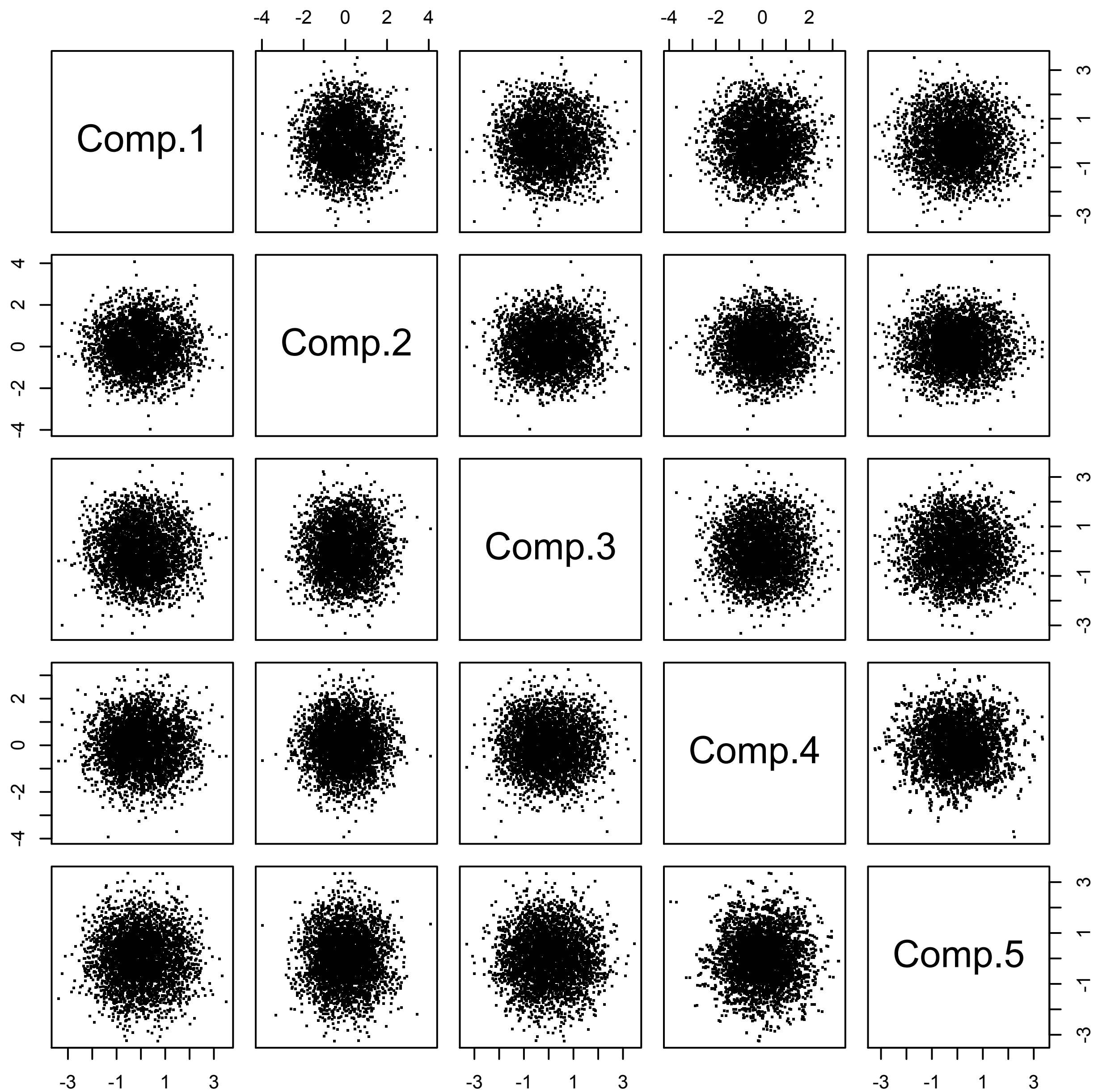
Finding Holes in Multivariate Data

Woollcott Smith
Temple University

http://astro.temple.edu/wksmith/

The goal is to find a hole in a k-dimensional cloud of points. By a hole we mean a local anomaly in the data where a k-dimensional elliptical region within the cloud contains fewer data points than expected, given the data points in the surrounding region.

