

Visualization, Feature Discovery, and Uncertainty*

David W. Scott, Dept. Statistics, Rice Univ., Houston

CATS Visualization Workshop
March 3, 2005 National Academies

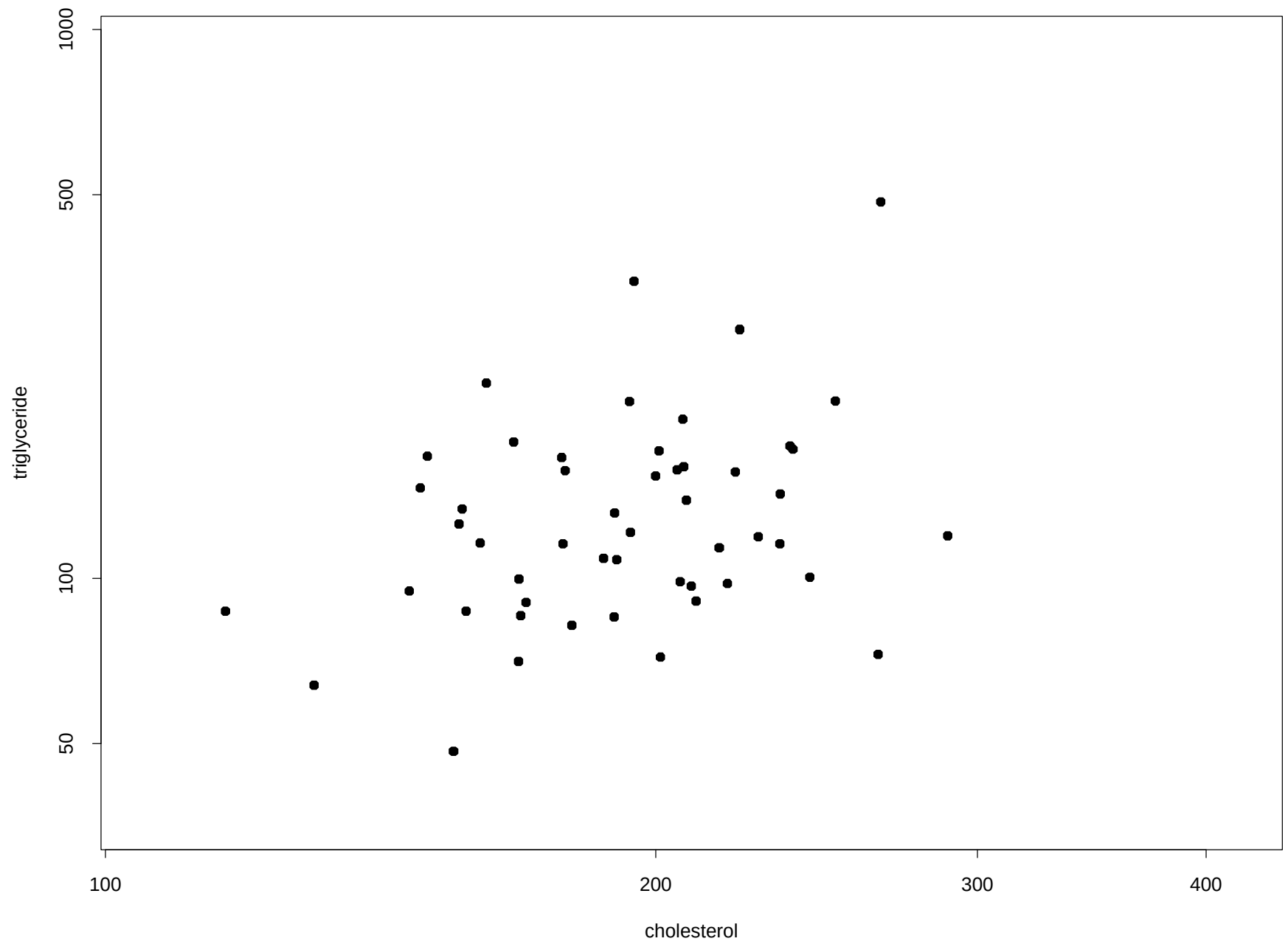
*Research supported in part by NSF grant DMS 02-04723 (non-parametric methodology) and NSF contract EIA-9983459 (digital government). Collaborators Bill Szewczyk and HG Sung.

1 Lipids: Is it normal?

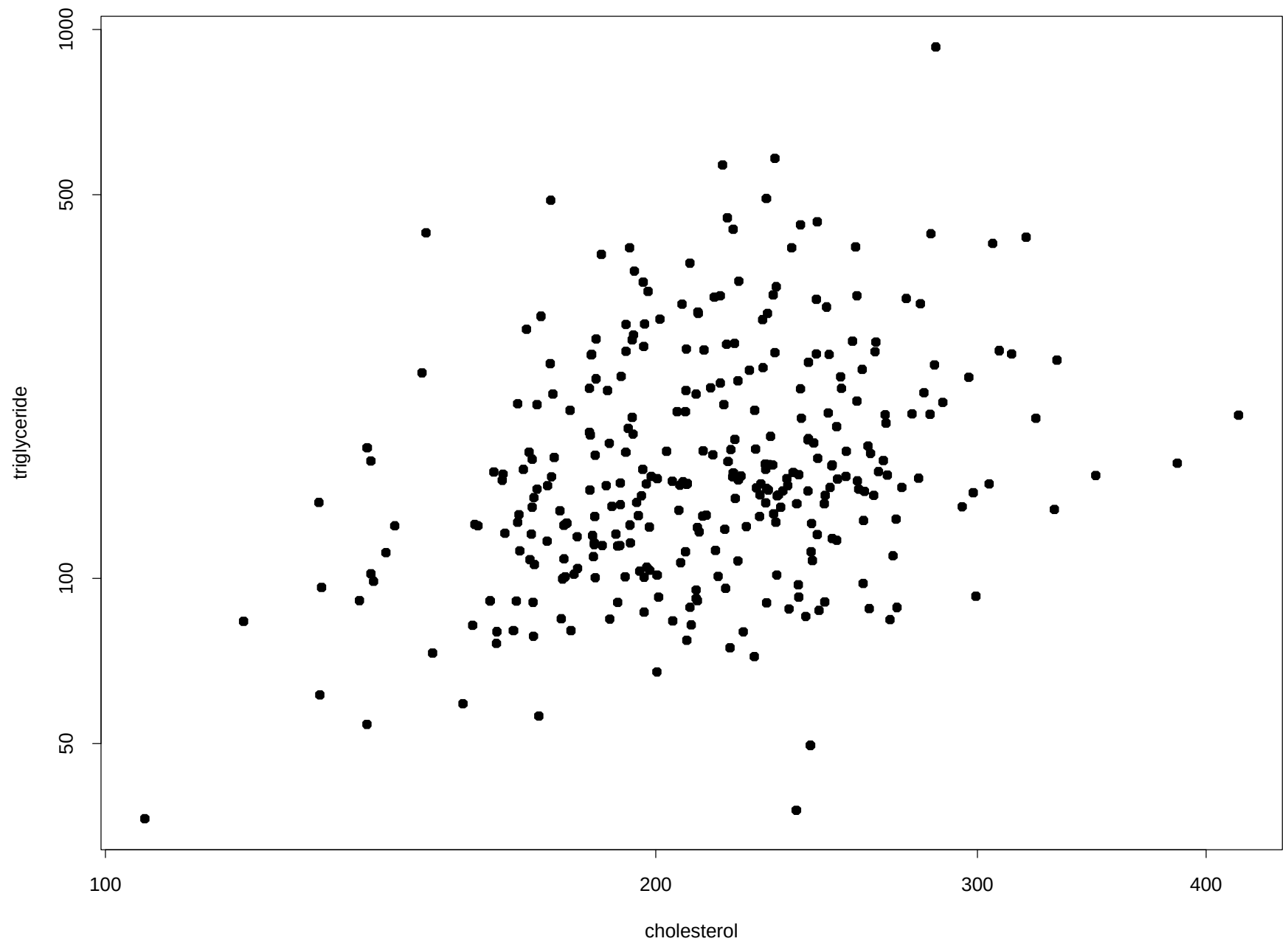
- remarks mixed with examples of nonparametric density estimation in practice (indirect data visualization)
- $f(x)$ conveys “all” information — visualize $f(x)$ of derived quantities
- as editor of JCGS past 4 yrs, noticeable decrease in number of graphics submissions
- scientists employ a few graphical types 99% of the time
- some basic tenets
- data $\neq$ structure
- smoothing $\Rightarrow \hat{f}(x)$ which contains structure (if can spot it)
- exploring feature spaces: easy structure (anything works) vs. subtle structure (smoothing helps)

- kernel estimates (ASH) provide an excellent summary of high-D structure (even if quite biased according to theory)

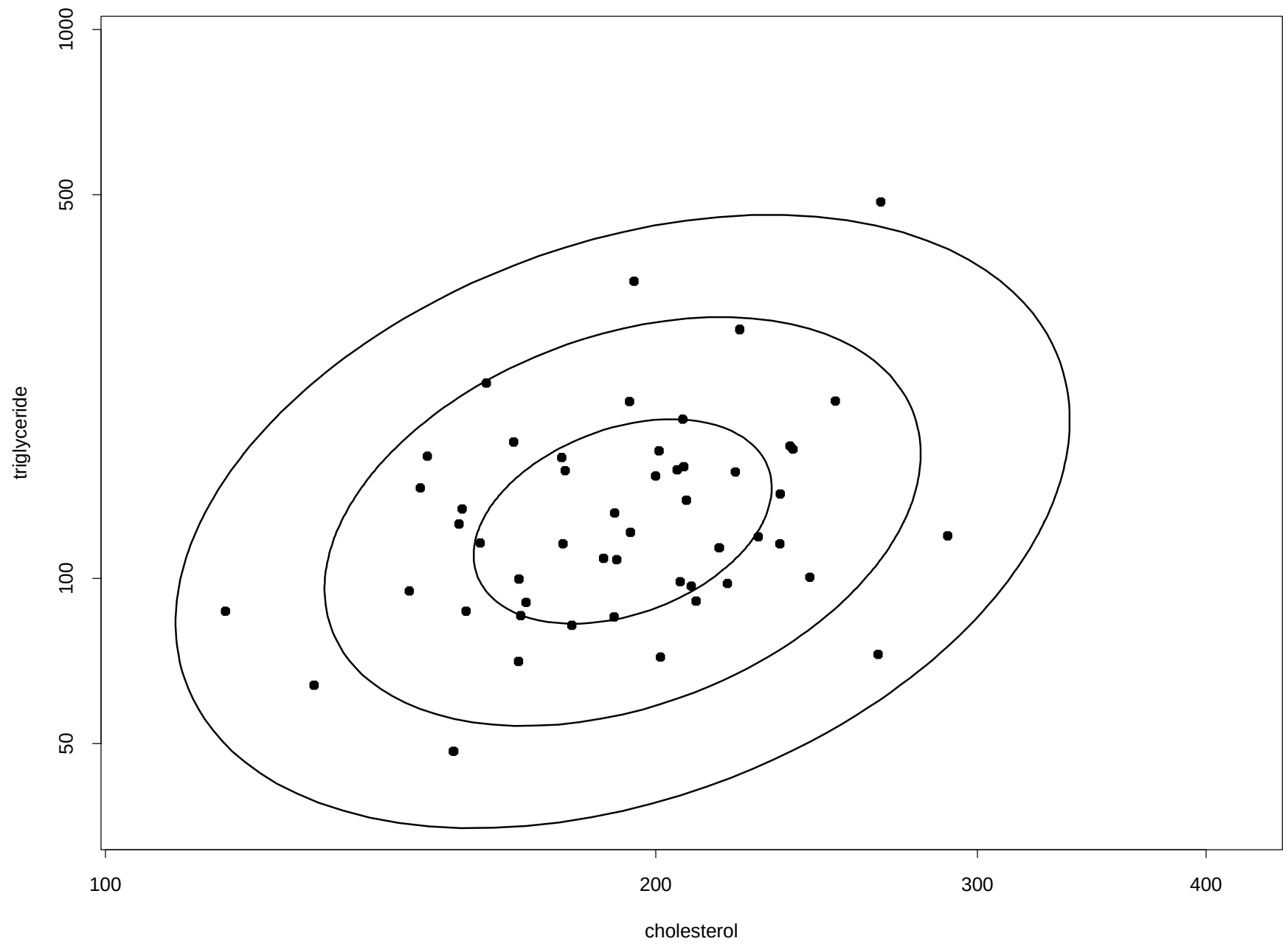
normal n = 51



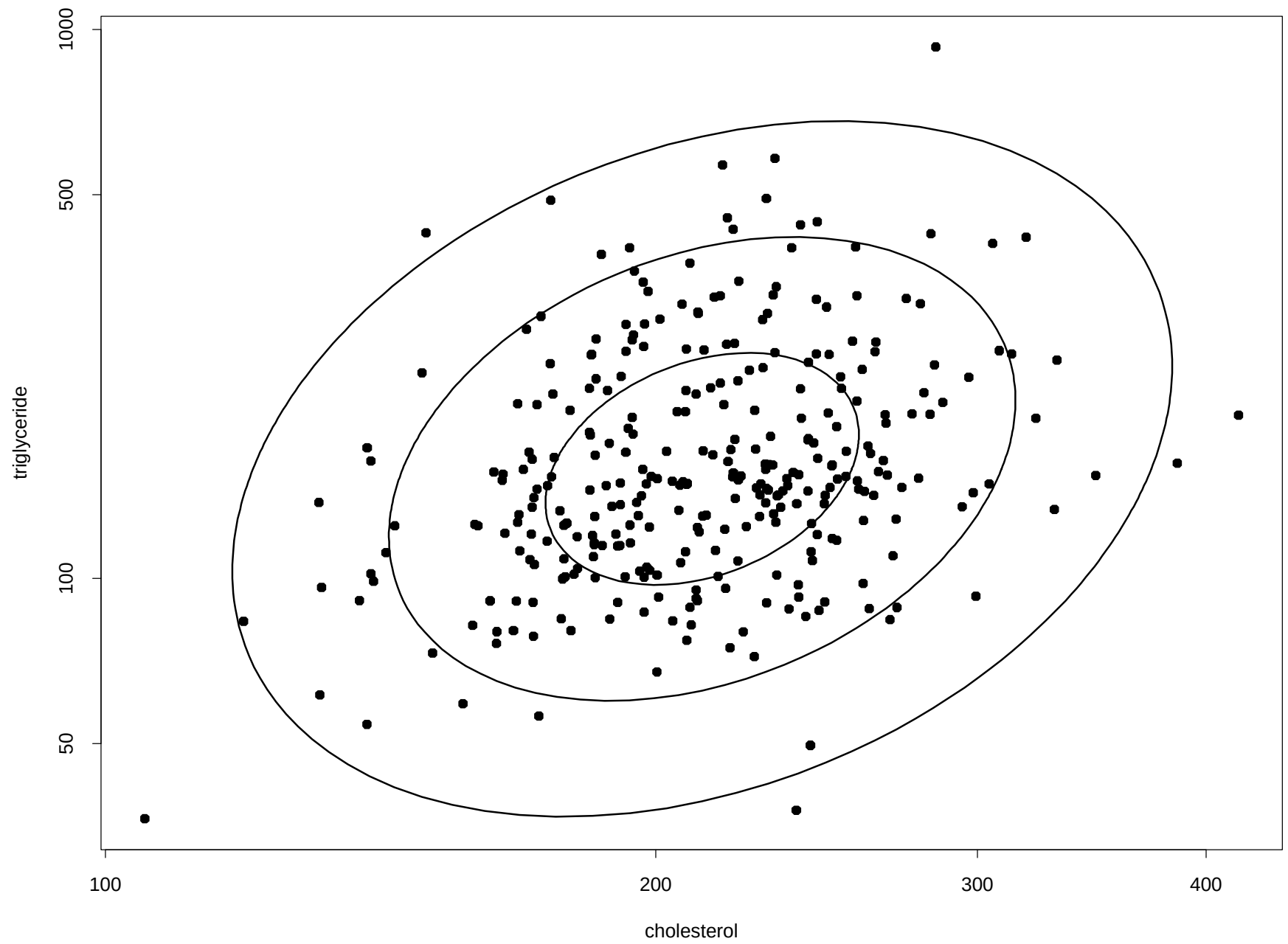
diseased n = 320



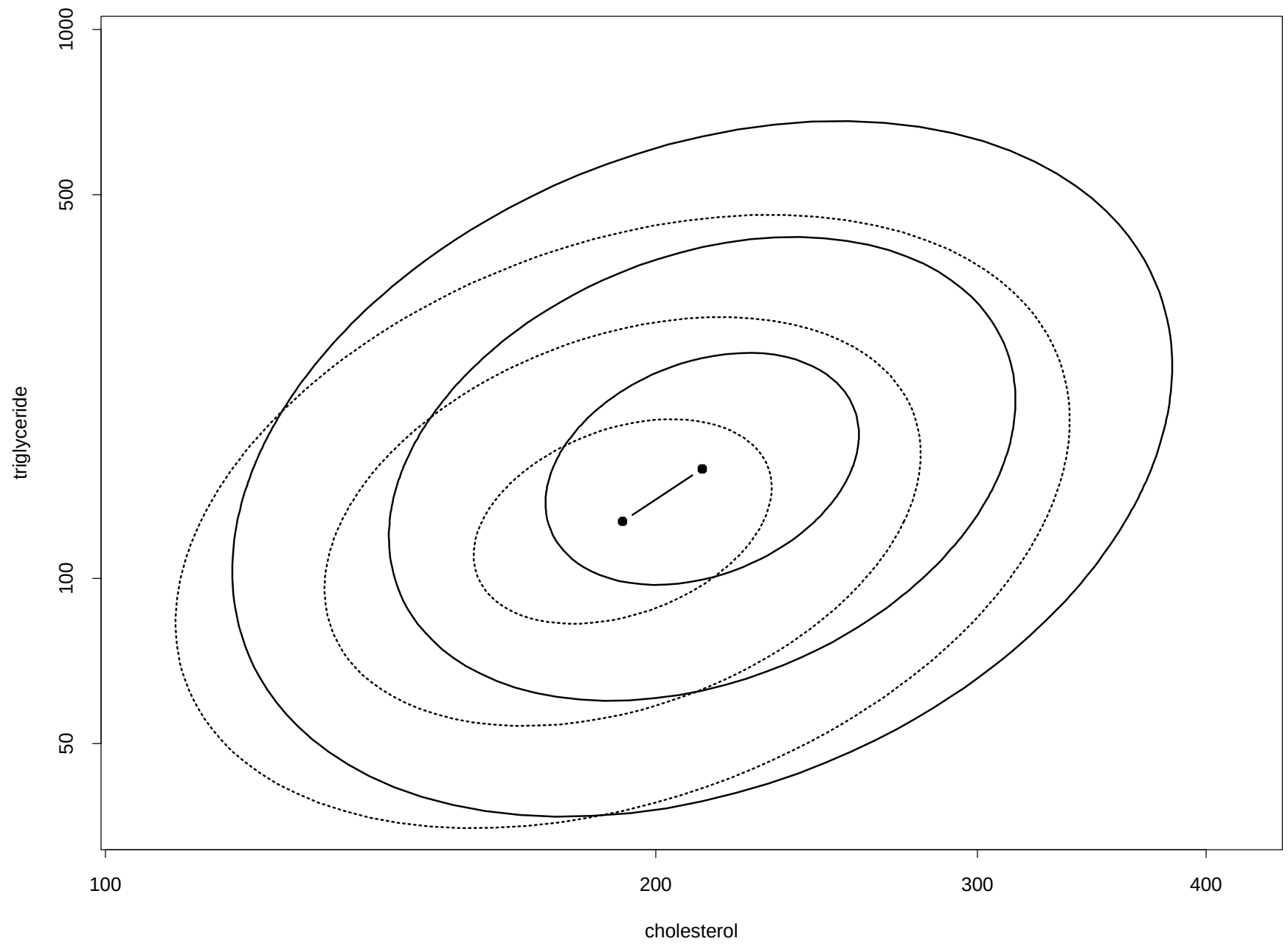
normal n = 51



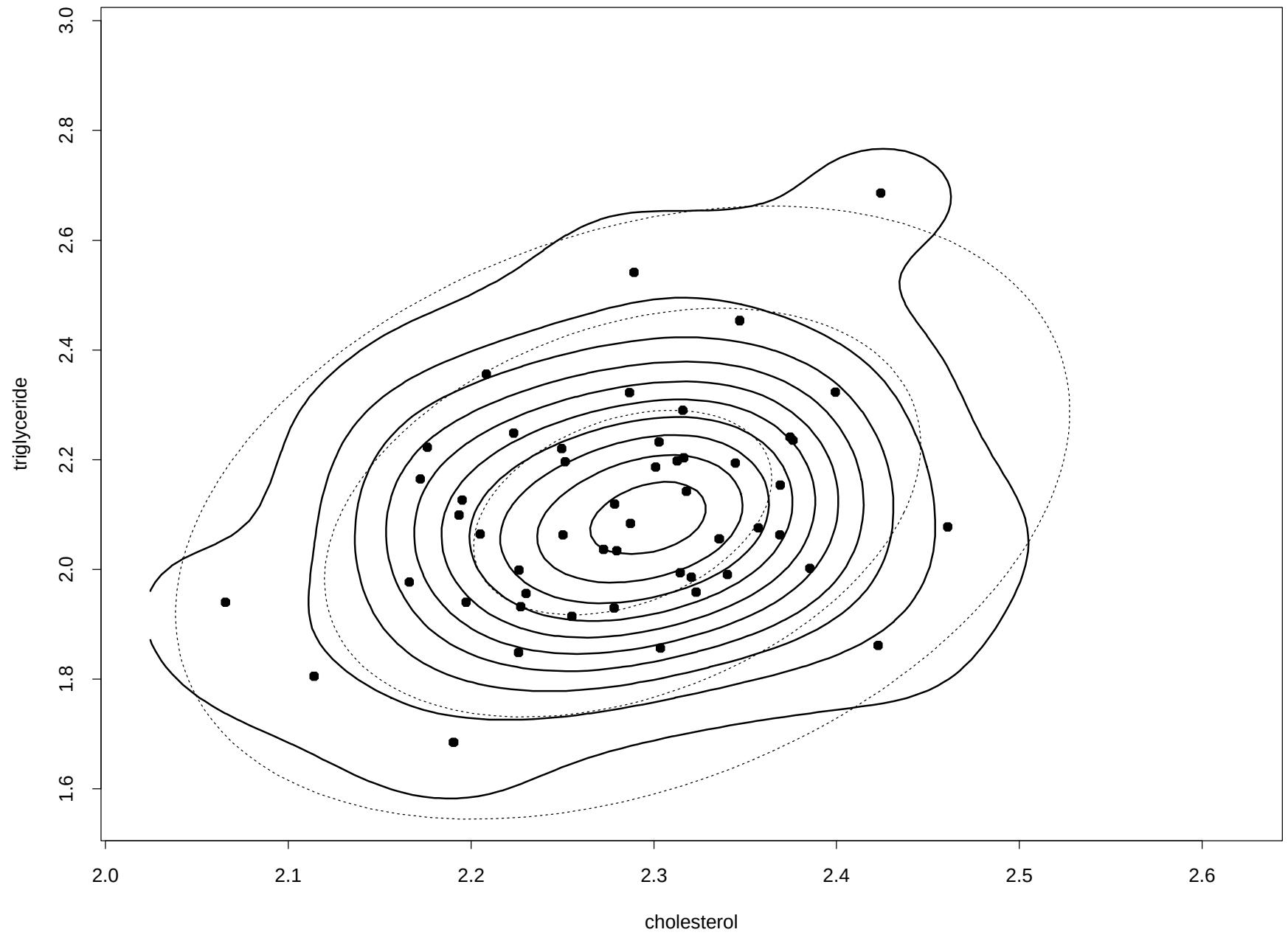
diseased n = 320



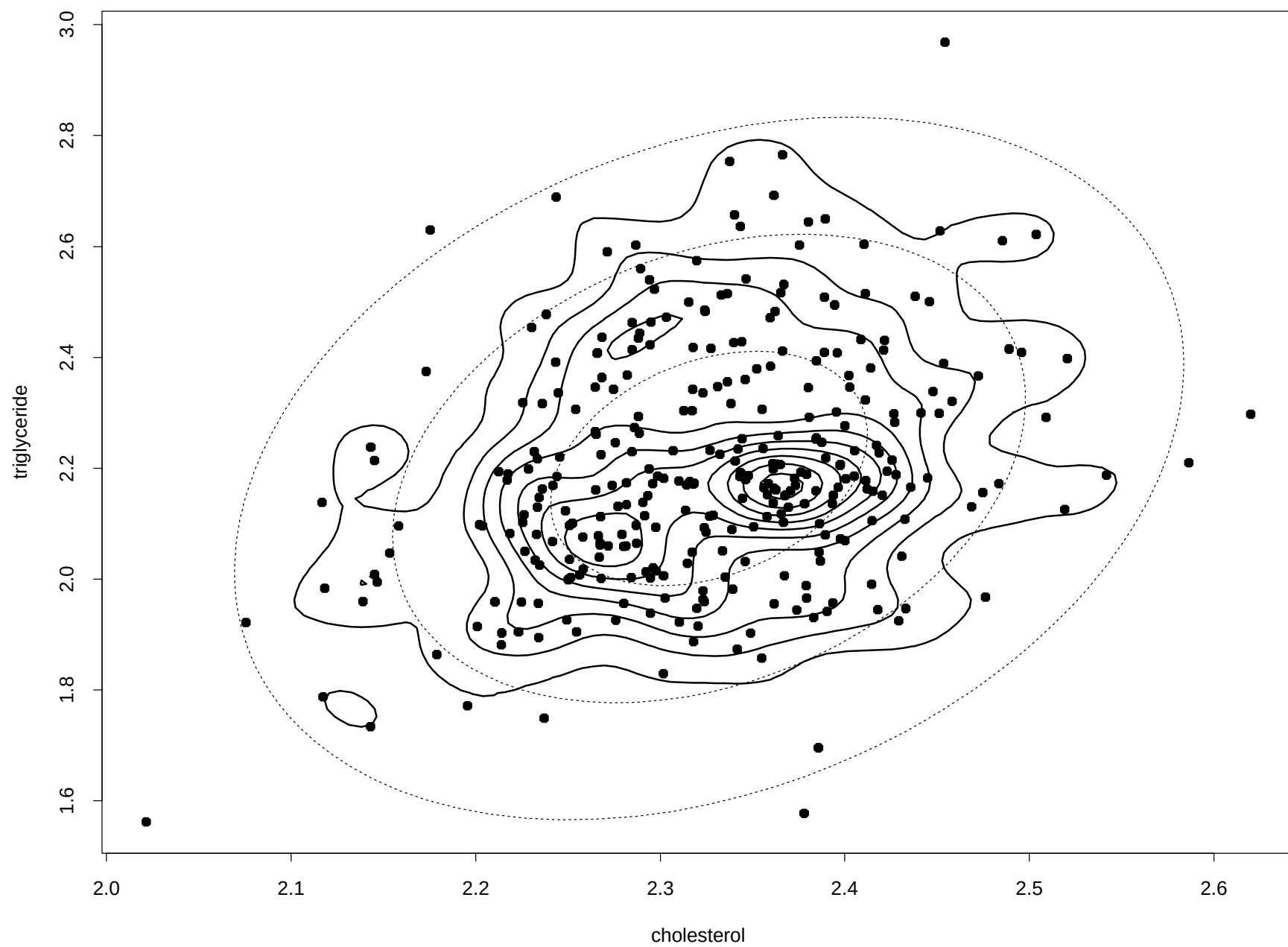
overlay of contours



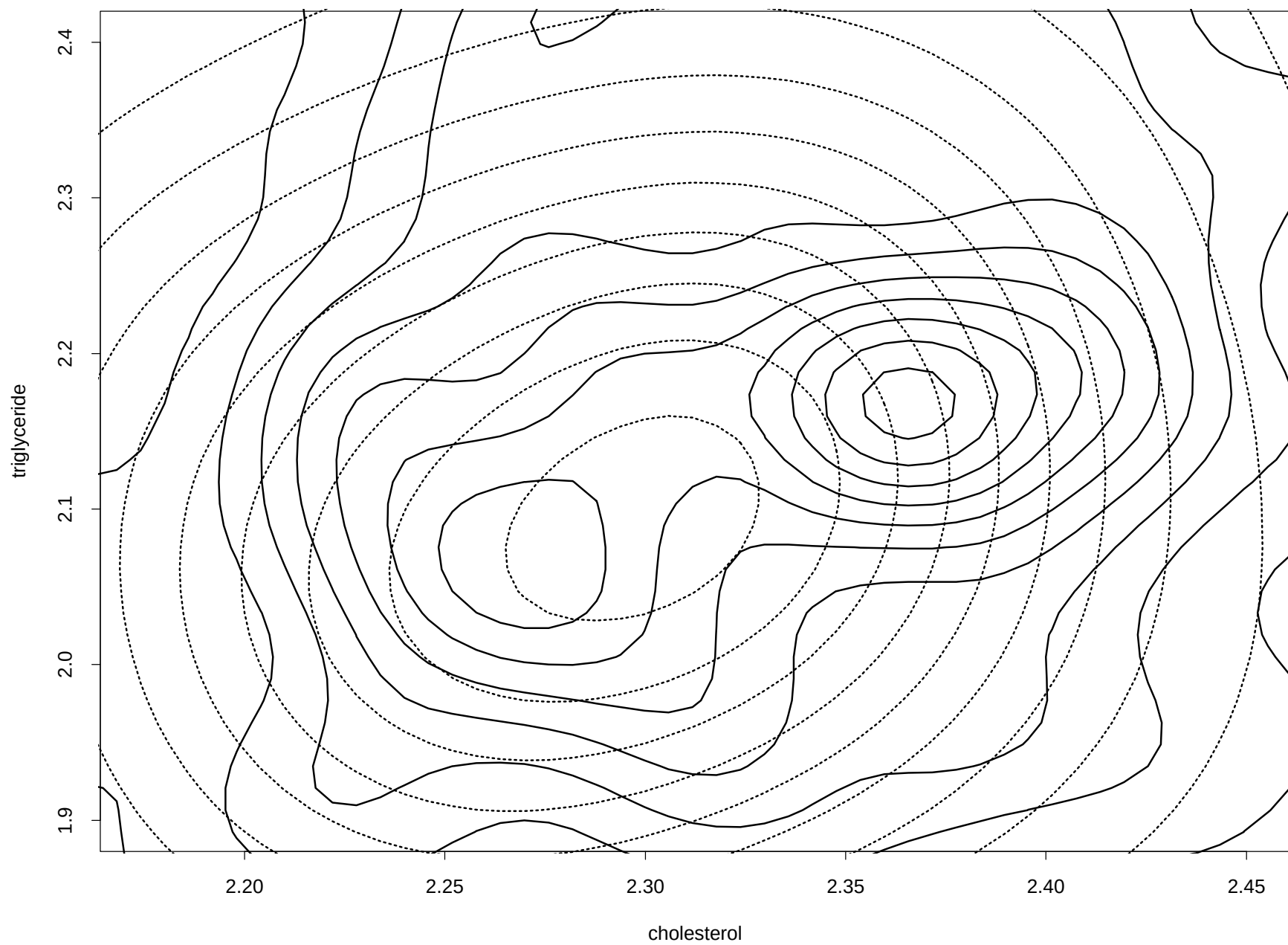
normal n = 51



normal n = 320



overlay of contours (modes at 186 and 233)



2 Old Faithful Data — Times Series

- easy to generate high-dimensional data
- one time series $\Rightarrow$ lagged data
- I am always struck how different the visualizations are for 1-D, 2-D, and 3-D data. Discontinuous. Agree?
- Education required for 3-D visualization and beyond. Not intuitive for all...
- $f(x, y) = f(y) f(x|y)$
- $f(x, y)$ gives an overview (gestalt)
- the pair, $f(y)$ and $f(y|x)$, give more precision
- extending $f(x, y, z) = f(z) f(x, y|z)$
- $f(x, y, z, t) = f(t) f(x, y, z|t)$

- cannot visualize LHS directly in any case
- $f(x, y, z, t_1, t_2) = f(t_1, t_2) f(x, y, z | t_1, t_2)$

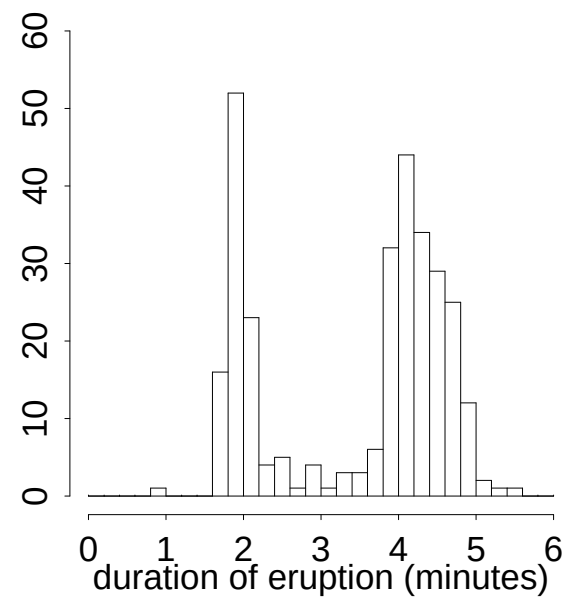
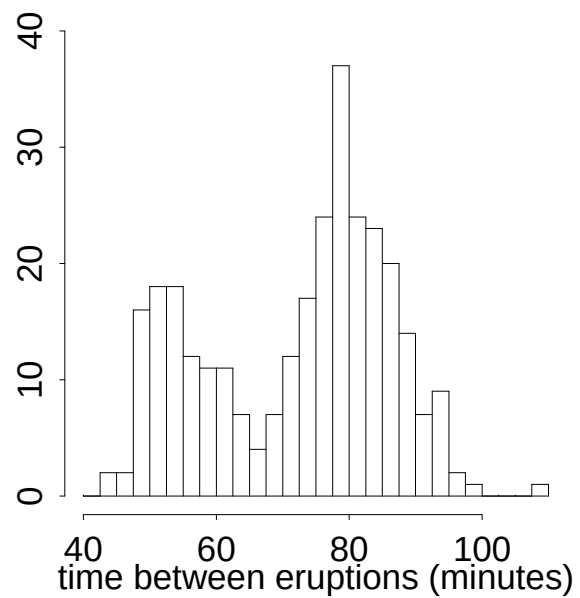


Figure 1: Old Faithful geyser data.

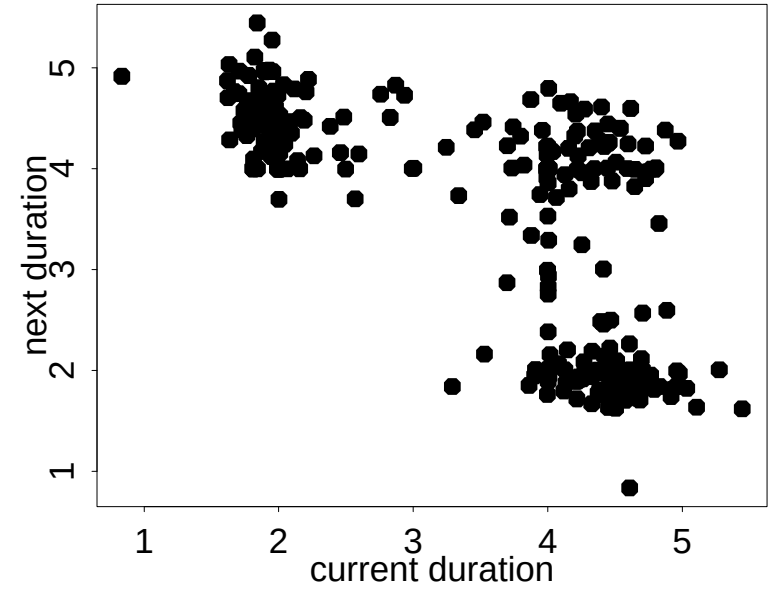
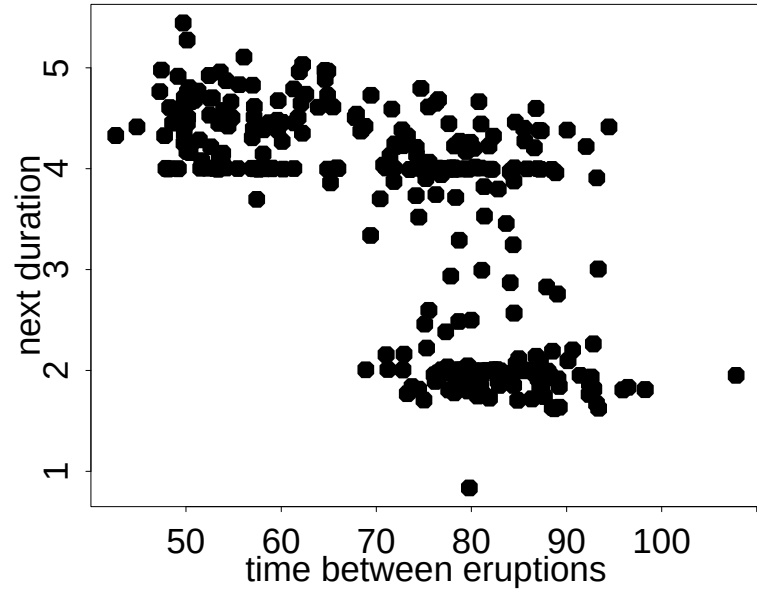


Figure 2: Old Faithful geyser data.

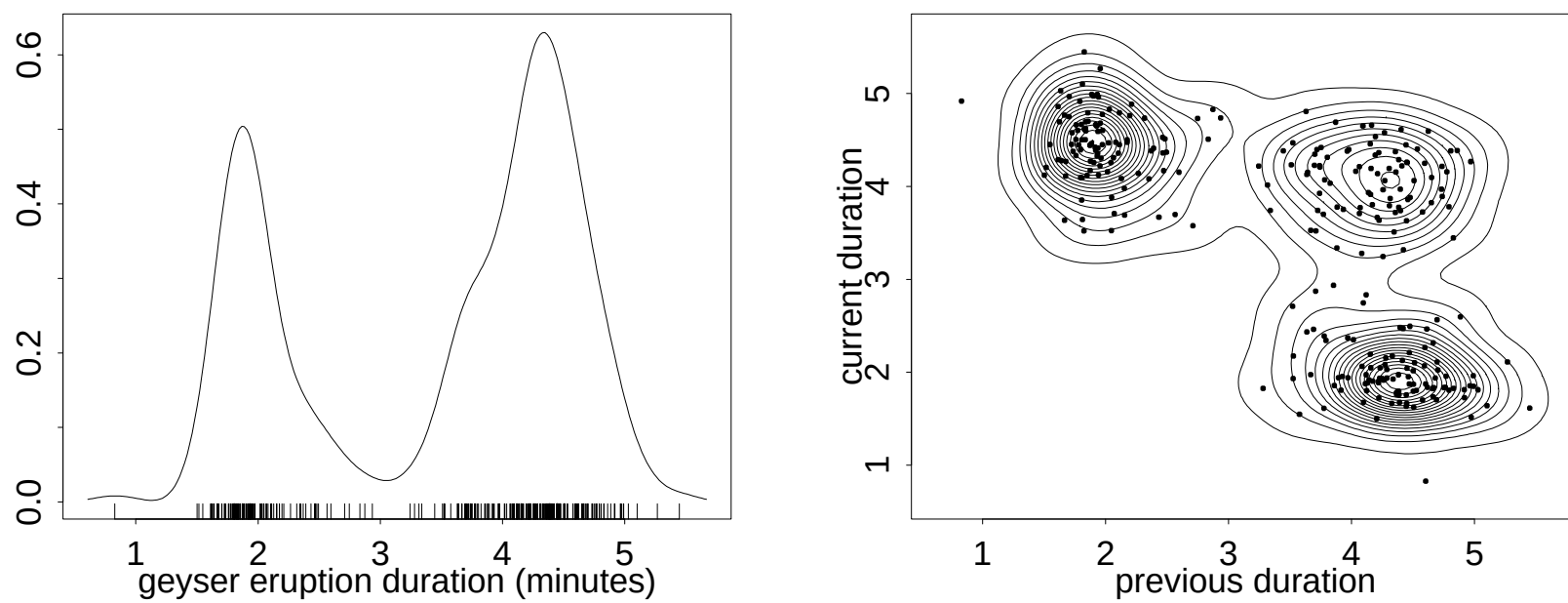


Figure 3: Old Faithful geyser data — lag 1.

