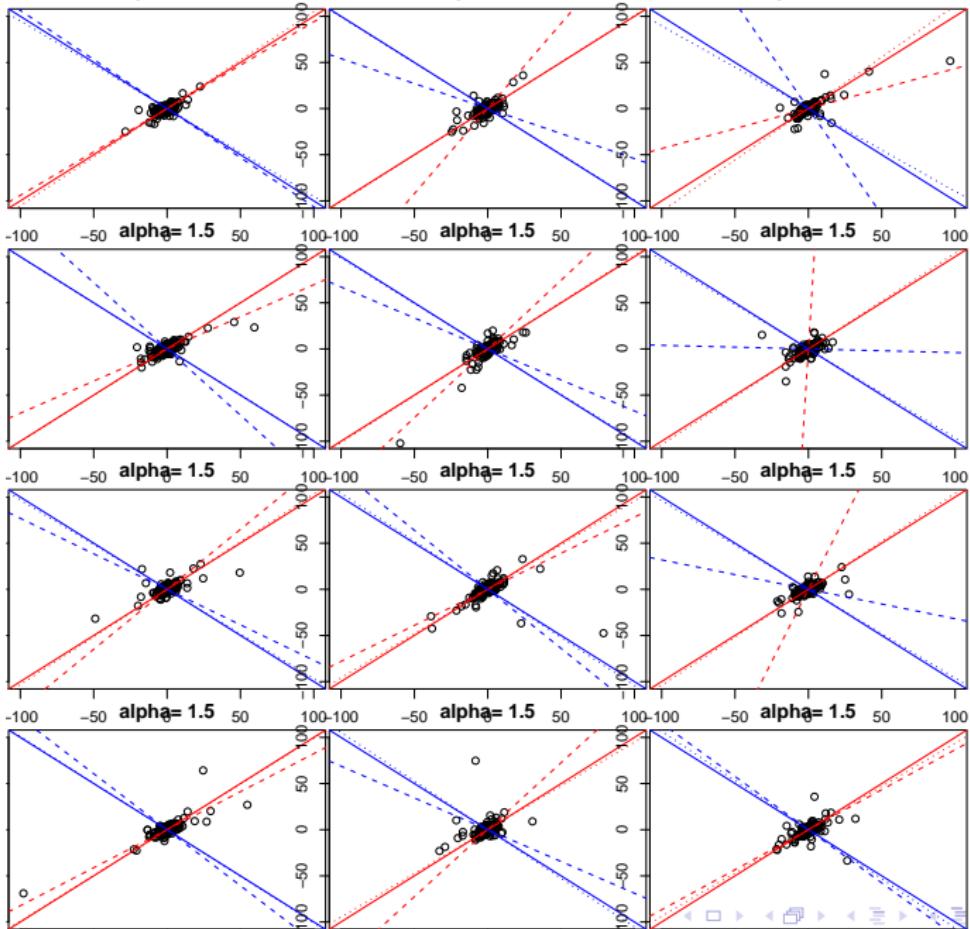
Computing the sample covariance requires calculating the means of all the components, and also the second moment. The mean is heavily influenced by outliers, and second moments are even more heavily influenced by outliers.

Next slide shows simulations with elliptical stable data, $\alpha = 1.5$. Solid lines show exact “principal components”, dashed lines show estimates from traditional PCA.

alpha= 1.5

alpha= 1.5

alpha= 1.5



Robust PCA in the elliptical stable case

When \mathbf{X} is elliptical stable, there is a corresponding shape matrix R , a $d \times d$ matrix that determines the shape. Two representations:

$$\mathbf{X} = R^{-1/2} \mathbf{Z} + \delta,$$

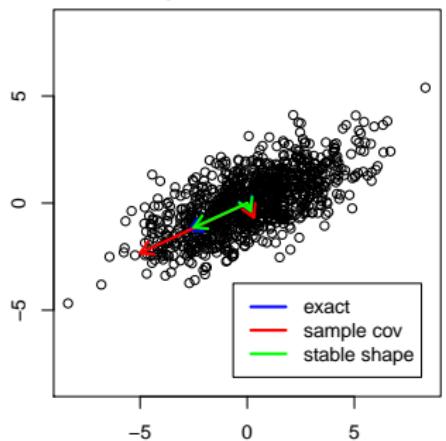
where \mathbf{Z} is isotropic/radially symmetric α -stable.

