A Minimum Distance Partial Mixture Approach To Finding Clusters

David W Scott
Department of Statistics
Rice University

September 10, 2018

Introductory Thoughts

- Since any minimum distance criterion is inherently robust (Donoho and Liu, 1998), this talk is really about robustness
- Robustness once was an esoteric topic
- Claim that these ideas are becoming central in "real" data sciences
- The old statistical paradigm was model, fit, criticize, adjust, repeat
- However, if the model is in fact correct, but some significant fraction of contamination (or mixing) has occurred, the chances of correctly discovering that fact, appropriately marking, etc, are not good
- Residuals are not always helpful as one might imagine
- Clustering is still a grand challenge, and fertile ground for robust methods
- However, initialization, whether k-means or Gaussian mixtures of k components is said to be an NP-hard problem

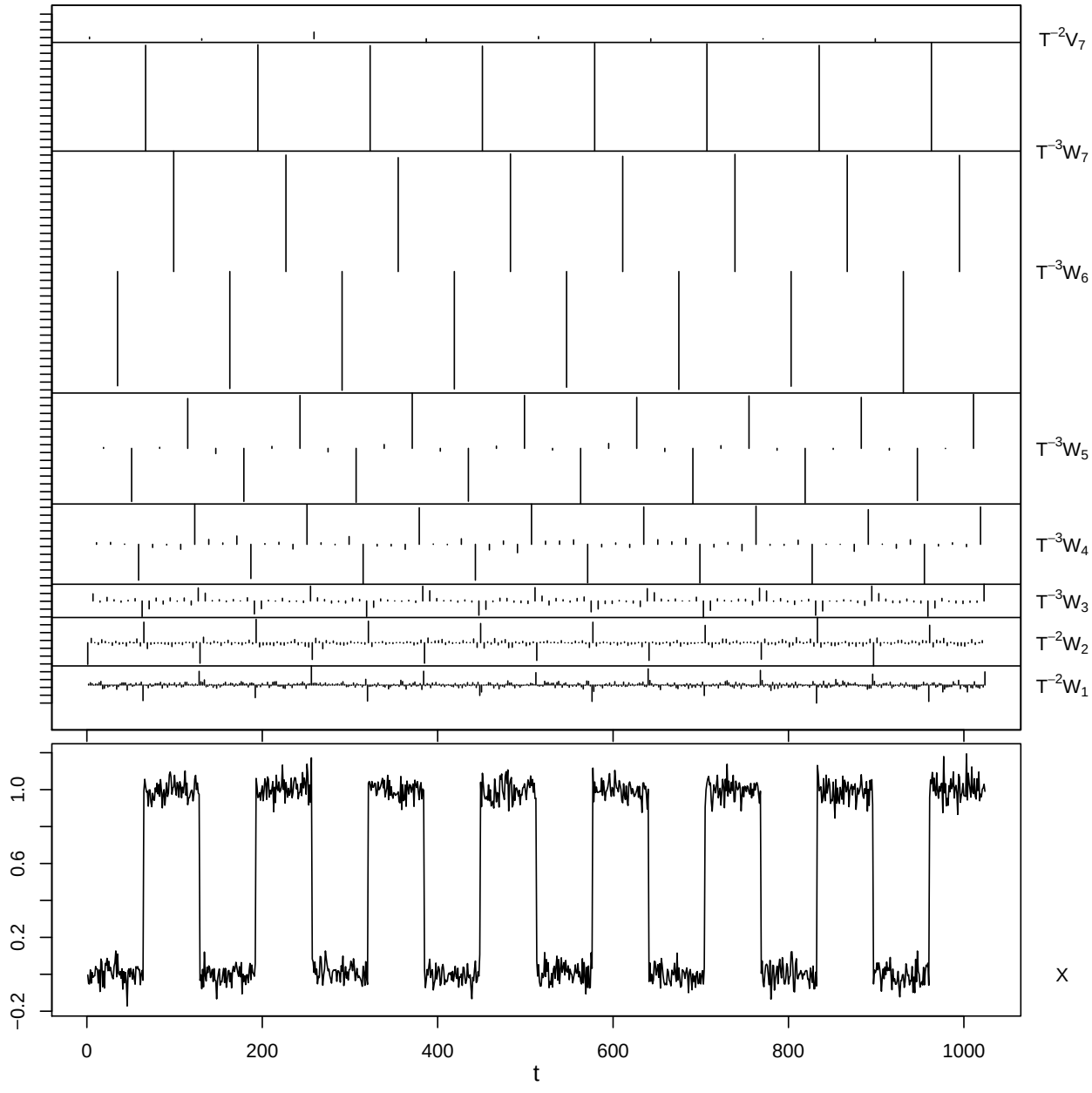
- Tukey's original PRIM9 system illustrated the idea of sequentially picking off clusters (or subsets thereof); we like that idea

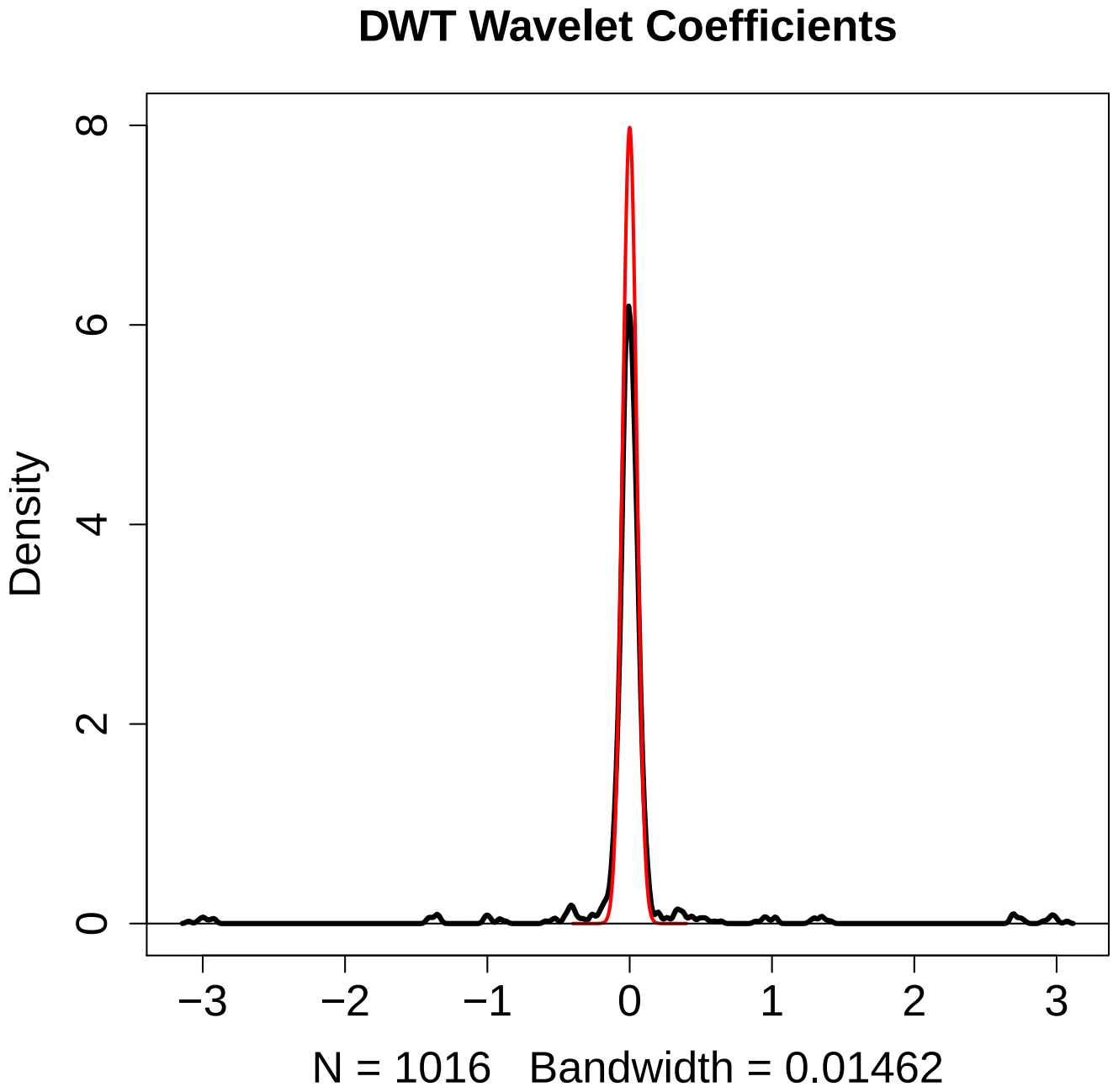
- Tukey $\rightarrow$ Thompson $\rightarrow$; my academic grandfather

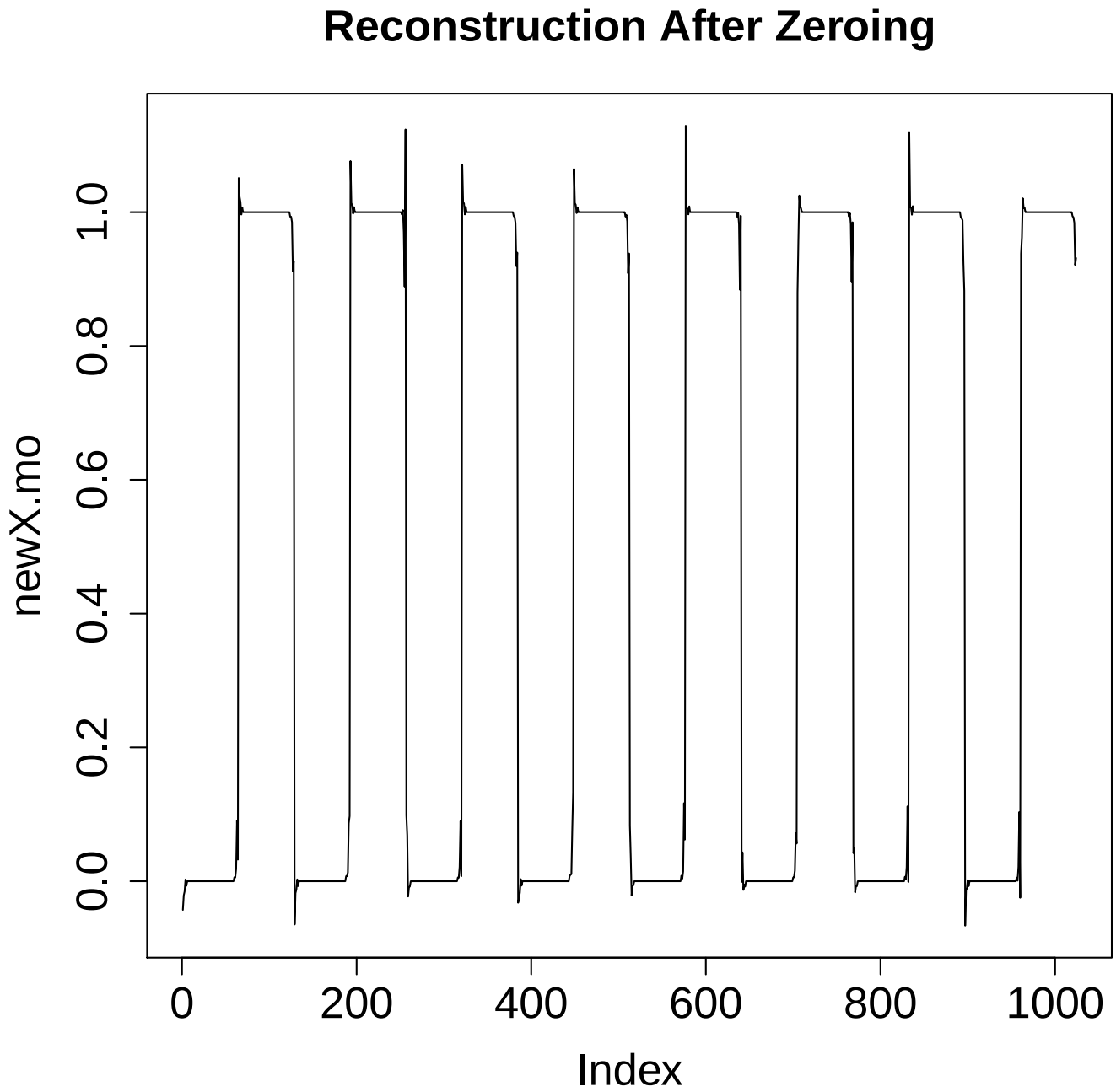
- Robust regression; the real workhorse
- Image analysis using wavelets (DWT); known that most coefficients are zero (and Gaussian!); what sense does it make to think of the nonzero coefficients following any pdf? (Bayesian spike and slab prior?)
- Sometimes know the signal distribution, but not the contamination pdf
- Are we ready for ultra-high dimensions???
- Today, we will not focus on theory—there are many good books and papers for that purpose
- We will focus on the exploratory and diagnostic power of robust procedures

DWT Example

- For a time series, you might know the statistics of the signal
- But you might not know the statistics of the outliers/noise?
- If you take the discrete wavelet transform (DWT) of a highly structured time series, the signal may become sparse
- Good theory to suggest that most of the sample DWT coefficients are normal with 0 mean and small variance
- (Same is often true when analyzing images.)
- In the transformed space, you know the statistics of the noise!!
- The wavelet coefficients of the signal follow no particular distribution; they are just nonzero (and large)
- A quick example using the R wavelet package: Fit a normal mixture (by L2E) centered at 0 and zero those coefficients.
- Kills almost 97% of the coefficients.


$T^{-2}V_7$
$T^{-3}W_7$
$T^{-3}W_6$
$T^{-3}W_5$
$T^{-3}W_4$
$T^{-3}W_3$
$T^{-2}W_2$
$T^{-2}W_1$
X
1.0
0.6
0.2
−0.2
0
200
400
600
800
1000
t


DWT Wavelet Coefficients
Density
0
2
4
6
8
−3
−2
−1
0
1
2
3
N = 1016 Bandwidth = 0.01462


Reconstruction After Zeroing
newX.mo
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Index

Does Artificial Intelligence/Data Sciences Need Robustness?

- how much consideration so far at NIPS conference?
- papers about robustness to label errors in clusters
- claims of robustness if the fraction of outliers much less than the number of nodes (e.g. On Robustness of Kernel Clustering, Yan and Sarkar, 2016)

- Deep Learning relies on truly massive datasets and large networks to capture "everything" in the data
- What if there are huge subsets of junk, or poorly understood/labeled portions?
- What is "learned" about those? How are decisions influenced?
- The black box nature of these nets makes that quick difficult to answer.

- Tesla crash in construction zone perhaps caused by original painted line dividers becoming uncovered over time?

NTSB investigating whether damaged crash barrier contributed to fiery fatal Tesla wreck

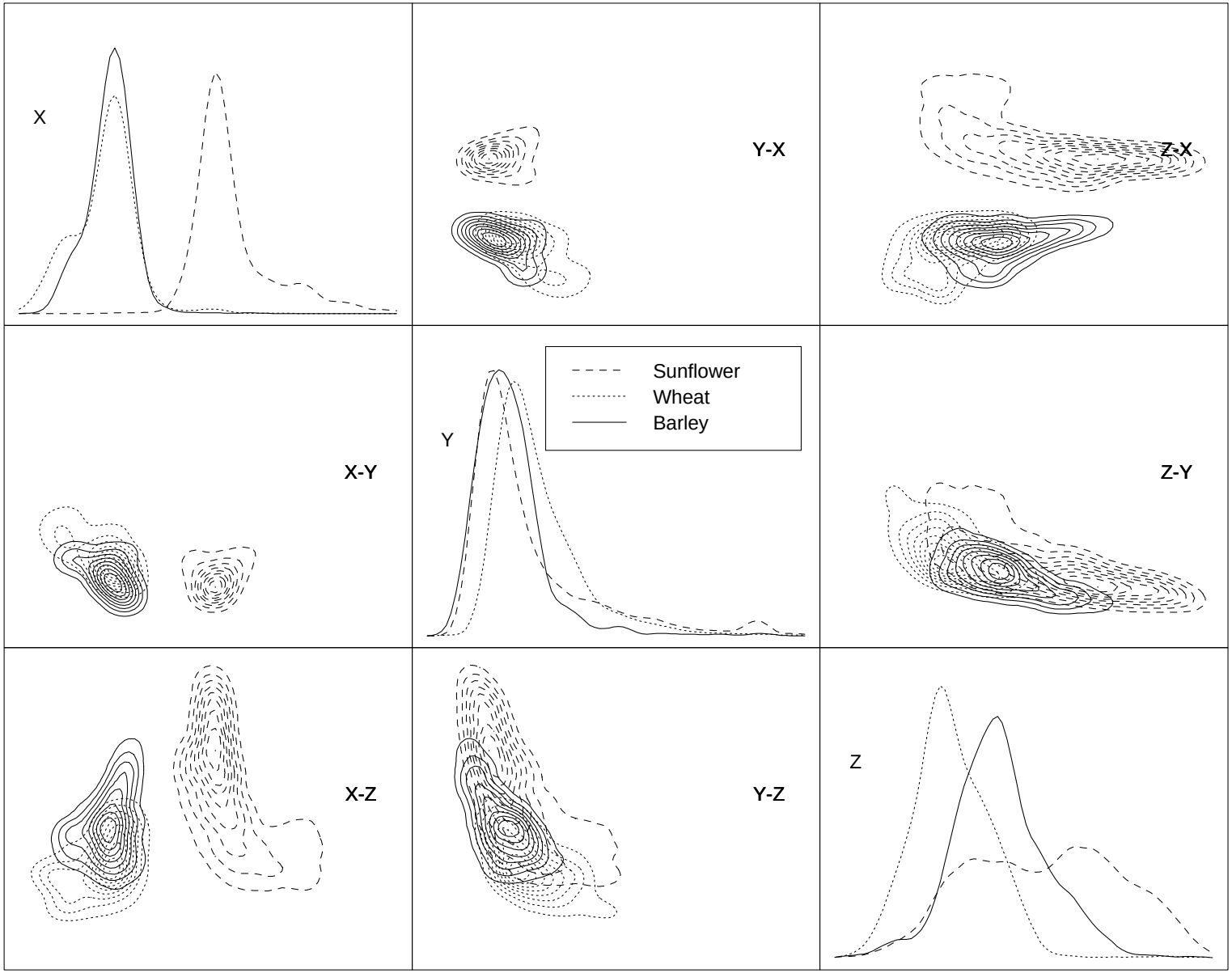
By **Faiz Siddiqui** March 28 Email the author

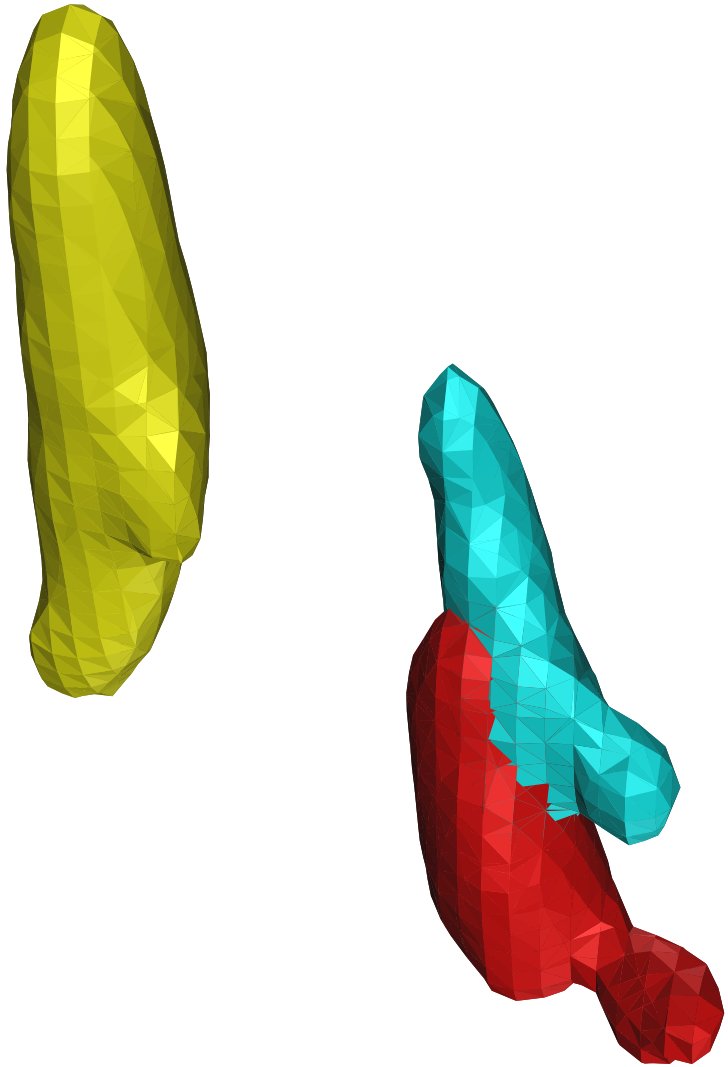
Emergency personnel work March 23 at the scene where a Tesla electric SUV crashed into a barrier on U.S. Highway 101 in Mountain View, Calif. (KTVU/AP)

Entry into Topic of Robustness via Nonparametric Density Estimation: Real Data Has Outliers and Strange Features

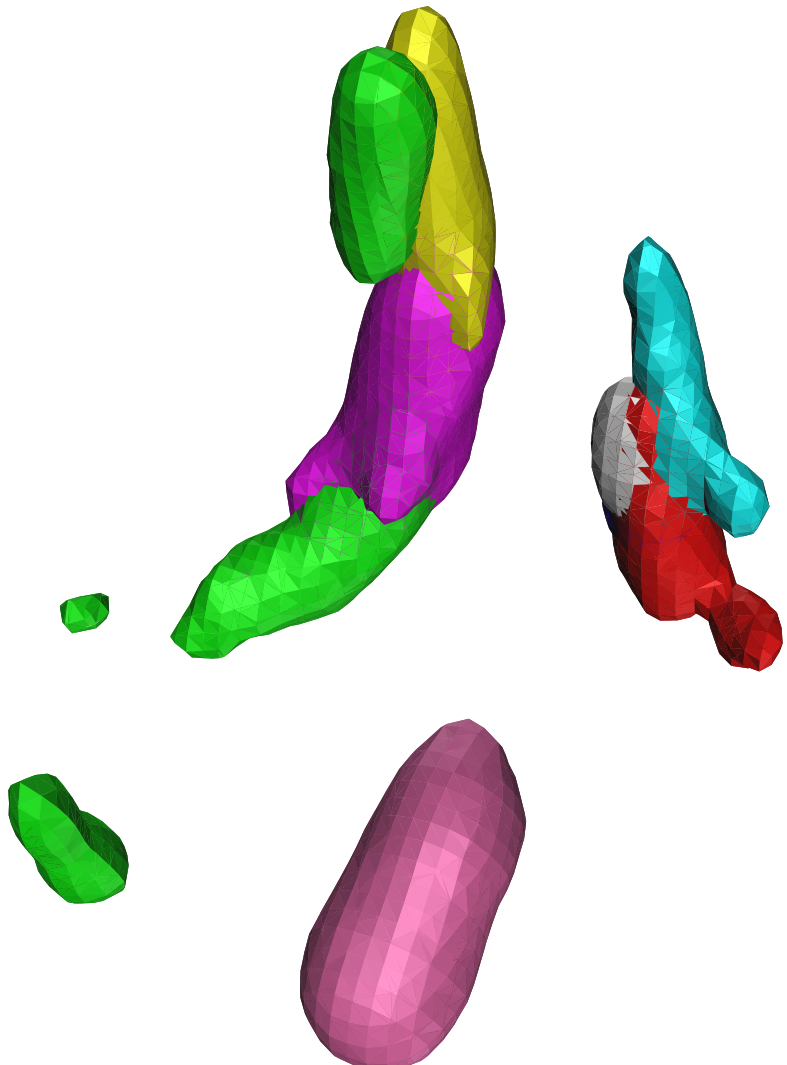
- LANDSAT IV (labelled) sample from North Dakota
- $x \in \Re^{24}$, $n \approx 23,000$; nonlinear transformation $\Re^{24} \longrightarrow \Re^3$
- Yes, "real" data has outliers and strange features

Three Most Common Crops: Sunflower, Wheat, Barley





Including Other Crops



TBC

Clarifies modern data analysis through nonparametric density estimation for a complete working knowledge of the theory and methods

Featuring a thoroughly revised presentation, *Multivariate Density Estimation: Theory, Practice, and Visualization, Second Edition* maintains an intuitive approach to the underlying methodology and supporting theory of density estimation. Including new material and updated research in each chapter, the *Second Edition* presents additional clarification of theoretical opportunities, new algorithms, and up-to-date coverage of the unique challenges presented in the field of data analysis.

The new edition focuses on the various density estimation techniques and methods that can be used in the field of big data. Defining optimal nonparametric estimators, the *Second Edition* demonstrates the density estimation tools to use when dealing with various multivariate structures in univariate, bivariate, trivariate, and quadrivariate data analysis. Continuing to illustrate the major concepts in the context of the classical histogram, *Multivariate Density Estimation: Theory, Practice, and Visualization, Second Edition* also features:

- Over 150 updated figures to clarify theoretical results and to show analyses of real data sets
- An updated presentation of graphic visualization using computer software such as R
- A clear discussion of selections of important research during the past decade, including mixture estimation, robust parametric modeling algorithms, and clustering
- More than 130 problems to help readers reinforce the main concepts and ideas presented
- Boxed theorems and results allowing easy identification of crucial ideas

Multivariate Density Estimation: Theory, Practice, and Visualization, Second Edition is an ideal reference for theoretical and applied statisticians, practicing engineers, as well as readers interested in the theoretical aspects of nonparametric estimation and the application of these methods to multivariate data. The *Second Edition* is also useful as a textbook for introductory courses in kernel statistics, smoothing, advanced computational statistics, and general forms of statistical distributions.

