Nonparametric Function Estimation
Stat 550[1] Chapters 7-8
Curse of Dimensionality & Regression

David W Scott[2]

Rice University

November 2
Fall 2023
Rice University

[1] A course based upon the 2nd edition of *Multivariate Density Estimation; Theory, Practice, and Visualization*, *John Wiley & Sons, 2015*

[2] www.stat.rice.edu/∼scottdw/

Chapter VII: The Curse of Dimensionality and Dimension Reduction

- The practical focus of most of this book is on density estimation in "several dimensions" rather than in very high dimensions.
- Assumption that a very detailed view of high-dimensional data is difficult, so work in a subspace.
- Chicken and egg problem: don't really know what we are missing in 6-12 dimensions
- Consider standard multivariate regression

$$y_i = \sum_{j=1}^{d} a_j x_{ij} + \epsilon_i = \mathbf{a}^T \mathbf{x}_i + \epsilon_i$$

- the relevant structure of the solution data space is precisely 2-dimensional:

$$\{(w_i, y_i) : 1 \leq i \leq n\}, \text{ where } w_i \equiv \mathbf{a}^T \mathbf{x}_i$$

Projection of high-dimensional data

- The choice is whether to work with techniques in the full dimension, or to first project and then work in the subspace.
- My choice is to project, if you are trying to understand structure
- Recent progress in "deep learning" has lead to huge neural networks that produce impressive results, but hard to interpret
- Can a kernel estimate of 20-dimensional data be constructed to aid projection? Almost no one believes so.
- But even a few decades ago, many held the belief that bivariate nonparametric estimation required prohibitively large samples to be sufficiently accurate.
- Yet we saw in Chapter I that data are very spread out in high dimensions

Unusual Structure: Volcano

- We show an example where two different kinds of structure have similar contours, namely, densities that are spherically symmetric with "holes"

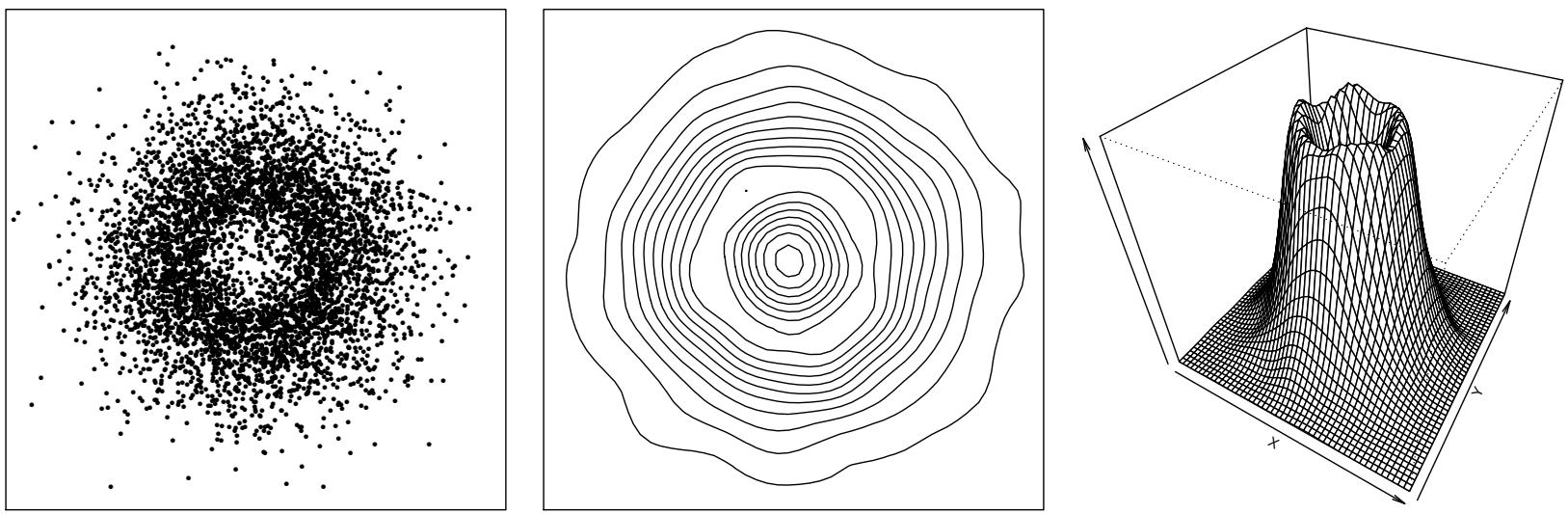


Figure: Scatter diagram and ASH representations of bivariate data from a density with a hole ($n = 5{,}000$). Observe how some contours occur in nested pairs.

Equivalent Sample Sizes (multivariate normal data)

- No perfect way of comparing errors in different dimensions
- Several suggestions:
  - dimensionless form of *MISE*

$$[\sigma_1\sigma_2\cdots\sigma_d \times MISE]^{1/2}$$

  - alternative

$$[MISE/R(f)]^{1/2}$$

  - focus on hardest point to estimate

$$RRMSE(0) = \sqrt{\text{MSE}\{\hat{f}(0)\}}/f(0)$$

  - root coefficient variation

$$RCV(0) = \sqrt{\text{var}\ \{\hat{f}(0)\}}/\text{E}\hat{f}(0)$$

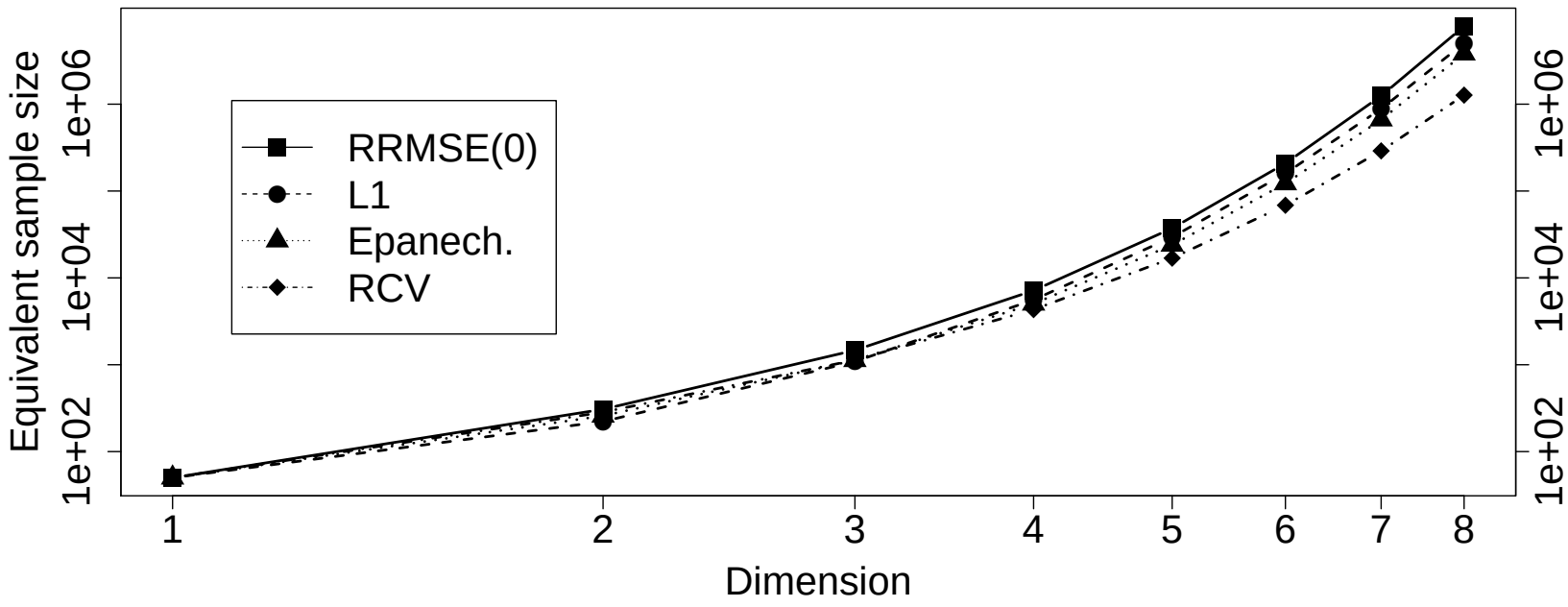


Figure: Equivalent sample sizes for several criteria that have the same value as in one dimension with 50 sample points. The density and kernel are both $N(0, I_d)$. The criterion values for RRMSE(0), AMIAE, Epanechnikov, and RCV(0) are 0.145, 0.218, 0.176, and 0.127, respectively.

Multivariate Pointwise L_1 Kernel Error

- Peter Hall showed some clever L_1 approximations.
- Scott and Wand applied to pointwise estimation in $\Re^d$. They showed the optimal L_1 and L_2 bandwidths were always close:

$$0.9635 \leq \frac{h_1^*(\mathbf{x})}{h_2^*(\mathbf{x})} \leq 1$$

- Remarkably, this ratio of asymptotically optimal bandwidths does not depend on the particular choice of kernel, the underlying density function, or the point of estimation.
- The global L_1 and L_2 bandwidths will be more different, since place different weights on different regions in space.

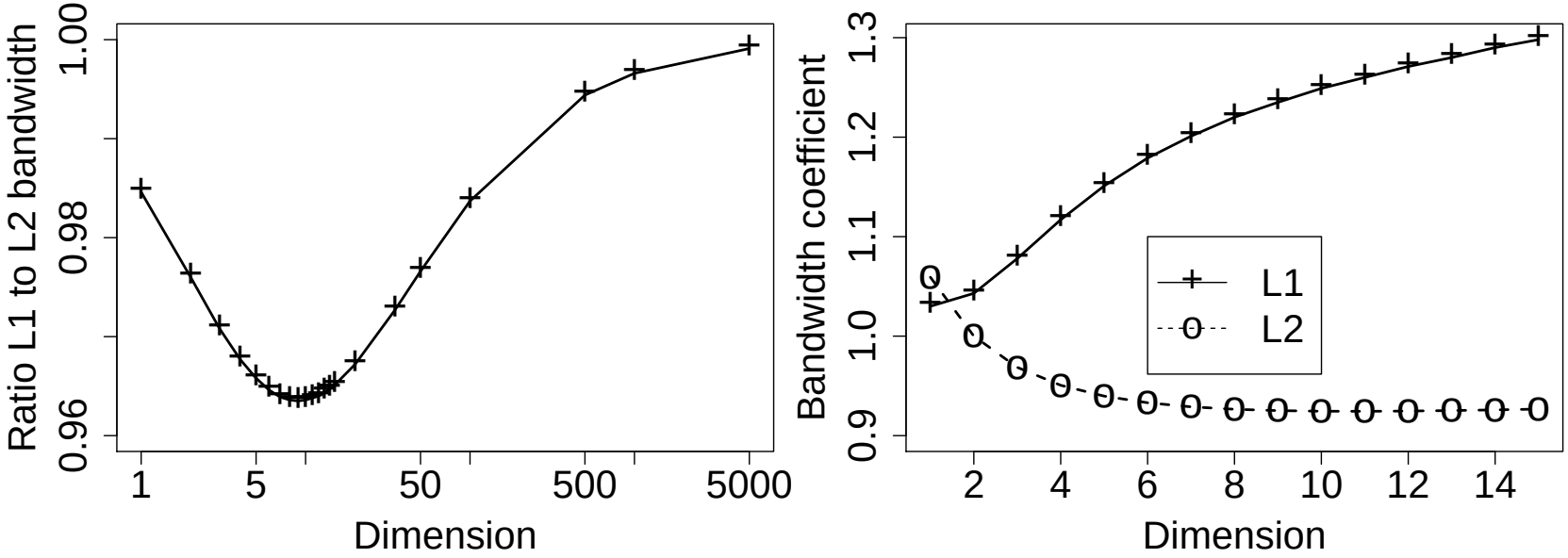


Figure: Ratio of pointwise optimal L_1 and L_2 bandwidths for all situations. The right frame displays the coefficients of $n^{-1/(d+4)}$ for the global L_1 and L_2 bandwidths for normal data.