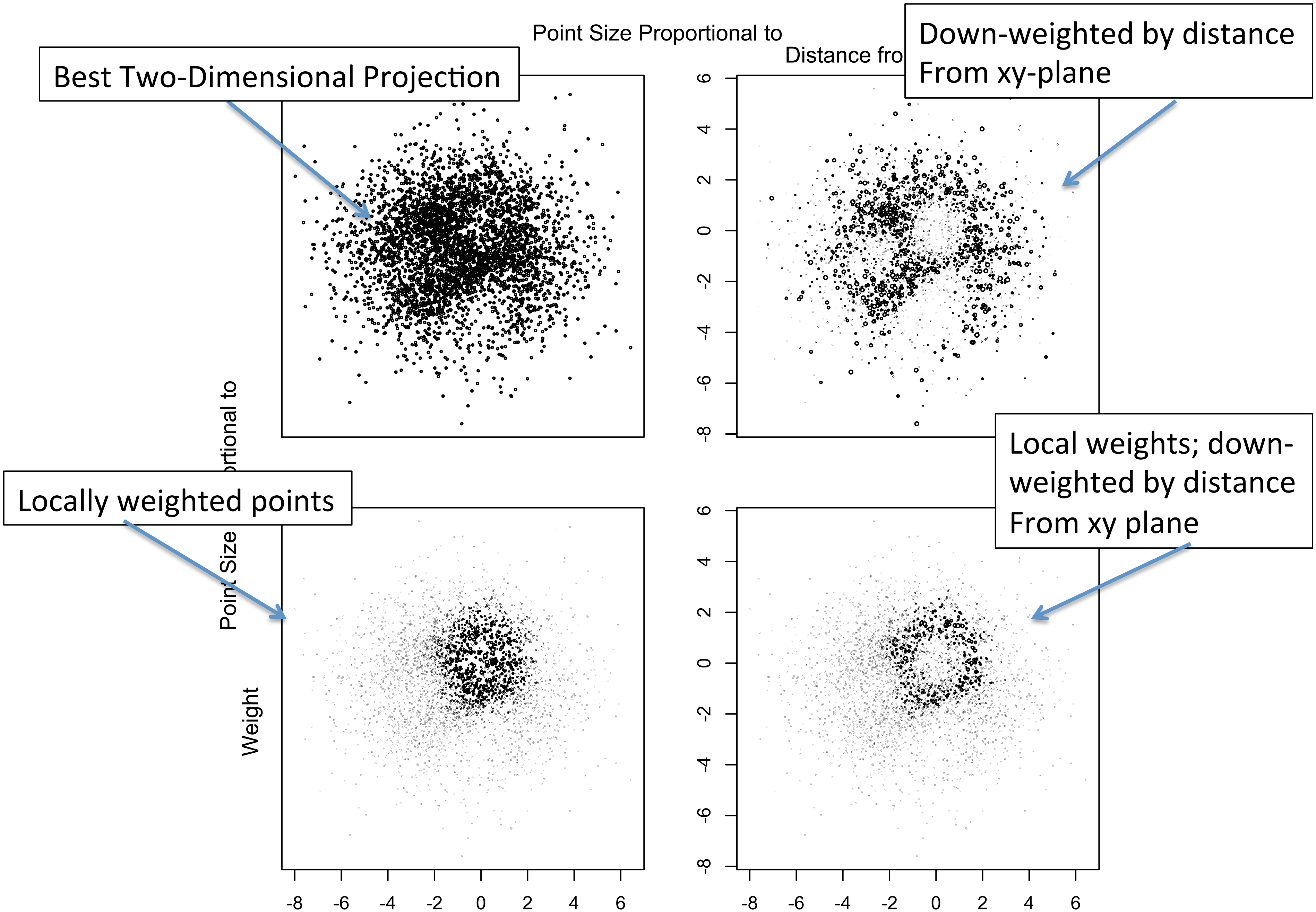
First a famous example: The Pollen Data Set.
An artificial data set created by
David Coleman for the 1987 JSM data competition.


Comp.1
Comp.2
Comp.3
Comp.4
Comp.5
-4 -2 0 2 4
-4 -2 0 2
-3 -1 1 3

Two basic ideas

- Increase the data weight in the neighborhood of the hole.
- Scatterplots that down-weight points by distance from the x-y plane.

The Pollen data again:


Point Size Proportional to
Distance fro
Best Two-Dimensional Projection
Down-weighted by distance
From xy-plane
Local weights; down-
weighted by distance
From xy plane
Locally weighted points
Point Size Proportional to
Weight
6
4
2
0
-2
-4
-6
-8
-8 -6 -4 -2 0 2 4 6

Outline

- Local weighting of data points near a hole: change of measure and importance sampling
- An objective function
- Optimization procedure
- Representing an *m*-dimensional hole in two dimensions
- Woods Hole tide data

Local Weighting

Using Classical Multivariate Normal Results

1. Multivariate normal density

$$n(\mathbf{x};\mu,\Sigma)=\frac{1}{(2\pi)^{k/2}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^t\,\Sigma^{-1}(x-\mu)\right)$$

2. Weighting function

$$g(\mathbf{x};\mu_w,\Sigma_w)=\exp\left(-\frac{1}{2}(x-\mu_w)^t\,\Sigma_w^{-1}(x-\mu_w)\right)$$

Change of Measure
Importance Sampling

Standard Bayes calculation for the multivariate normal

Local normal density

$$n(\mathbf{x};\mu_l,\Sigma_l) = C \quad n(\mathbf{x};\mu,\Sigma)\, g(\mathbf{x};\mu_w,\Sigma_w)$$

Local Covariance Matrix

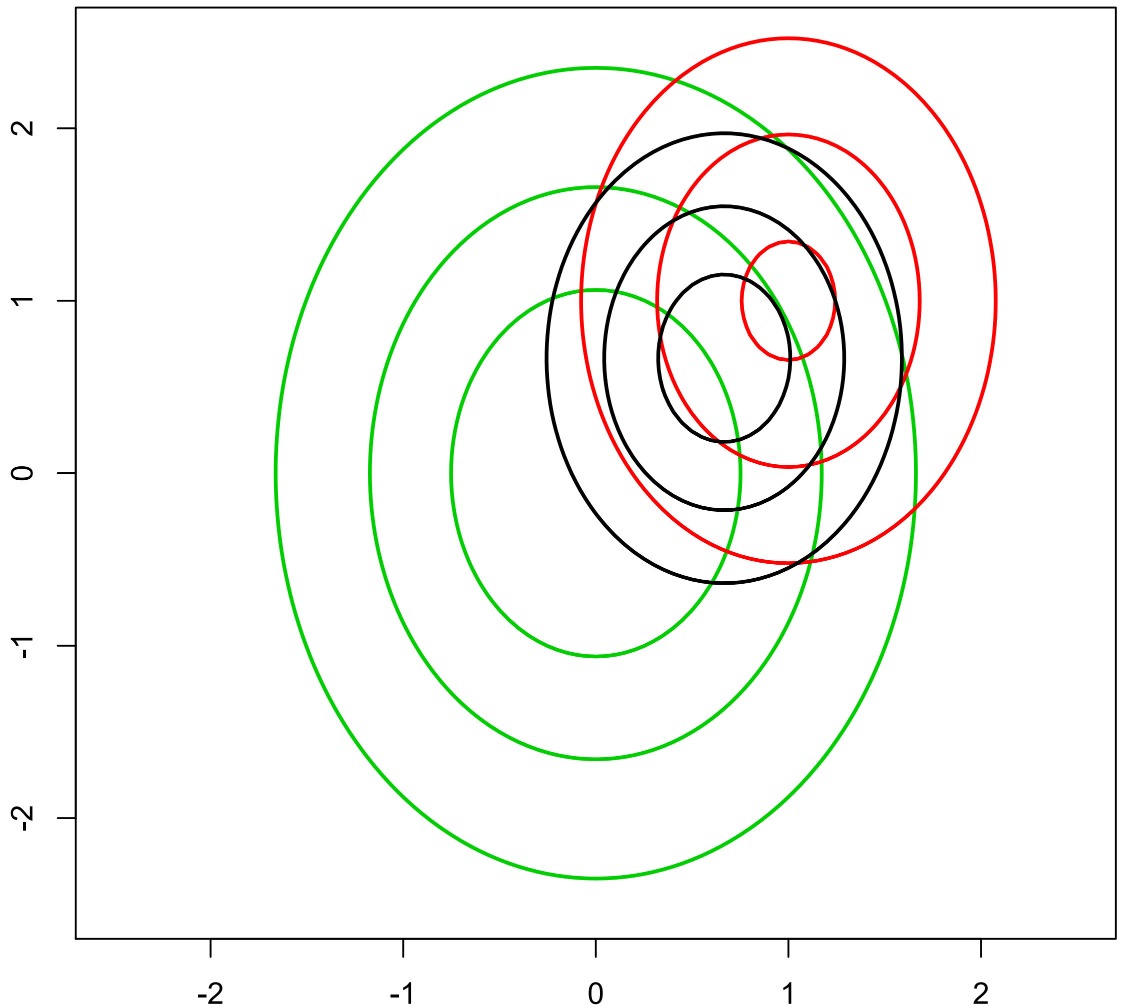
$$\Sigma_l = \left(\Sigma^{-1} + \Sigma_w^{-1}\right)^{-1}$$