DAVID W. SCOTT, PhD, is Noah Harding Professor in the Department of Statistics at Rice University. The author of over 100 published articles, papers, and book chapters, Dr. Scott is also Fellow of the American Statistical Association (ASA) and the Institute of Mathematical Statistics. He is recipient of the ASA Founder's Award and the Army Wilks Award and his research interests include computational statistics, data visualization, and density estimation. Dr. Scott is coeditor of *Wiley Interdisciplinary Reviews: Computational Statistics* and former editor of the *Journal of Computational and Graphical Statistics*.

with website

Cover Image: (background) iStockphoto © Studio-Pro
(figure) Courtesy of the author

WILEY

Also available as an e-book

Scott

MULTIVARIATE DENSITY ESTIMATION

SECOND EDITION

with website

WILEY

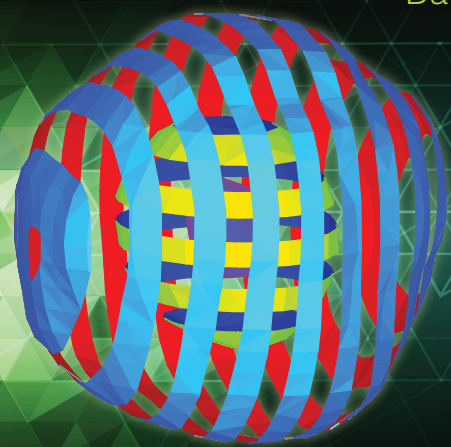
Wiley Series in Probability and Statistics

MULTIVARIATE DENSITY ESTIMATION

Theory, Practice, and Visualization

David W. Scott

SECOND EDITION



with website

WILEY

Entry into Topic of Robustness via Nonparametric Density Estimation: One Motivation for *L2E* Minimum Distance

- My general interest in robustness was certainly born of my advisor's (Jim Thompson) advisor (John Tukey); however, it was thru NPDE that I pursued a different path
- **Remark: L2E refers to the use of the L2 norm as the criterion in a minimum distance estimator (E)**
- Consider the usual kernel density estimator (KDE)

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- In one sense, the kernel estimator is very robust to outliers
  - if add outliers to the data, then change in $\hat{f}(x)$ is very small
  - however, some functionals of the KDE are not robust
  - for example, $\int x \hat{f}(x)\, dx = \bar{x}$
- use KDE as a graphical tool for identifying outliers?
- choice of kernel will influence perception of outliers
- do Cauchy data have outliers? Black Swan events?

- **Strategy:** Borrowing ideas from parametric estimation for nonparametric estimation and **vice-versa**
- Consider problem of choosing histogram $\hat{f}_h(x)$ bin width h:
  - MLE approach: (Ordinary and LOO)

$$\hat{h} = \arg\max_h \prod_{i=1}^{n} \hat{f}_h(x_i) \qquad \text{degenerate solution } \hat{h} = 0$$

$$\hat{h} = \arg\max_h \prod_{i=1}^{n} \hat{f}_{h,-i}(x_i) \quad \text{Habbema, Hermans, van den Broek}$$

  - Asymptotics approach: (BCV, Plug-in)

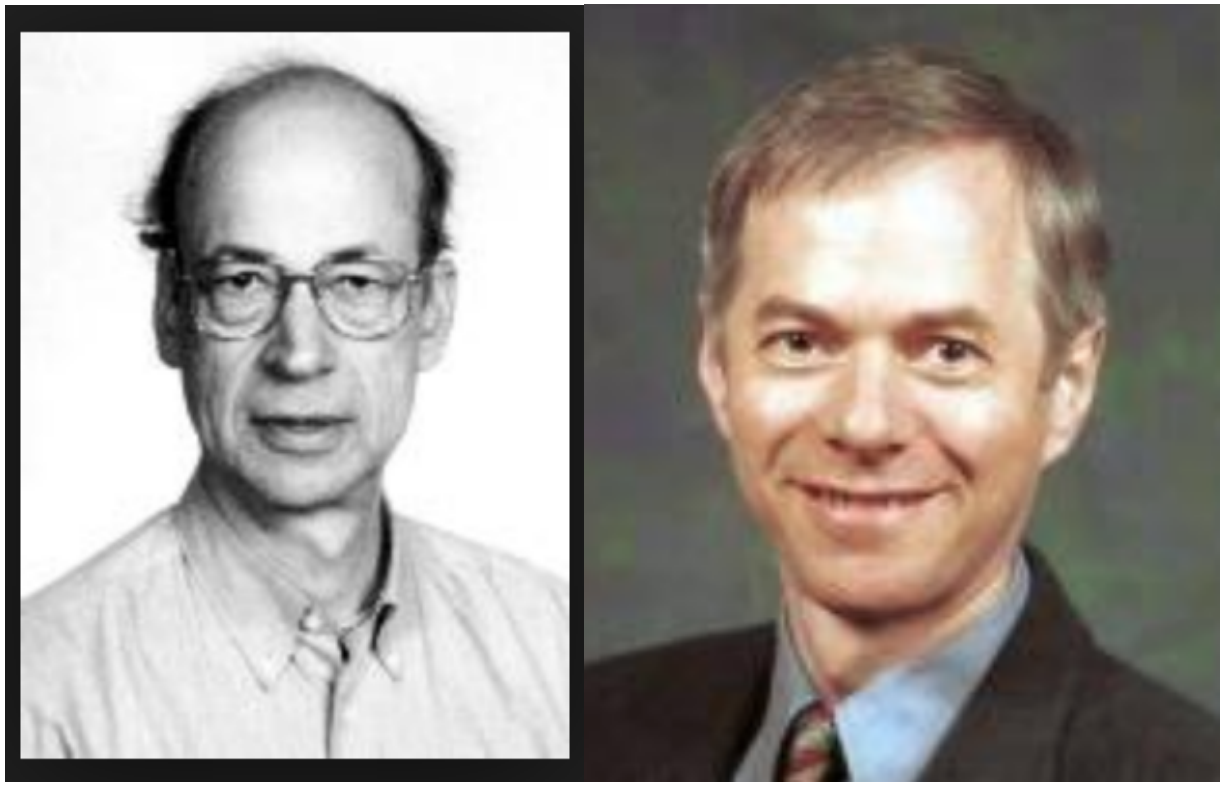
$$AMISE(h) = \frac{1}{nh} + \frac{1}{12} h^2 \int f'(x)^2 dx$$

$$h^* = \left[\frac{6}{n \int f'(x)^2 dx} \right]^{1/3} \qquad \text{Scott (1979) } \textit{Biometrika}$$

$$\boxed{h^* = 3.5\, \hat{\sigma}\, n^{-1/3} \qquad \text{normal reference rule}}$$

Mats Rudemo's Least-Squares Cross-Validation (LSCV)

- M Rudemo (1982), **Empirical Choice of Histograms & Kernel Density Estimators**, *Scand J of Stats*, 9:65–78.



- Adrian Bowman (1984), **An alternative method of cross-validation for the smoothing of density estimates**, *Biometrika*, 71:353–360.

$$\hat{h} = \arg\min_h \int_{x=-\infty}^{\infty} \left[\hat{f}_h(x) - f(x)\right]^2 dx$$

$$= \arg\min_h \int_{x=-\infty}^{\infty} \left[\hat{f}_h(x)^2 - 2\,\hat{f}_h(x)\,f(x) + f(x)^2\right] dx$$

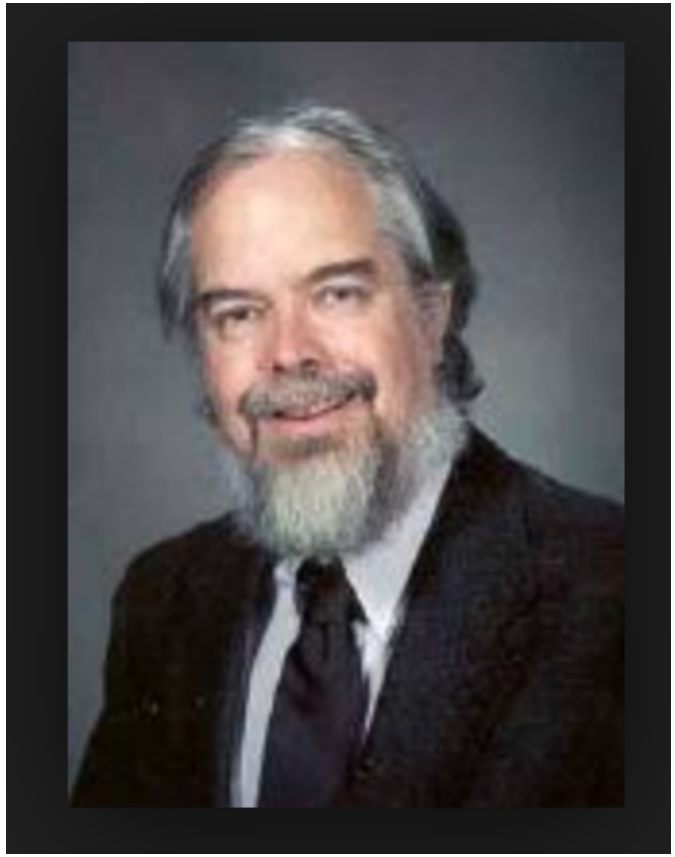
$$= \arg\min_h \left[\int \hat{f}_h(x)^2\,dx - 2\int \hat{f}_h(x)\,f(x)\,dx + \cancel{\int f(x)^2\,dx}\right]$$

$$= \arg\min_h \left[\int \hat{f}_h(x)^2\,dx - 2\,E\hat{f}_h(X)\right] \quad \text{where } X \sim f(x)$$

$$= \arg\min_h \left[\int \hat{f}_h(x)^2\,dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(x_i)\right] \quad \text{as an unbiased estimator}$$

- Note the use of the leave-one-out (LOO) density estimator again.
- This is nearly the *L2E* minimum distance criterion for θ

Joint Work w/ George Terrell in NPDE



- Education: 1977 Phd Mathematics, Rice University
  "Nonsplitting of H-Space Sequences," Morton Curtis, Director
- 1977–1986 BCM, UH, NASA/Lockheed, Rice
- 1986–present Associate Professor, Dept Stats, Virginia Tech

George Terrell, Joint Publications (All Dealt With Bandwidth Variational Problems in Some Fashion)

- Terrell/Scott (1980). **On improving convergence rates for nonnegative kernel density estimates.** Annals of Statistics, 8:1160–1163.
- Terrell/Scott (1985). **Oversmoothed nonparametric density estimates.** JASA, 80:209–214.
- Scott/Terrell (1987). **Biased and unbiased cross-validation in density estimation.** JASA, 82:1131–1146.
- Terrell/Scott (1992). **Variable kernel density estimation.** Annals of Statistics, 20:1236–1265.

Significant Solo Publication:

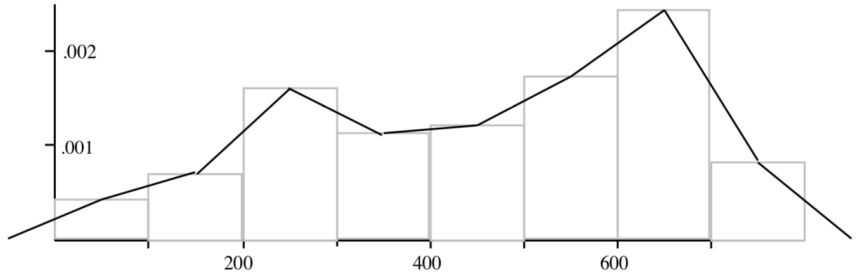
Terrell, George R. (1990). **Linear density estimates.** Proceedings of the Statistical Computing Section, Joint Statistical Meetings, pp. 297 - 302. (Unpublished book in ASA/SIAM series.)

- George examined nonparametric density estimators that solve the following variational problem:

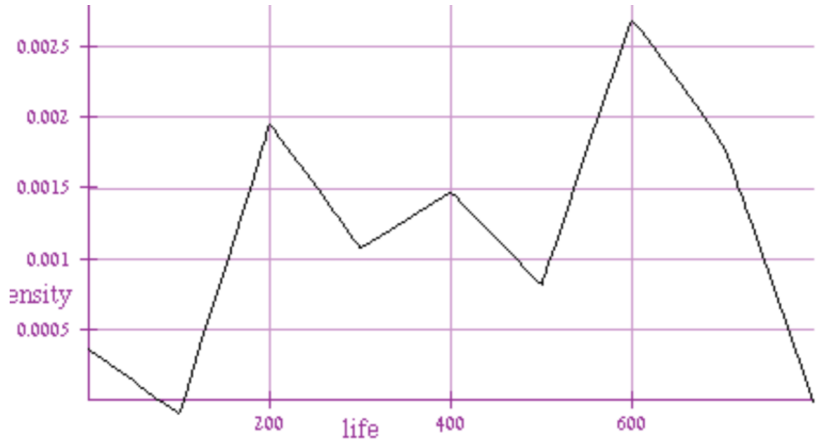
$$\hat{f} = \arg\min_{f \in H} \left[\int f(x)^2 \, dx - \frac{2}{n} \sum_{i=1}^{n} f(x_i) + R(f) \right],$$

- where H is a class of functions (first order splines, for example)
- and the penalty function $R(f)$ is selected by the investigator
- Note: LOO is not required for $\sum f(x_i)$ term

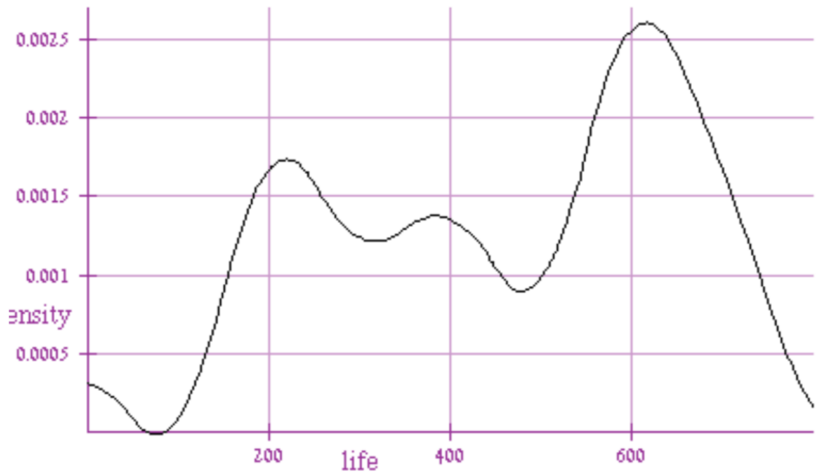
Three examples using mouse survival data: (1) ordinary frequency polygon; (2) linear spline; (3) quadratic spline.



Frequency Polygon for Mouse Survival Times



$l = 1$ Least-Squares Spline Density Estimate of Mouse Survival



$l = 2$ Mouse Survival Density

The *L2E* Criterion

- Just a simple extension of LSCV and Linear Density Estimation to *L2E*: Scott (2001), *Technometrics*
- Given a model $f_\theta(x) = f(x|\theta)$,

$$\hat{\theta} = \arg\min_\theta \left[\int f_\theta(x)^2 \, dx - \frac{2}{n} \sum_{i=1}^{n} f_\theta(x_i) \right] \qquad x \in \Re^d, \quad \theta \in \Re^p$$

- Secret to practical success is
  - the integral is available in closed form (numerical integration severely limits dimension and numerical optimization)
  - density also easy to evaluate
  - notice: does not require LOO in parametric setting
- For example, mixture of normals works well, since

$$\int_{\Re^d} \phi(x|\mu_1, \Sigma_1) \, \phi(x|\mu_2, \Sigma_2) \, dx = \phi(0|\mu_1 - \mu_2, \Sigma_1 + \Sigma_2)$$

The *L2E* Criterion

- I was using L2E to model the stochastic frontier while visiting Leopold Simar, Wolfgang Härdle, and Irene Gijbels in summer 1998
- While giving a talk at Texas A&M on 10/1/1998, Manny Parzen pointed out the relationship of L2E to the divergence work of Basu et al (1998). Ian Harris was up the road at SMU.
- I finally published the L2E paper in a special issue of Technometrics (2001) dedicated to John Tukey. Karen Kafadar was editor.
- Focus was on extensions to partial mixture estimation:

$$f_\theta = N(\mu, \sigma^2) \quad \text{versus} \quad f_\theta = w \cdot N(\mu, \sigma^2)\,,$$

the 3-parameter normal: more later.

Related Work

- Donoho, D.L. and R.C. Liu (1988). **The "Automatic" Robustness of Minimum Distance Functionals**. *The Annals of Statistics*, 16: 552–586.
- Beran, R. (1977). **Minimum Hellinger Distance Estimates for Parametric Models**. *The Annals of Statistics*, 5:445–463.
  - Impressive first attempt
  - Fails the practical test of not requiring numerical integration
- Basu, A., Harris, I.R., Hjort, N.L., & Jones, M.C. (1998). **Robust and Efficient Estimation by Minimising a Density Power Divergence.** *Biometrika*, 85:549–559.

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

  - $\alpha = 1$ is *L2E*; $\alpha \to 0$ is MLE
  - For normal mixtures, other values of α fail the practical test of not requiring numerical integration

Related Work

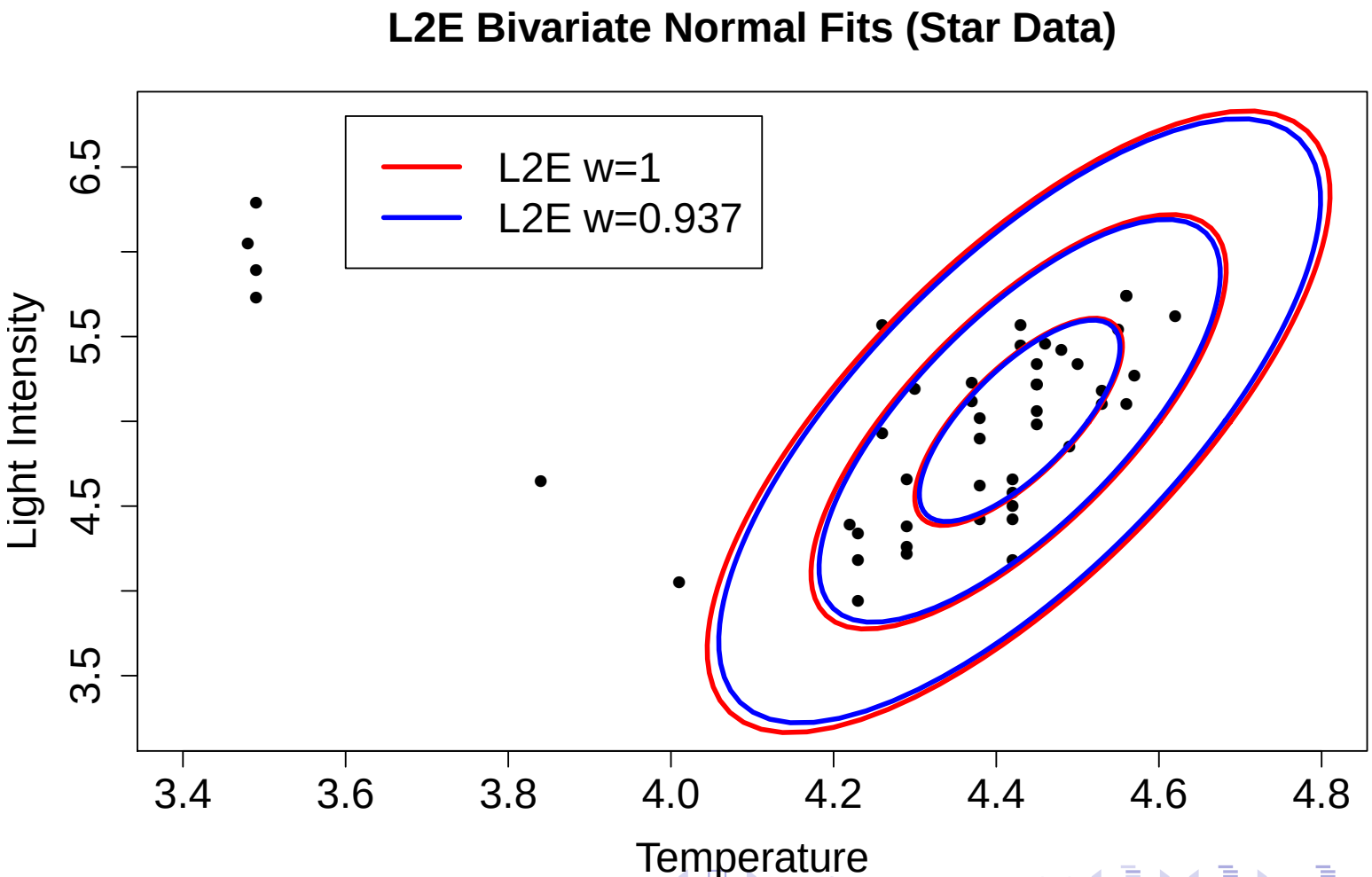
- Basu, Ayanendranath, Hiroyuki Shioya, and Chanseok Park (2011), **Statistical Inference: The Minimum Distance Approach.** CRC Press.
- Hjort, H.L. (1994), **Minimum L2 and Robust Kullback-Leibler Estimation,** Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, eds. P. Lachout and J.Á. Víšek, Prague Academy of Sciences of the Czech Republic, pp. 102 – 105.
- Huber, P.J. (1981), *Robust Statistics.* John Wiley & Sons, New York.
- Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.
- Scott, D.W. (2001), **Parametric Statistical Modeling by Minimum Integrated Square Error,** Technometrics, **43**, 274 – 285.

L2E Weaknesses

- Integrated squared error is not dimensionless
- Scaling of the criterion becomes important even in several dimensions
- Numerical optimization can overflow or underflow
- Represent a general covariance matrix by the Cholesky decomposition, but lots of parameters and numerical optimization limitations also limit p and d
- L2E is similar to MLE in that singularities are a potential attractor (infinite likelihood; negative infinite L2E)
- Use R's generic **nlminb** or **nlm** to perform optimization

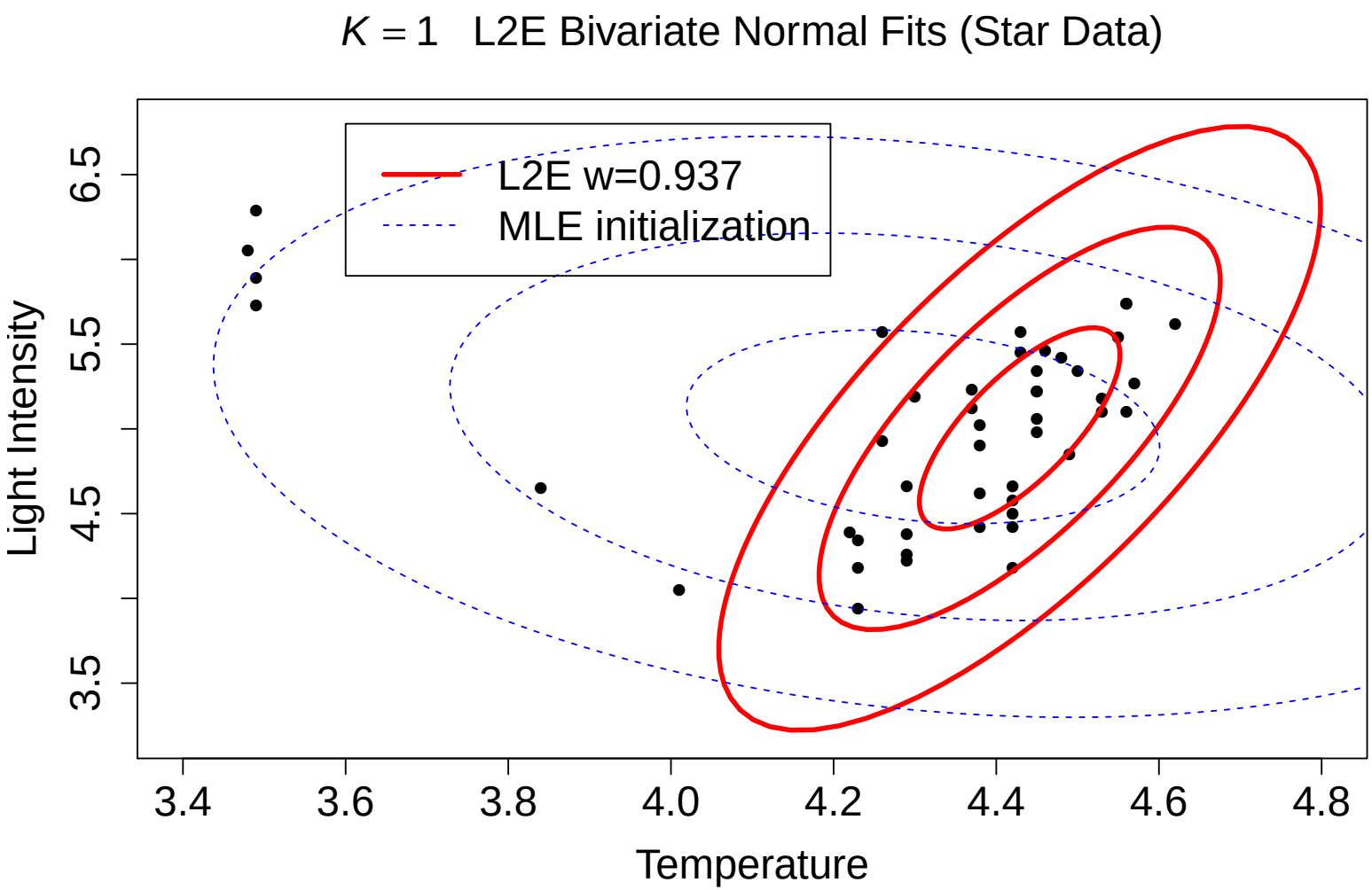
L2E Scaling Example

- Trying to fit a second mixture to the 4 giant stars fails
- Fails either with $K = 2$ or $K = 1$ starting at the 4 stars (singular covariance estimate — tries to fit only 2 of the 4)