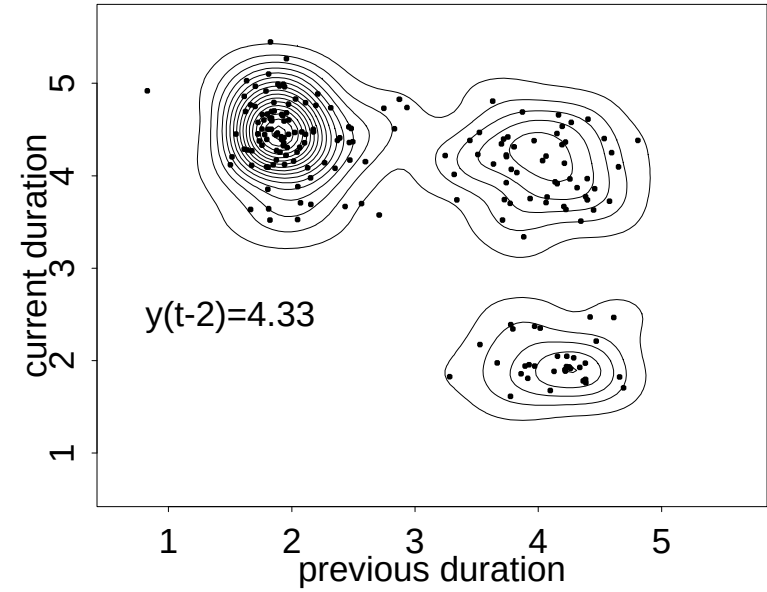
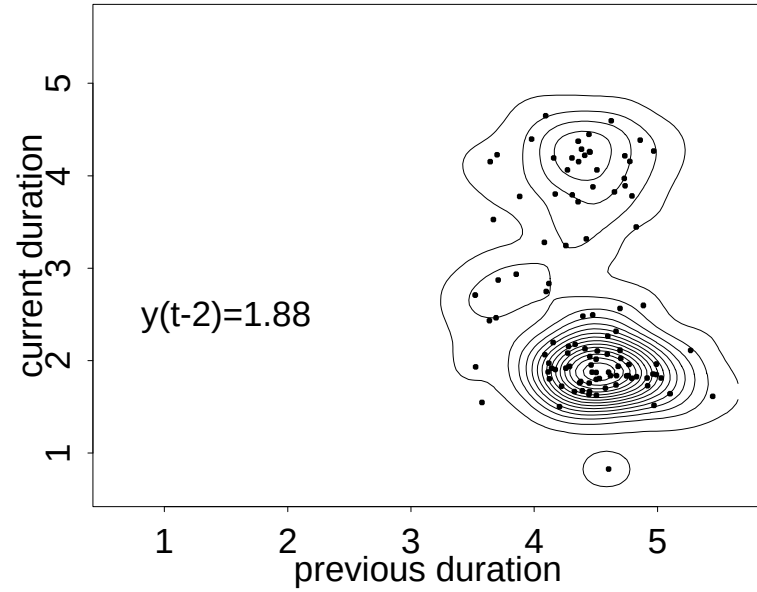


Figure 4: Old Faithful geyser data — lags 1 and 2.

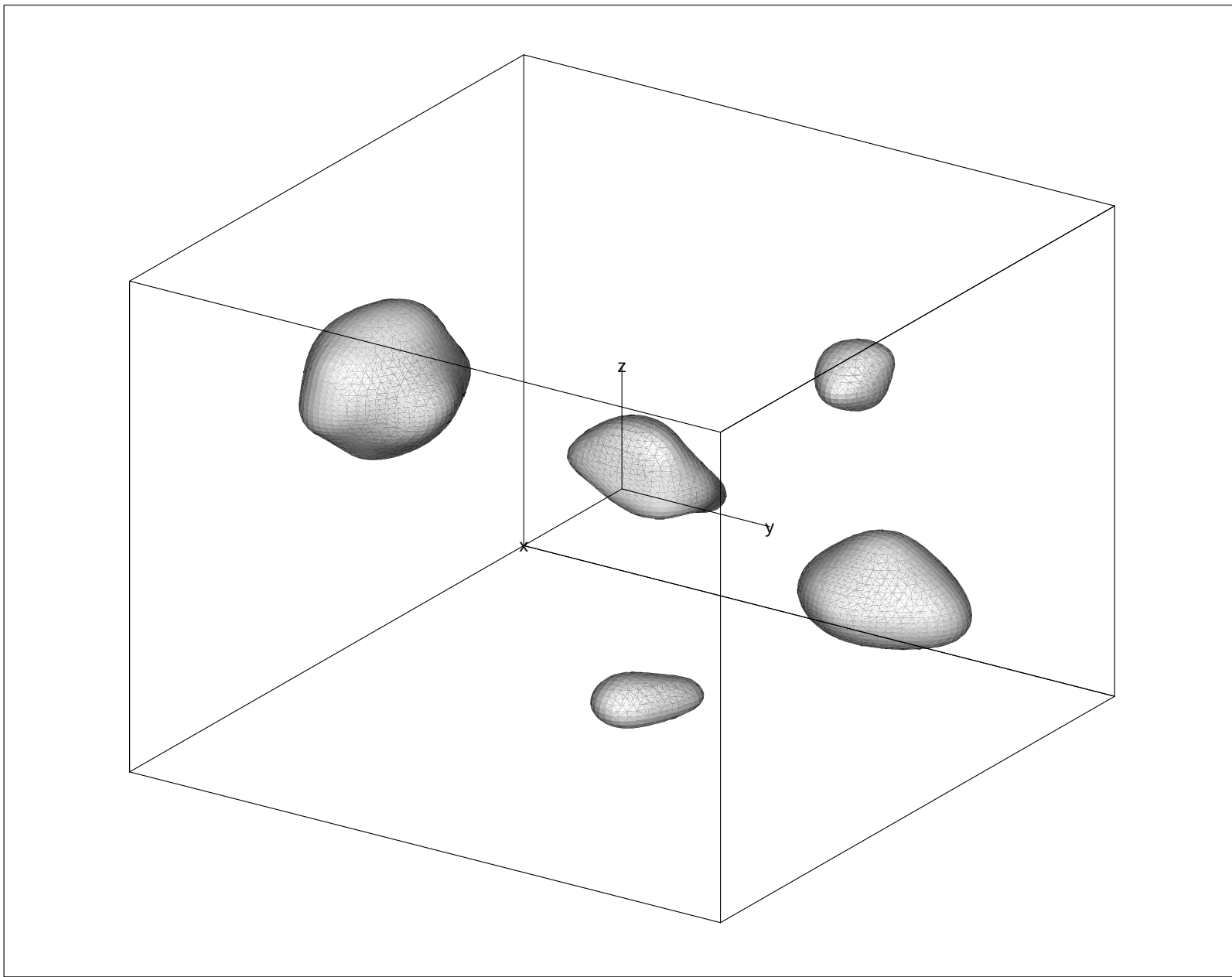


Figure 5: Old Faithful geyser data — lags 1 and 2 (58% contour).

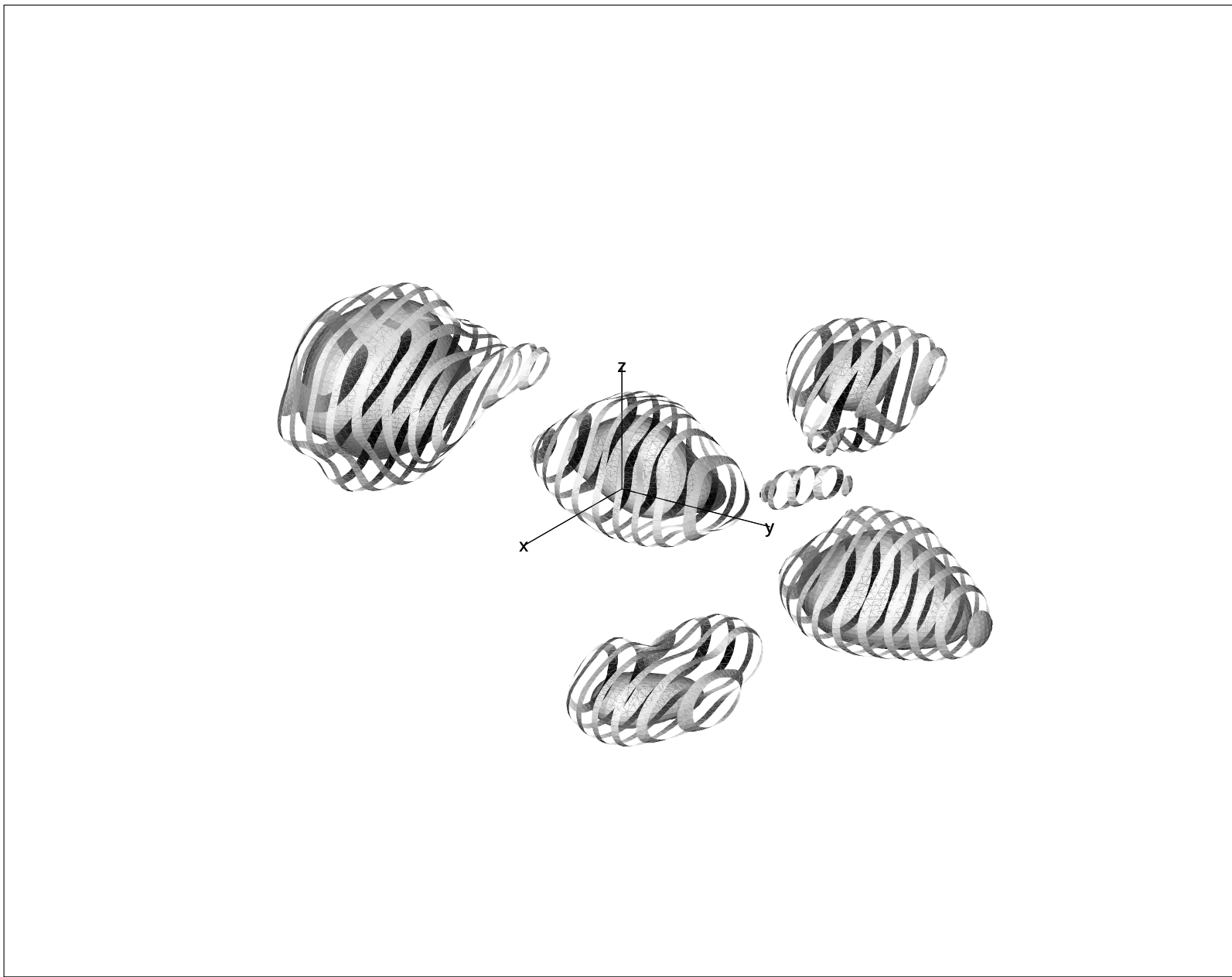


Figure 6: Old Faithful geyser data — 28% and 58% contours.

3 Landsat – Visual Clustering/Discrimination

- more data is good, but not necessarily for visualization
- but more data *is* good for smoothing
- $n \rightarrow \infty$ smoothing becomes exact vs. data becomes dense
- eye can be trained but easily tricked/fooled
- more features $\Rightarrow$ better separability of classes (good)
- $p \rightarrow n$ (or $p > n$) $\Rightarrow$ many spurious “summaries” that seem to separate classes; which are stable? (microarrays)
- discovery vs. confirmation

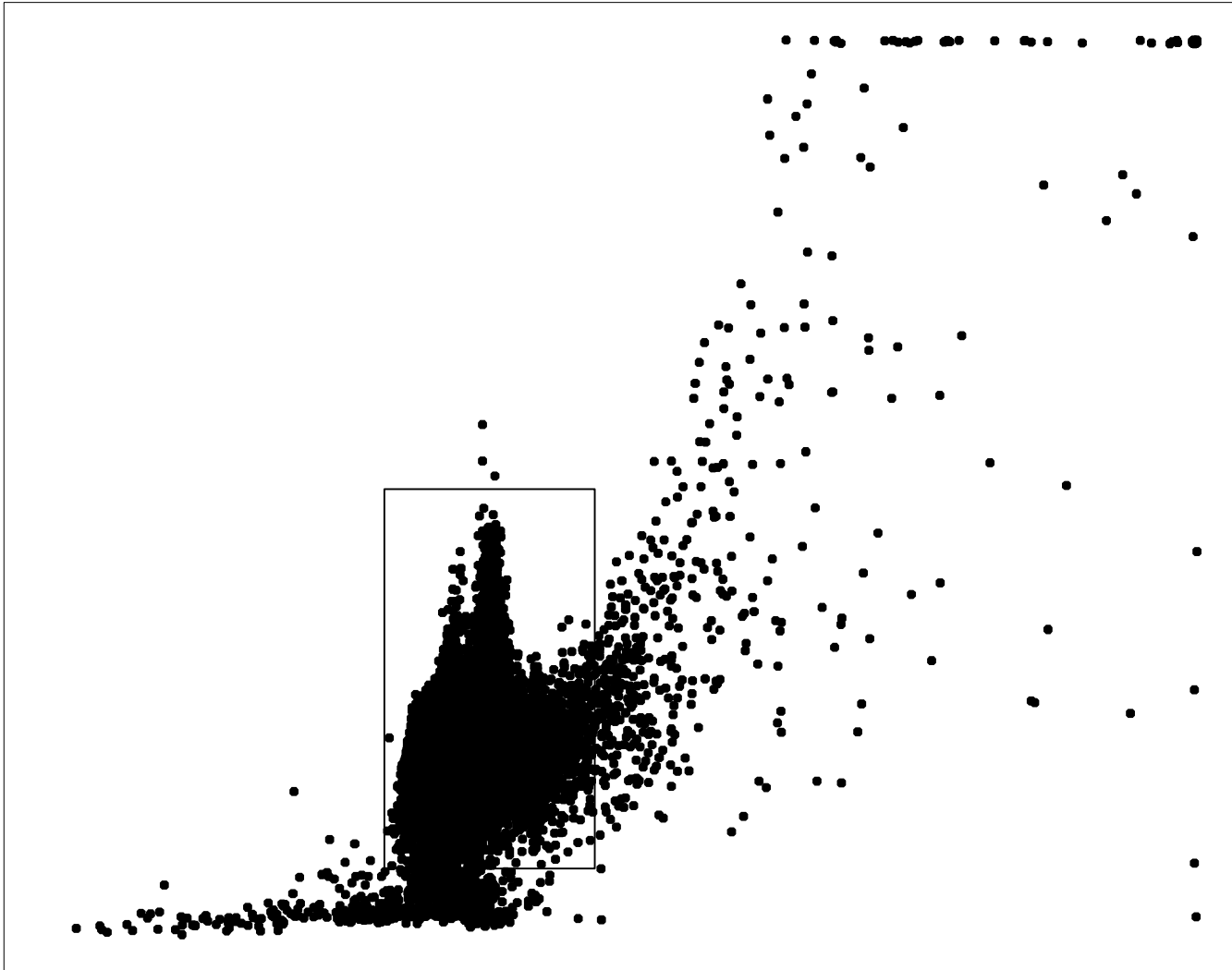


Figure 7: Landsat IV: scatterplot $n=23932$ pixels

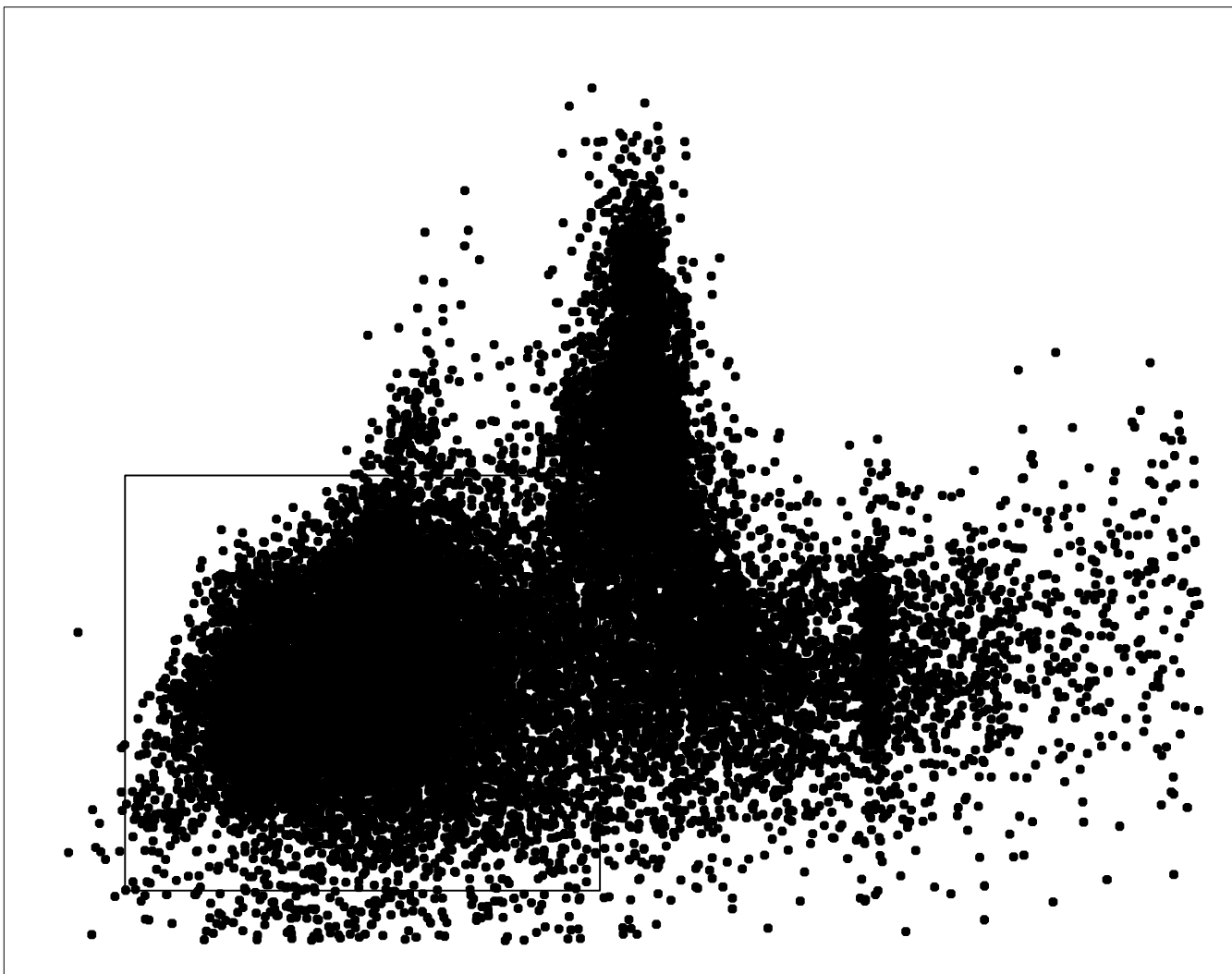


Figure 8: Landsat IV: scatterplot (blowup)

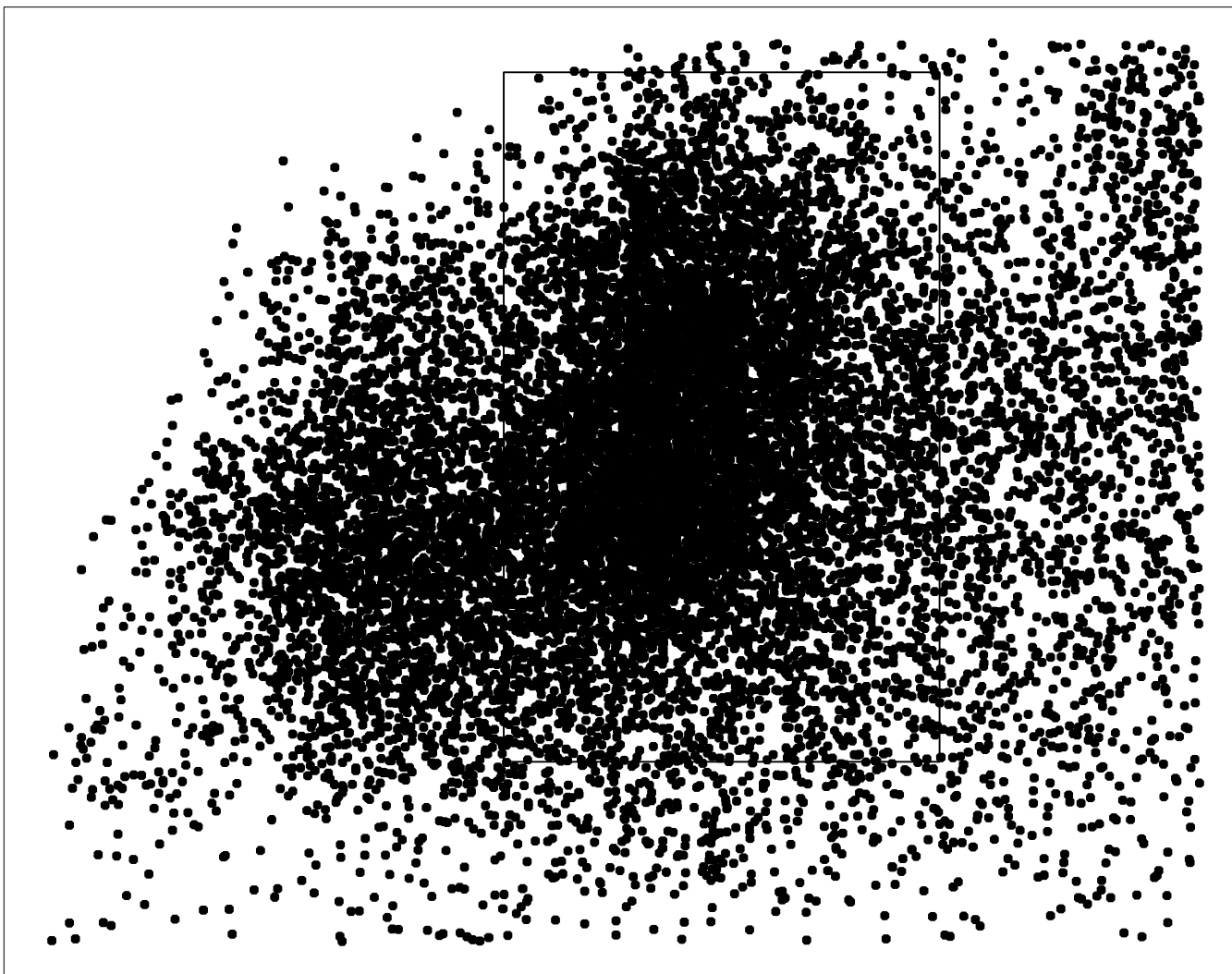


Figure 9: Landsat IV: scatterplot (blowup)

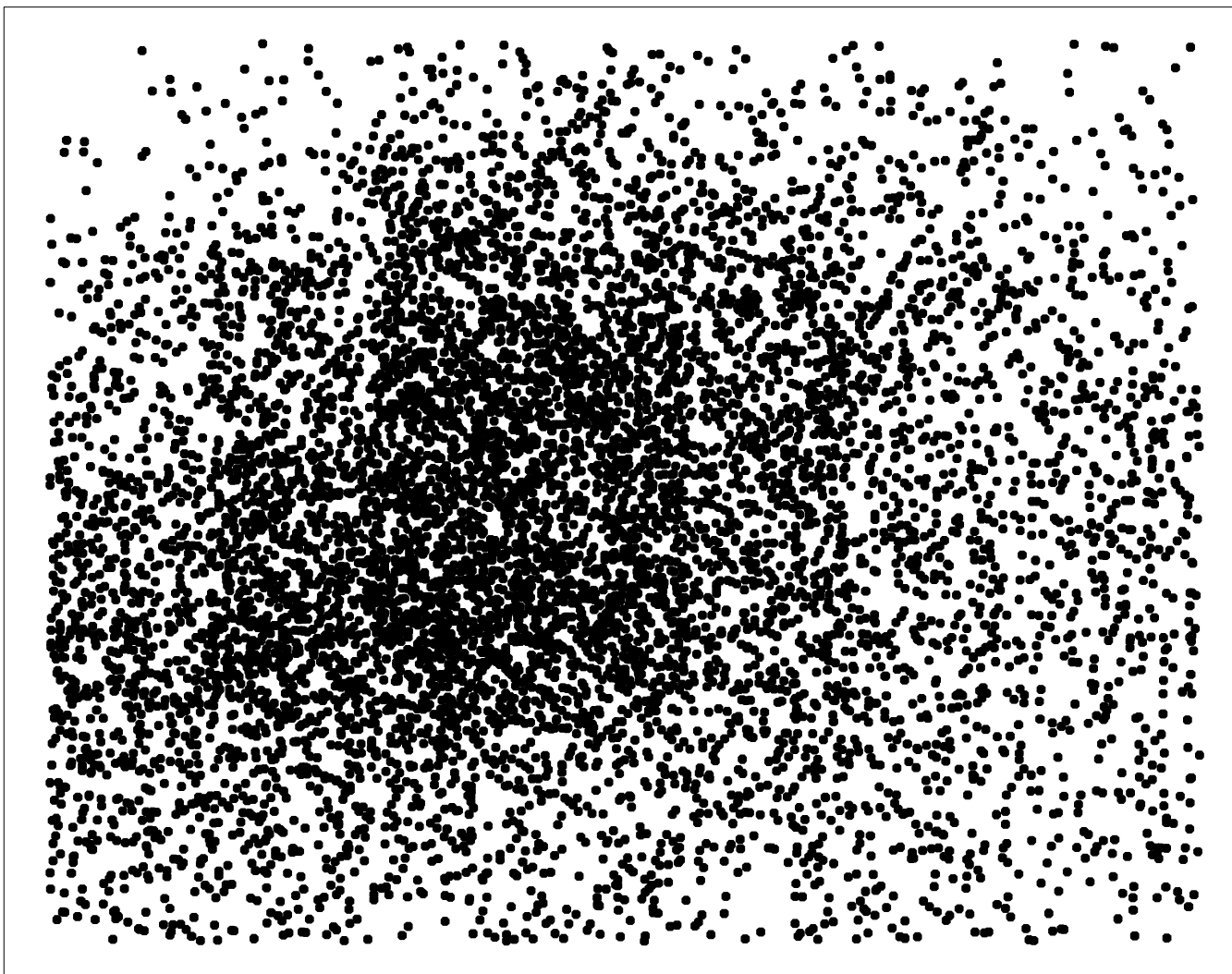


Figure 10: Landsat IV: scatterplot (blowup)

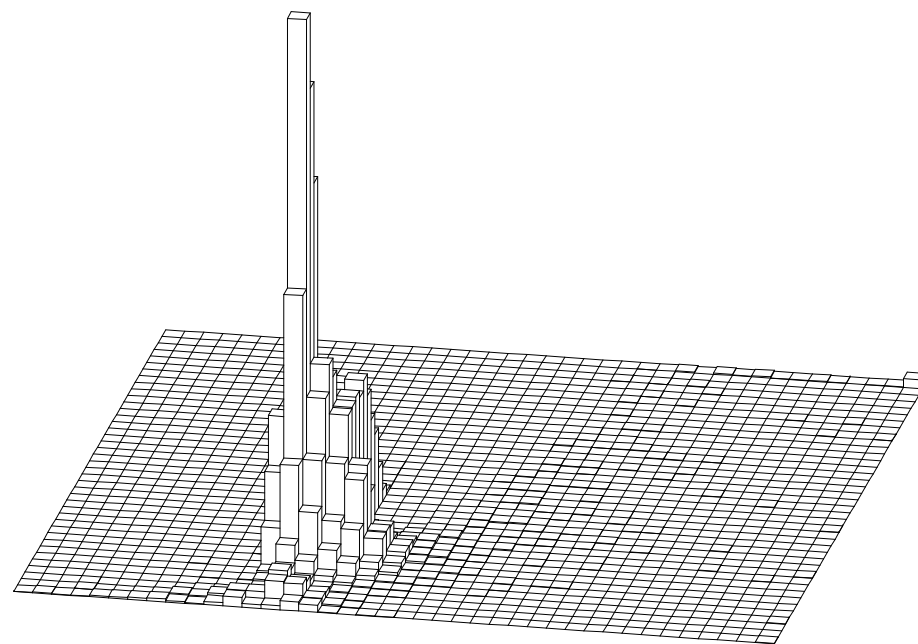


Figure 11: Landsat IV: Histogram of first view

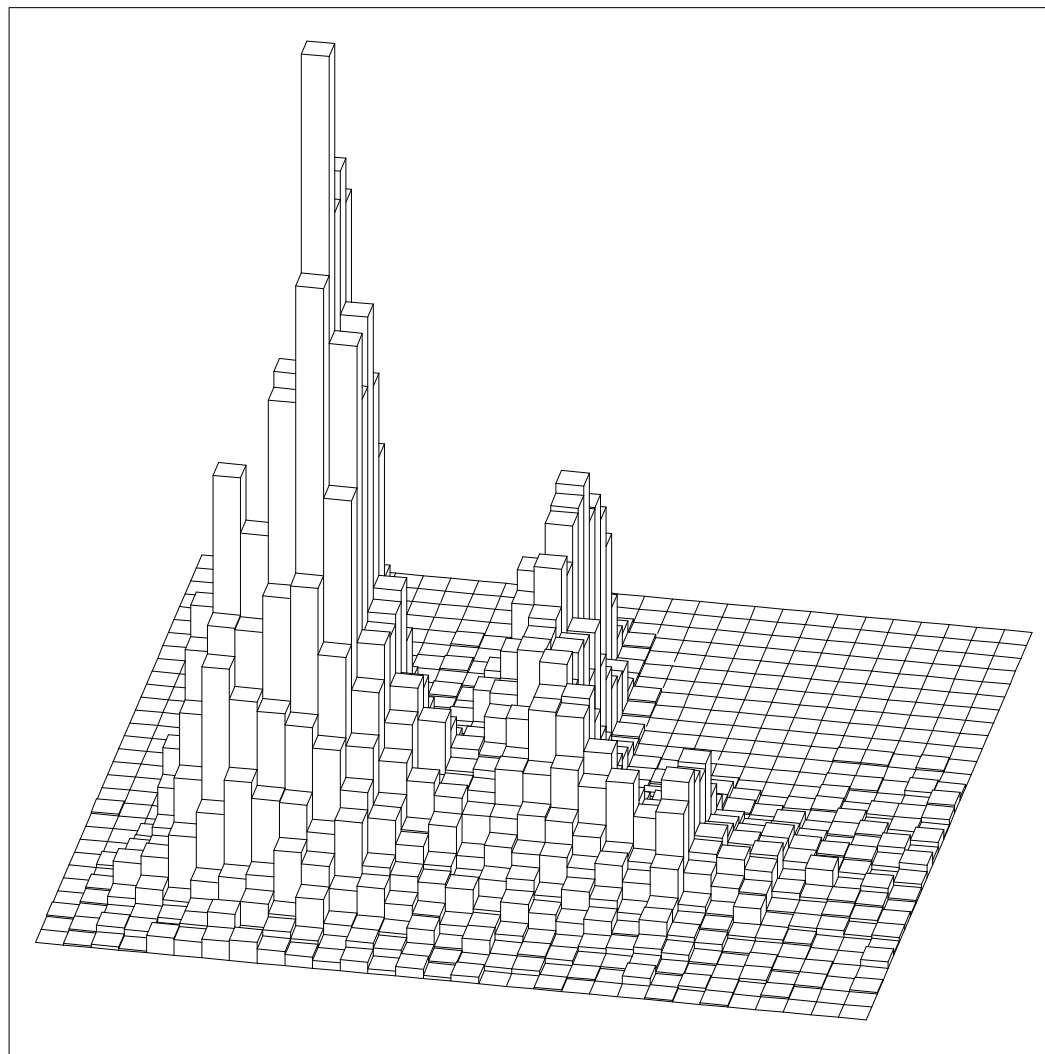


Figure 12: Landsat IV: Histogram of first blowup

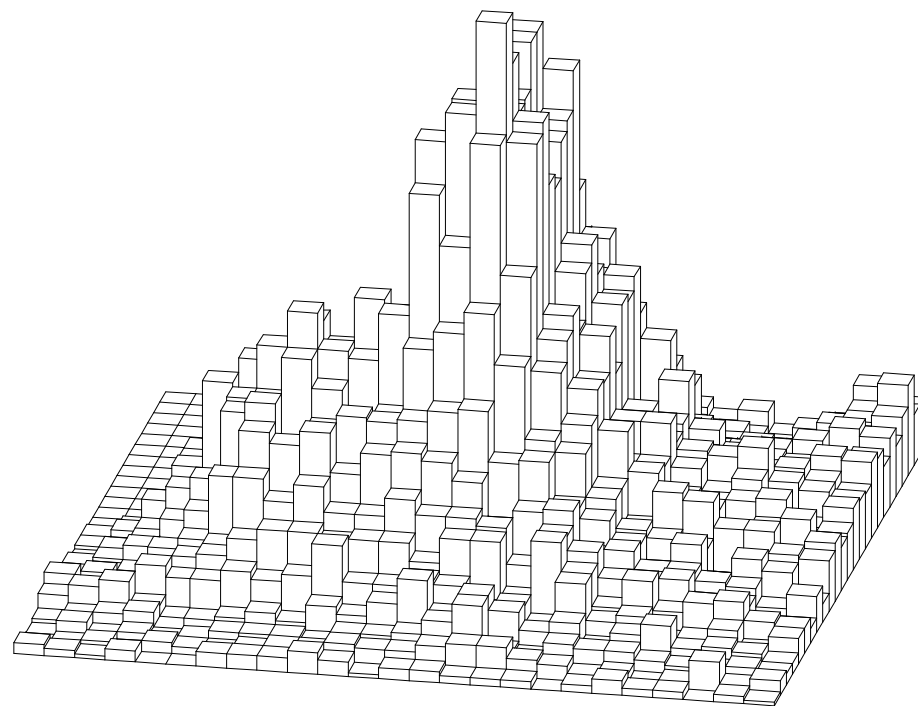


Figure 13: Landsat IV: Histogram of second blowup

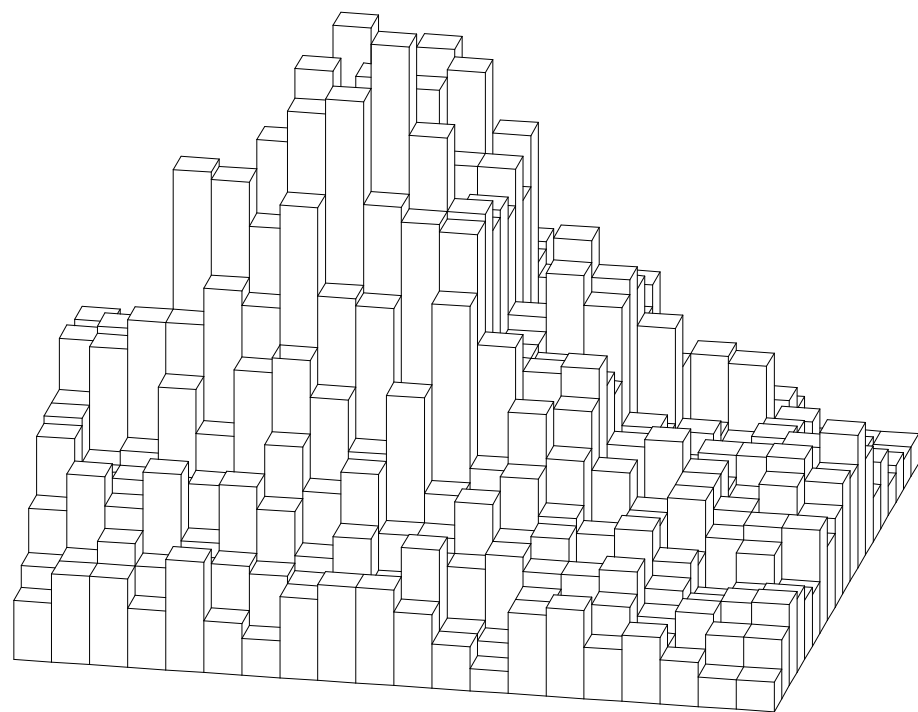


Figure 14: Landsat IV: Histogram of third blowup

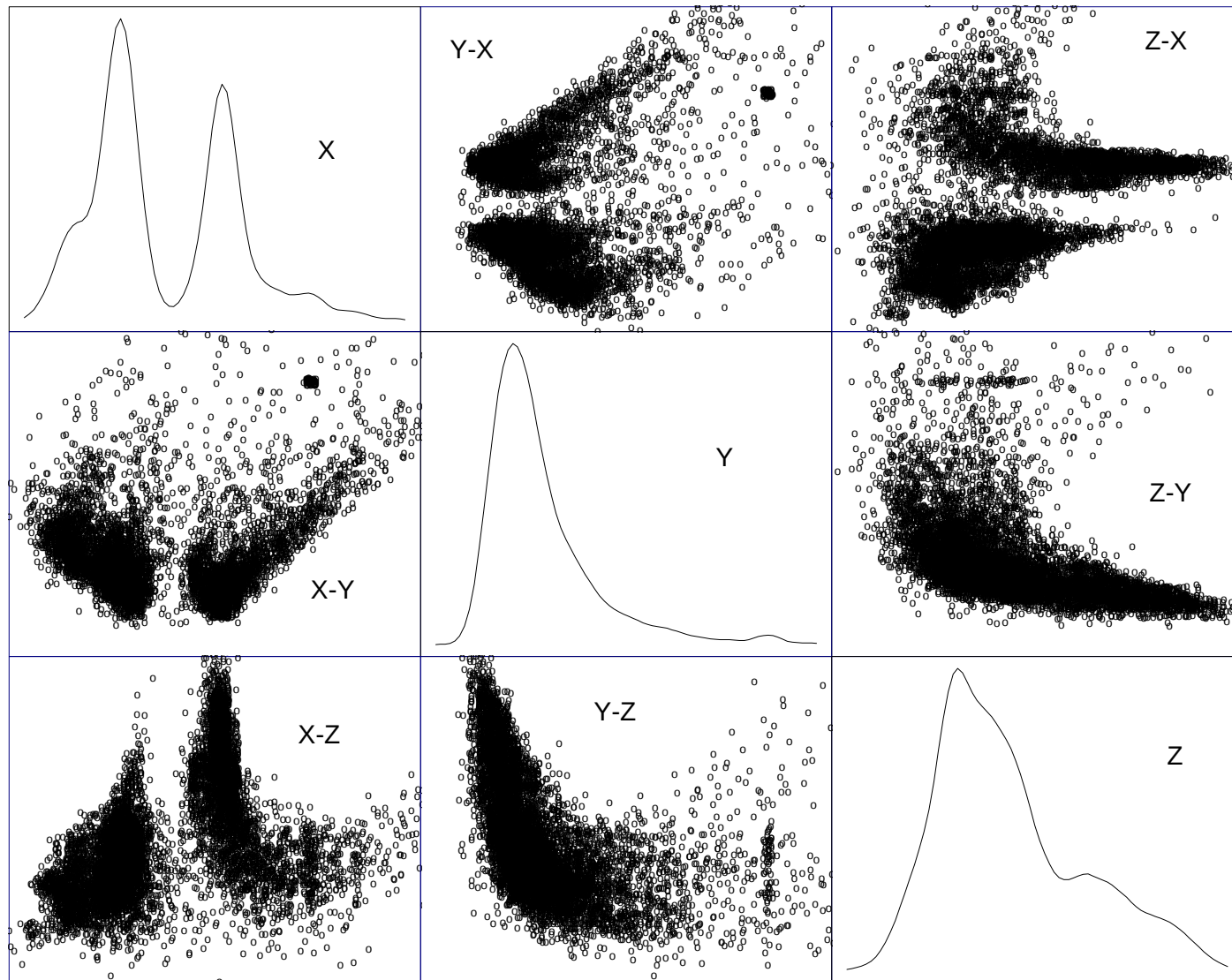


Figure 15: Landsat IV: 3 features of 3 crops (sunflower, spring wheat, spring barley)