$$\mathbf{X} = A^{1/2} \mathbf{G} + \delta,$$

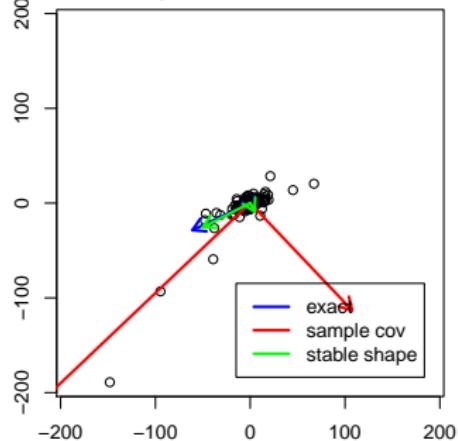
where $A > 0$ is a positive $(\alpha/2)$ -stable univariate stable r.v. and $\mathbf{G} \sim N(0, R)$.

Robust PCA: estimate center δ and shape matrix R by a method that takes into account the heavy tails in the data. N. (2013) provided methods to do this.

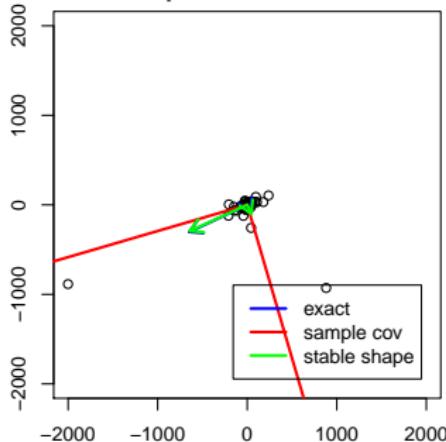
$\alpha = 2 \ n = 1000$



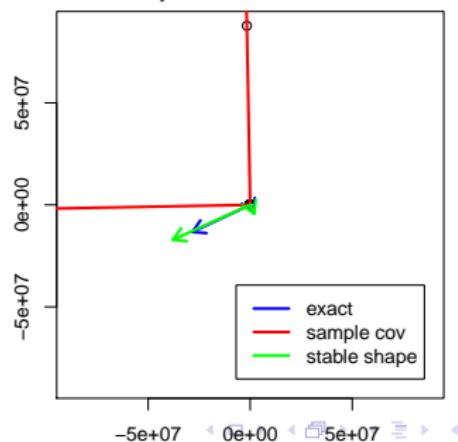
$\alpha = 1.5 \ n = 1000$



$\alpha = 1 \ n = 1000$



$\alpha = 0.5 \ n = 1000$



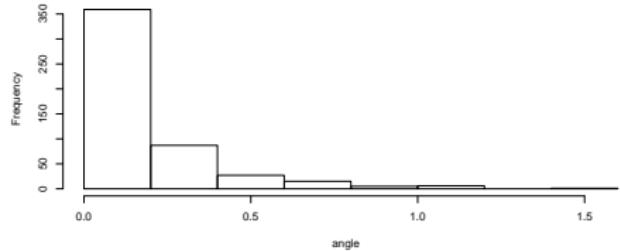
angle between exact & estimated first eigenvector

alpha= 1.5

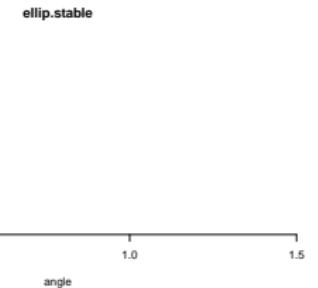
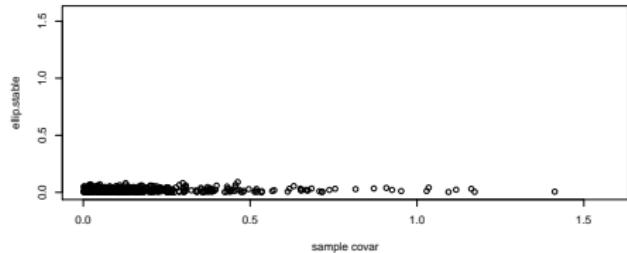
n= 1000 m= 500