Examples and Discussion

- Are these theoretical arguments too pessimistic?
- We look at some empirical evidence, following an example of Friedman, Stuetzle, and Schroeder in their projection pursuit density estimation paper.
- They considered a small sample in $\Re^{10}$ with $n = 225$
- The (x_1, x_2)-space contains the signal, which is an equal mixture of 3 shifted bivariate $N(0, I_2)$ densities with means $(\mu, 0)$, $(-\mu, 3)$, and $(-\mu, -3)$ with $\mu = 3^{3/2}/2$. The marginal variances are both 7.
- Next, eight pure normal noise dimensions are added, each with variance 7. The authors investigated the task of estimating the following slice of $f(\mathbf{x})$:

$$f(x_1, x_2, 0, 0, 0, 0, 0, 0, 0, 0)$$

- Will the trimodal structure be apparent in the slice?

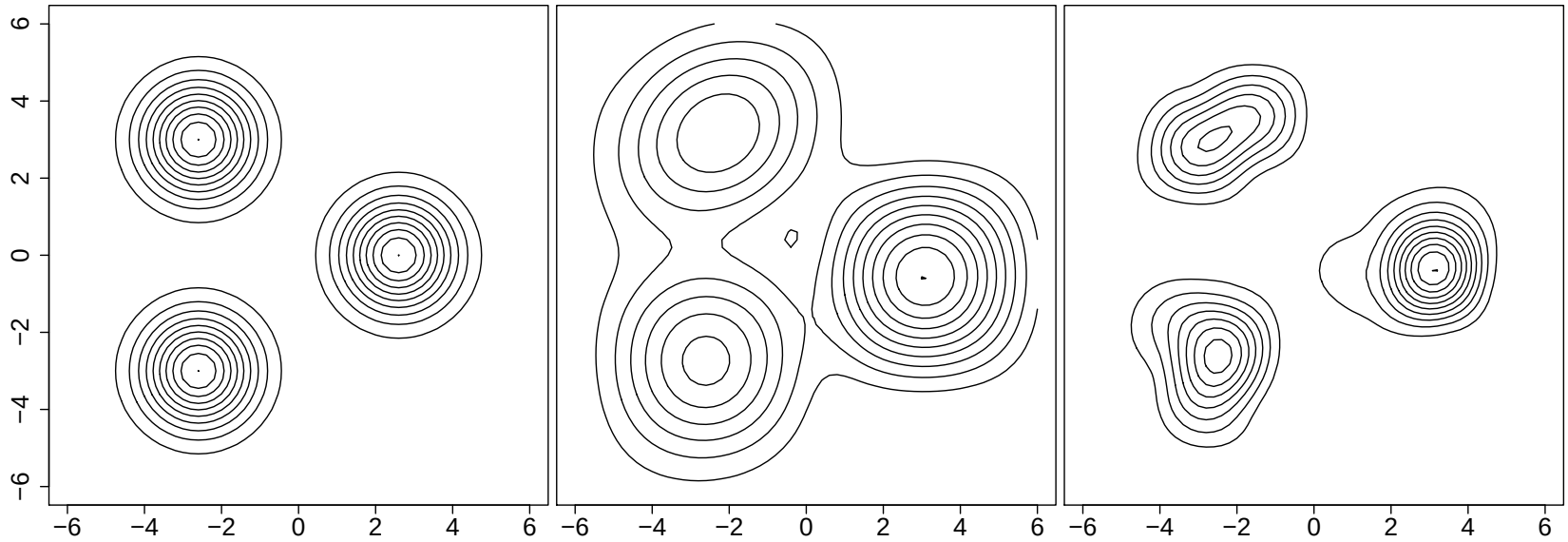


Figure: Bivariate slice of true density and slices of 2 10-D triweight product kernel estimates. The 9 contour levels are equally spaced up to the mode. The bandwidths for the middle frame were $h_i = 4.0$; for the right frame $h_1 = h_2 = 2.0$ and $h_3 = \cdots = h_{10} = 5.25$

- Not quite as impressive as it appears, since the estimates are biased down quite a bit. Nevertheless, the *structure* is quite visible.

Example PRIM4 Raw and Transformed

- Consider the PRIM4 data set ($n = 500$), one of several sets originating from the Stanford Linear Accelerator Center during development of the original PRIM system (related data sets include PRIM7 and PRIM9).
- Each of the raw variables is very strongly skewed
- By successive application of Tukey's transformation ladder to each variable, the skewness can be reduced.
- The 4 transformations were

$$\log_{10}(x_1),\ \sqrt{\log_{10}(1+x_2)},\ -\sqrt{\log_{10}(1-x_3)},\ -\sqrt{\log_{10}(1-x_4)}$$

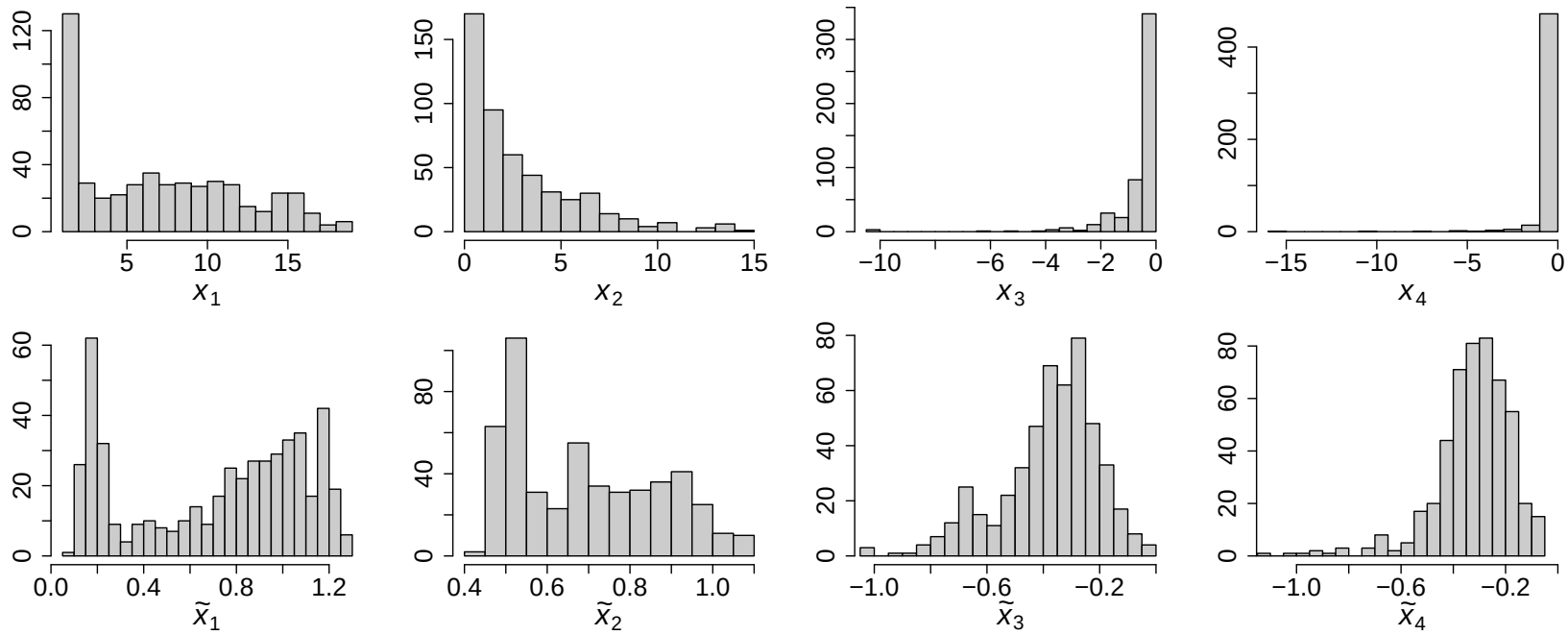


Figure: Histograms of the original 4 variables in PRIM4 before (top row) and after (bottom row) transformation; $n = 500$. Note some the histograms are multimodal after the transformation.

Effect of Transformation on ASH

- Since the raw data are so skewed, much of the data in $\Re^4$ is on or near an edge.
- After transformation, the data occupy are more central position.
- The raw data are very close to being dimension deficient, so harder to estimate well
- We compare the same slice $\hat{f}(x_1, x_2, x_3|x_4)$ of the ASH estimator before and after transformation.

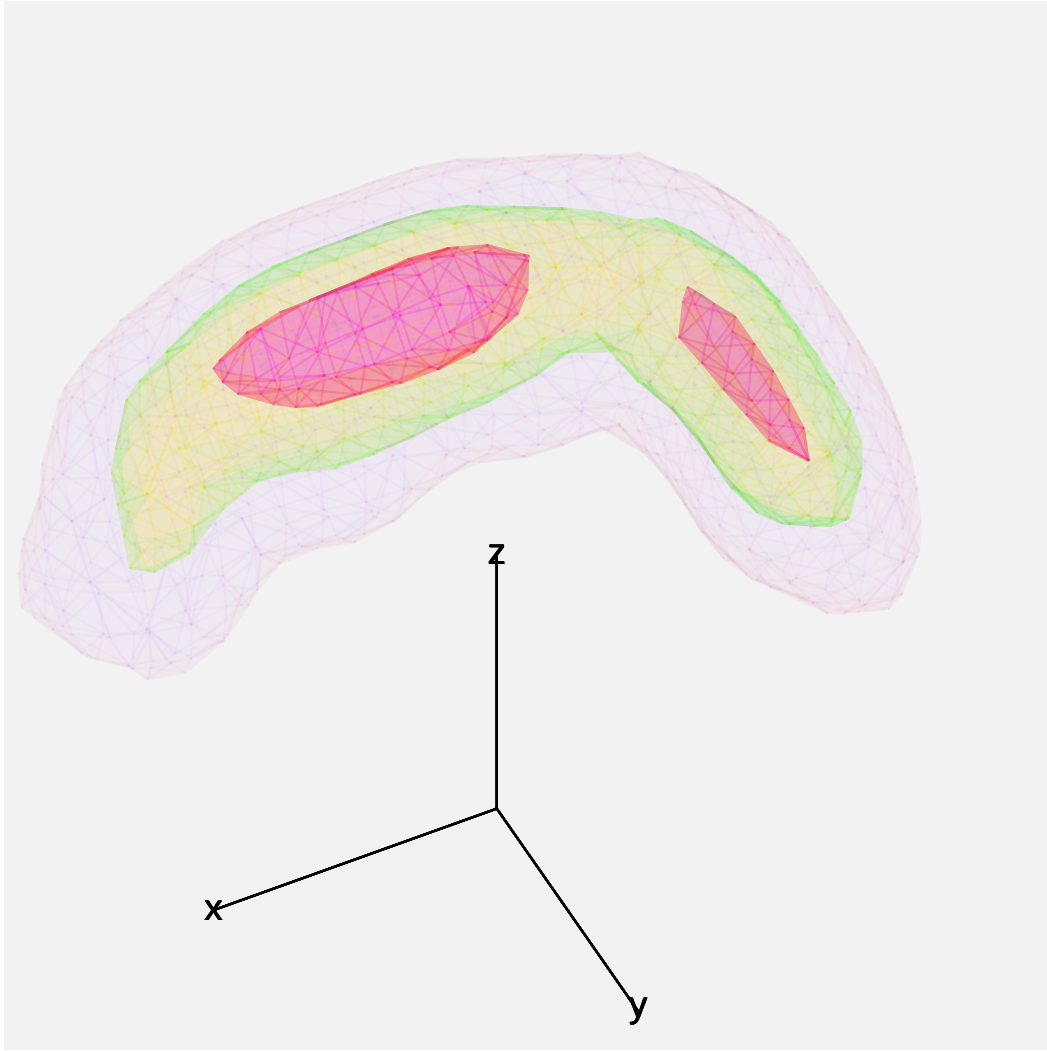


Figure: Three contour shells ($\alpha = 10\%$, 30%, and 60%) of a slice of the averaged shifted histogram of the four-dimensional PRIM4 data set with 500 points. These variables are heavily skewed, and the resulting density estimation problem more difficult.