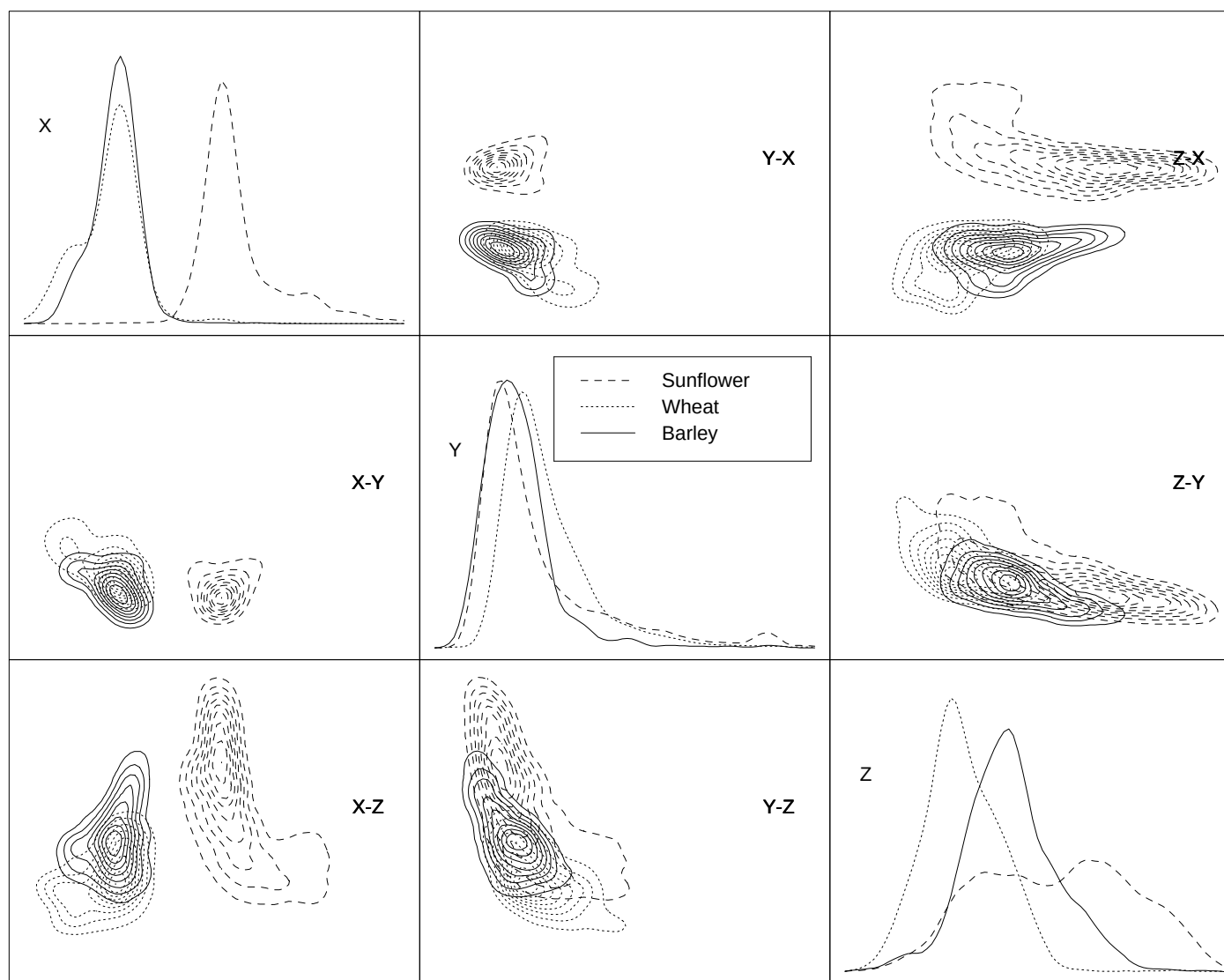


Figure 16: Landsat: 3 crops and 3 features

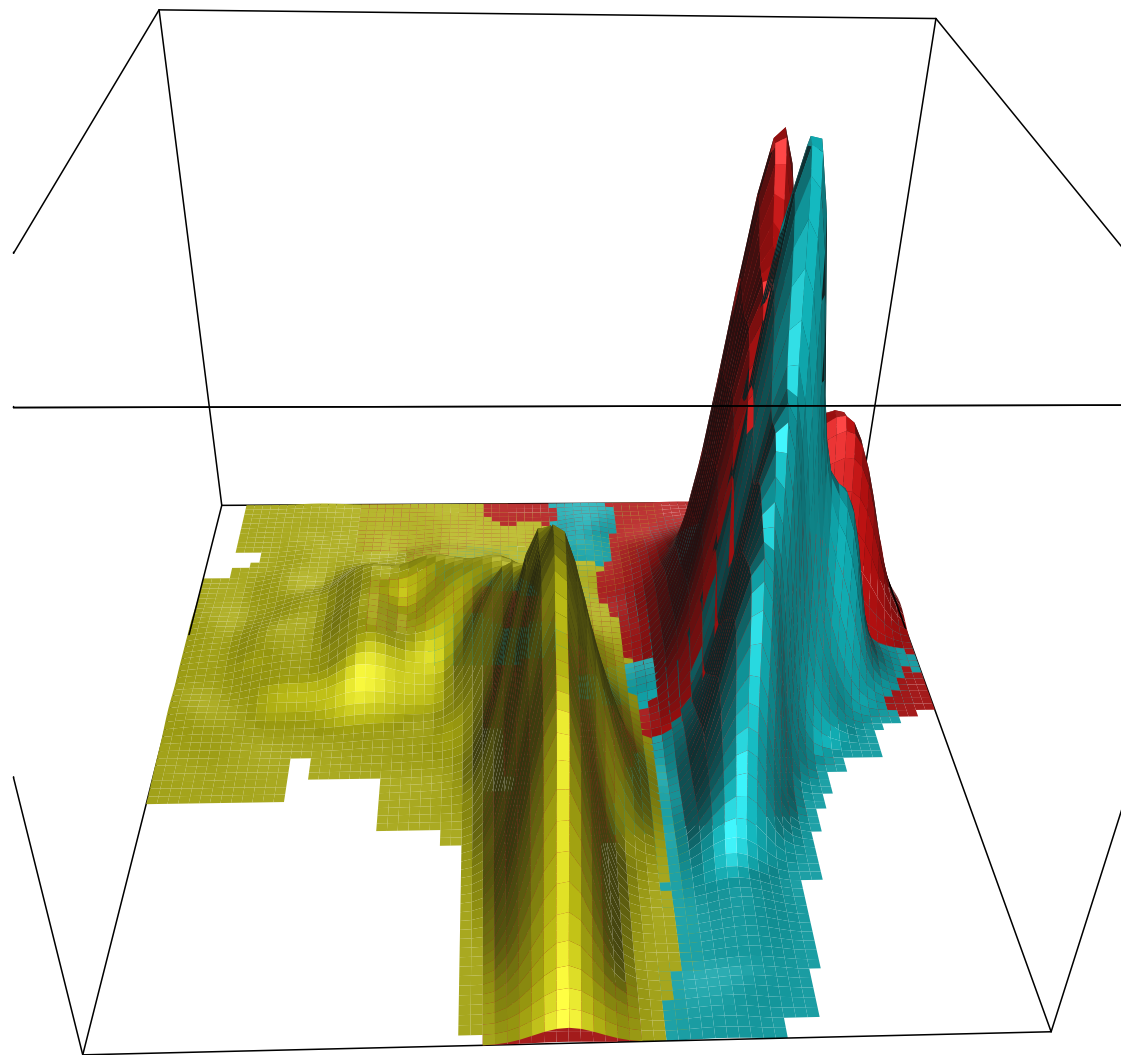


Figure 17: Landsat: 3 crops and pairwise features

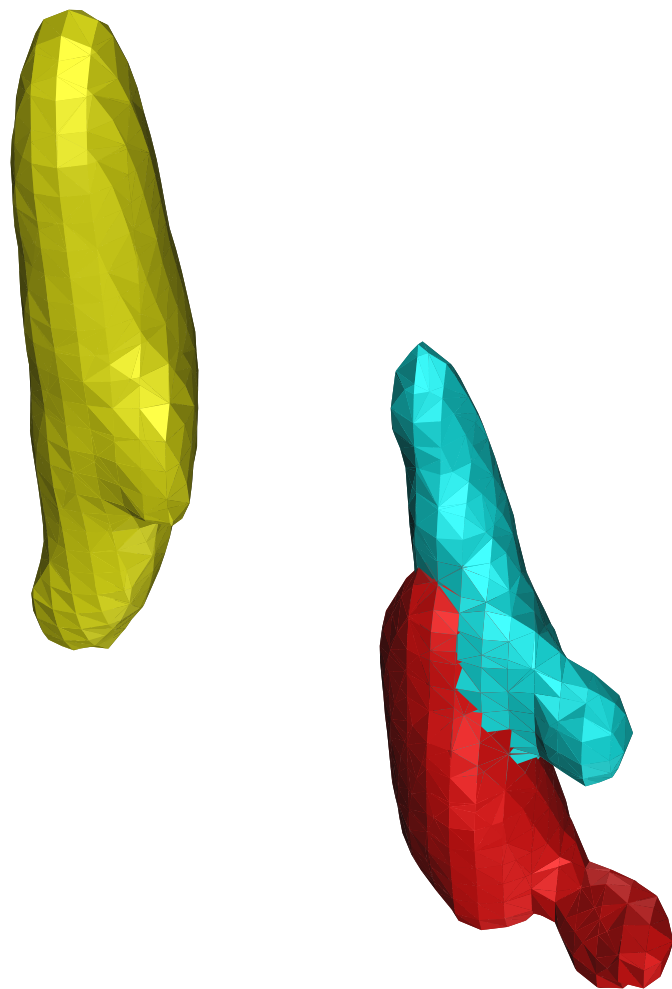


Figure 18: Landsat: 3 crops and trivariate features

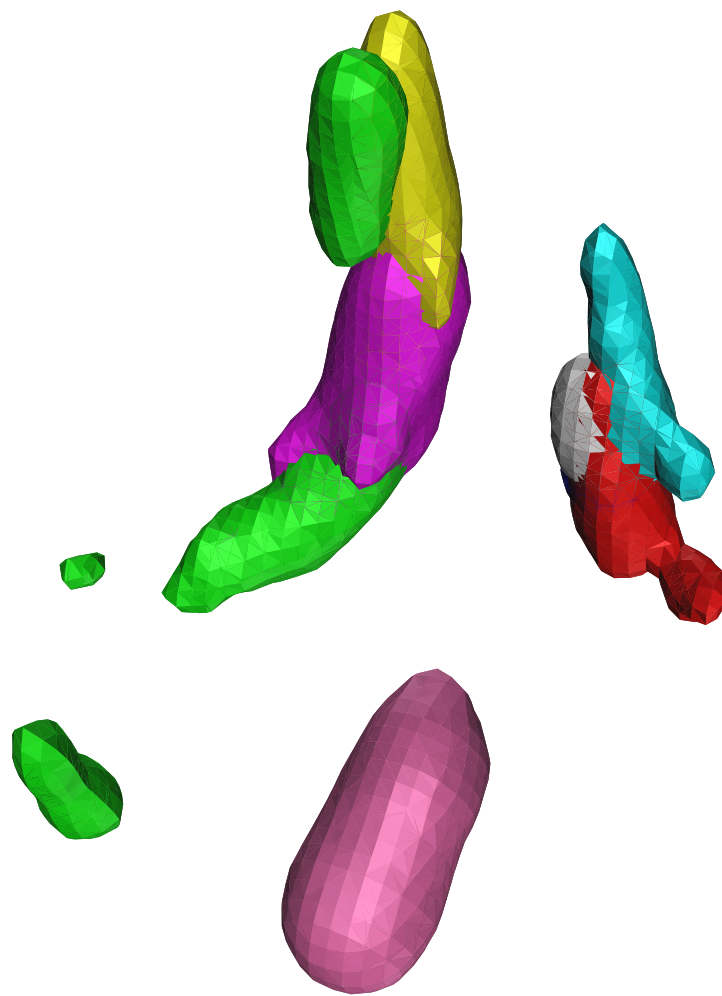


Figure 19: Landsat: all crops and trivariate features

4 Lynx and Earthquake Densities – 3 and 4 dimensions

- visualization strongly rooted in “real” word (1–3 dimensions)
- how to extrapolate 1,2,3-D experience to 20,50,1000-D ?
- parallel coordinates allows easy representation of 2-25 dimensions, but still only able to find low-dimensional features
- limited language for high-D features
- limited to features that can be displayed in a low-D subspace/projection (eg. PP)
- data with “holes”

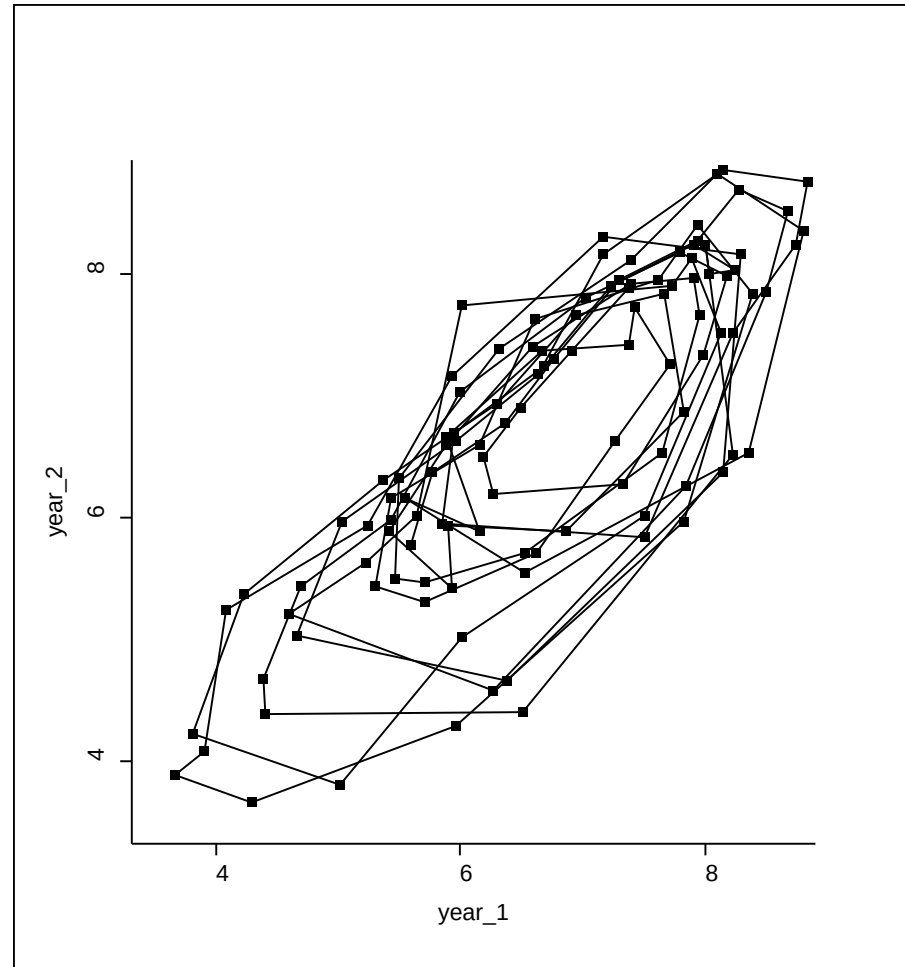


Figure 20: Lynx

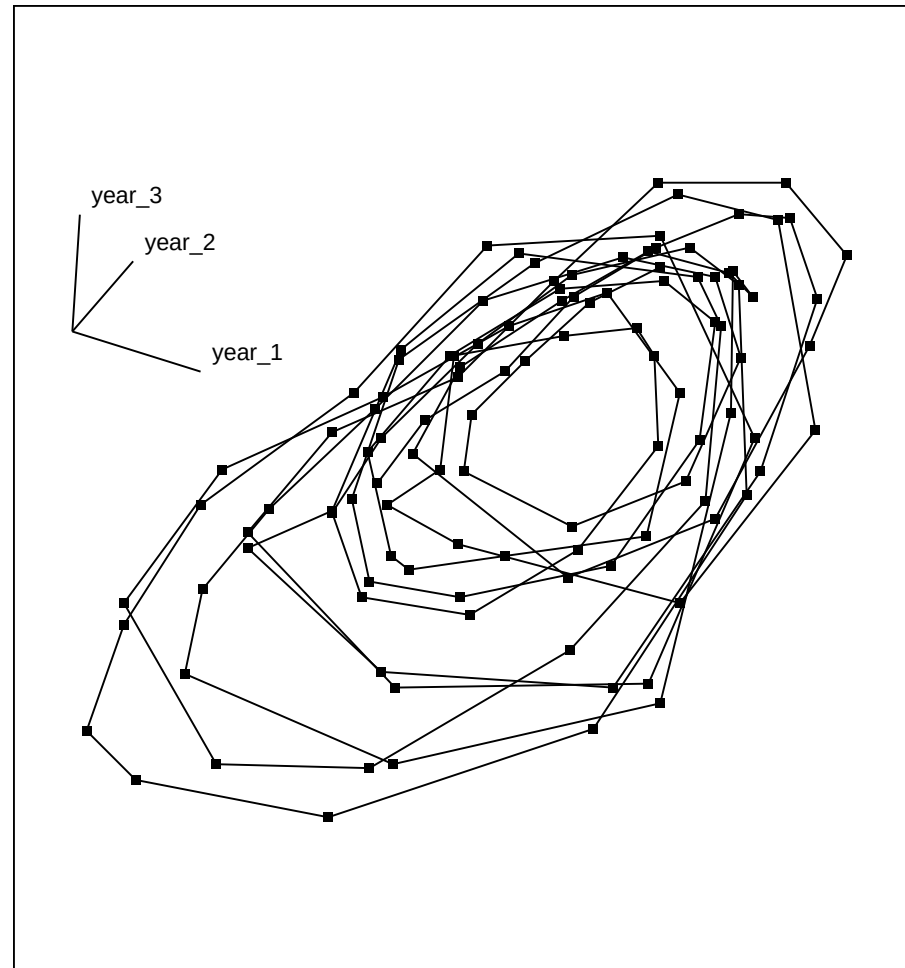


Figure 21: Lynx

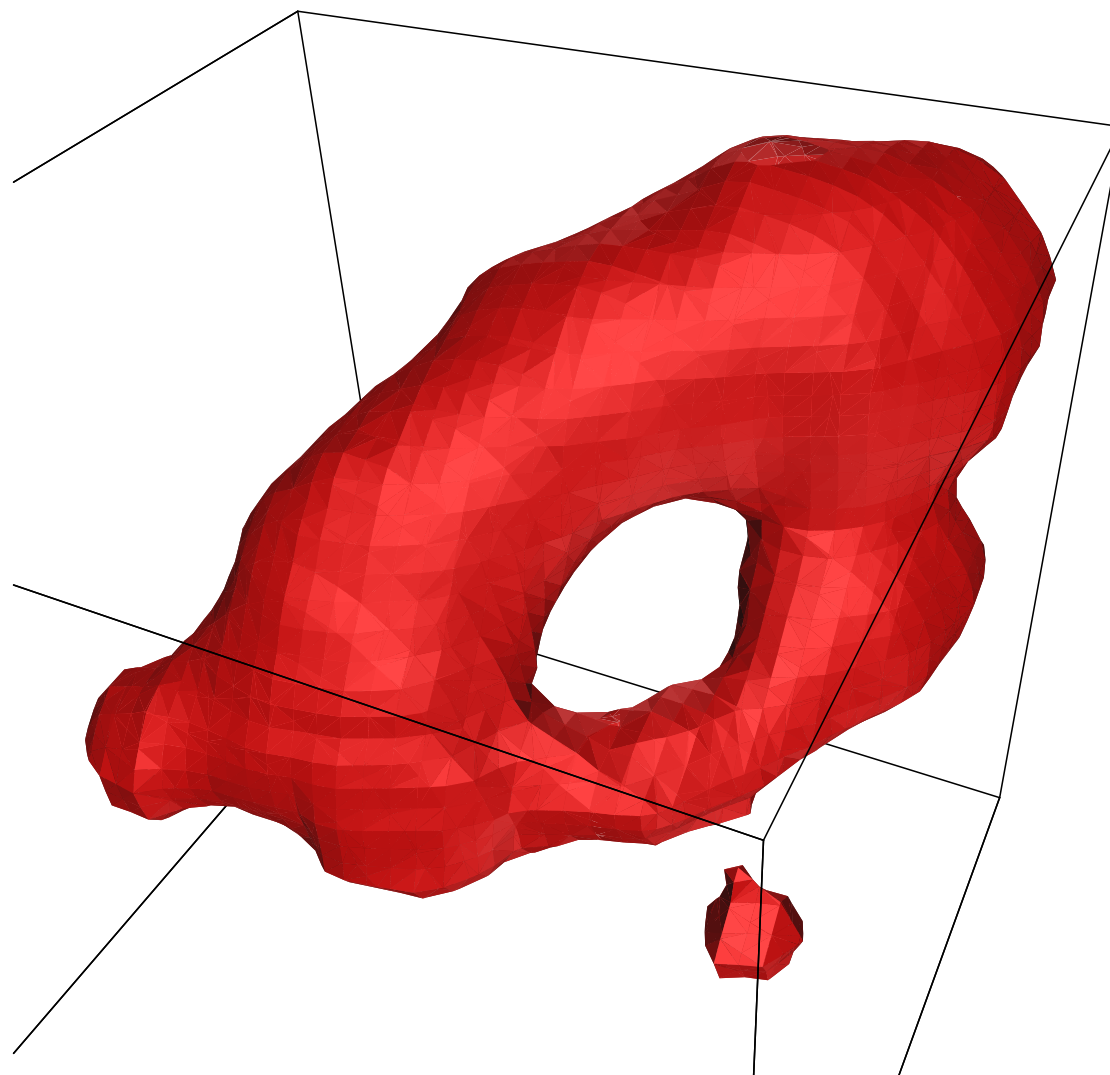


Figure 22: Lynx

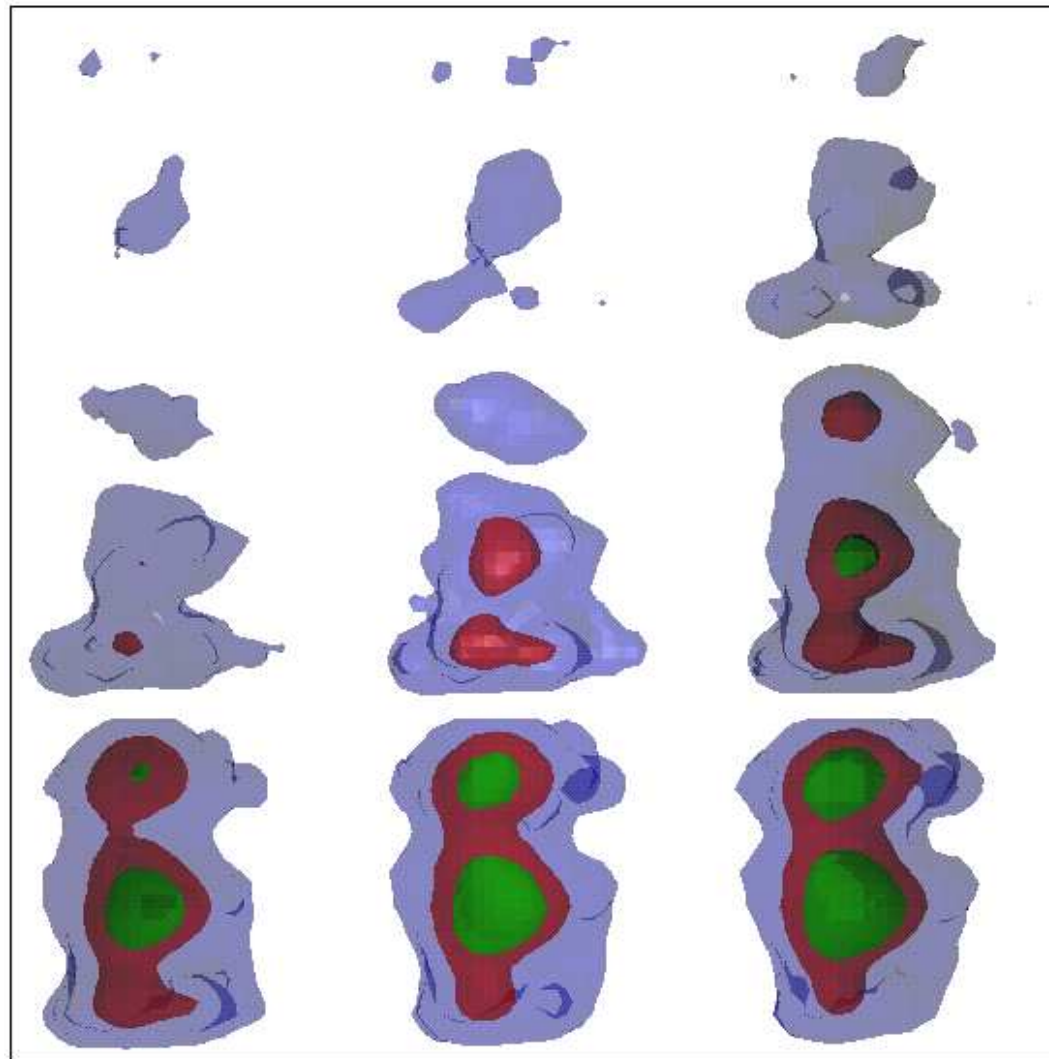


Figure 23: Mount St. Helen's Earthquake

5 Robust Regression: Mapping Residuals

- often advantageous to visualize data indirectly through the filter of a model
- eg. are data normal? plotting data on “normal” paper — easy to see a straight line and departures from normality (vs. cdf)
- likewise, models provide context for data
- in “real” world, model may not be known (so nonparametric approach)
- more commonly, “know” model for: (1) noise but not the signal; (2) signal but not the noise; (3) subset of signal only (4) subset of noise only (*partial model knowledge*)
- eg. (1) wavelet noise in images $\Rightarrow$ normal coefficients
- eg. (2) physics says shape of regression curve for lightning is quadratic (Tom Burr LANL)

- mixture models attractive/effective (or a subset of a mixture model)
- minimum distance estimation (vs. EM/MLE) has advantages: (1) curve matching criterion; (2) robustness against outliers; (3) fitting incomplete models (Scott, 2001, L2E, Technometrics)
- extension to regression straightforward

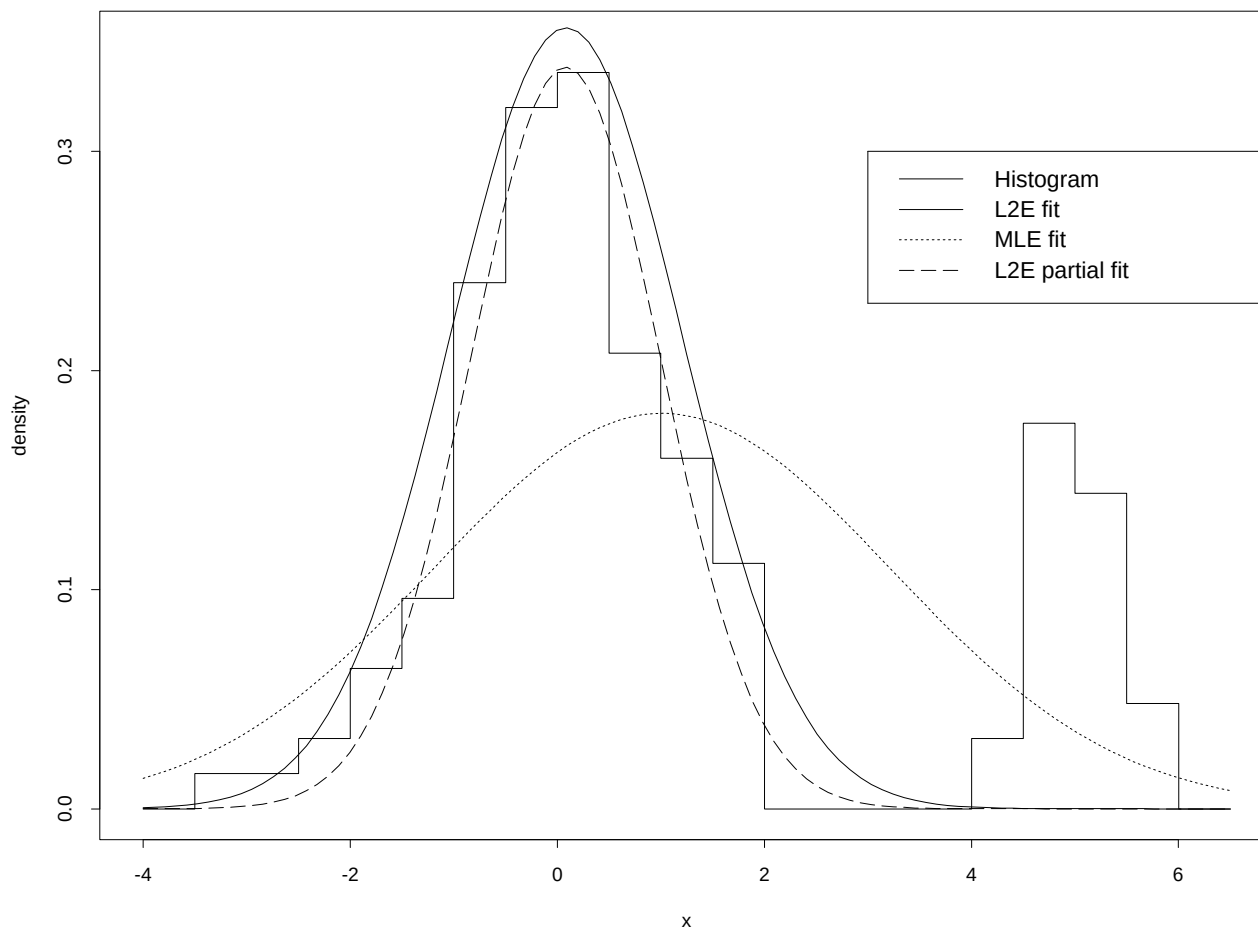


Figure 24: Histogram of 125 points from the mixture $0.8 N(0, 1) + 0.2 N(5, 1)$. Also shown are the maximum likelihood and L2E fits using the incorrect model $N(\mu, \sigma^2)$. Finally, the L2E fit of the 3-parameter model $w \cdot N(\mu, \sigma^2)$ is shown.

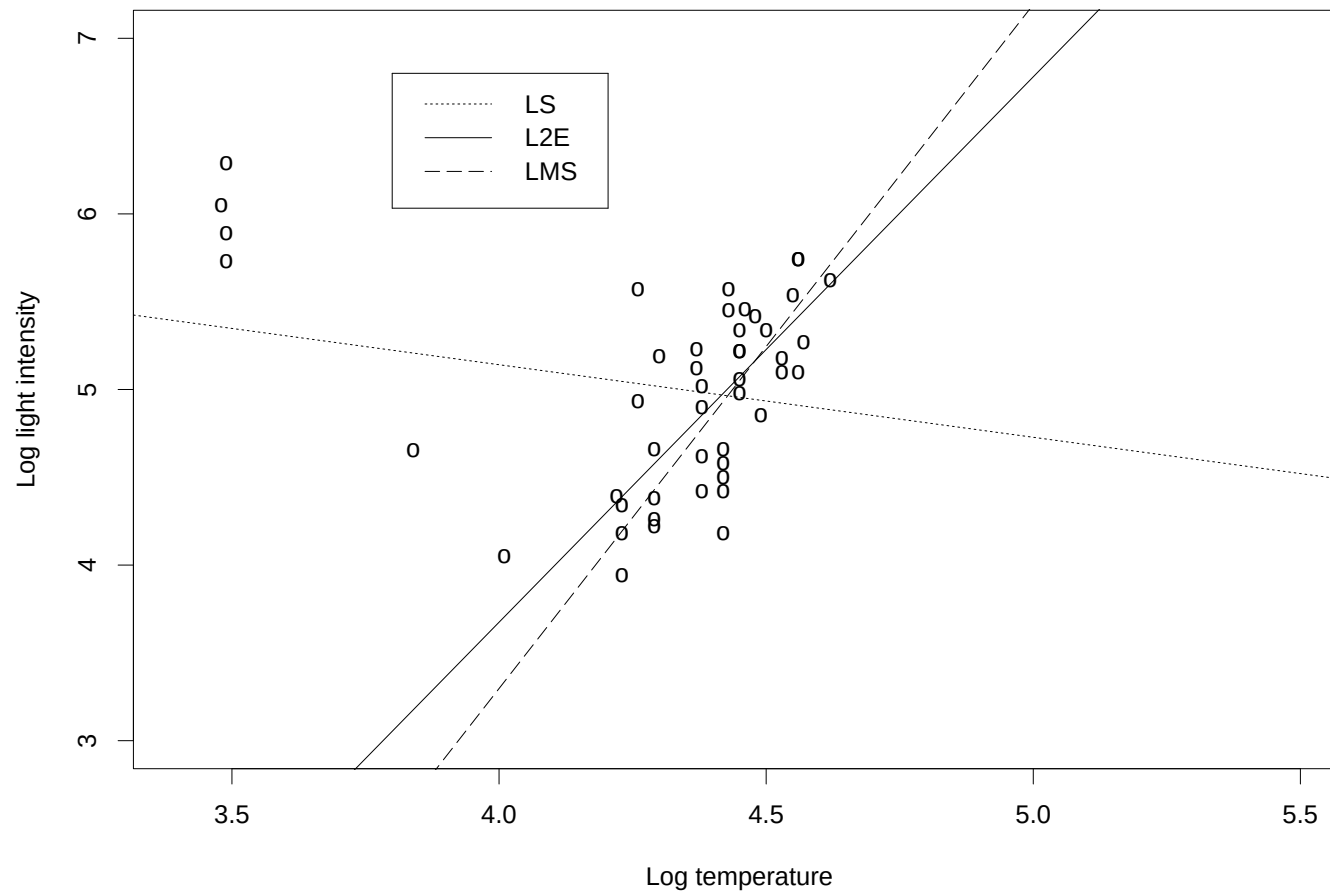


Figure 25: For the star data, straight-line fits by least squares, least median squares, and L2E.

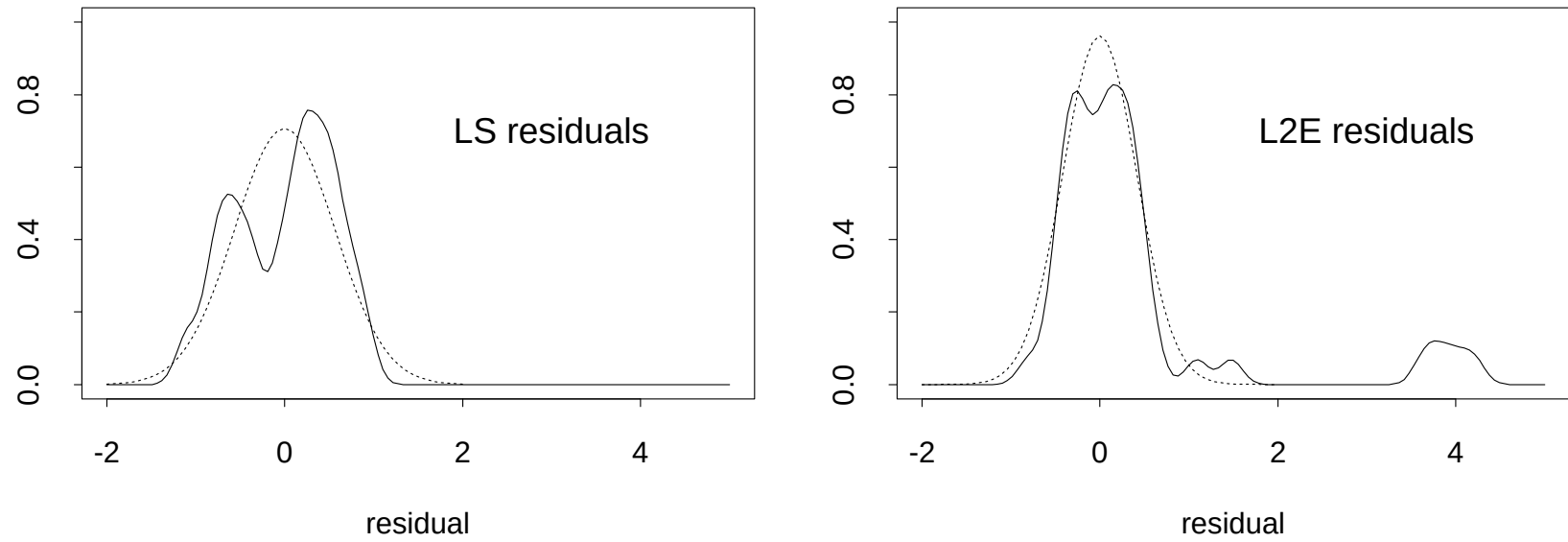


Figure 26: Residual plots for the star data. The assumed $N(0, \sigma_c^2)$ fit

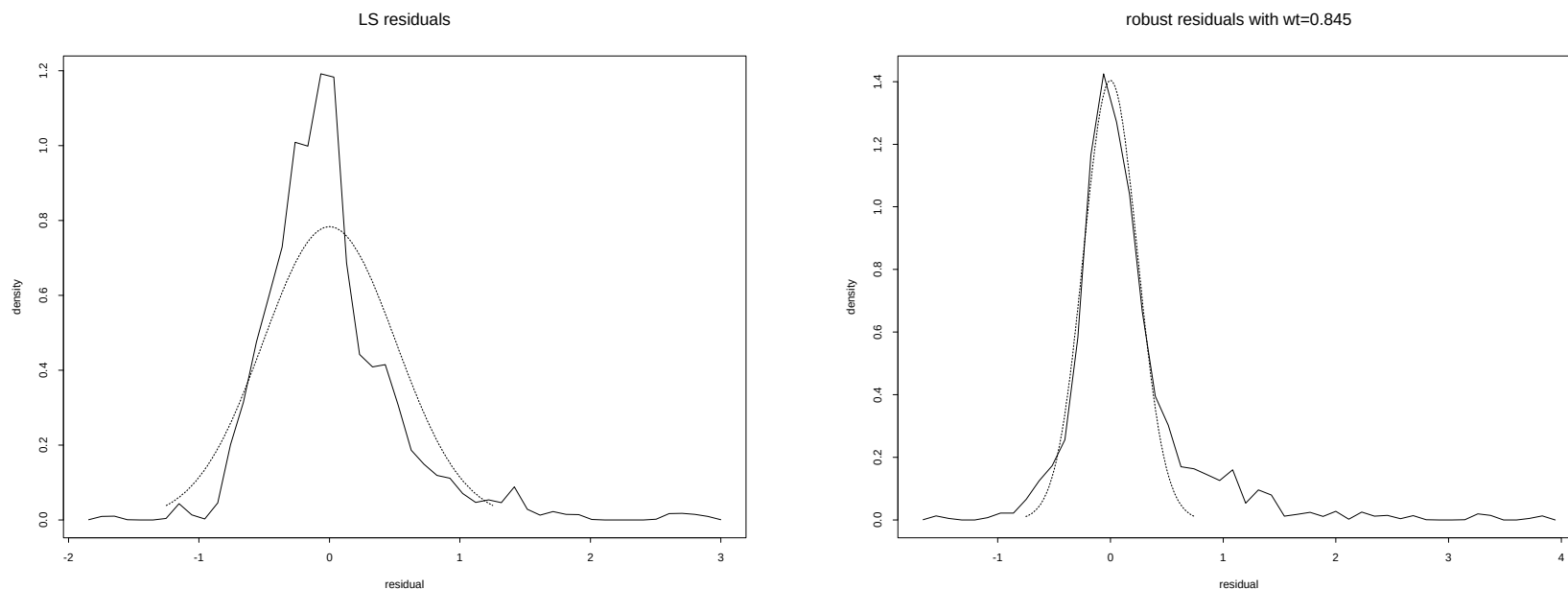


Figure 27: Left frame: Kernel estimate of the residuals for a least-squares fit of the Boston Housing data, together with the $N(0, \sigma_\epsilon^2)$ fit. Right frame: Kernel estimate of the residuals for the L2E fit of the Boston Housing data, together with the $w \cdot N(0, \sigma_\epsilon^2)$ fit, where $\hat{w} = 0.845$.