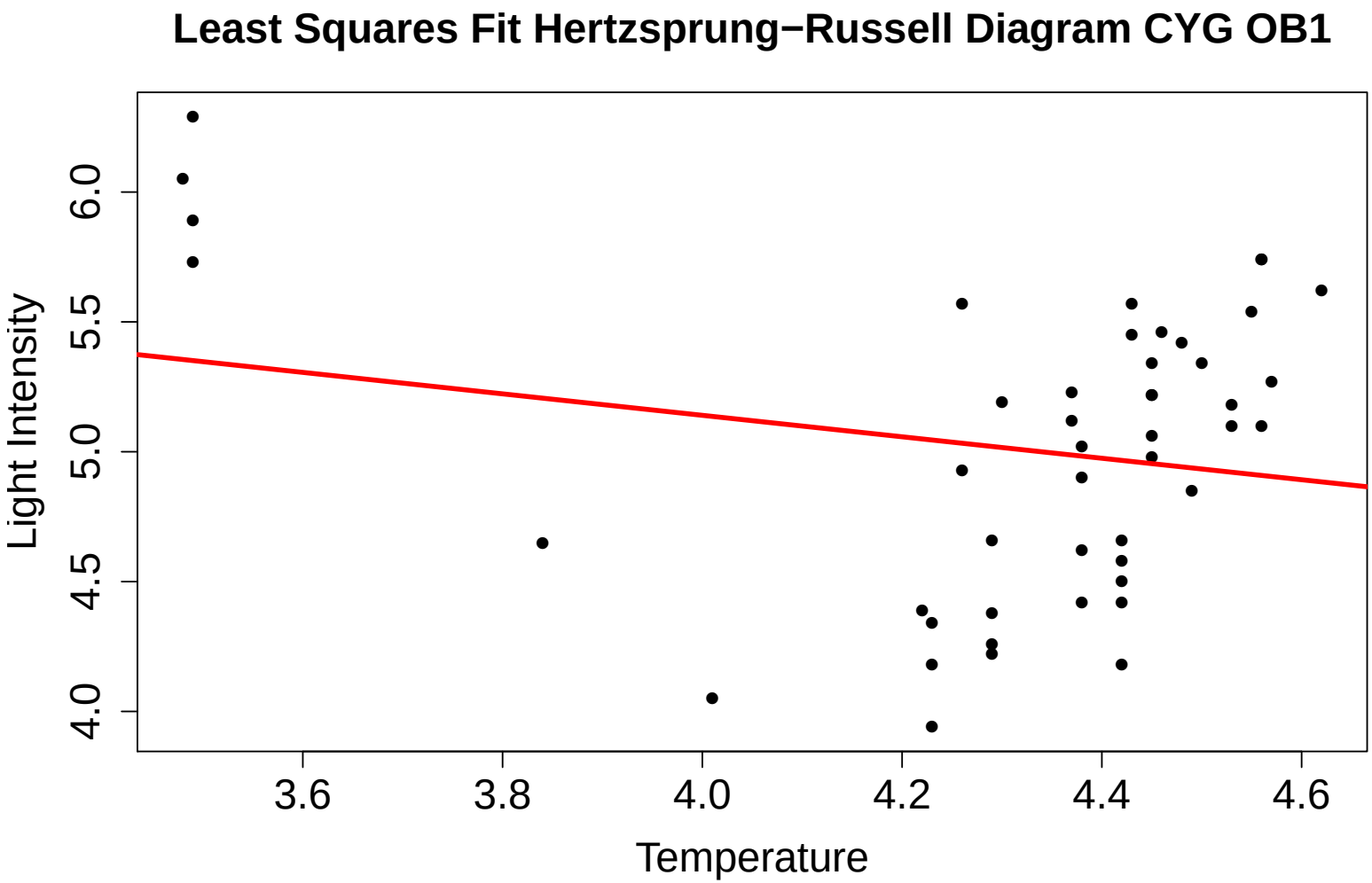
L2E Initialization Example

- There is not a local L2E solution starting at the MLE



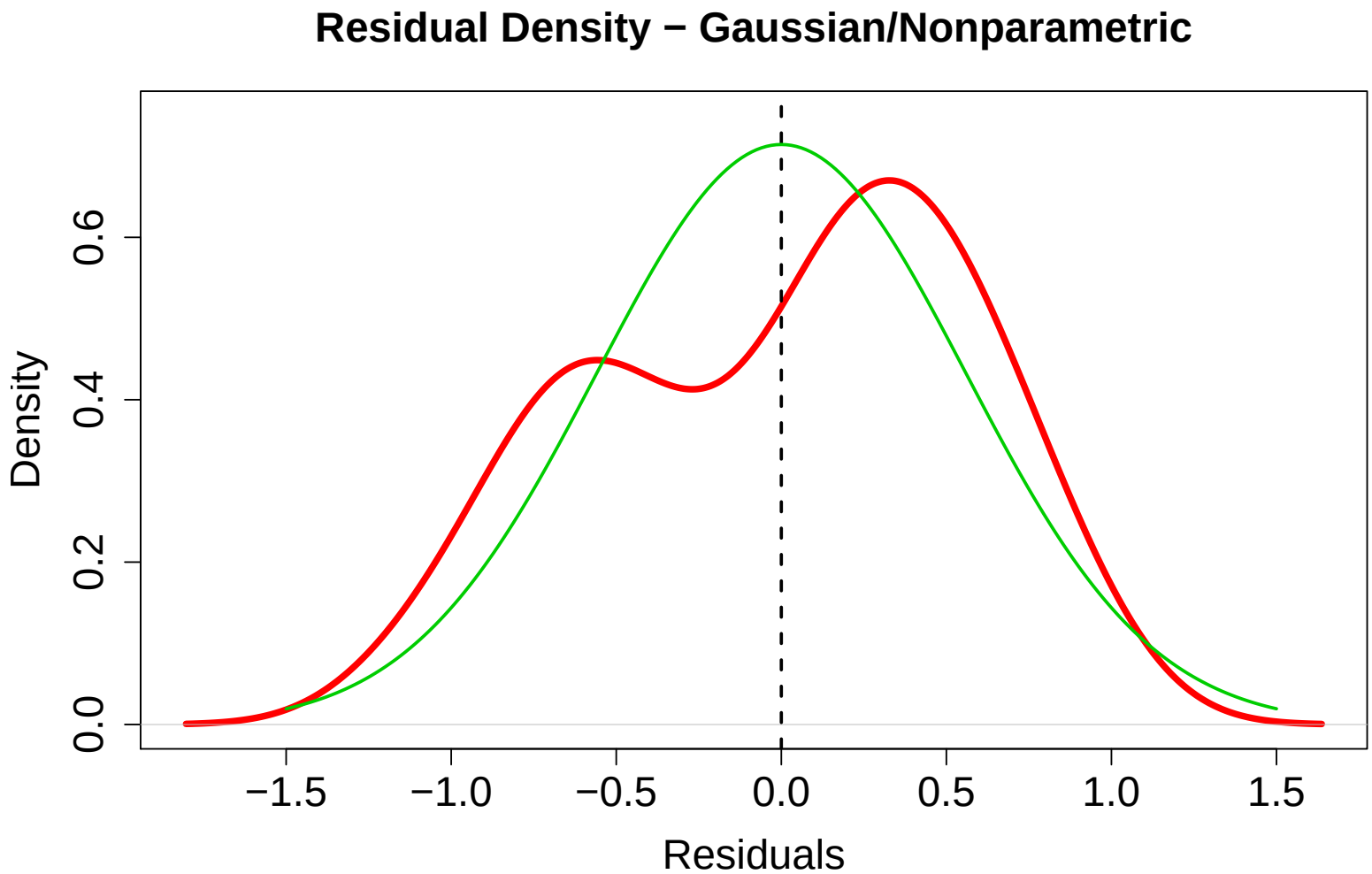
Difficulty Diagnosing Presence and Effect of Outliers

- Least squares fit (but note line does not go through outliers)



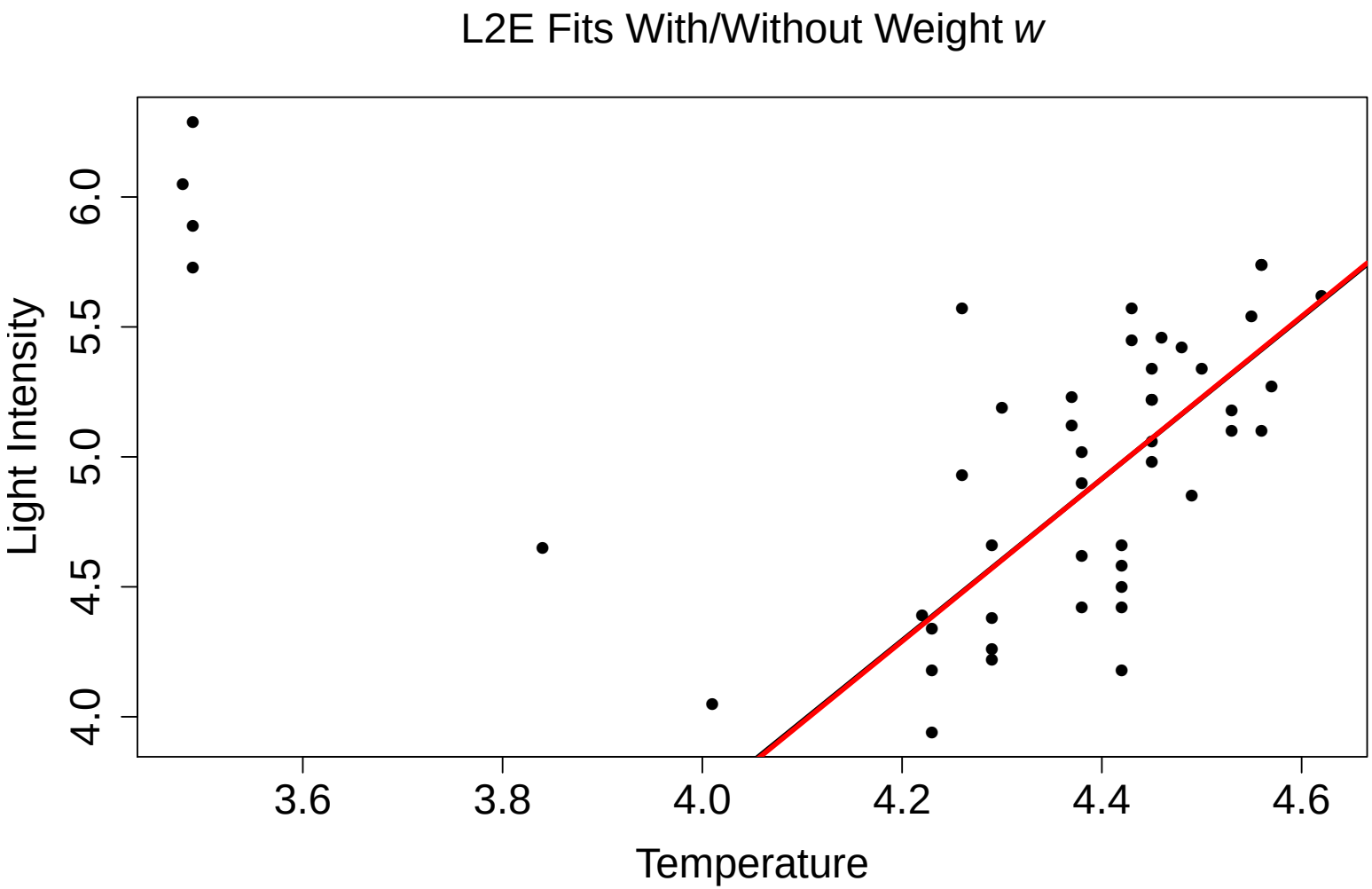
Least-Squares Residuals — Interpretation?

- Rough guess as to fraction of residuals? Model modification?



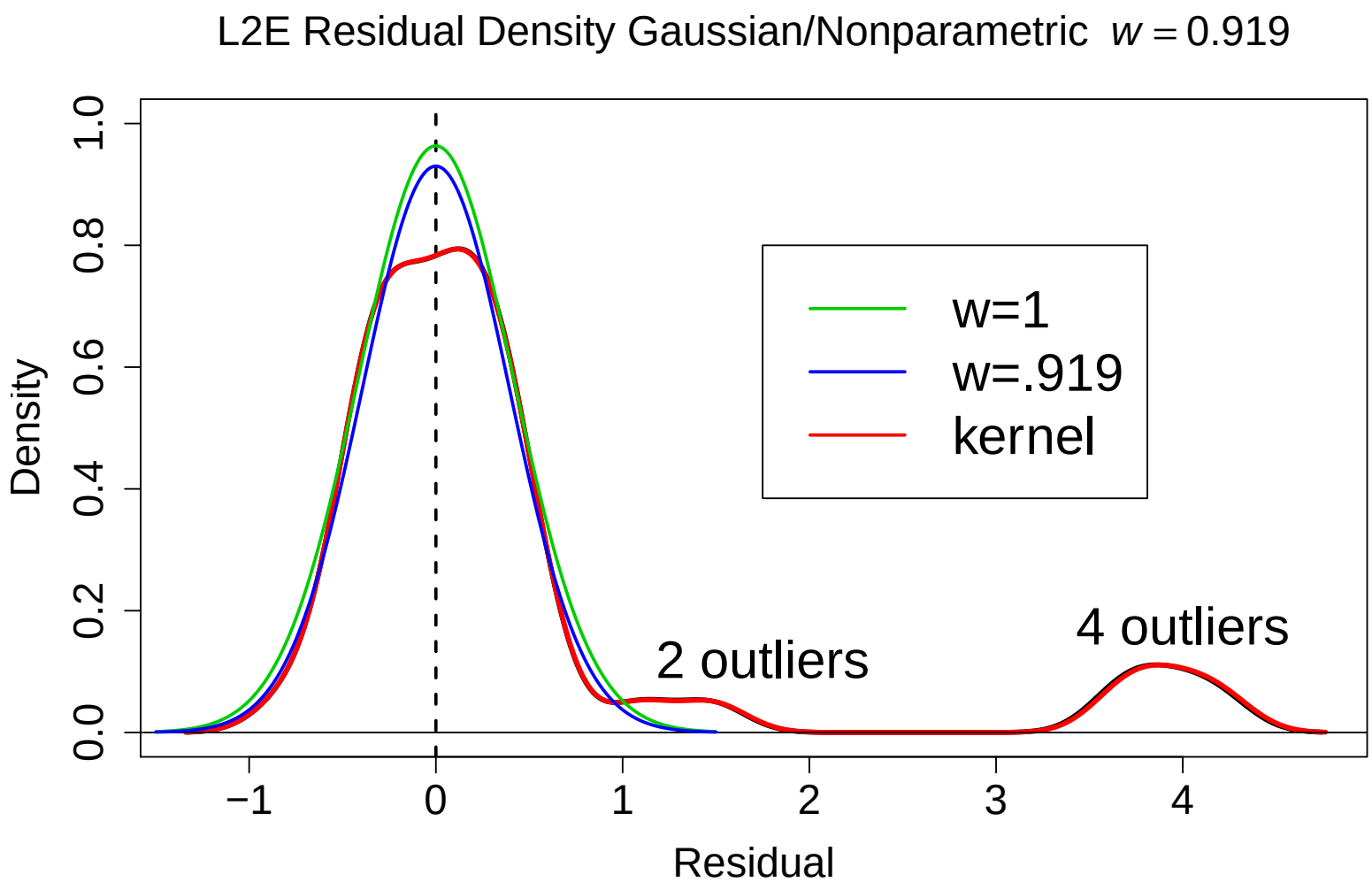
L2E Fit (With "Good" Initialization)

- Fit similar to that of a robust covariance estimate



L2E Residuals — Interpretation?

- $47 \times w = 43.2$, although visually 4-5 outliers?



Remarks: Difficulty Diagnosing Presence and Effect of Outliers

- Easy enough to "see" what is going on here
- Only 2-D
- In higher dimensions, can be quite challenging
- Trying many initial guesses for θ and including w as a parameter can greatly increase odds of effective diagnosis
- Often least-squares and L2E fits are very similar (Good! The model may fit all the data and robustness is not an issue)
- Return to this with Boston Housing data later (14-D)

General M-Estimation: Robustifying MLE

- For smooth PDF's, MLE maximizes

$$\hat{\theta}_{MLE} = \arg\min_{\theta} \sum_{i=1}^{n} \log\, f(x_i|\theta)$$

or, more generally,

$$\hat{\theta}_{MLE} = \arg\min_{\theta} \sum_{i=1}^{n} \rho(x_i|\theta)$$

- Thus we are looking for a root of

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{n} \rho(x_i|\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \rho(x_i|\theta) = 0$$

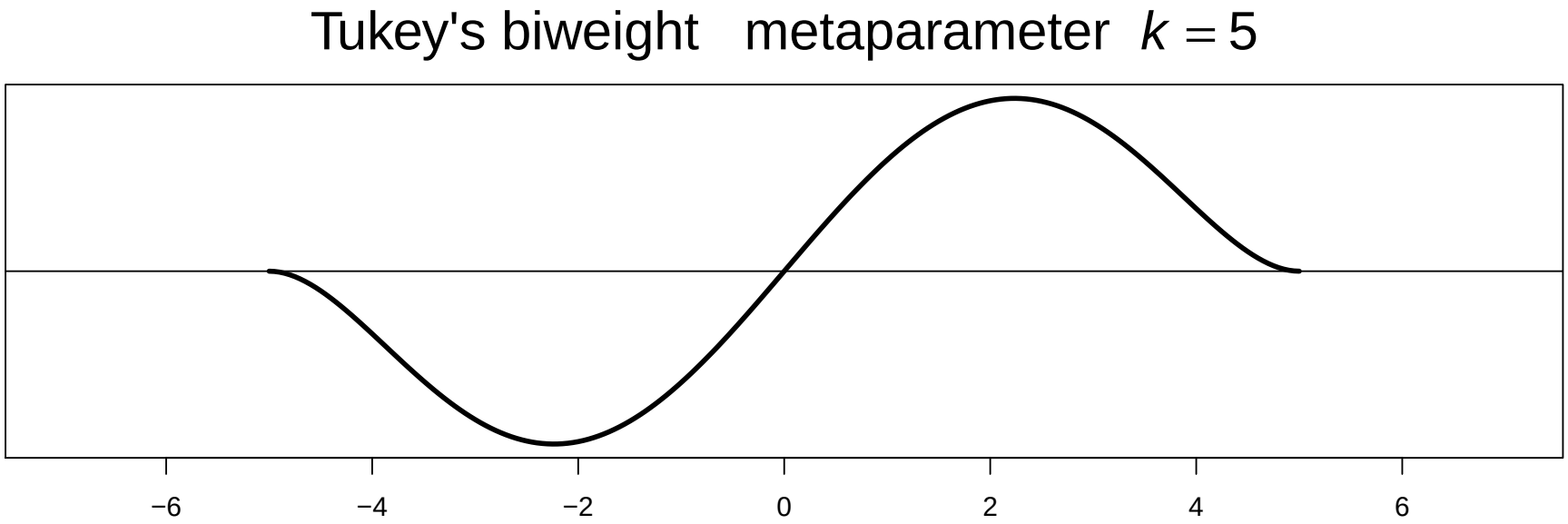
General M-Estimation: Robustifying MLE

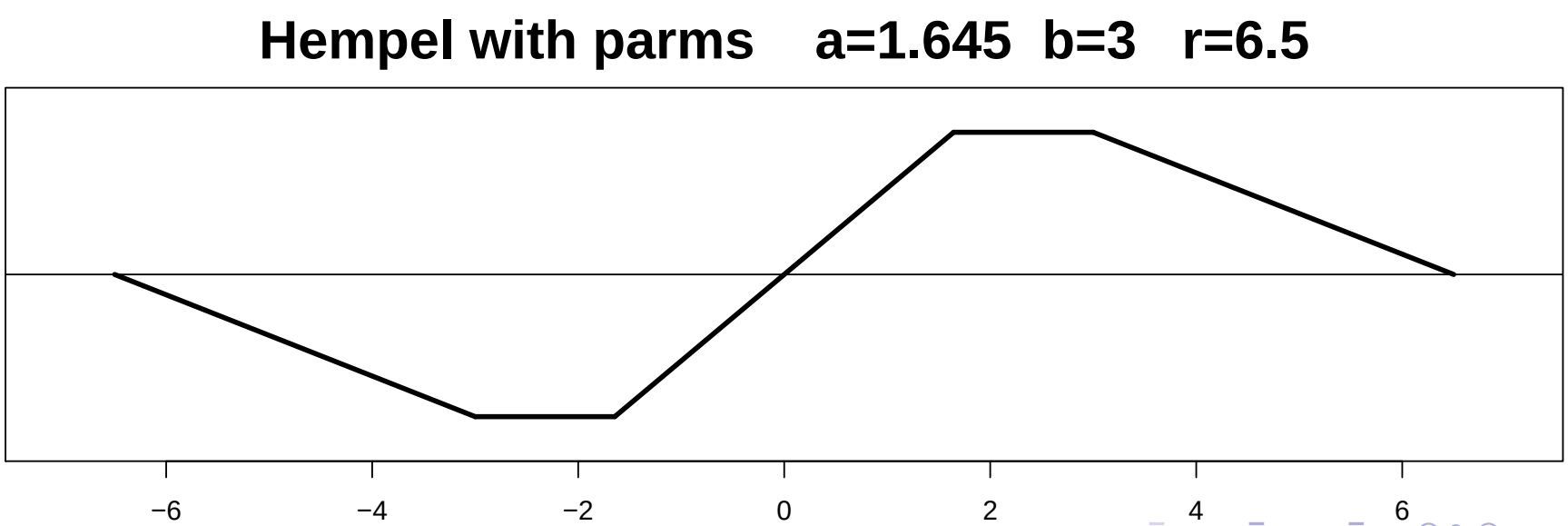
- Often, extreme features as $|x| \to \infty$ in the function

$$\psi(x|\theta) = \frac{\partial}{\partial\theta}\rho(x|\theta)$$

  result in estimators that are heavily influenced by outliers
- Cottage industry in modifying the shape of $\psi(x|\theta)$:
  - bounded influence
  - redescending
  - egs: Tukey, Huber, Hampel, ...
- One "cost" is getting the "scale" right — metaparameter
- L2E is an M-Estimator as well, but once the PDF $f(x|\theta)$ is specified, the $\rho(x,\theta)$ and $\psi(x,\theta)$ are also completed specified (implicitly; no metaparameters)

Tukey and Hempel Influence Function Examples





Consequences to M-Estimation of Choosing Density Model

- The robustness of L2E versus MLE is easiest to see with the location-shift pdf model $\phi(x|\mu, 1)$

$$\hat{\mu}_{L2E} = \arg\min_{\mu} \left[\frac{1}{2\sqrt{\pi}} - \sum_{i=1}^{n} \phi(x_i|\mu, 1) \right]$$

$$= \arg\max_{\mu} \sum_{i=1}^{n} \phi(x_i|\mu, 1)$$

$$\hat{\mu}_{MLE} = \arg\max_{\mu} \sum_{i=1}^{n} \log \left[\phi(x_i|\mu, 1) \right]$$

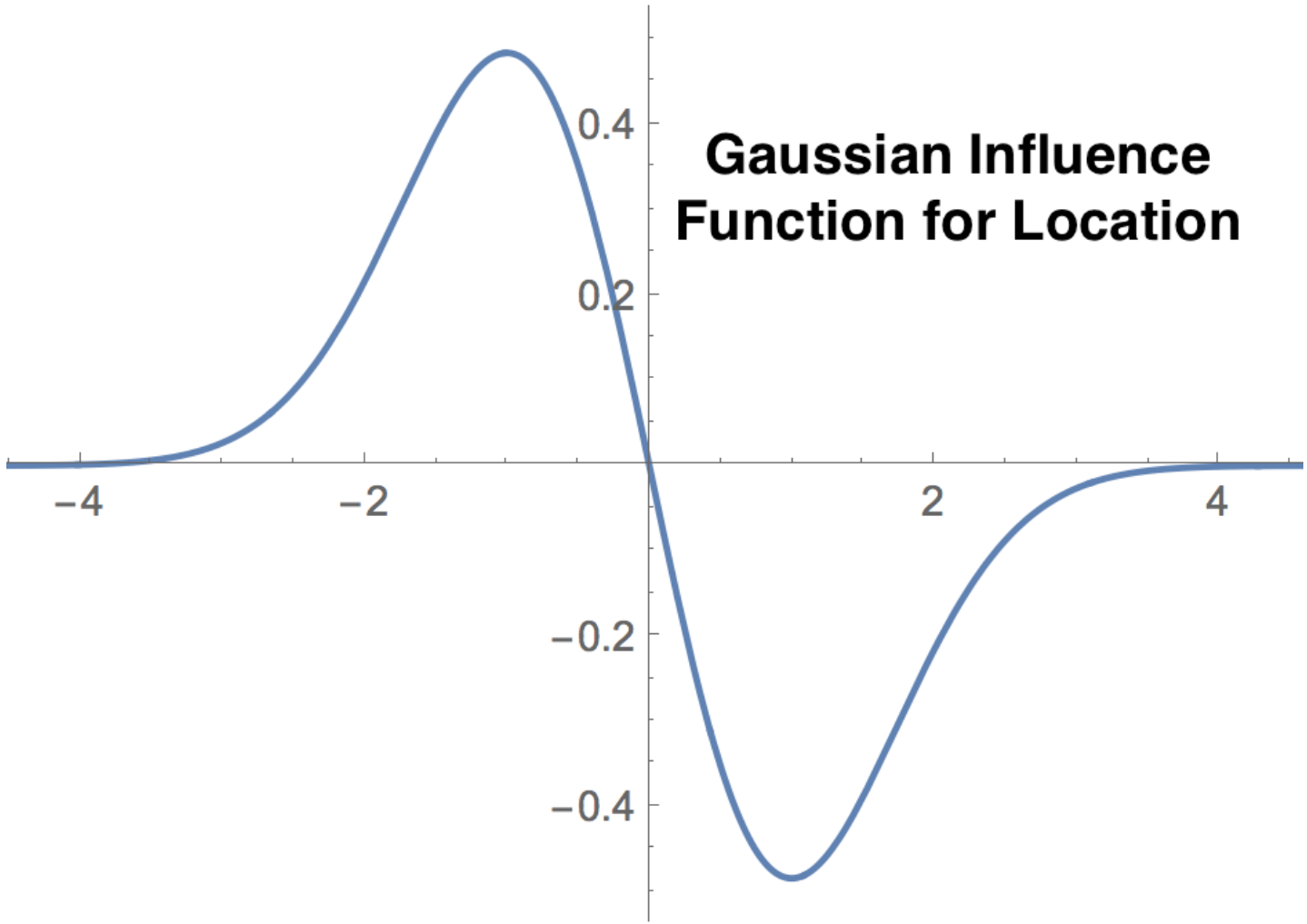
- L2E removes the logarithm, allowing for the impact of outliers to be diminished (bounded influence)
- that is, zero density values at x_i do not disqualify potential $\hat{\mu}$'s