Local Mean

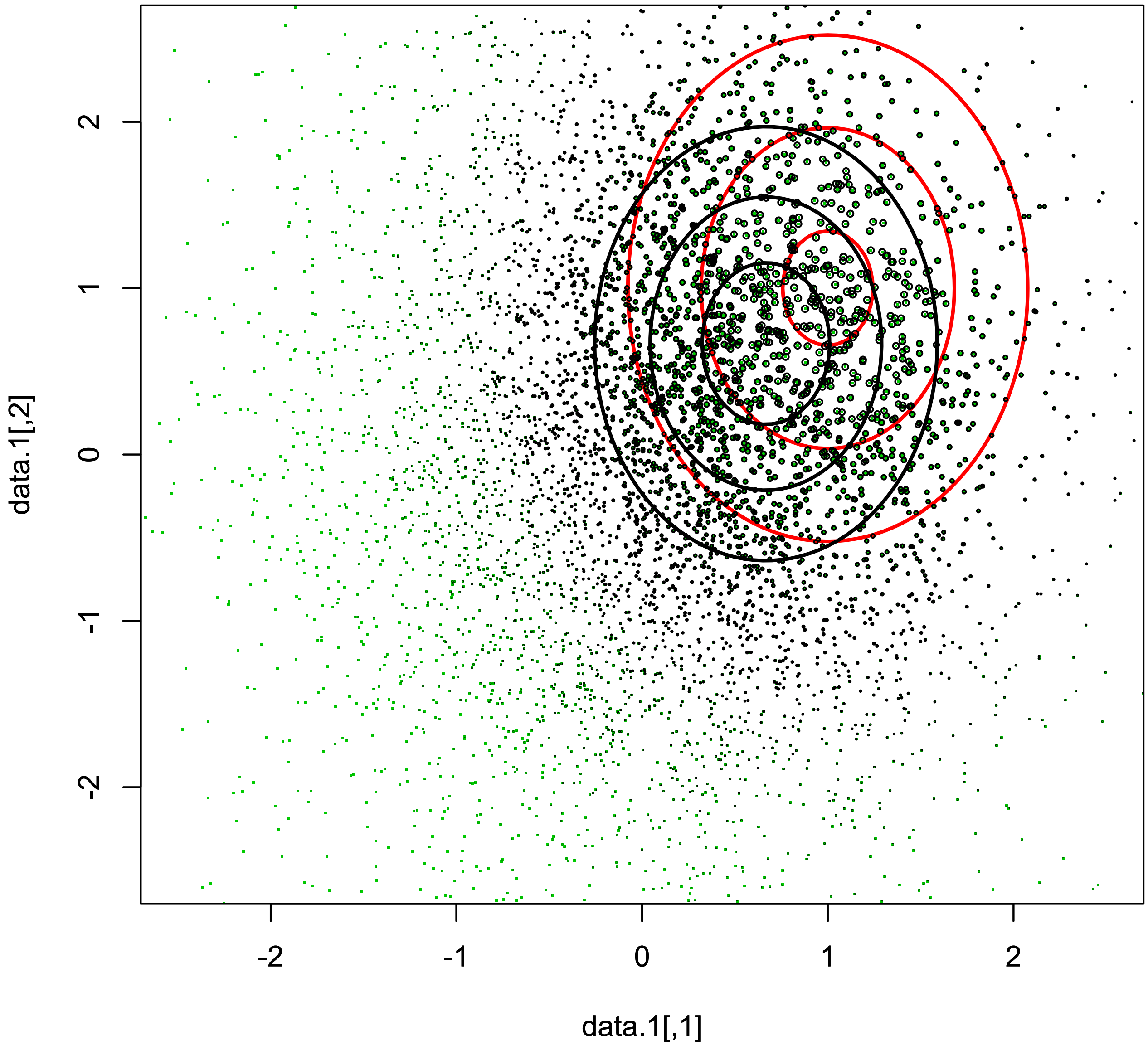
$$\mu_l = \Sigma_l \left(\Sigma^{-1}\mu + \Sigma_w^{-1}\mu_w\right)$$

Normalizing constant

$$C^{-1} = \frac{\left|\Sigma_l\right|^{\frac{1}{2}}}{\left|\Sigma\right|^{\frac{1}{2}}} \exp\left(+ \frac{1}{2}\left(\mu_l^t \Sigma_l^{-1} \mu_l - \mu^t \Sigma^{-1} \mu - \mu_w^t \Sigma_w^{-1} \mu_w\right)\right)$$



$$n(\mathbf{x};\mu_l,\Sigma_l) = C \;\; n(\mathbf{x};\mu,\Sigma)\, g(\mathbf{x};\mu_w,\Sigma_w)$$



$$n(\mathbf{x};\mu_l,\Sigma_l) = C \; n(\mathbf{x};\mu,\Sigma)\, g(\mathbf{x};\mu_w,\Sigma_w)$$