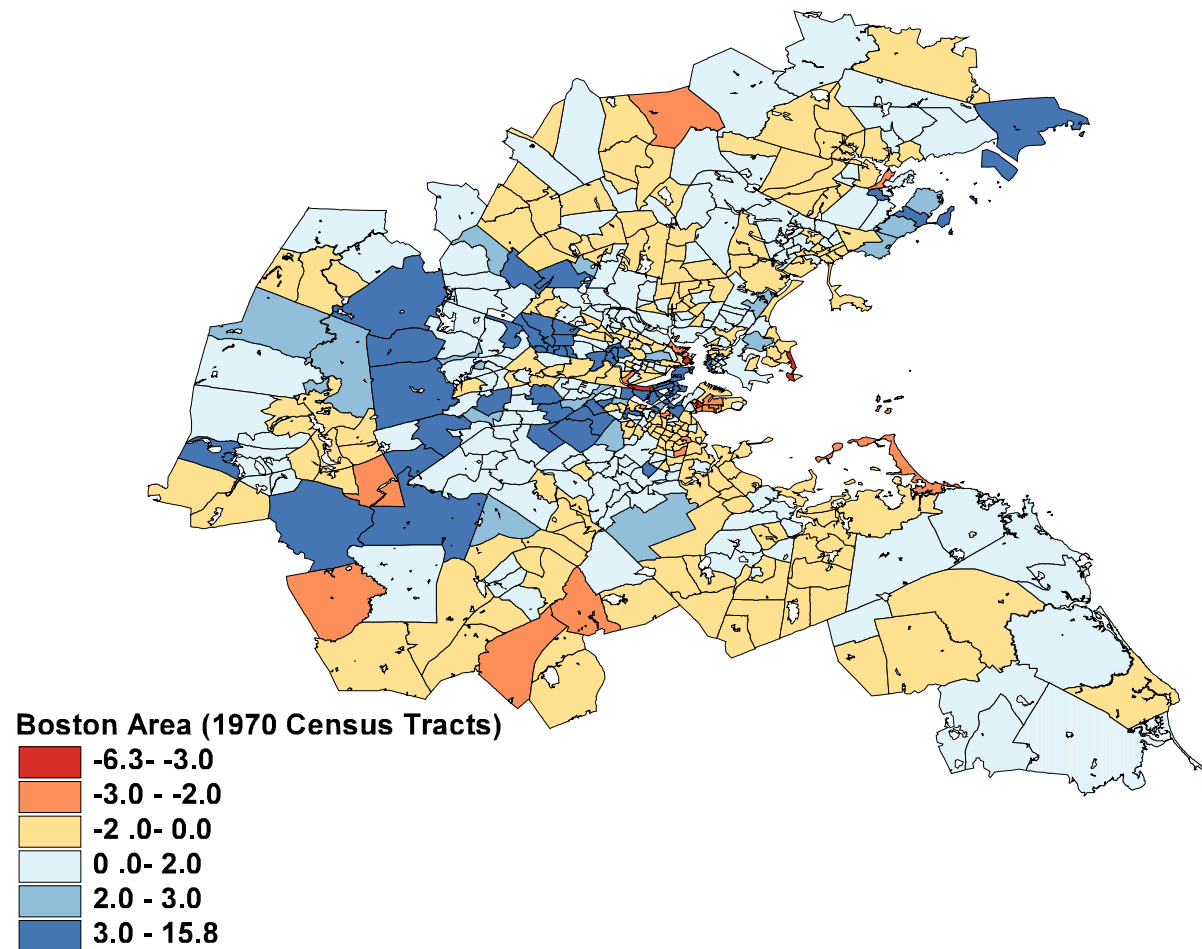


Figure 28: Standardized L2E residuals for the Boston Housing data. Census tracts colored dark red are more than 3 standard deviations below the predicted median housing value, while dark blue are more than 3 σ 's above the predicted value.

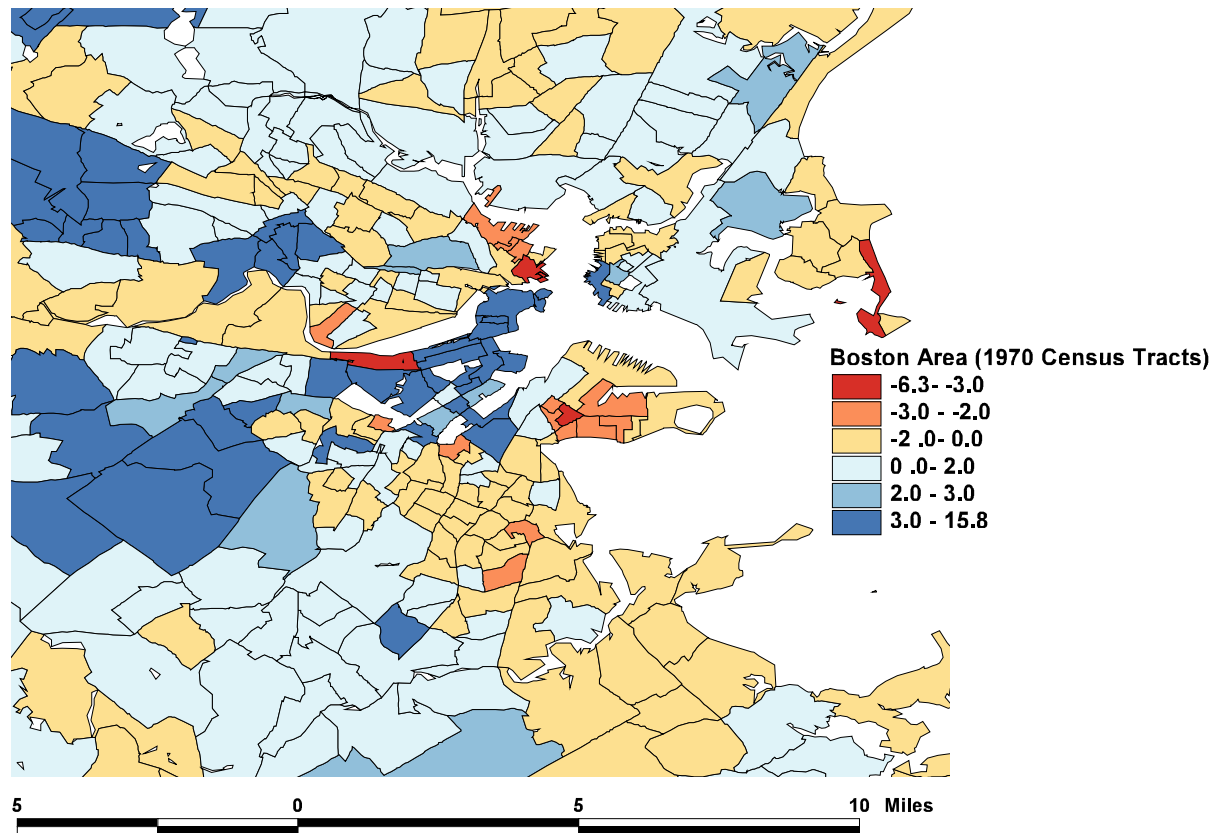


Figure 29: Detail of previous figure.

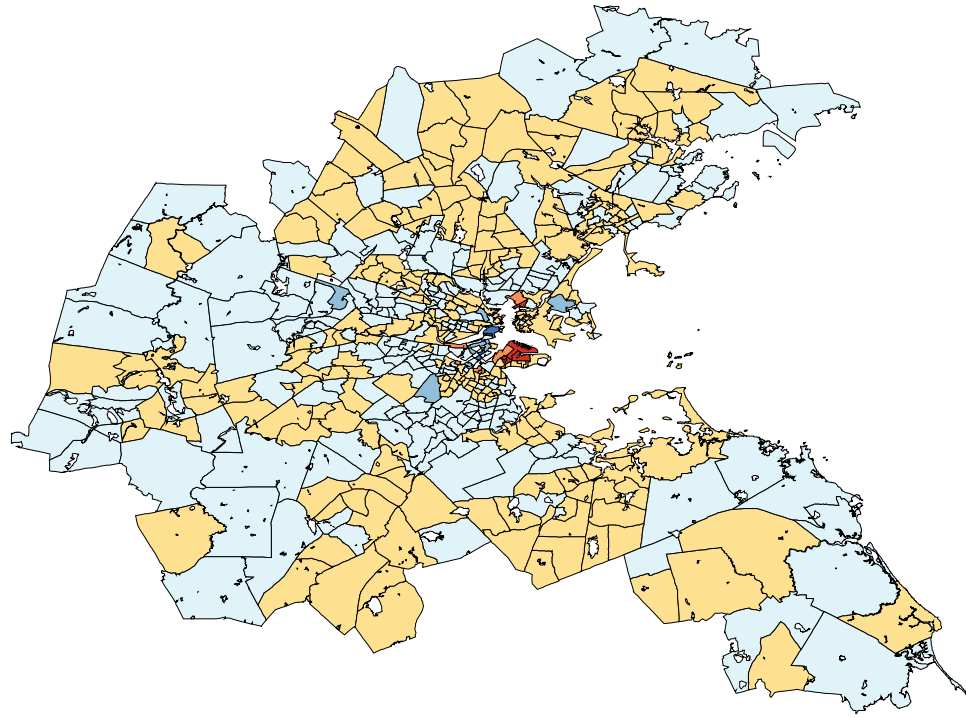


Figure 30: Standardized least squares residuals Boston Housing data.

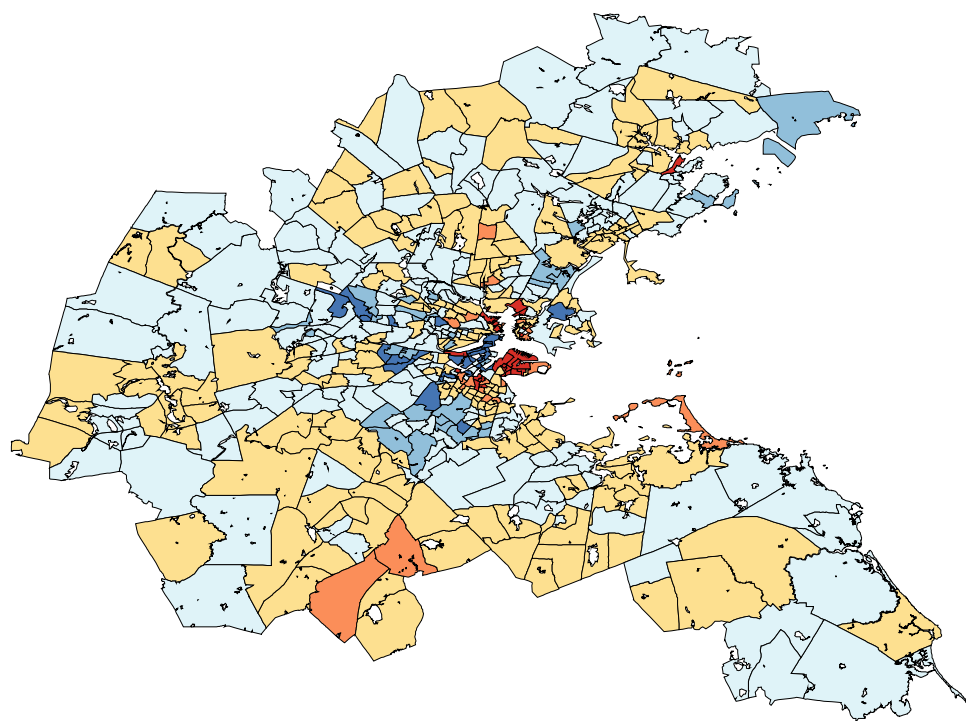


Figure 31: L2E residuals Boston.

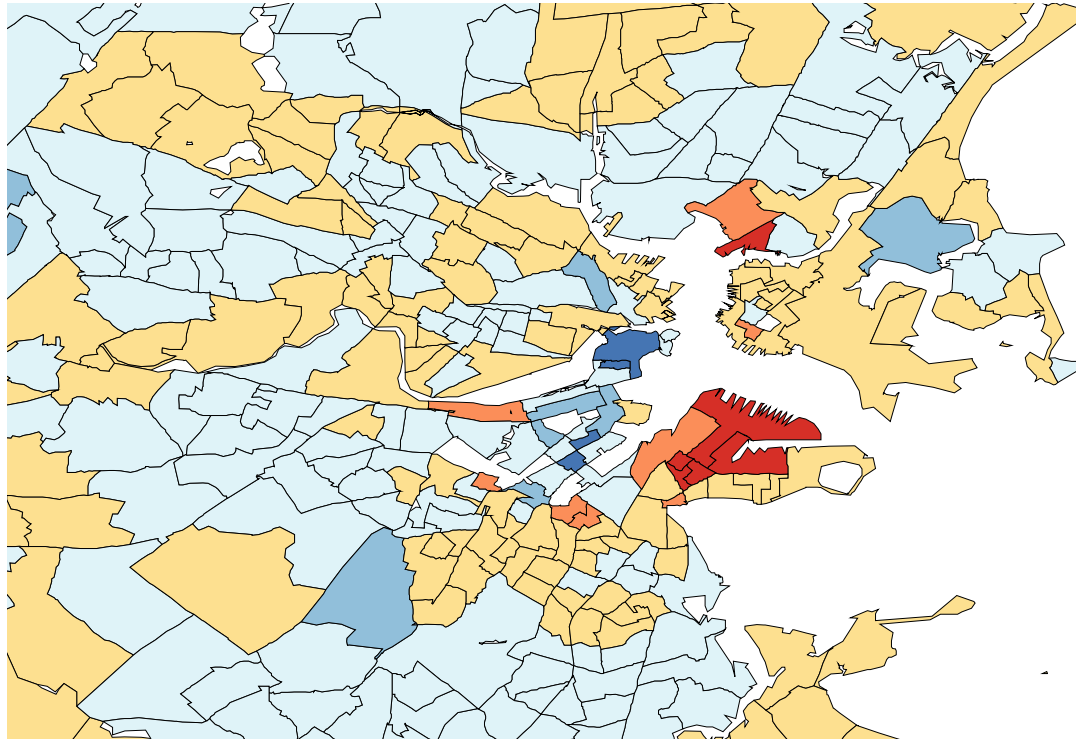


Figure 32: Least squares residuals central Boston.

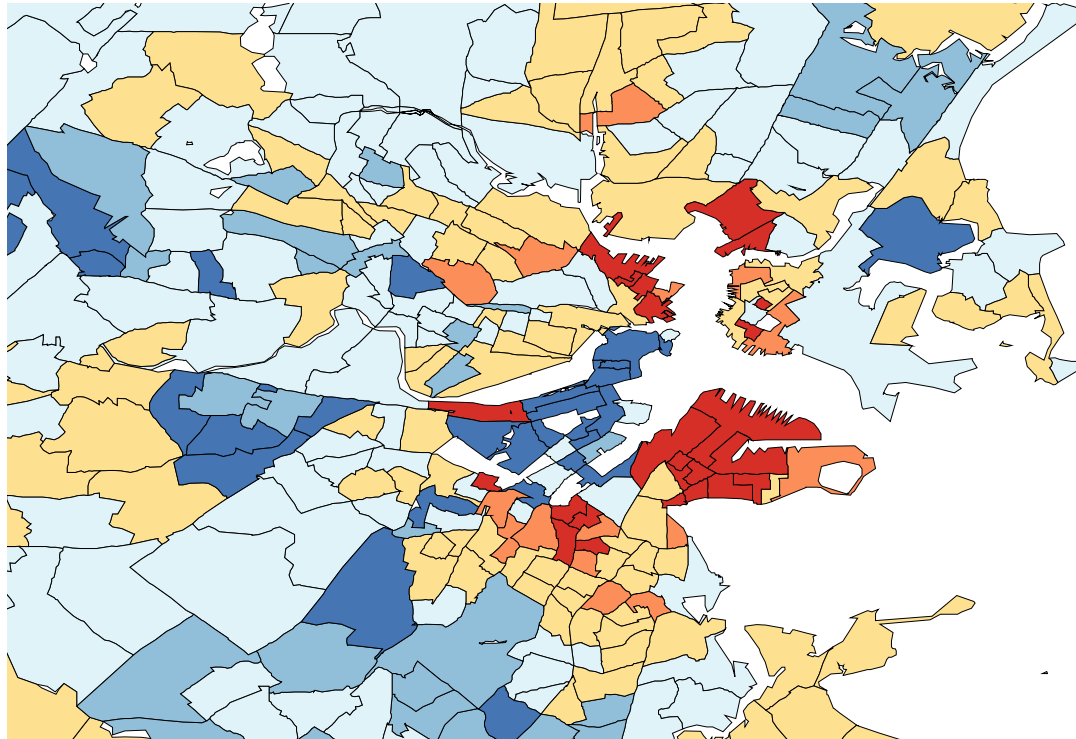


Figure 33: L2E residuals central Boston.

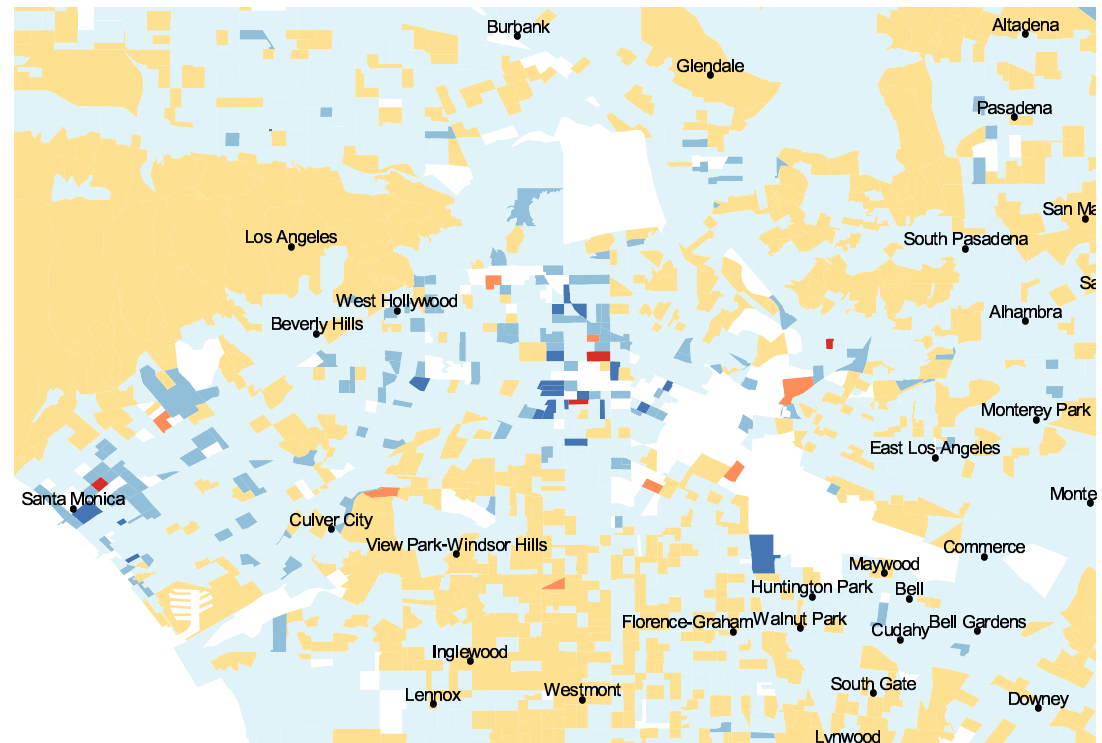


Figure 34: Detail of least squares residuals HOLLYWOOD area.

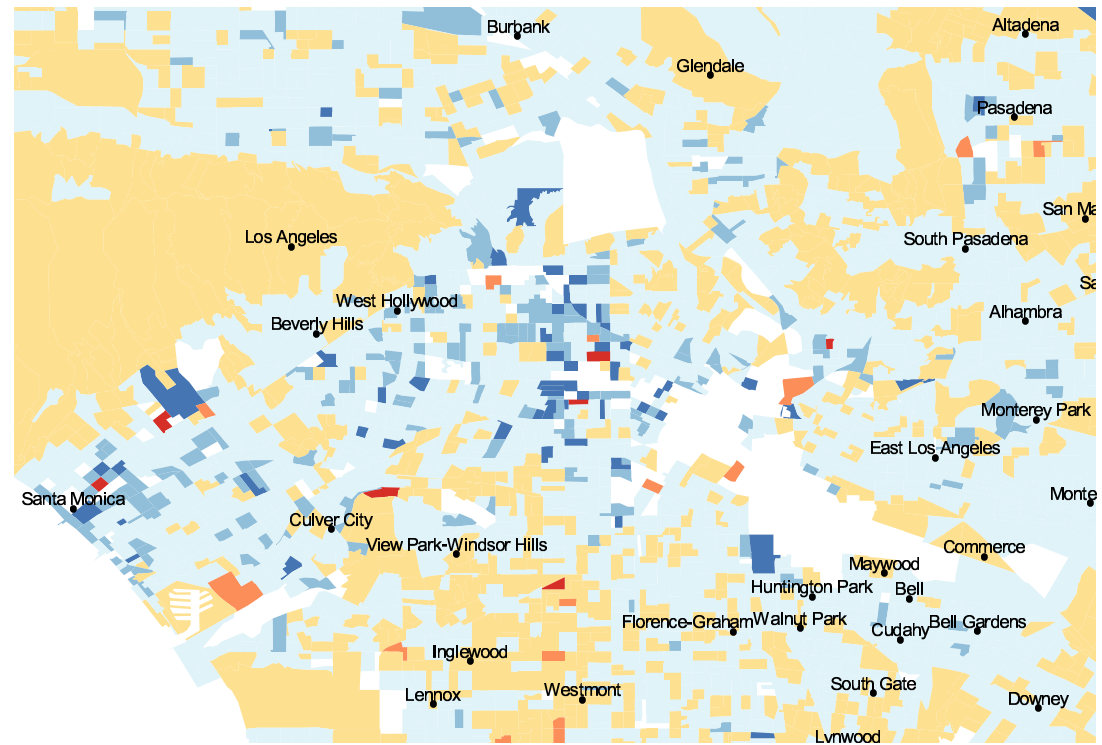


Figure 35: Detail of L2E residuals HOLLYWOOD area.

6 Gaussian Mixture Regression and IPRA (Iterated Pairwise Replacement Algorithm)

- semiparametric (eg normal mixtures) probably always superior to kernels beyond 4 or 5 dimensions
- note that kernel estimators are also normal mixtures ($K = n$)
- find $K \ll n$ with mixture “close” to the kernel estimate
- IPRA (Scott and Szewczyk, 2001, Technometrics) iterated pairwise replacement algorithm
- visualize parameters of mixture or projections
- if $K = 17$ in $\mathcal{R}^{25}$ with EM and components $N(\mu_k, D_k)$, is real $f(x)$ just one $N(\mu, \Sigma)$ or not?
- GMR - gaussian mixture regression/classification (Hsi-Guang Sung, 2004 thesis): IPRA + peeling

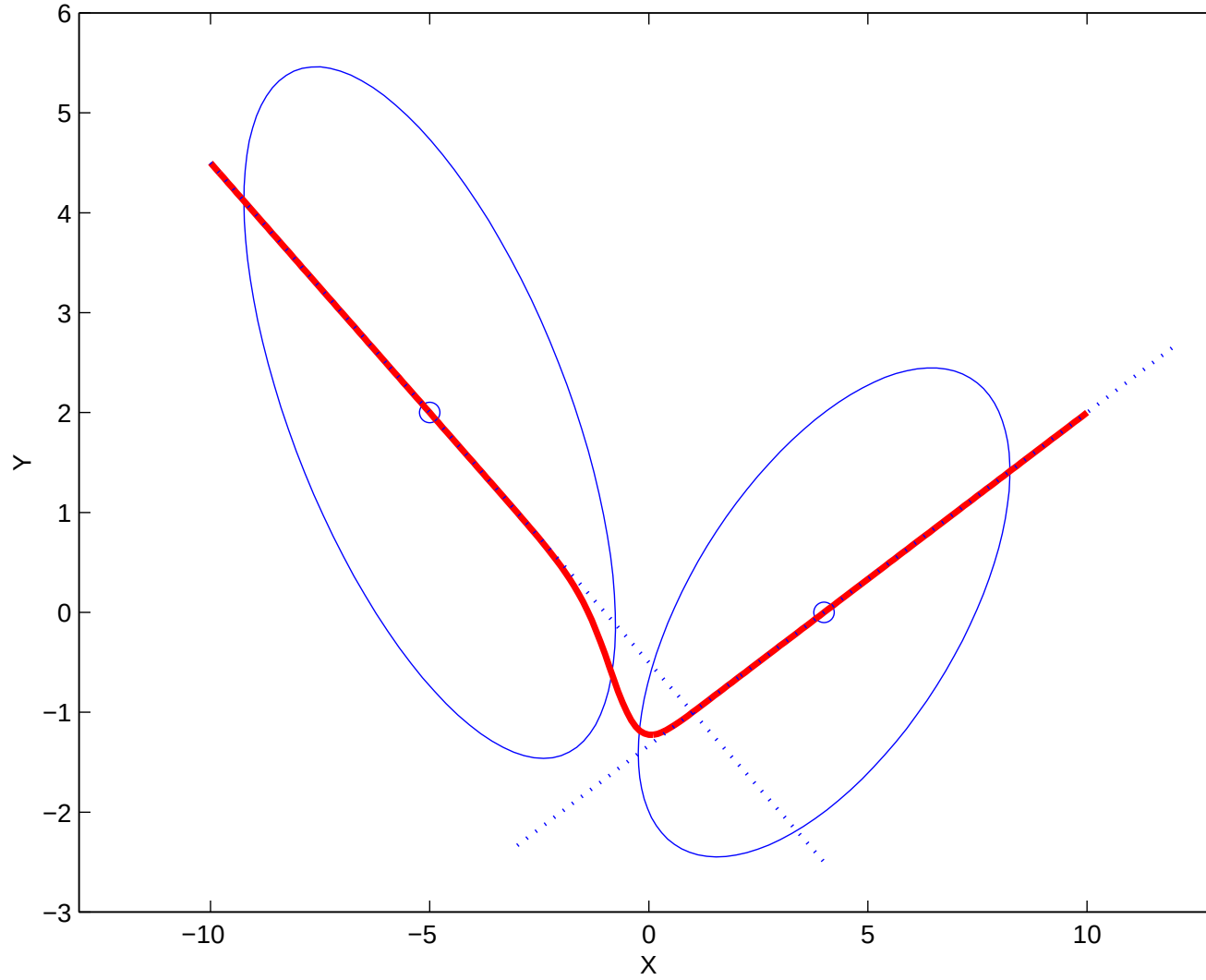


Figure 36: Gaussian mixture regression example (GMR).

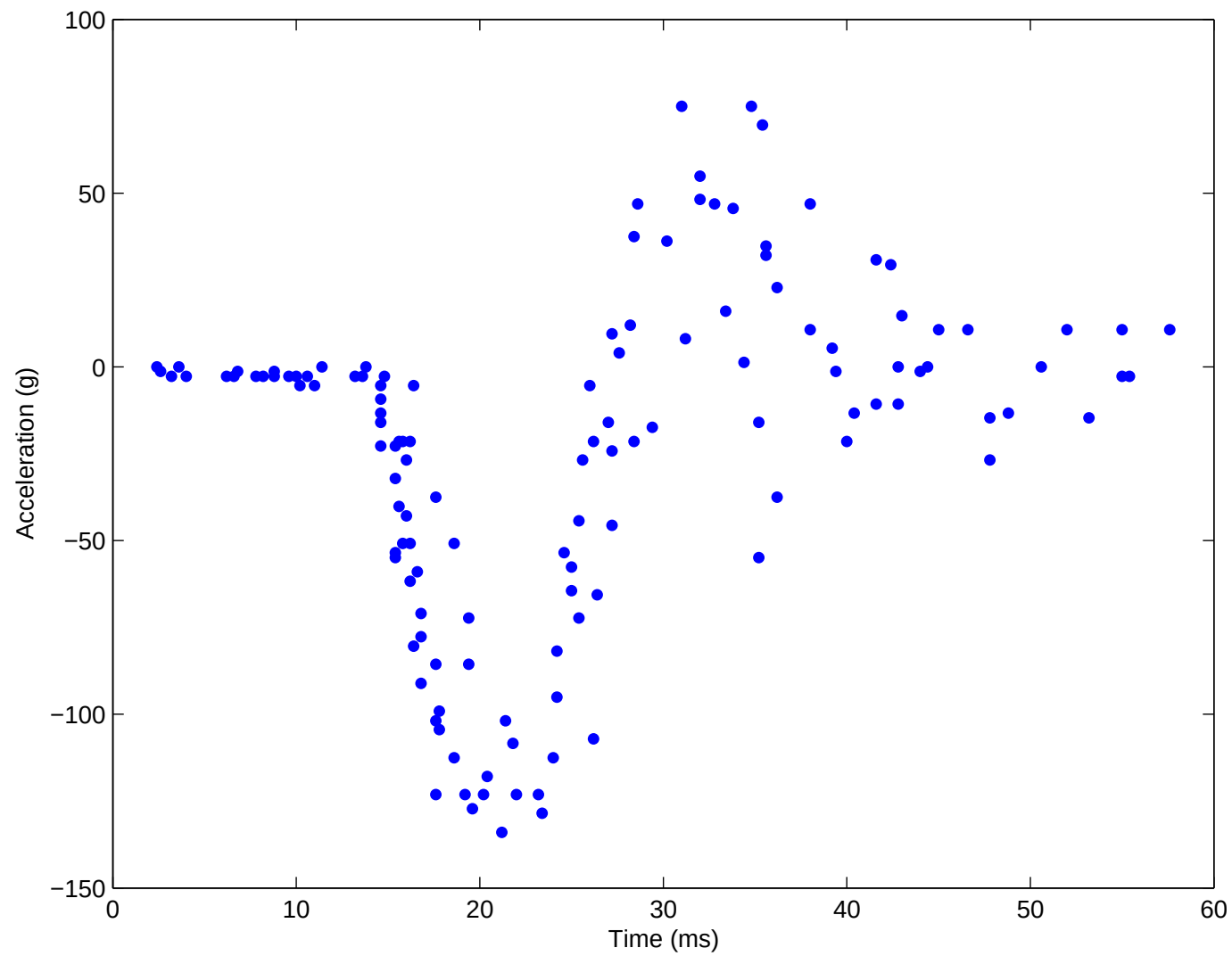


Figure 37: Motorcycle data.

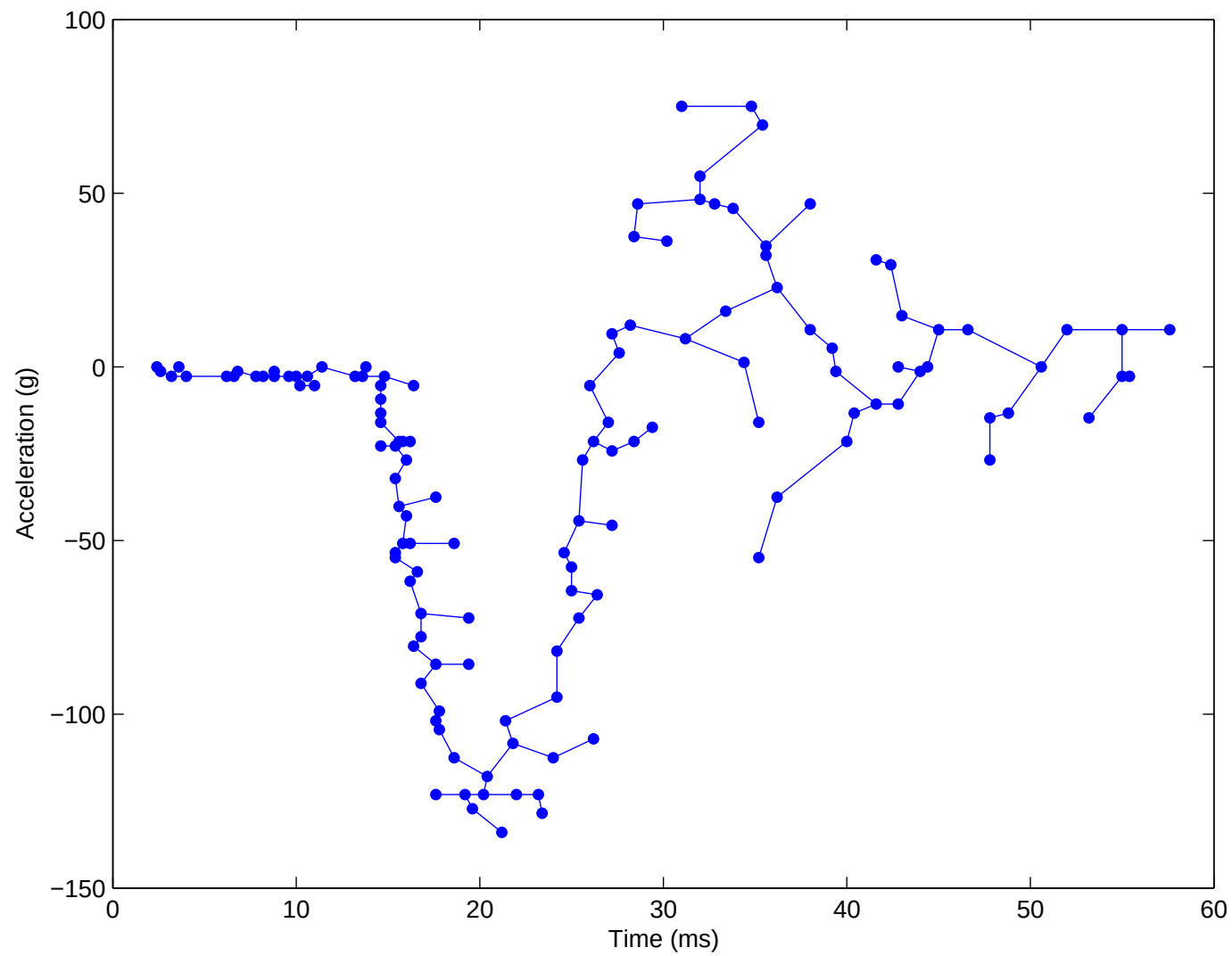


Figure 38: Motorcycle data with minimum spanning tree.

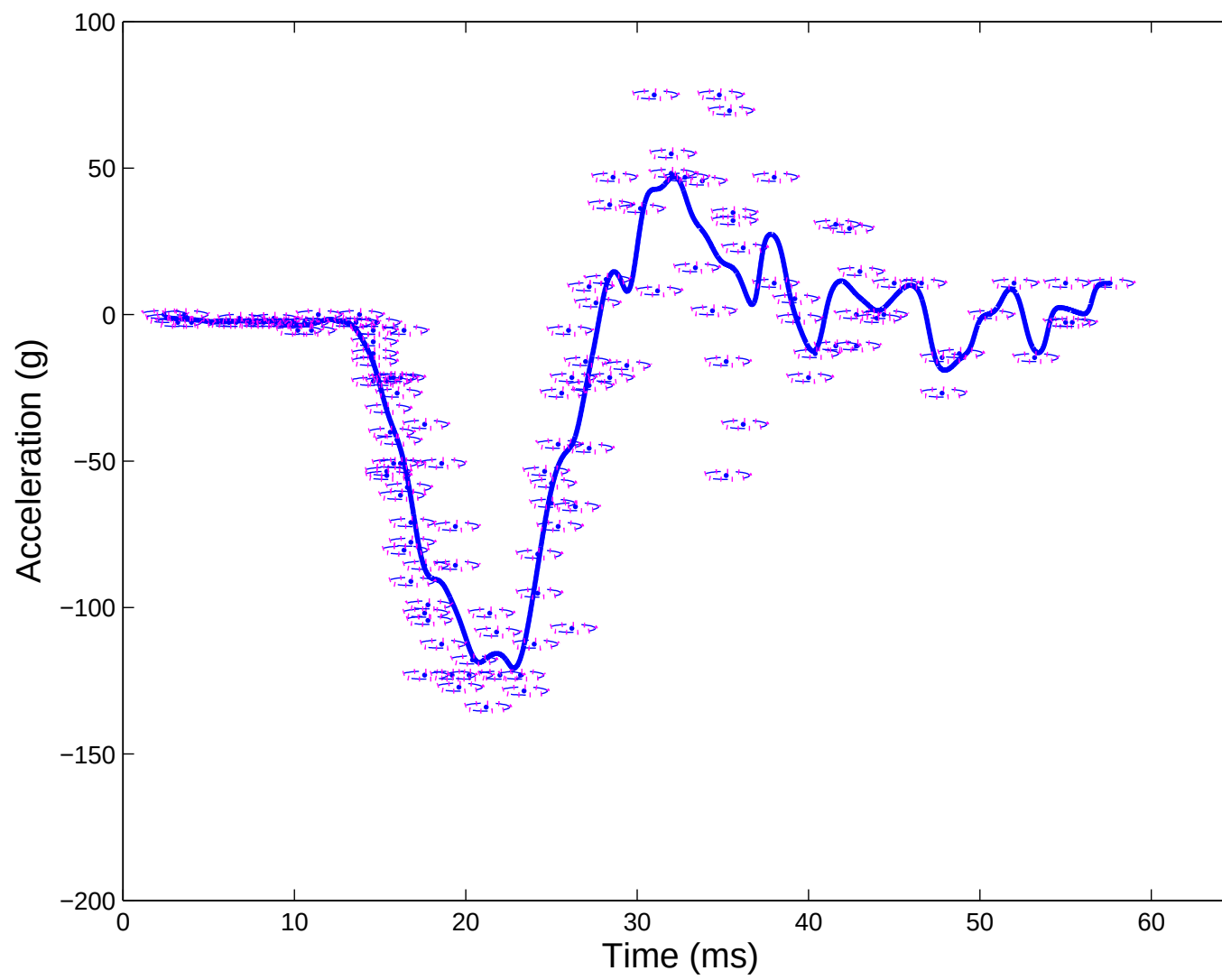


Figure 39: GMR(133).

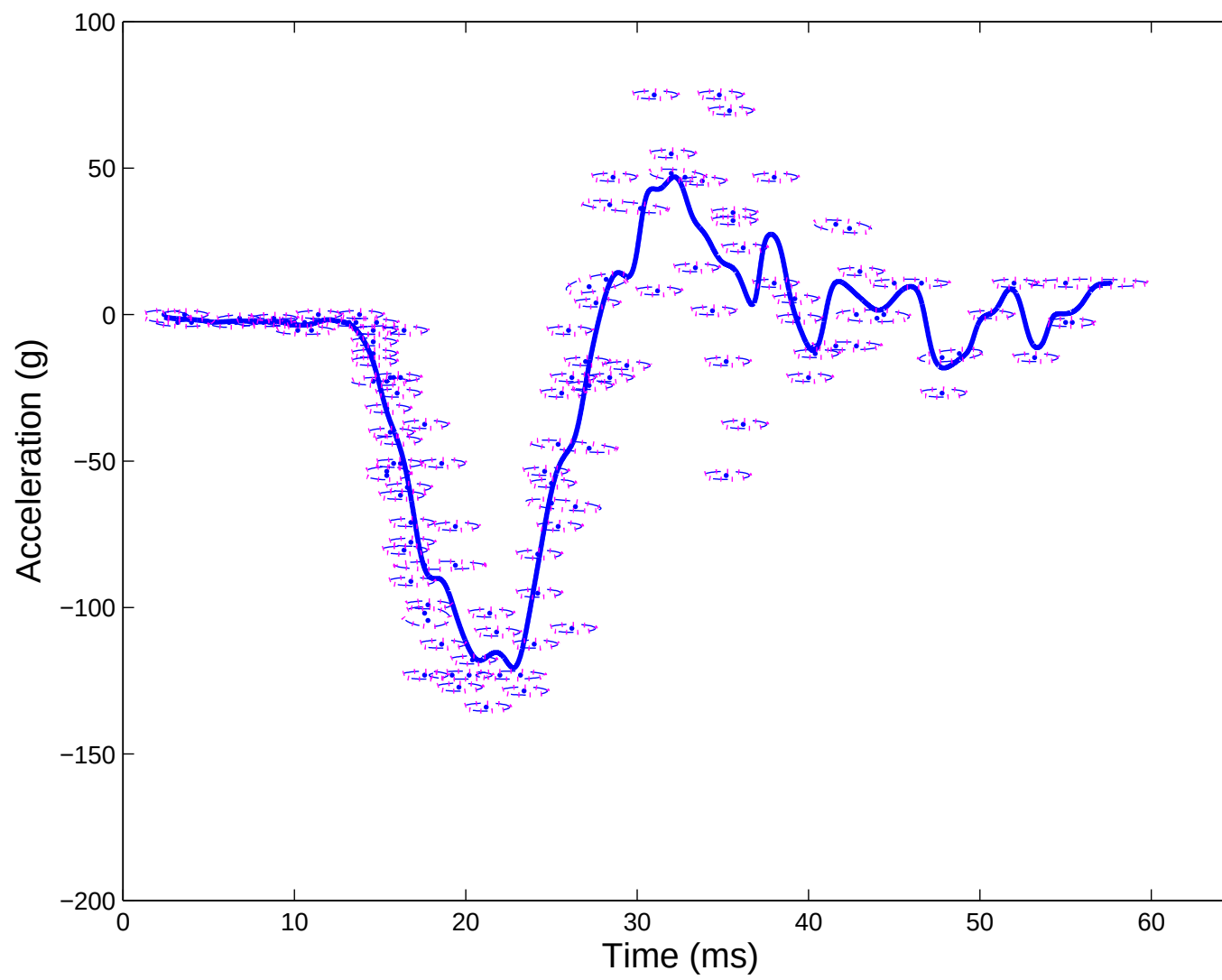


Figure 40: GMR(100).