Implicit L2E Influence Function Examples

- show the normal influence function shapes for μ (σ known)
- show the normal influence function shapes for σ
- do for the t distribution (as a scale parameter)
- do for the *negativeexponential* (scale parameter)
- *gamma* for skewness (fixed scale parameter)

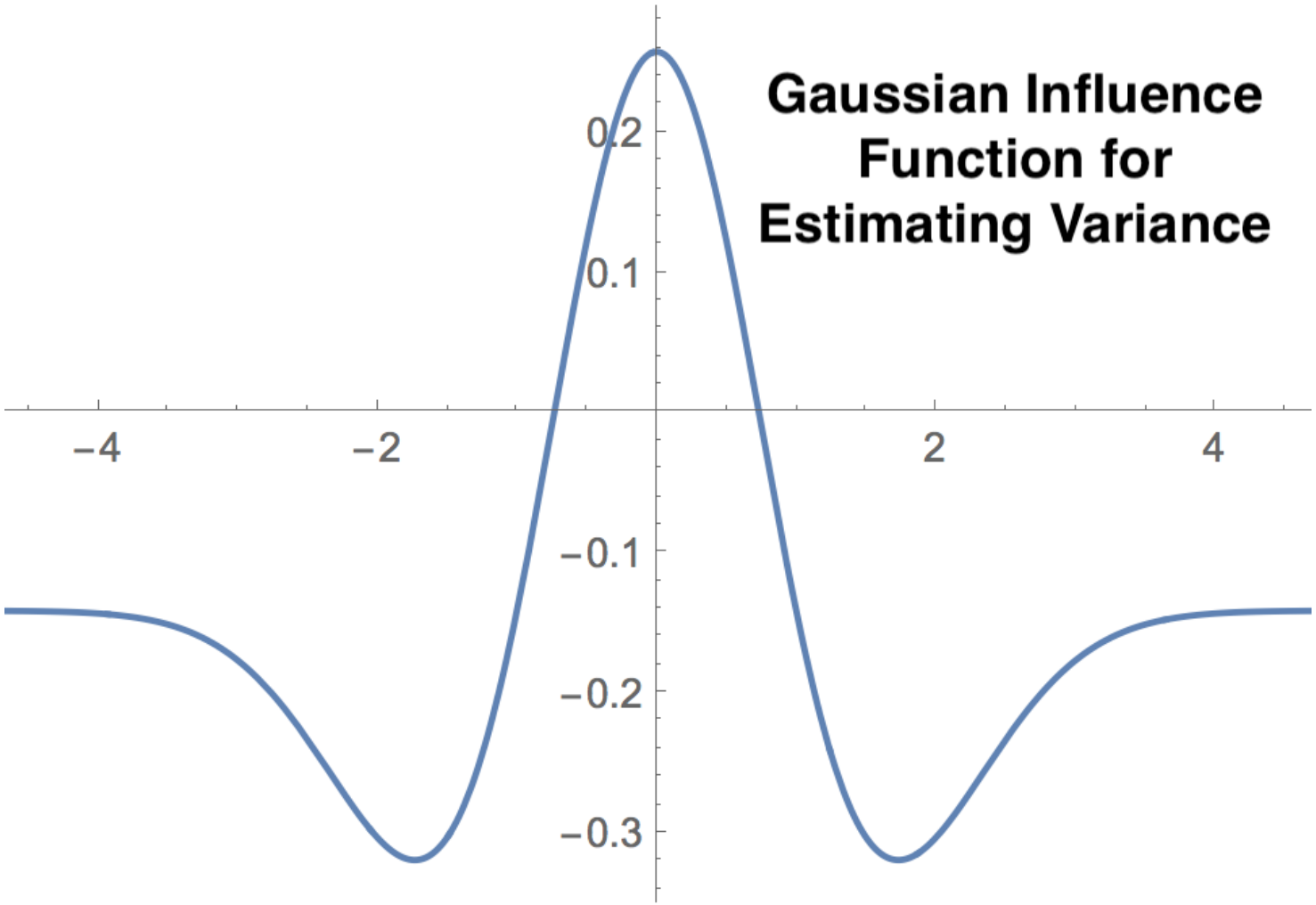
- All done symbolically using *Mathematica*

L2E Influence Function for μ: Normal w/ Known Variance



Redescending

L2E Influence Function for Variance



Partially redescending, but bounded influence

Thoughts

- It would be hard to imagine constructing by hand so many different choices for the influence function
- The estimates are not so different (I imagine) from a well-scaled Tukey, Huber, Hampel, or other influence function
- However, do not need to perform the extra step of specifying the hyper-parameters

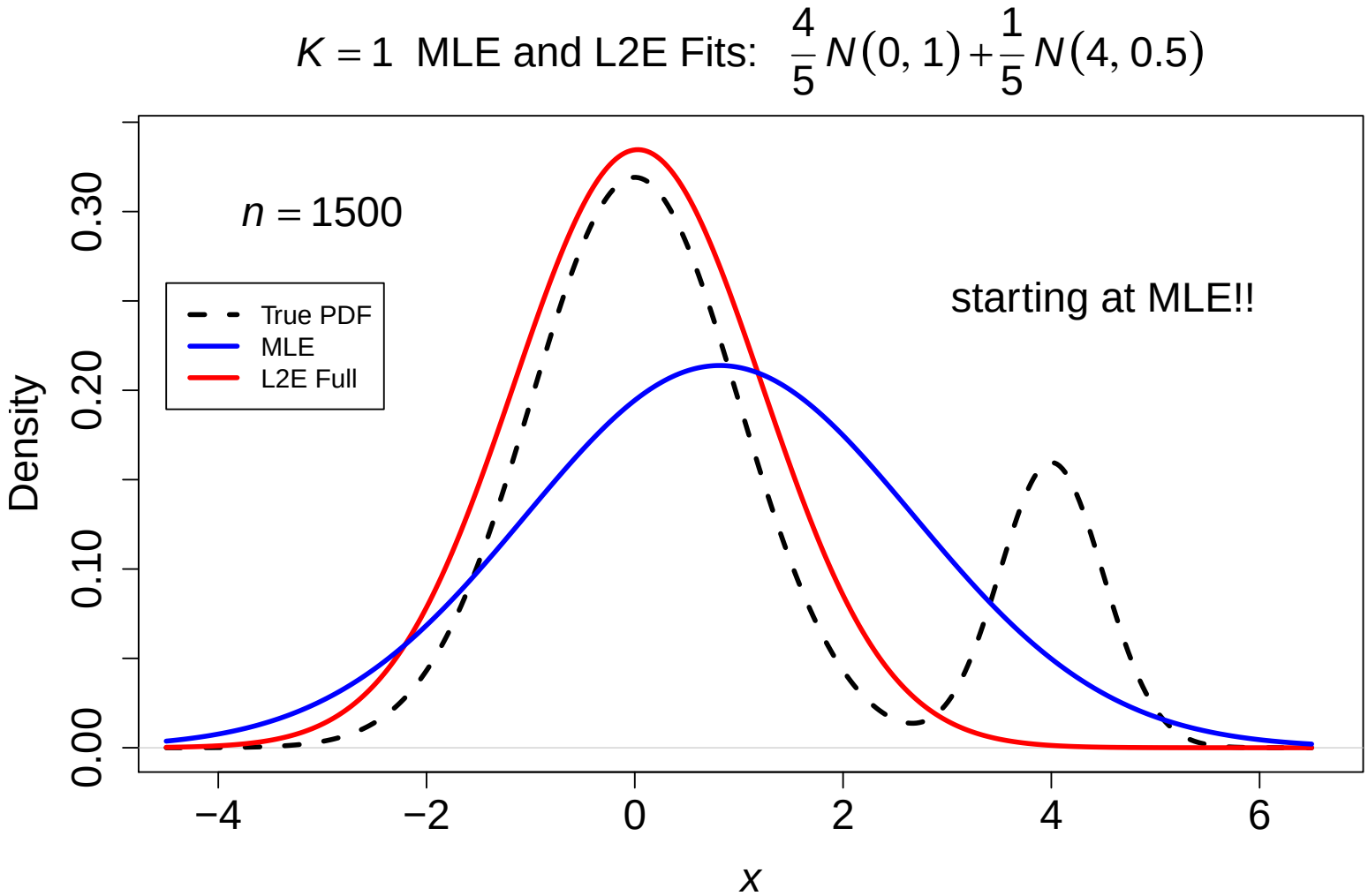
Introductory Thoughts on Gaussian Mixtures

- Use of Gaussian mixtures to find clusters has a rich history
- The idea of including extra components to find outliers is a clever idea due to Aitkin & Wilson (1980)
- However, the initialization problem is paramount
- The EM algorithm is sure but slow
- Rice PhD Hathaway (1986) showed how EM does not usually go to infinite likelihood even if the parameterization of the covariance matrix allows for such:

  Hathaway, R. J. (1986). Another interpretation of the EM algorithm for mixture distributions. Statistics & probability letters, 4(2), 53-56.
- The constraints $w_k \geq 0$ and $\sum_{k=1}^{K} = 1$ are active with MLE. Relaxing the constraints results in useless estimates.
- With L2E, those constraints may be enforced; in particular, relaxing the sum constraint while retaining the positivity constraints is often quite illuminating when the mixture model is not perfect (or its initialization)

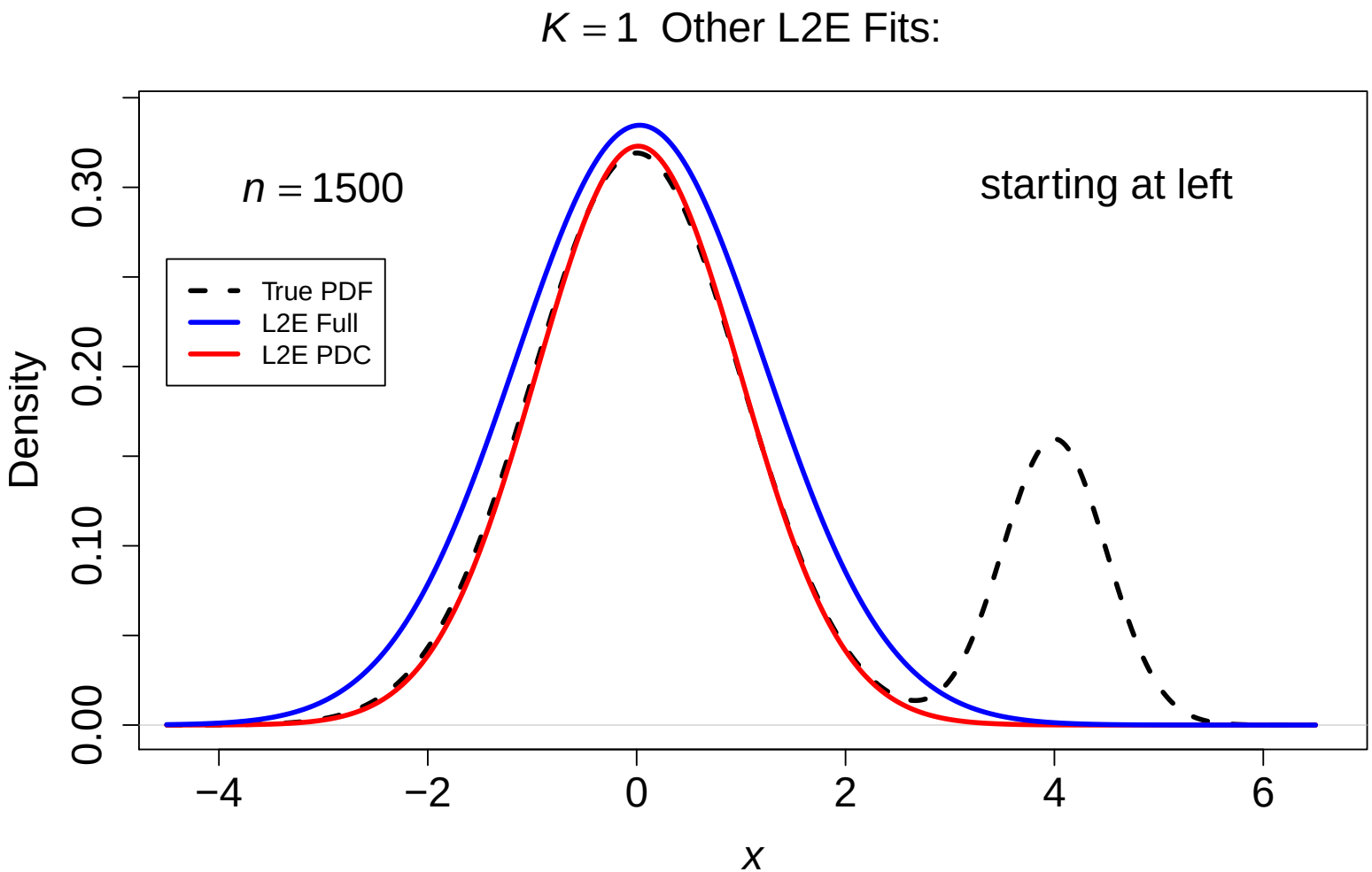
Understanding Full and Partial Gaussian Mixtures

- True $K = 2$ mixture model: $0.8 \cdot N(0, 1) + 0.2 \cdot N(4, .5)$



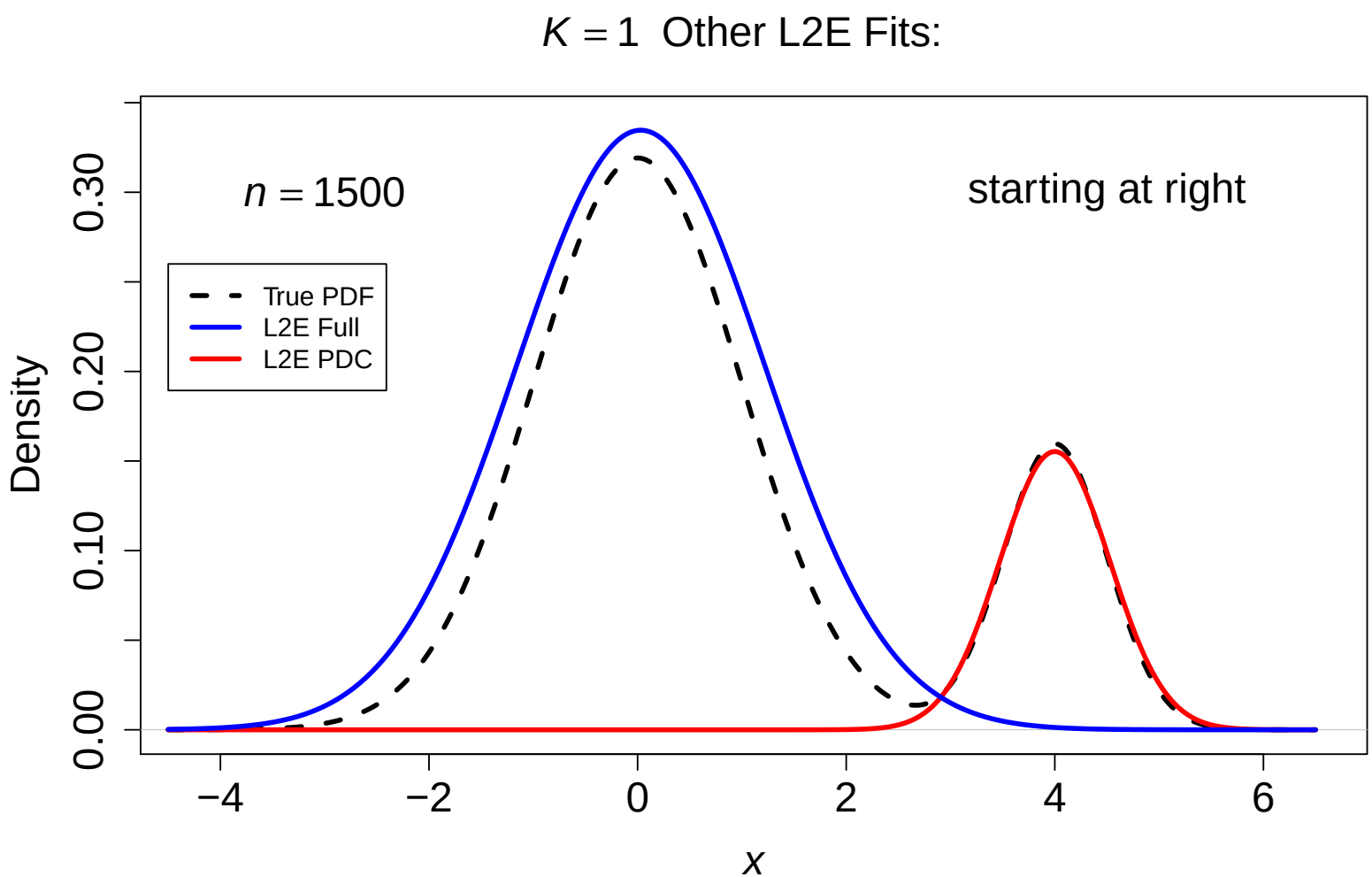
Understanding Full and Partial Gaussian Mixtures

- L2E fits with $w = 1$ enforced, and with w a "free" parameter:



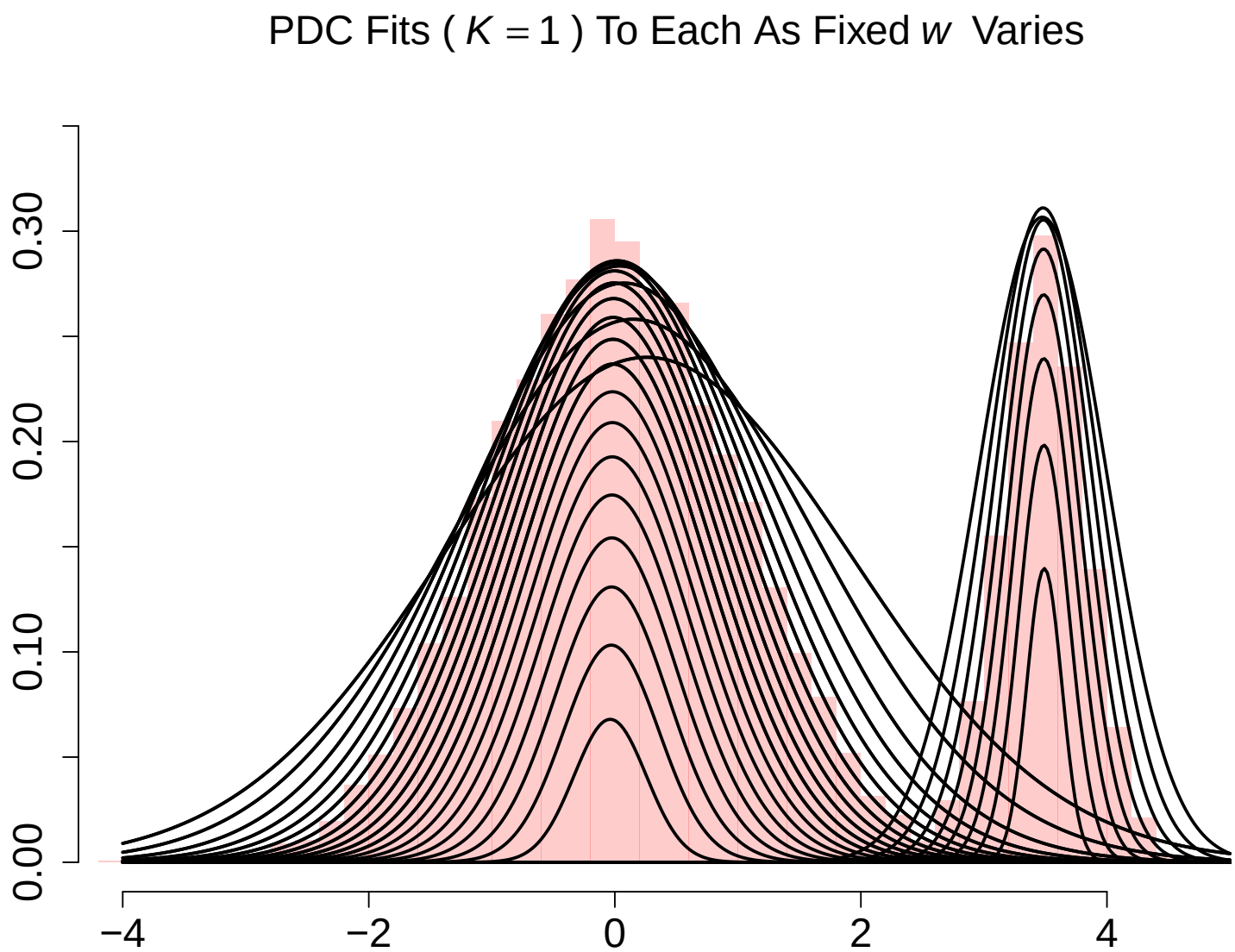
Understanding Full and Partial Gaussian Mixtures

- Note that the small right component is a local attractor to L2E and PDC (observe that $\hat{w} < 1/2$)



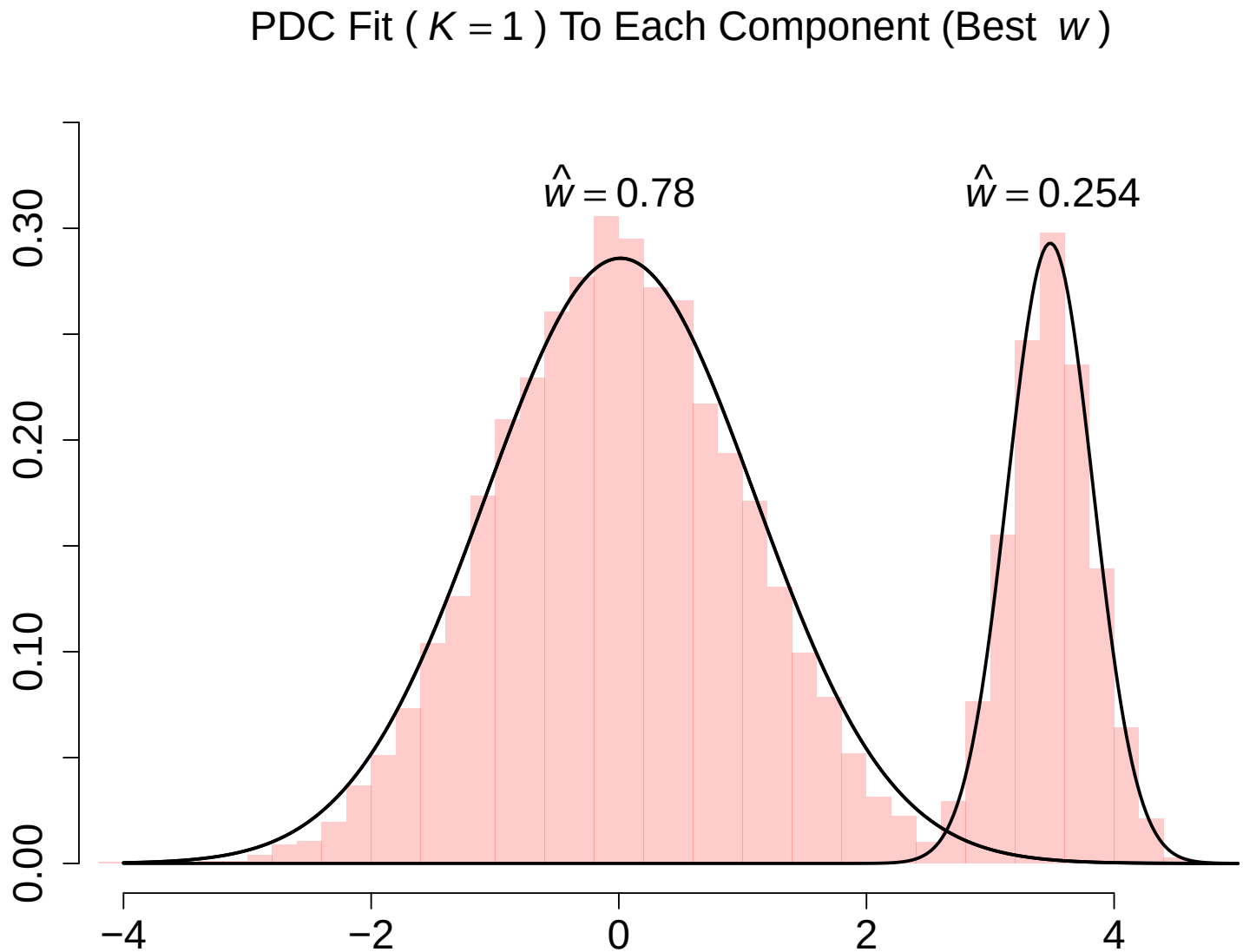
Understanding Full and Partial Gaussian Mixtures

