Nonparametric Function Estimation
Stat 550[1]

David W Scott[2]

Rice University

August 24
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/

Introduction

- It is a pleasure to stand before you and share this material. I hope this will be a highly interactive format, and that you will feel free to raise your hand with questions of clarification.
- Classical parametric statistics has evolved to handle nonparametric forms.
- Now we have entered the realm of big data, data sciences, massive datasets, data mining, machine learning, and now deep learning.
- Hal Varian, the chief economist at Google, opined:

  The sexy job in the next 10 years will be statisticians. People think I am joking, but who would have guessed that computer engineers would have been the sexy job of the 1990s?

- The material in this course is always the beginning of the new statistics.

Introduction (continue)

- Our general approach is to build on the notion that all statistical techniques are straightforward, if the underlying density function is known — we shall estimate the unknown multivariate pdf *nonparametrically*
- Our computer power increases the demand for tool that can detect and summarize the structure in multivariate data
- NPDE is recognized as a useful tool in 1-D and 2-D; we will demonstrate that is a powerful tool in high dimensions, with particular emphasis on $\Re^3$ and $\Re^4$
- We will introduce the major ideas via the classical histogram, the most widely applied and intuitive nonparametric density estimator — then develop links between the histogram and more statistically efficient method

Introduction (concluded)

- The nonparametric world is more complex that its parametric counterpart — we have selected material that is representative of the broad spectrum of theoretical results available, with an eye on the potential user
- Visualization is a key aspect of effective multivariate nonparametric analysis — most analyses and results can be presented graphically (cf Tukey's "exploratory data analysis")
- Background assumed? I'll try to keep it light.
  - Mathematics: Taylor's series, approximations, optimization
  - Statistics: basic moments, variance, mean squared error, Binomial, normal
  - Advanced Statistics: will introduce clustering, classification

Chapter I: Representation and Geometry Data in $\Re^d$

Key ideas:

- parametric analysis is most powerful (if model correct)
- nonparametric analysis is most flexible
- graphical analysis for discovering the unexpected
- options for graphical tools for visualizing structure in multidimensional data:
  - depicting the data points themselves
  - displaying functions estimated from those points

Introduction

- classical linear multivariate statistical methods reply primarily upon analysis of the covariance matrix — effective for datasets with hundreds of variables
- "unparametric" methods may be loosely collected under the heading of exploratory data analysis — a graph may provide a more concise representation than a parametric model, because hundreds of parameters may be involved
- nonparametric approaches are intermediate — but calibration is hard since nonparametric estimates must be optimized separately for each application
- on the other hand, we take the point of view that no nonparametric estimate is wrong, just different aspects of the solution are emphasized
- The "curse of optimality" might suggest that this is an illogical point of view?

Historical Perspective

- Sir Francis Galton discovered the correlation coefficient empirically, which had a strong influence on Karl Pearson
- Pearson is best known for his goodness-of-fit tests, frequency curves, biometry, but Pearson was also a strong proponent of the geometrical representation of data
- From his lectures in November, 1891, at a job interview at Gresham College in London, we read from his syllabus this cryptic note:
  - Erroneous opinion that Geometry is only a means of popular representation: *it is a fundamental method of investigating and analysing statistical material*. (his italics)
  - He also coined the words "histogram", "chartograms", "sterograms", "stigmograms", ...

Fisher's point of view

- But Fisher held a different point of view; see his comments on diagrams in *Statistical Methods for Research Workers* (1932):

  > *The preliminary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them.*

- Fisher and Pearson had long-standing feuds, which can be traced to fundamental disagreements about parametric versus nonparametric modeling and the underlying assumptions.
- An emphasis on optimization and the efficiency of statistical procedures has been a hallmark of mathematical statistics ever since.

Graphical Display of Multivariate Data Points

- one problem is coping with massive data sets
- becomes exponentially more difficult as the dimensionality of the data increases, a phenomenon known as the *curse of dimensionality*
- the goal of statistical data analysis is to extract the maximum information from the data, and to present a product that is as accurate and as useful as possible
- the data may, for example, be strongly non-normal, fall onto a nonlinear subspace, exhibit multiple modes, or be asymmetric.

Multivariate Scatter Diagrams

- pairwise scatter diagrams — multivariate structure?

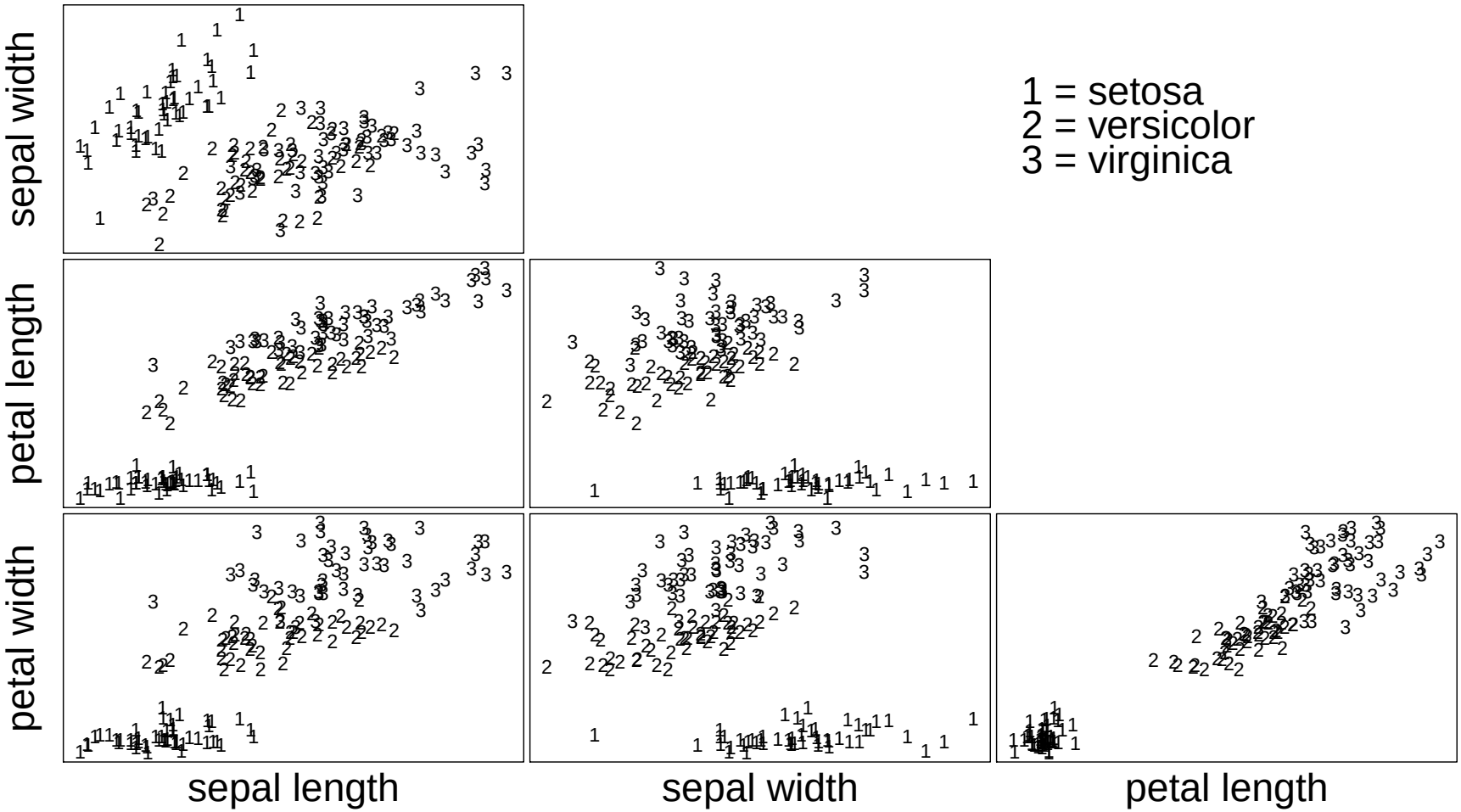
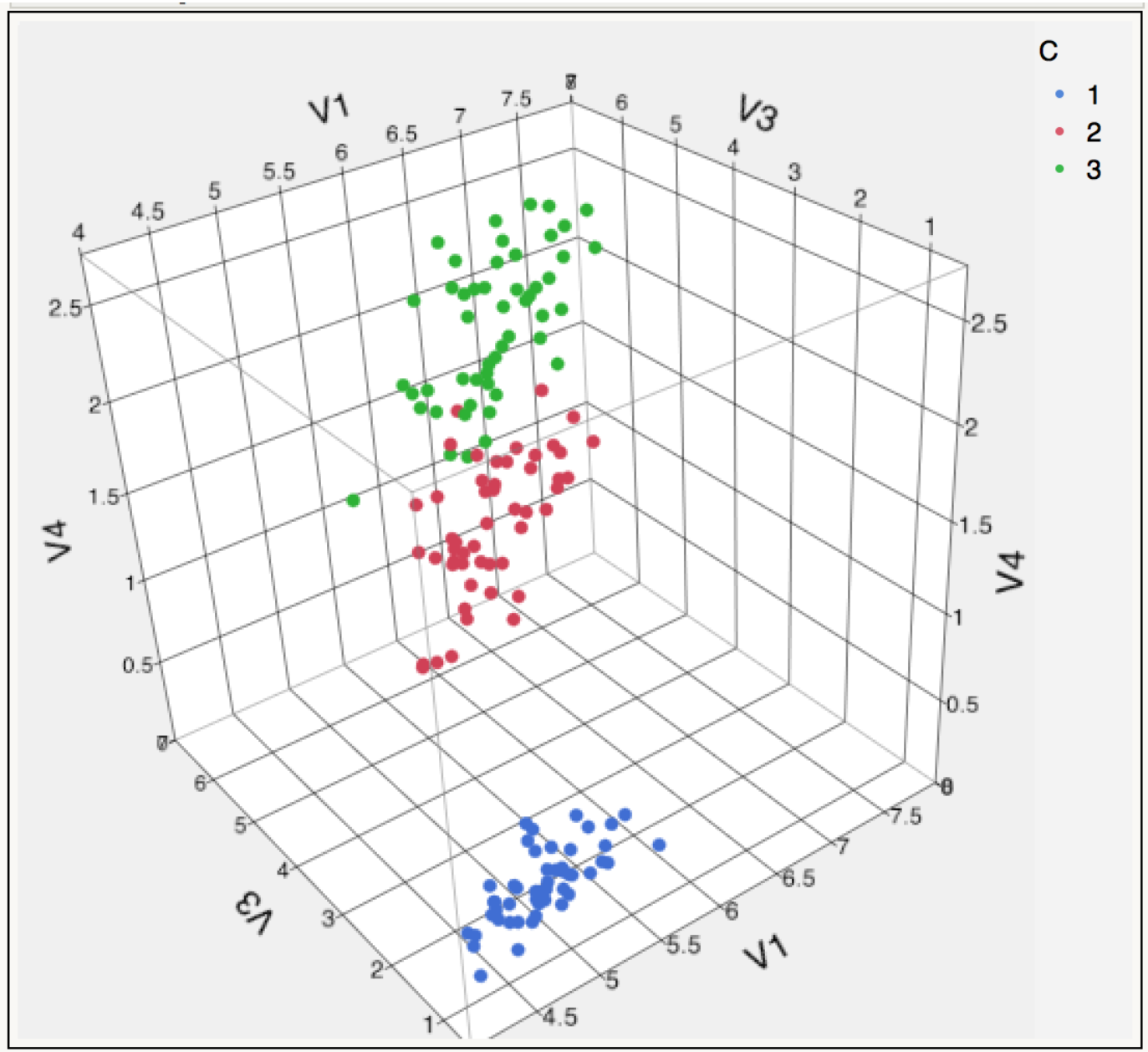
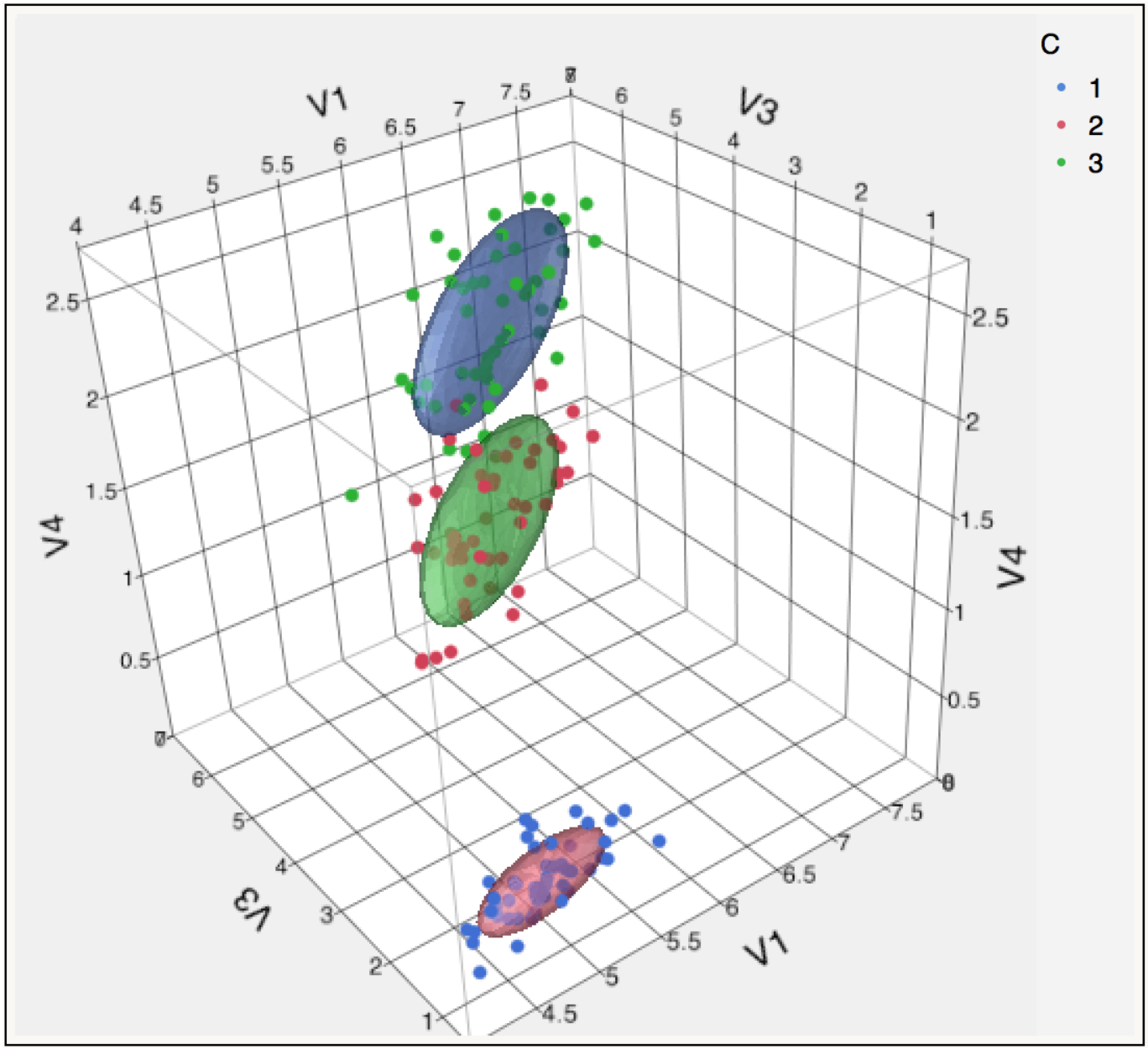


Figure: Pairwise scatter diagrams of the *Iris* data with the 3 species labeled.

JMP Visualizations of Iris Data



JMP Visualizations with Normal Fits



Heart Disease Example: Substantial overlap of classes

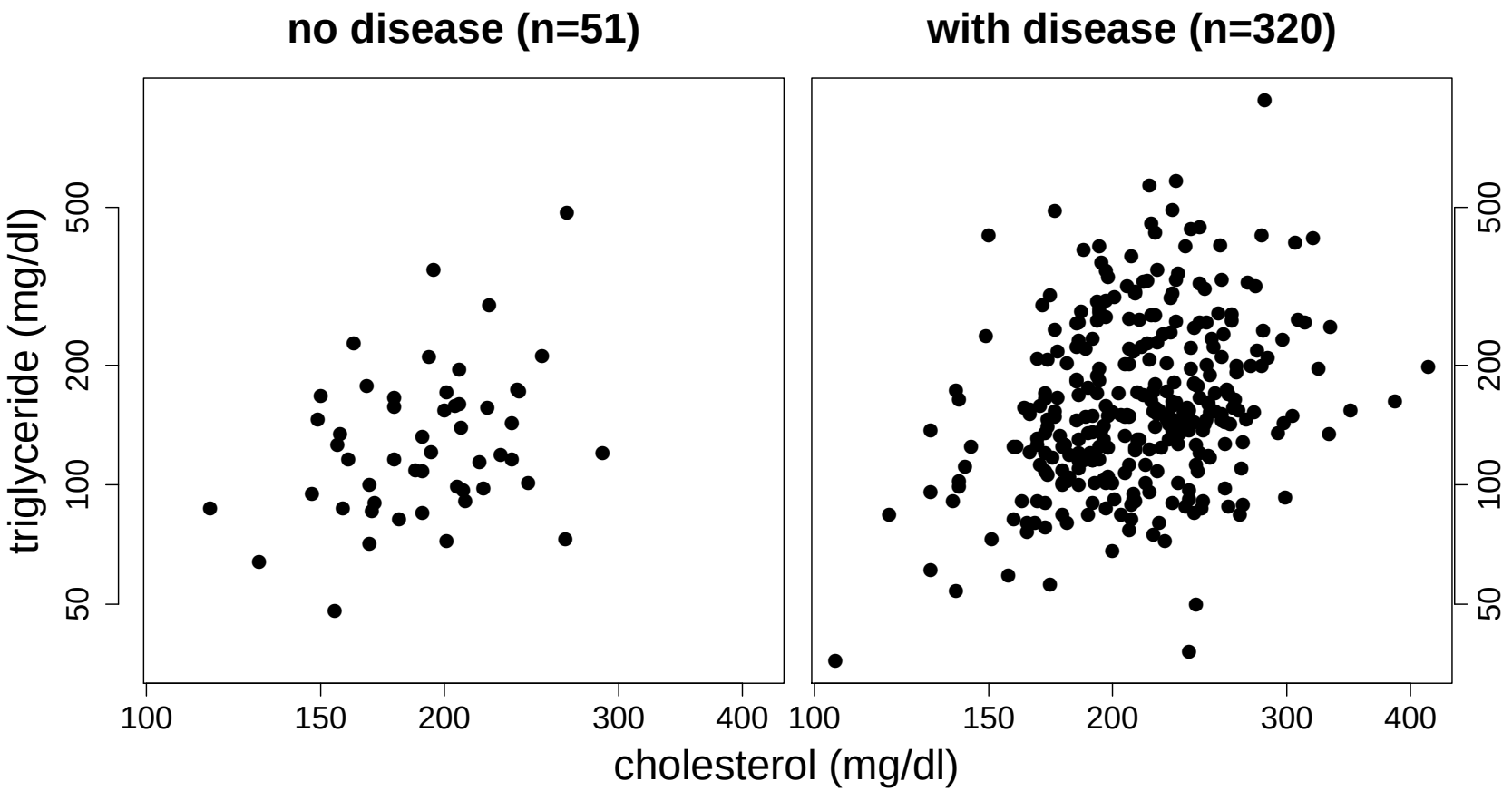


Figure: Scatter diagrams of blood lipid concentrations for 320 diseased and 51 nondiseased males.