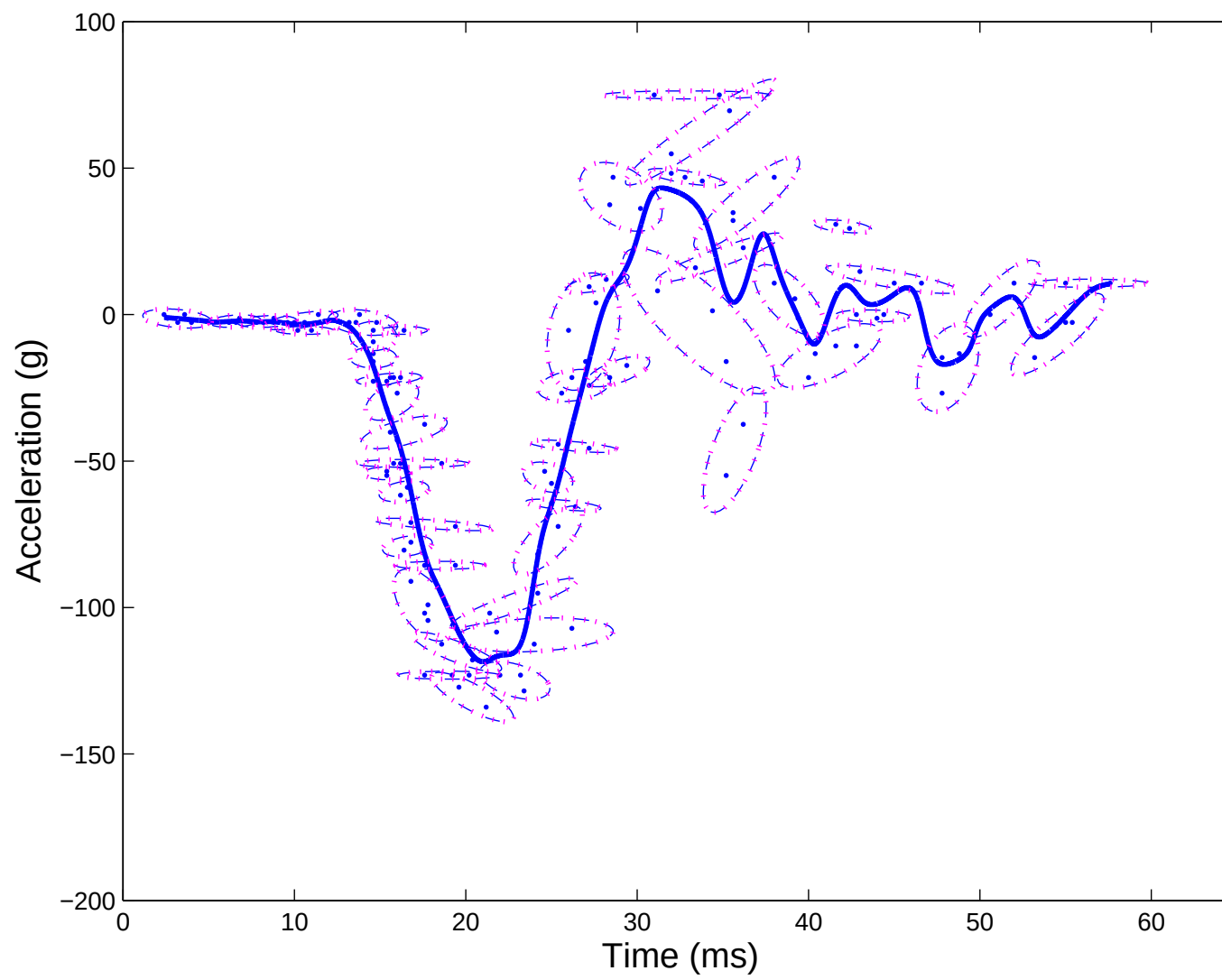


Figure 41: GMR(50).

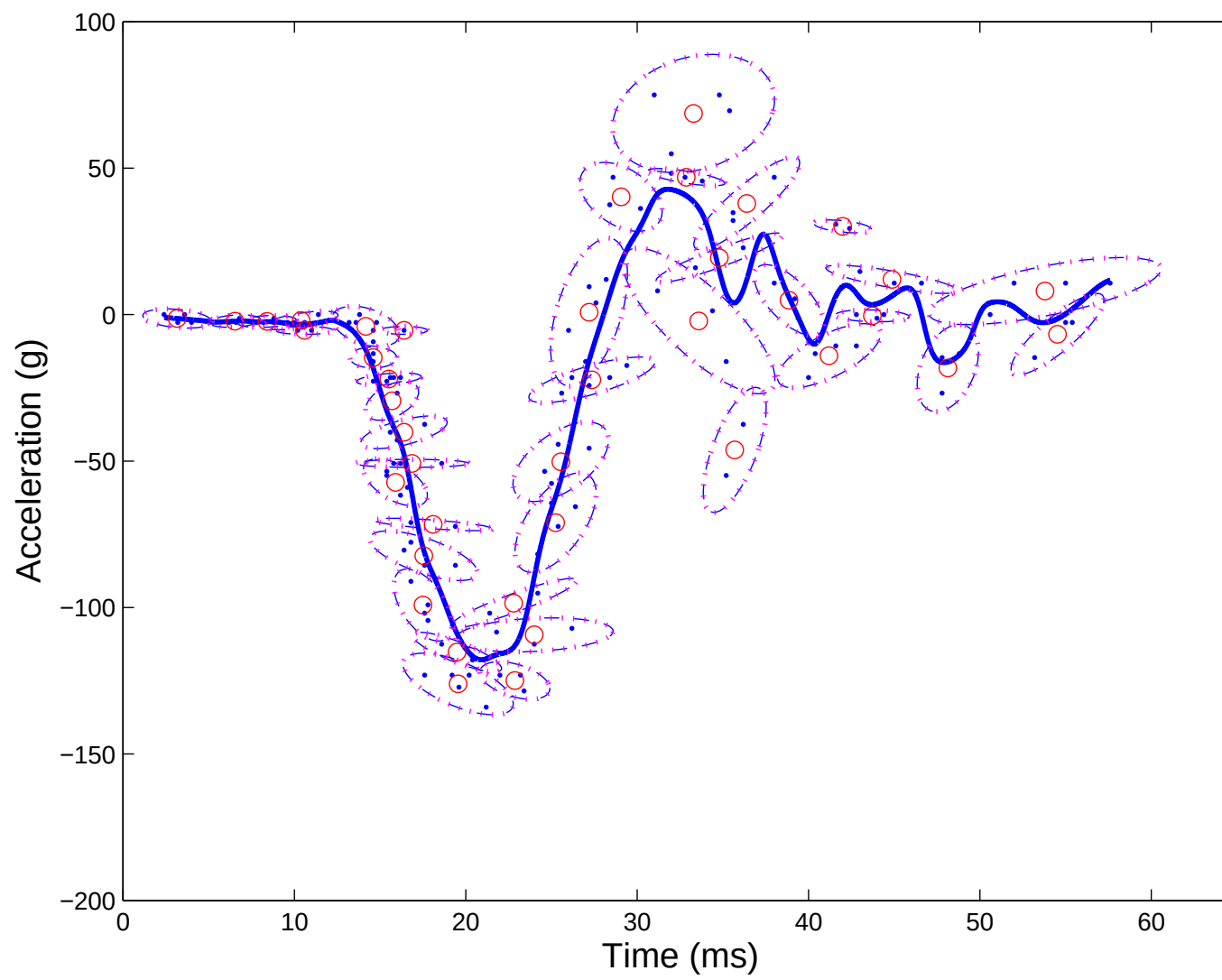


Figure 42: GMR(40).

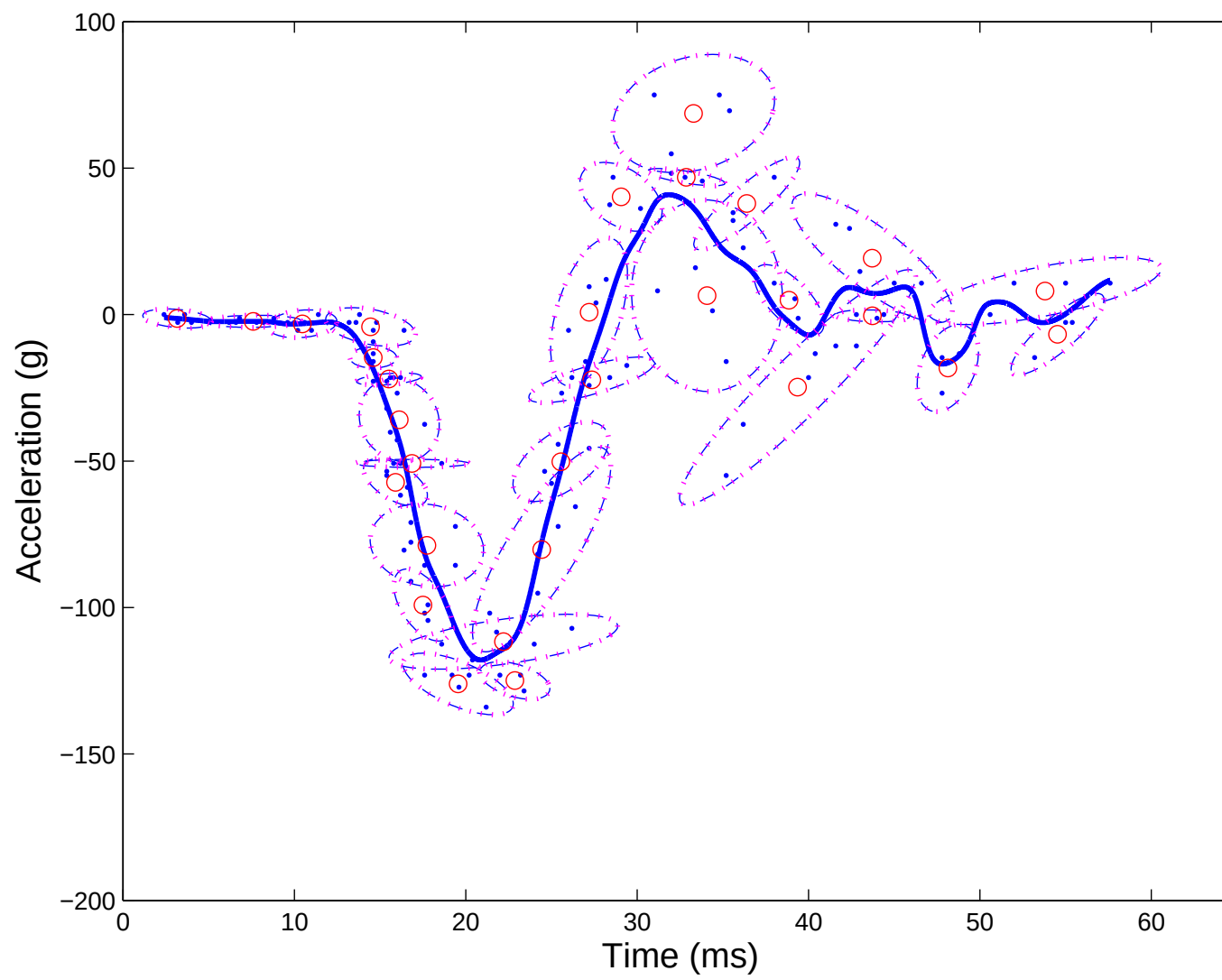


Figure 43: GMR(30).

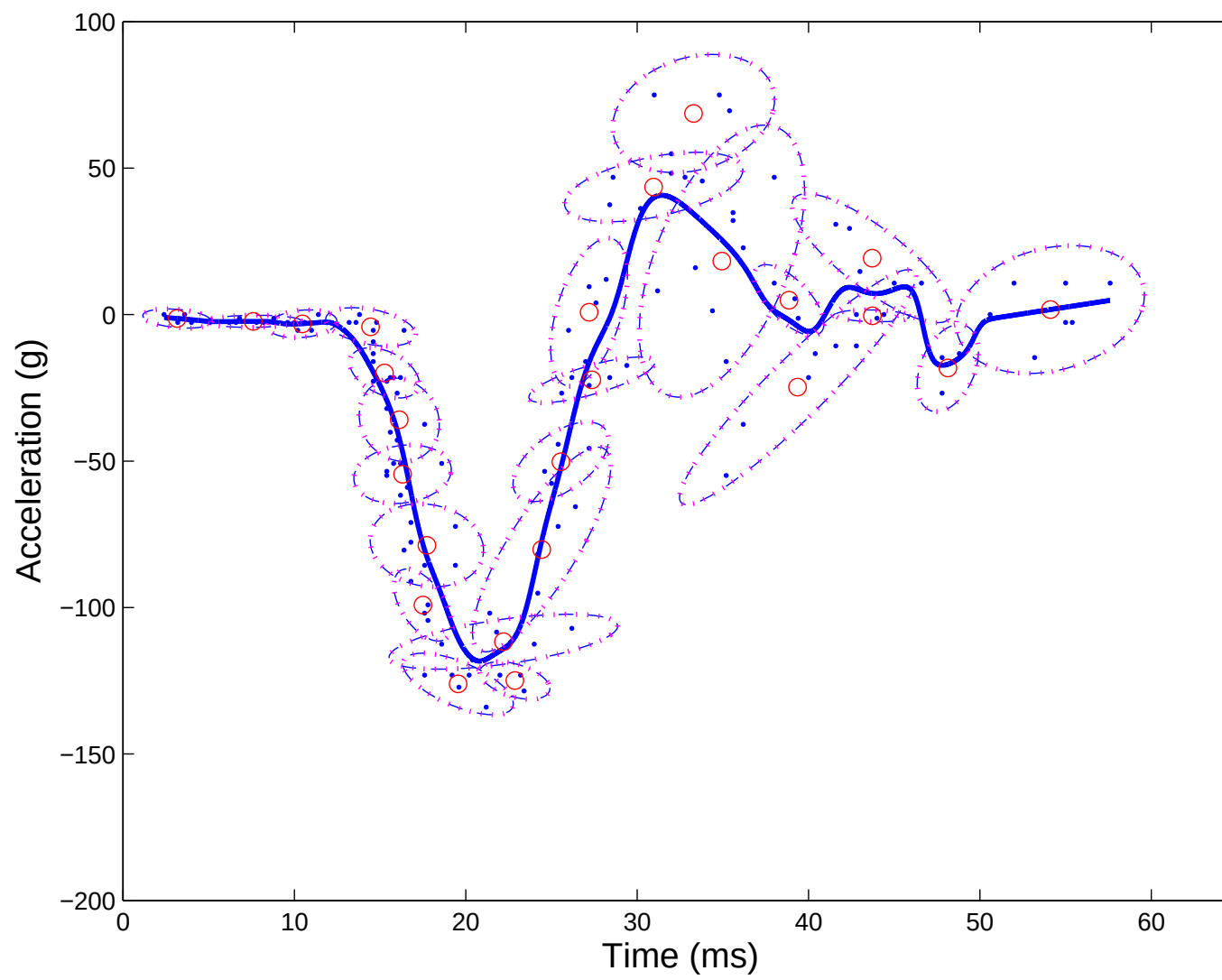


Figure 44: GMR(25).

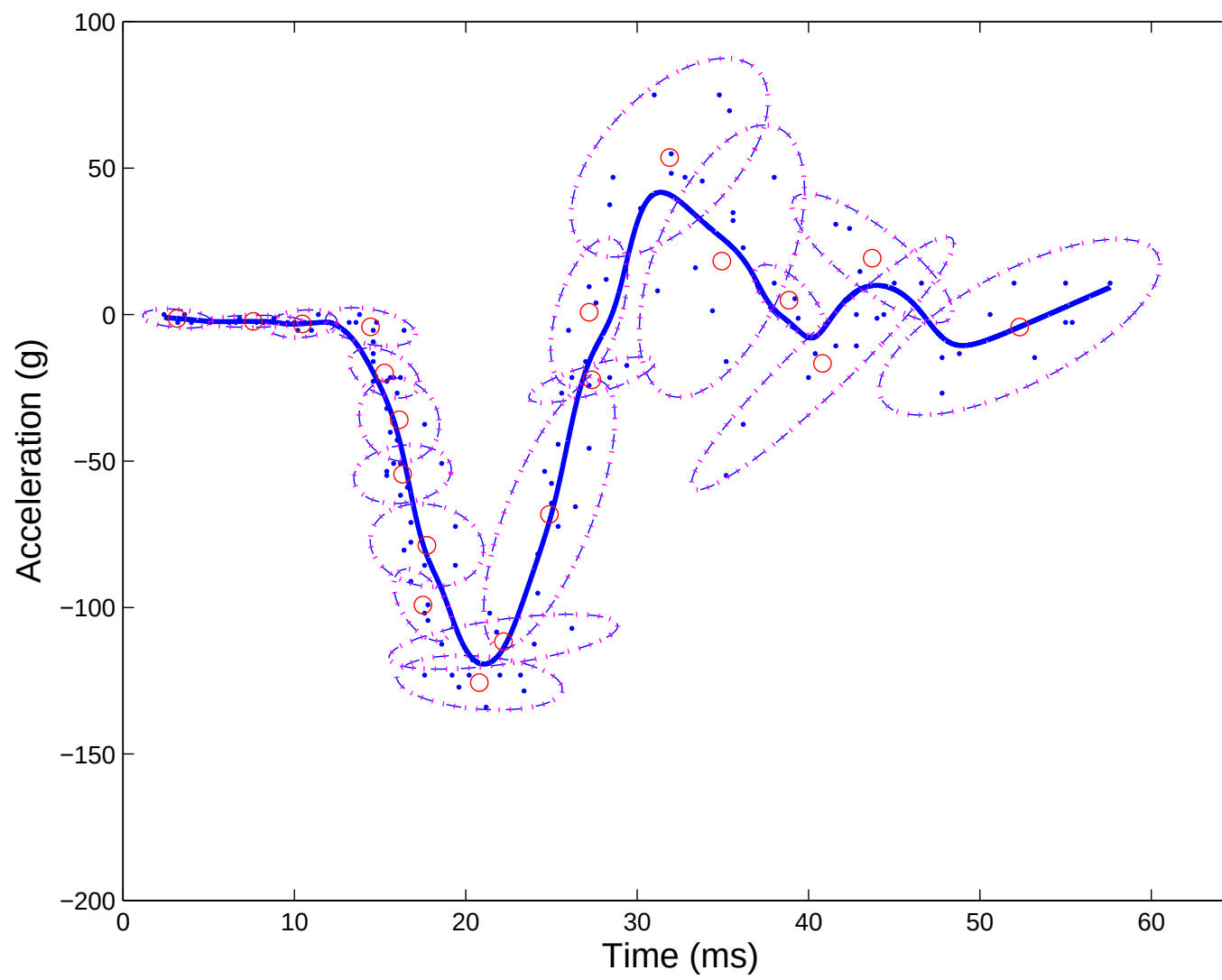


Figure 45: GMR(20).

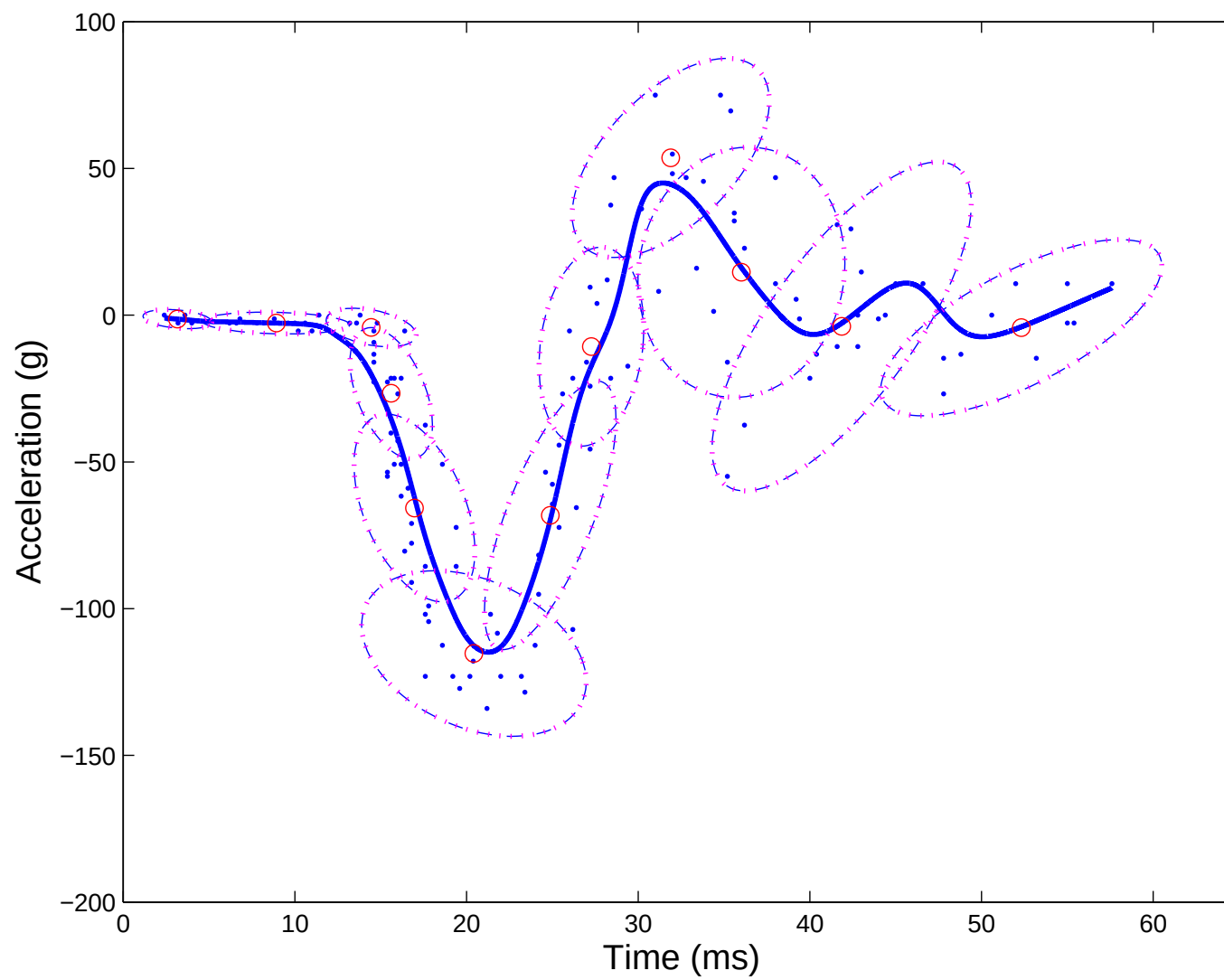


Figure 46: GMR(12).

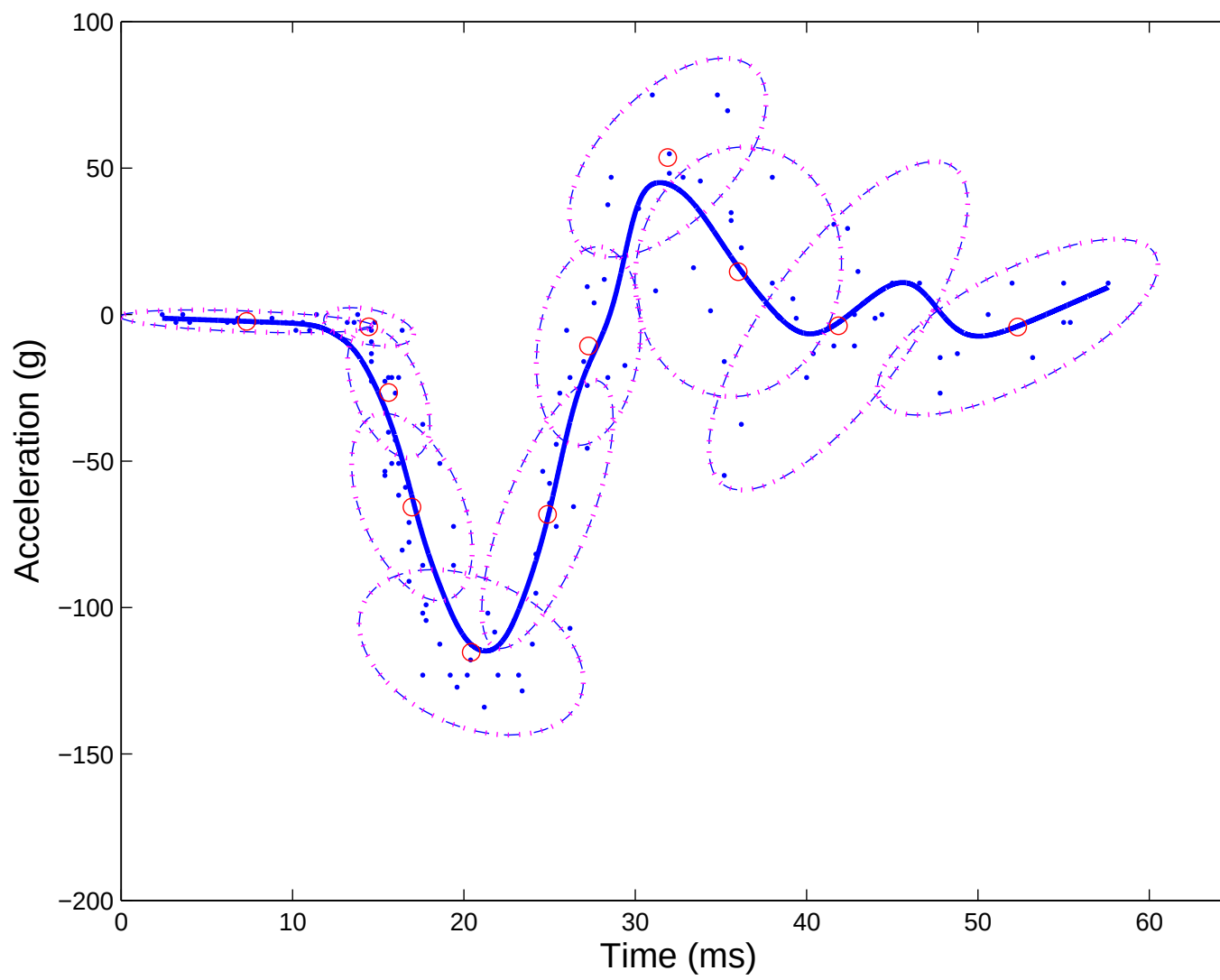


Figure 47: GMR(11).

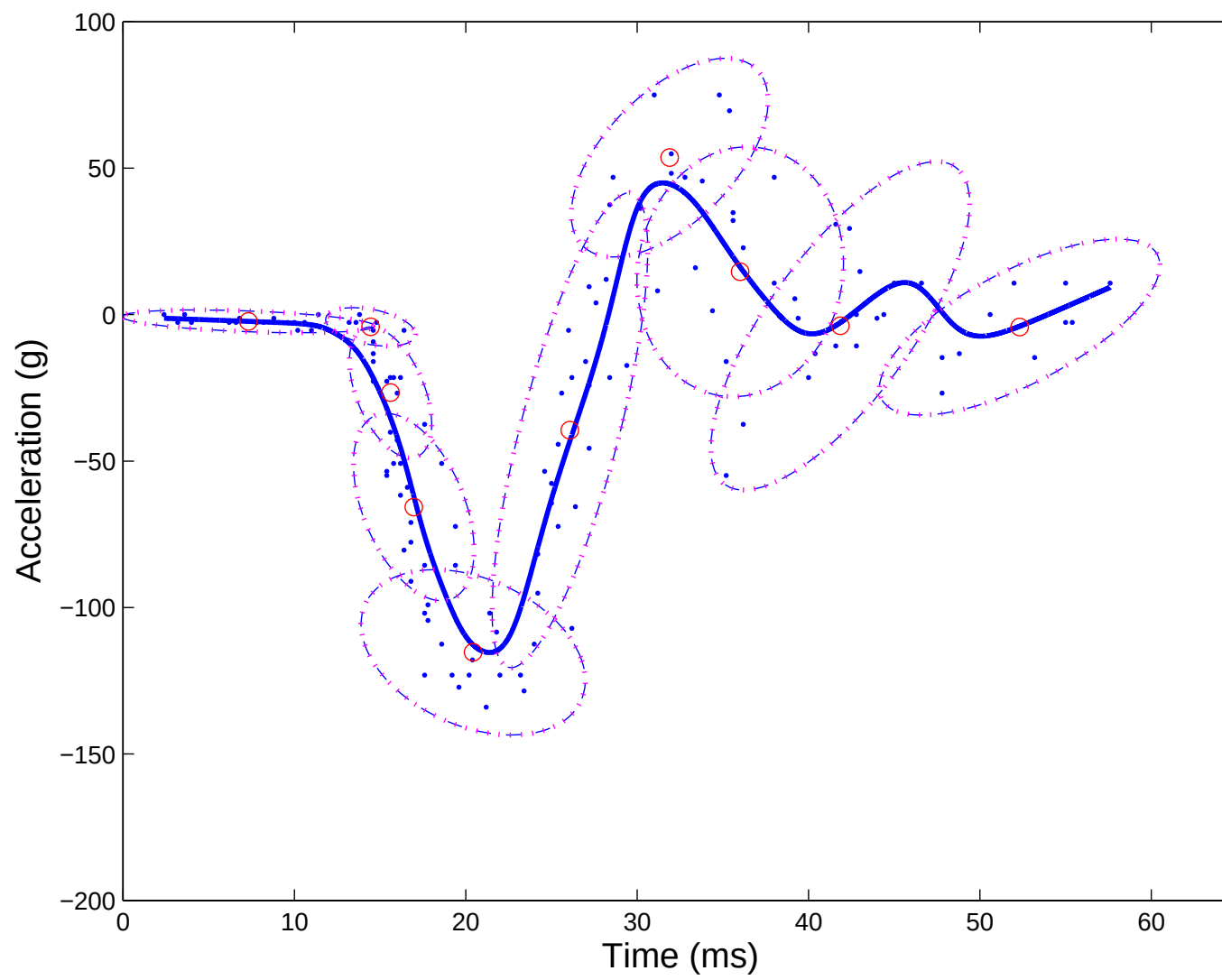


Figure 48: GMR(10).

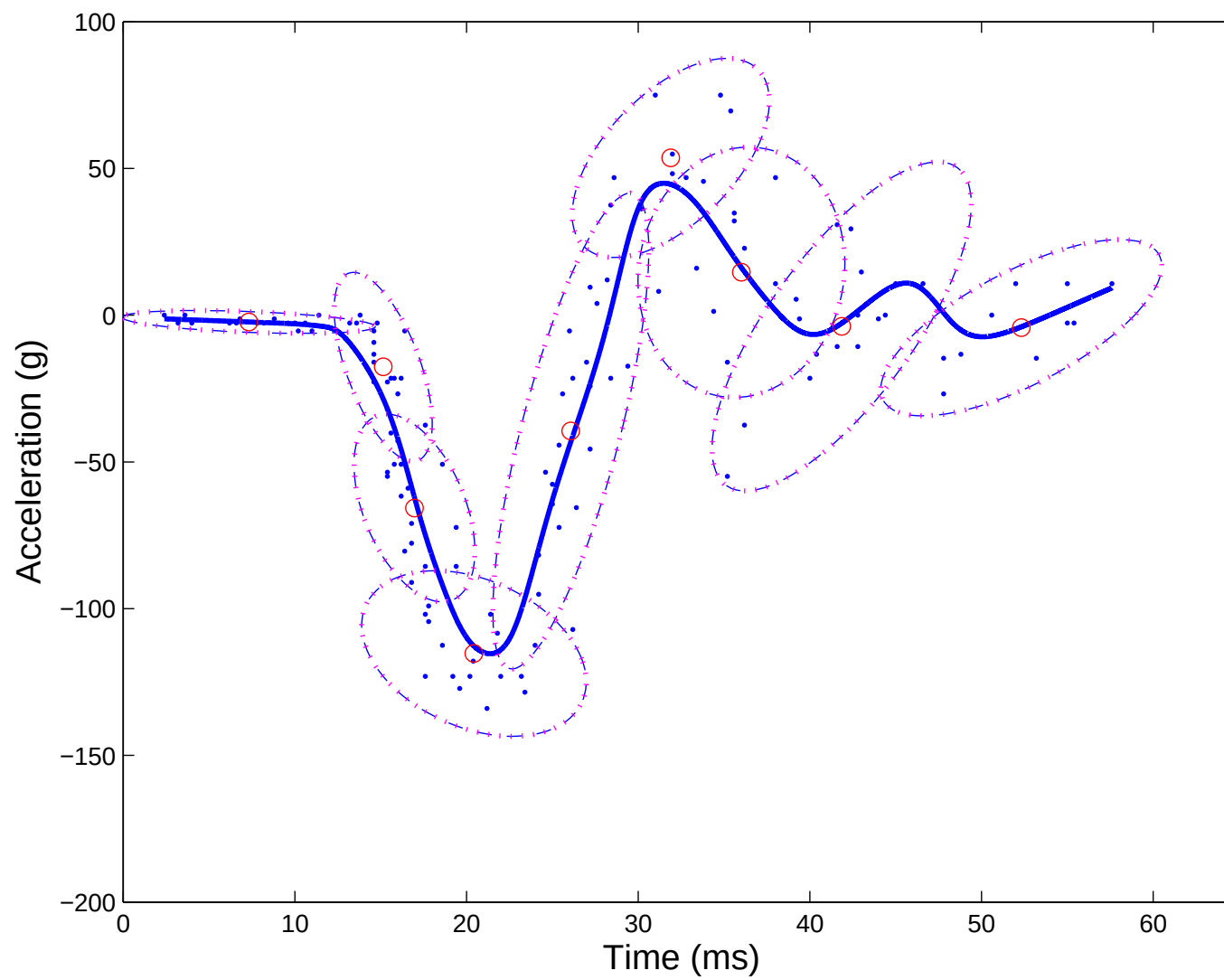


Figure 49: GMR(9).

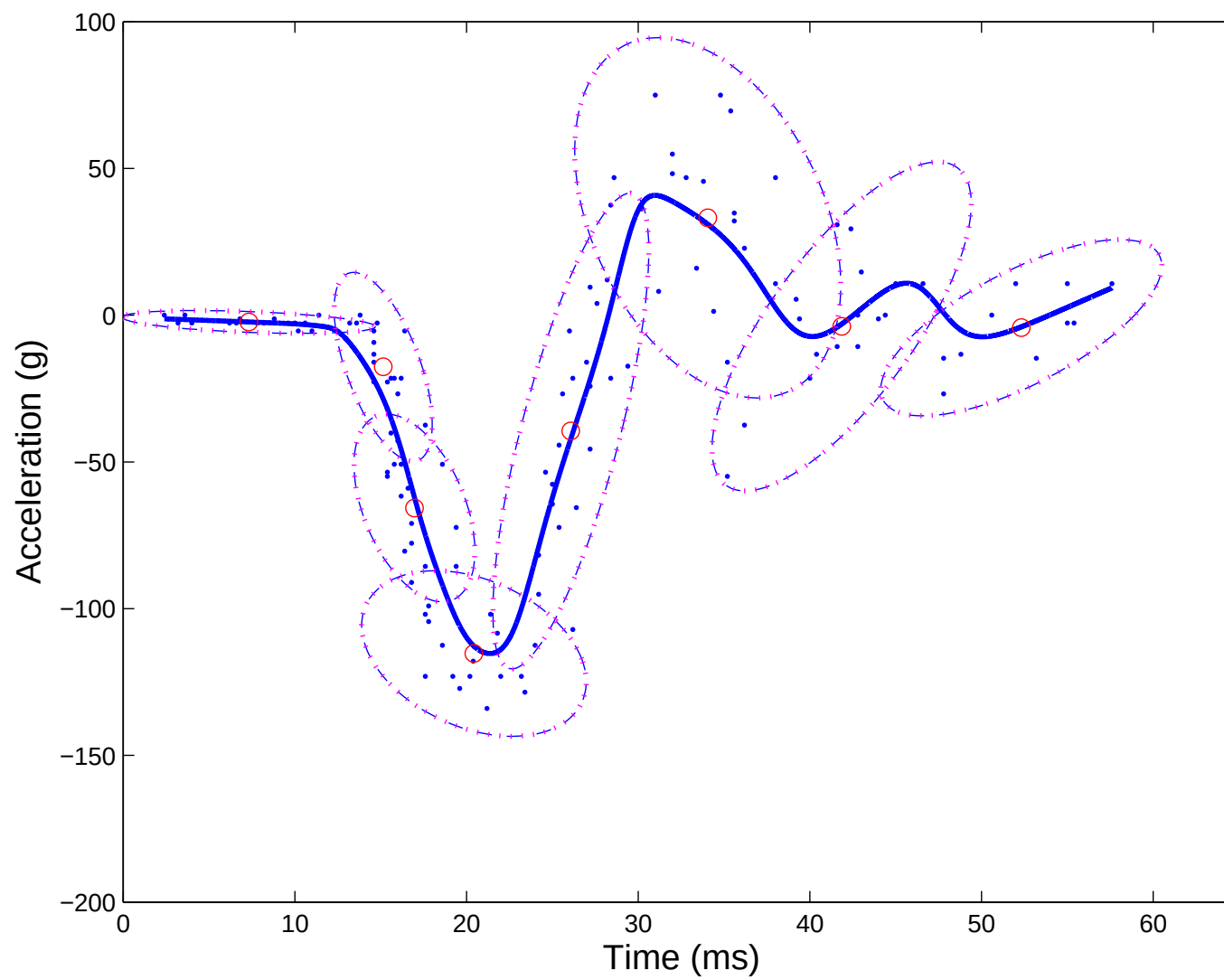


Figure 50: GMR(8).