General Hole Finding Procedure

1. For fixed weight parameters μ_w and Σ_w compute the weights

$$w_i = \frac{g(x_i, \mu_w, \Sigma_w)}{\sum_{i=1}^{n} g(x_i, \mu_w, \Sigma_w)}$$

2. Estimate the local parameters

$$\hat{\mu}_L = \sum_{i=1}^{n} w_i \mathbf{x_i} \quad \text{and} \quad \hat{\Sigma}_l = \sum_{i=1}^{n} w_i \left(\mathbf{x_i} - \hat{\mu}_l\right)\left(\mathbf{x_i} - \hat{\mu}_l\right)^t$$

3. Compute the objective function.

$$\overline{p}\left(\mu_w, \Sigma_w\right) = \sum_{i=1}^{n} g(\mathbf{x}_i; \hat{\mu}_l, \alpha\,\hat{\Sigma}_l)\; w_i$$

4. Minimize $\overline{p}\left(\mu_w, \Sigma_w\right)$ with respect to μ_w and Σ_w.

Numerical Details

1) Parameters are constrained so that

$$\sum_{i=1}^{n} g\left(x_i, \mu_w, \Sigma_w\right) = n_w << n$$

In these examples n_w is less than 10% of the data.

2) The Cholesky decomposition is used to parameterize the matrix

$$\Sigma_w^{-1} = LL^t$$

3) We use R, nlminb and selected starting values to search for the minimum.

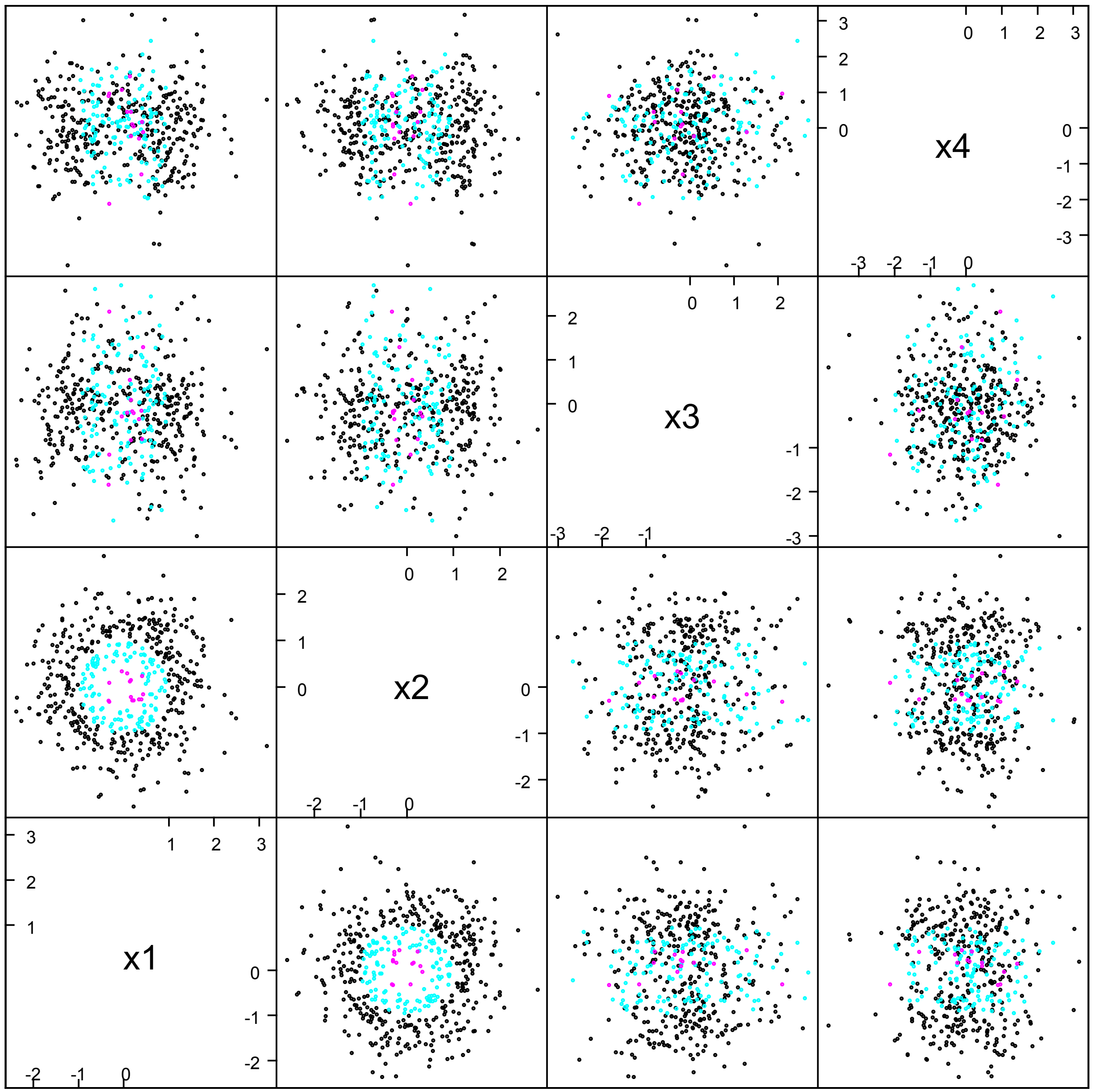
Scatterplot Displays for *k*-dimensional Holes

- 2-dimensional holes in *k*-dimensional space is just a projection-pursuit problem.