rho= 0.7

sample covar



angle



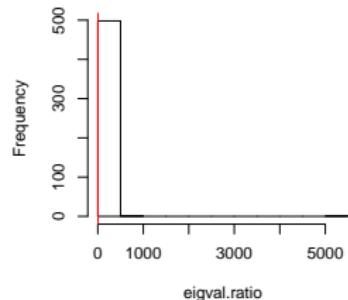
sample covar

eigenvalue ratio, exact= 5.667

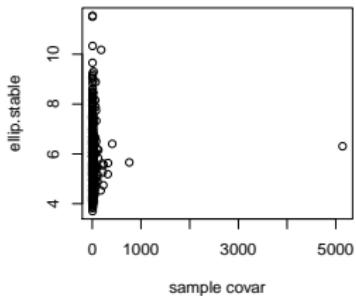
alpha= 1.5

n= 1000 m= 500

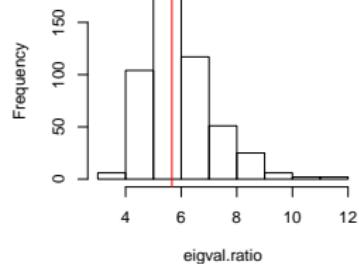
rho= 0.7



eigval.ratio



ellip.stable



Non-elliptical stable case

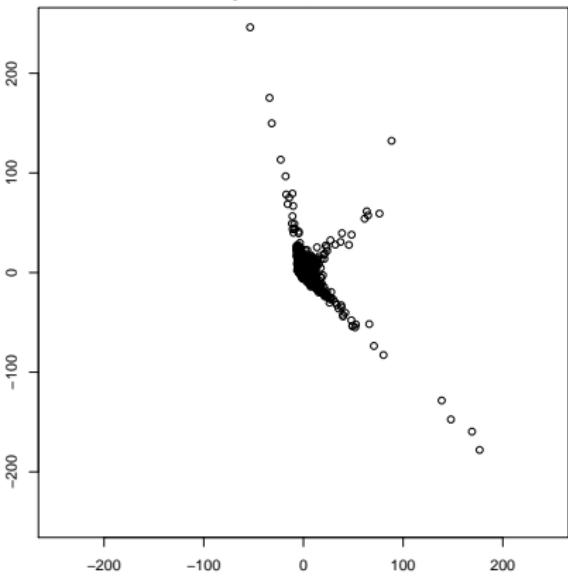
Here it can be meaningless to use PCA, even the robust PCA described above: there may be very non-elliptical dependence structures. One interesting case is when there are independent components:

$$\mathbf{X} = A\mathbf{Z},$$

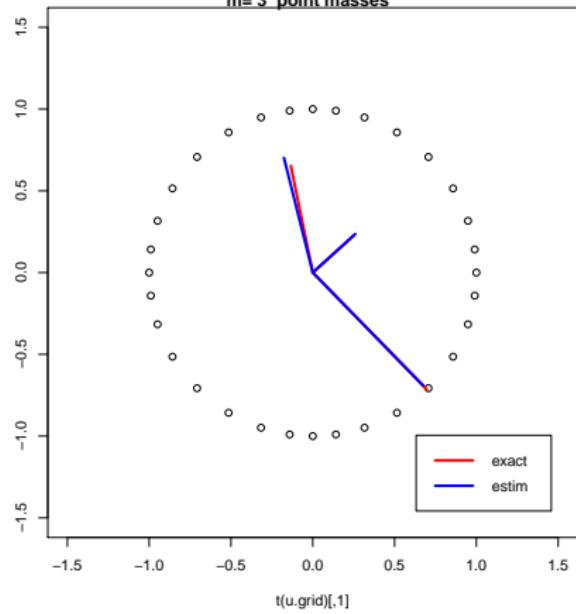
where A is a $d \times m$ matrix of coefficients and $\mathbf{Z} = (Z_1, \dots, Z_m)$ are i.i.d. stable. (Equivalently, the spectral measure of \mathbf{X} is discrete.)

Want an robust ICA Note that we can have $m < d$, $m = d$, or $m > d$.

$\alpha = 1.5, n = 5000$



32 gridpoints
m=3 point masses



ICA in the stable case: m known

Let \mathbf{X}_i , $i = 1, \dots, n$ be a sample from a multivariate stable distribution. Use the fact that linear combinations of a multivariate stable r.v. are univariate stable.

Pick a grid $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{n\text{grid}}$. For each $j = 1, \dots, n\text{grid}$, calculate the univariate data set: $y_{j,i} = \langle \mathbf{u}_j, \mathbf{X}_i \rangle$, $i = 1, \dots, n$. Let $\hat{\gamma}_j$ and $\hat{\beta}_j$ be the univariate scale and skewness of this projection in direction \mathbf{u}_j .

$$A^* := \arg \min_A \sum_{j=1}^m (\hat{\gamma}_j^\alpha - \gamma_{j,A}^\alpha)^2 + \sum_{j=1}^m (\hat{\beta}_j \hat{\gamma}_j^\alpha - \beta_{j,A} \gamma_{j,A}^\alpha)^2,$$

where $\gamma_{j,A}$ and $\beta_{j,A}$ are the exact scale and skewness for the projection in direction \mathbf{u}_j for the ICA model given by $\mathbf{X} = A\mathbf{Z}$.