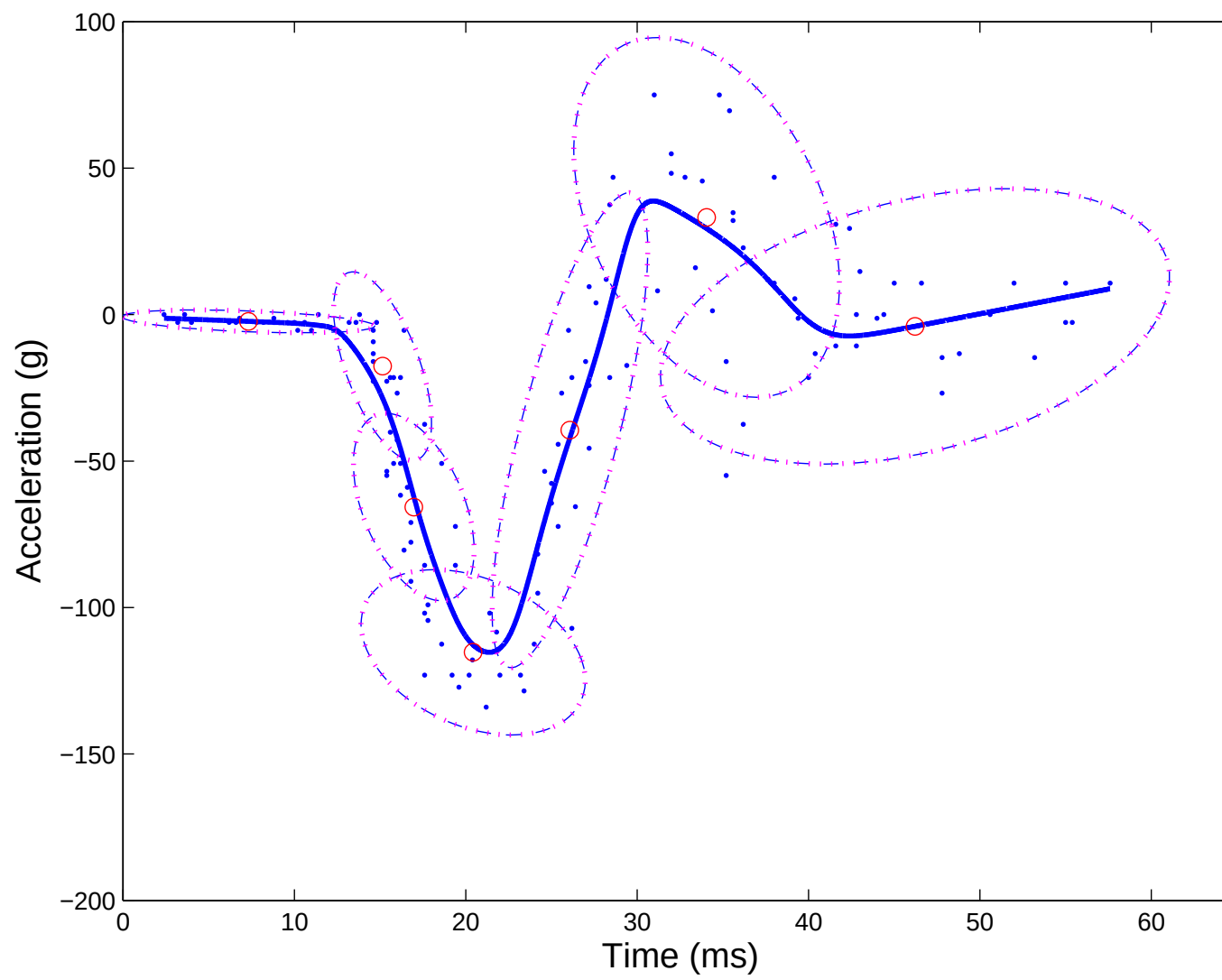


Figure 51: GMR(7).

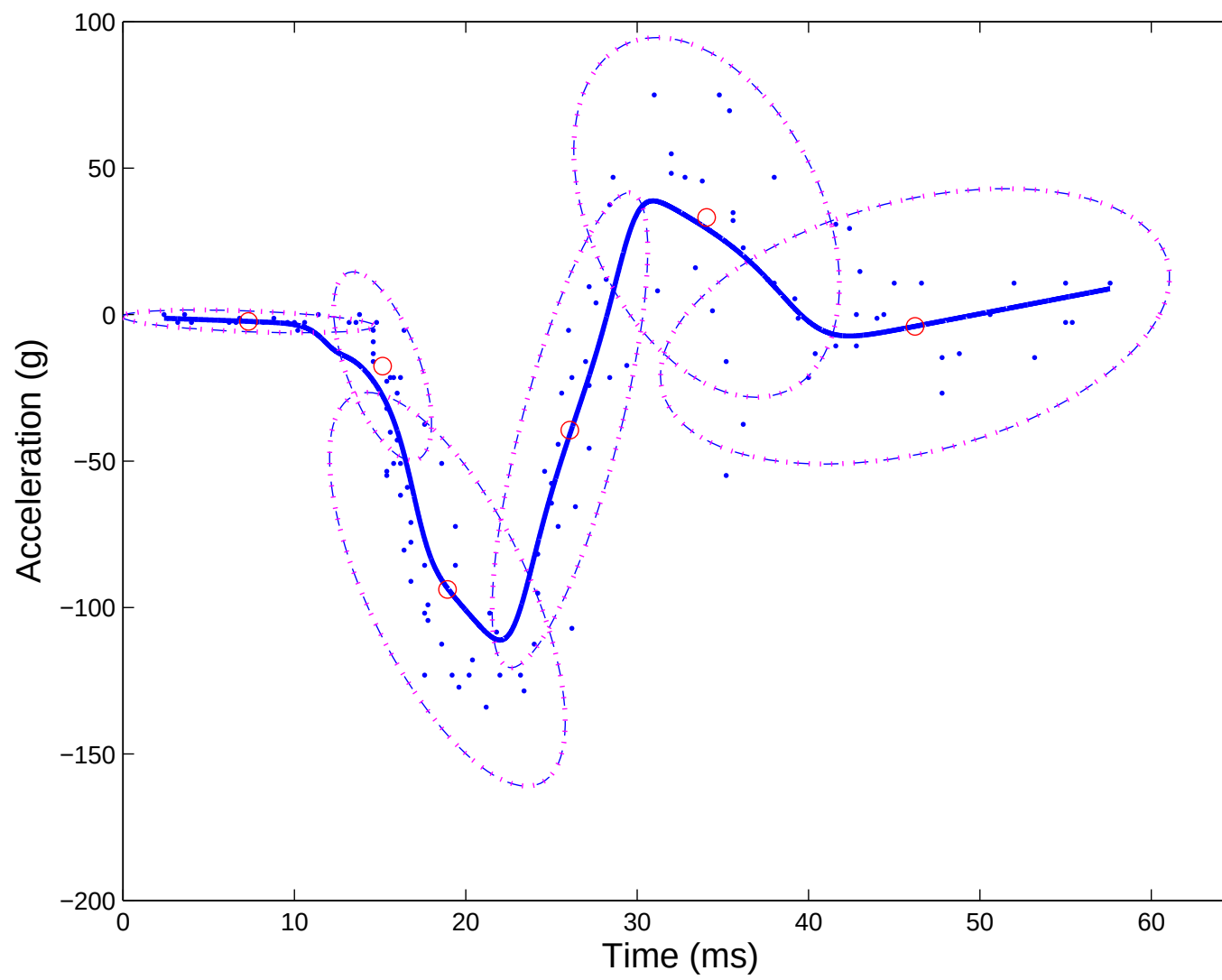


Figure 52: GMR(6).

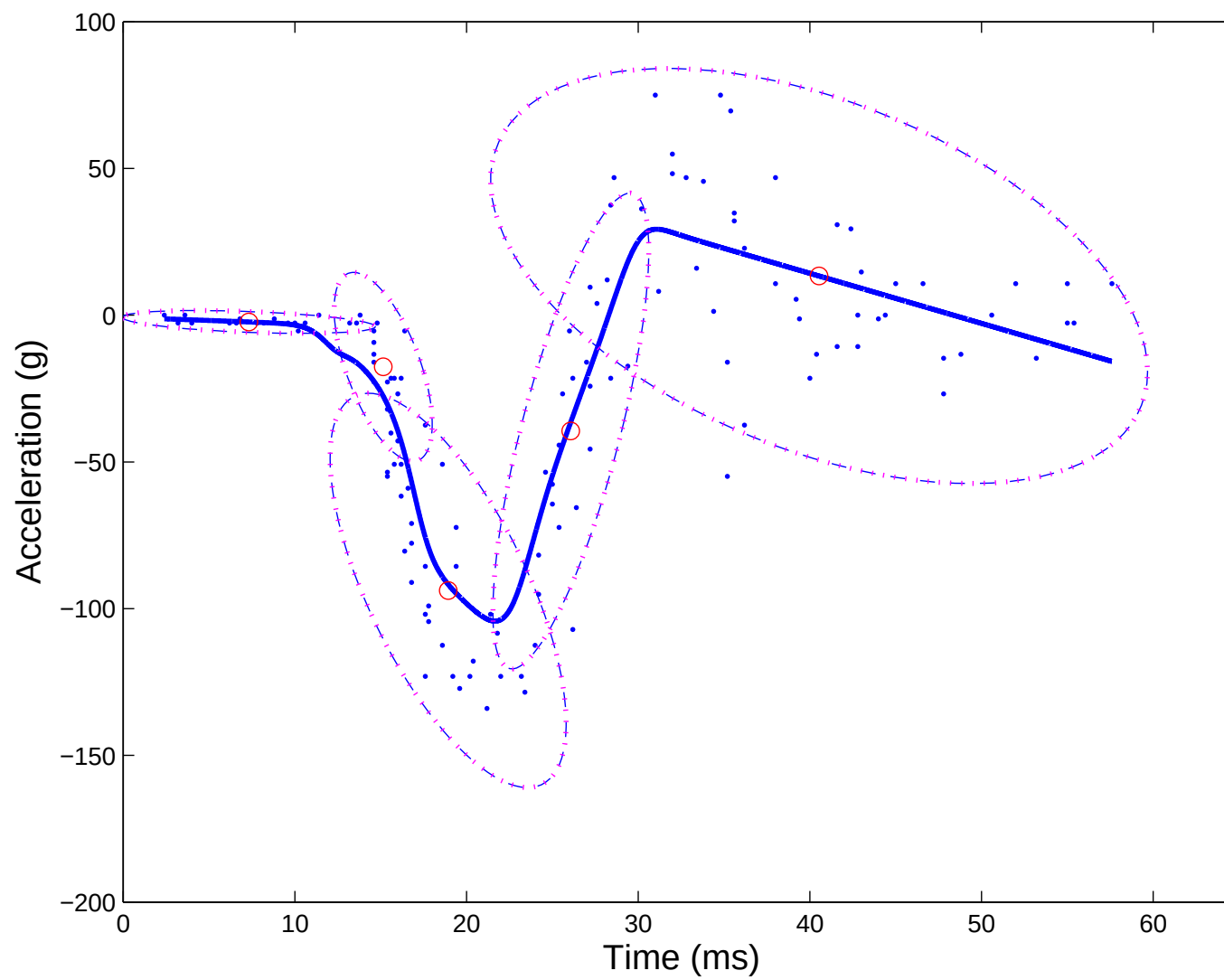


Figure 53: GMR(5).

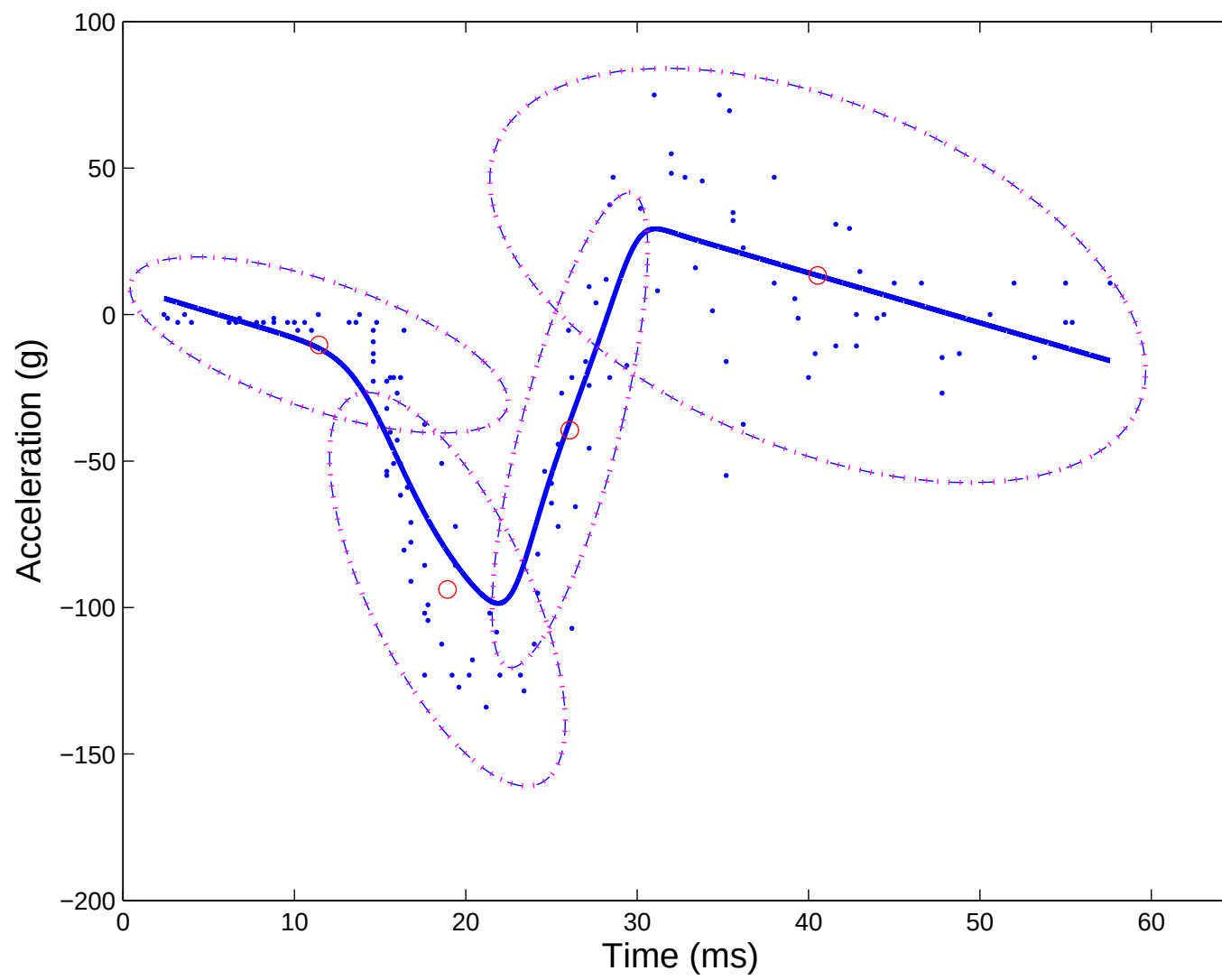


Figure 54: GMR(4).

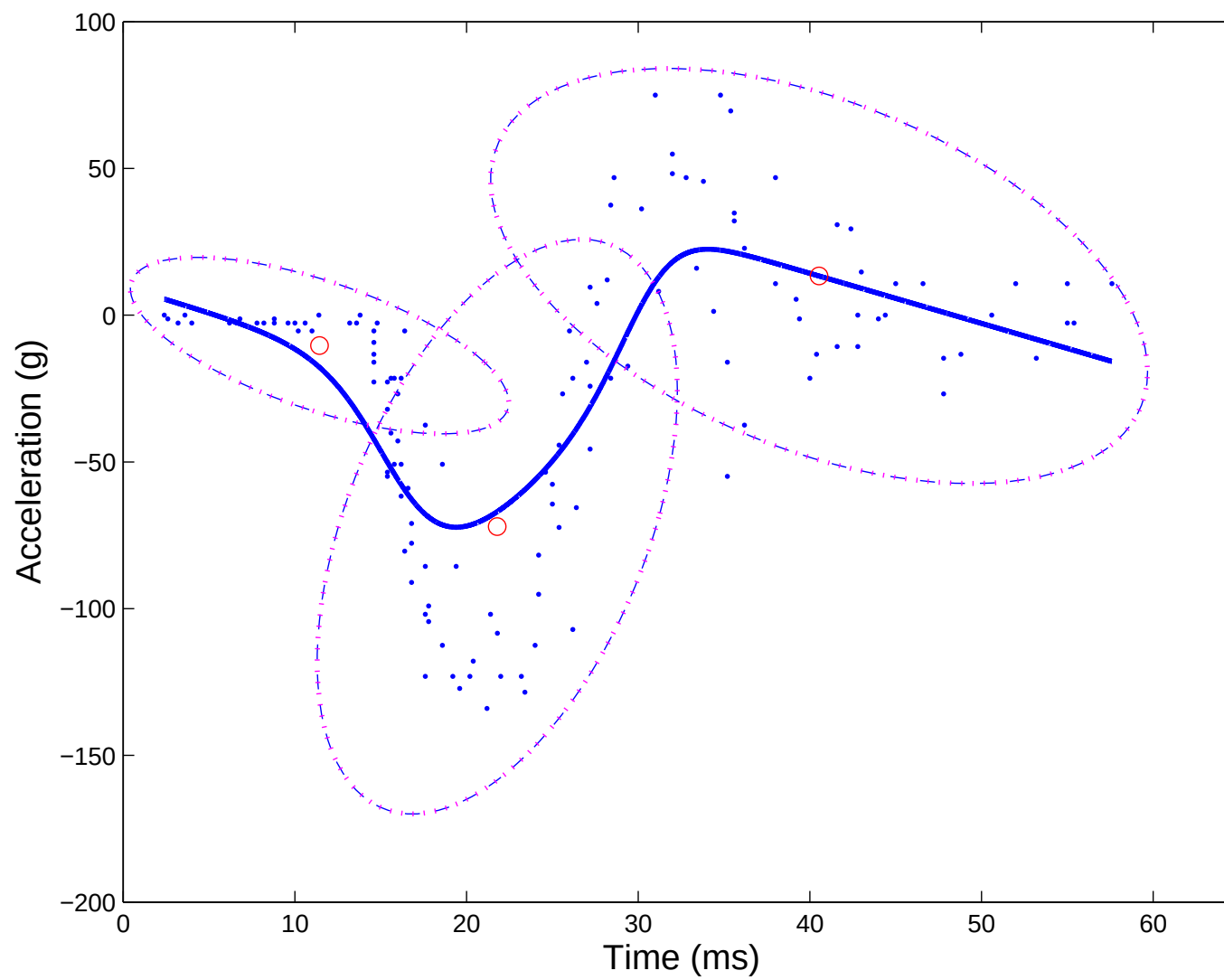


Figure 55: GMR(3).

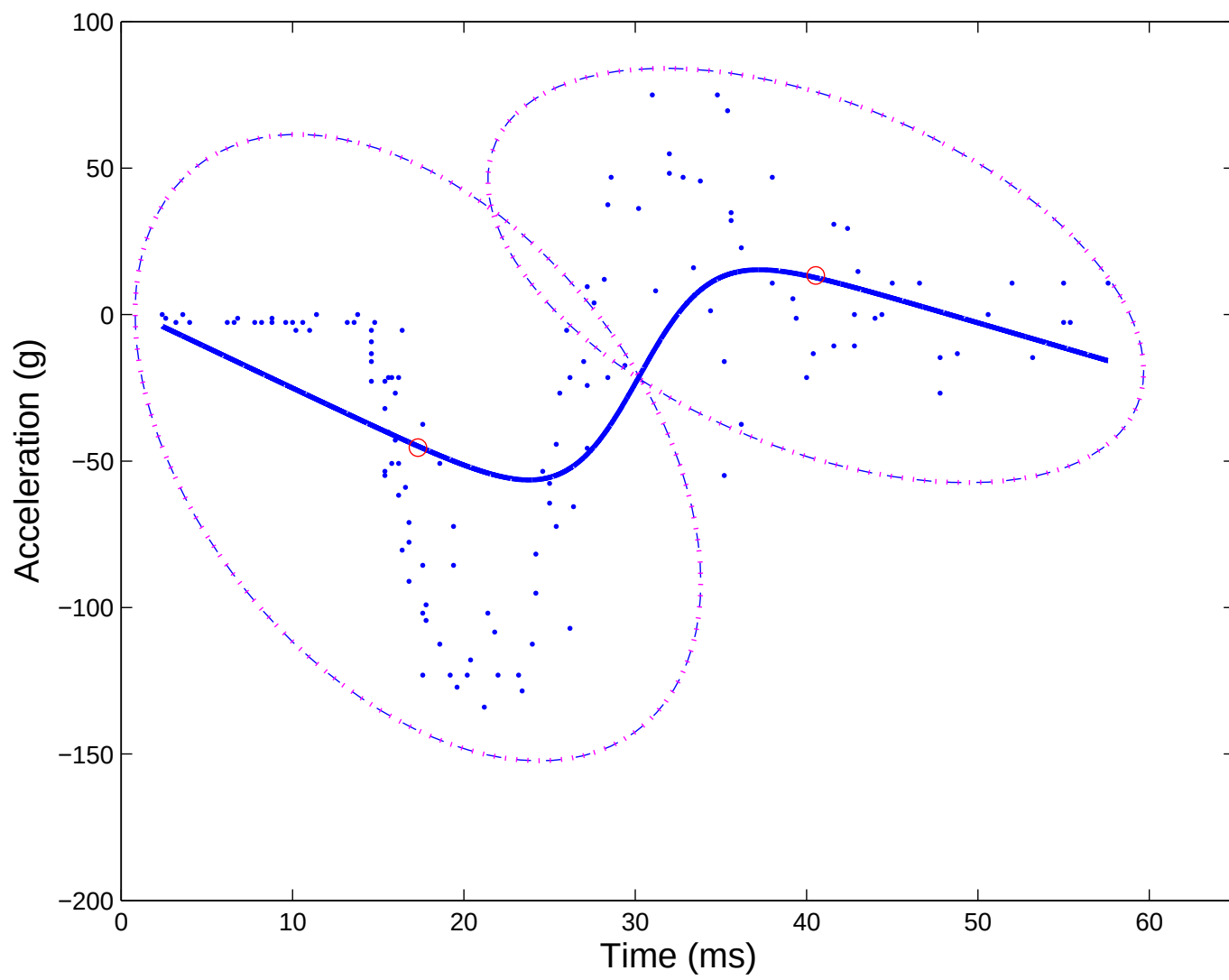


Figure 56: GMR(2).

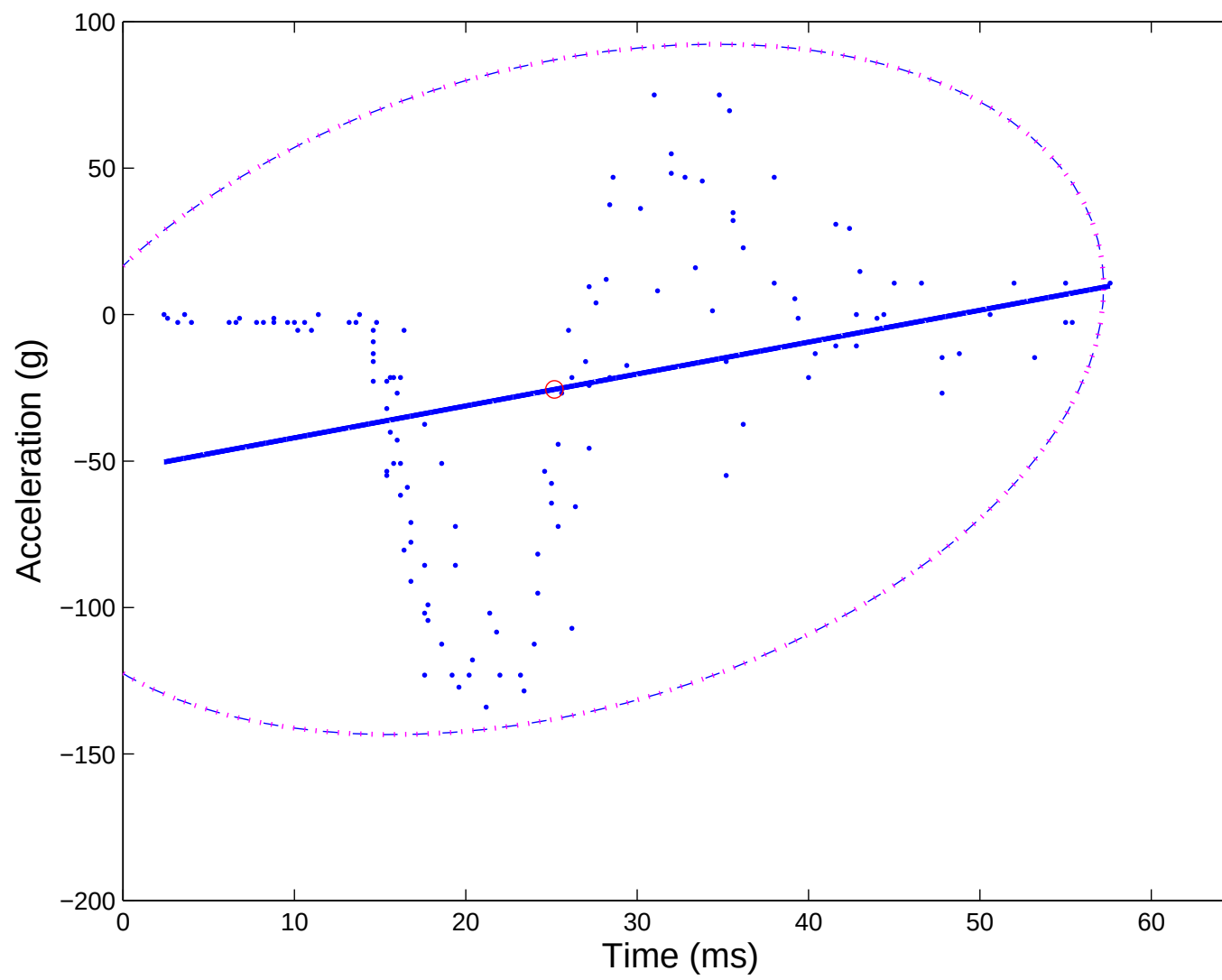


Figure 57: GMR(1).

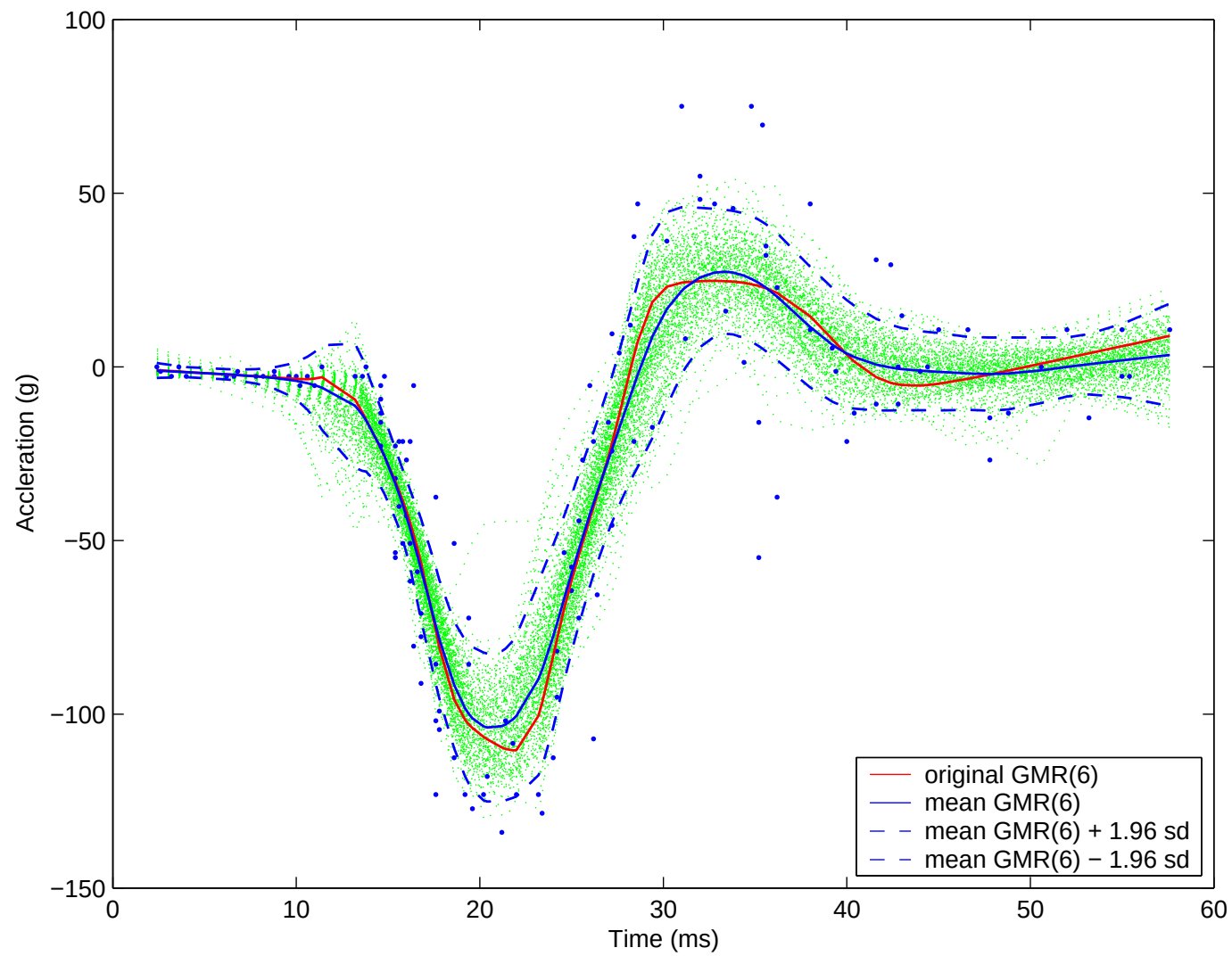


Figure 58: Bootstrap samples using GMR(6).

7 Handwritten Zip Code Data

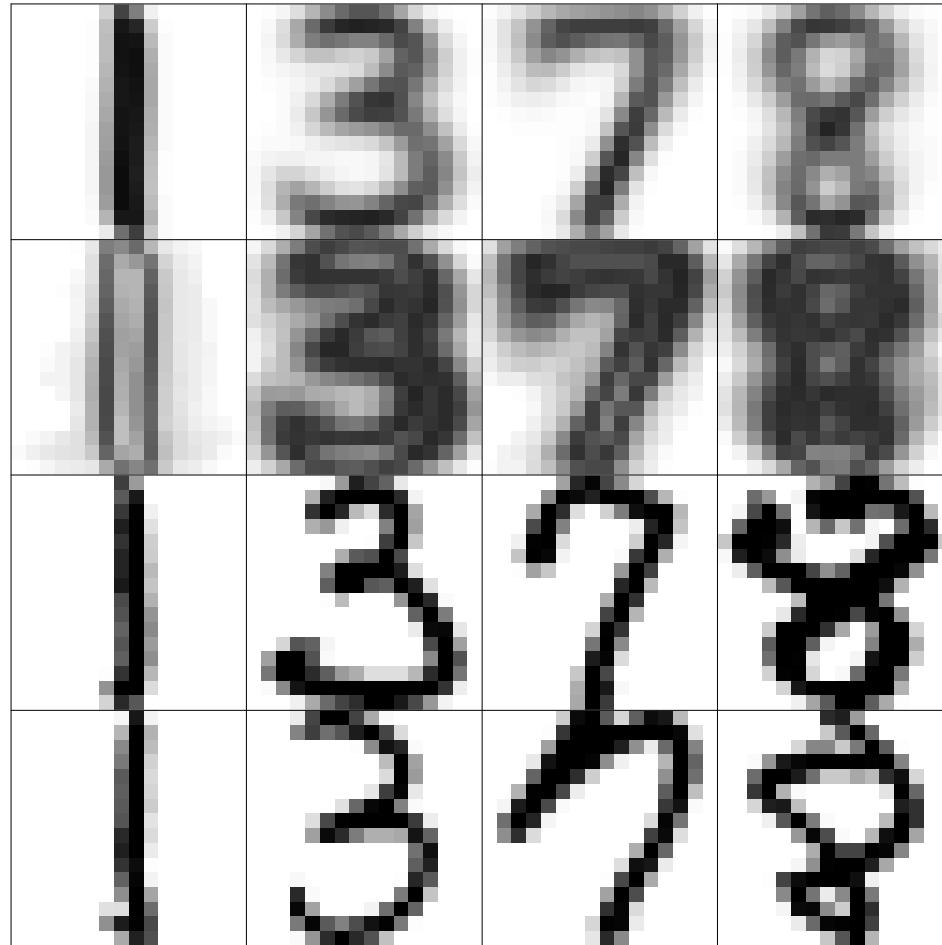


Figure 59: Mean, standard deviation, and examples of zip code digits 1, 3, 7, and 8.

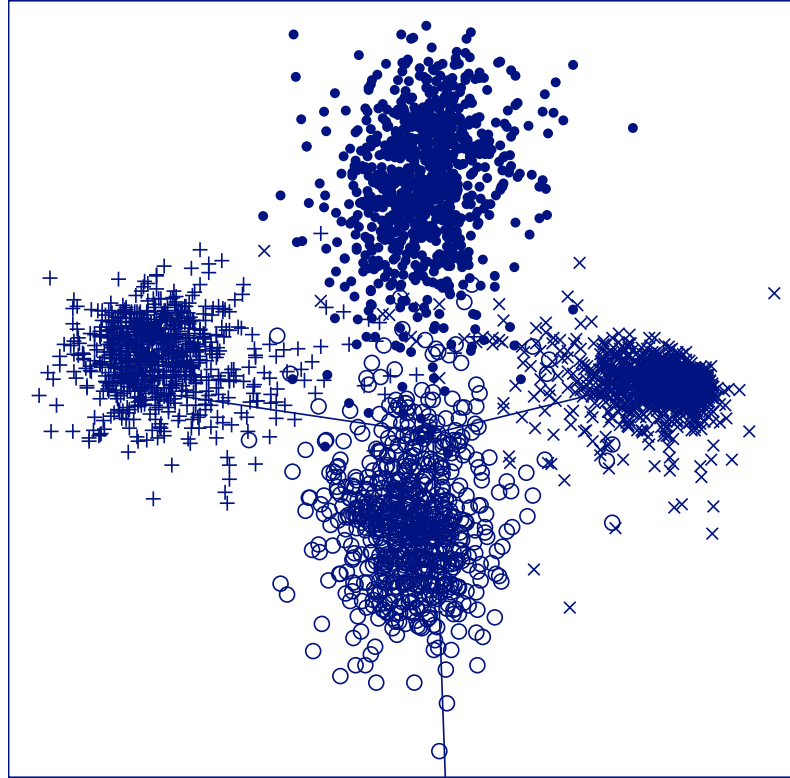


Figure 60: LDA subspace of zip code digits 1 ($\times$), 3 ($\bullet$), 7 (+), and 8 (O).

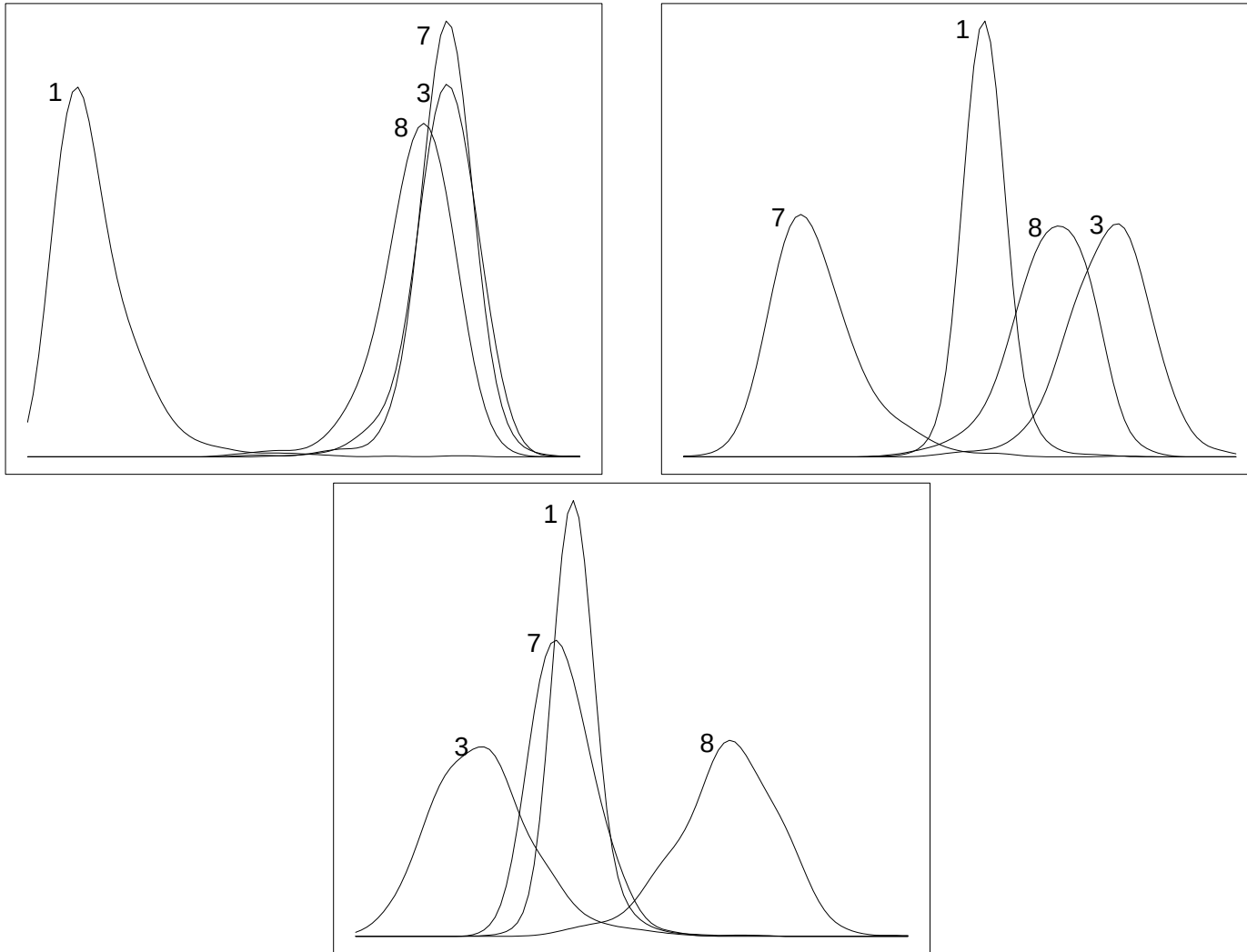


Figure 61: ASH's for each of the 4 digits for the 1st, 2nd, and 3rd LDA variable (L-R).

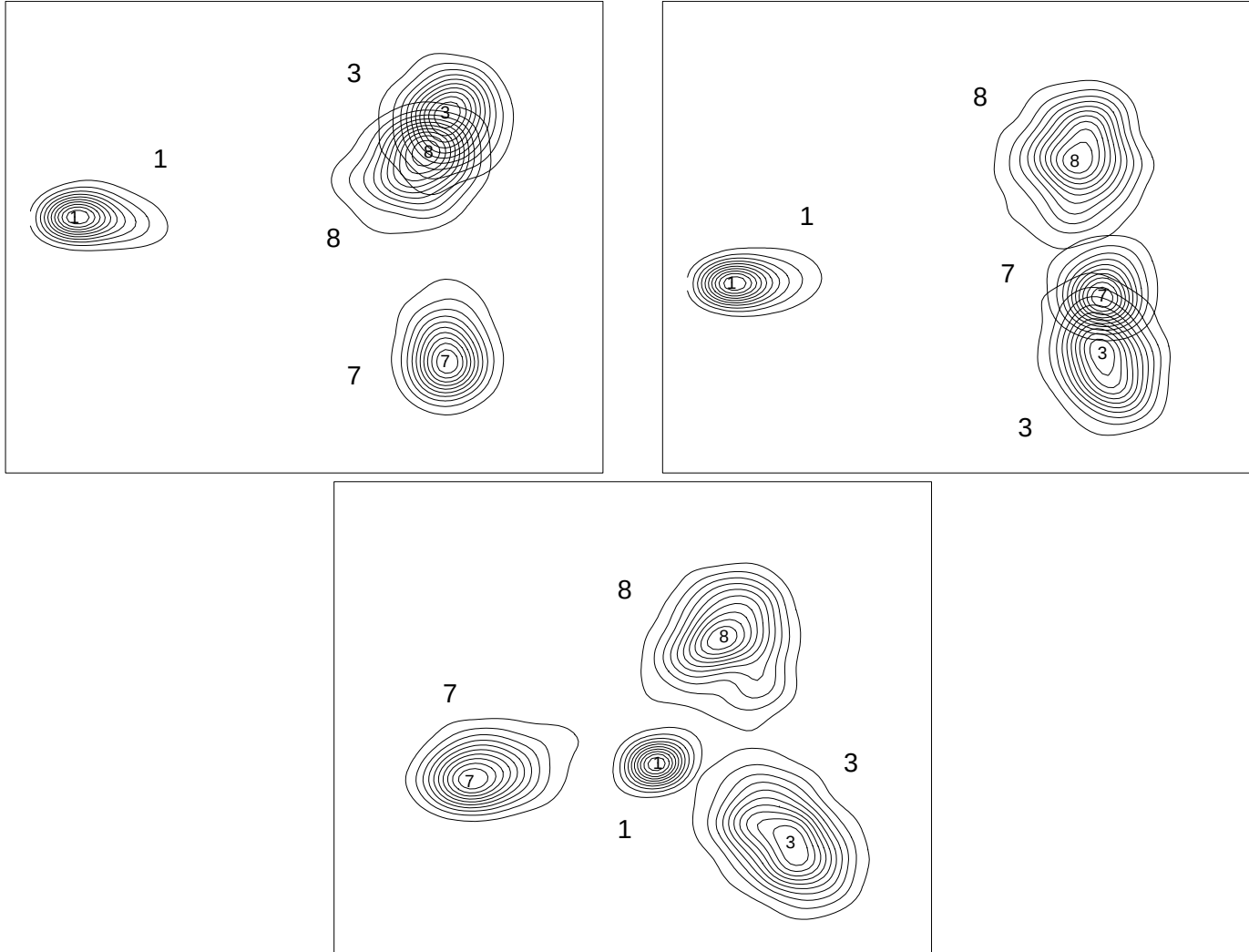


Figure 62: ASH's for each of the 4 digits for the pairs of LDA variables.