Landsat IV Example: Overplotting problems $n > 3000$

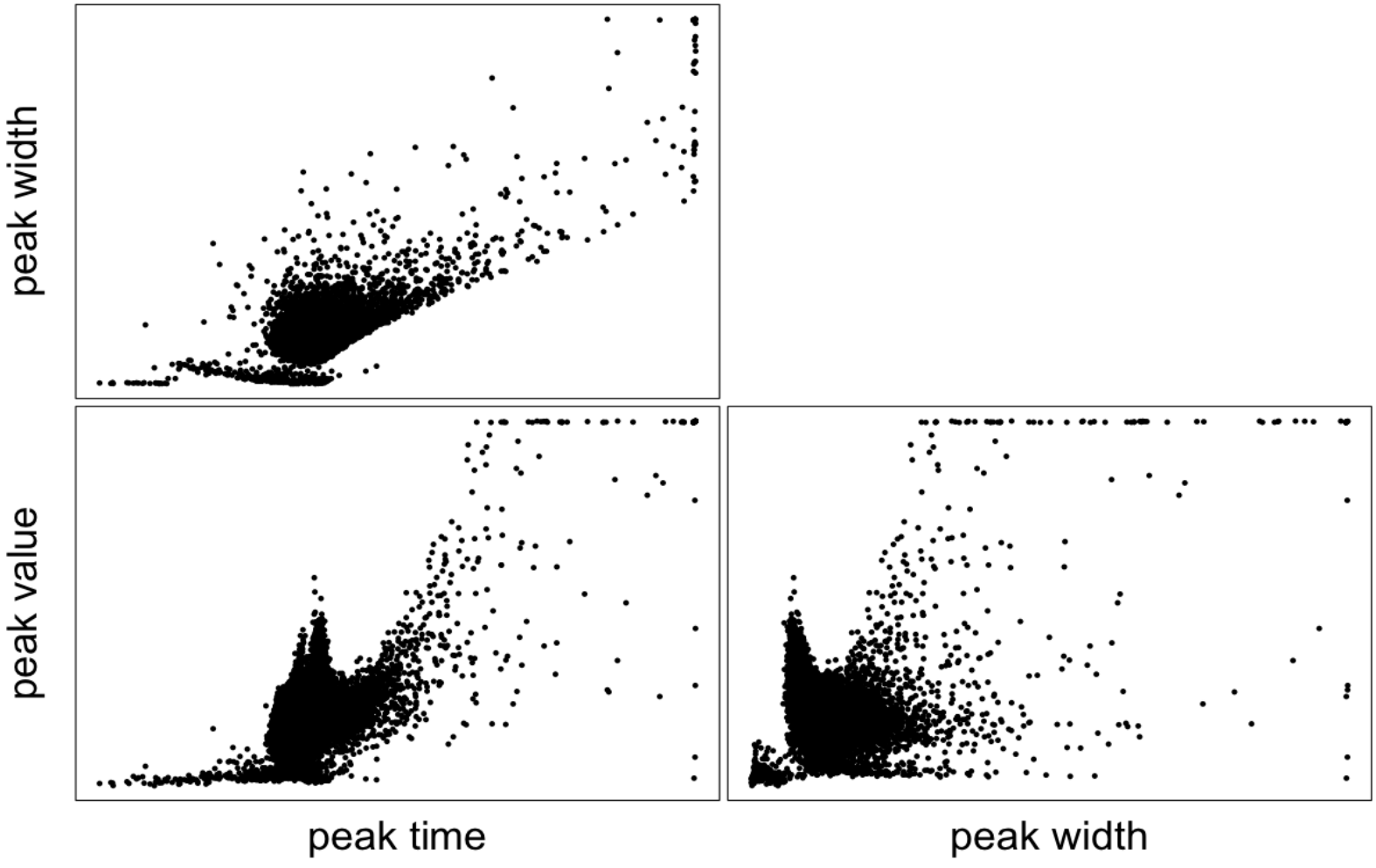
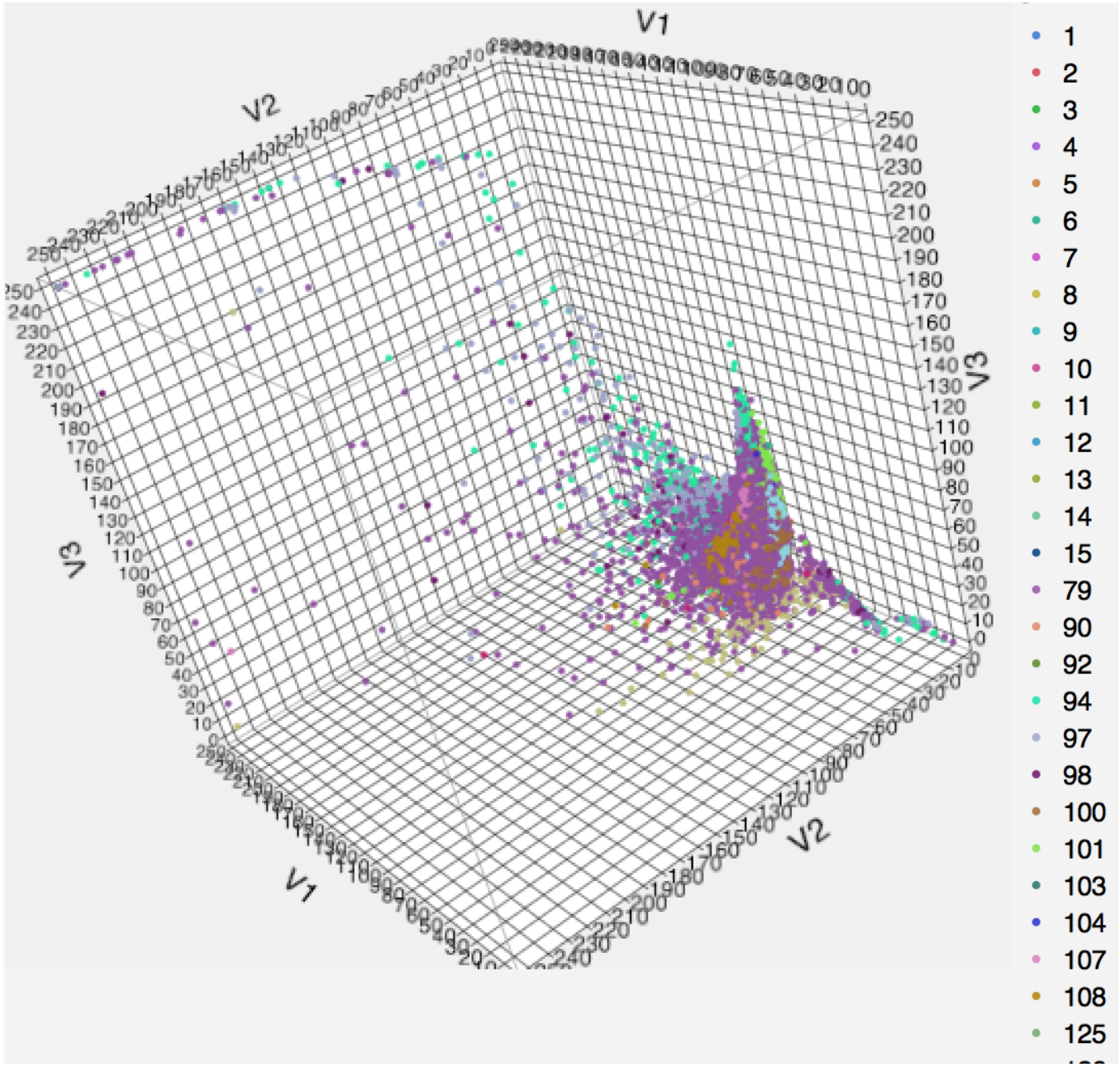
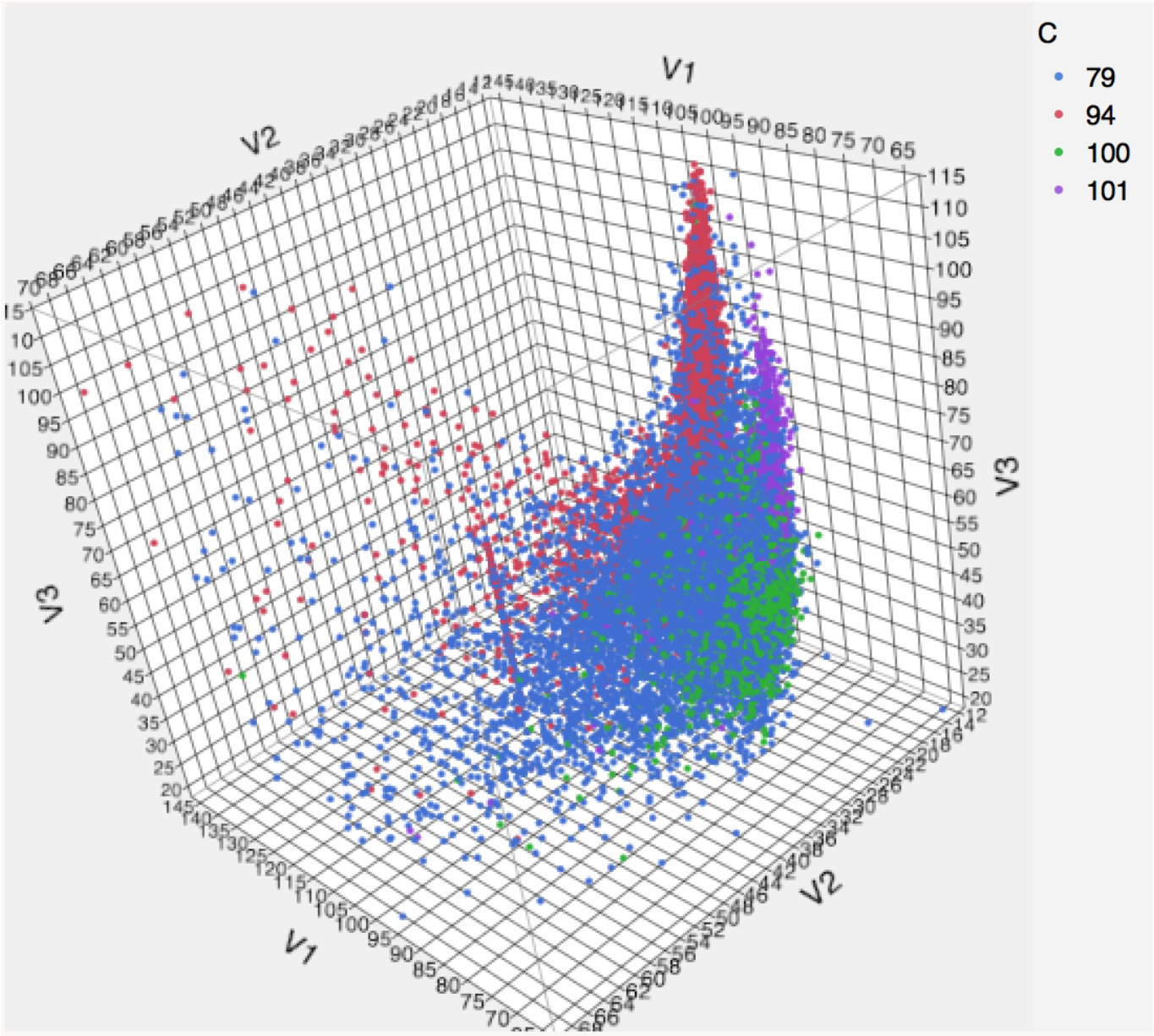


Figure: Pairwise scatter diagram of transformed Landsat data from 22,932 pixels over a 5 by 6 nautical mile region. The range on all the axes is (0, 255).

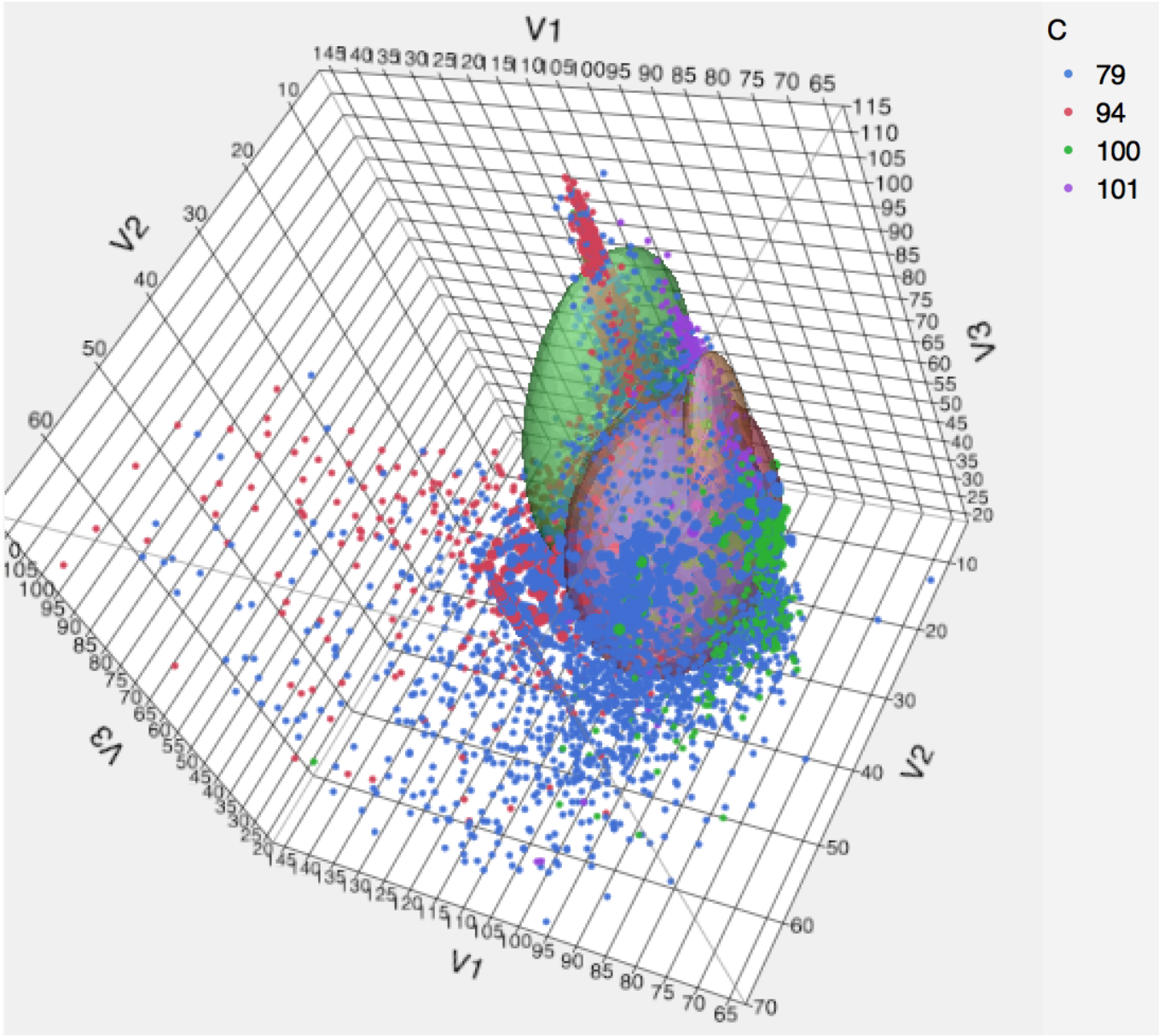
JMP Visualizations with Normal Fits



JMP Visualizations with Normal Fits



JMP Visualizations with Normal Fits



PRIM4 Example: Rotating Scatterplots (w/ brushing)

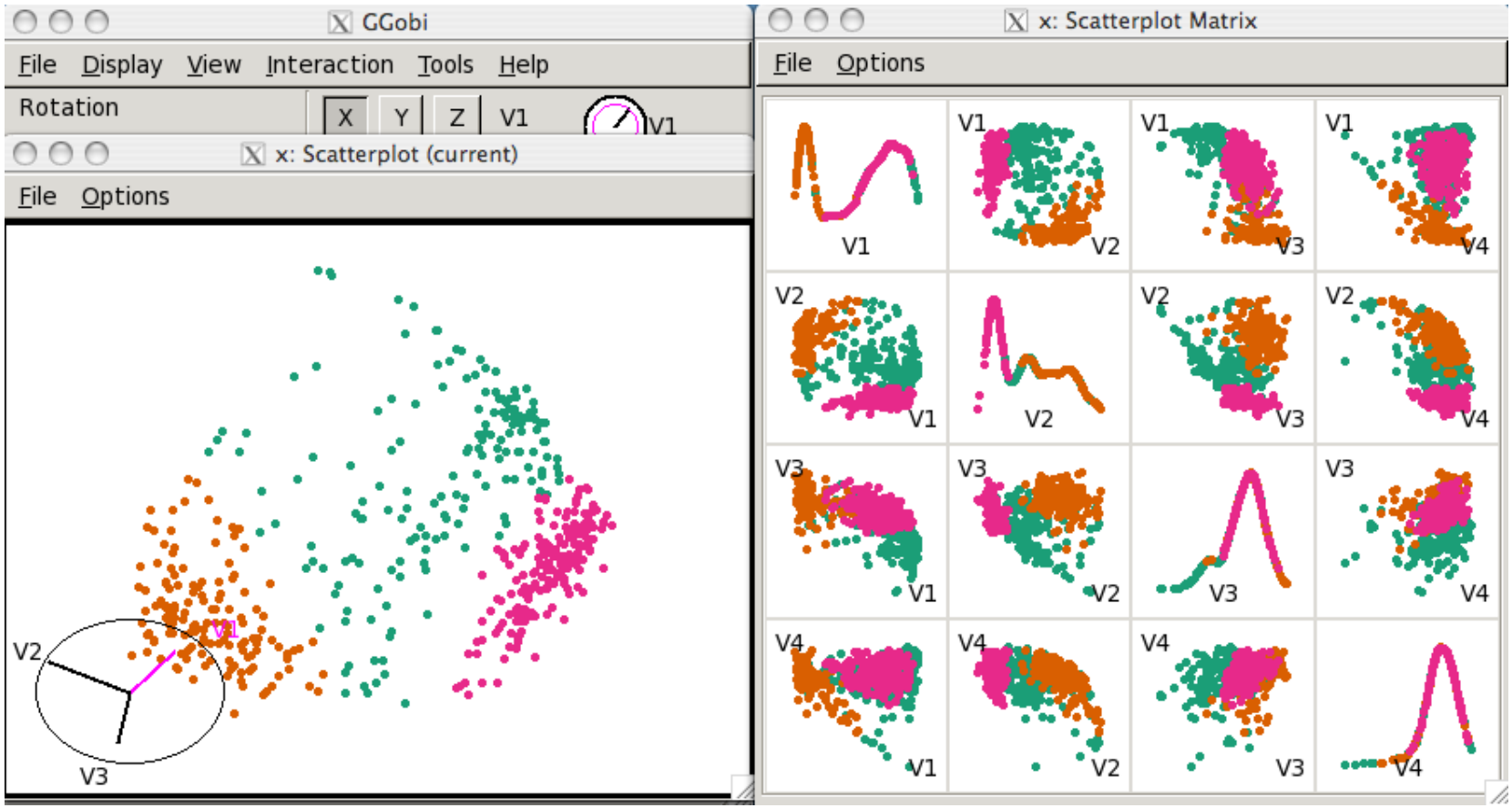
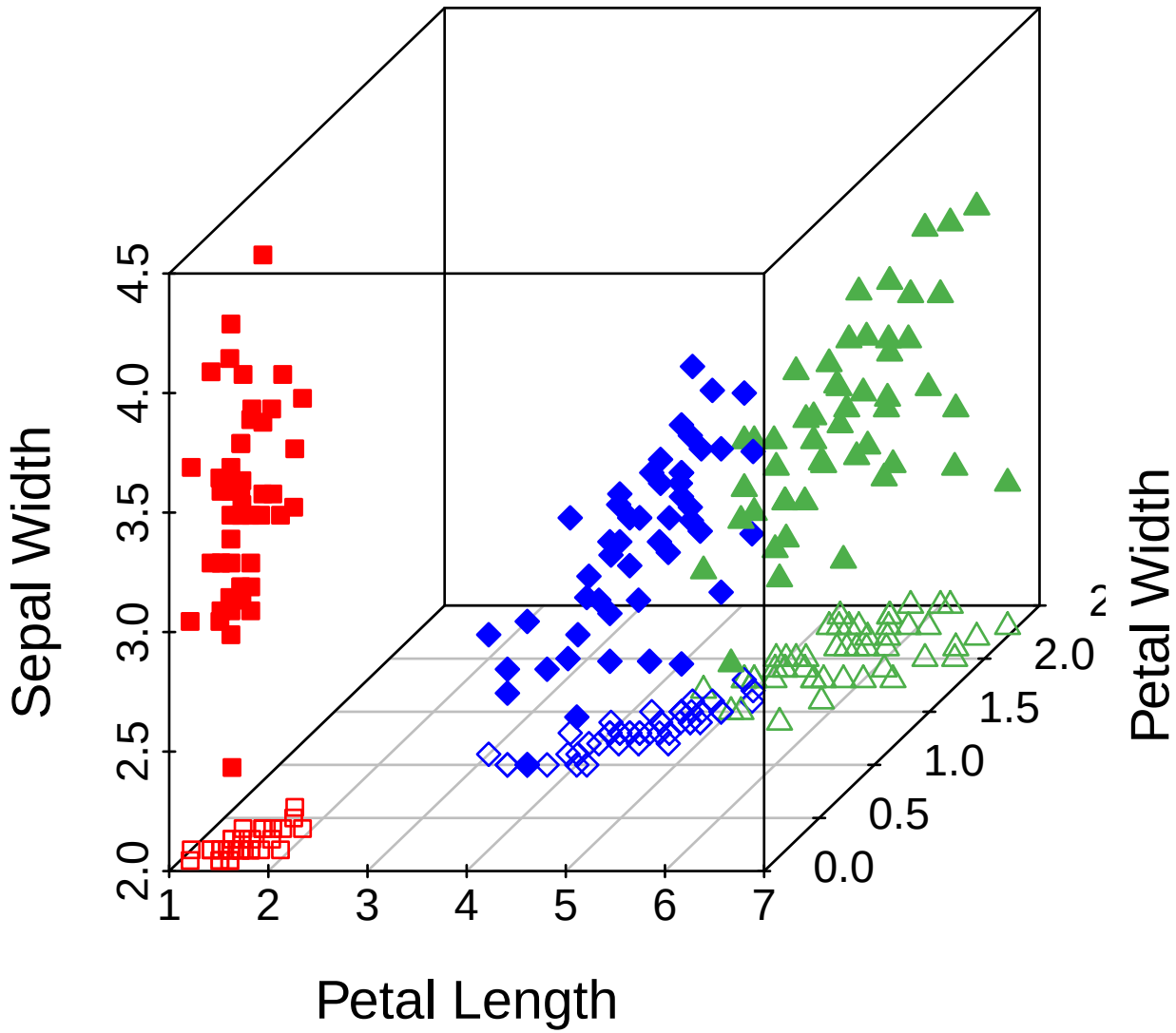


Figure: Pairwise scatterplots of the transformed PRIM4s data using the ggobi visualization system. Two clumps of points are highlighted by brushing. (Note: ggobi and xgobi and rggobi are to be replaced by ggvis.)