ICA in the stable case: m unknown

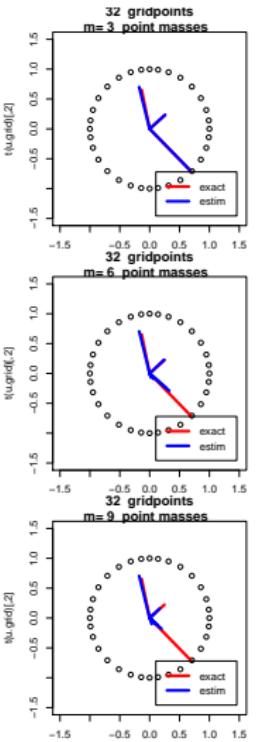
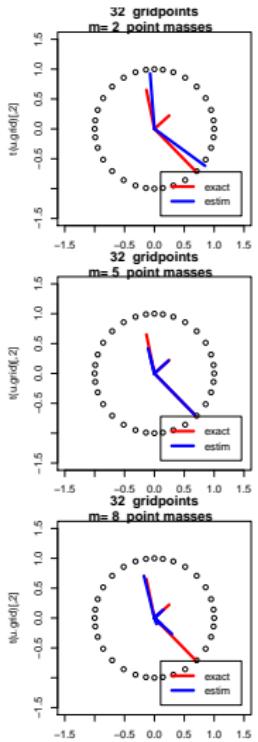
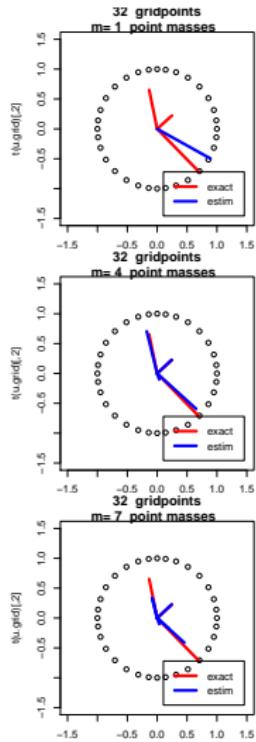
Vary m in some range and look for point where larger values of m don't add much to the fit.

AICc seems to do a good job selecting correct m .

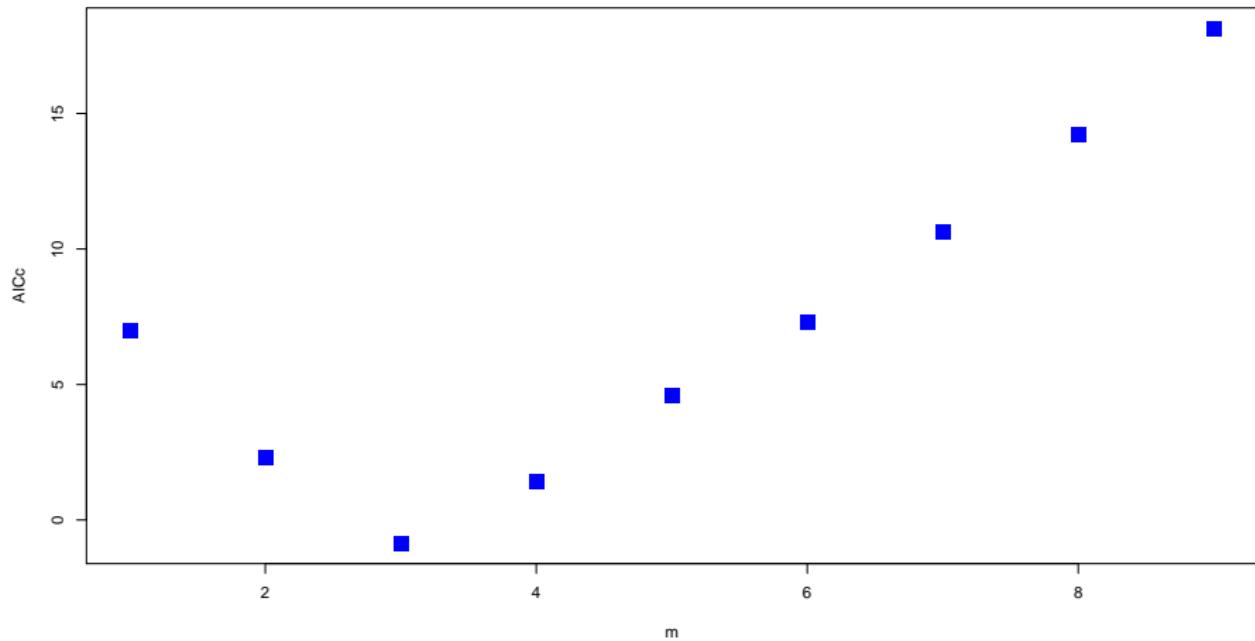
$$\text{AICc} = 2m + 2 \log(\text{ObjFn}) + \frac{2m(m+1)}{ngrid - m - 1},$$