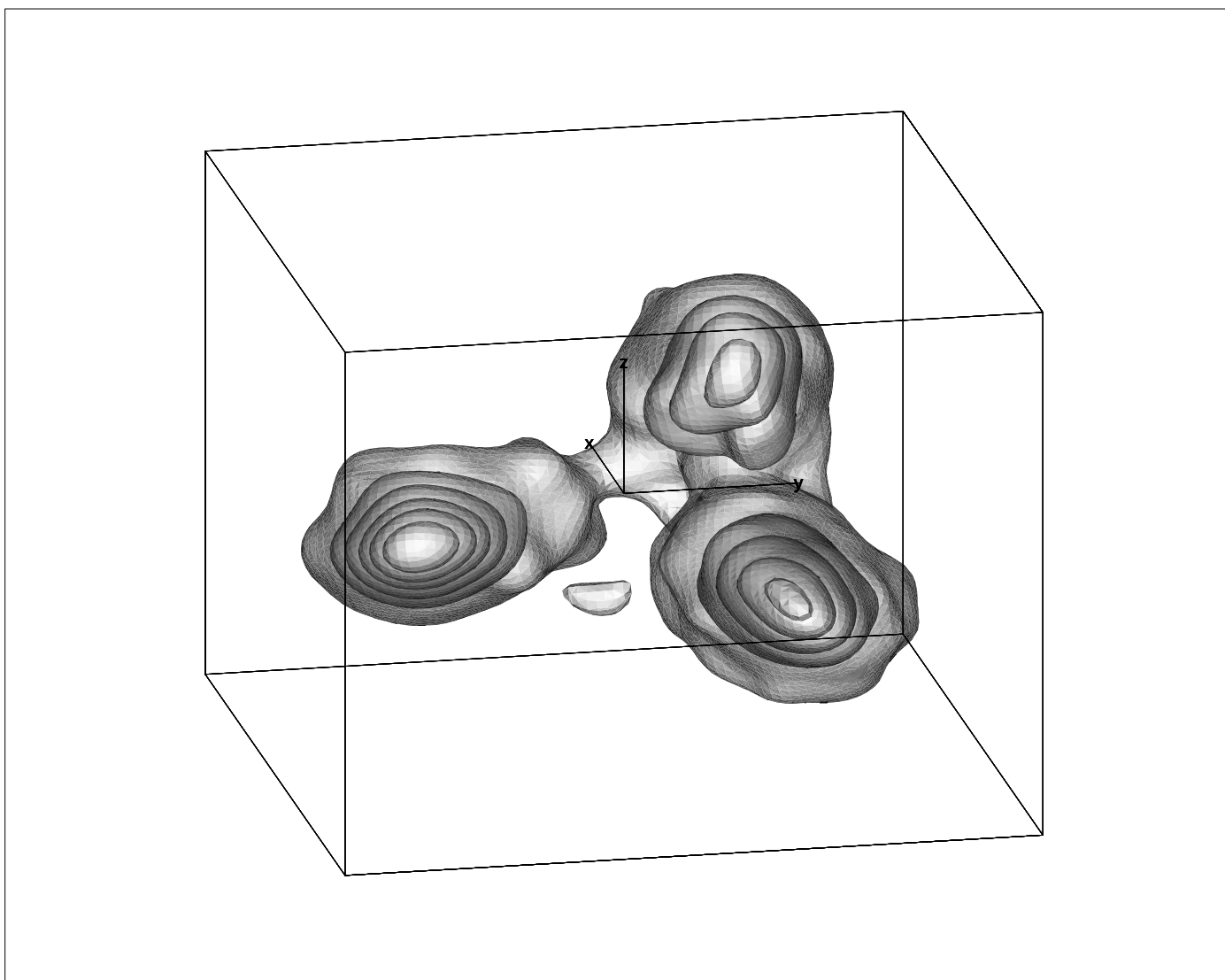


Figure 63: Trivariate ASH of 3 LDA variables for digits 3, 7, and 8. The digit 7 is in the left cluster; the digit 8 in the top cluster; and the digit 3 in the lower right cluster.

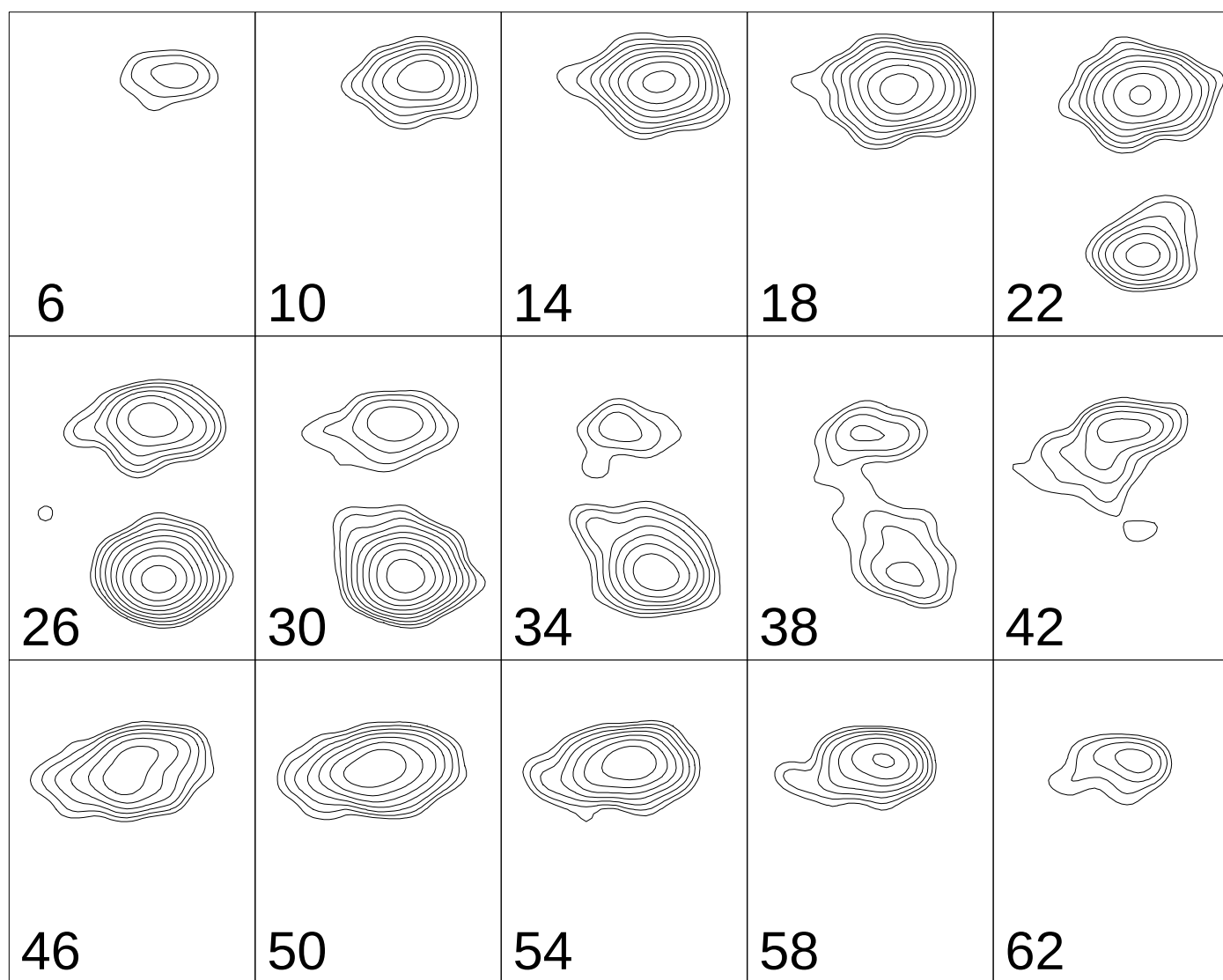


Figure 64: A sequence of slices of the three-dimensional ASH of the digits 3, 7, and 8 . The z -bin number is shown in each frame from the original 75 bins.

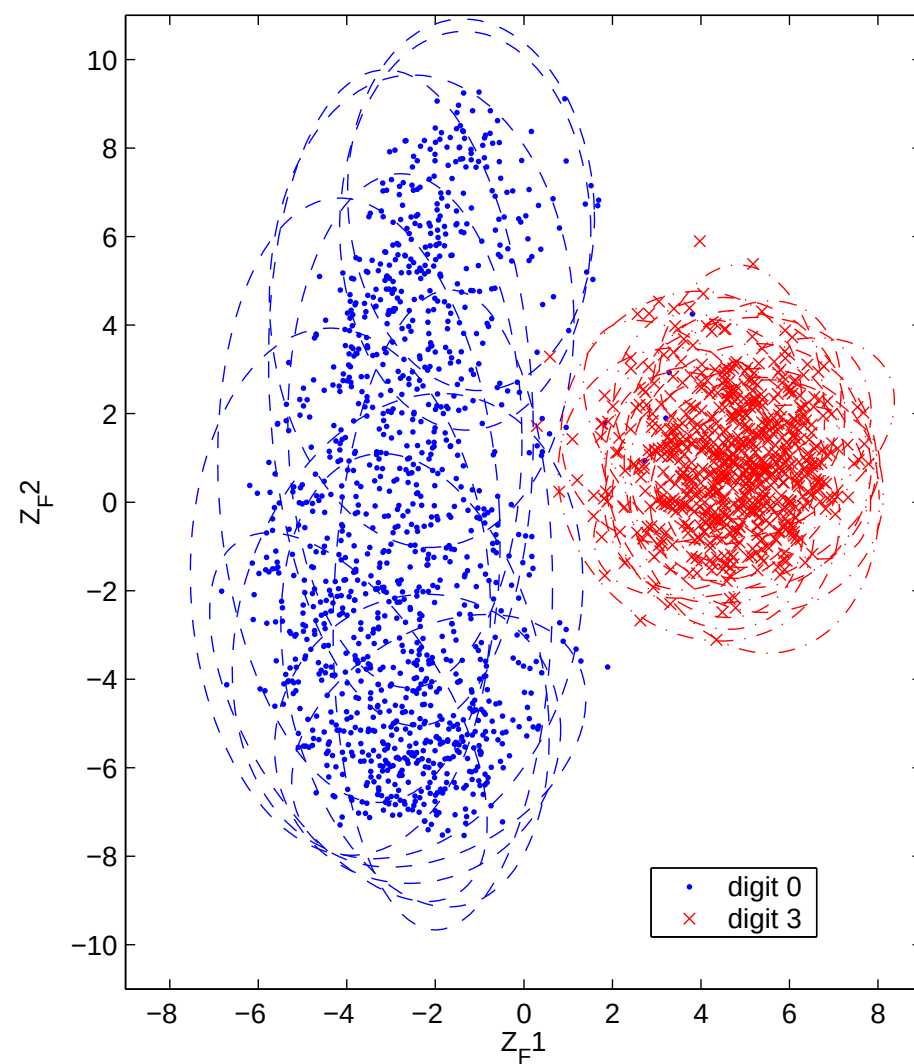


Figure 65: GMR $256 \rightarrow 10 \rightarrow 2$ dimensions. Forward best feature subspace $K=2$ (digits 0 and 3).

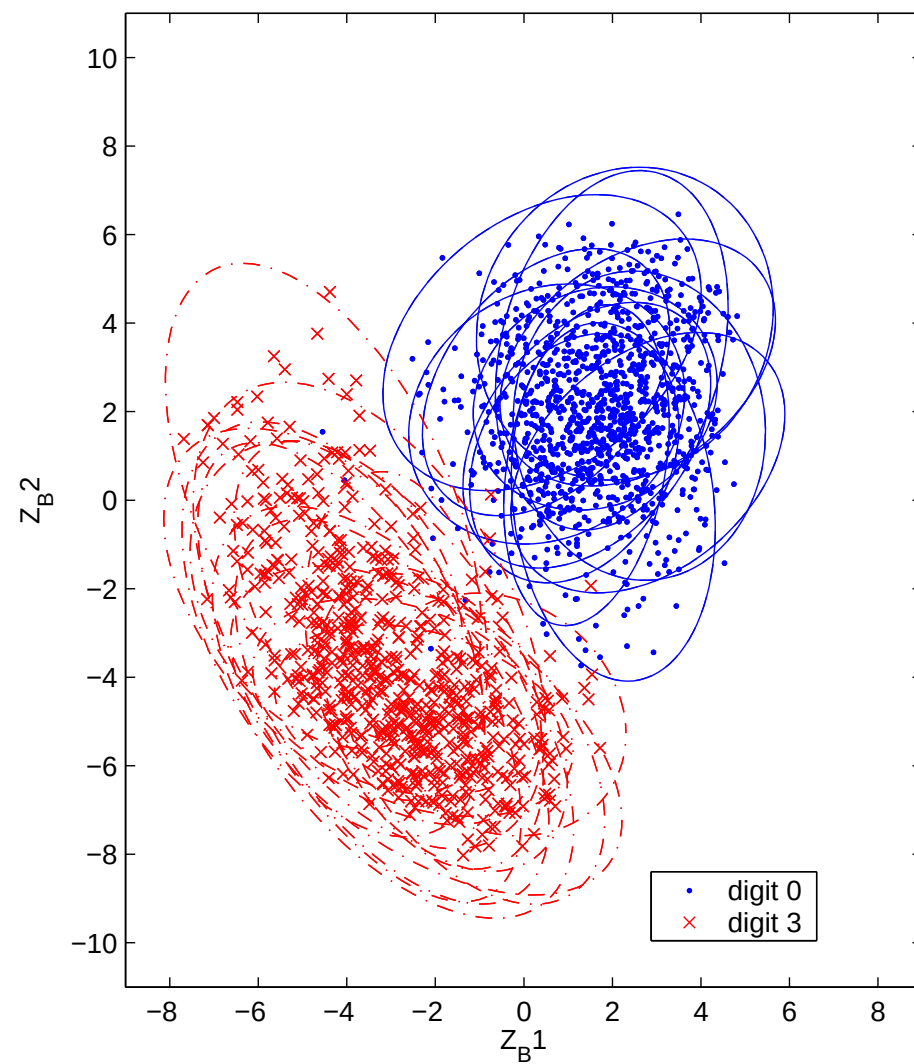


Figure 66: Backwards best feature subspace $K=2$ (digits 0 and 3).

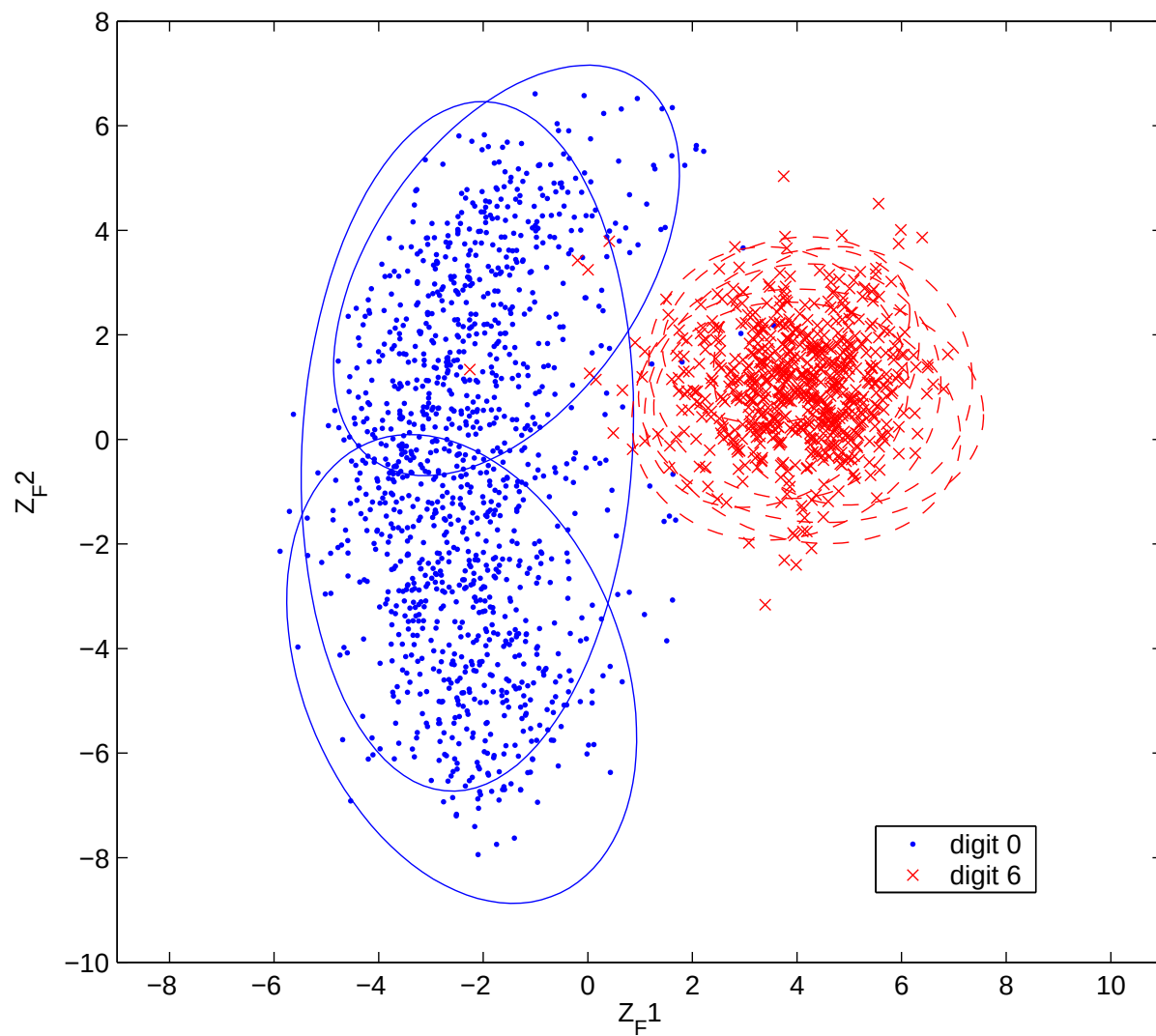


Figure 67: Forward best feature subspace $K=2$ (digits 0 and 6).

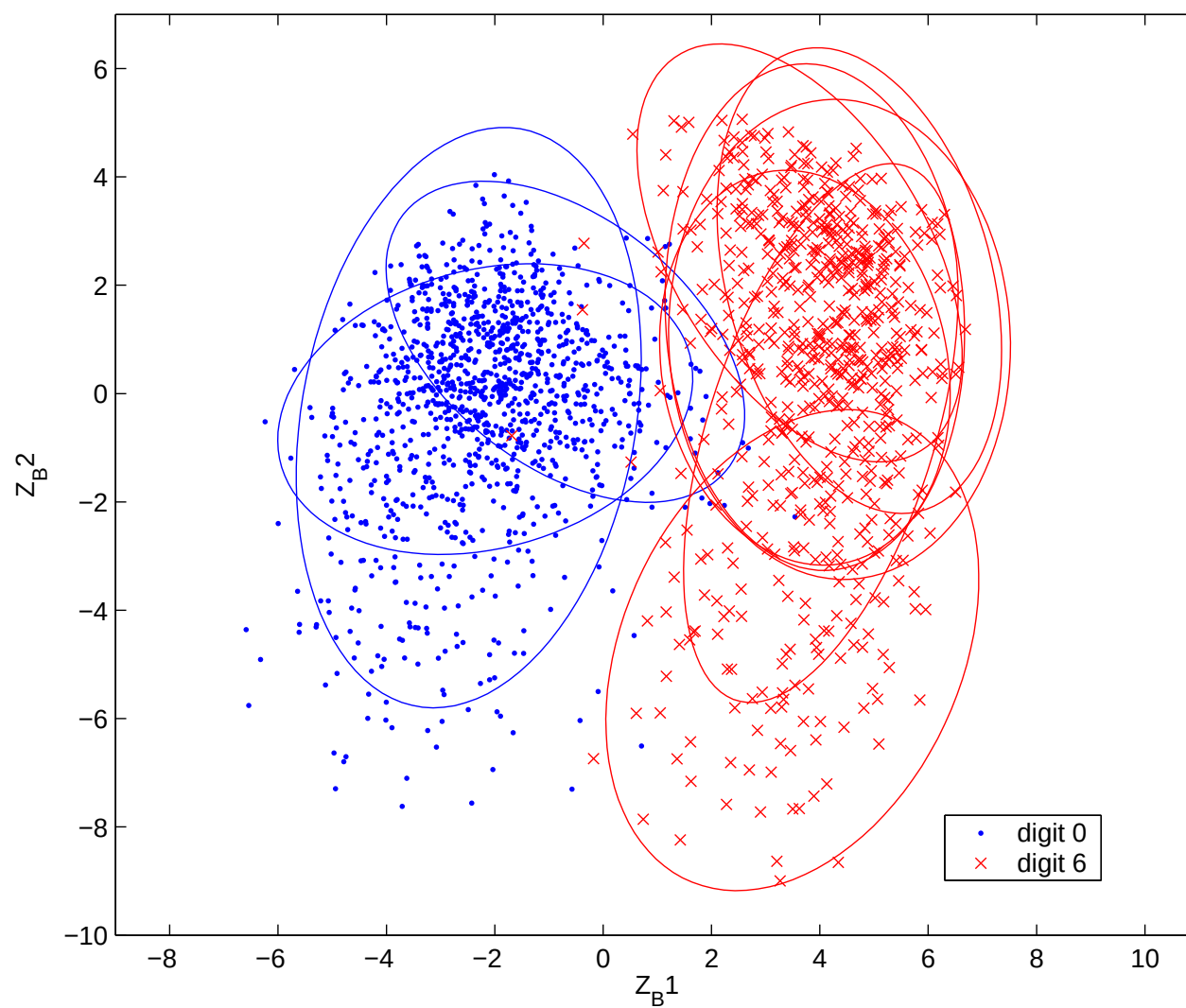


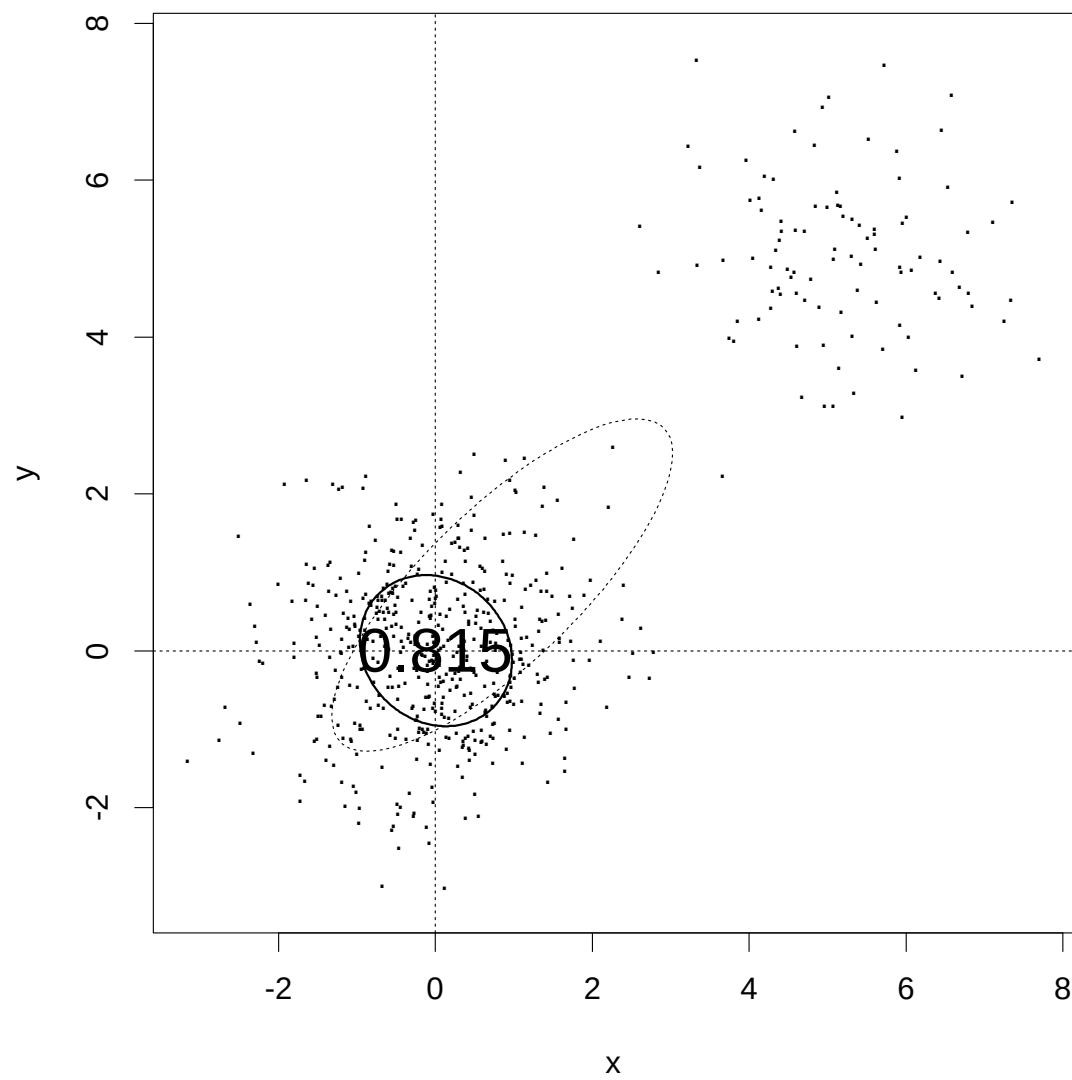
Figure 68: Backwards best feature subspace $K=2$ (digits 0 and 6).

8 Partial Mixture Density Estimation

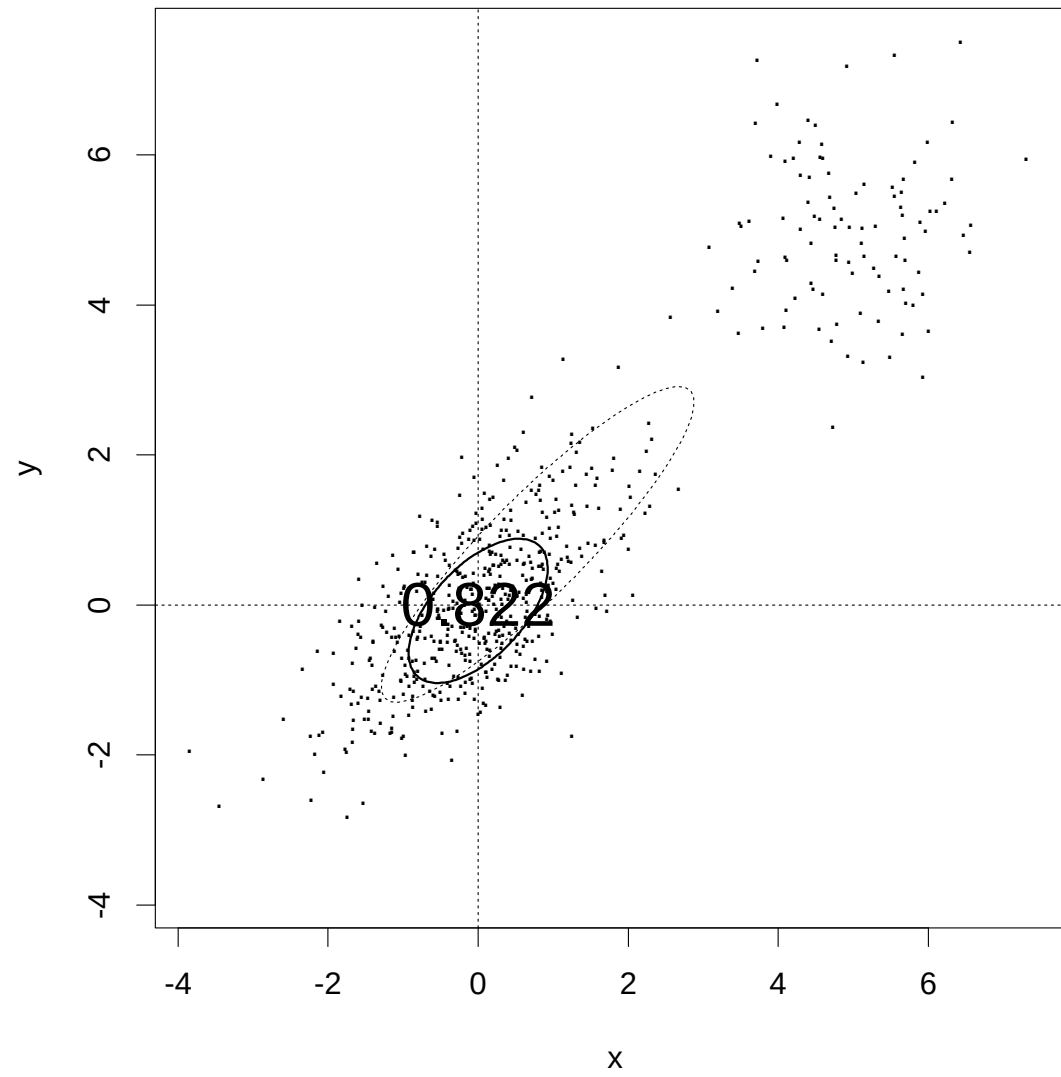
- cool hardware $\Rightarrow$ cool animations
- grand tour/grand density tour
- compare raw grand tour with invoking projection pursuit (PP)
- pure exploration vs. guided/aided exploration
- very important for pushing visualization to high-D features is to introduce models in between the data and the visualization engine
- new criteria !?
- ggobi has a selection of PP criteria, hardwired for 2-D
- limited work on PP criteria for higher dimensions
- 3-D and 4-D important since can “see” a GT in 3-D (stereo glasses) and even 4-D (array of 2-D views or stereo 3-D slices)

- 3-D density grand tour
- p -D parallel coordinates tour (no dimension reduction) — opt criteria here?
- mixture models may be viable
- PMDC fits of partial model $w N(\mu, \Sigma)$
- useful for “probing” high-D surfaces/data
- models that aid the *process* of exploration
- eg. Australian athlete data

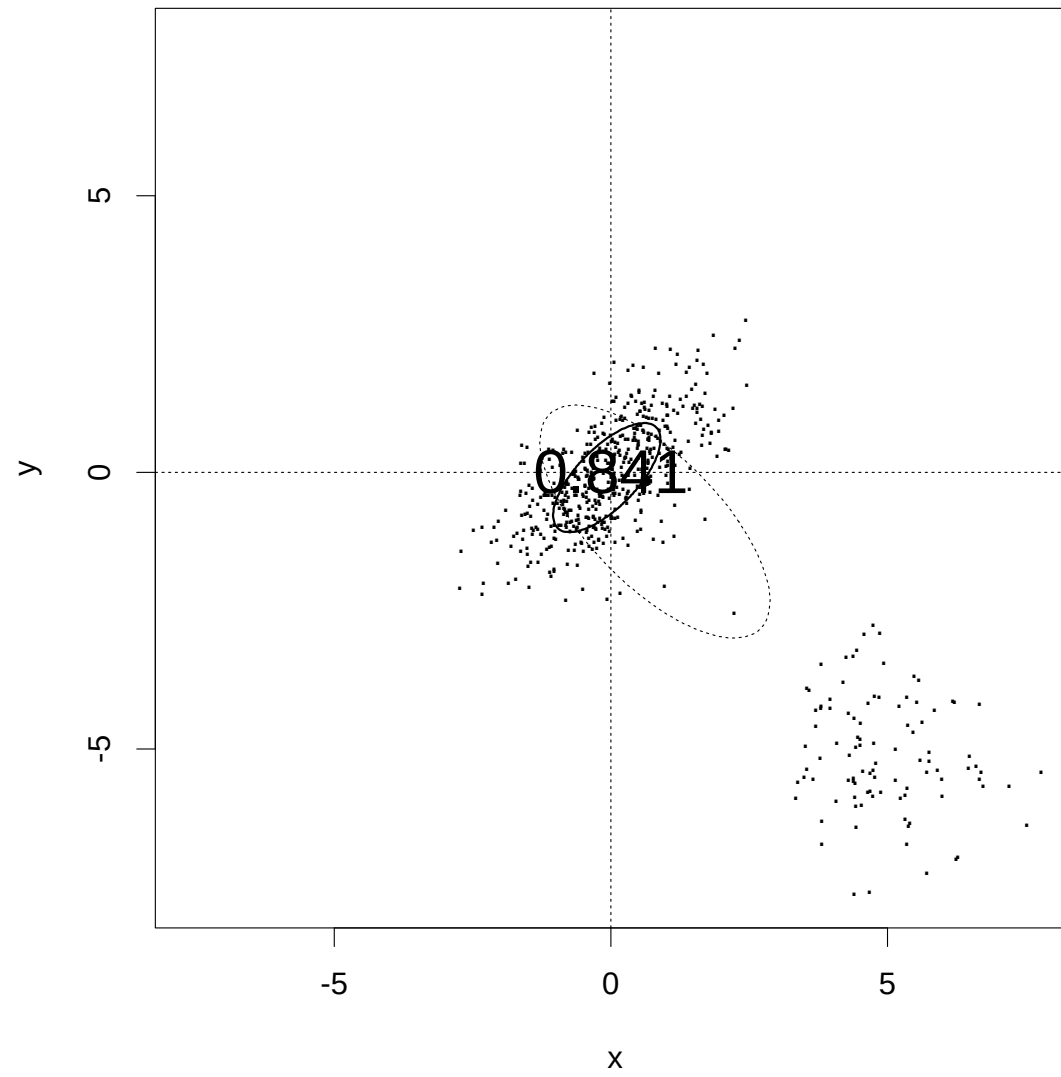
500+100 $\mu=5$



500+100 $\mu=5$ $\rho=.7$



500+100 $\mu=5$ $\rho=.7$



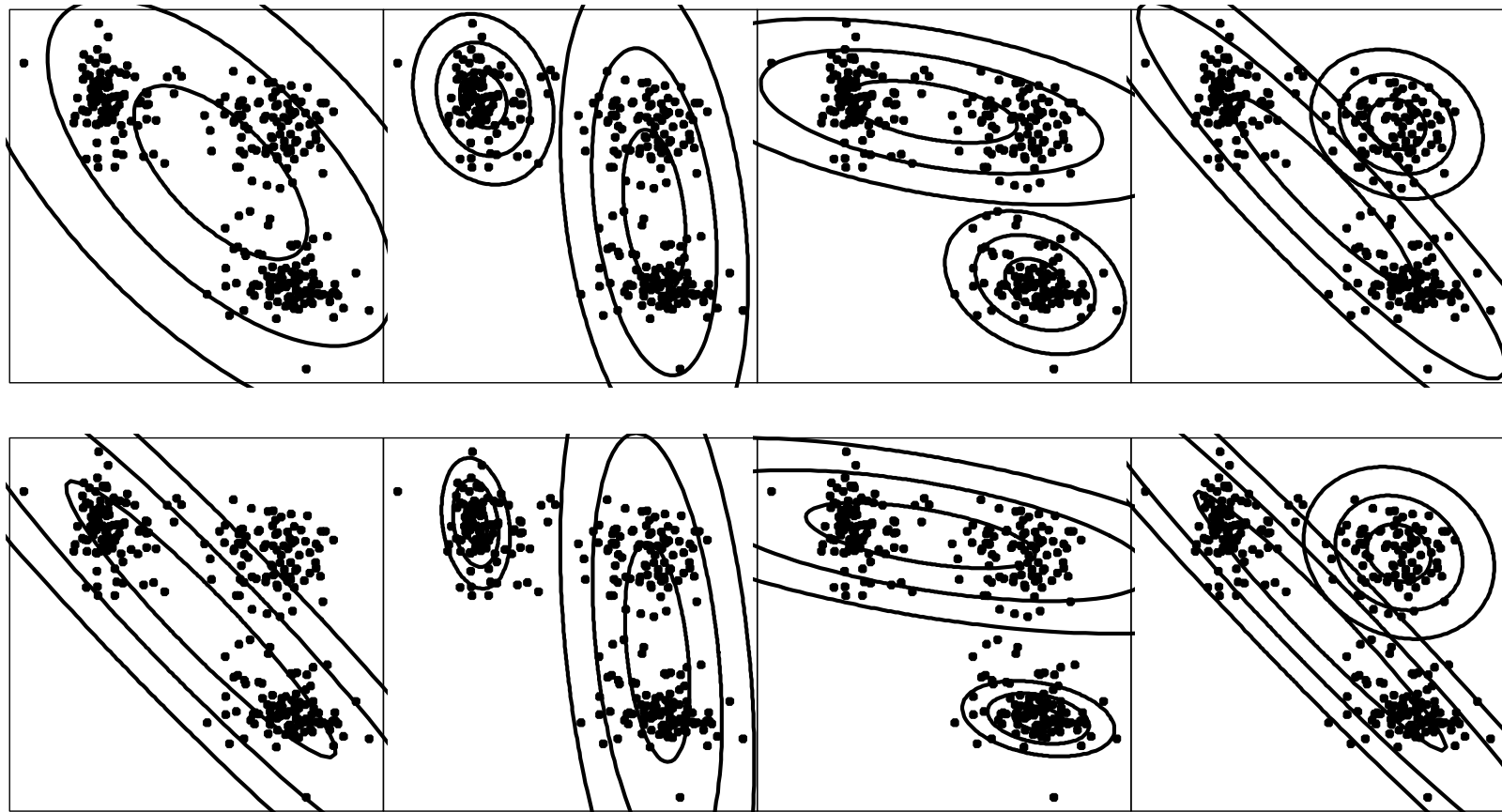


Figure 69: (top) MLE normal mixture fits to lagged Old Faithful geyser eruption data with $K = 1$ and $K = 2$. The weights in each frame from L to R are (1.0) , $(.350, .650)$, $(.645, .355)$, and $(.728, .272)$. (bottom) L2E mixture fits, with weights (1.0) , $(.258, .742)$, $(.714, .286)$, and $(.711, .289)$.