- Fitting $K = 1$ w/ L2E to same data as w ranges from .05 – 1



Understanding Full and Partial Gaussian Mixtures

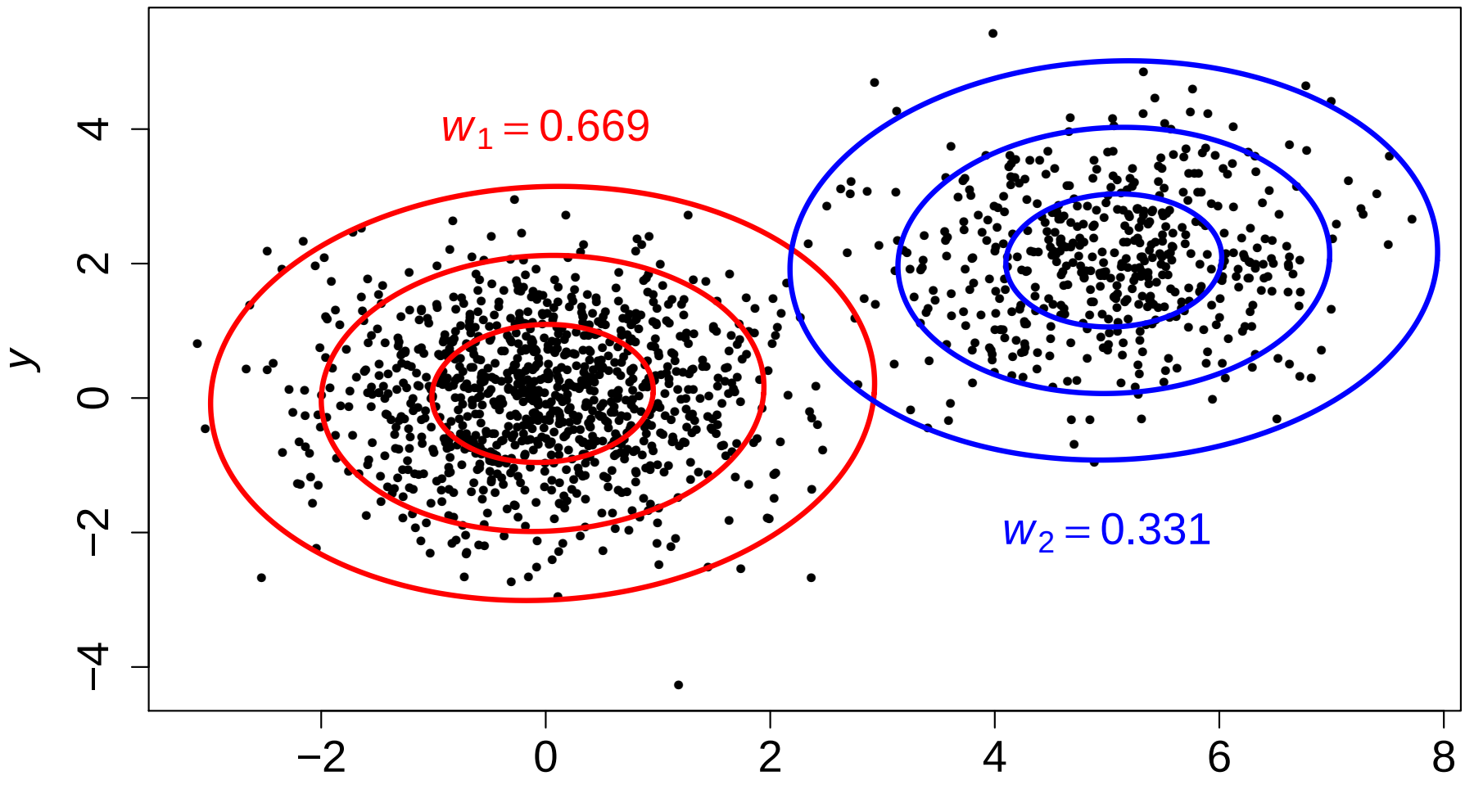
- Fitting $K = 1$ w/ L2E to same data with w a parameter



Understanding Full and Partial Gaussian Mixtures $\Re^d$

- Fitting $K = 2$ mixture to a true $K = 2$ normal mixture

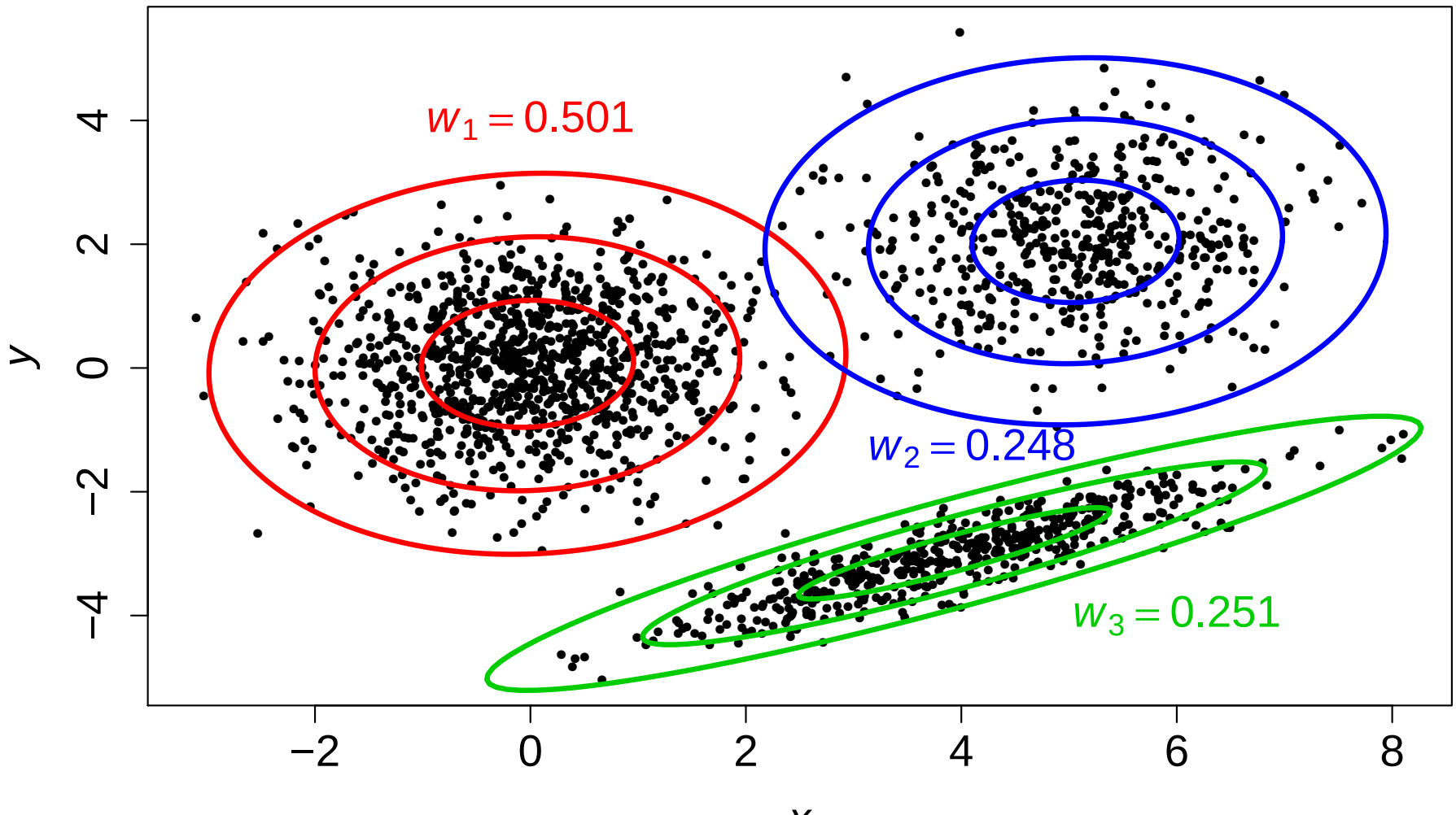
$$f_\theta(x) = \frac{2}{3}\phi(\,(0,0),\, I_2) + \frac{1}{3}\phi(\,(5,2),\, I_2)$$



Understanding Full and Partial Gaussian Mixtures $\Re^d$

- Fitting $K = 3$ mixture to a true $K = 3$ normal mixture

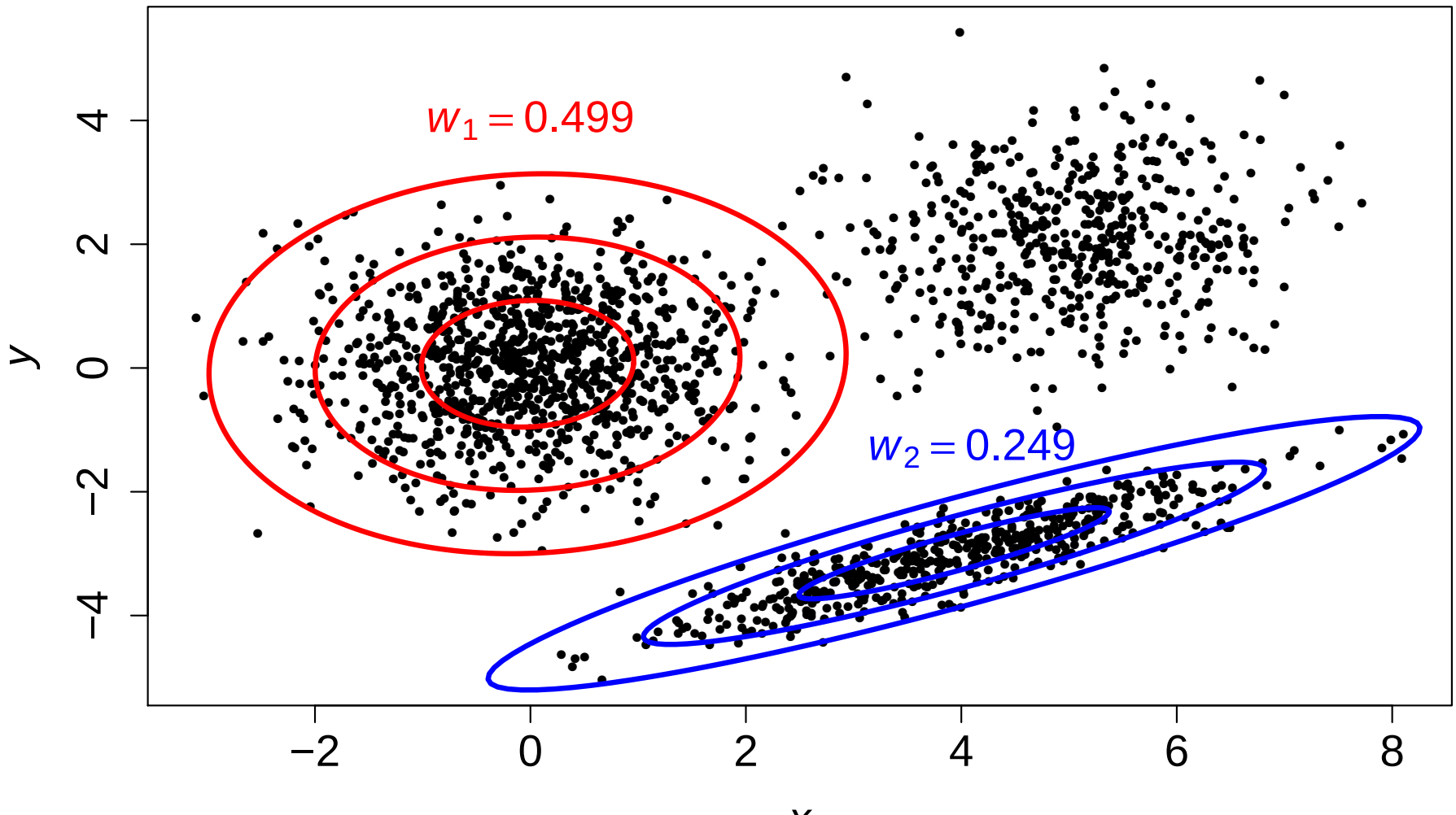
$$f_\theta(x) = \frac{1}{2}\phi(\,(0,0),\, I_2) + \frac{1}{4}\phi(\,(5,2),\, I_2) + \frac{1}{4}\phi(\,(4,-3),\, \Sigma_2)$$



Understanding Full and Partial Gaussian Mixtures $\Re^d$

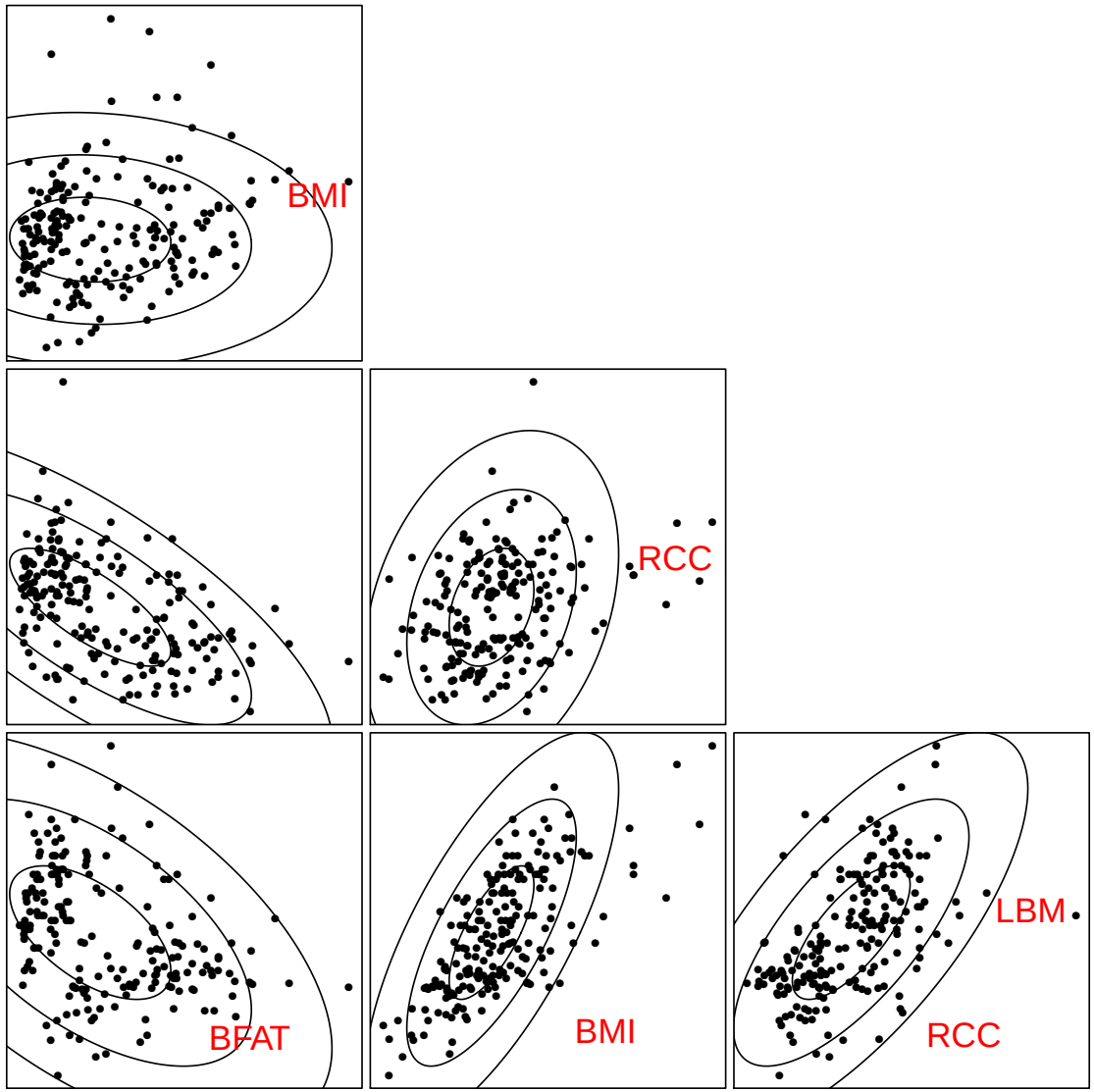
- Fitting $K = 2$ mixture to a true $K = 3$ normal mixture

$$K = 2 \text{ fit to } f_\theta(x) = \frac{1}{2}\phi(\mu_1, I_2) + \frac{1}{4}\phi(\mu_2, I_2) + \frac{1}{4}\phi(\mu_3, \Sigma_2)$$



Application to Australian Athlete Data ($n = 202$)

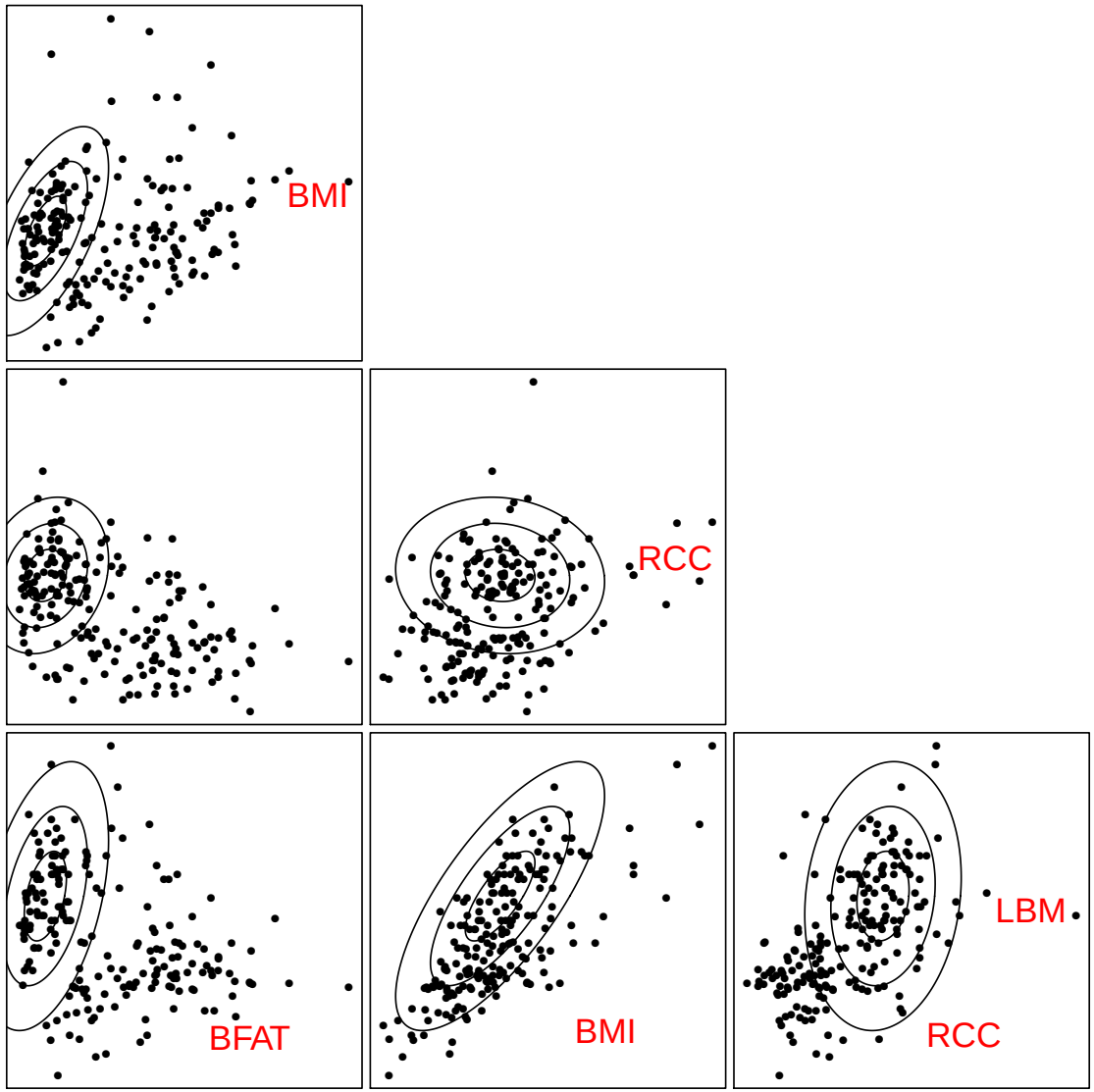
MLE fit to $x \in \Re^4$: body mass index (BMI), red cell count (RCC), body fat (BFAT), lean body mass (LBM)



Application to Australian Athlete Data ($n = 202$)

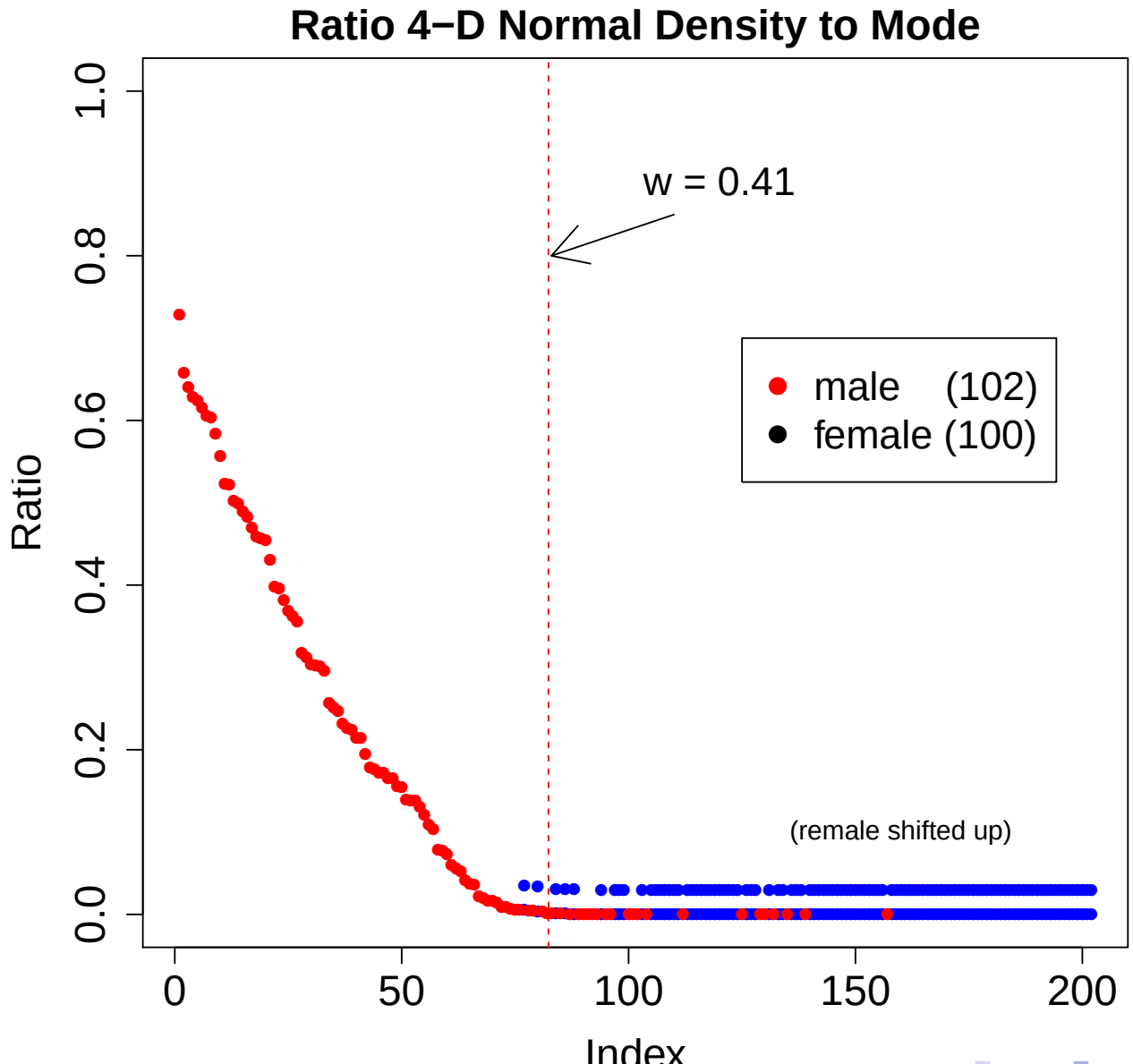
Partial L2E Fit With Initialization at MLE for 102 Male Athletes

— $w\, N(\mu_4, \Sigma_4)$, where $\hat{w} = 0.41$



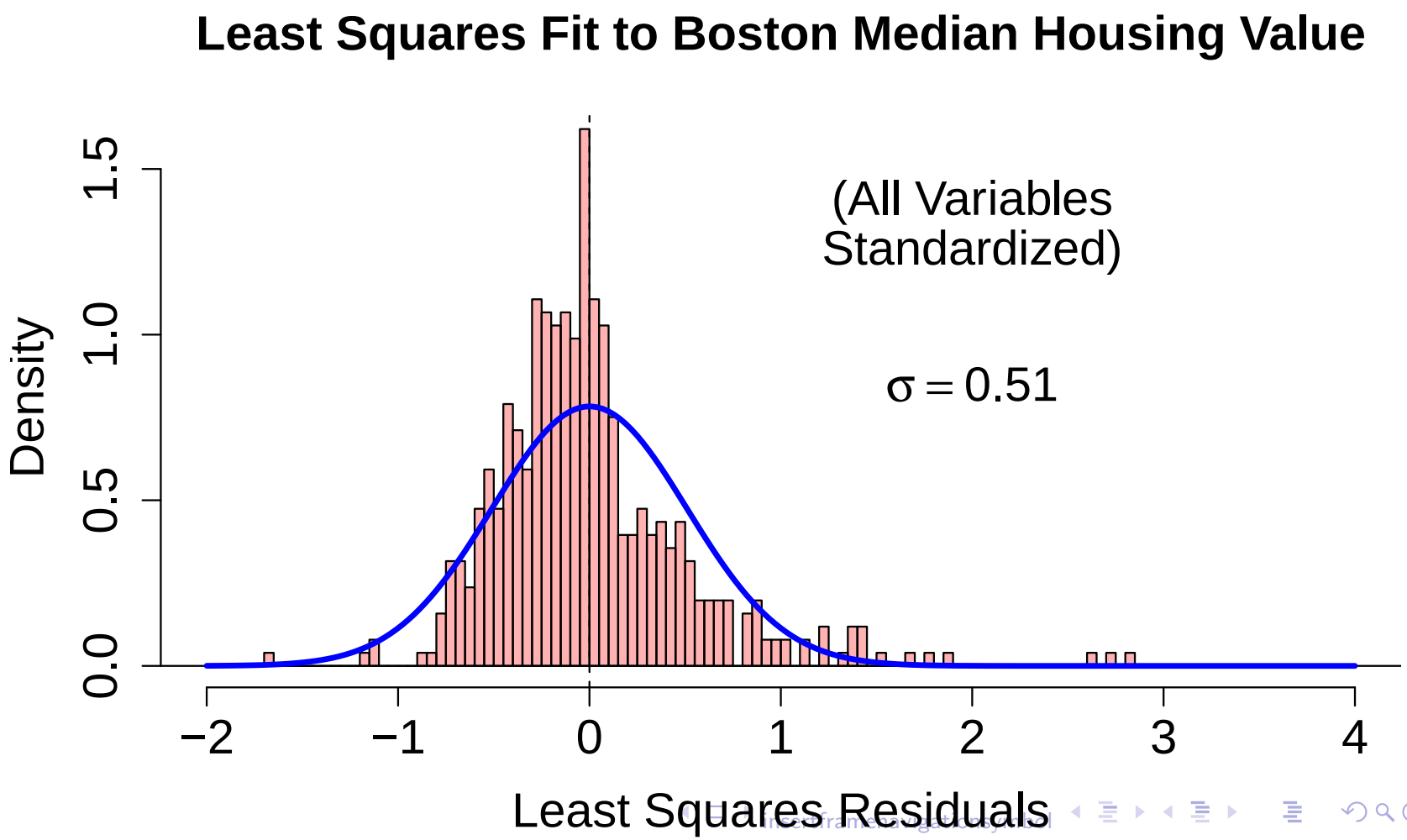
Application to Australian Athlete Data ($n = 202$)

Sorted model density fits (height) using L2E PDC Fit $w\phi(x|\mu_4, \Sigma_4)$



Back to Regression: Boston Housing Data

- How hard is diagnosing a linear model from the residuals?
- Consider the (standardized) Boston Housing data, where y is the median home value, and $\mathbf{x} \in \Re^{13}$ is a vector of 13 predictors.