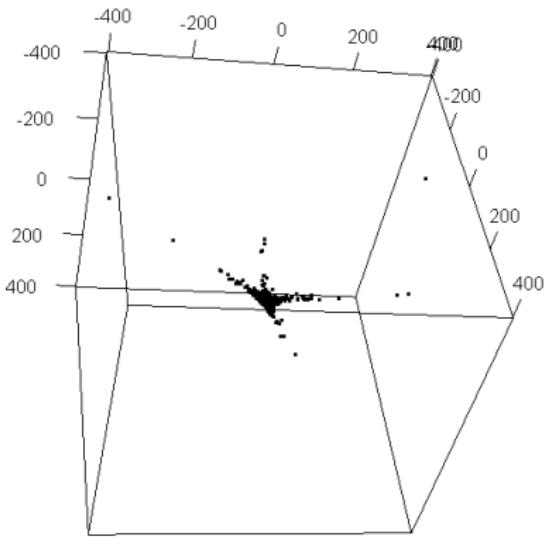
where ObjFn is the optimal value of the objective function on previous page.

Varying m

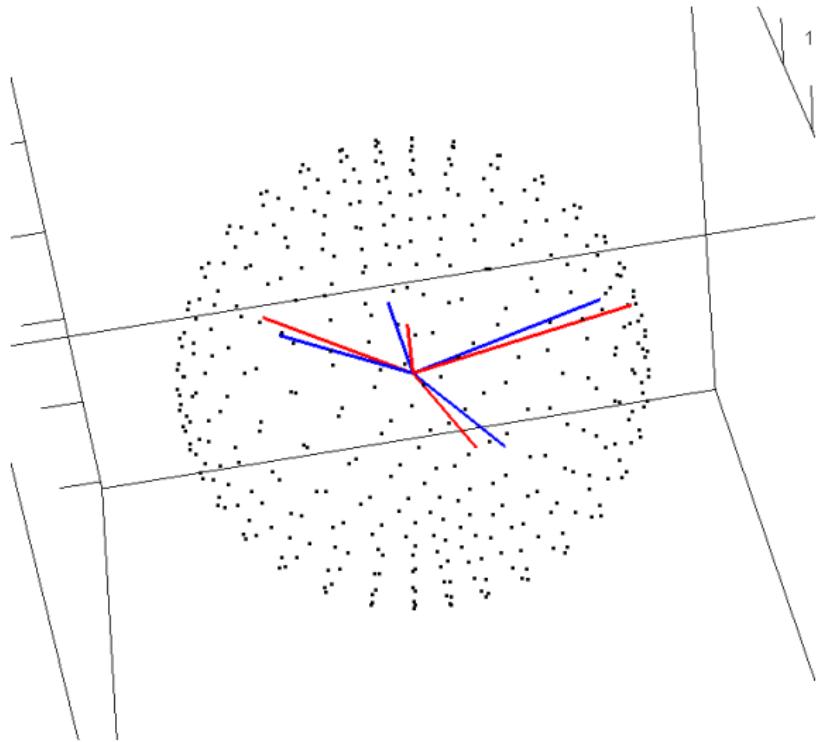


determining m





$\alpha = 1.5, n = 5000$



Future

- Non-stable case?
PCA should work - perhaps as is, perhaps replace scale estimate with some robust measure of scale, e.g. interquartile range.
ICA - not clear
- Measure of independence for multivariate stable laws
- Computational methods for multivariate stable distributions
- ICA parallels with multivariate extreme value distributions