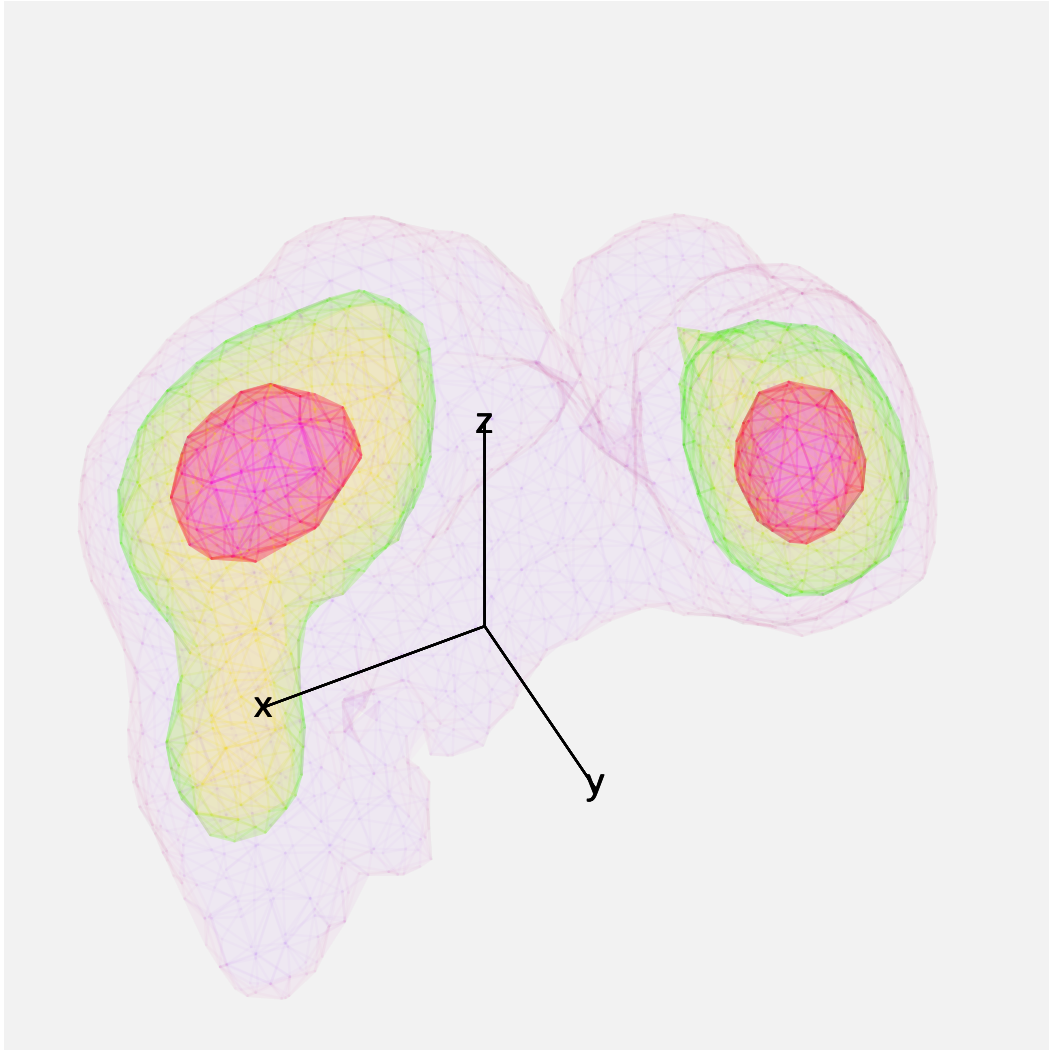


Figure: Three contour shell levels based on an ASH of the transformed PRIM4 data. The transformation was chosen to reduce skewness in each marginal variable. Such marginal transformations can greatly improve the quality of density estimation in multiple dimensions.

Example: Raw and Transformed Mount St. Helens Data

- The earthquake data (x, y, z) preceding the 1982 eruption of Mount St. Helens
- The epicenters of 512 earthquakes in a 6-week period prior were located
- No reason to transform the longitude or latitude variables
- However, the depth variable hides some interesting clusters
- Used the transform $-\log_{10}(-z)$.

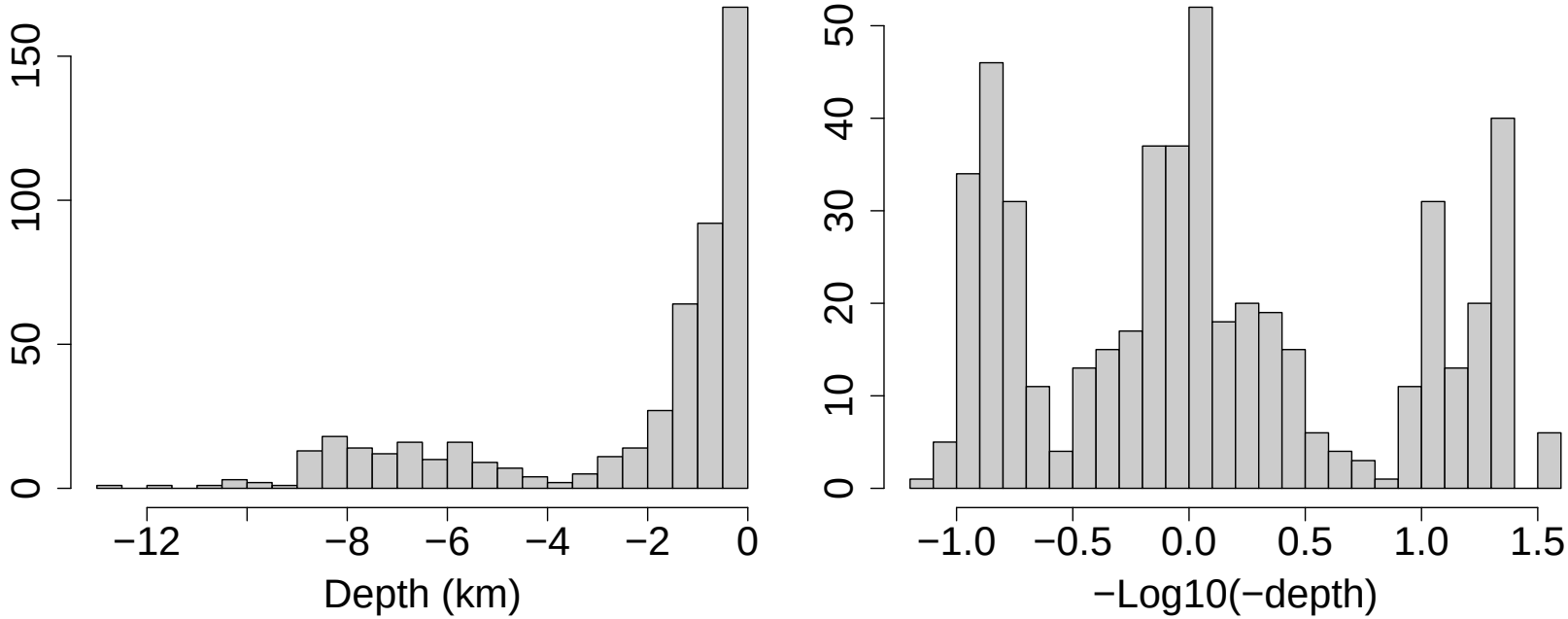


Figure: Histograms of the depths of 510 earthquake epicenters and the transformed depths.

Some Trivariate ASH Contours

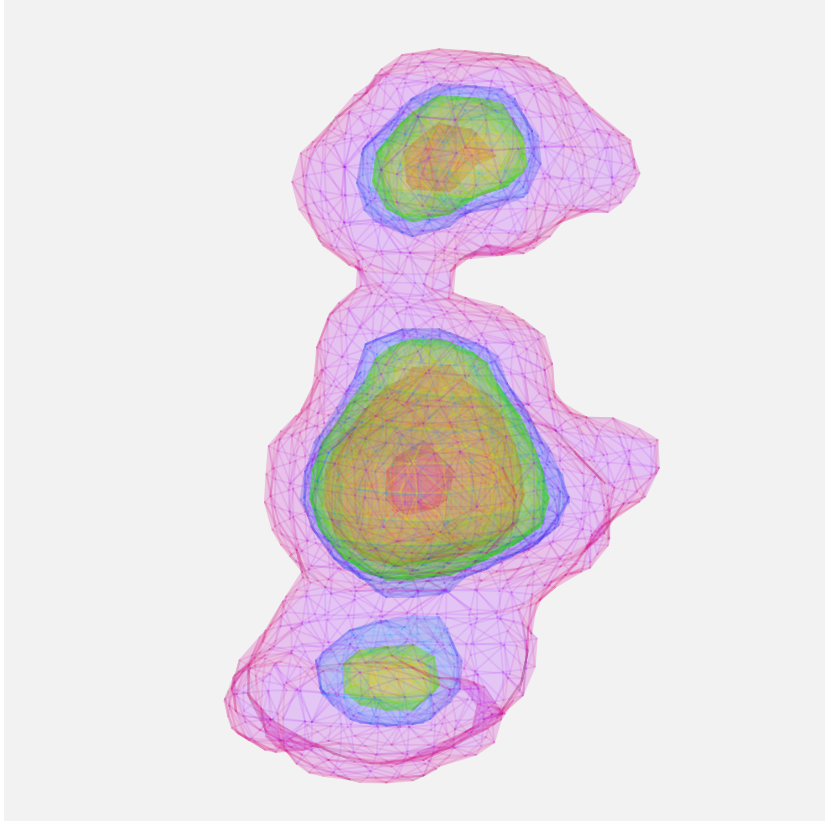


Figure: ASH of the location of 510 earthquake epicenters and the transformed depths.

Some 4-D ASH Contours Sliced on Time Before Eruption

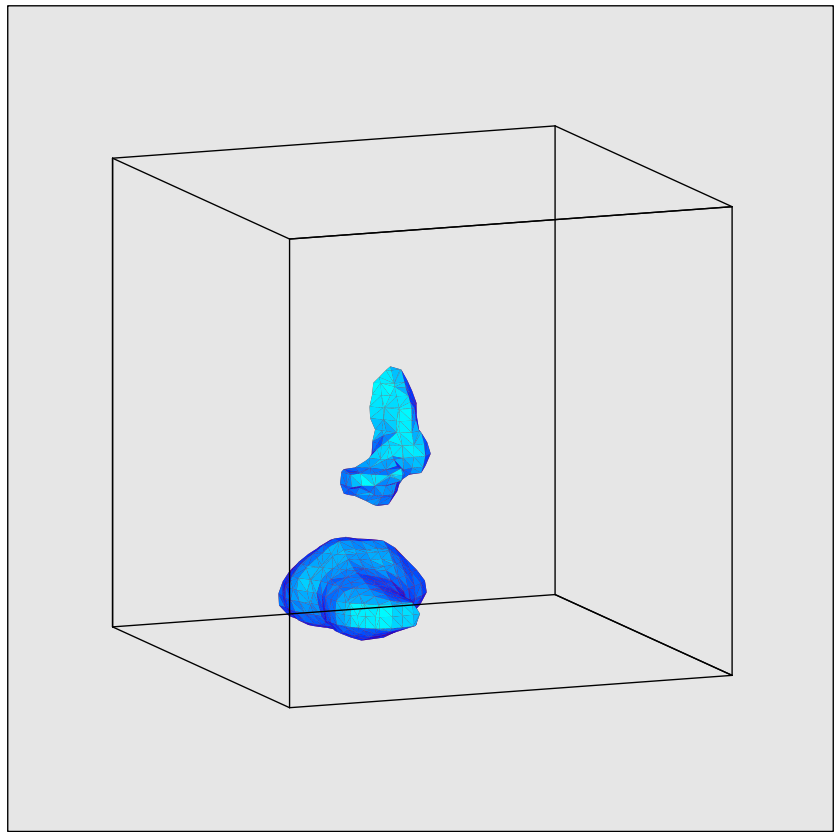


Figure: (Top left) ASH of the location of 510 earthquake epicenters and the transformed depths. The next three frames show the space-time ASH at approximately one week intervals leading up to the eruption (all at the same α level.)