L2E Regression (Using R Function nlminb)

Usual model:

$$Y = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \epsilon\,, \qquad \epsilon \sim N(0, \sigma_\epsilon^2)$$

L2E fit driven by assumption of parametric form of error distribution:

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}_\epsilon) = \arg\min \left[\frac{1}{2\sqrt{\pi}\sigma_\epsilon} - \frac{2}{n} \sum_{i=1}^{n} \phi(\hat{\epsilon}_i | 0, \sigma_\epsilon^2) \right]$$

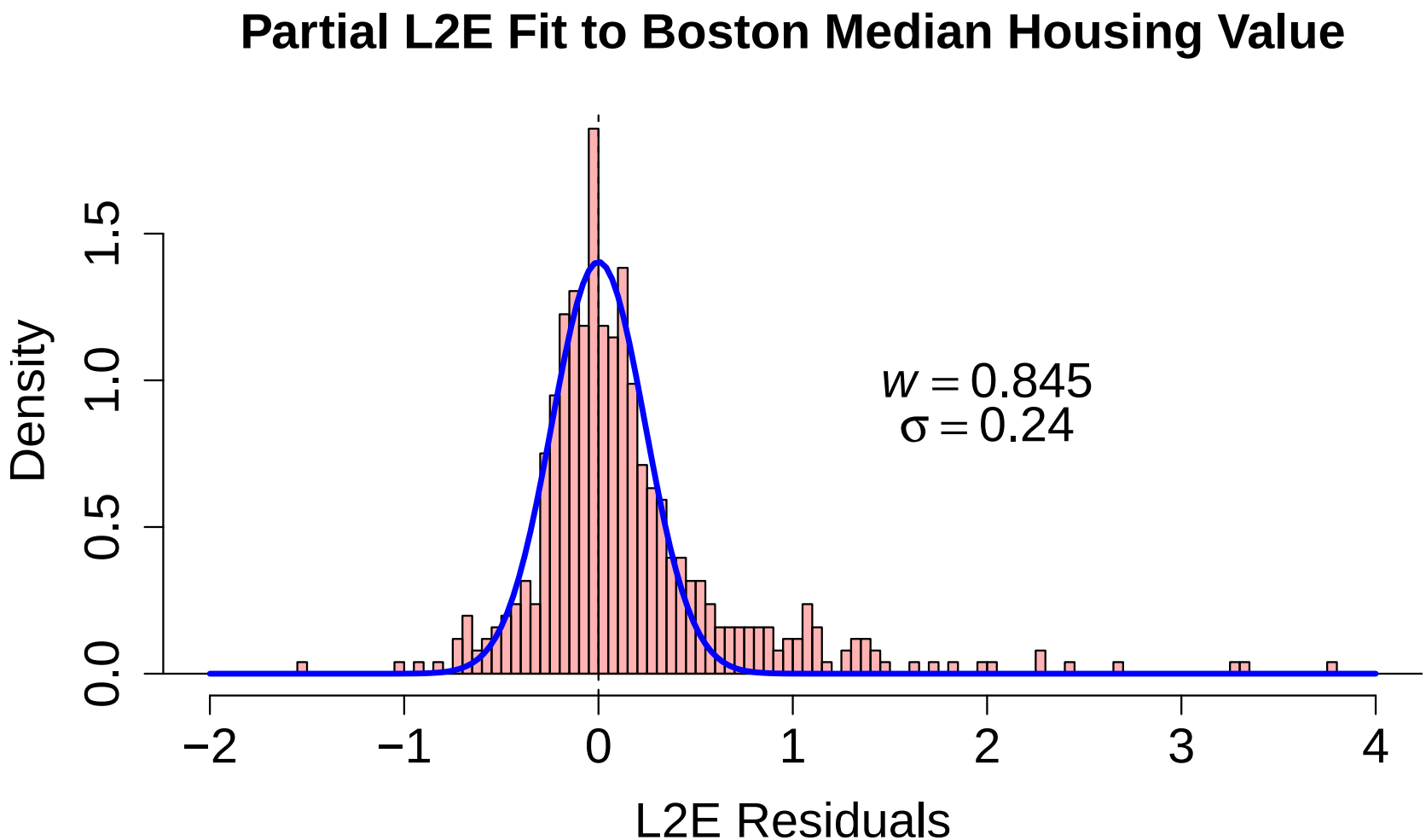
To check for group(s) of outliers, can use model

$$\epsilon \sim w \cdot N(0, \sigma_\epsilon^2)$$

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}_\epsilon, \hat{w}) = \arg\min \left[\frac{w^2}{2\sqrt{\pi}\sigma_\epsilon} - \frac{2w}{n} \sum_{i=1}^{n} \phi(\hat{\epsilon}_i | 0, \sigma_\epsilon^2) \right]$$

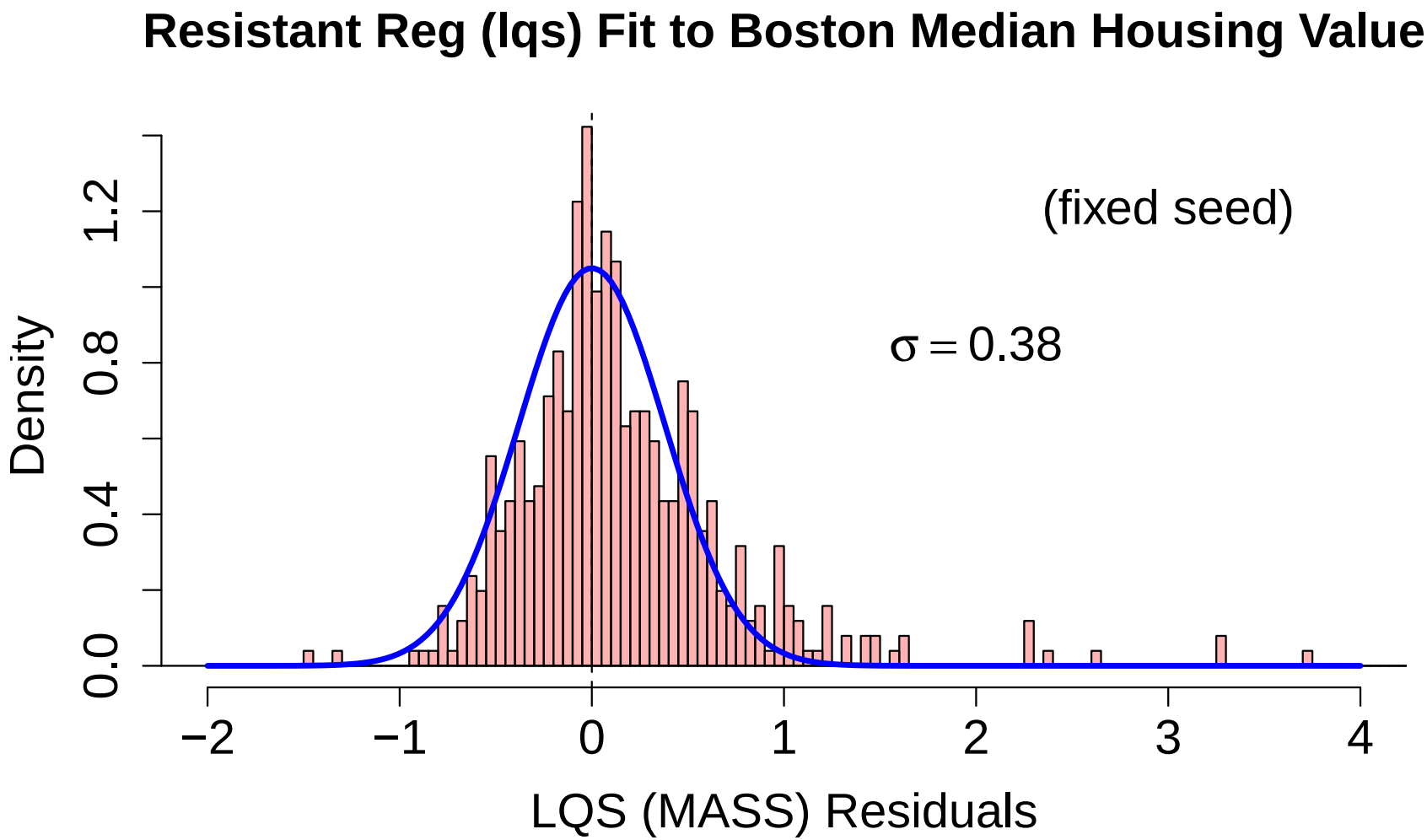
Back to Regression: Boston Housing Data

- Well, the residuals do not seem to follow the normal pdf. A few outliers?
- Now consider the residuals from a $w \cdot N(0, \sigma^2)$ L2E fit.



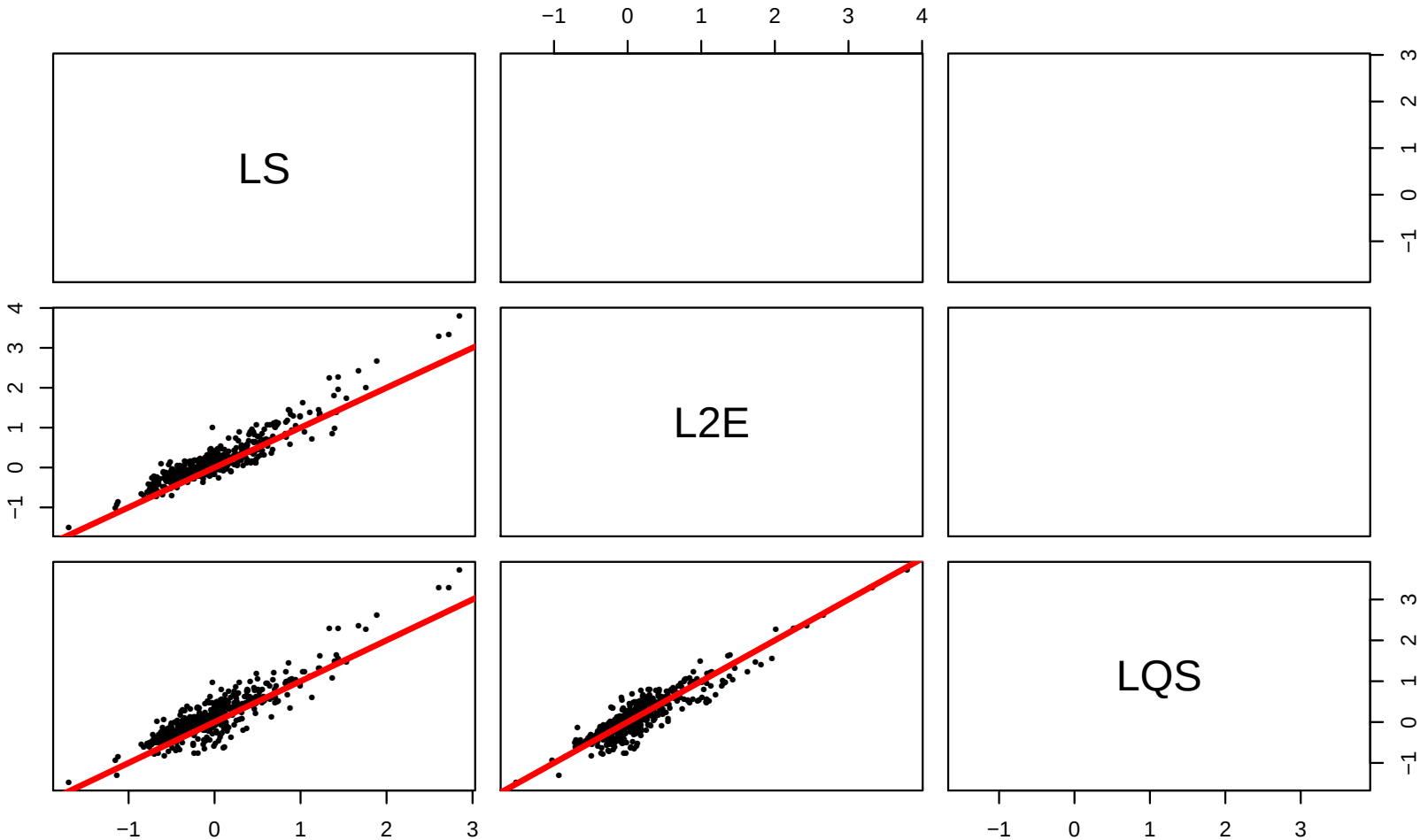
Resistant Regression (Function lqs in MASS Library)

- Is forcing residuals to look "as normal as possible" effective?



How Similar are the Residuals?

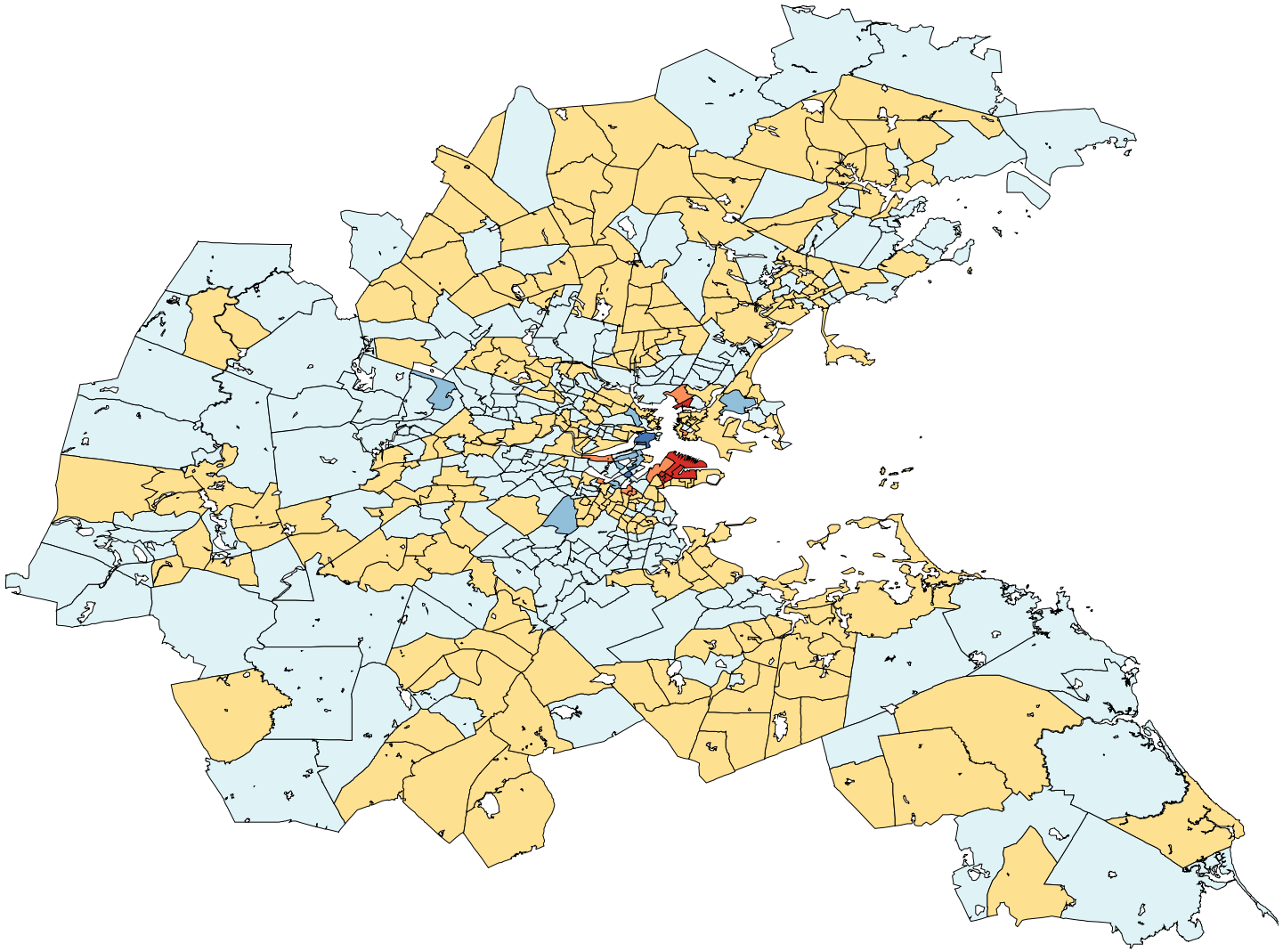
- Is forcing residuals to look "as normal as possible" effective?



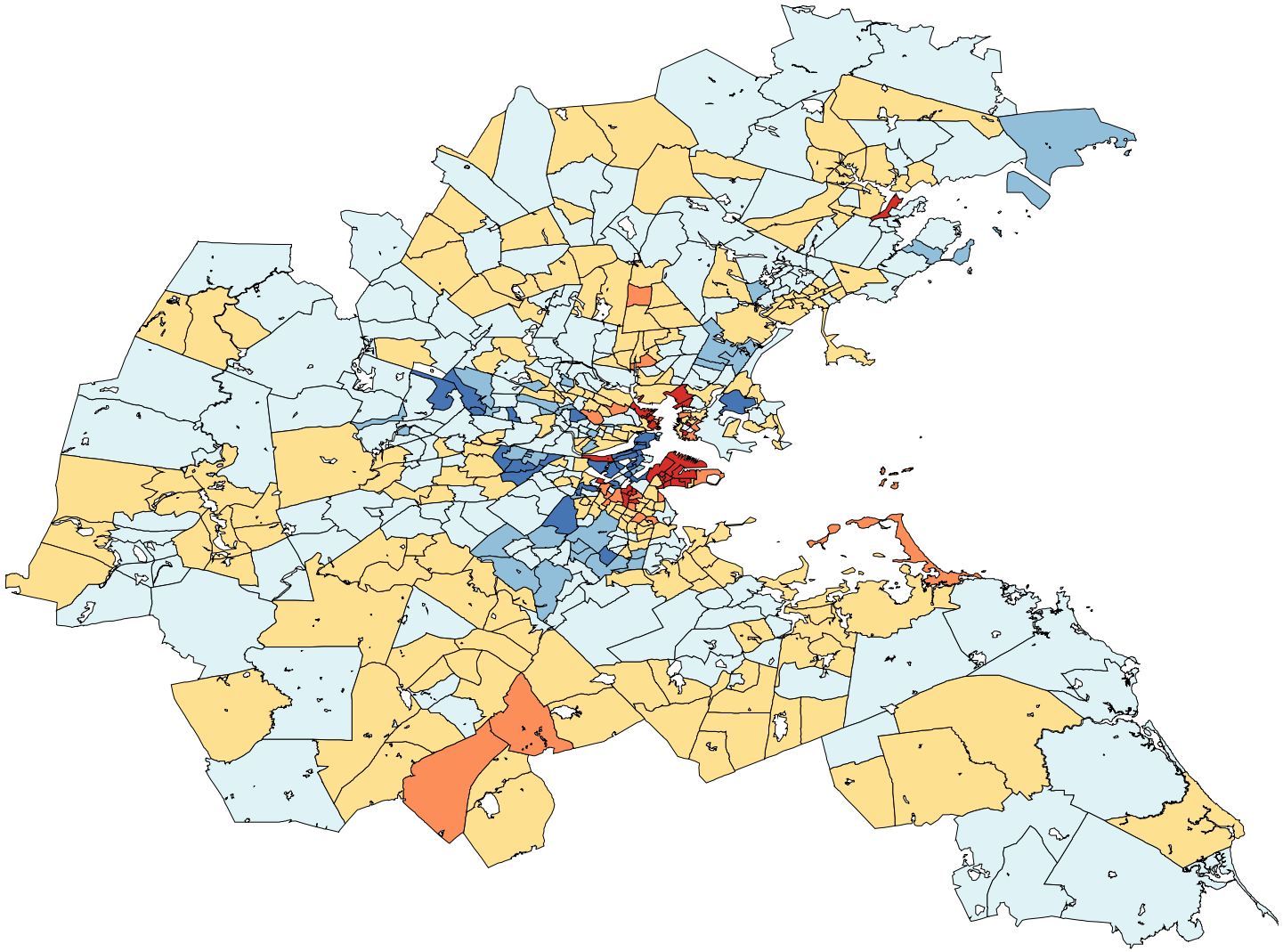
Comparing Model Coefficients

Var	LS	L2E	LQS
Intercept	0.000	-0.155	-0.156
CRIM	-0.101	-0.140	-0.131
ZN	0.118	0.078	0.241
INDUS	0.015	0.015	0.115
CHAS	0.074	0.048	0.040
NOX	-0.224	-0.061	-0.033
RM	0.291	0.394	0.398
AGE	0.002	-0.135	-0.043
DIS	-0.338	-0.177	-0.391
RAD	0.290	0.105	0.156
TAX	-0.226	-0.135	-0.461
PTRATIO	-0.224	-0.135	-0.125
B	0.092	0.133	0.133
LSTAT	-0.407	-0.180	-0.256

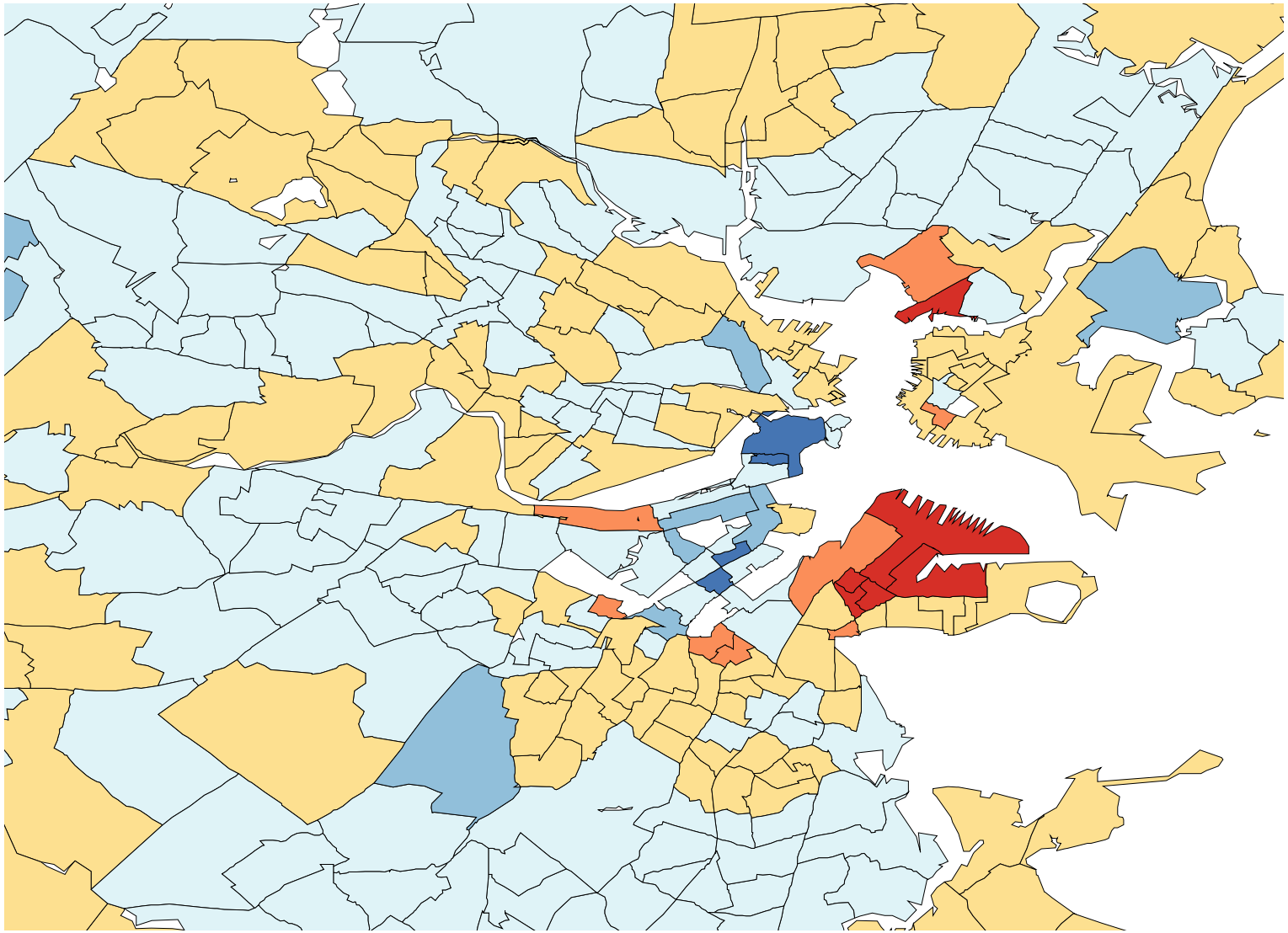
Comparing Residuals (LS Standardized)