Scatterplots using glyphs ($3 \leq p \leq 7$): Iris example



Scatterplots using glyphs ($3 \leq p \leq 7$): Iris example

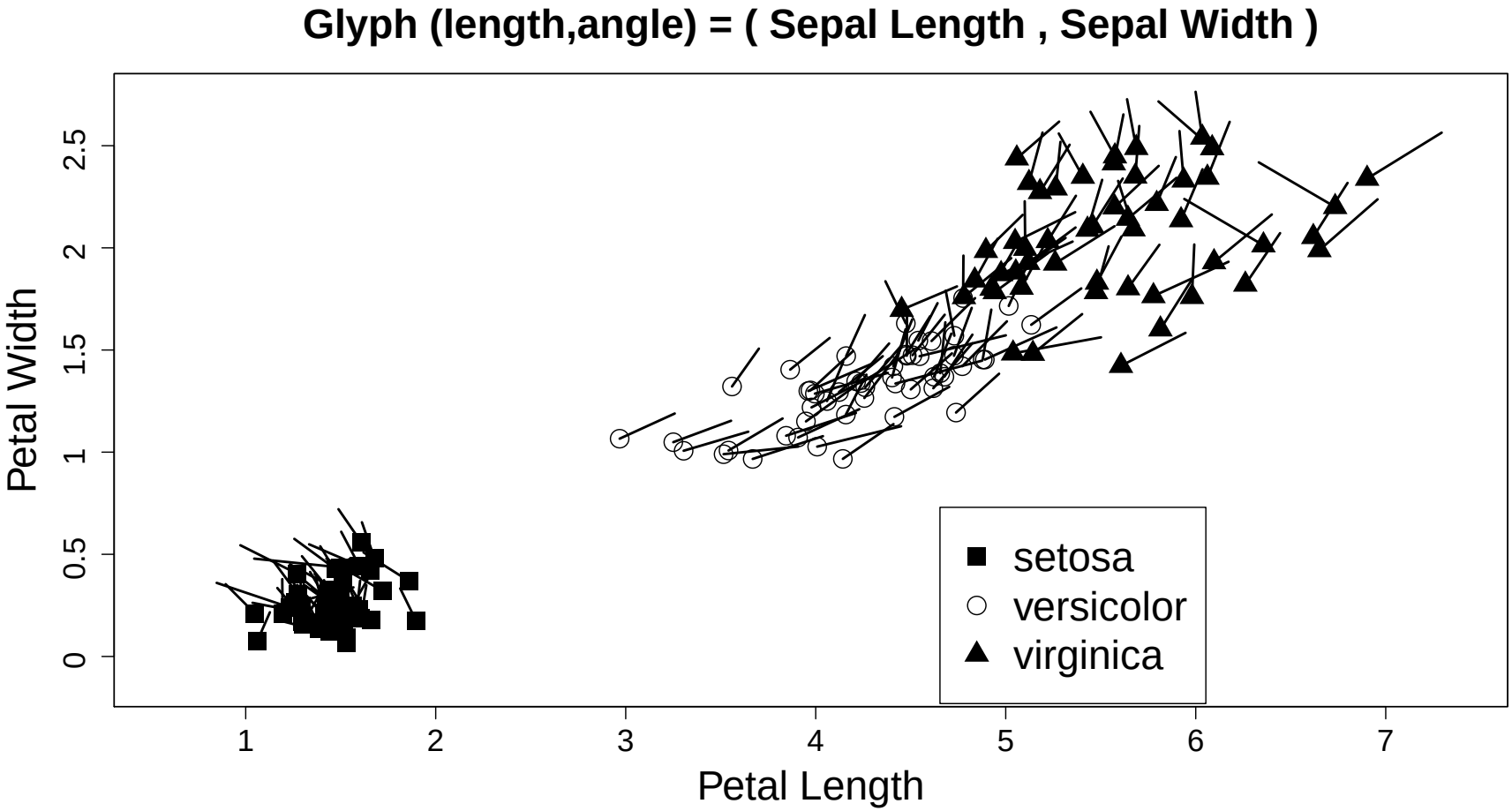
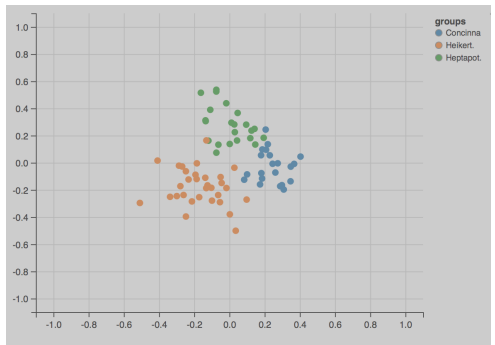


Figure: Glyph scatter diagram of the *Iris* data.

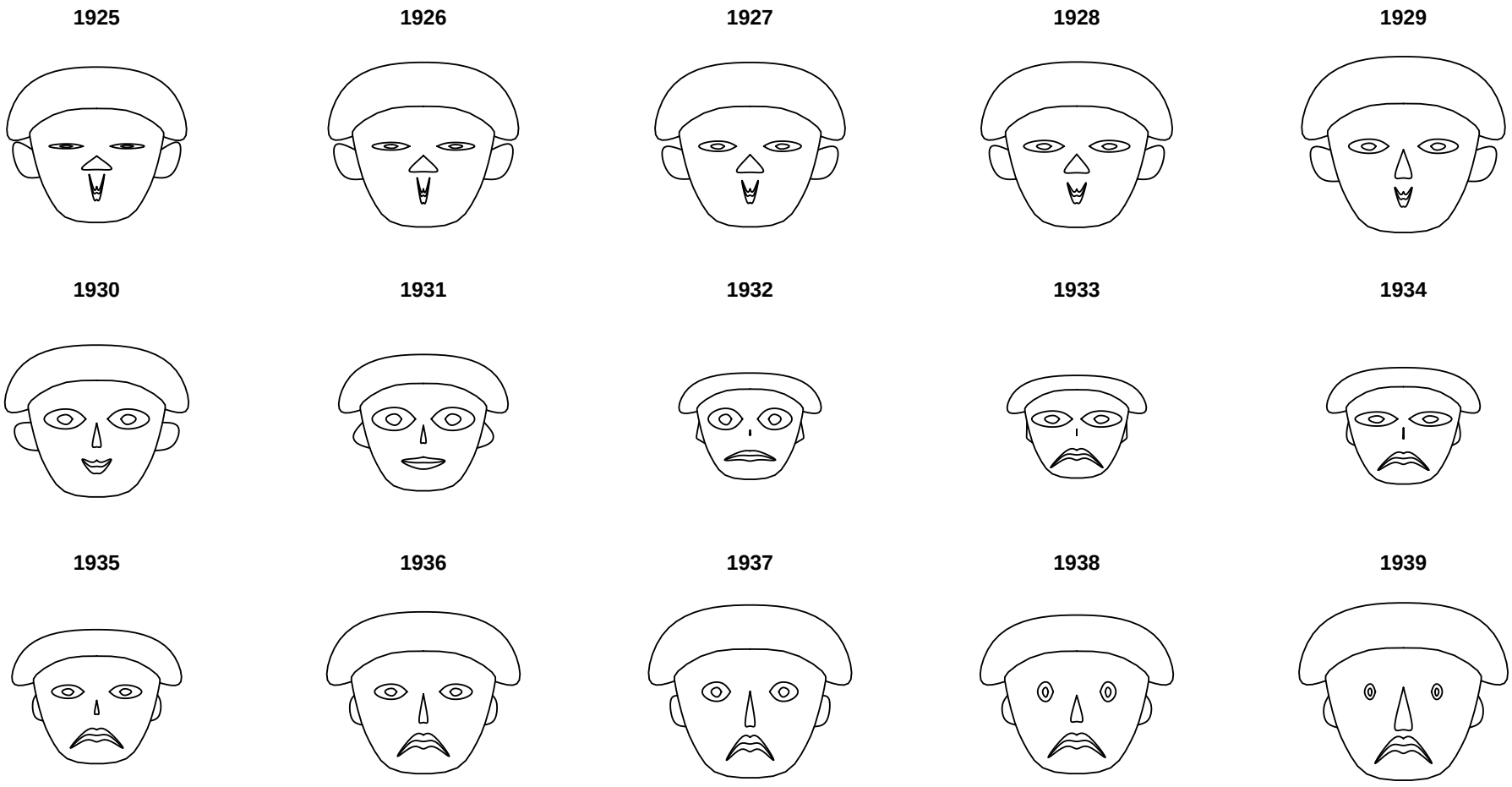
Other Ideas

- The Grand Tour — continuous projection from $\Re^p \rightarrow \Re^2$

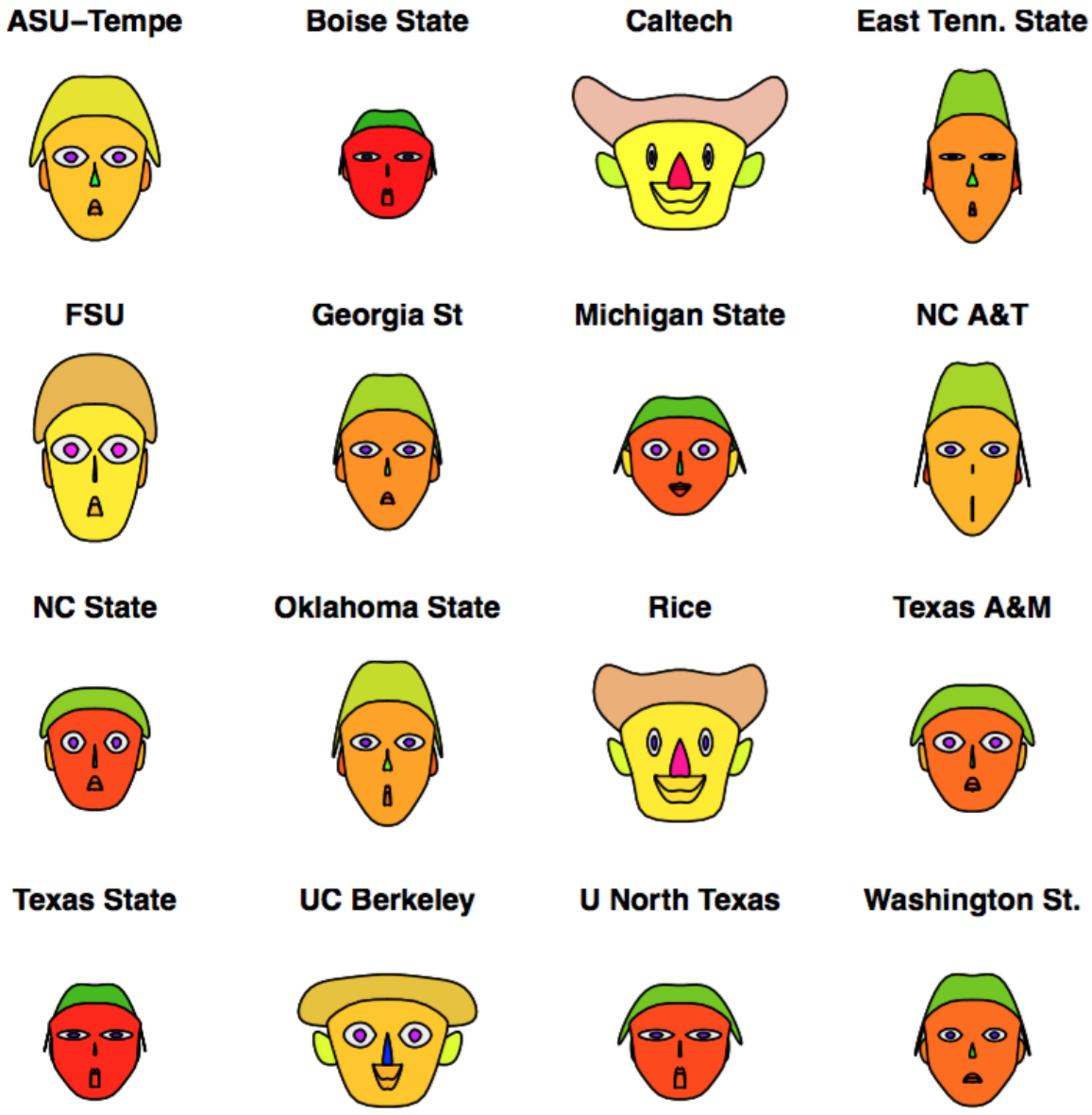


- Chernoff Faces — when each $\mathbf{x}_i \in \Re^p$ is important
- Star diagrams
- Parallel Coordinates — abandon Euclidean coordinates

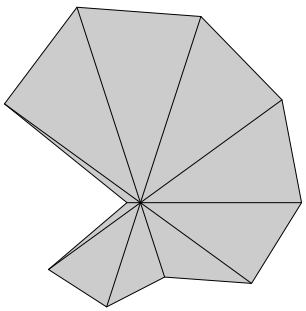
Chernoff Faces: Economic Data 1925-1939



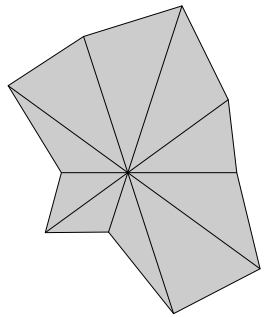
Chernoff Faces: American Universities



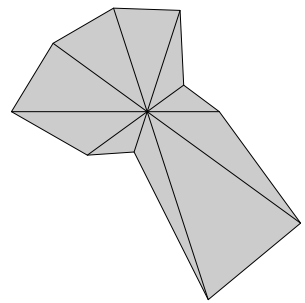
Chernoff Faces: Economic Data 1929-1932



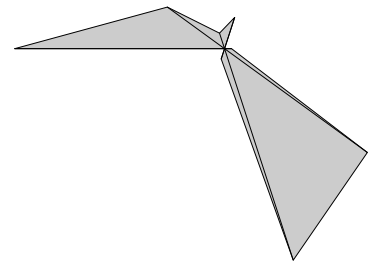
1929



1930



1931



1932

Parallel Coordinates

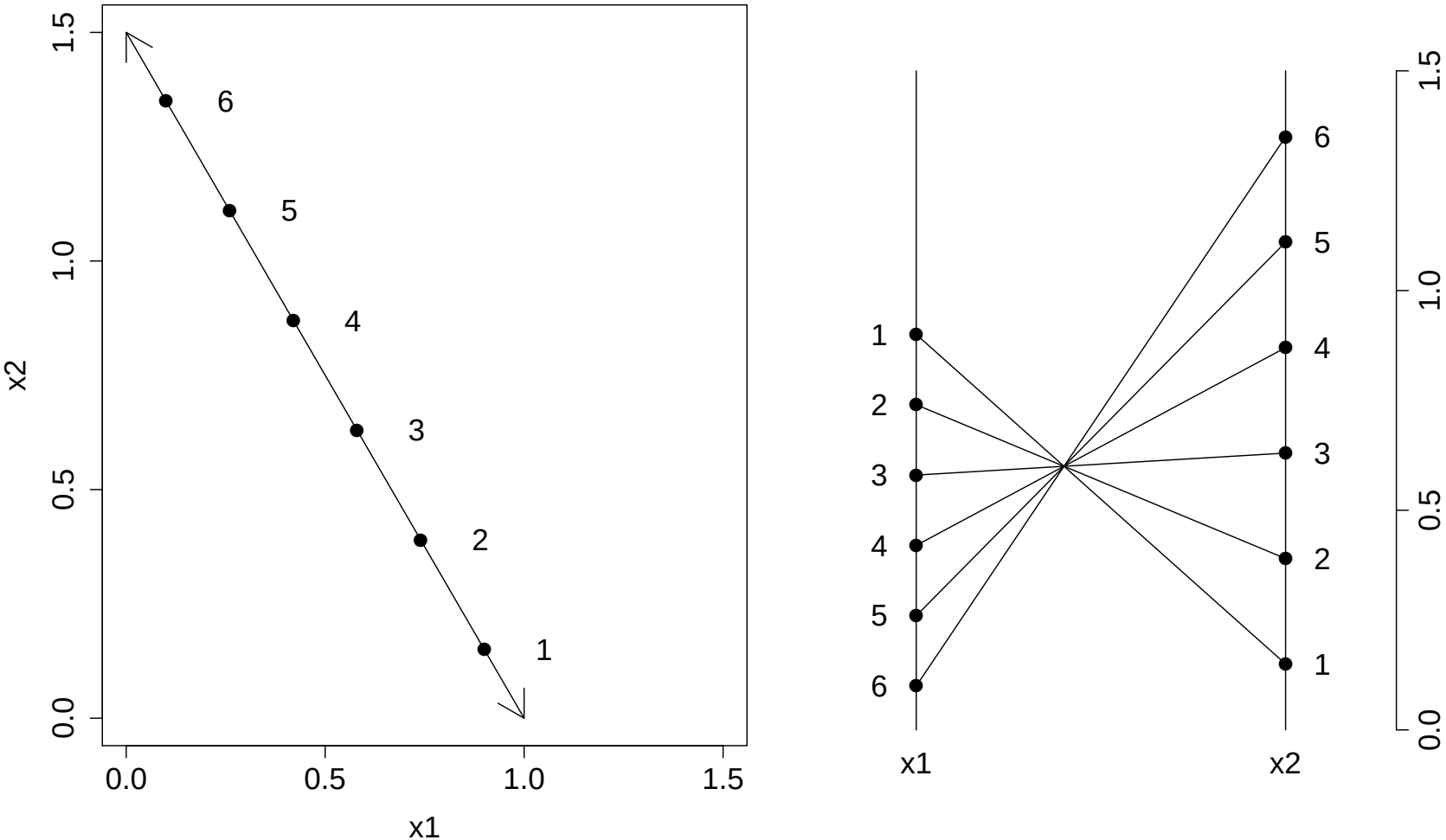
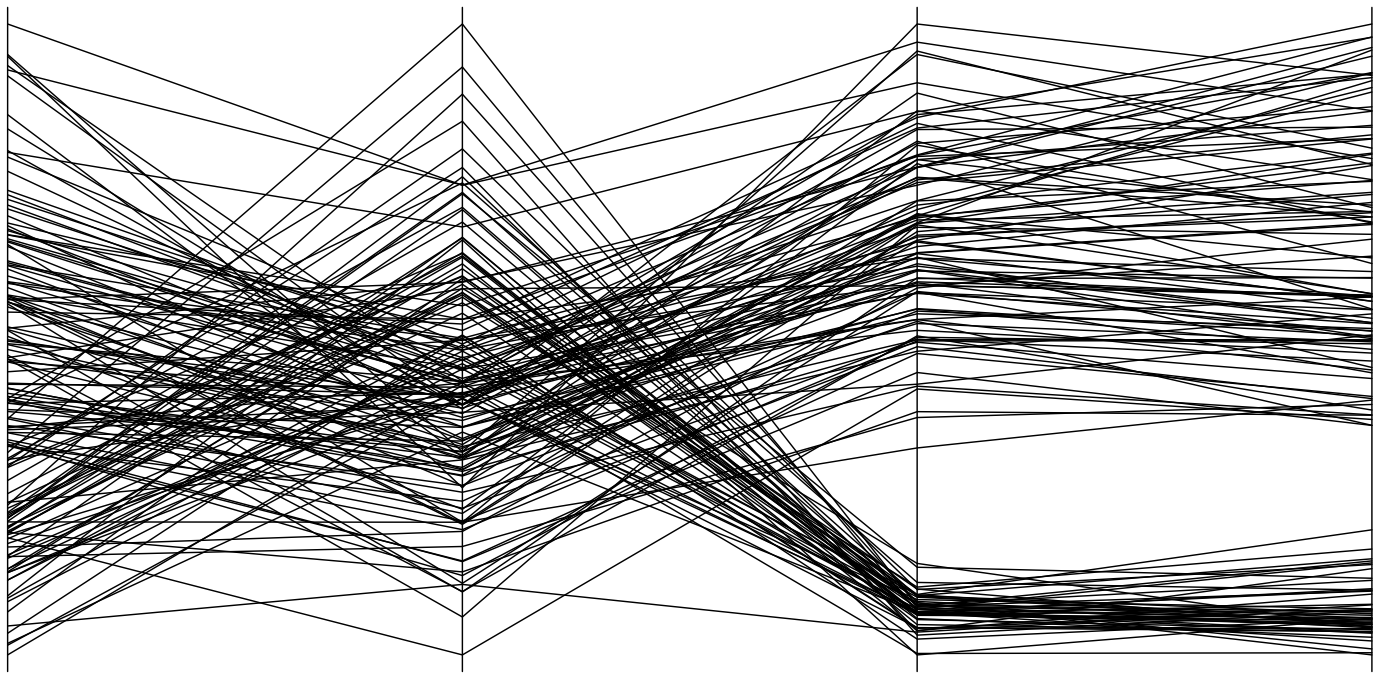
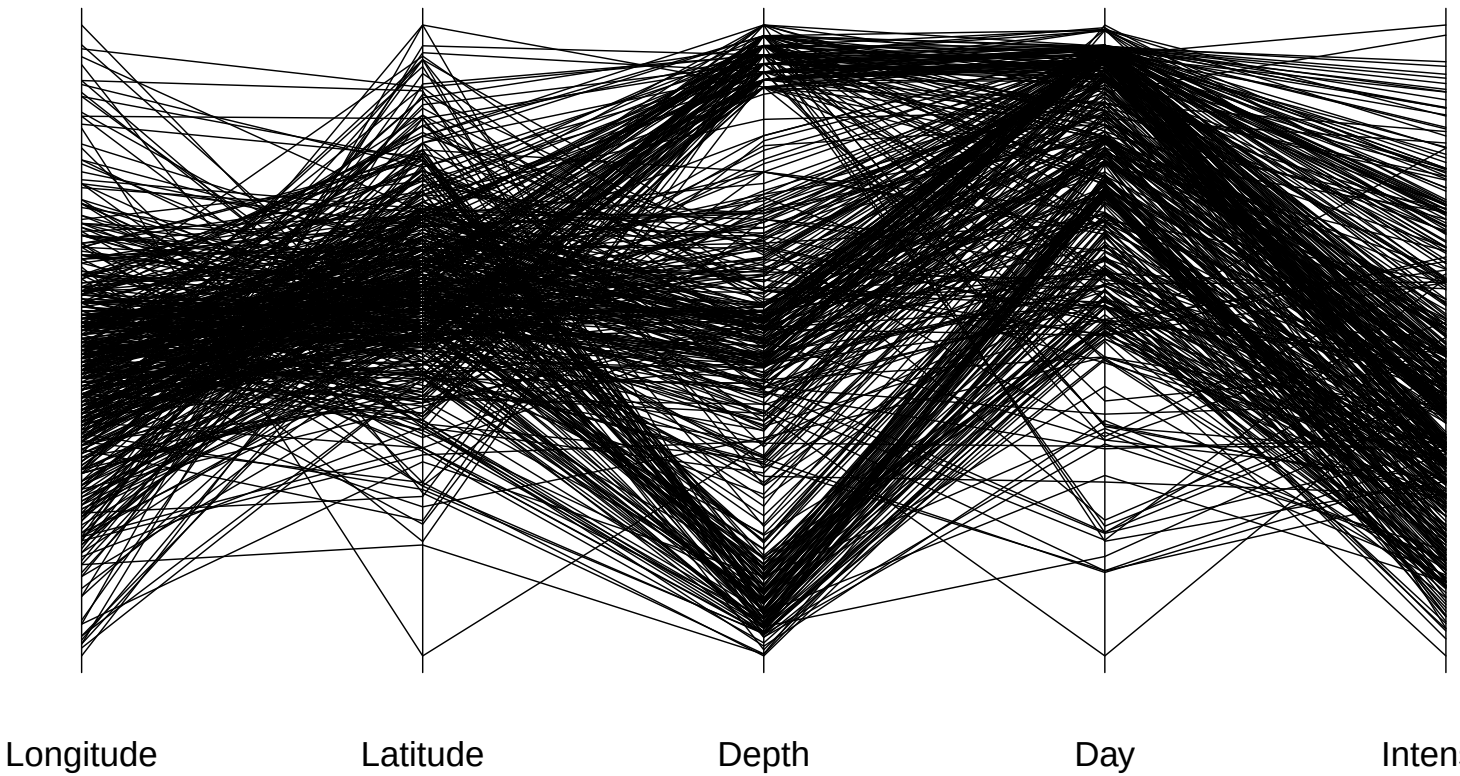


Figure: Example of duality of points and lines between Euclidean and parallel coordinates. The points are labeled 1 to 6 in both coordinate systems.