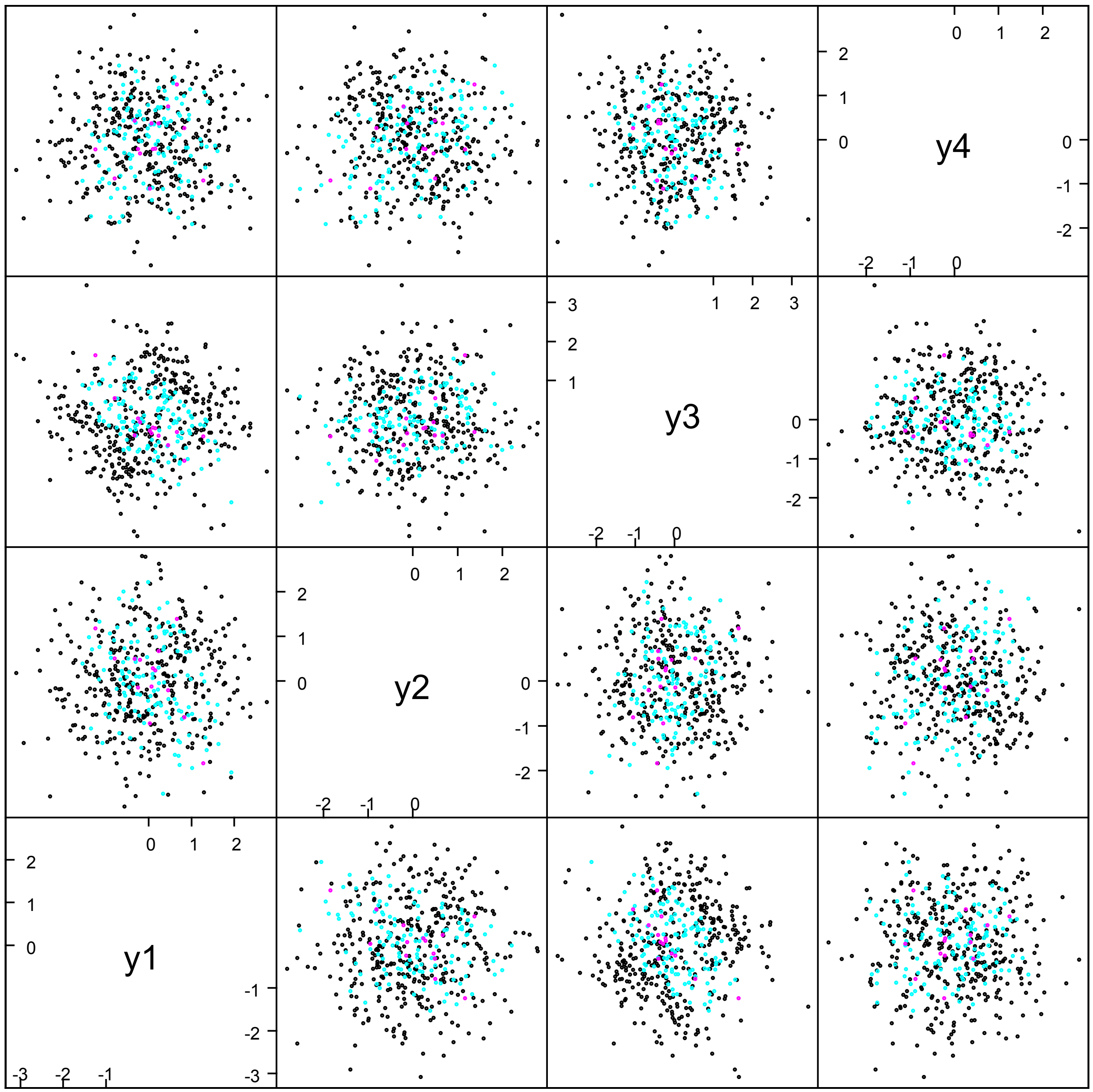
- Random points on a *k*-dimensional sphere.

- Scatterplots for for *k*-dimensional holes.

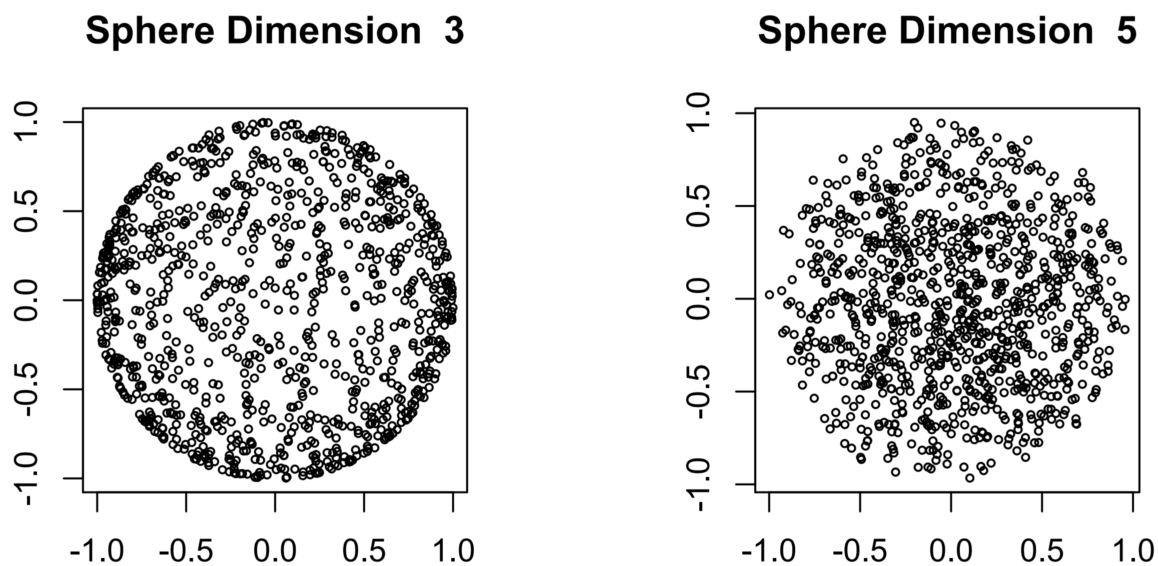
Model Hole in X1 X2 Plane: Hole labeled by distance