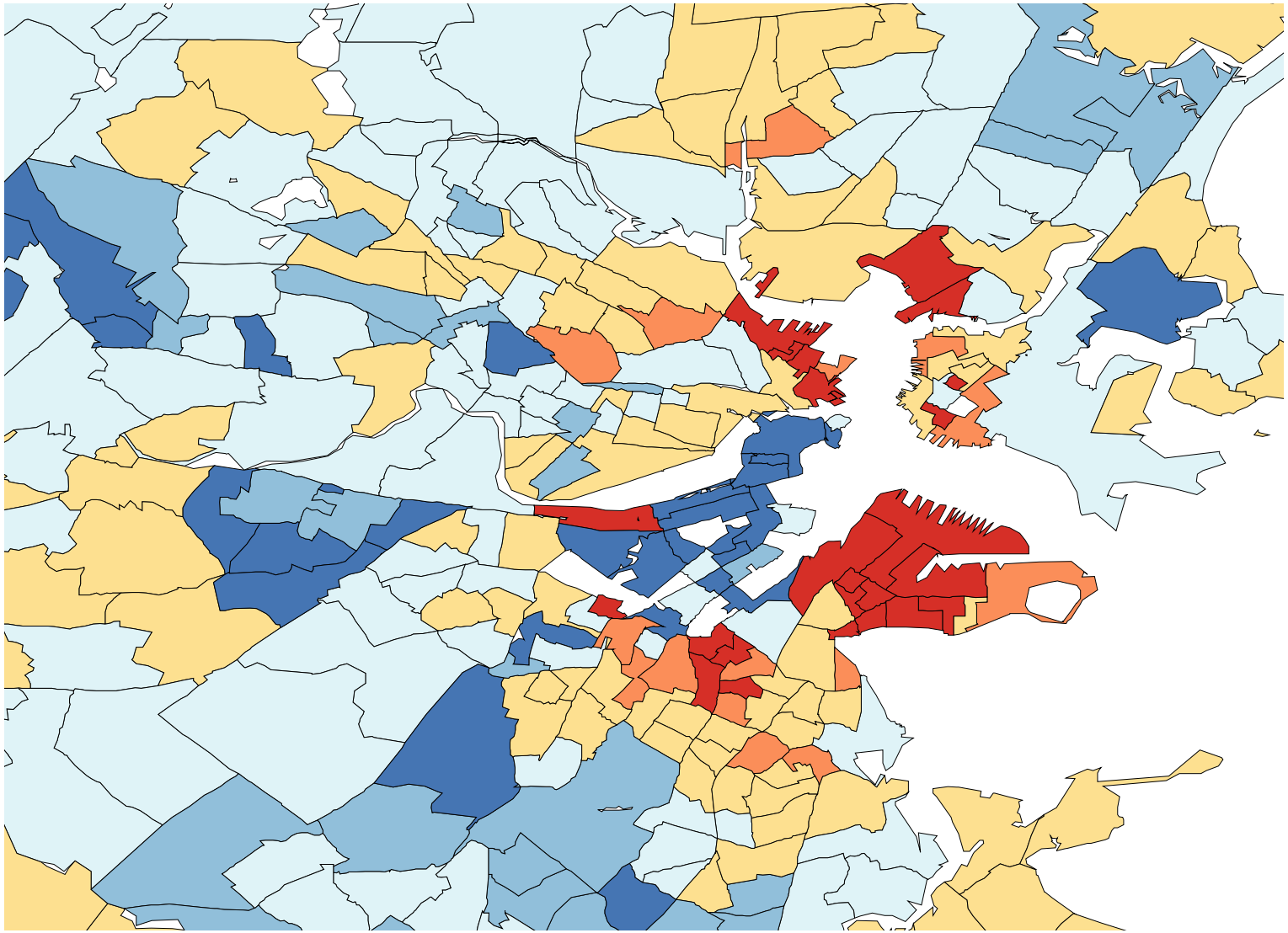
Comparing Residuals (L2E Standardized)



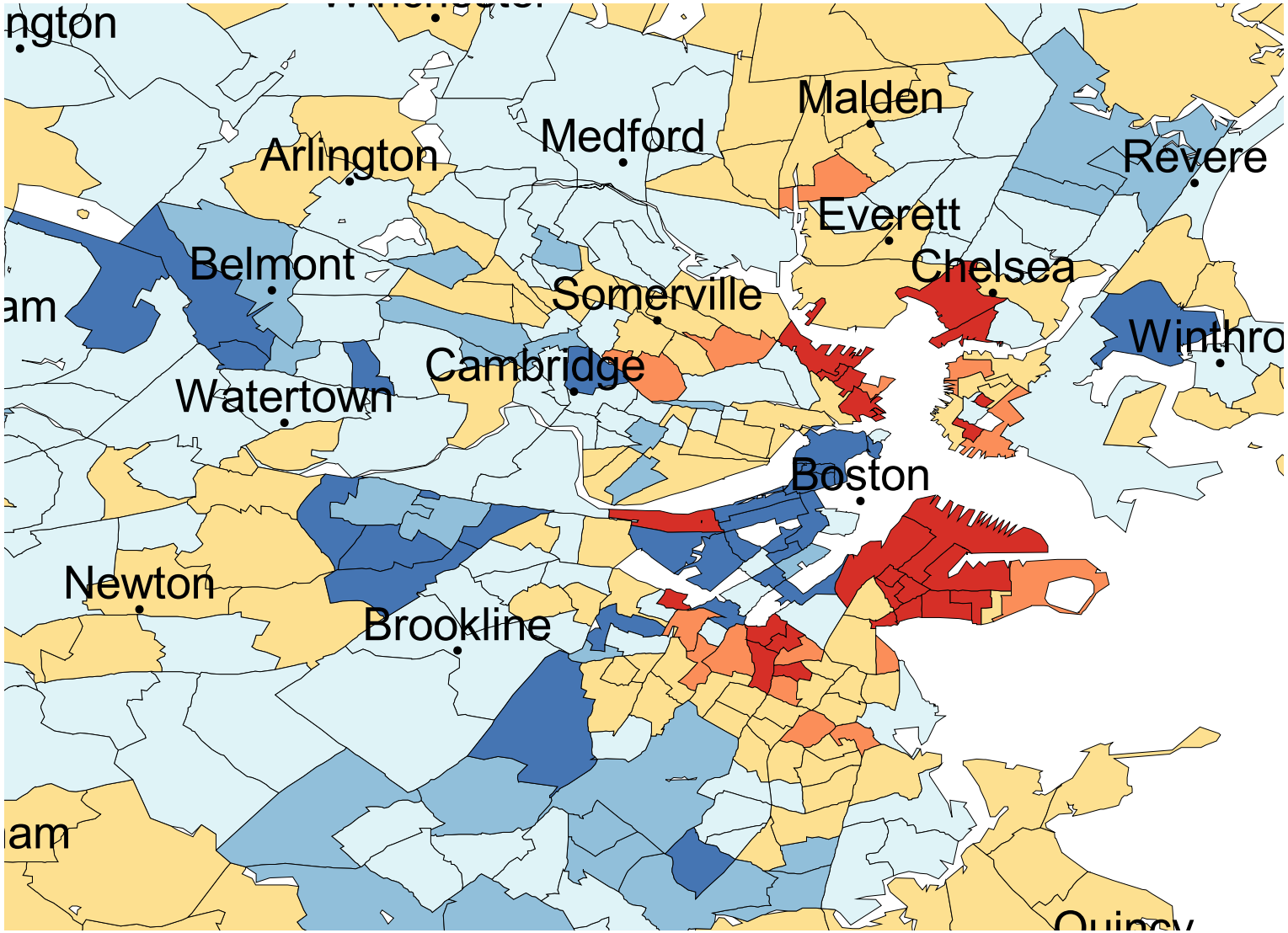
Comparing Residuals (LS Standardized Blow-Up)



Comparing Residuals (L2E Standardized Blow-Up)



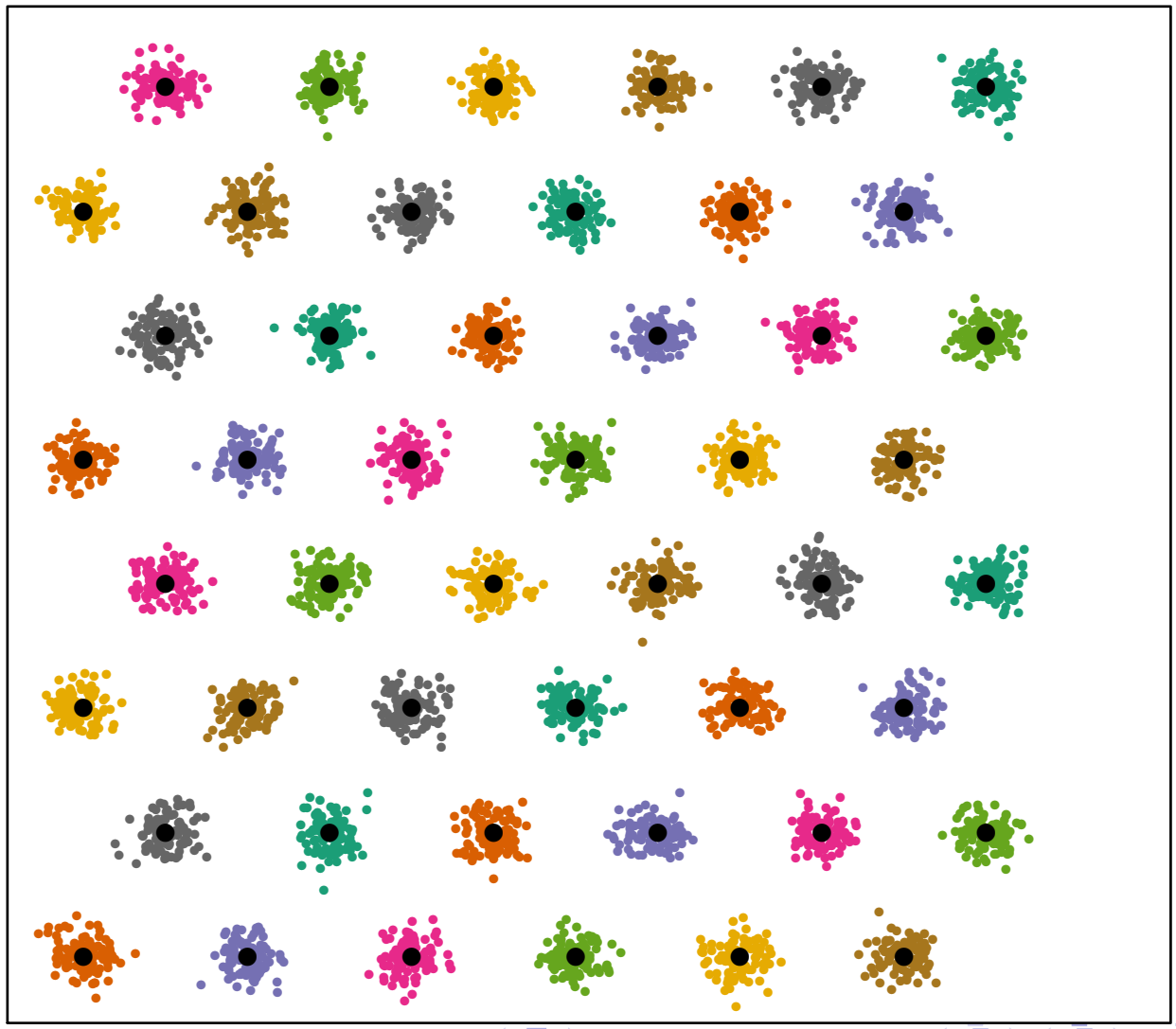
Comparing Residuals (L2E Labelled)



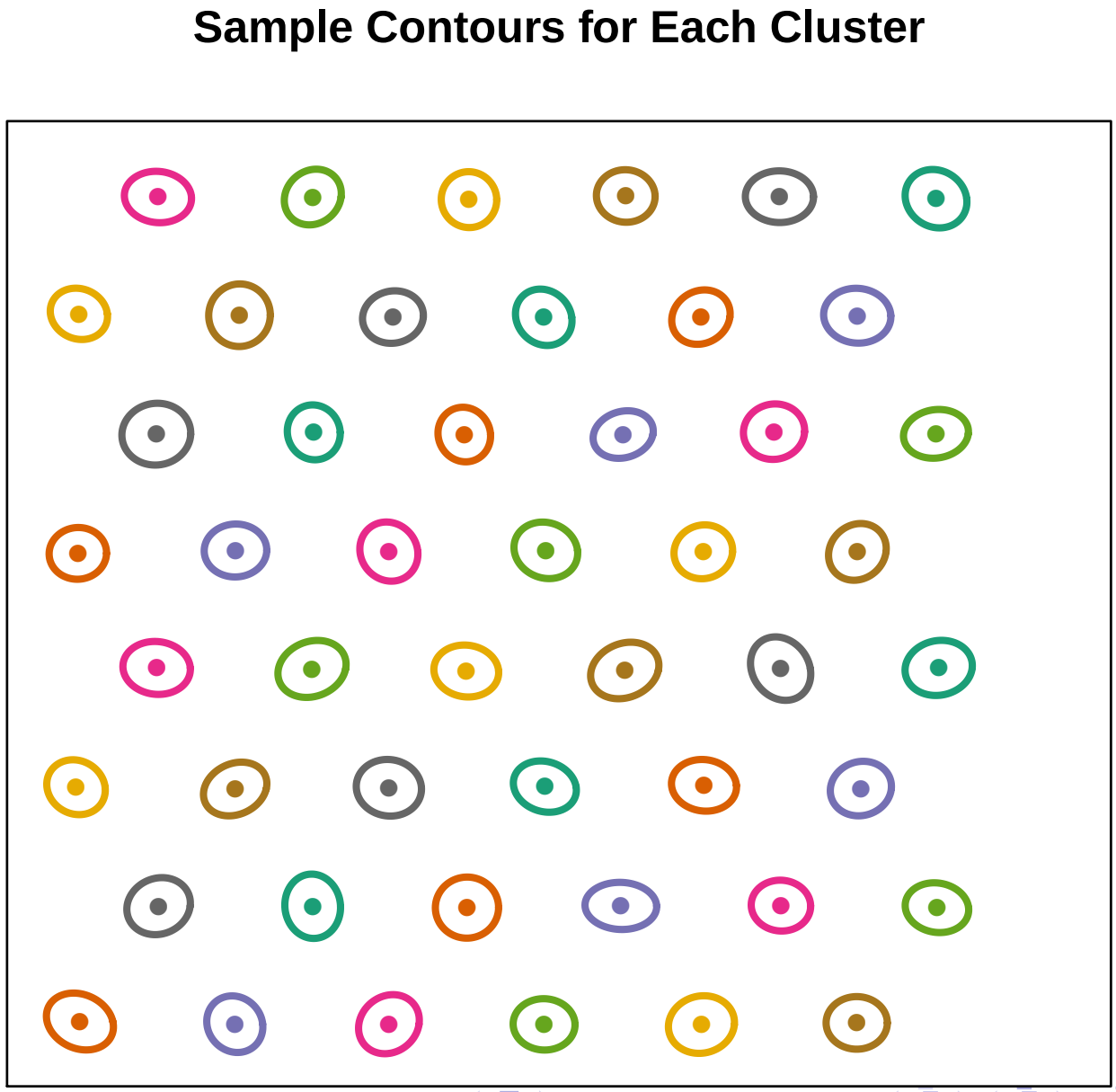
Probing for Clusters: A Small Study of Initialization

- Finding good initializations for mixture problems is said to an NP hard problem by computer scientists
- Here, we investigate in a simple problem in $\Re^2$ with $K = 48$
- We bias the solutions by giving it certain information
- The true cluster means are on a hexagonal grid 6×8
- The true cluster shape is circular
- The K-means algorithm favors clusters that are roughly circular, so we use this as a surrogate for fitting a normal mixture
- We give the R routine `kmeans` the true value of K
- Simultaneously, we "probe" the data using a singular PDC using L2E; we give the algorithm the true value of σ and let it find the point (μ_x, μ_y)

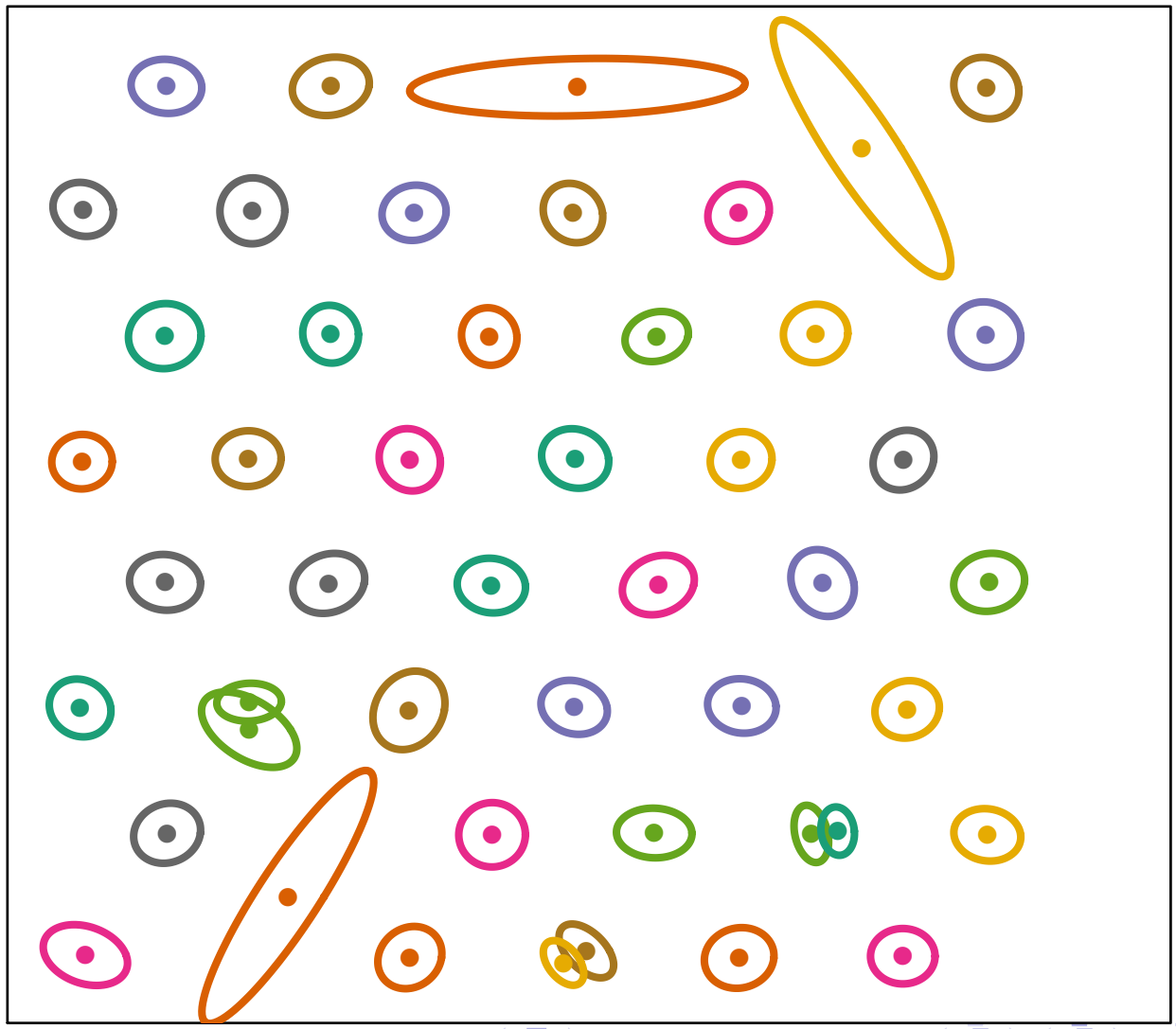
Example Run With $\sigma = 0.1$



Example Run With $\sigma = 0.1$

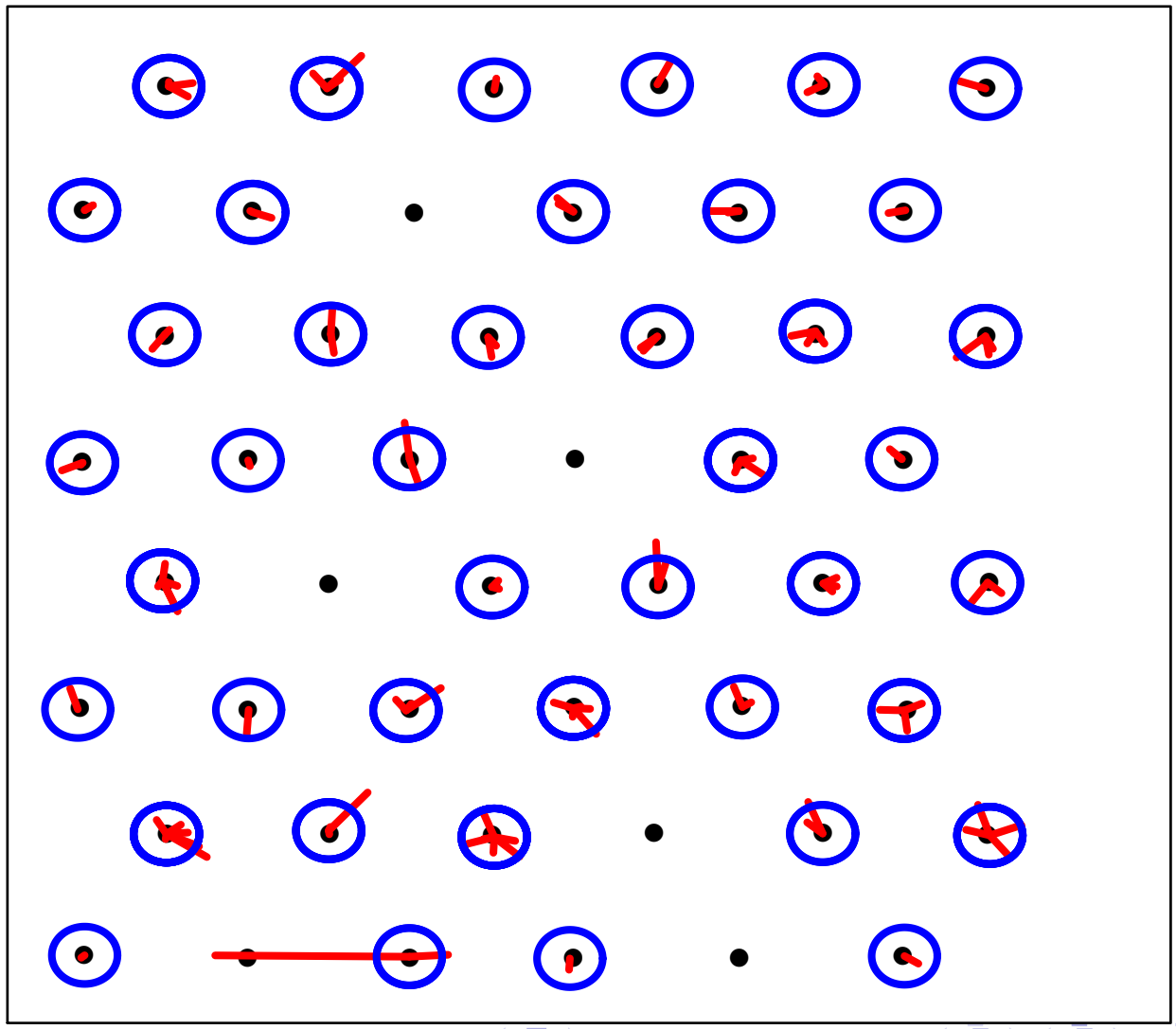


Example Run With $\sigma = 0.1$



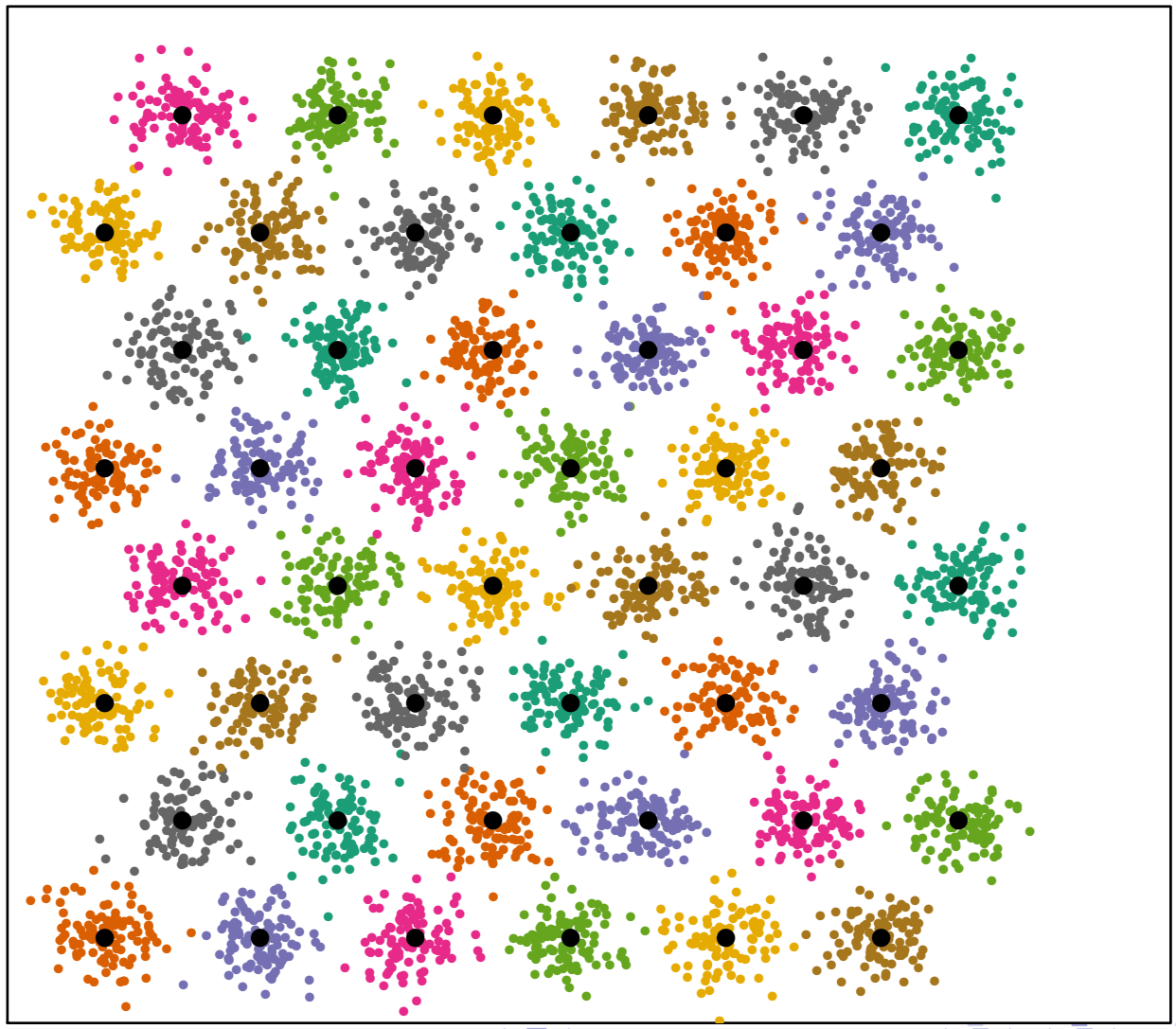
Example Run With $\sigma = 0.1$

Sample Contours for PDC with 100 Random Starts (σ known)