Some 4-D ASH Contours Sliced on Time Before Eruption

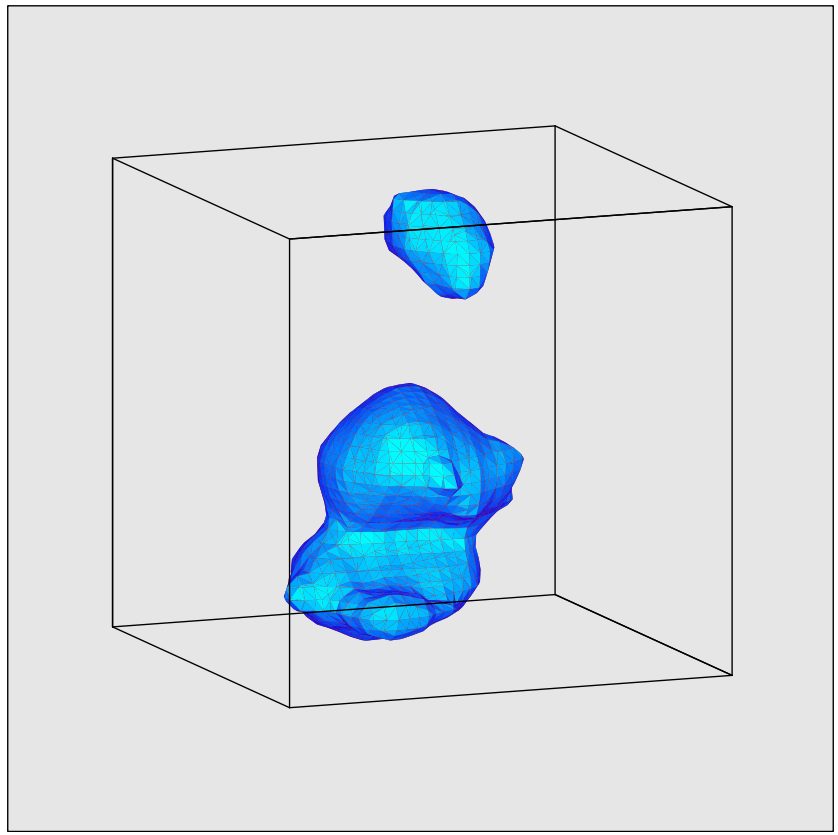


Figure: (Top left) ASH of the location of 510 earthquake epicenters and the transformed depths. The next three frames show the space-time ASH at approximately one week intervals leading up to the eruption (all at the same α level.)

Some 4-D ASH Contours Sliced on Time Before Eruption

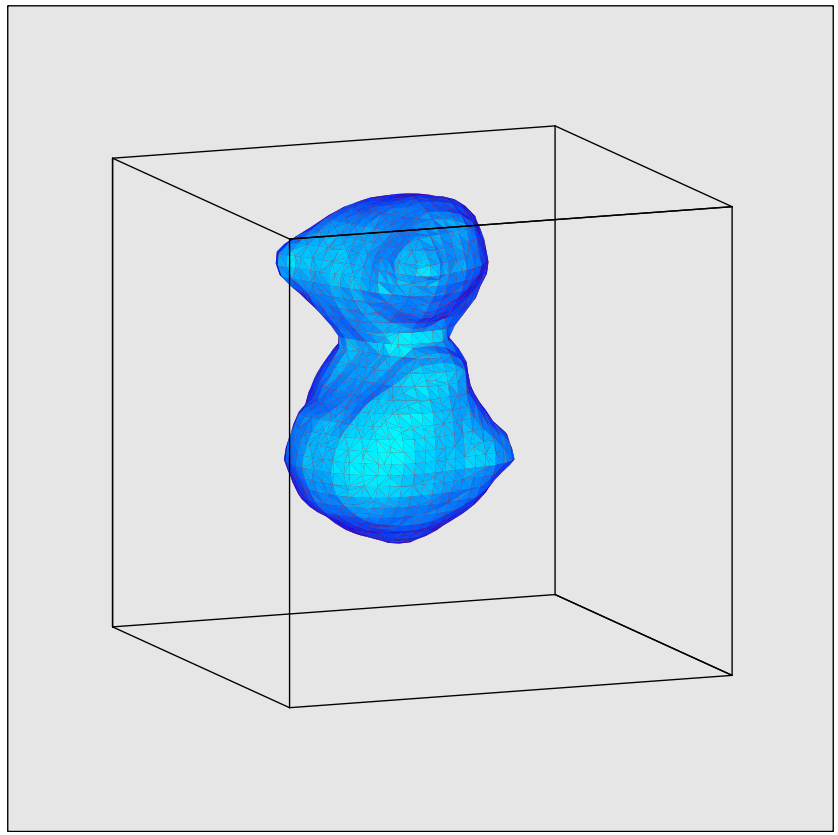


Figure: (Top left) ASH of the location of 510 earthquake epicenters and the transformed depths. The next three frames show the space-time ASH at approximately one week intervals leading up to the eruption (all at the same α level.)

Projection Algorithms

- Principal Components is the most widely used: project onto the first k eigenvectors
- However, its optimization criterion is *variance*, which may or may not capture *structure*
- We use PCA to initially remove any zero (or small) variance dimensions, perhaps retaining 90% of the variance
- Then do something fancy with information theory
- Friedman and Tukey tried a projection-pursuit criterion that showed peaks (grand tour example)
- Huber proposed negative Shannon entropy
- Jee and Scott tried Fisher information as the criterion

Simple Example from $\Re^2 \rightarrow \Re^1$

- With a simple bivariate mixture of 3 normals, only Fisher information prefered a direction (135 degrees) that reveals the trimodality.
- The other criteria superimpose 2 of the 3 components.

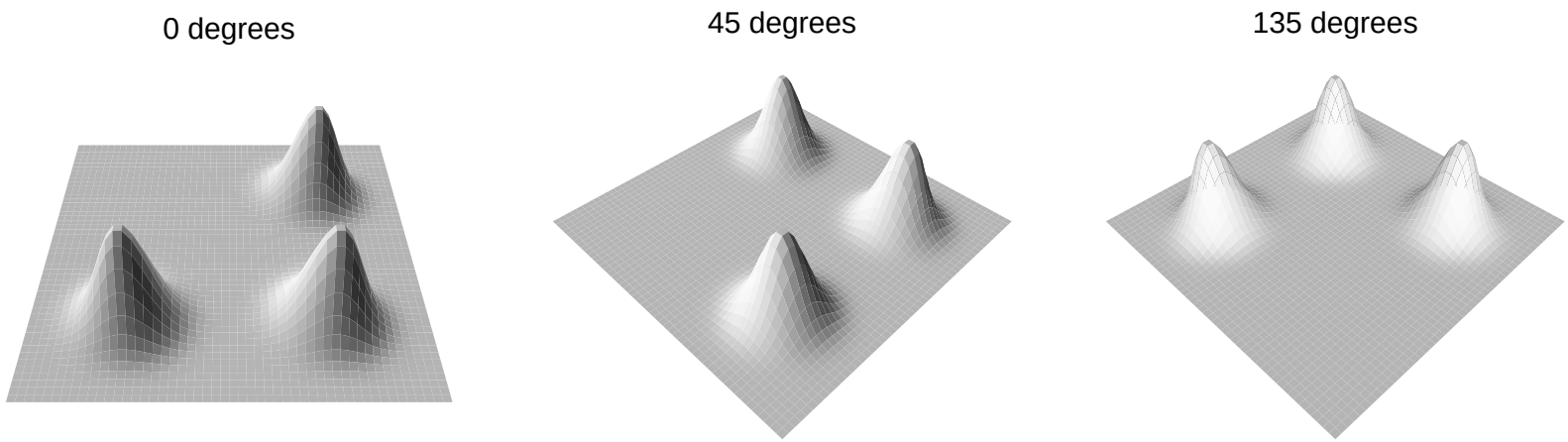


Figure: Three possible projection angles ($\theta = 0°, 45°, 135°$) for a bivariate mixture of 3 normal densities.