Orthogonal Transform of Model Data : Hole labeled by distance



Classic Example, Random points on the unit Sphere



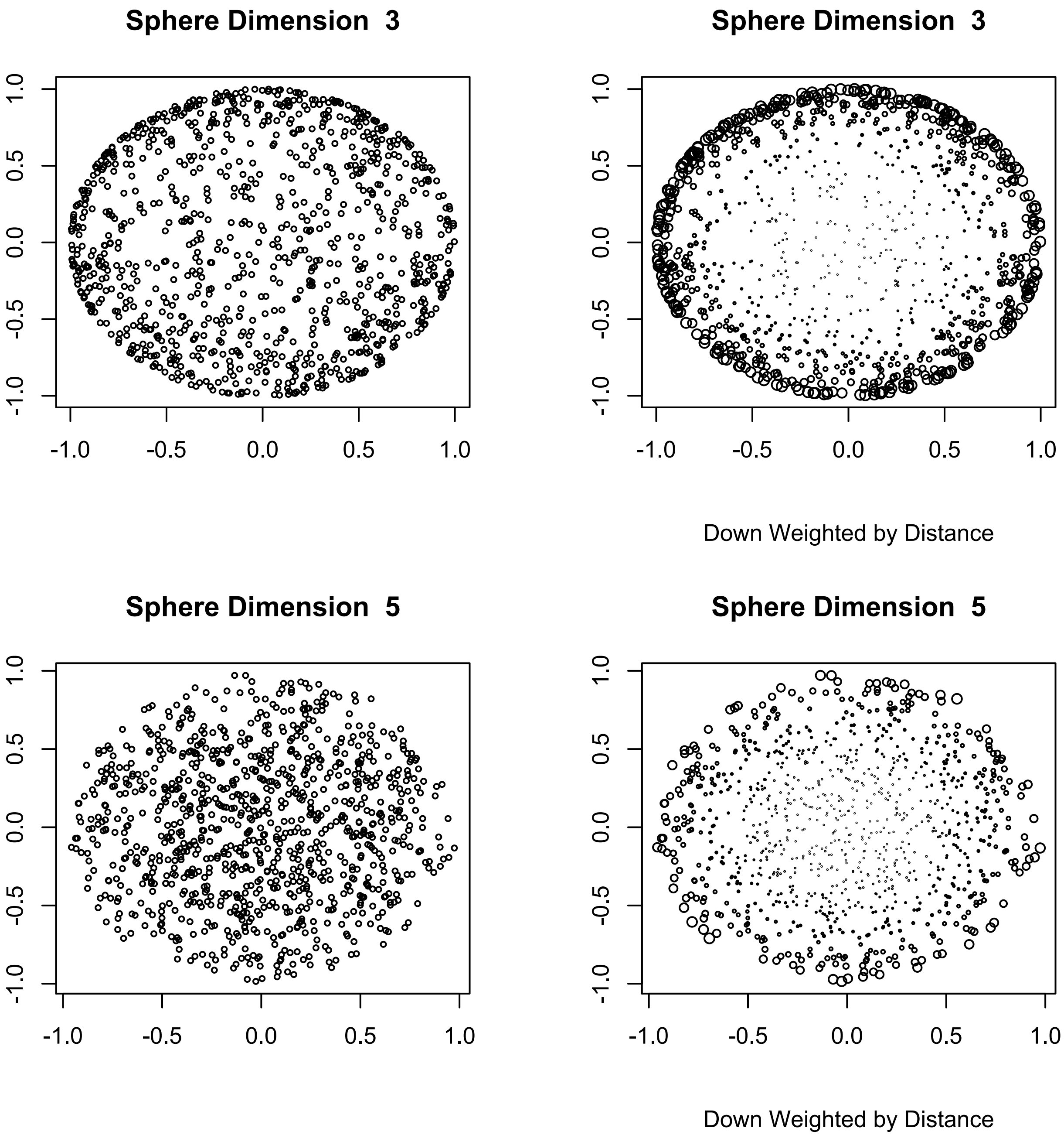
Solution: Distance from the x-y plane

$$d_i = \sum_{j=3}^{k} x_{ij}^2 \ .$$

Down-weight points by

$$\eta_i = \exp(-c\, d_i),$$

where c denotes a constant dependent on dimension and point density.



Sphere Dimension 3
Sphere Dimension 3
Down Weighted by Distance
Sphere Dimension 5
Sphere Dimension 5
Down Weighted by Distance
-1.0
-0.5
0.0
0.5
1.0

Sphere the Locally Weighted Data

Weighted mean and covariance

$$\hat{\mu}_l = \sum_{i=1}^{n} w_i \mathbf{x_i} \quad \text{and} \quad \hat{\Sigma}_l = \sum_{i=1}^{n} w_i \left(\mathbf{x_i} - \hat{\mu}_l\right)\left(\mathbf{x_i} - \hat{\mu}_l\right)^t$$

Sphered data, centered at 0 and with unit sample covariance matrix

$$\mathbf{x}_i^* = \left(x_i - \hat{\mu}_l\right) V_l \Lambda^{-1/2},$$

V_l denotes the matrix of eigen vectors of $\hat{\Sigma}_l$.

Λ denotes the diagonal matrix of eigen values of $\hat{\Sigma}_l$.

Final step: Standard Projection Pursuit

Find a k-by-2 orthonormal matrix M that minimizes the weighted projected points in the center of the (y_1, y_2) plane.