Parallel Coordinates: Iris example



Parallel Coordinates: Mount St. Helens Earthquakes



Limitations of Visualization of Points

- most valuable with small data sets, where individual points are identifiable and interesting
- limited ability to display high-dimensional structure
- for large n, scatter diagrams emphasize tails/outliers
- point-based graphics do not provide a consistent picture of the data as $n \to \infty$?

Graphical Display of Multivariate Functionals

Scatterplot smoothing by density functions

- the scatter diagrams points to the bivariate density function
- i.e. the raw data need to be smoothed in order to obtain a consistent view
- as $n \to \infty$, the bivariate density function estimate gets better!
- the histogram is the simplest example of a scatterplot smoother

United States Mortality Histogram

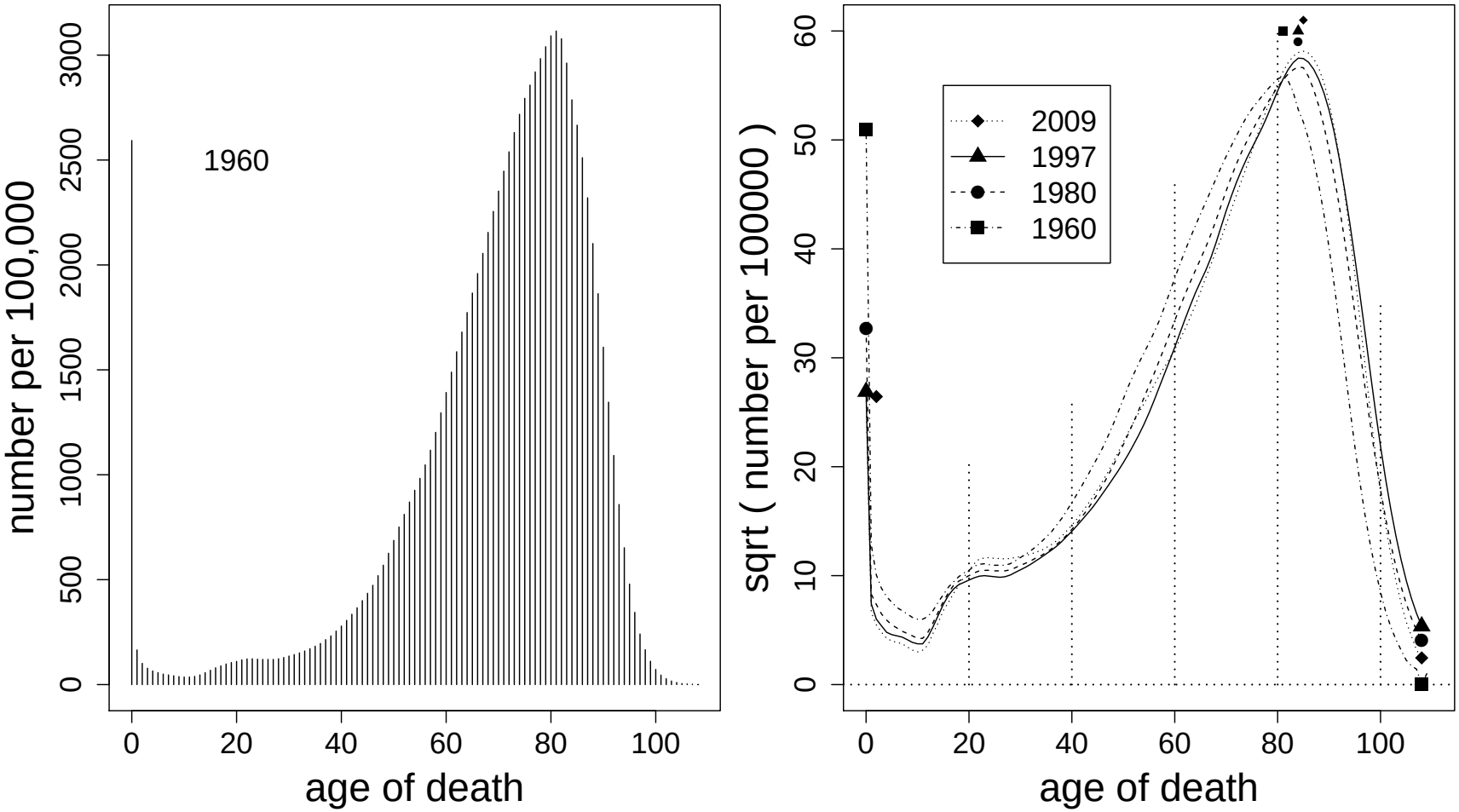


Figure: Histogram of the U.S. mortality data in 1960. Rootgrams (histograms plotted on a square-root scale) of the mortality data for 1960 — 2009. History: John Graunt's Bills of Mortality (1662),

When the raw data are binned? (Because n large.)

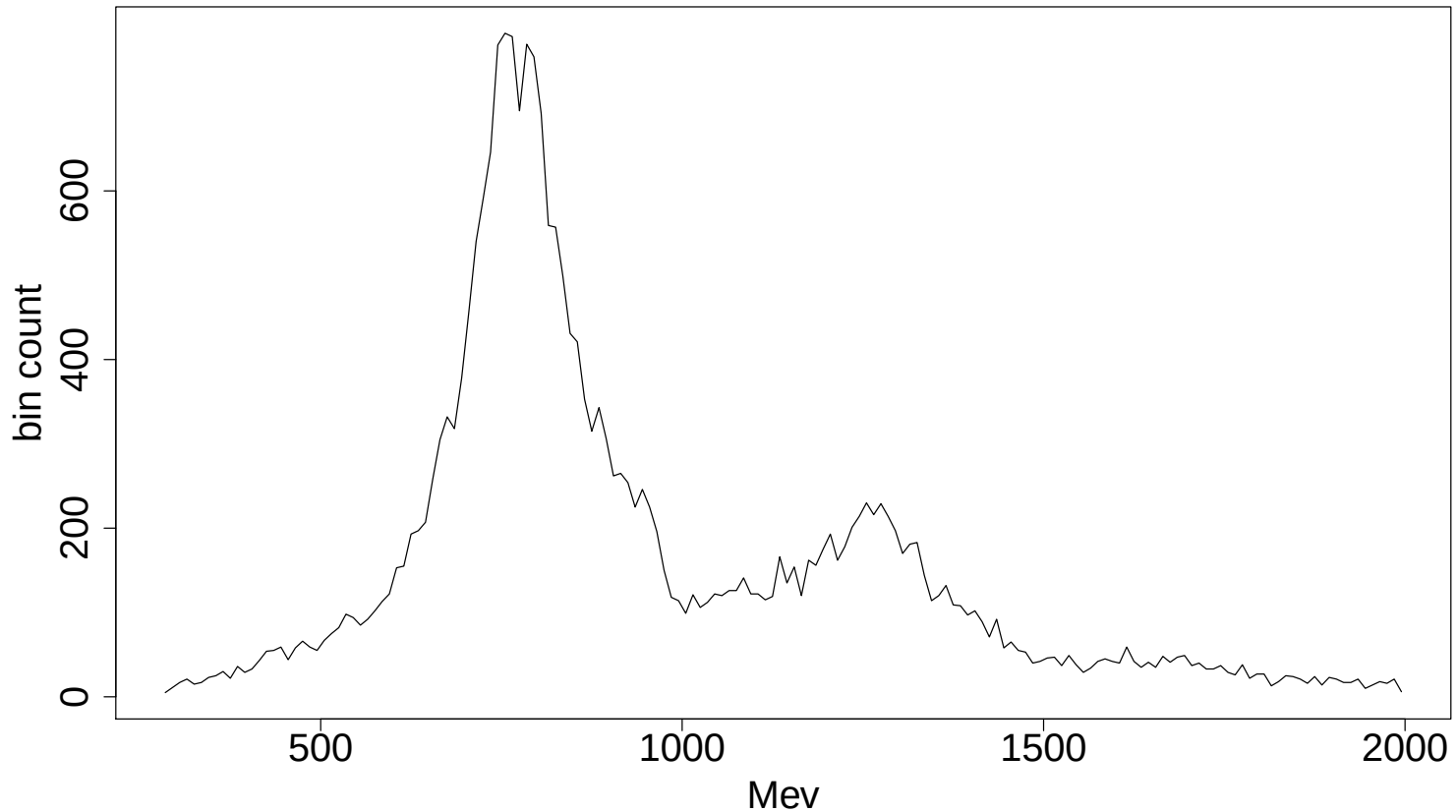


Figure: Histogram of LRL dataset. (I.J. Good's "bump" example, with $n = 25{,}752$ and 300 bins.)

Lagged Time Series Data (Old Faithful eruptions)

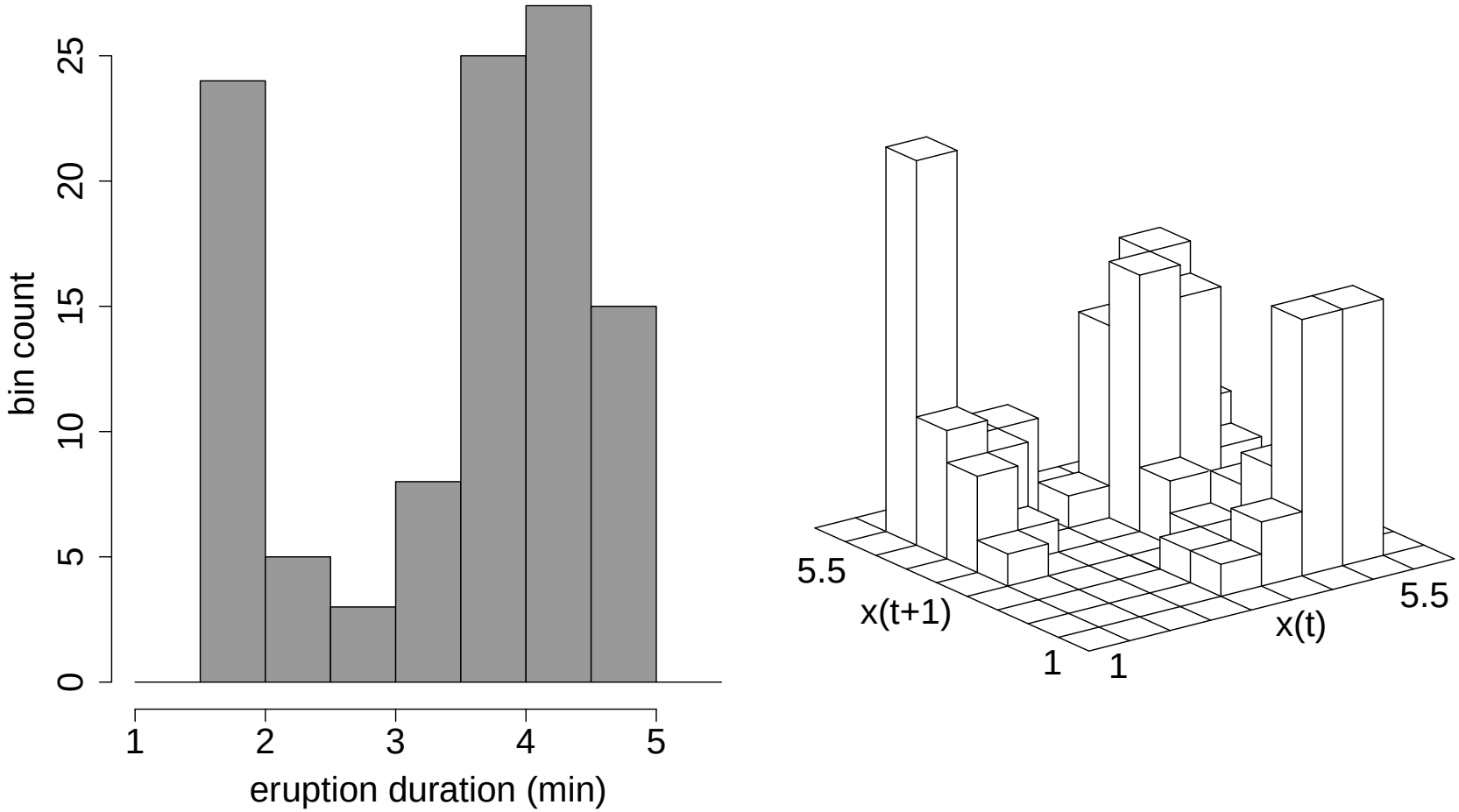


Figure: Histogram of $\{x_t\}$ for the Old Faithful geyser dataset, and a bivariate histogram of the lagged data (x_t, x_{t+1}).

Slice Trivariate Data (LANDSAT IV): $\hat{f}(x_1, x_2, x_3)$

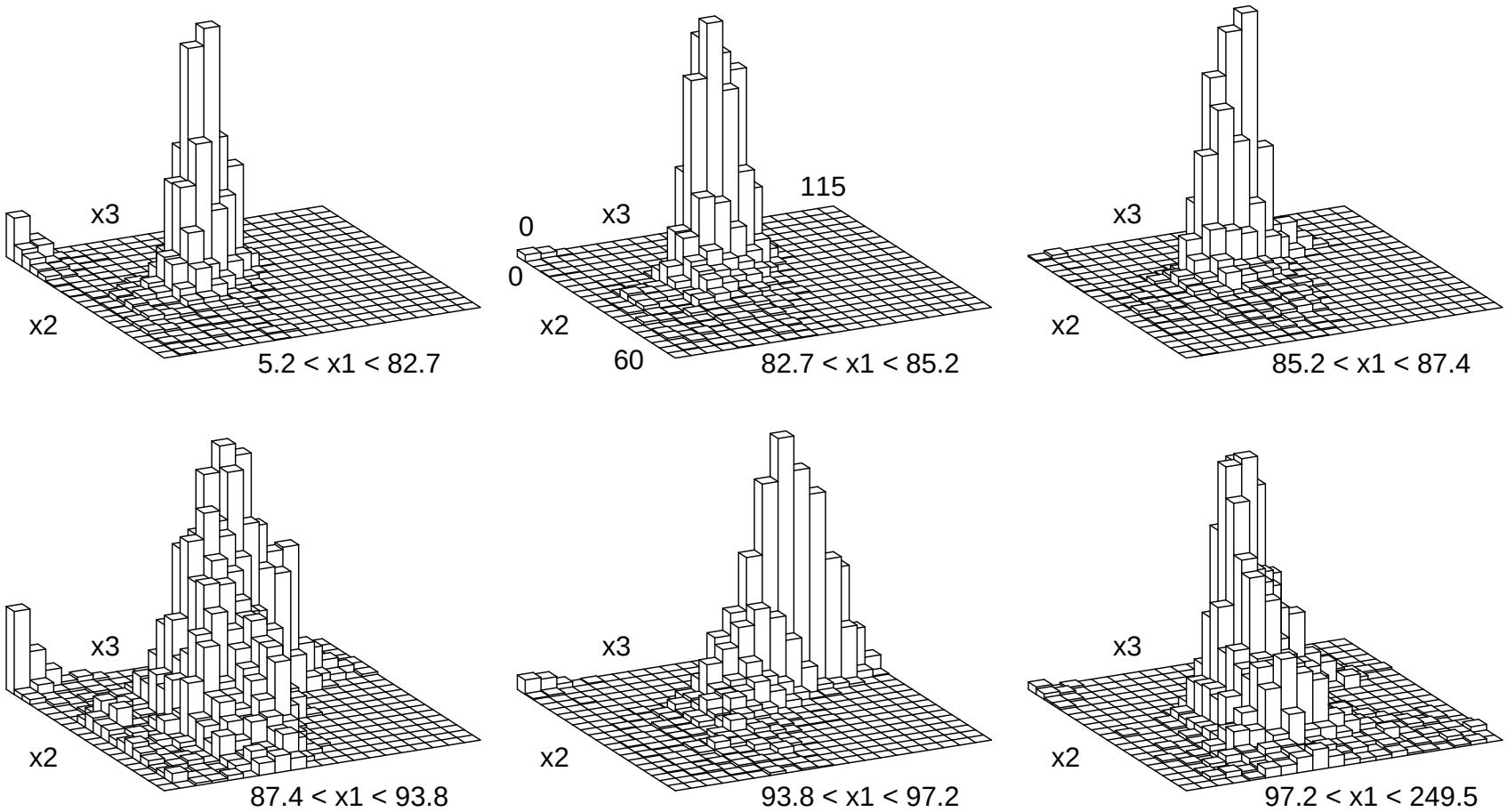


Figure: Bivariate histogram slices of the trivariate Landsat data. Slicing was performed at the quantiles of variable x_1.