Fisher Information Applied to PRIM7 Data

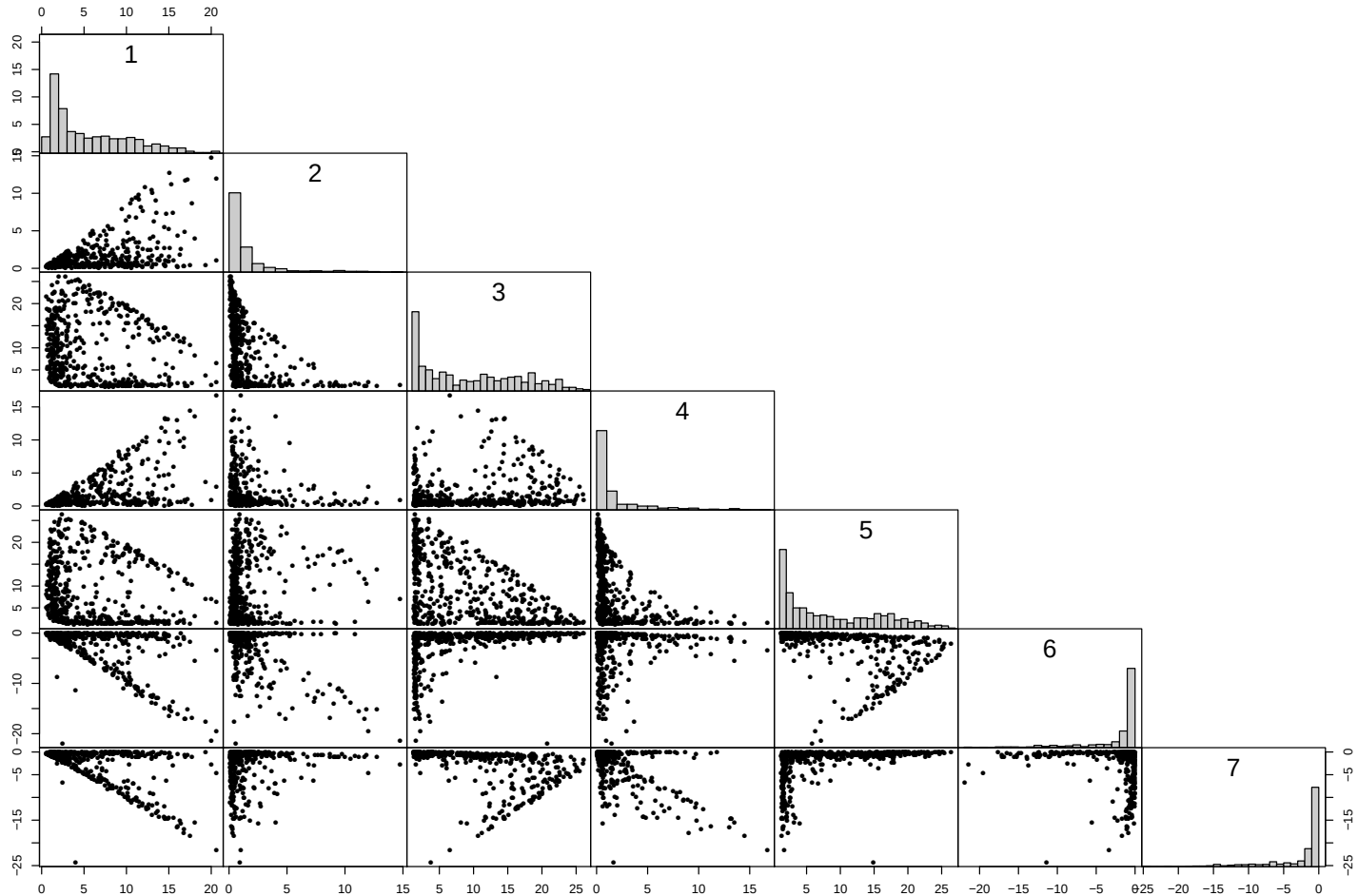
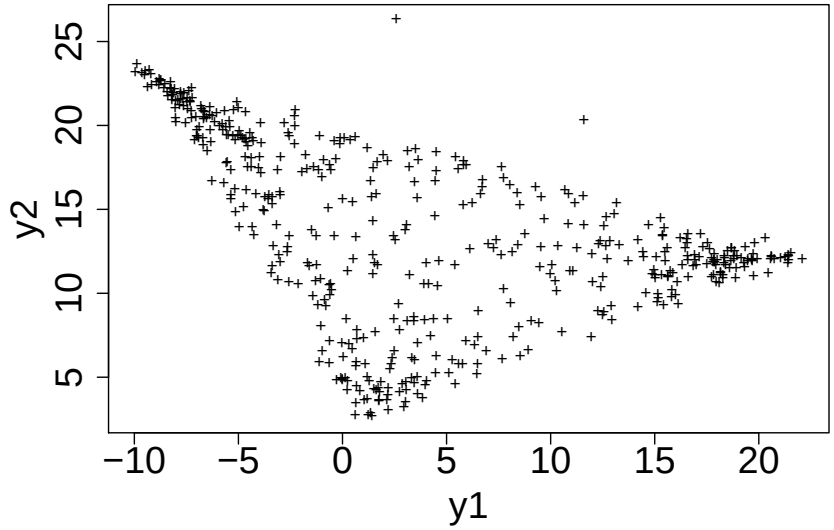
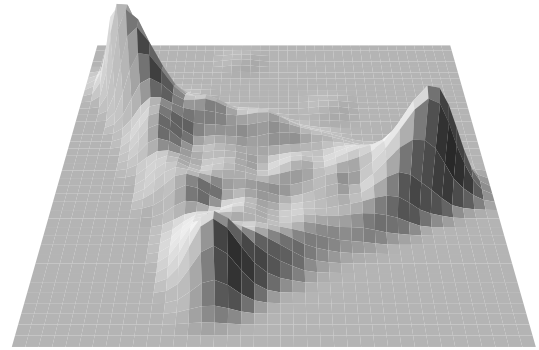
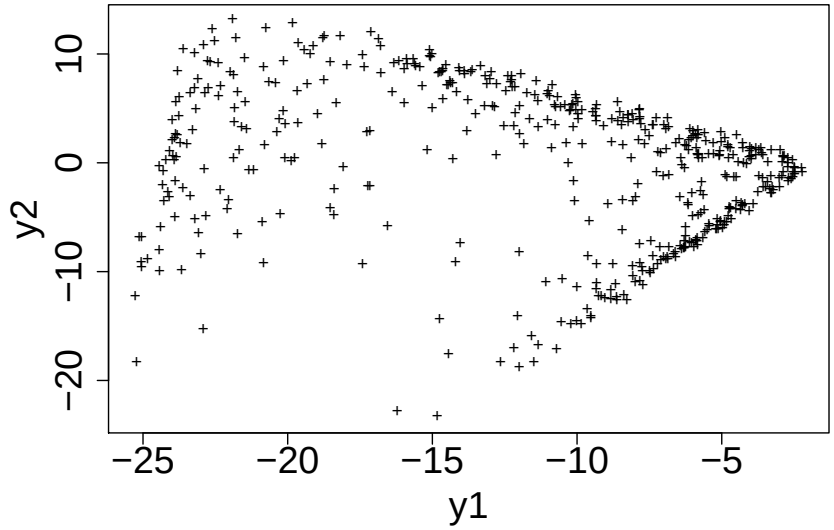


Figure: Pairs of the 7 variables in the PRIM7 data.







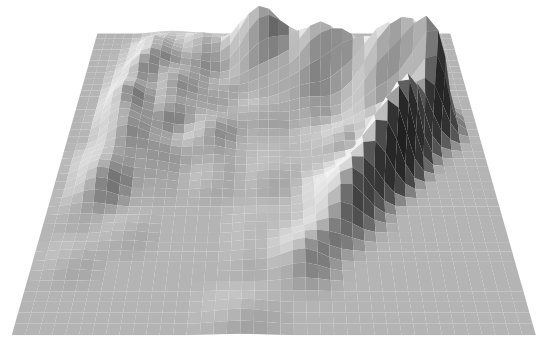


Figure: Bivariate projections of the PRIM7 data using the global and best local maxima of Fisher information. The ASH estimates for each are shown beside.

LANDSAT IV Example: Model-Based Nonlinear Projection

- The raw data are 6 4-dimensional observations at 23 day intervals
- These were non-linearly transformed into 3 dimensions by a specialist at NASA Houston (knowledge domain expert)
- These data model the "greenness" of a pixel over the growing season

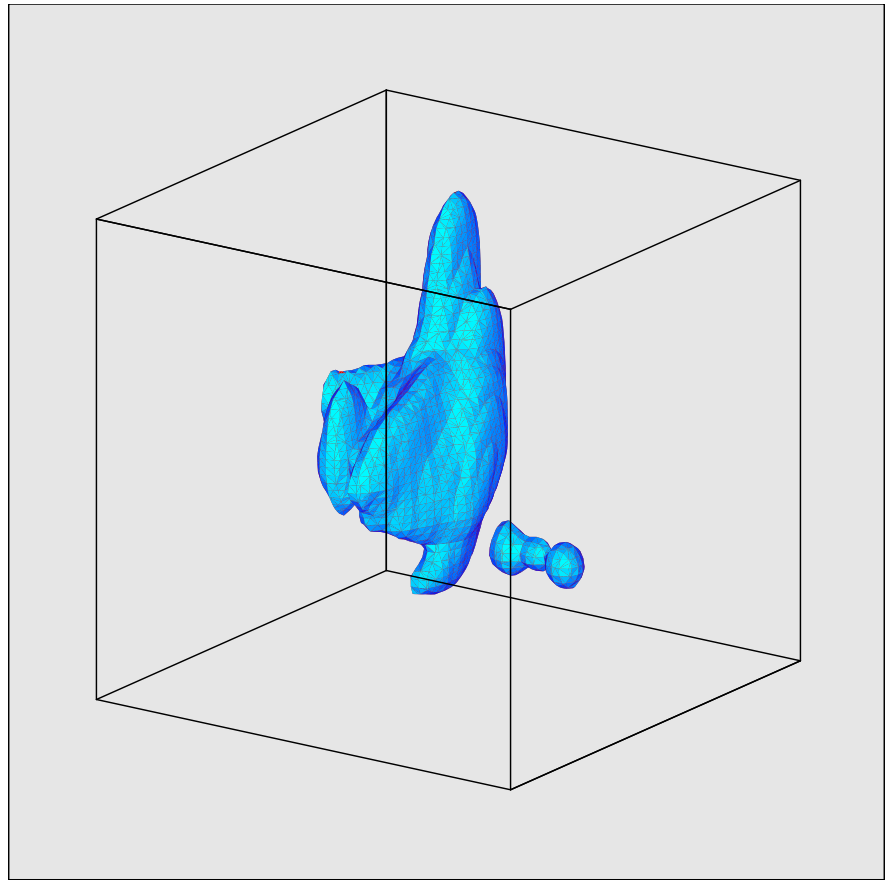


Figure: The $\alpha = 1\%$ level contour of a trivariate averaged shifted histogram of the Landsat data set of 22,932 points. The small disjoint second shell in the bottom right corner represents some of the outliers in the data set.

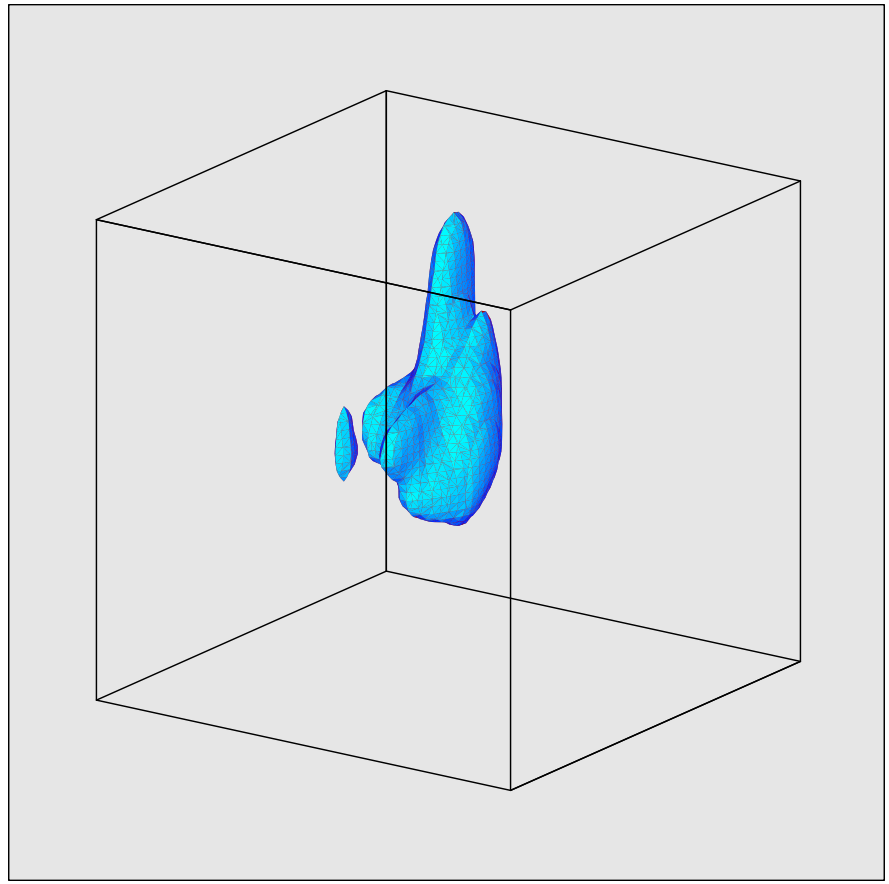


Figure: The $\alpha = 3.5\%$ level contour of a trivariate averaged shifted histogram of the Landsat data set of 22,932 points. The outliers resulted from singularities in the model-based data transformation algorithm from the original 24-dimensional Landsat data and were recorded at the minimum or maximum values.