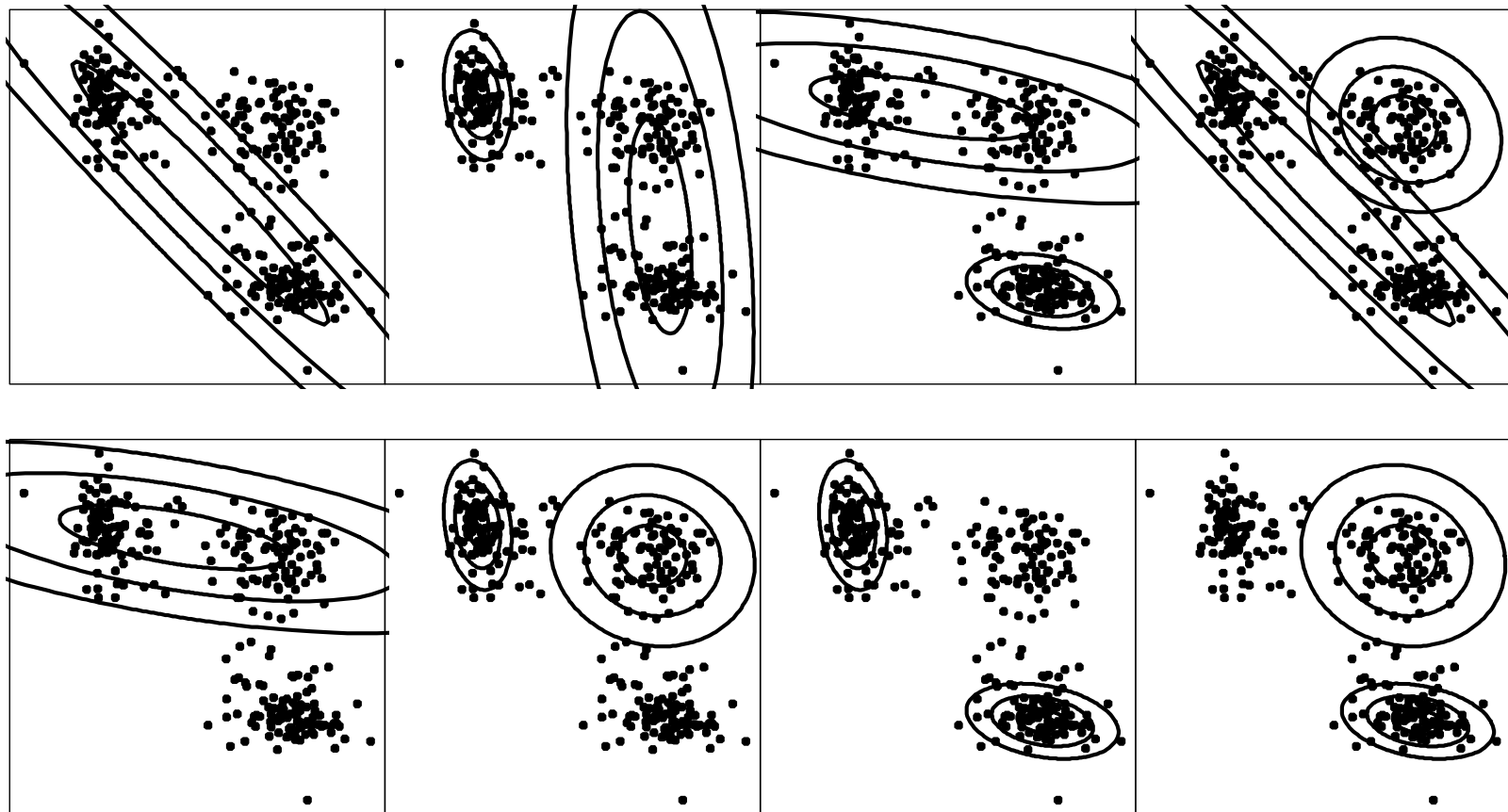


Figure 70: L2E fits without weight constraints. **(top)** The weights in each frame are (.783), (.253, .694), (.683, .283), and (.751, .297). **(bottom)** The weights in each frame are (.683), (.253, .316), (.253, .283), and (.316, .283).

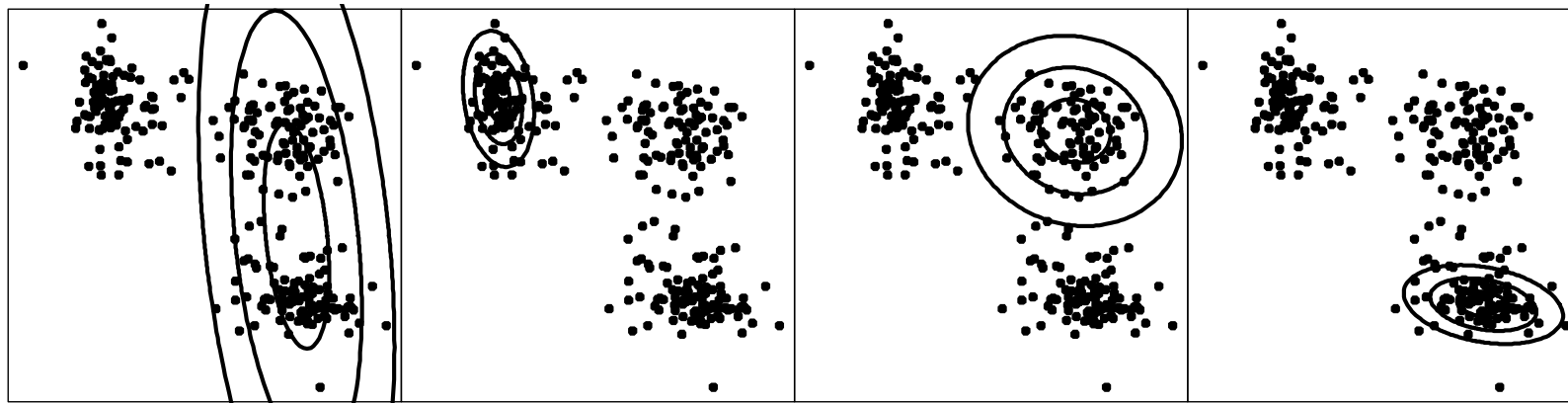


Figure 71: Four more $K = 1$ partial mixture fits to the geyser data. The weights in each frame are (.694), (.253), (.316), and (.283).

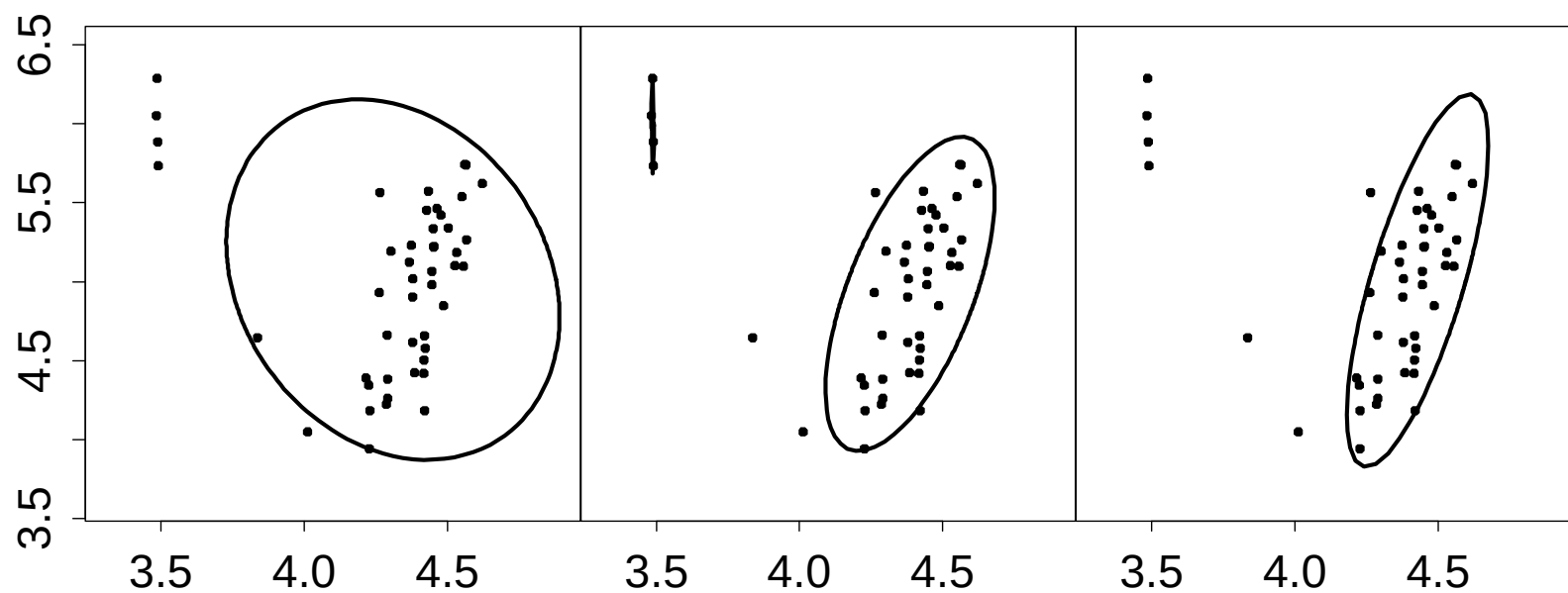


Figure 72: Two- σ contours of MLE ($K = 1$), MLE mixture ($K = 2$), and partial L2E mixture ($K = 1$) fits to the blurred star data.

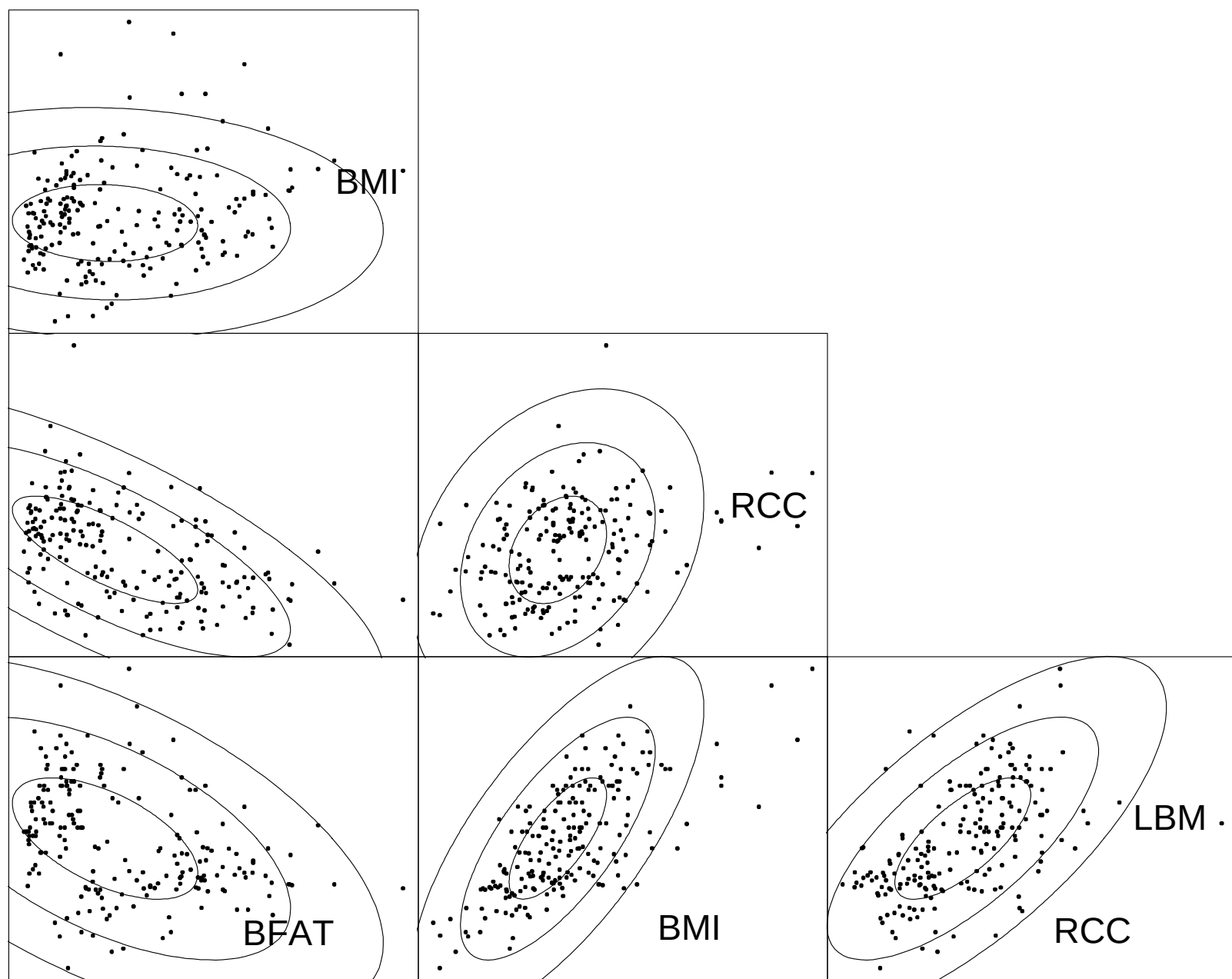


Figure 73: Ellipses representing the $(1,2,3)$ - σ contours of a L2E partial mixture estimate of the Australian athlete data.

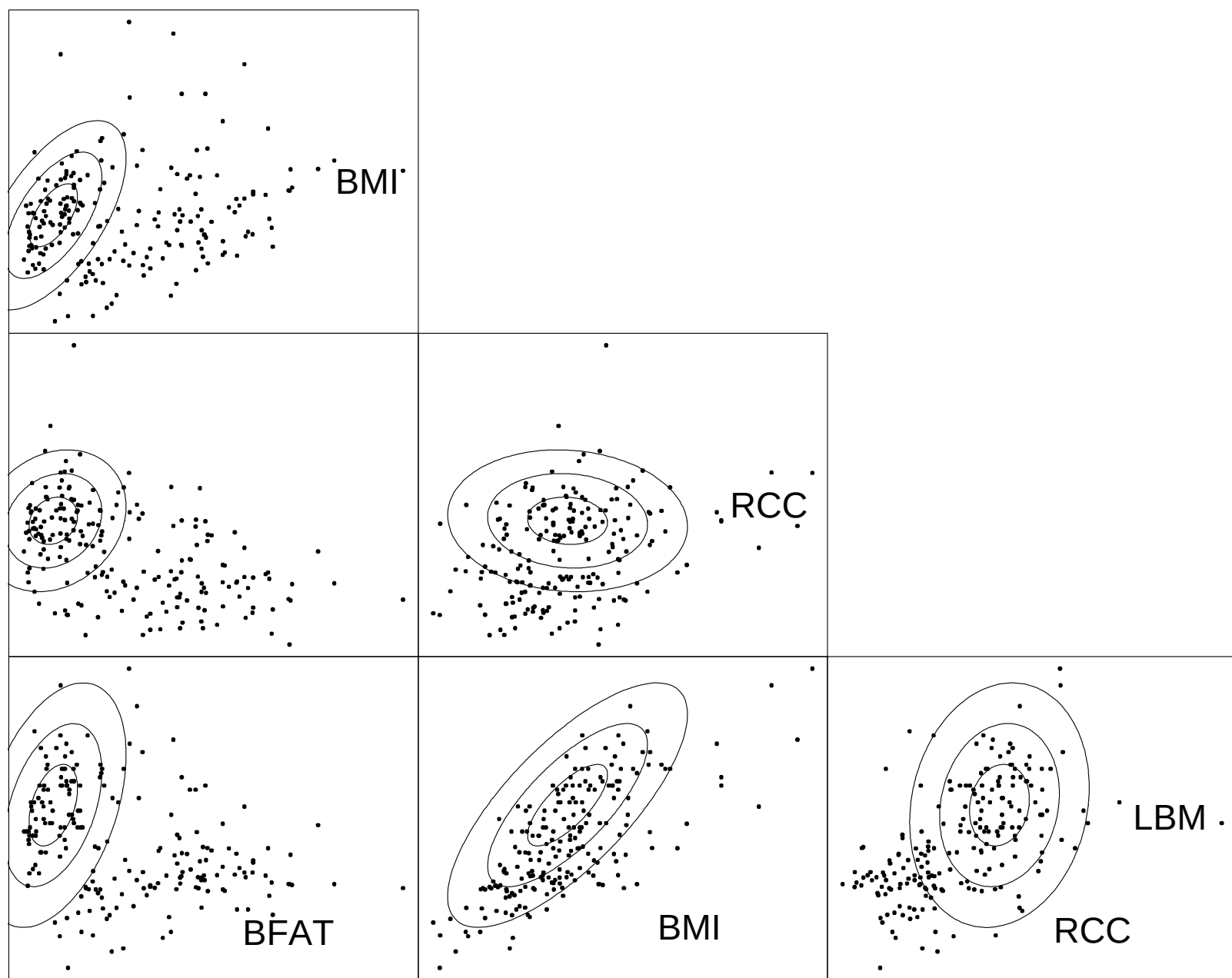


Figure 74: Ellipses representing the (1,2,3)- σ contours of a second L2E partial mixture estimate of the Australian athlete data; $\hat{w} = .43$.

9 Skewers: Principal Components for Unlabeled Mixtures

- complements
- principal components (same as the SVD-method) — important techniques
- Dan Sorensen's algorithms for finding K largest singular values and singular vectors
- useful for massive datasets (collection of images in $\Re^{10,000,000}$)
- BUT — what if data are actually from a mixture of normals
- Def: skewer = 1st principal component re-located to pass through data cloud
- $\bar{x} + \alpha v_1$ if one component
- can we define an algorithm that is attracted to a skewer if $K > 1$?

- partial L2E method works, if look at distribution of distance to a proposed skewer

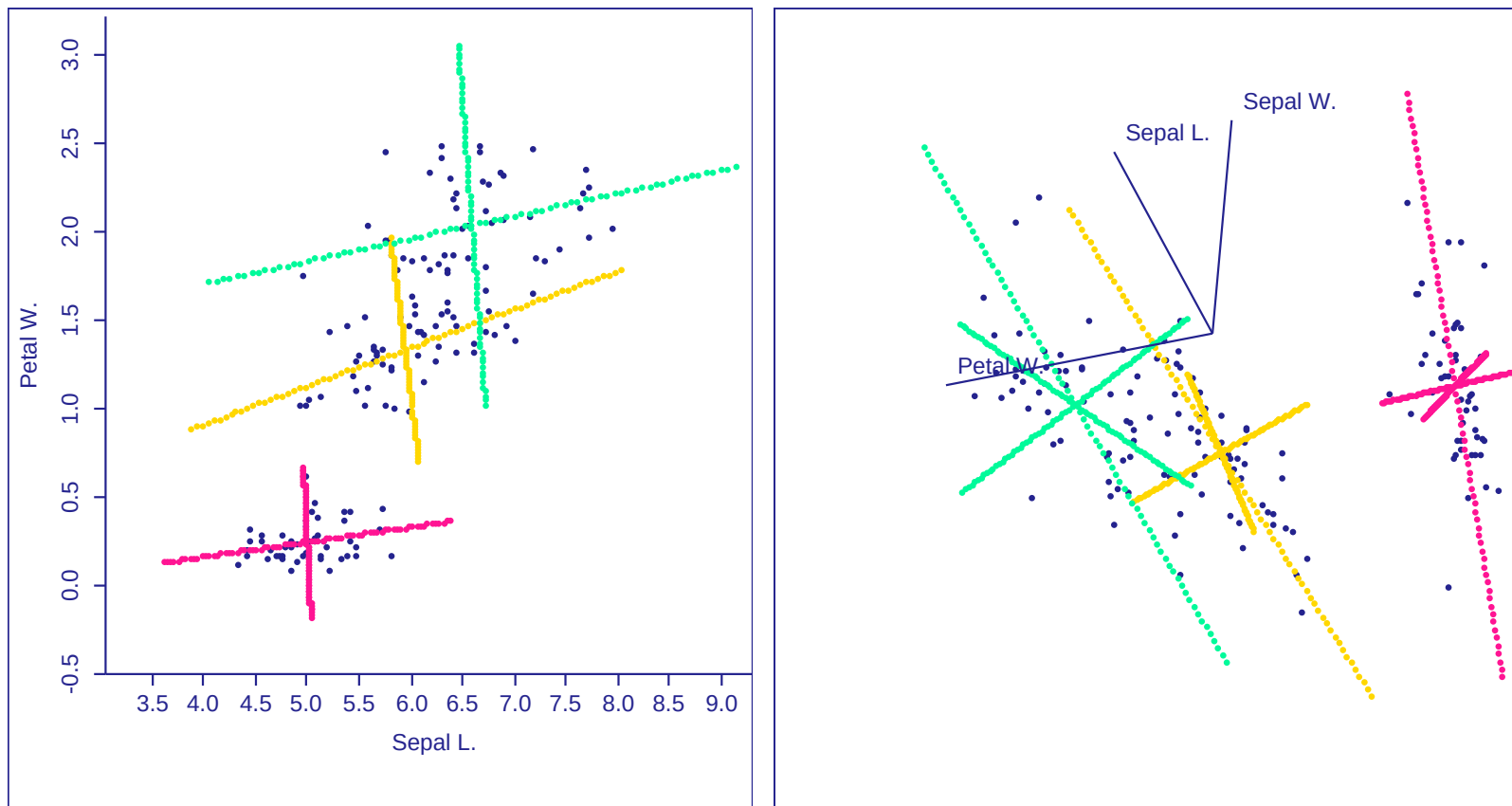


Figure 75: Eigenvectors for each of the three Iris species in $\mathbb{R}^2$ (Sepal Length and Petal Width) and $\mathbb{R}^3$ (Sepal Width added).

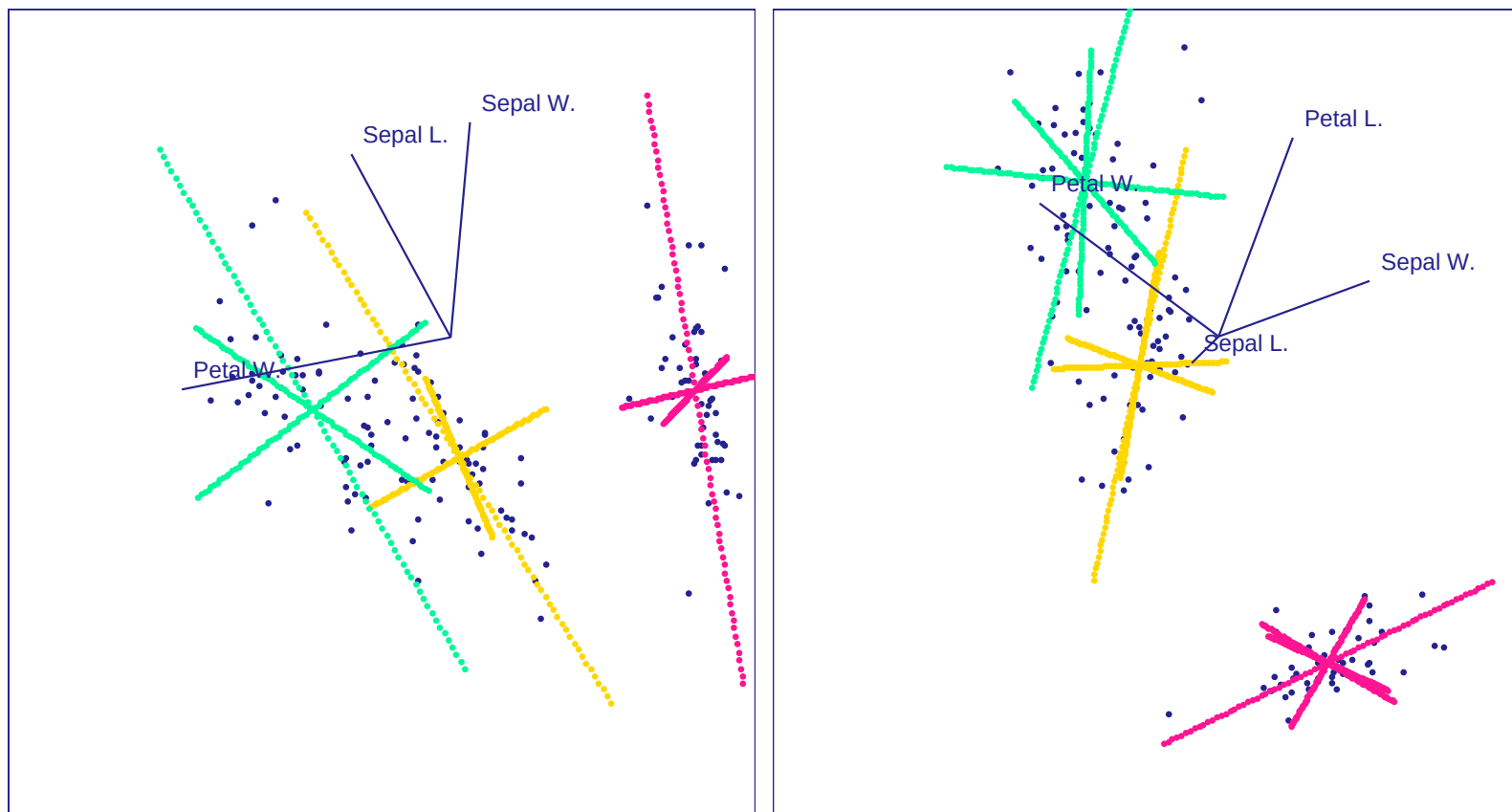


Figure 76: Eigenvectors for each of the three Iris species in $\mathbb{R}^3$ and $\mathbb{R}^4$ (grand tour view).

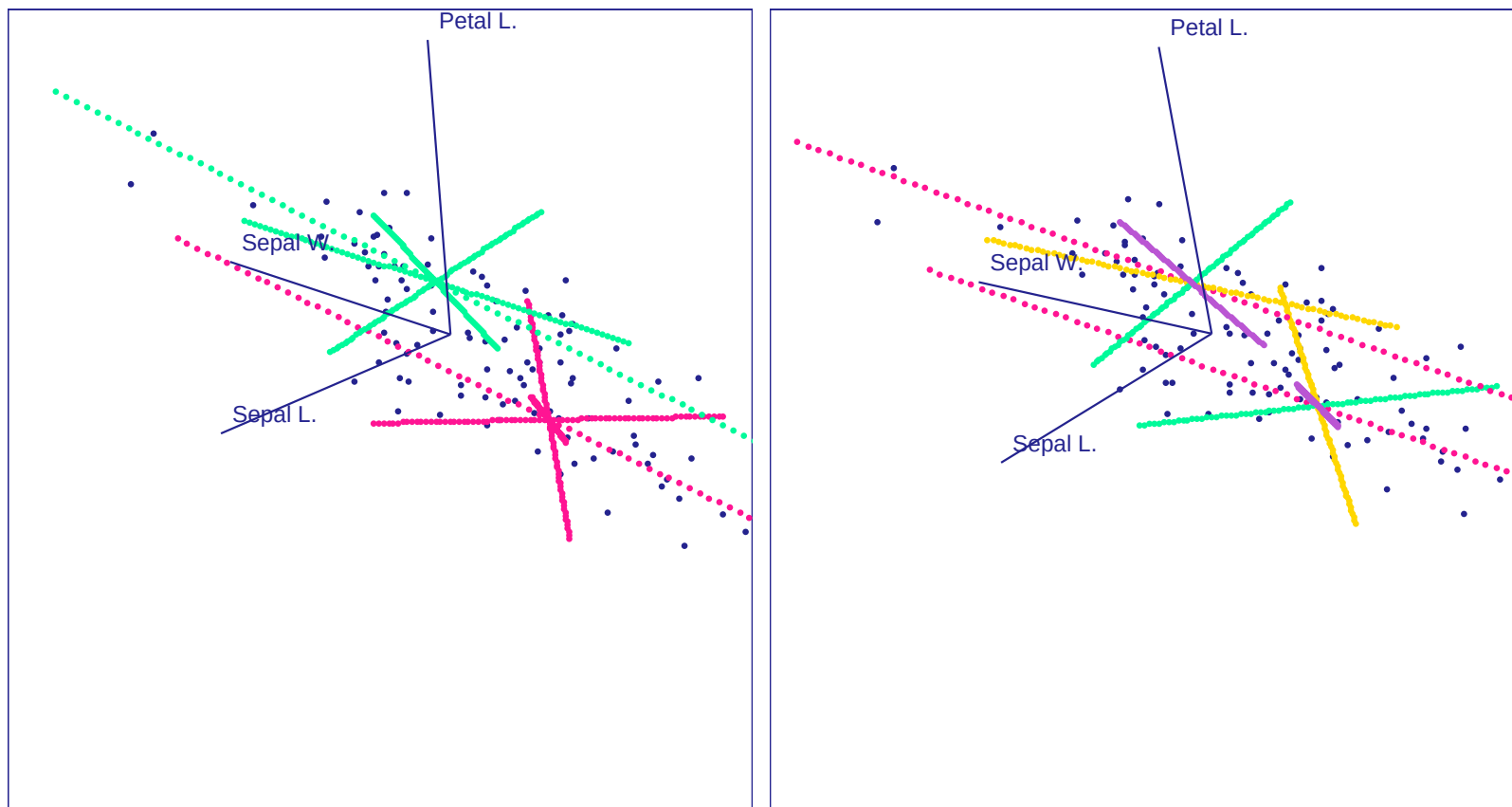


Figure 77: Eigenvectors in $\mathbb{R}^3$ of *Versicolor* and *Virginica* Iris species: colored by species (left frame); colored by size of eigenvalue (right frame)

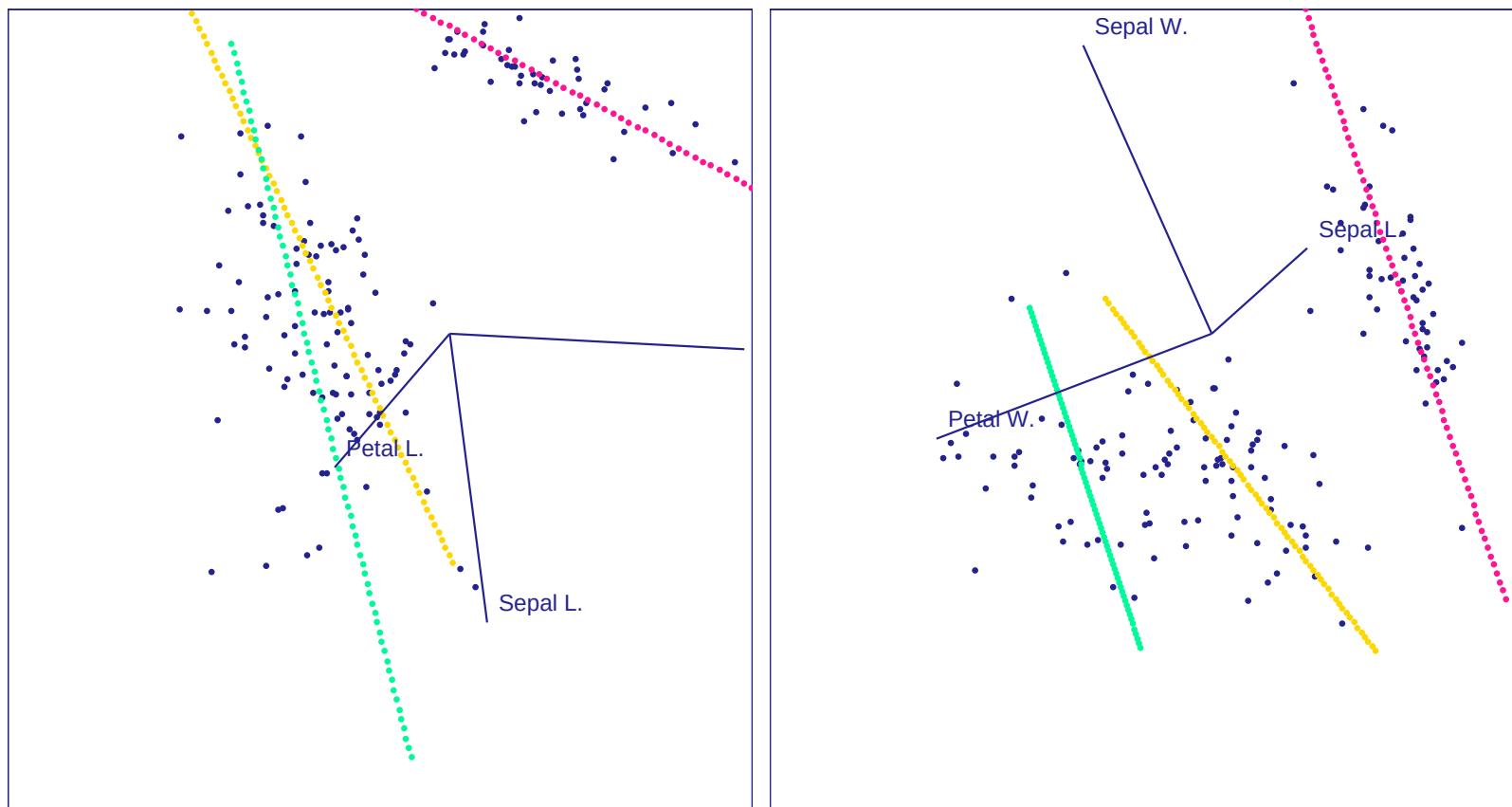


Figure 78: “4-D Skewers” of 3 Iris species in $\mathbb{R}^3$ (left frame: variables-123; right frame: variables-124).

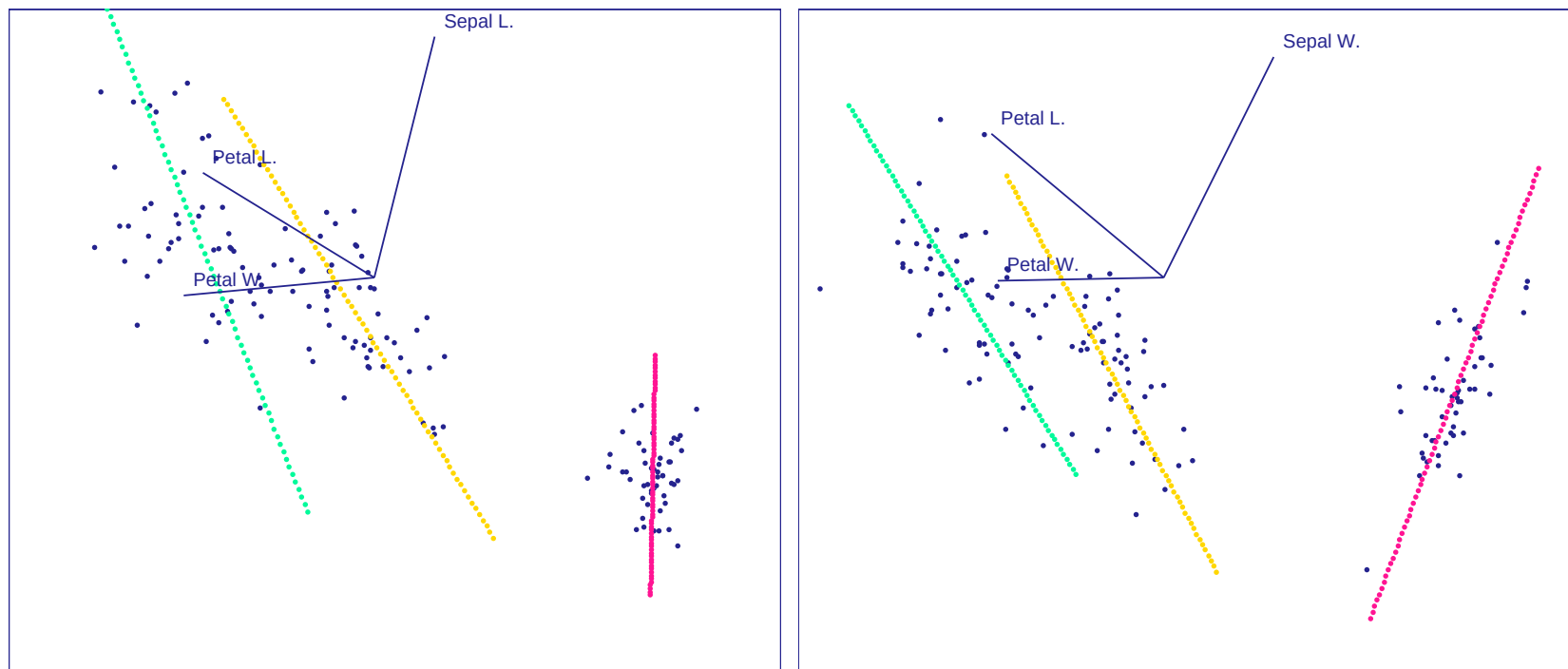


Figure 79: “4-D Skewers” of 3 Iris species in $\mathbb{R}^3$ (left frame: variables-134; right frame: variables-234).

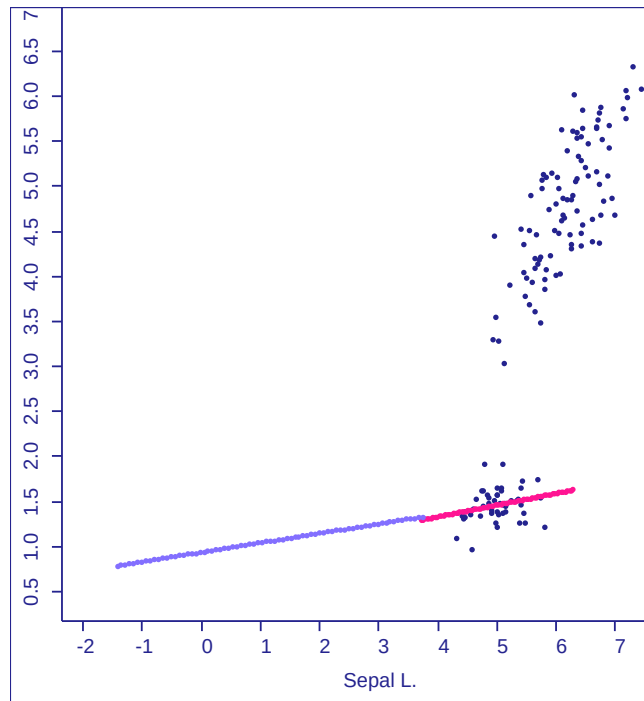


Figure 80: 2-D Skewer of Iris Data

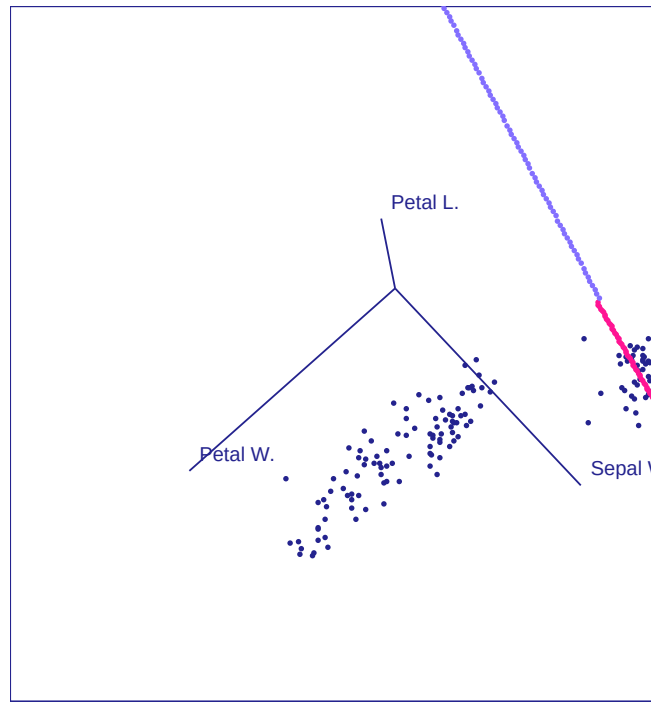


Figure 81: 3-D Skewer of Iris Data

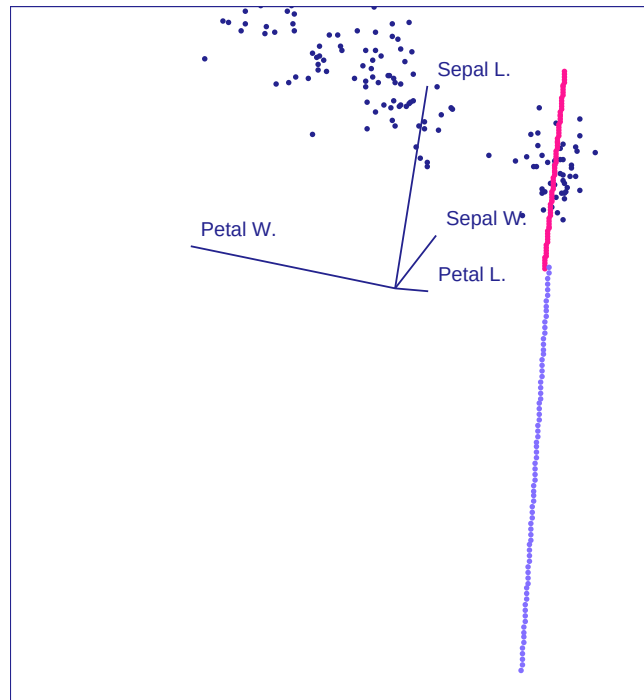


Figure 82: 4-D Skewer of Iris Data

10 Conclusions

- visualizing uncertainty ideas
- for complex visualization, may be too much to try to add uncertainty cues
- need a “grand tour” of the “confidence region”, using the same view/visualization of the data/model
- continuity helps learn areas where the model is more/less stable
- where structure is *not* preserved, etc
- a “PP” might search for the most extreme curves in the confidence set
- note: hurricane confidence intervals seemed much too narrow this past summer over 24 hours out, and *discontinuous*
- (ensemble of deterministic models vs. stochastic models?)

- with data mining (complete data), density useful for seeing relative frequency in universe
- conveying uncertainties when not a random sample still useful exercise; no real sense of “size” of confidence region; nor properly centering (biased); nor sure if shape of confidence region is reliable (Σ wrong)
- *thank you* www.stat.rice.edu/~scottdw