Scatterplot smoothing: Regression case (gas flow data)

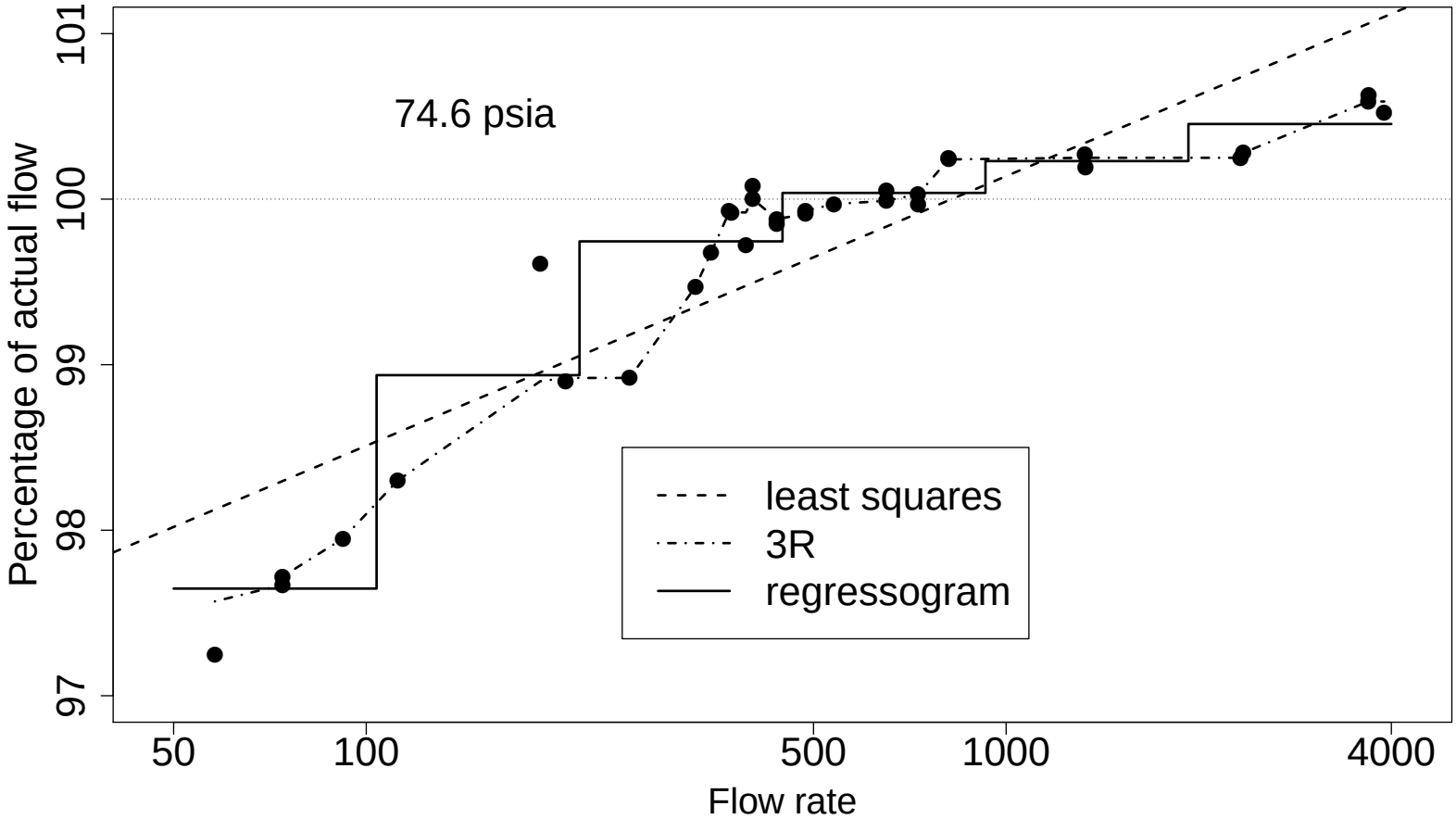


Figure: Accuracy of a natural gas meter as a function of the flow rate through the valve at 74.6 psia. The raw data ($n = 33$) are shown by the filled points. The 3 smooths (least squares, Tukey's 3R, and Tukey's regressogram) are superimposed.

Complete Gas Flow Data: Function of Pressure

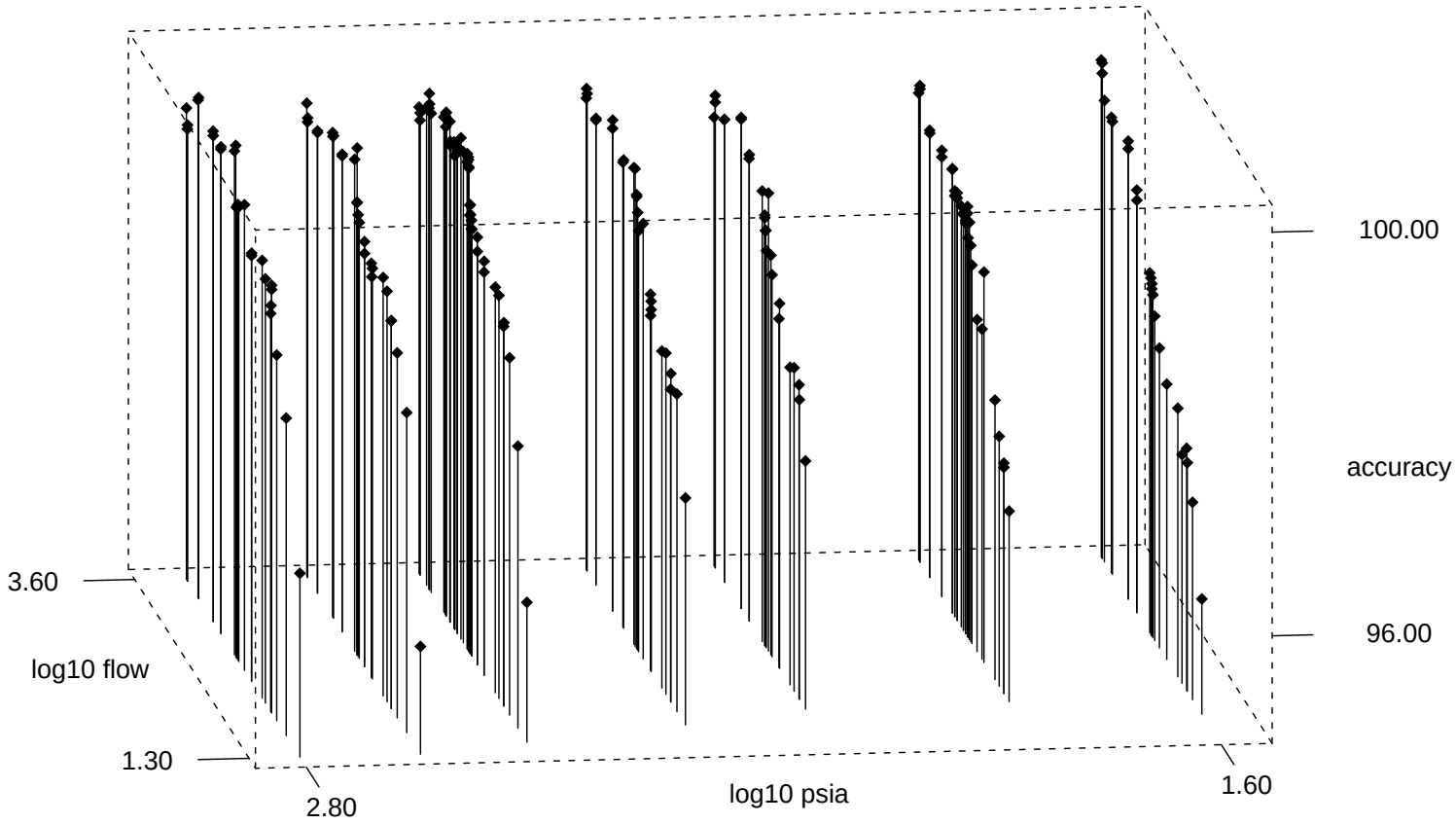
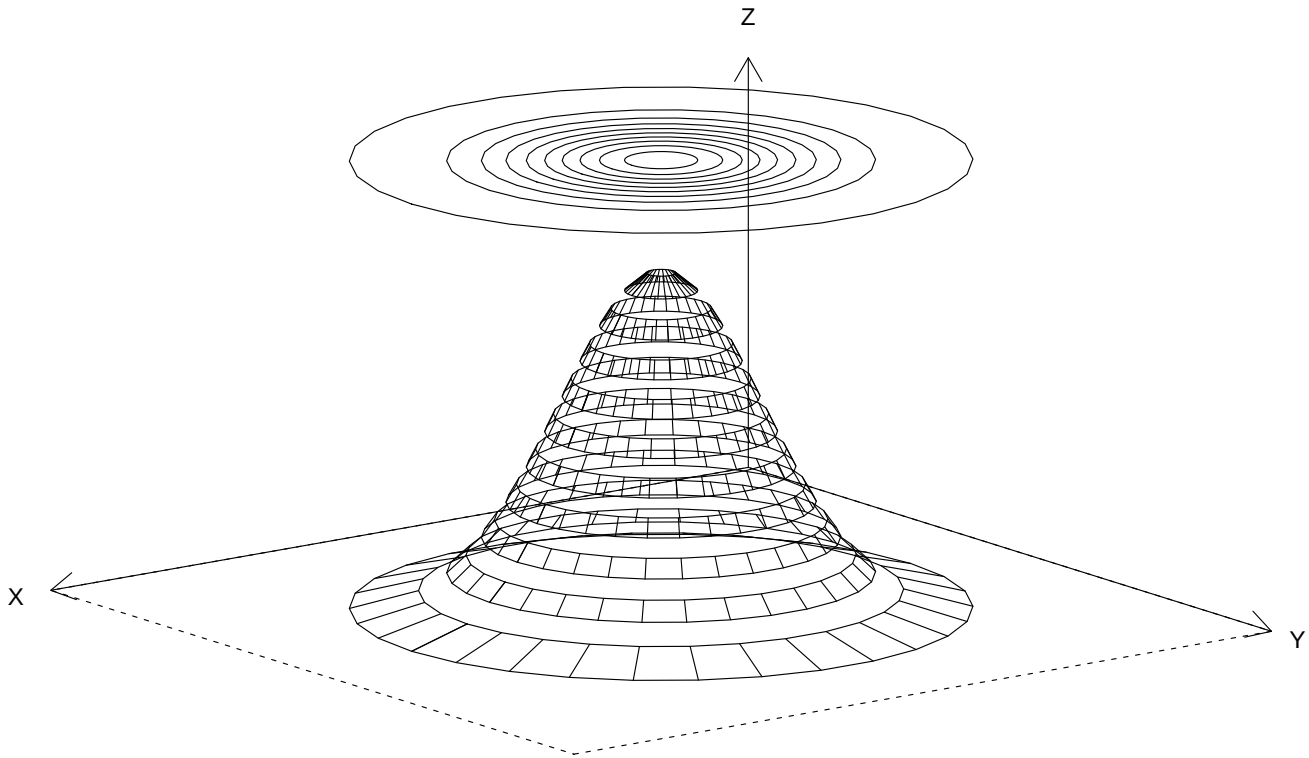


Figure: Complete 3-D view of the gas flow dataset (cf 2nd from right).

Visualization of Multivariate Functions

Contour versus Perspective: preference?



- **contour view:** data in $\Re^2$ and so are contours (equally spaced vertical slices)
- **perspective view:** in $\Re^3$, but only see $2\frac{2}{3}$ dimensions?
- what if data in $\Re^3$?

Contour Surface and Level Sets

- Introduce a notation for a particular level set:

$$\boxed{\alpha\text{-Contour}: \qquad S_\alpha = \{\mathbf{x} : f(\mathbf{x}) = \alpha f_{\max}\}, \qquad 0 \leq \alpha \leq 1}$$

- For normal data, the general contour surfaces are hyper-ellipses defined by the equation

$$(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) = -2 \log \alpha$$

- A representation of trivariate contour plot of $f(x_1, x_2, x_3)$ would generally contain several "nested" surfaces, for example, $\{S_{0.1}, S_{0.3}, S_{0.5}, S_{0.7}, S_{0.9}\}$

Trivariate Normal Example: Stereo

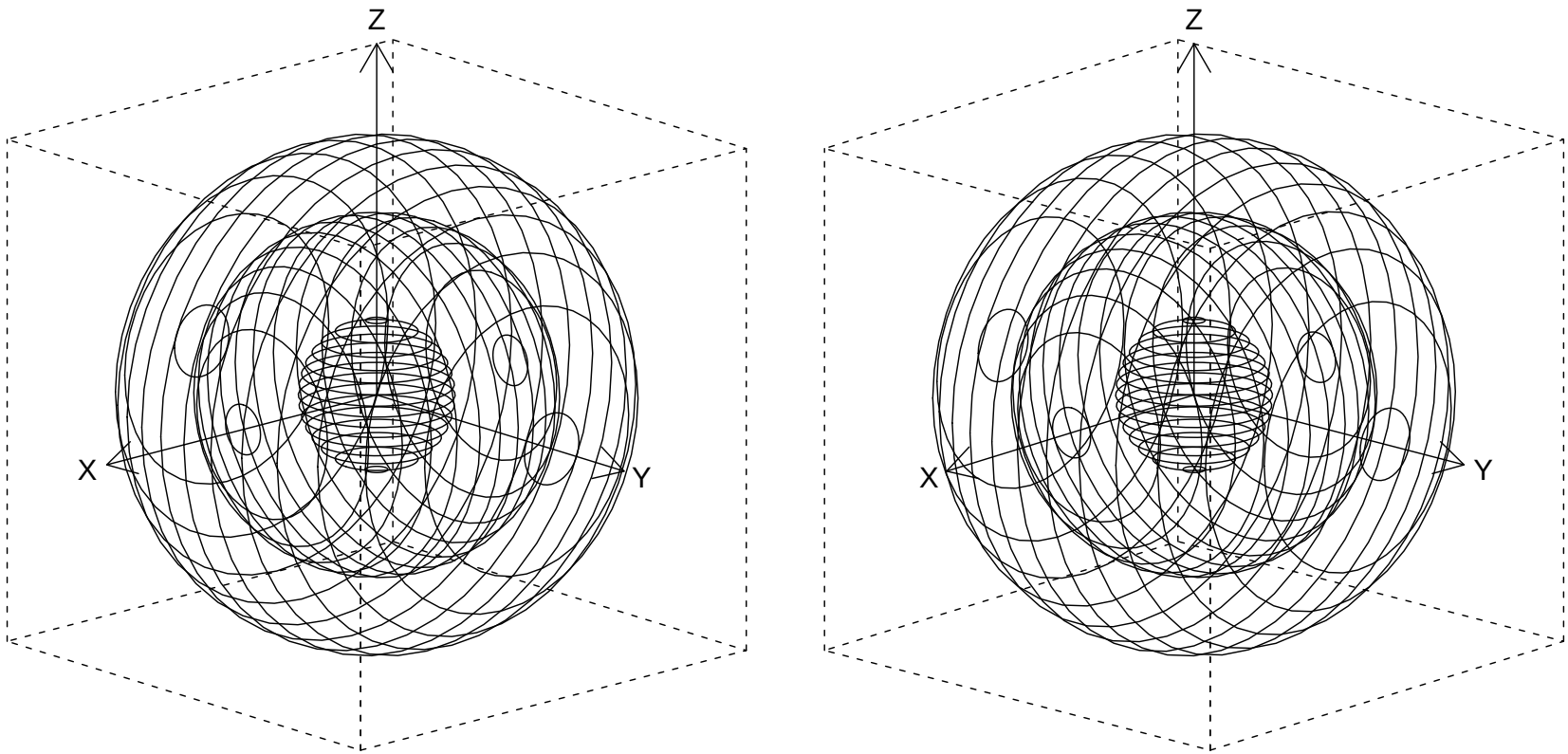
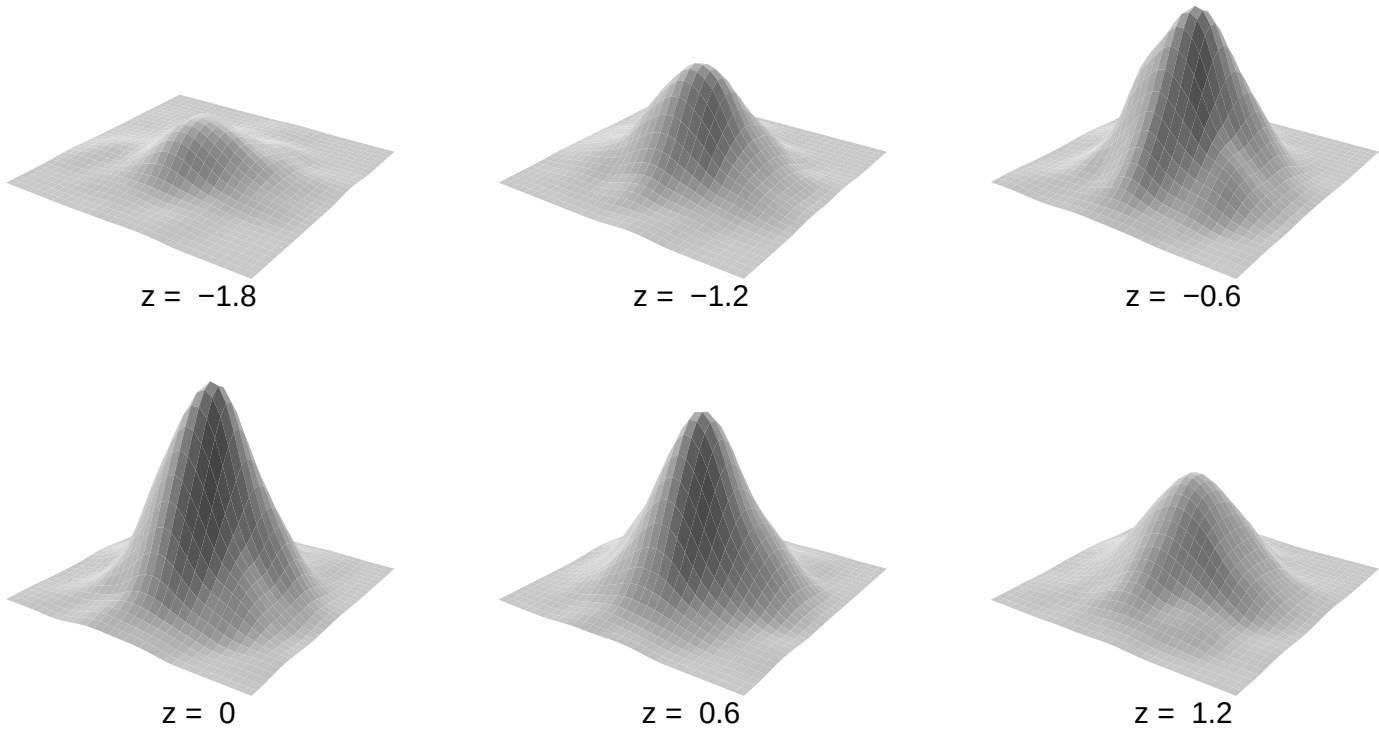


Figure: Stereo representation of 3 α-contours of a trivariate normal density. Gently crossing your eyes should allow the 2 frames to fuse in the middle.

Bivariate Slices of Smoothed Histogram $n = 1000$, $N(\mathbf{0}, \boldsymbol{I}_3)$



Note: These are not the conditional densities,

$$\hat{f}(x, y, z|z = z_0) \quad \text{but the slices} \quad \hat{f}(x, y, z_0)$$

which are the un-normalized slices.

What if normalize?

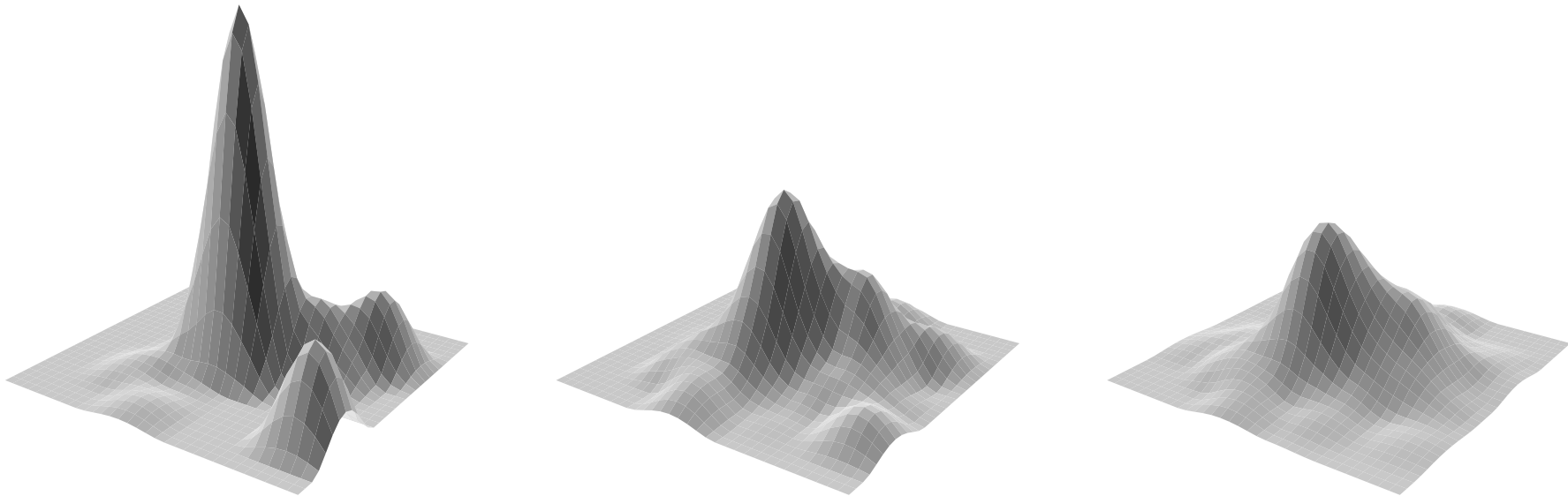
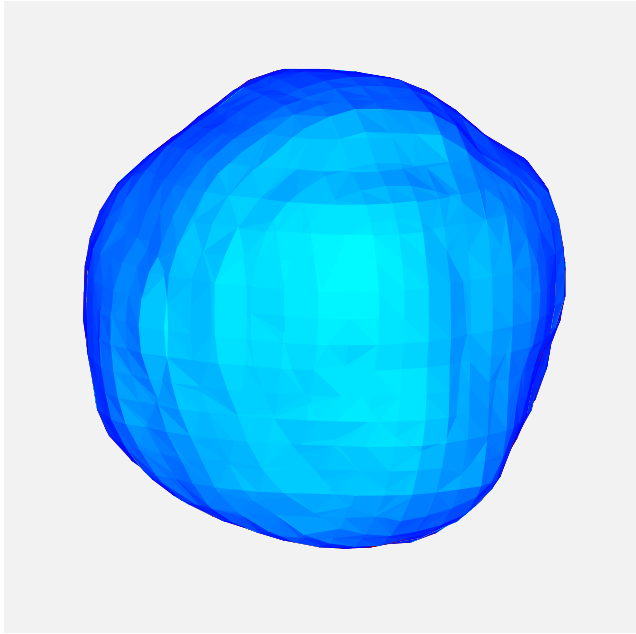
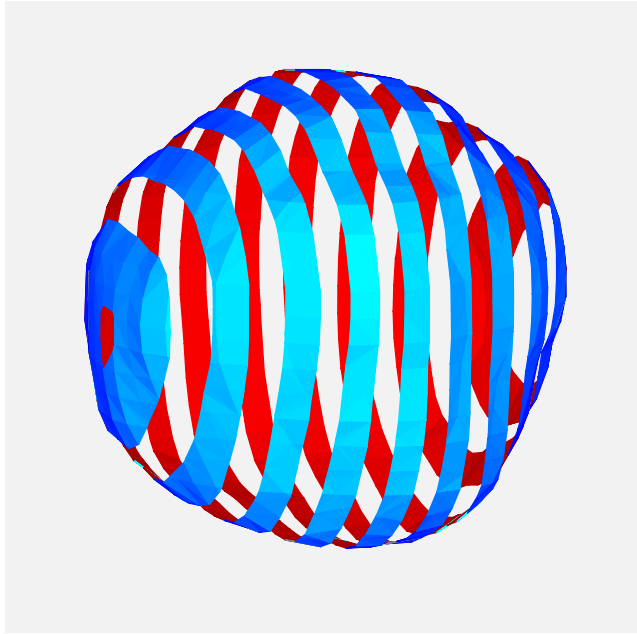


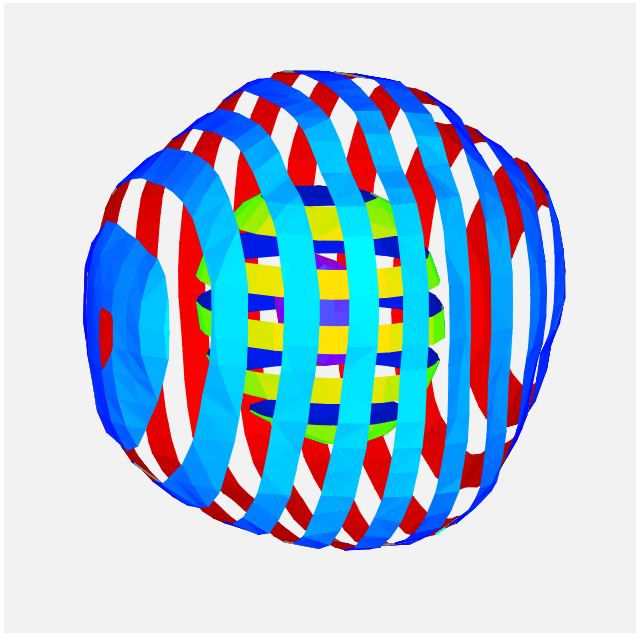
Figure: Normalized slices in the left tail of the smoothed histogram.