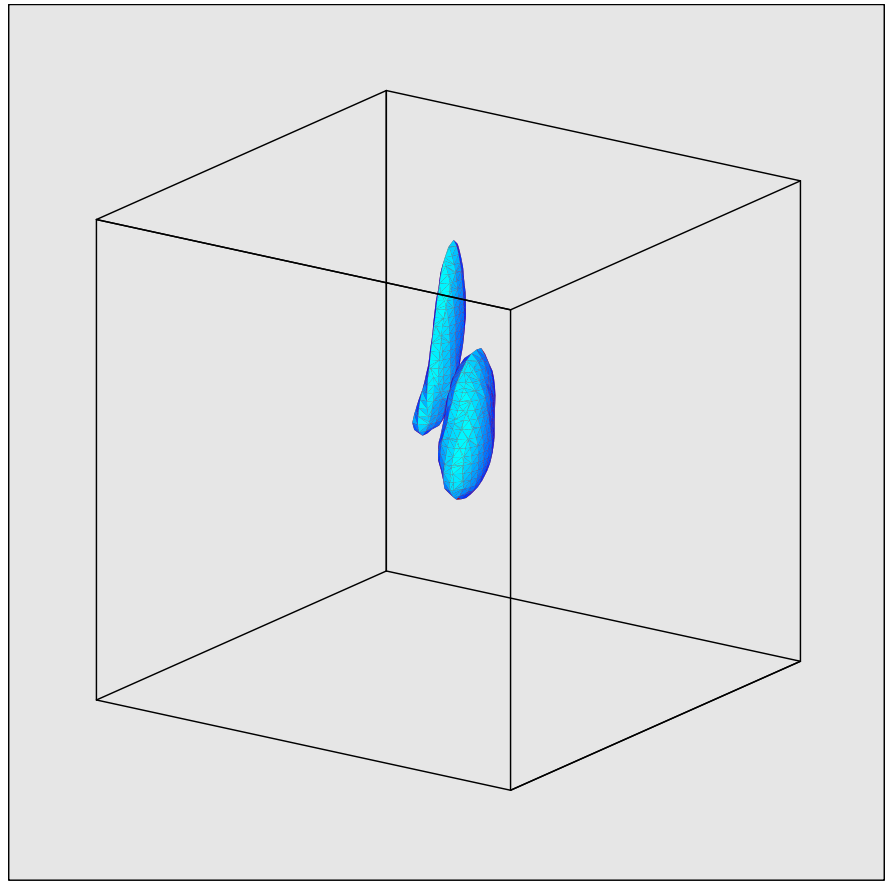


Figure: The $\alpha = 15\%$ level contours of a trivariate averaged shifted histogram of the Landsat data set of 22,932 points.

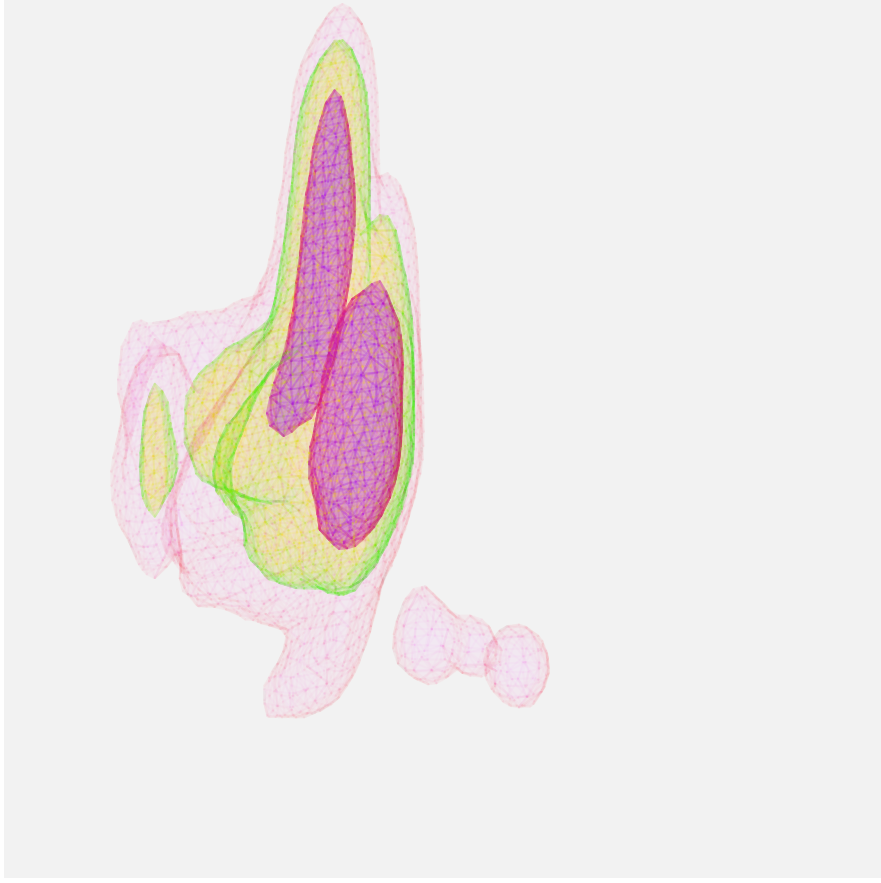


Figure: A composite of the first three frames using transparency. An examination of the crops being grown in this region reveals that the the tall cluster in the middle represents sunflowers; and the largest cluster on the right represents small grains including wheat, and the small small cluster on the far left represents sugar beets.

Using the Crop Labels (3 Crops)

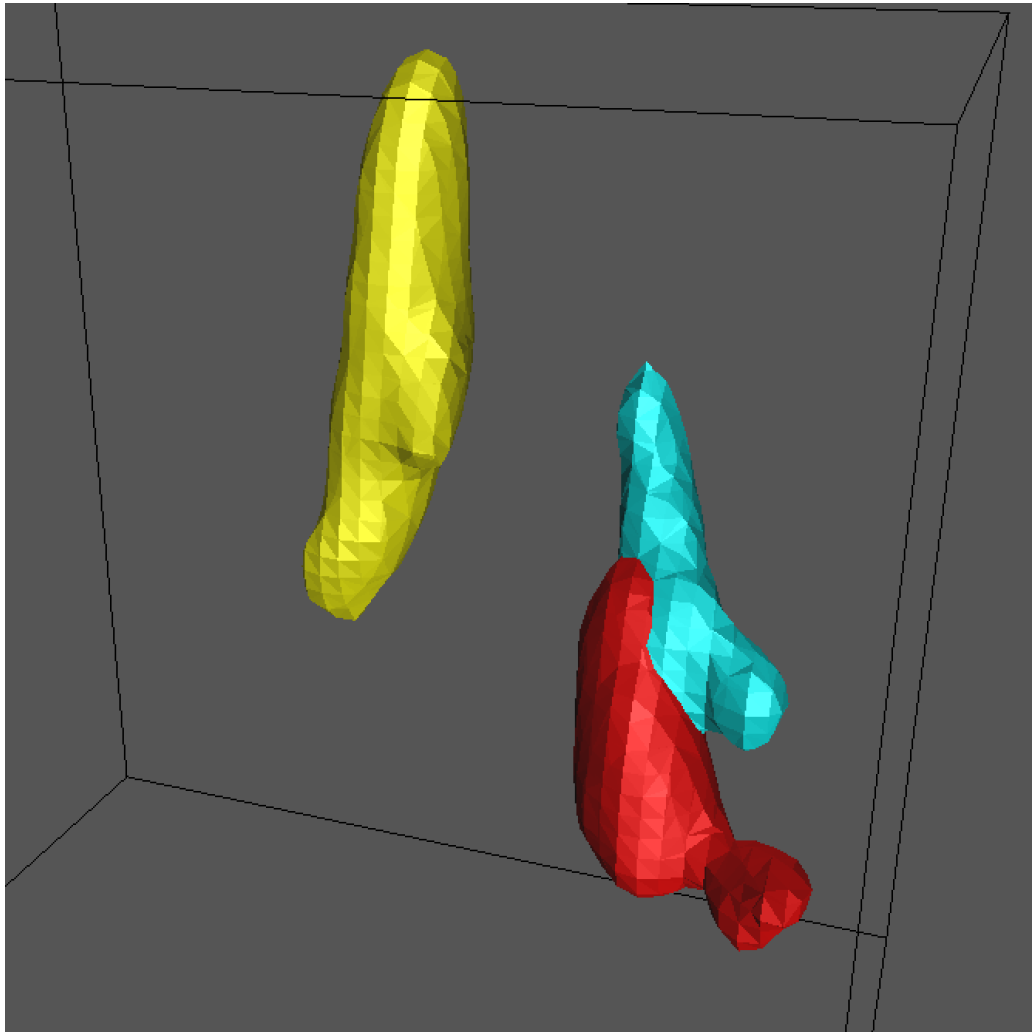


Figure: Median trivariate contours for sunflower ($n = 3694$, yellow), spring wheat ($n = 3811$, red), and barley ($n = 892$, cyan). The median contour contains 50% of the labelled data.

Using the Crop Labels (More Crops)

Figure: Median contours as shown before with spring oats ($n = 459$, white), peanuts ($n = 304$, purple), soybeans ($n = 731$, magneta), and sugar beets ($n = 506$, green).

Nonparametric Regression and Additive Models

- We use the multivariate product kernel estimator to compute the conditional mean, rather than assume a parametric form.
- The theoretical regression function is defined to be

$$r(x) = E(Y|X = x) = \int y\, f(y|x)dy = \frac{\int yf(x,y)dy}{\int f(x,y)dy}$$

- In the 2-D case plug in the product kernel for $f(x,y)$ in the above

$$\begin{aligned}\hat{f}(x,y) &= \frac{1}{nh_xh_y}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right) K\left(\frac{y-y_i}{h_y}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x-x_i)\, K_{h_y}(y-y_i)\end{aligned}$$

Nadarya-Watson Nonparametric Regression

- The denominator is the marginal density function, which here becomes

$$\int \hat{f}(x,y)\,dy = \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x-x_i)\int K_{h_y}(y-y_i)\,dy$$
$$= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x-x_i) \quad \text{(the usual 1-D KDE)}$$

- Since $\int y\,K_h(y-y_i)\,dy = y_i$, the numerator becomes

$$\int y\,\hat{f}(x,y)\,dy = \frac{1}{n}\sum_{i=1}^{n} y_i\,K_{h_x}(x-x_i)$$

Nadarya-Watson Nonparametric Regression

- Therefore,

$$\hat{r}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i \, K_{h_x}(x - x_i)}{\frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x - x_i)} = \sum_{i=1}^{n} w_{h_x}(x, x_i) \, y_i$$

- The weights (which sum to 1) are given by

$$w_{h_x}(x, x_i) = \frac{K_{h_x}(x - x_i)}{\sum_{j=1}^{n} K_{h_x}(x - x_j)}$$

Local Polynomial Fits

- By piecing together *local polynomial fits*, one can obtain a nonparametric regression estimate
- Simpliest is a local **constant** fit

$$\bar{y} = \arg\min_a \sum_{i=1}^{n} (a - y_i)^2 \qquad \textit{but not local}$$

- Use the kernel factor to focus on a neighborhood of x:

$$\hat{r}(x) = \arg\min_a \sum_{i=1}^{n} \left[K_h(x - x_i) \times (a - y_i)^2 \right]$$

- Easy to show the minimizer is

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} y_i \, K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)}$$

which is precisely the Nadaraya-Watson again!

Example

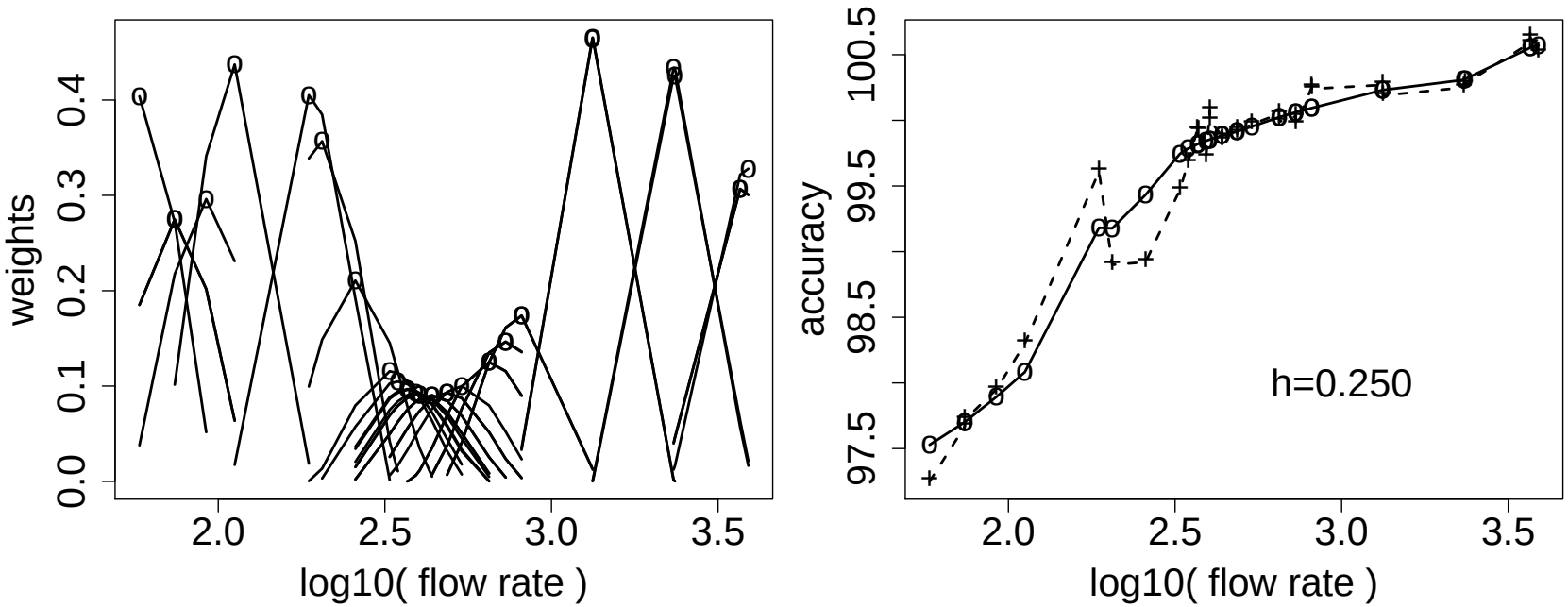


Figure: Biweight Nadaraya-Watson kernel weights and estimate for the gas flow dataset.