project $\quad \mathrm{y}_i = x_i^* M$

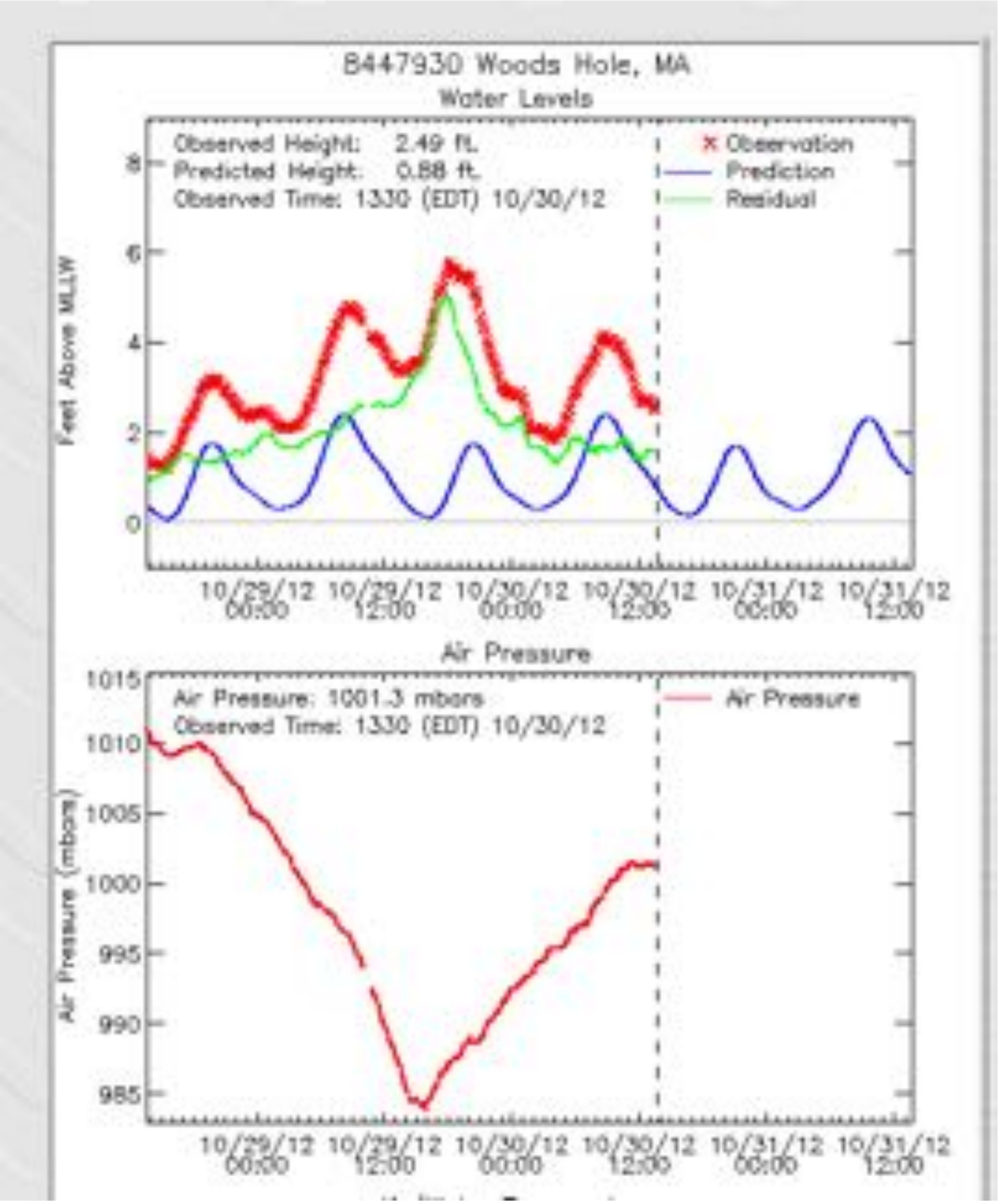
minimize $\quad \sum_{i=1}^{n} w_i \exp(-a\, y_i\, y_i^t)$