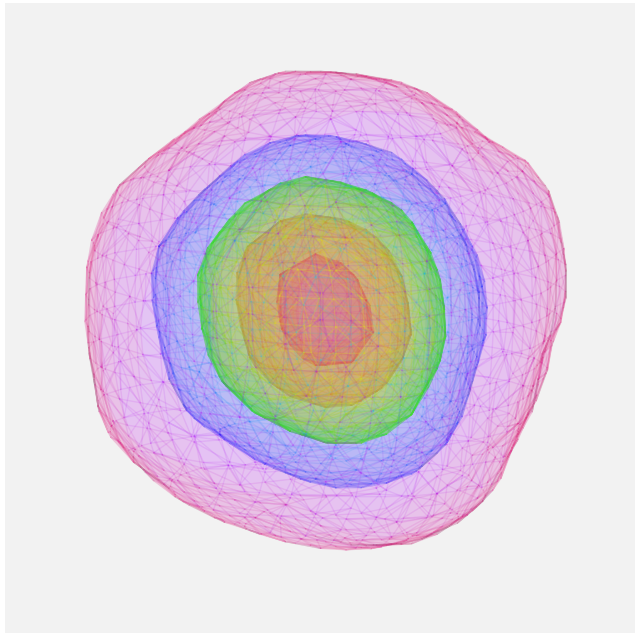
- These are very rough in the tails, since not much data
- If do not normalize, easy to tell if in the middle or in the tails

Other Visualization Examples of these data.









Density Slicing with 4 Variables

In $\Re^4$, the α-level contours of interest are based on the slices:

$$S_{\alpha,t} = \{(x, y, z) : f(x, y, z, t) = \alpha f_{\max}\},$$

- For a fixed choice of α, as the slice value t changes continuously, the contour shells will expand or contract smoothly, finally vanishing for extreme values of t.
- A single theoretical contour of the $N(\mathbf{0}, I_4)$ density is a sphere who radius is greatest when $t = 0$, the shrinks and vanishes as t moves away from 0.
- With several α-shells displayed simultaneously, the contours would be nested spheres of different radii, appearing at different values of t, but of greatest diameter when $t = 0$.

Example of Iris Data, Sliced at Sepal Width = 3.4 cm

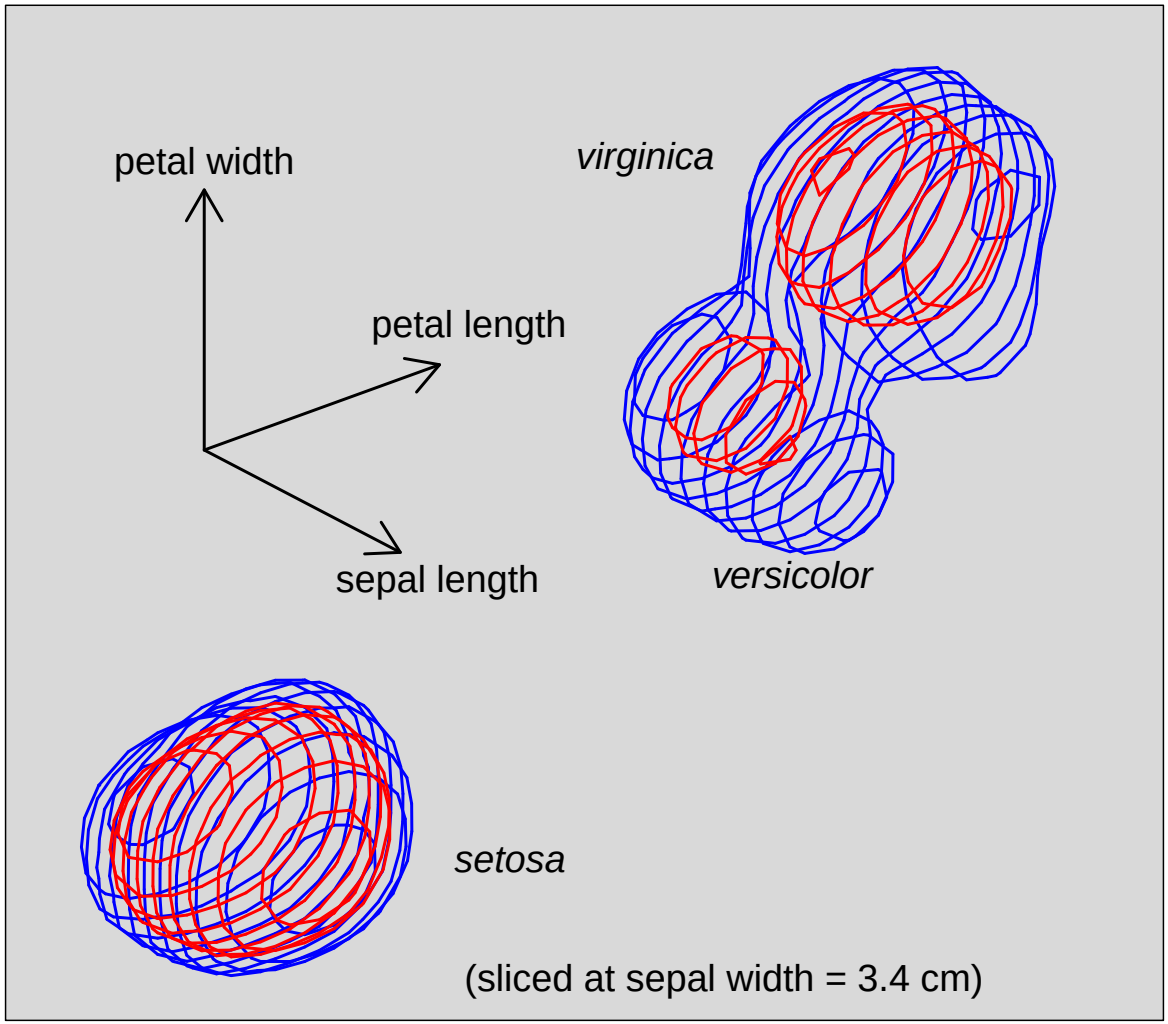


Figure: Two α-level contour surfaces from a slice of a five-dimensional smoothed histogram (ASH) for $\alpha = 4\%$ *and* 10%.

Detailed slices of previous figure

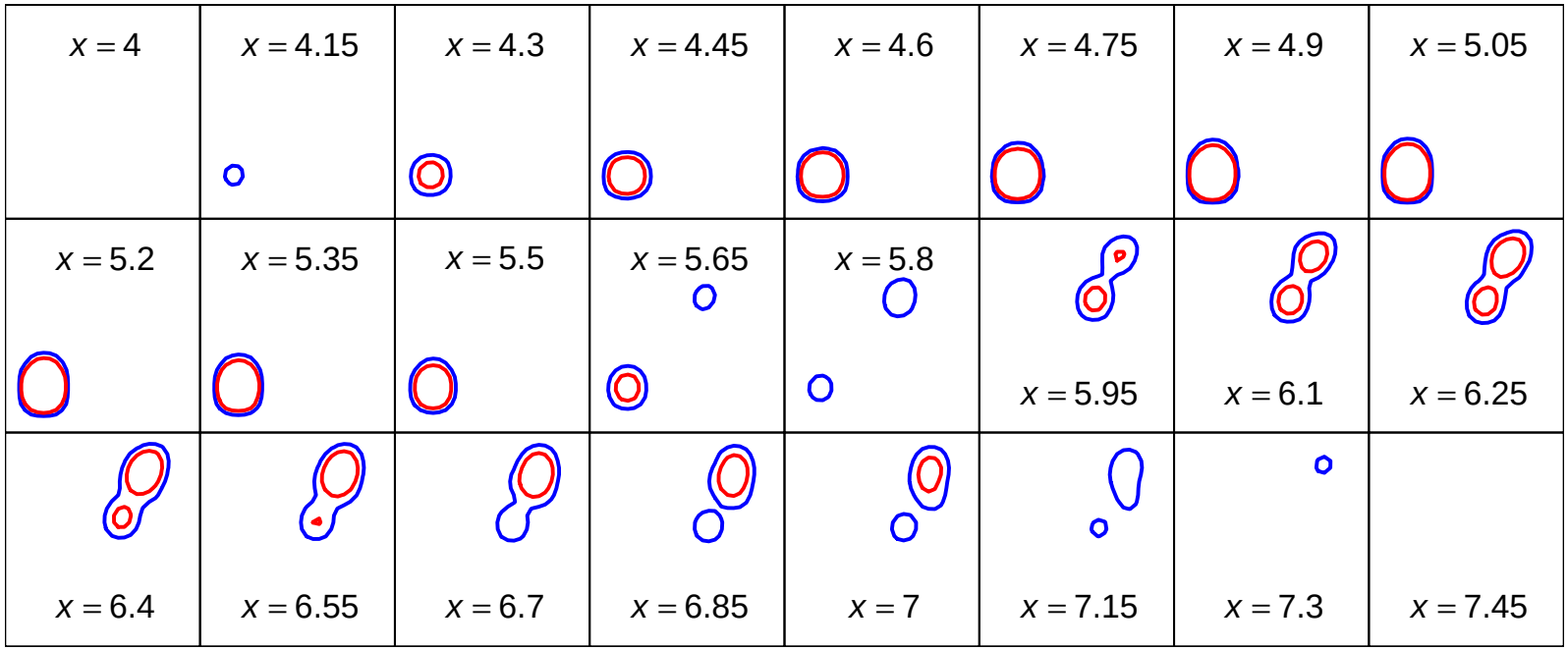


Figure: A detailed breakdown of the 3-D contours taken from the ASH estimate $\hat{f}(x, y, z, t = 3.4)$ as the sepal length, x, ranges from 4.00 to 7.45 cm.

Iris data, excluding sepal width and classes.

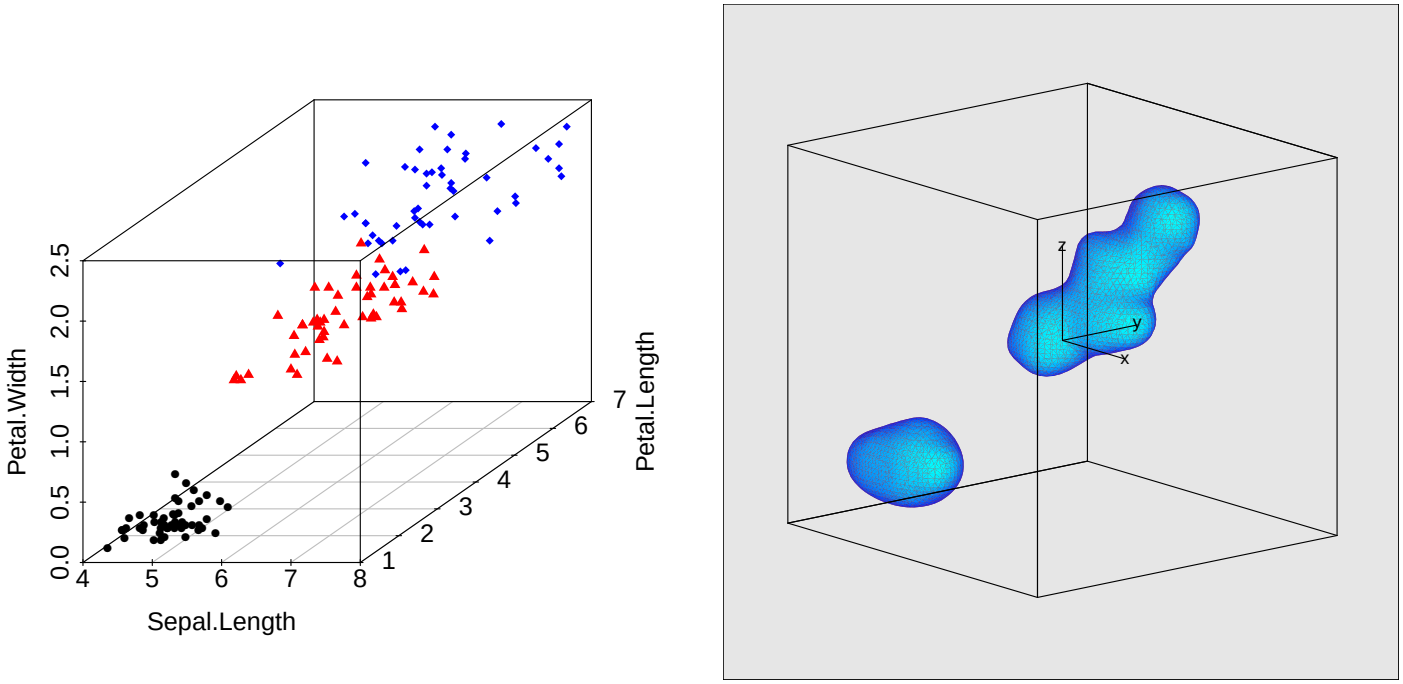


Figure: Analysis of 3 of the 4 *Iris* variables, omitting sepal width entirely. The contour ($\alpha = 0.17$) does not clearly show 3 groups?

Iris data, excluding sepal width and classes.

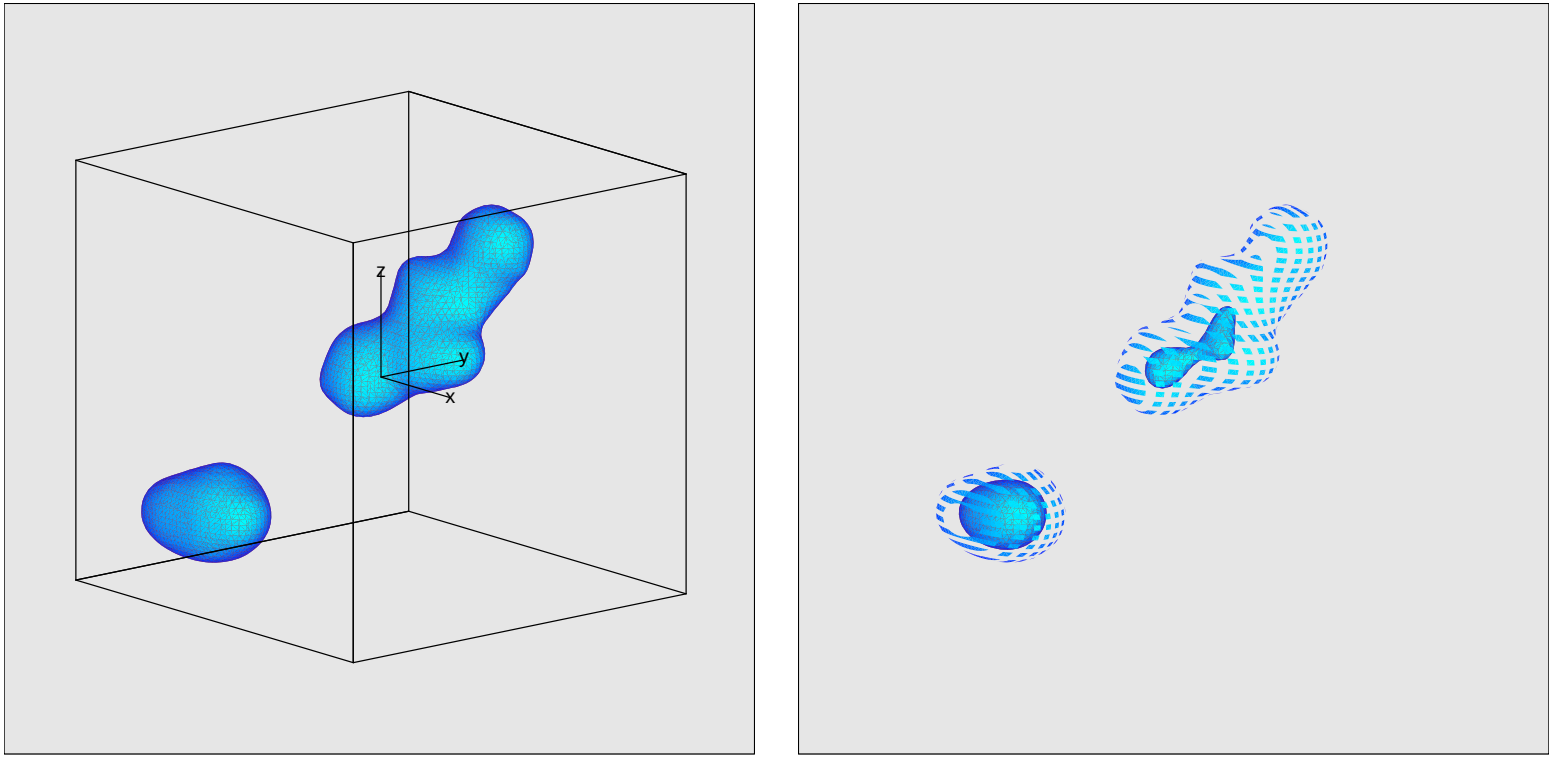


Figure: The higher contour level, $\alpha = 0.44$, perhaps hints at 3 groups?