Modal Regression

- In introductory statistics classes, we often mention alternatives to the sample mean —- the median and the mode
- Rather than using the KDE to compute the conditional mean or the conditional median (more robust?), we propose to compute the conditional **mode(s)**

$$\text{Modal regression curve : } \quad \hat{r}(x) = \arg(s) \max_y \hat{f}(y|x),$$

- Examples: old faithful data and the penny thickness data

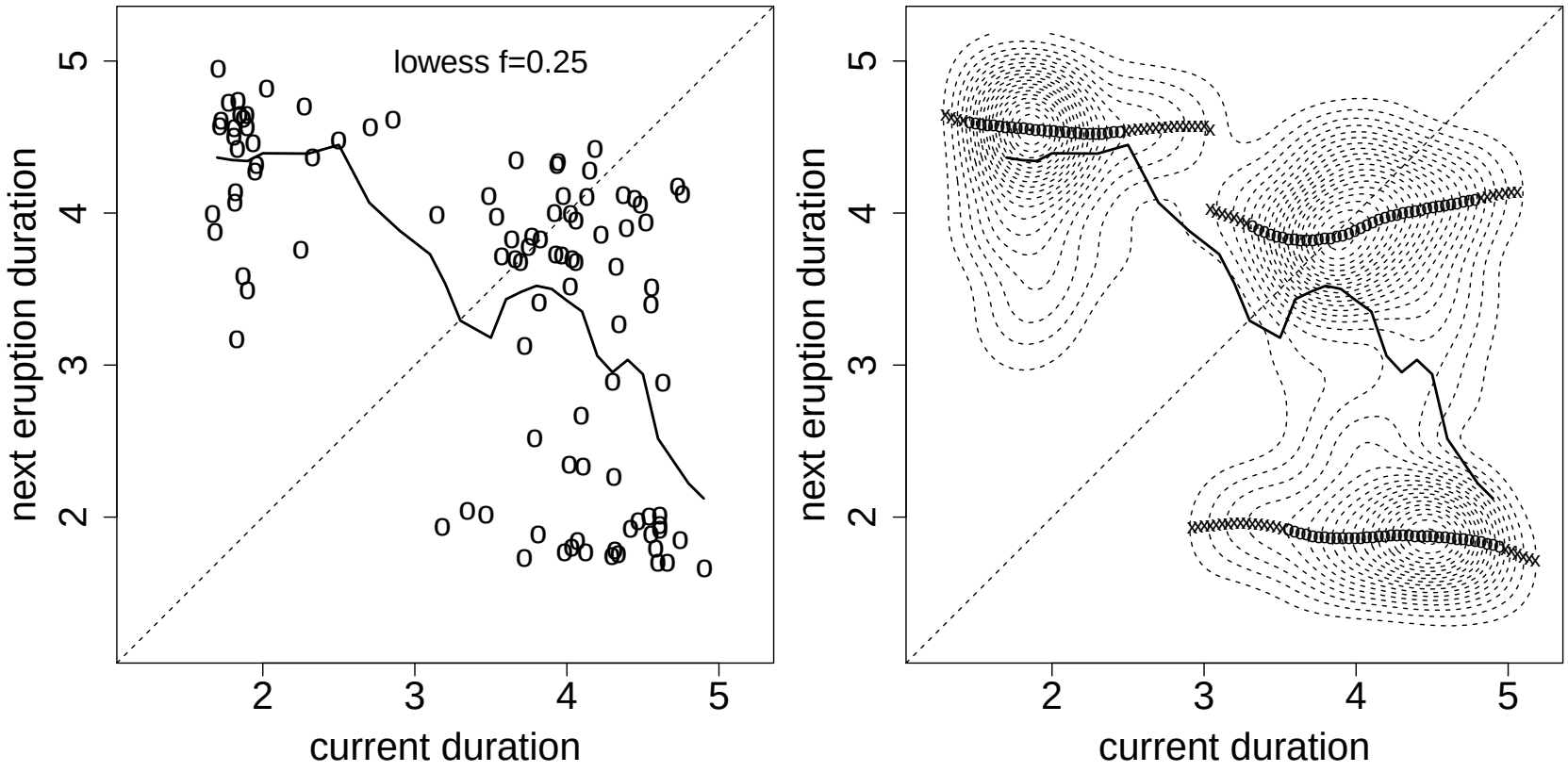


Figure: Conditional mean (LOWESS) and conditional mode smoothers of the lagged Old Faithful duration dataset. The conditional mode is displayed with the symbol "o" when above the 25% contour and with an "x" between 5% and 25%. The 45° line is also shown.

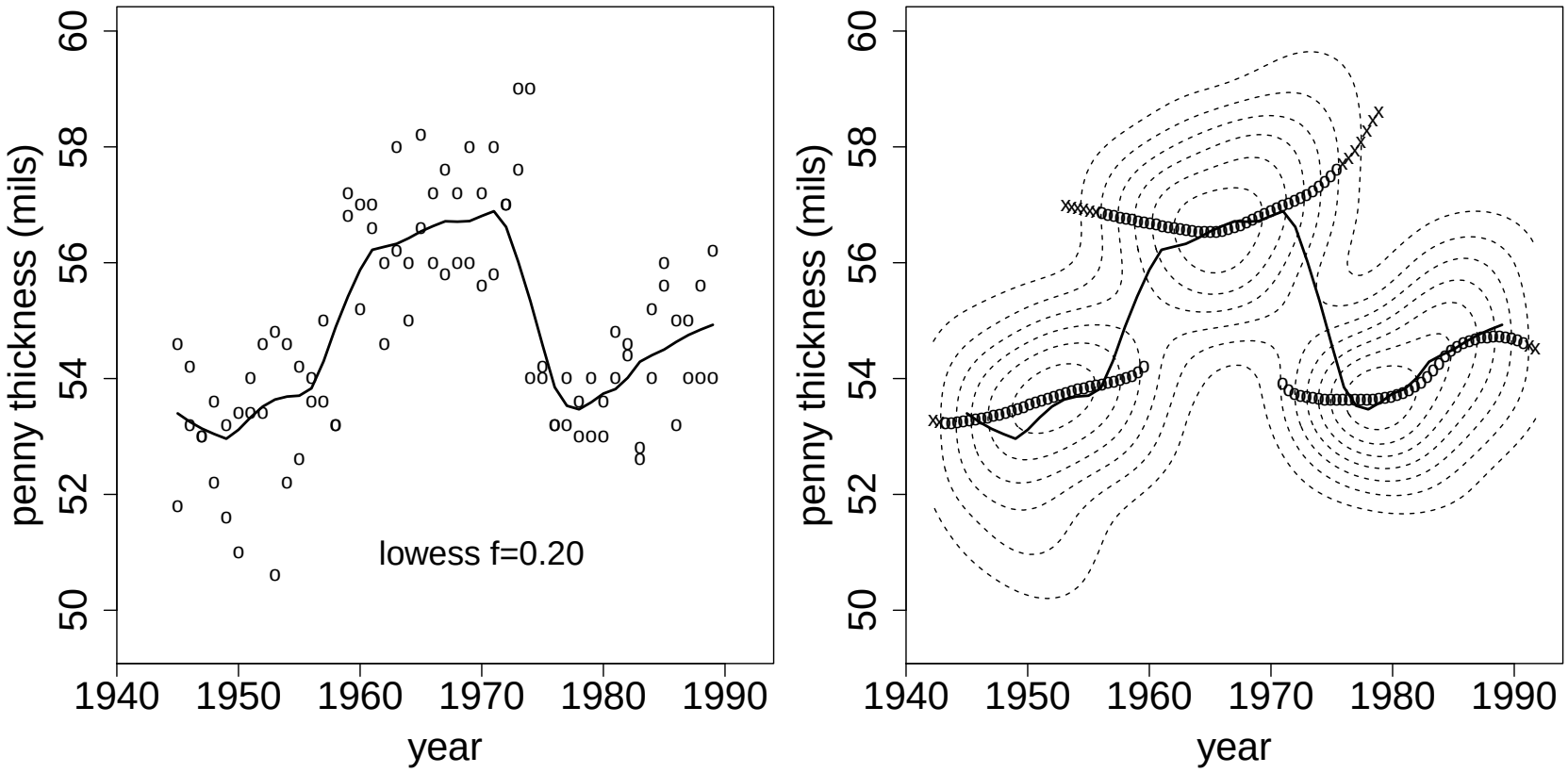


Figure: Conditional mean and conditional mode smoothers of the U.S. penny thickness dataset.

Local L_1 Regression: Wang and Scott

- Recall L_1 regression is inherently robust
- Examples: motorcycle data and Cauchy noise

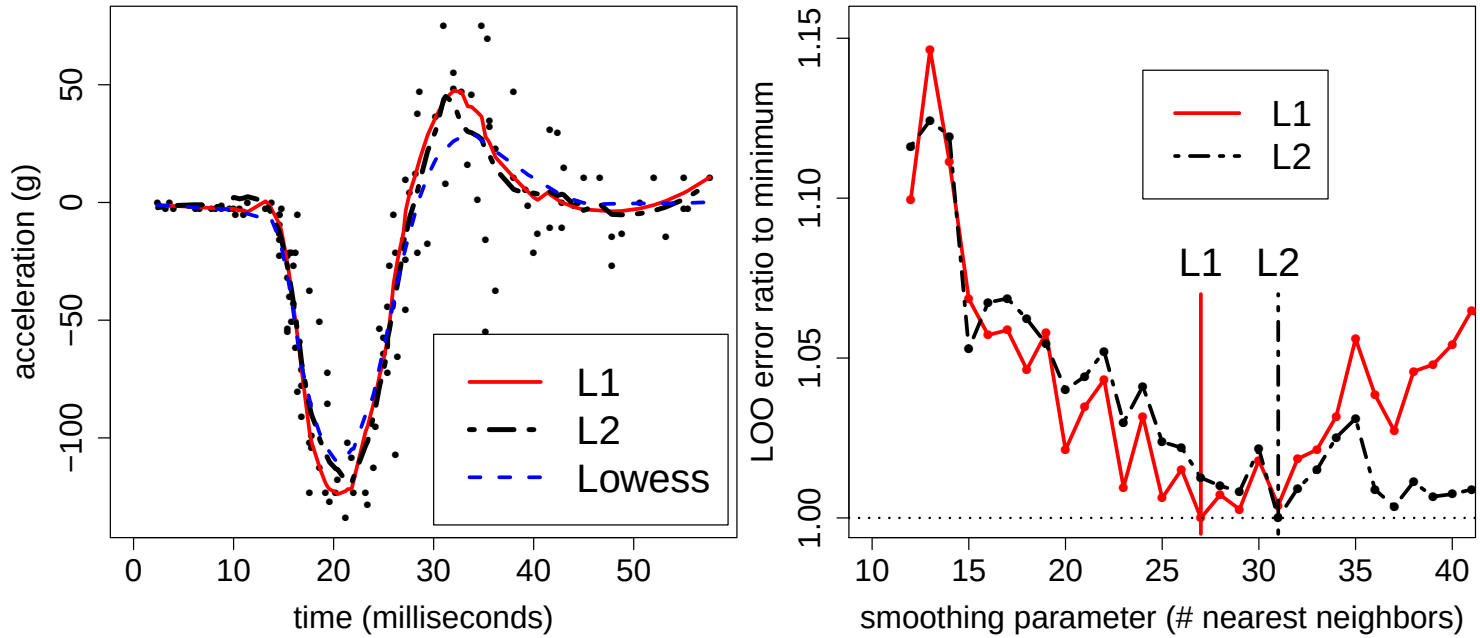


Figure: (Left) Local L_1 and L_2 quadratic fits to the motorcycle data, together with a LOWESS fit with $f = 0.25$. (Right) Normalized Leave-One-Out (LOO) cross-validation criteria, namely, the mean absolute error and the standard deviation for the L_1 and L_2 fits, respectively. The raw values range from $(16.2, 18.6)$ and $(18.5, 20.8)$, respectively. The best L_1 fit occurred with 27 points in each local neighborhood.