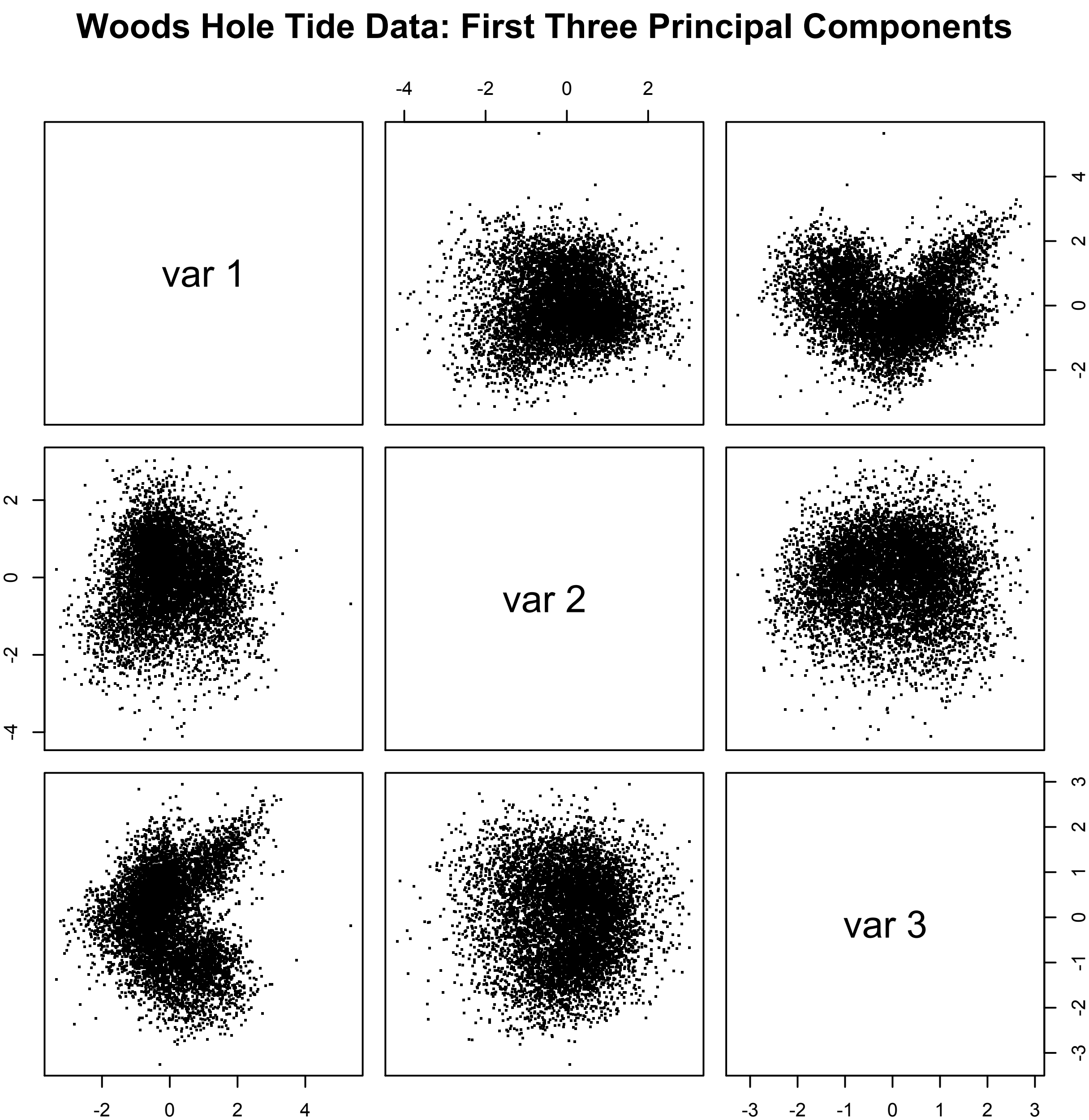
Woods Hole Tide Gauge data

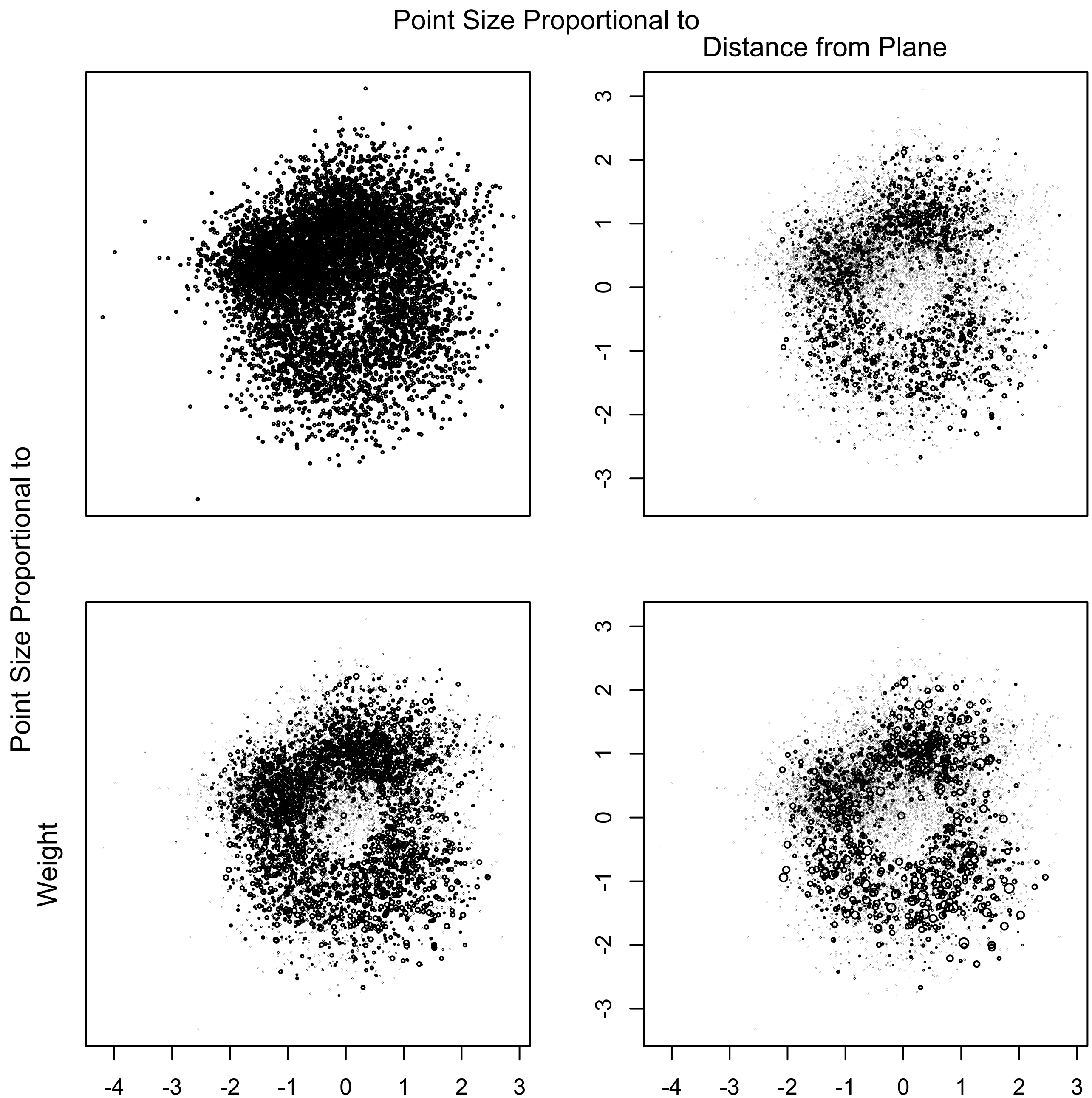
1. Woods Hole hourly water level measurements.
2. Measurements lagged by 1 through 5 hours.
3. The data was thinned by using every third observation.


Tide gauge

Hurricane Sandy Data
NOAA web




Woods Hole Tide Data: First Three Principal Components
var 1
var 2
var 3
-4 -2 0 2
4 2 0 -2
2 0 -2 -4
3 2 1 0 -1 -2 -3
-2 0 2 4
-3 -2 -1 0 1 2 3



Point Size Proportional to
Distance from Plane
Point Size Proportional to
Weight
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3

Hole Models are essentially random truncation models.
For one-dimensional holes see
Morrell, C. H. and Johnson R. A.(1991), "Random truncation and neutrinos", Technometrics:,33, 429-440.

Pollen Data

Becker, R. A., L. Denby, R. McGill and A. R. Wilks(1986). Datacryptanalysis: A Case Study. *Proceedings of the Section on Statistical Graphics of Amer. Statistical Association*, 1987, pp. 92-97. (And a Bell Labs Tech Report)

Found the three holes. But only because of a special property of the simulation.

- In the simulation all data points were paired.
- Truncated points were then the missing points in a pair.
- The missing points formed three clusters.

Finding Holes in Data using two- and three-dimensional non parametric density estimation.

Scott, D.W. (2009). Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley, New York.

Nested α-level density contours indicates a hole.

The general consensus seems to be that finding holes in data is

Too Hard
and
Too Esoteric.

Locally-Weighted Hole Finder

- Is robust to departures from multivariate normality.
- Is based on classical multivariate analysis.
- Solves a rather hard problem
  in a transparent and straightforward way.
- Generalizable to other departures from "Local Normality".

These slides are on my web site at Temple University

*astro.temple.edu/wk**smith**/*