Slicing with 5 Variables (or more)

- With more than 4 variables, the most appropriate sequence of slicing is not clear.
- With exactly 5 variables, bivariate contours of (x_4, x_5) may be drawn; then a sequence of trivariate slices may be examined tracing along one of these bivariate contours.
- With more than 5 or 6 variables, deciding where to slice at all is a difficult problem because the number of possibilities grows exponentially.
- Will try projection-based methods later

Overview of Contouring and Surface Display ($\Re^2$)

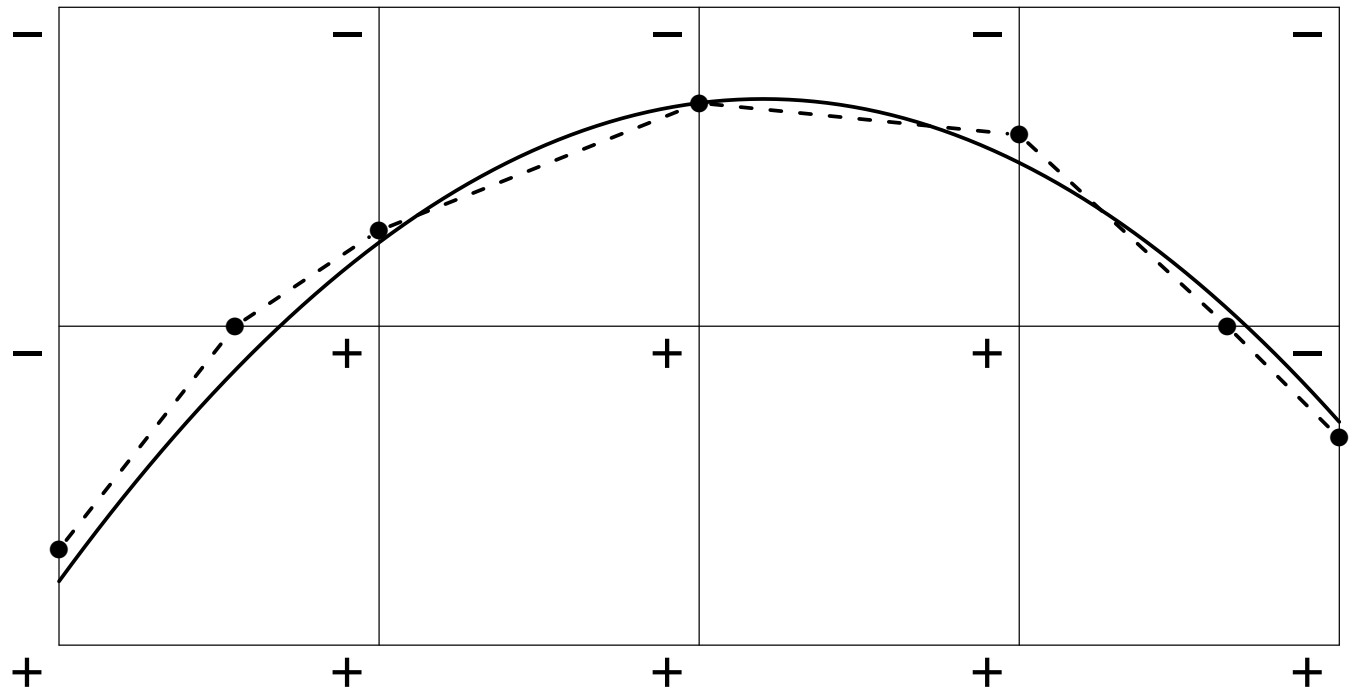


Figure: A portion of a bivariate contour at the $\alpha = 0$ level of a smooth function measured on a regular grid and using linear interpolation (dotted lines).

Contouring in $\Re^3$ with 2-D slices: Stereo View

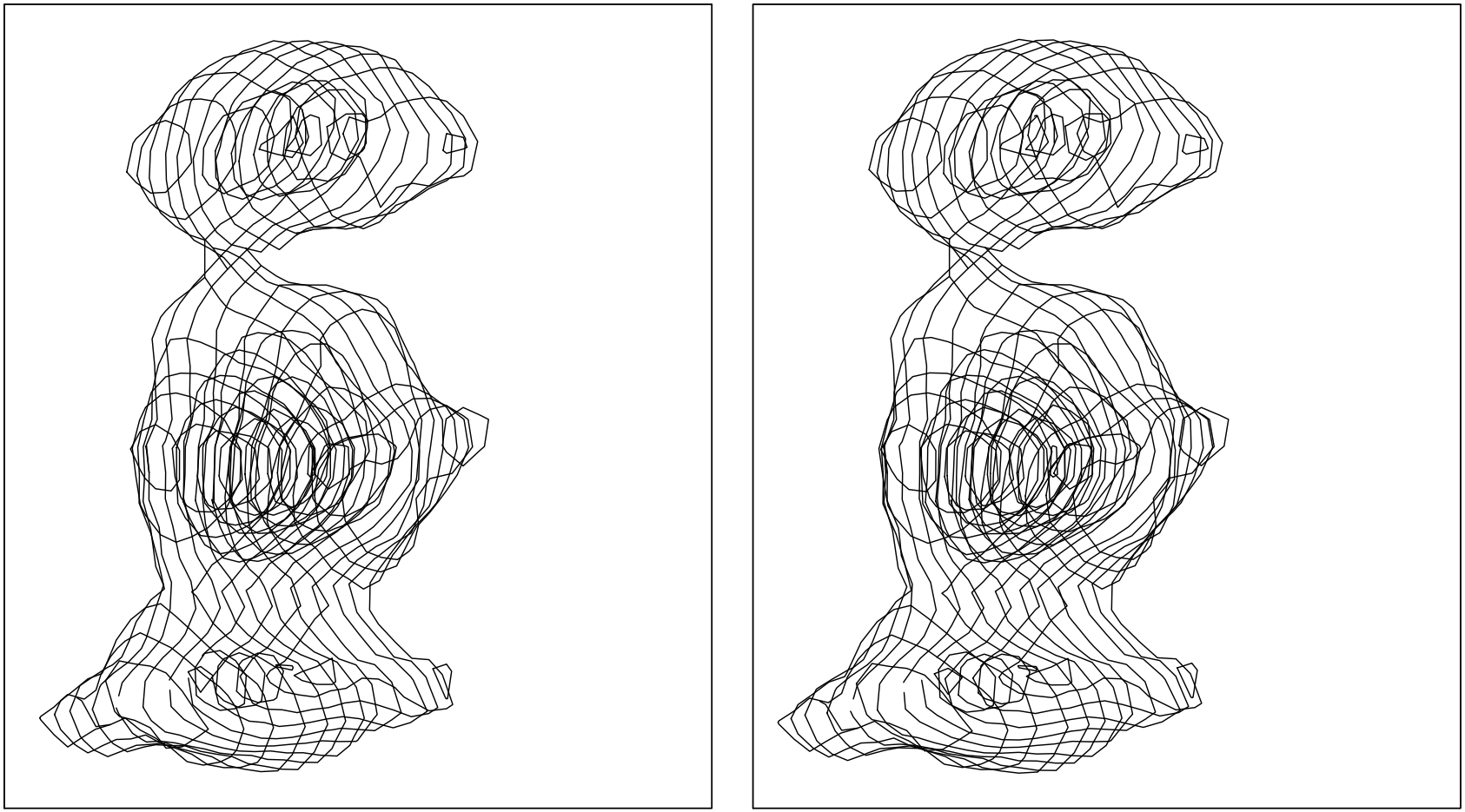


Figure: Simple stereo representation of four 3-D nested shells of the Mount St. Helens earthquake data.

Contouring in $\Re^3$ by Marching Cubes

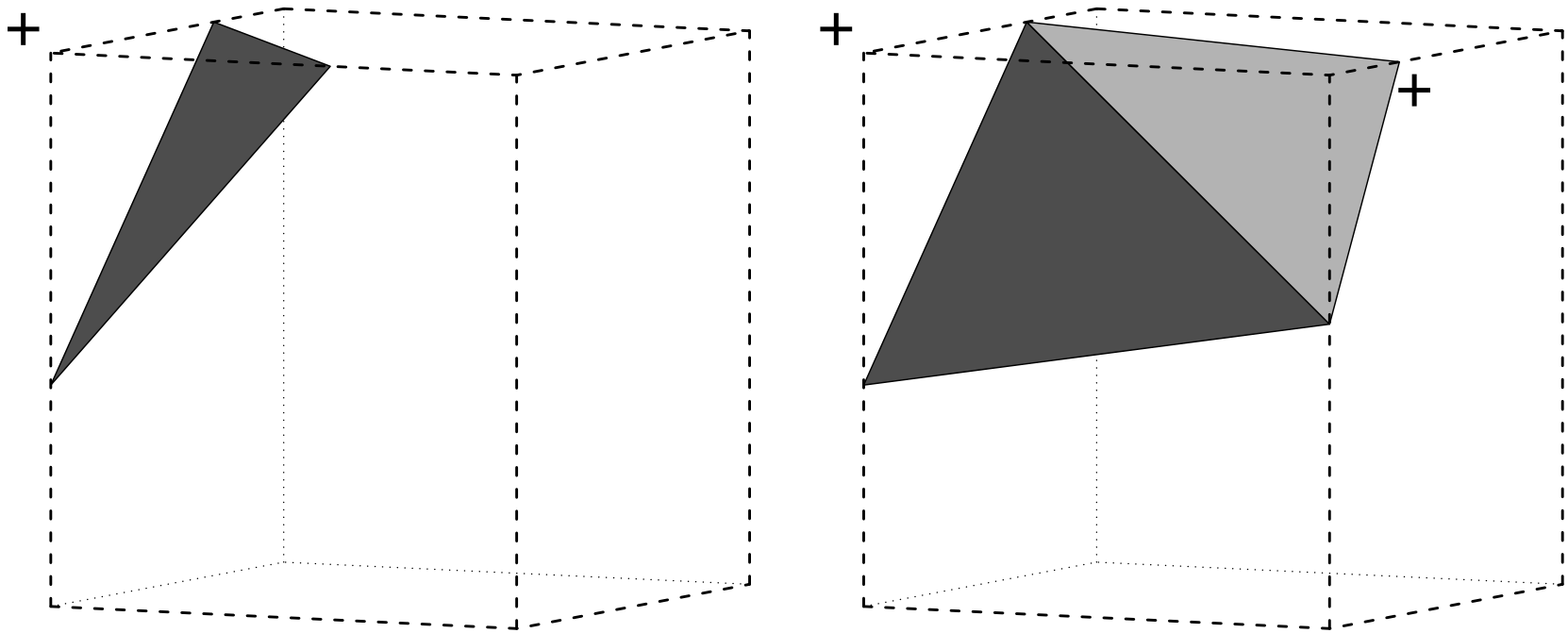


Figure: Examples of marching cube contouring algorithm. The corners with values above the contour level are labeled with a + symbol.

The Normal Bell-Shaped Curve in Higher Dimensions

- Volume of a sphere in $\Re^d$ of radius a:

$$V_d(a) = \frac{a^d \pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$$

- Fraction Hypercube Volume in an Inscribed Hypersphere:

Dim d	1	2	3	4	5	6	7
Fraction	1	0.785	0.524	0.308	0.164	0.081	0.037

- Fraction of volume between hypersphere of radius r and $r-\epsilon$:

$$\frac{V_d(r) - V_d(r-\epsilon)}{V_d(r)} = \frac{r^d - (r-\epsilon)^d}{r^d} = 1 - \left(1 - \frac{\epsilon}{r}\right)^d \xrightarrow[d\to\infty]{} 1$$

(Hence all volume concentrated at hyper-surface.)

Tail Probabilities of Multivariate Normal

Table: Probability Mass *Not* in the "Tail" of a Multivariate Normal Density, i.e., inside the 1% contour, $S_{0.01}$

d	1	2	3	4	5	6	7
$1000p$	998	990	973	944	899	834	762

d	7	8	9	10	15	20
$1000p$	762	675	582	488	134	20

Distance from Origin of Normal Data

- If $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{I}_d)$, then the origin is "most likely"
- All directions are equally likely by symmetric
- The distance (squared) of a point from the origin

$$D^2 = \sum_{j=1}^{d} X_j^2 \sim \chi(d)$$

Hence, a little approximation gives

$$D \approx N\left(\sqrt{d}, \tfrac{1}{2}\right)$$

For example, for $X \in \Re^{100}$, 99% of the data points satisfy

$$8.2 < D < 11.8 \quad \text{surprising!!??}$$

Distribution of point closest to the origin

$$\begin{aligned}\Pr(D \leq c) &= 1 - \Pr(D > c) = 1 - \Pr(D_1 > c, D_2 > c, \ldots, D_n > c) \\ &= 1 - \Pr(D_1 > c)^n = 1 - \Pr(D_1^2 > c^2)^n = 1 - \Pr(\chi_d^2 > c^2)^n \\ &= 1 - \left(1 - \Pr(\chi_d^2 \leq c^2)\right)^n .\end{aligned}$$

Thus applying Leibniz's rule,

$$f_D(c) = \frac{d}{dc}\Pr(D \leq c) = n\left(1 - \Pr(\chi_d^2 \leq c^2)\right)^{n-1} \times 2c f_{\chi_d^2}(c^2)$$

Examples (Implications for Biometrics?)

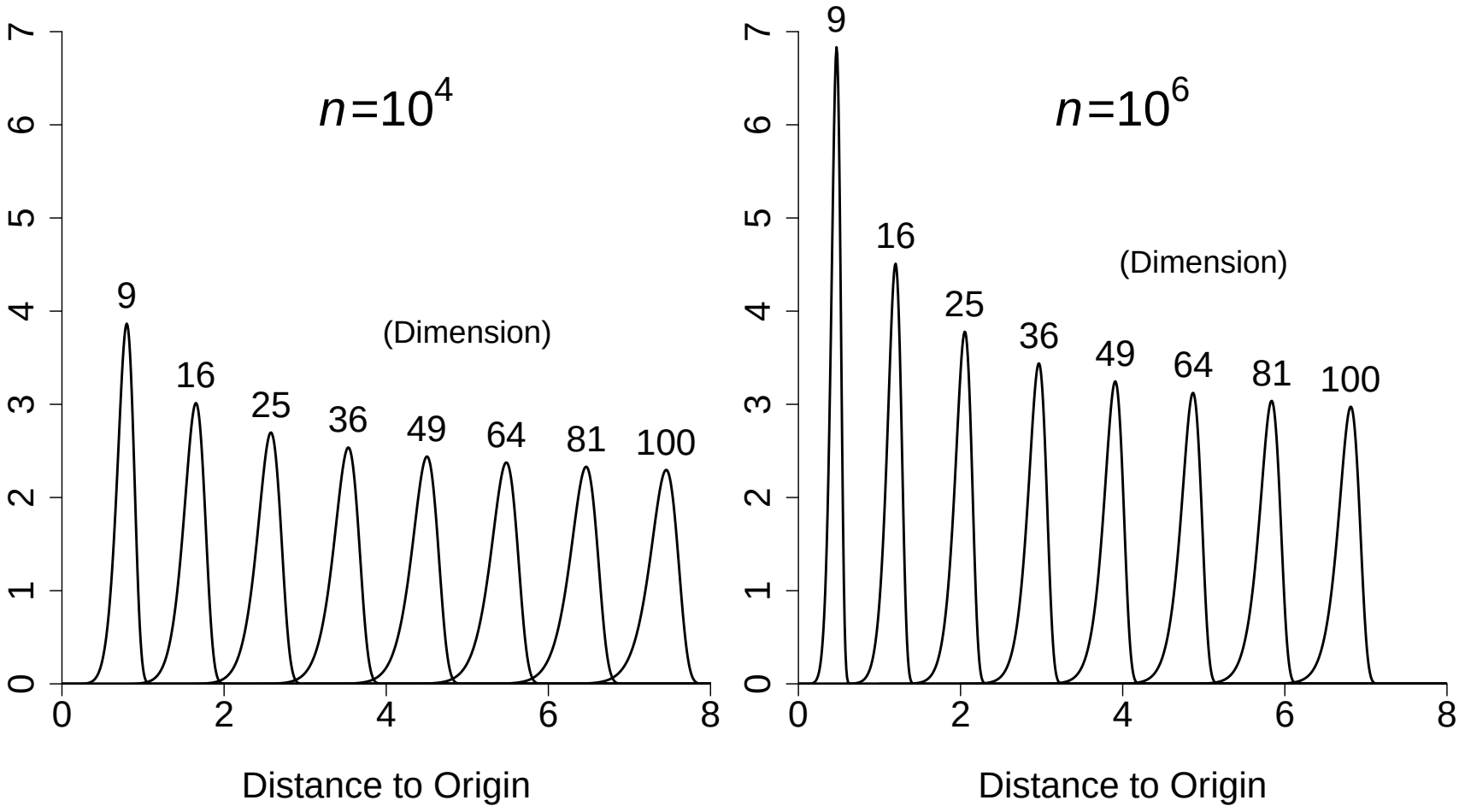


Figure: Densities of distance of closest point to the origin for sample sizes $n = 10^4$ and 10^6, for various dimensions $9 \leq d \leq 100$.

Chapter II: Nonparametric Estimation Criteria

- The focus of nonparametric estimation is different from that of parametric estimation $f(\cdot|\theta)$:

$$\theta \text{ versus the entire density function } f(\cdot)$$

- A good value $\hat{\theta}$ should result in $f(\cdot|\hat{\theta}) \approx f(\cdot) = f(\cdot|\theta)$
- Pearson-Fisher debate on the problem of **specification** and **estimation**
- An incorrectly specified pdf results in a bias that will not disappear as $n \to \infty$ (curse of optimality/robustness)
- nonparametric methods eliminate the need for model specification — loss of efficiency need not be too great

Estimation of the Cumulative Distribution Function

- $F(x) = \Pr(X \leq x)$
- *empirical cumulative distribution function* (ecdf), defined as

$$F_n(x) = \frac{\#\{x_i \leq x\}}{n} = \frac{\#x_i \in (-\infty, x]}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(x_i)\,,$$

- since $nF_n(x)$ is a binomial random variable with $p = F(x)$,

$$\mathrm{E}F_n(x) = \mathrm{E}I_{(-\infty,x]}(X) = 1 \times \Pr(X \in (-\infty, x]) = F(x)$$

- Furthermore, since $nF_n(x)$ is a binomial random variable, $B(n, p)$, with $p = F(x)$, then

$$\mathrm{Var}\{F_n(x)\} = p(1-p)/n \qquad \text{so } F_n(x) \text{ is best !!}$$

- Prefer cdf or pdf??? If cdf, all done !!! Answer clearer in $\Re^d$.

Empirical CDF in 1 Dimension

Can you see the major feature in these data?

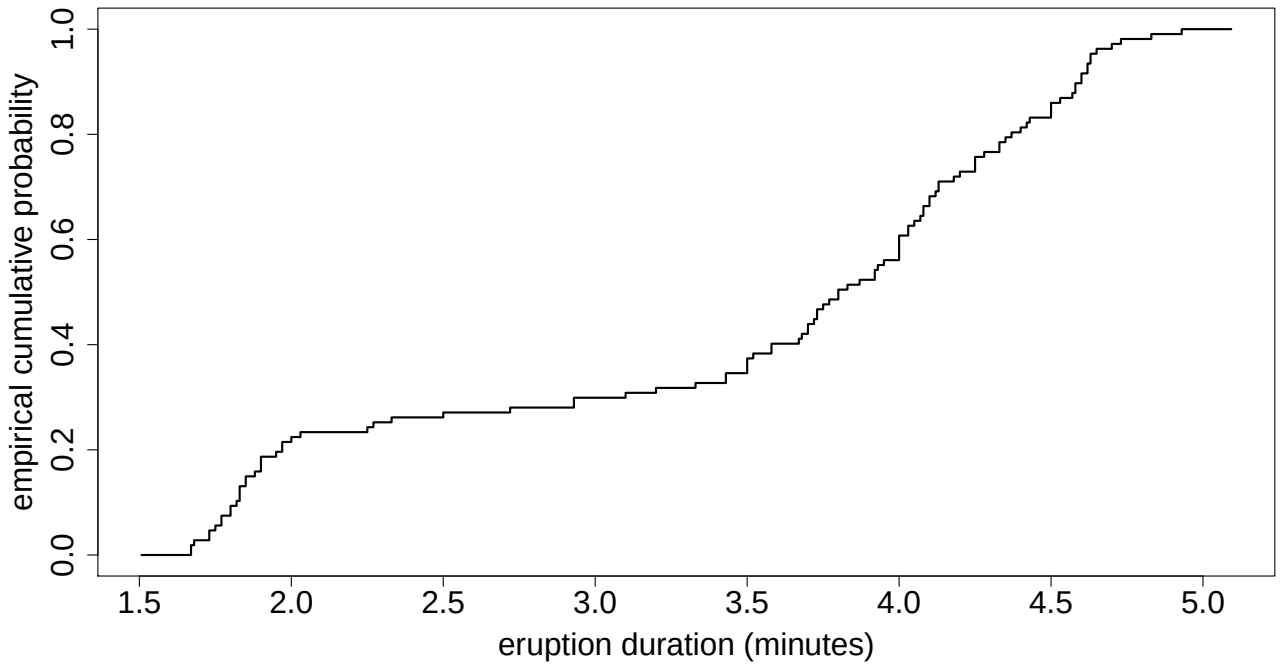


Figure: Empirical cumulative distribution function of the Old Faithful geyser dataset.

Empirical CDF in 2 Dimensions

Can you see the major feature in these data?