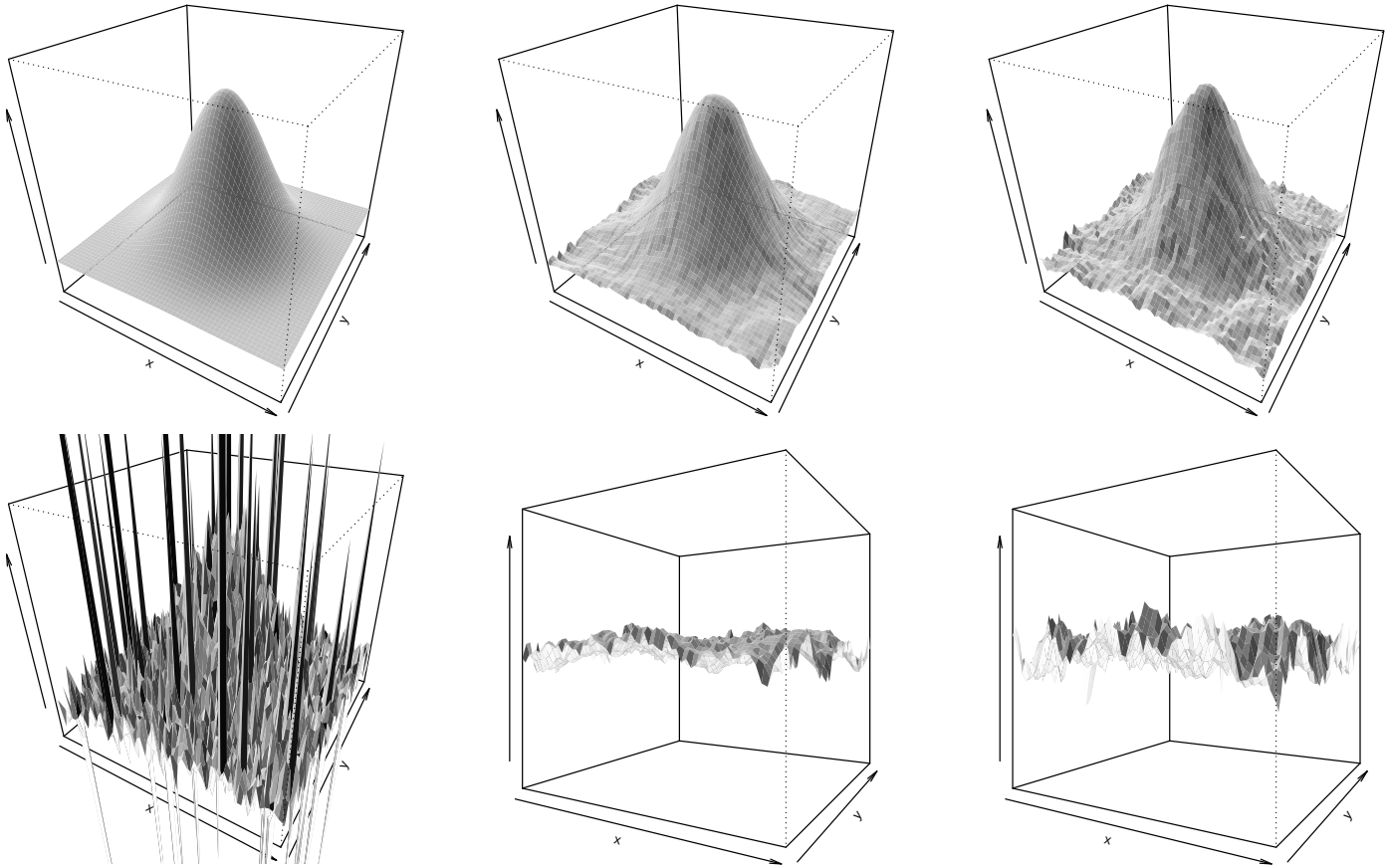


Figure: (Top row) Regression surface that is a $2\pi\times$bivariate normal on $[-3,3]\times[-3,3]$ on a 61×61 mesh; L_1 local full quadratic nonparametric estimate with $\hat{m}_x = 10$ and $\hat{m}_y = 8$; and L_2 local full quadratic nonparametric estimate with same smoothing parameters. (Bottom row) Surface contaminated with noise from the mixture density $.975\, N(0, 0.1^2)$ $+\ 0.025\, N(0, 2.5^2)$, and residual surfaces for estimates above. The residual plots are centered at $z = 0$ and the vertical scale is expanded by a factor of 3.

Additive Modeling: see book for details

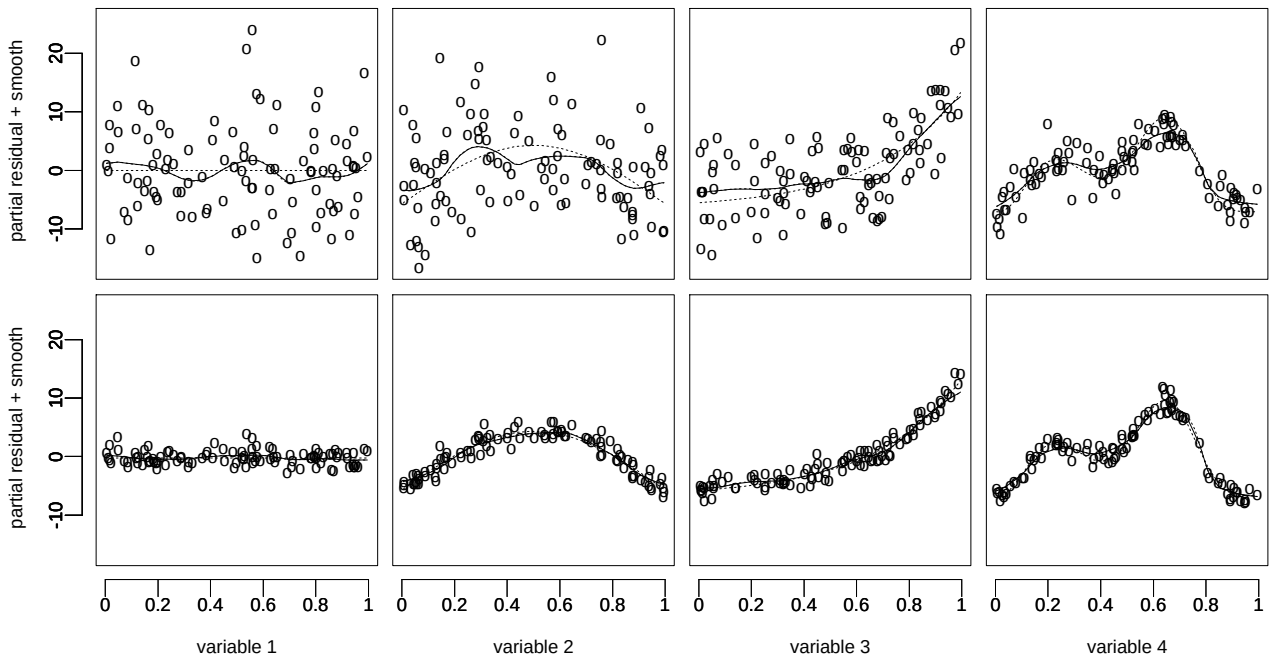


Figure: Additive model iteration for simulated data from (**??**). The true additive function is shown as a dotted line and the estimated additive function as a solid line. The top row gives the initial loop and the bottom row the final iteration (6 loops).

Sliced Inverse Regression (SIR)

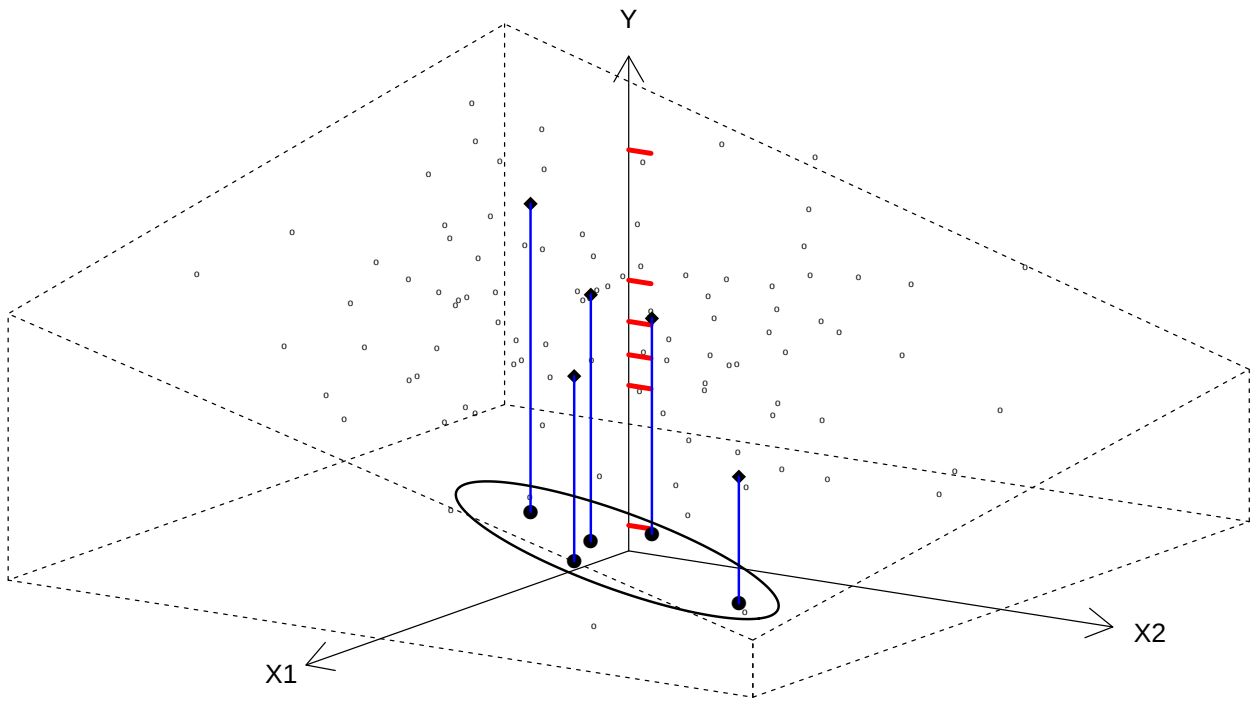


Figure: Example of the SIR dimension reduction technique.