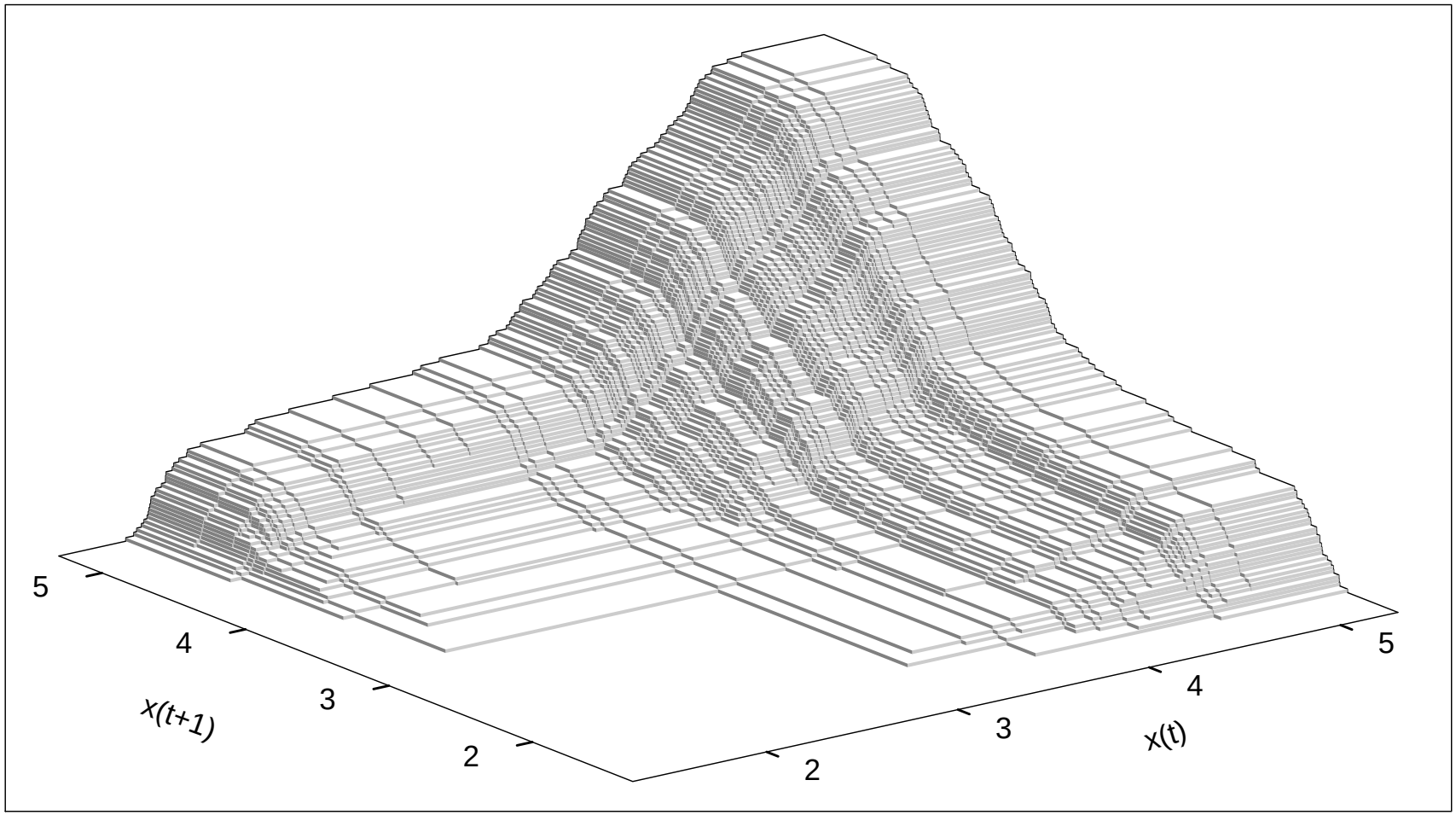


Figure: Empirical bivariate cdf of the lagged Old Faithful data $\{x_t, x_{t+1}\}$

Direct Nonparametric Estimation of the Density

- Conclusion: pdf much easier to understand graphically than the cdf:

$$f(x) = F'(x)$$

- Use the "best" nonparametric cdf estimator, $F_n(x)$, to obtain *empirical probability density function* (epdf):

$$f_n(x) = \frac{d}{dx} F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

where $\delta(t)$ is the Dirac delta function.

- this is the (discrete) "bootstrap" pdf — but graphically ugly
- Question: Does there exist a "best" minimum variance unbiased estimator of $f(x)$? Answer: Rosenblatt (1956): NO
- You may be familiar with some popular biased estimators: ridge regression, Stein estimators,...

Example: Thickness of U.S. pennies over 50 years.

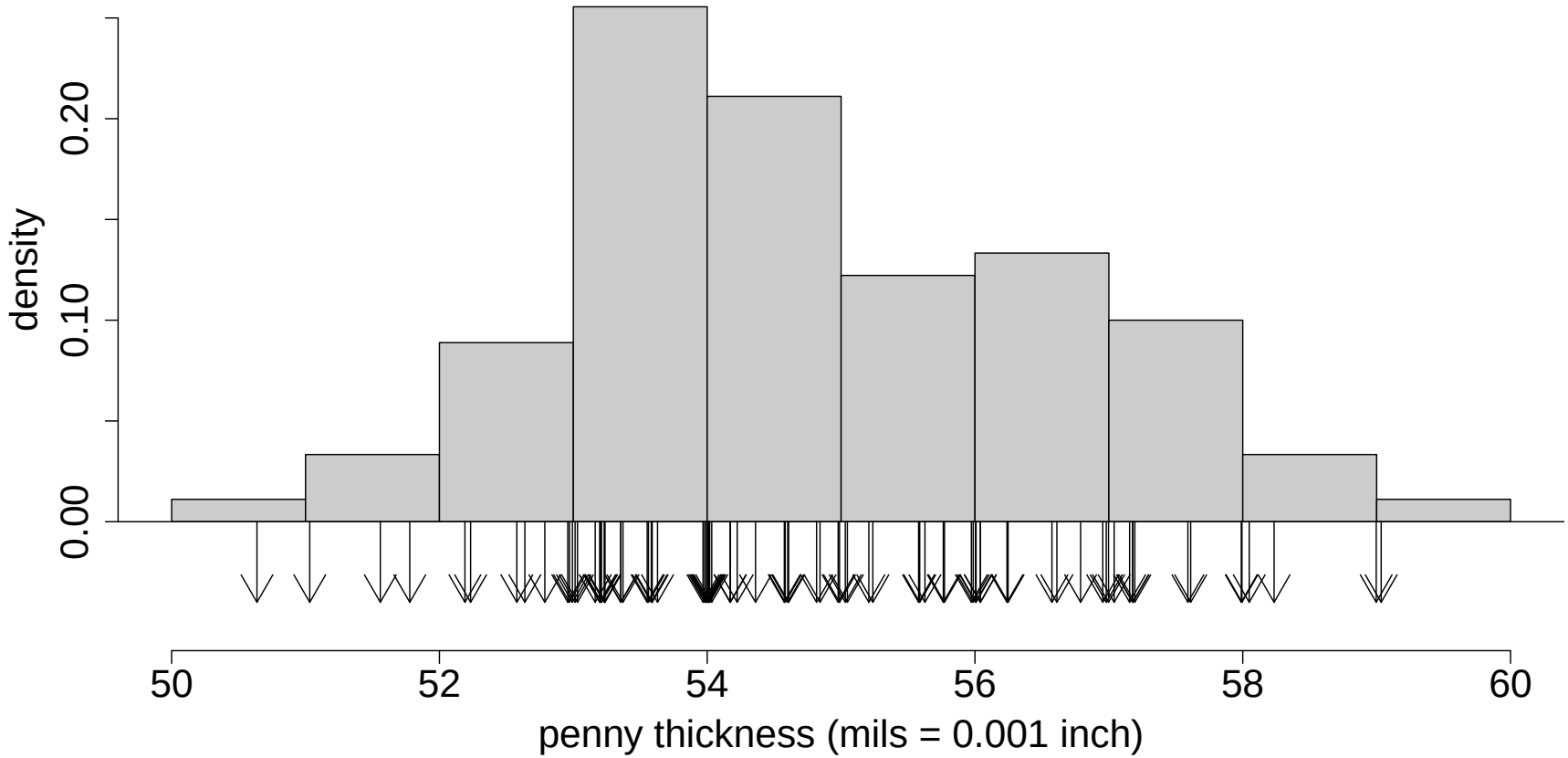


Figure: A histogram and empirical pdf (pointing down) of the U.S. penny thickness dataset.

Error Criteria for Density Estimates

- "optimality" is not an absolute concept, but intimately linked to the choice of a criterion
- criterion preference is largely subjective
- in the parametric world, use the mean squared error (MSE) criterion:

$$\text{MSE}\{\hat{\theta}\} = \text{E}[\hat{\theta} - \theta]^2 = \text{Var}\{\hat{\theta}\} + \text{Bias}^2\{\hat{\theta}\}$$

- can use this same criterion to evaluate $\hat{f}(x)$ for **fixed** x:

$$\text{MSE}\{\hat{f}(x)\} = \text{E}[\hat{f}(x) - f(x)]^2 = \text{Var}\{\hat{f}(x)\} + \text{Bias}^2\{\hat{f}(x)\}$$

- If do this for every x, then an infinite-dimensional problem!

Global Error Criteria for Density Estimation

- For some the most intuitively appealing global criterion is the L_∞ norm:

$$\sup_x \left| \hat{f}(x) - f(x) \right|$$

- At the other end of the spectrum is the L_1 norm:

$$\int \left| \hat{f}(x) - f(x) \right| \, dx$$

- Neither of these criteria is as easily manipulated as the L_2 norm, which in this context is referred to as the *integrated squared error* (ISE):

$$\text{ISE} = \int \left[\hat{f}(x) - f(x) \right]^2 \, dx$$

- Others: Helling distance, Akaike's informatin criterion, L_p norms, Kullback-Leibler divergence,...

ISE and MISE and IMSE Criteria

- The ISE is a complicated random variable that depends on the true unknown density function, the particular estimator, the sample size, and the particular realization of n points.
- It will be sufficient to examine the average of the ISE over these realizations; that is, the mean of the random variable ISE or *mean integrated squared error* (MISE):

$$\begin{aligned} \mathrm{MISE} \equiv \mathrm{E}\big[\mathrm{ISE}\big] &= \mathrm{E}\left\{ \int \big[\hat{f}(x) - f(x)\big]^2 \, dx \right\} \\ &= \int \mathrm{E}\big[\hat{f}(x) - f(x)\big]^2 \, dx = \int \mathrm{MSE}\{\hat{f}(x)\} \, dx \equiv \mathrm{IMSE}\,, \end{aligned}$$

where the interchange of the integral and expectation operators is justified by an application of Fubini's theorem.

IMSE and MISE Criteria

- The last quantity is the IMSE, which is an abbreviation for the *integrated mean squared error*.
- Thus the MISE error criterion has two different though equivalent interpretations:
  - it is a measure of the average global error;
  - it is also a measure of the accumulated pointwise error.
- Question: Is this equality true or false? (A bit advanced...)

$$\mathrm{E}\left\{\int \left[\hat{f}(x) - f(x)\right]^2 dx\right\} \stackrel{?}{=} \int \mathrm{E}\left[\hat{f}(x) - f(x)\right]^2 f(x)\, dx\,,$$

which is a weighted average MSE criterion.

MISE for Parametric Estimators

- Since MISE is not a familiar criterion, will do a few examples with parametric estimation: usual $O(n^{-1})$ convergence?
- Uniform density $f = U(0, \theta)$, where $\theta = 1$ is estimated by the maximum likelihood estimator $\hat{\theta} = x_{(n)}$, the nth-order statistic.
- $\text{ISE} = \left(\frac{1}{x_{(n)}} - 1\right)^2 \cdot x_{(n)} + (0-1)^2 \cdot \left(1 - x_{(n)}\right) = \frac{1}{x_{(n)}} - 1$
- Since $f(x_{(n)}) = n x_{(n)}^{n-1}$, for $0 < x_{(n)} < 1$, it follows that

$$\text{MISE} = \int_0^1 \left(\frac{1}{x_{(n)}} - 1\right) \cdot n\, x_{(n)}^{n-1}\, dx_{(n)} = \frac{1}{n-1} = O(n^{-1})$$

Best Estimators of $U(0, \theta)$

- Well known that the unbiased estimator of θ is

$$\hat{\theta} = \frac{n+1}{n} x_{(n)}$$

- If consider estimators of the form $\hat{\theta} = c\, x_{(n)}$, then

$$\text{MISE}(c) = \begin{cases} n/[(n-1)c] - 1 & c < 1 \\ \left[2 - nc^{n-1} + (n-1)c^n\right] / [(n-1)c^n] & c > 1 \end{cases}$$

for which

$$c^* = 2^{1/(n-1)} \approx 1 + \frac{\log(2)}{n} = \frac{n + 0.693}{n}$$

General Parametric MISE Method

- First use of Taylor Series:

$$\mathsf{I}(\hat{\theta}) = \int \left[f(t|\hat{\theta}) - f(t|\theta) \right]^2 \, dt = \sum_k \frac{1}{k!} (\hat{\theta} - \theta)^k \, \mathsf{I}^{(k)}(\theta)$$

- Easy to check that $\boldsymbol{I}(\theta) = 0$ and $\boldsymbol{I}'(\theta) = 0$; hence,

$$\mathsf{AMISE}(\hat{\theta}) = \mathsf{E}[\mathsf{I}(\hat{\theta})] = \frac{1}{2} \, \mathsf{var} \, (\hat{\theta}) \, \mathsf{I}''(\theta) + \cdots .$$

where A = asymptotic.

Parametric AMISE for $N(\mu, \sigma^2)$ pdf

- Start with a bivariate Taylor Series of the estimator $N(\bar{x}, s^2)$ around $N(0, 1)$ (see book for details)
- For standard normal, the result is that

$$\text{AMISE}\{\phi(\bar{x}, s^2)\} = \text{AMISE}\{\phi(\bar{x}, 1)\} + \text{AMISE}\{\phi(0, s^2)\}$$

- The numerical result

$$\text{AMISE}\{\phi(\bar{x}, s^2)\} = \frac{1}{4n\sqrt{\pi}} + \frac{3}{16n\sqrt{\pi}} = \frac{7}{16n\sqrt{\pi}}$$

- This is $O(n^{-1})$ as usual
- Thus knowing μ is better than knowing σ^2. Why?
- Notice that $\bar{x}$ and s^2 are maximum-likelihood estimators, and not derived from the AMISE criterion

Data-Based Parametric Estimation Criteria

- Where does maximum likelihood come from?
- Let the parametric estimator will be denoted by $f_\theta(x)$, but the true density will be denoted by $g(x)$. This reflects the fact that the model may not be (exactly) correct, only an approximation.
- Answer: MLE is an unbiased estimate of minimizer of the Kullback-Leibler divergence

$$
\begin{aligned}
d_{KL}(f_\theta, g) &= \int g(x) \log \frac{g(x)}{f_\theta(x)} dx \geq 0 \\
&= \int g(x) \log g(x)\, dx - \int g(x) \log f_\theta(x)\, dx \\
&= \textit{constant} - E[\log\, f_\theta(X)]\,; \quad \text{thus an estimate is} \\
\hat{d}_{KL}(f_\theta, g) &= \text{constant} - \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(x_i)\,,
\end{aligned}
$$

which when minimized, is exactly the MLE criterion.

Is There a Data-based Estimator for Hellinger distance?

- Hellinger distance is defined by

$$d_H(f_\theta, g)^2 = \int \left[\sqrt{f_\theta(x)} - \sqrt{g(x)}\right]^2 dx \geq 0\,.$$

- There is no obvious quantity to substitute for the unknown $g(x)$, save a nonparametric estimate such as a histogram.
- Aside from the computational complexity of performing the numerical minimization, different choices of the histogram would result in different estimates of θ. And no good for $\Re^d$.
- Same problems if focus on the L_1 criterion, $\int |f_\theta(x) - g(x)|\, dx$
- Dimensional analysis shows both Hellinger and L_1 are dimensionless, which would be nice to have.

Data-Based Criterion for integrated squared error

- Expanding the ISE into the sum of 3 integrals:

$$\begin{aligned}\mathrm{ISE}(\theta) &= \int [f_\theta(x) - g(x)]^2 \, dx \\ &= \int f_\theta(x)^2 \, dx - 2 \int f_\theta(x) \, g(x) \, dx + \int g(x)^2 \, dx \\ &= \int f_\theta(x)^2 \, dx - 2 \, \mathrm{E} f_\theta(X) + \text{constant}\end{aligned}$$

- Assuming that the model is square integrable in a convenient form, and choosing the obvious unbiased estimator for the expectation term, we arrive at the fully data-based criterion

$$\hat{\theta} = \arg\min_\theta \widehat{\mathrm{ISE}}(\theta) = \arg\min_\theta \left[\int f_\theta(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} f_\theta(x_i) \right]$$

- Called the L_2E estimator (Scott, 2001, Technometrics, 43, 274–285). Works equally well in $\Re^d$.

Example of L_2E

- Consider fitting a normal density, $\phi(x|\mu, \sigma^2)$, to the Rayleigh data ($n = 15$) that measured the weight of a standard volume of nitrogen (Tukey, 1977).
- The criterion is given by

$$\hat{\theta} = \arg\min_{\theta} L_2E(\theta) = \arg\min_{\theta} \left[\frac{1}{2\sqrt{\pi}\sigma} - \frac{2}{n} \sum_{i=1}^{n} \phi(x_i|\mu, \sigma^2) \right]$$

- Also considered a 4-parameter mixture model for $f_\theta(x)$

$$f_\theta(x) = w\,\phi(x|\mu_1, \sigma^2) + (1 - w)\,\phi(x|\mu_2, \sigma^2)$$

Example continued (Discovery of the nobel gas argon)

Lord Rayleigh's recognition of the 2nd cluster resulted in his winning the 1904 Nobel Prize in Physics.

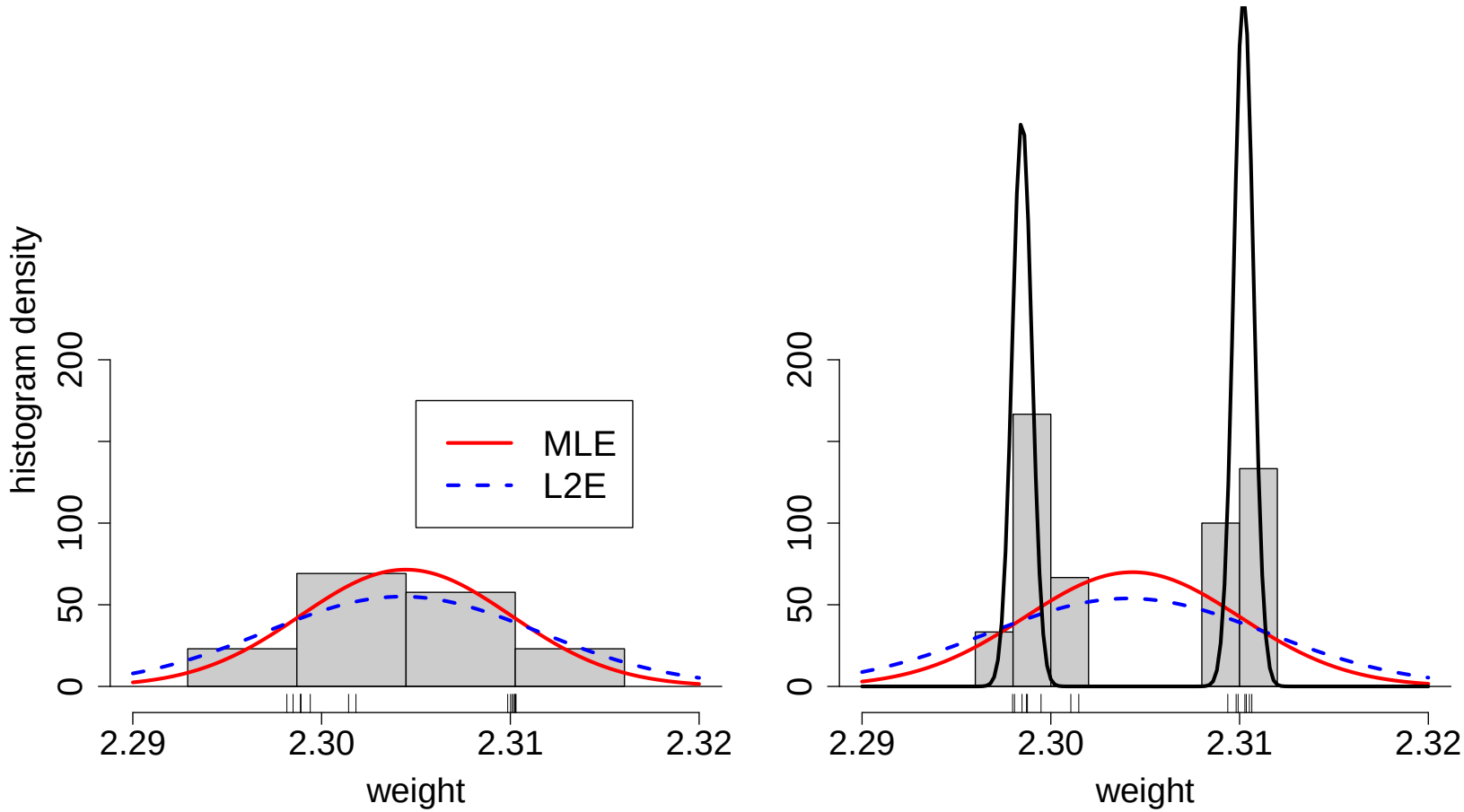


Figure: (L) Histogram with MLE and L_2E normal fits to the Rayleigh data. (R) L_2E normal mixture fit to blurred data with common variance.

Nonparametric Families of Distributions

Pearson Family of Distributions: Pearson's influence on nonparametric density estimation is twofold:

- He coined the word *histogram*
- He introduced and studied density functions that are solutions to the differential equation

$$\frac{d \log f(x)}{dx} = \frac{x - a}{b + cx + dx^2}$$

- Identified 7 types of solutions to this equation, depending on the roots of the denominator and which parameters were 0.
- Interestingly, this class contains most of the important classical distributions: normal, Student's t, Beta, and Snedecor's F.
- Pearson proposed using the first 4 sample moments to estimate the unknown constants (a, b, c, d); i.e. data-based

Nonparametric Families of Distributions (continued)

- Johnson family
- Marshall and Olkin
- Question: **When is an estimator nonparametric?**
  - Is the Pearson family nonparametric? (4 parameters?)
  - It has proven surprisingly difficult to formulate a working definition for what constitutes a nonparametric density estimator.
  - A heuristic definition may be proposed based on the necessary condition that the estimator "work" for a "large" class of true densities.
  - Does a nonparametric estimator require an infinite number of parameters? (Answer later.)
  - A histogram *is a nonparametric density estimator*. But a histogram has only 1 parameter (the bin width)?