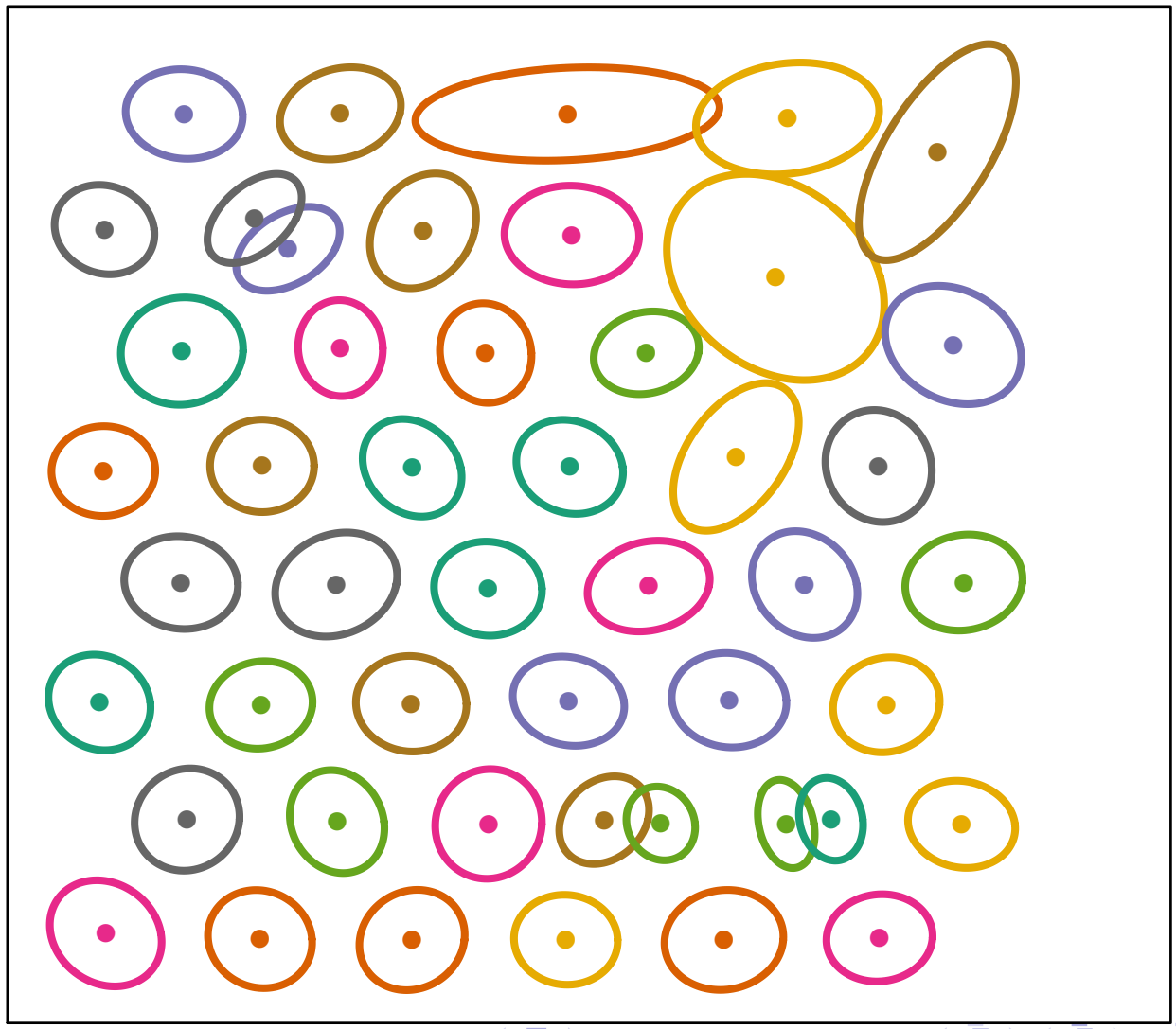


Example Run With $\sigma = 0.175$

Hexagonal 8x6 Cluster Pattern $n_k = 100$

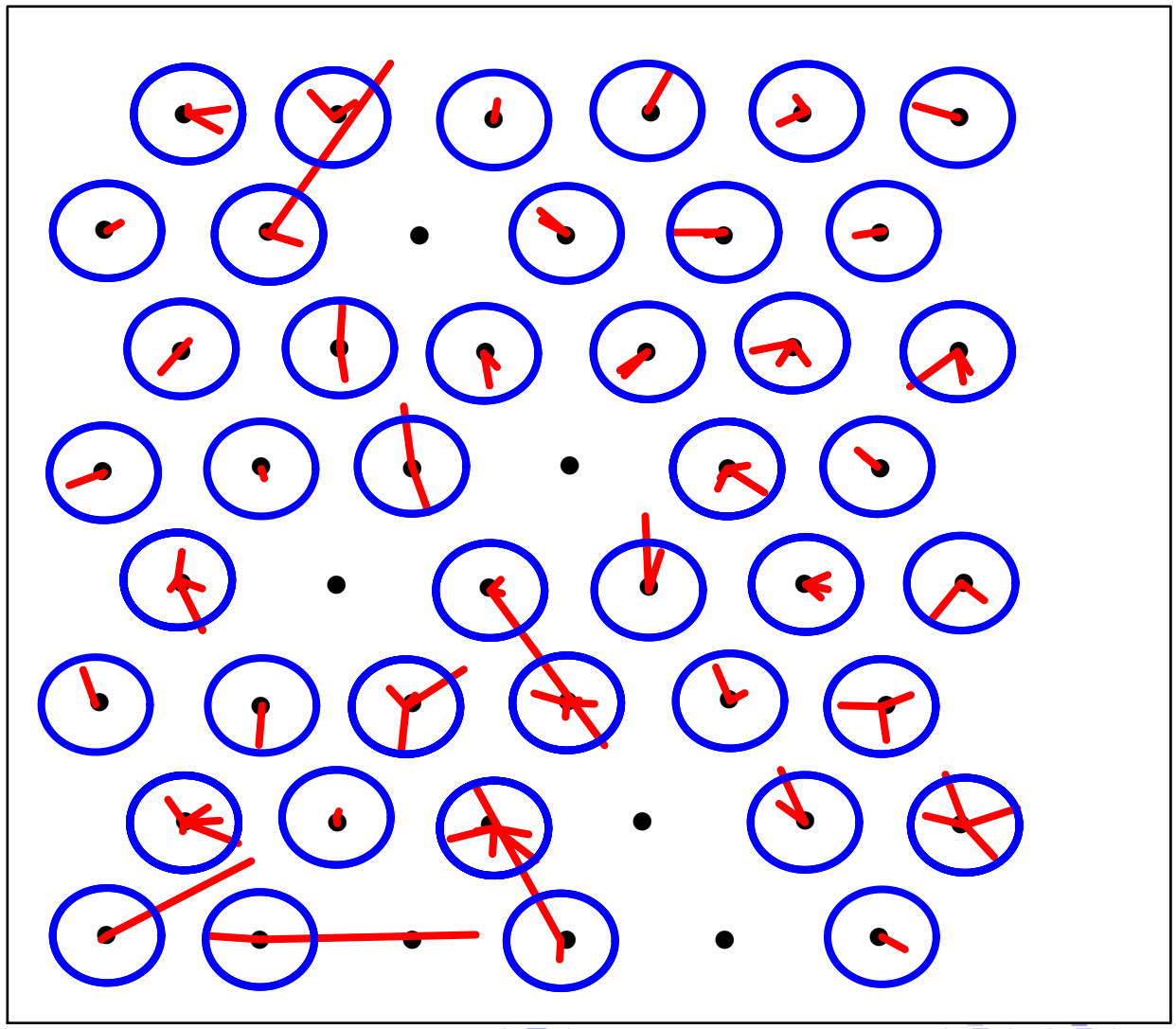


Example Run With $\sigma = 0.175$



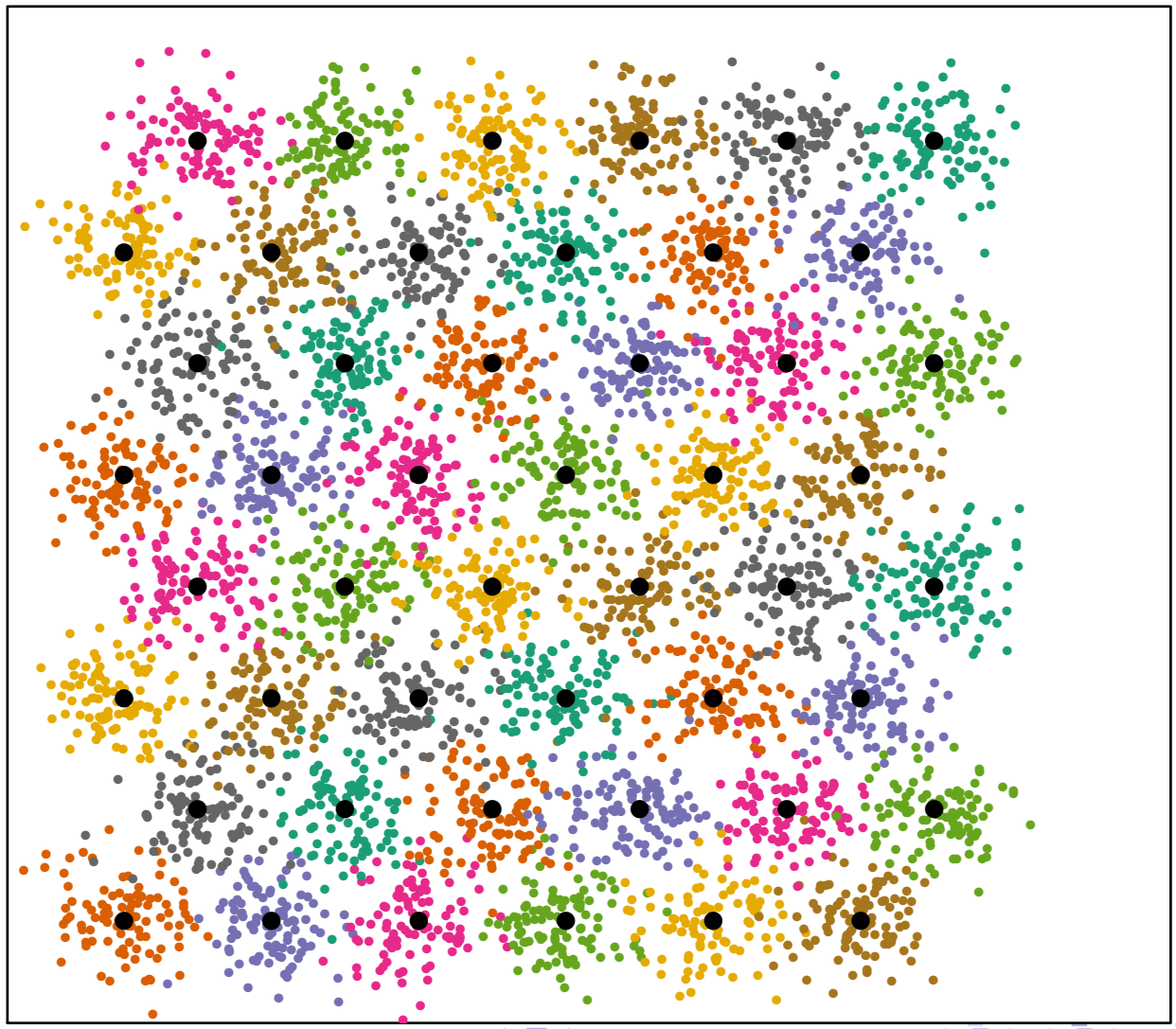
Example Run With $\sigma = 0.175$

Sample Contours for PDC with 100 Random Starts (σ known)

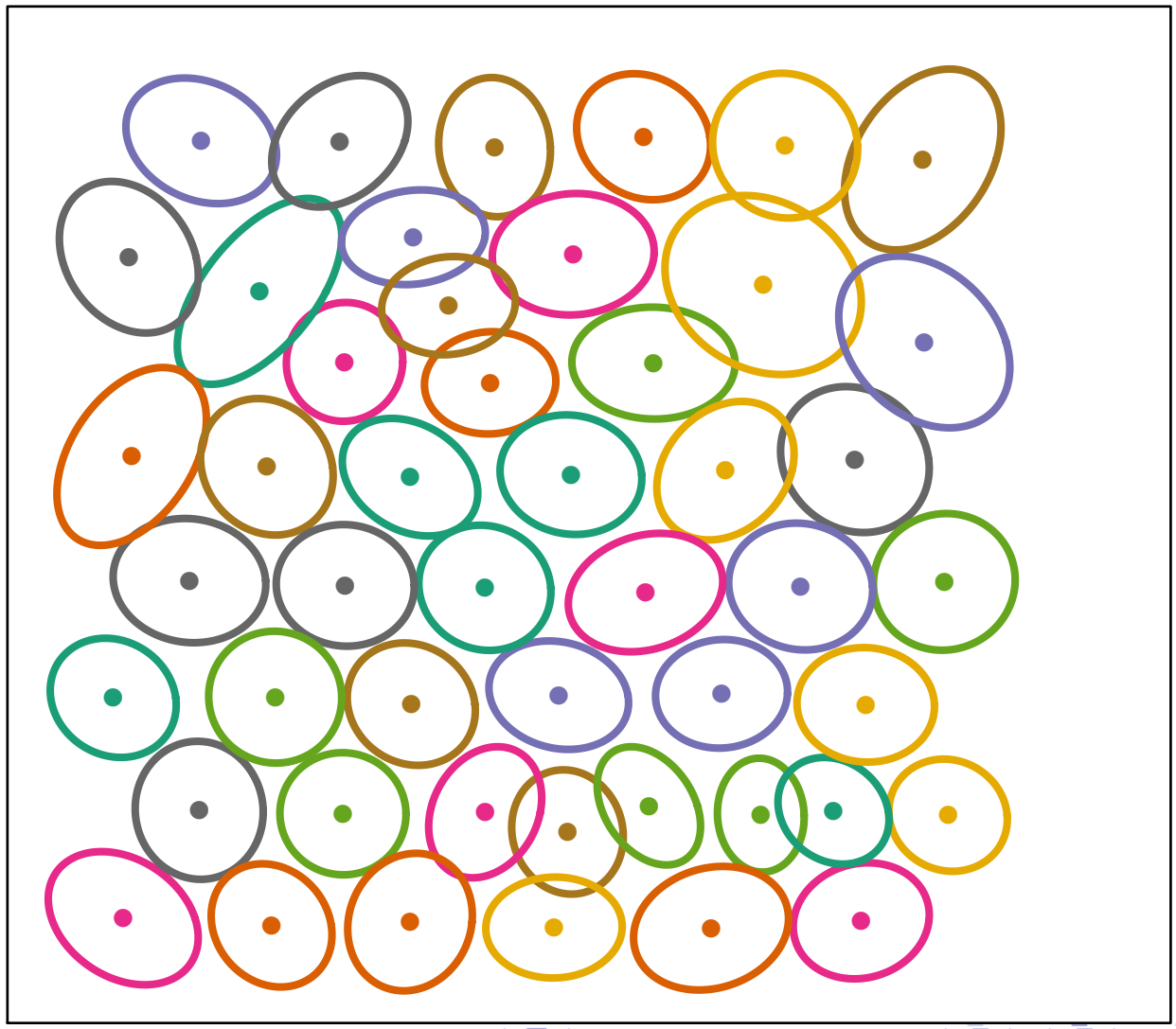


Example Run With $\sigma = 0.25$

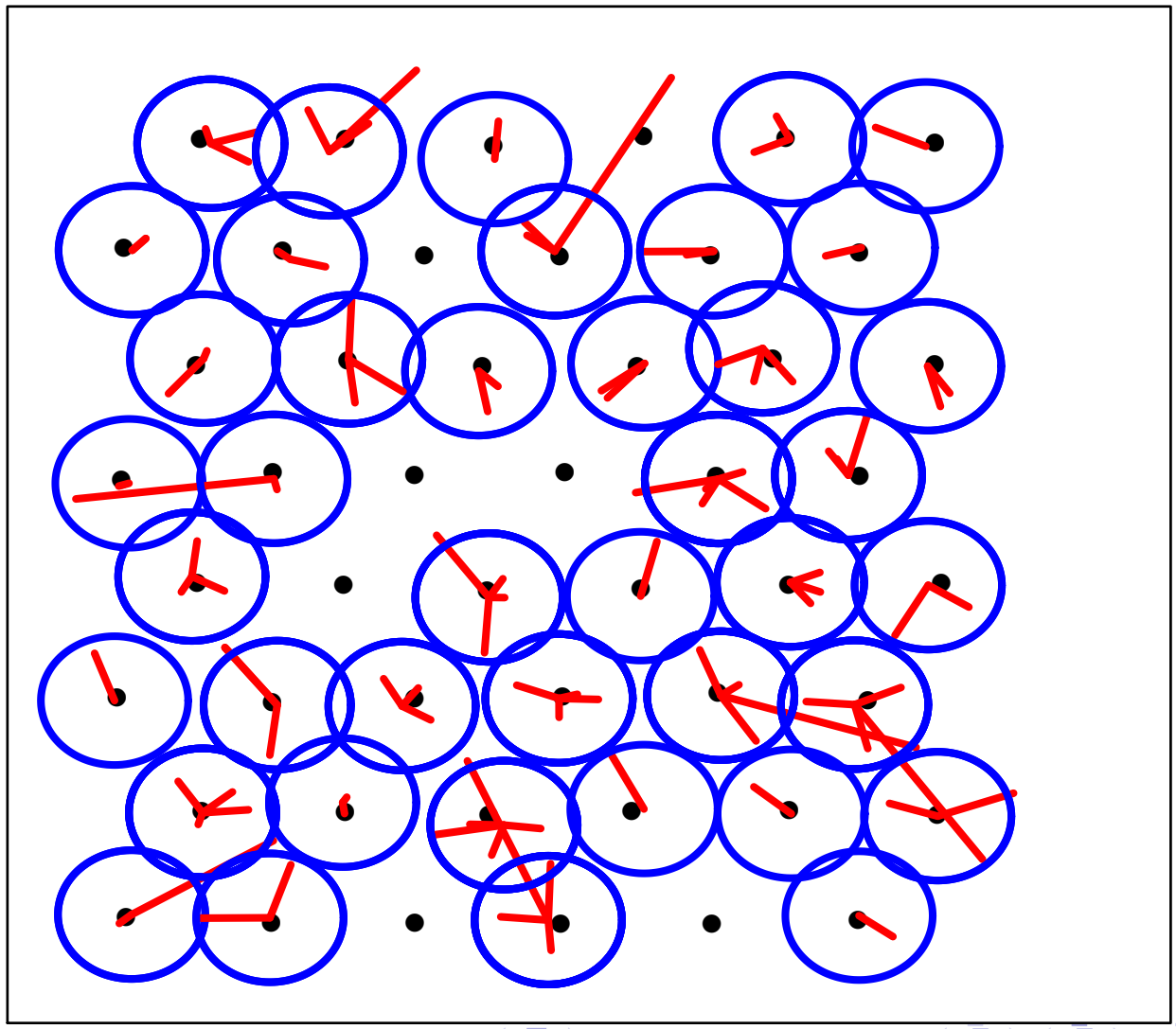
Hexagonal 8x6 Cluster Pattern $n_k = 100$



Example Run With $\sigma = 0.25$

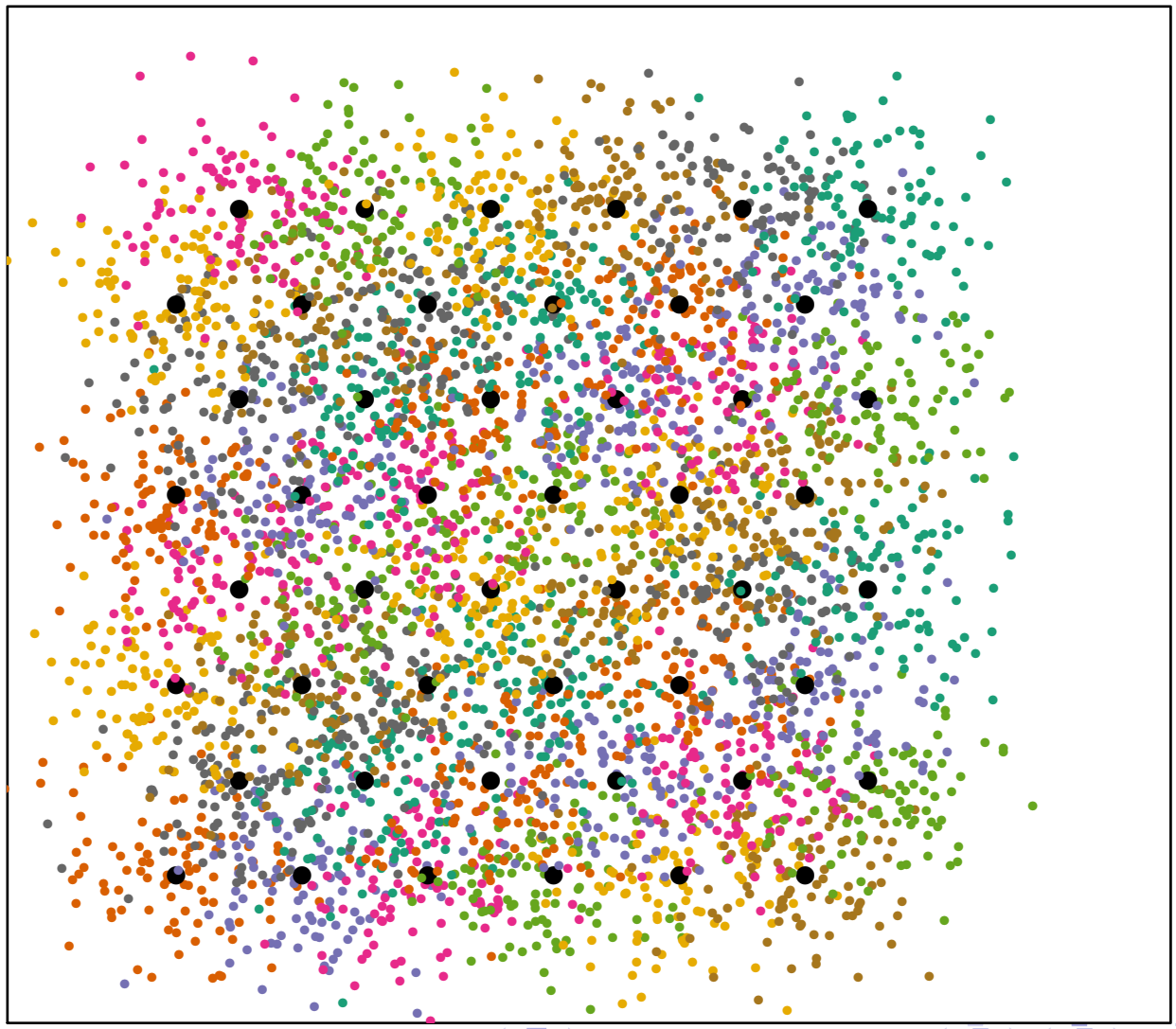


Example Run With $\sigma = 0.25$

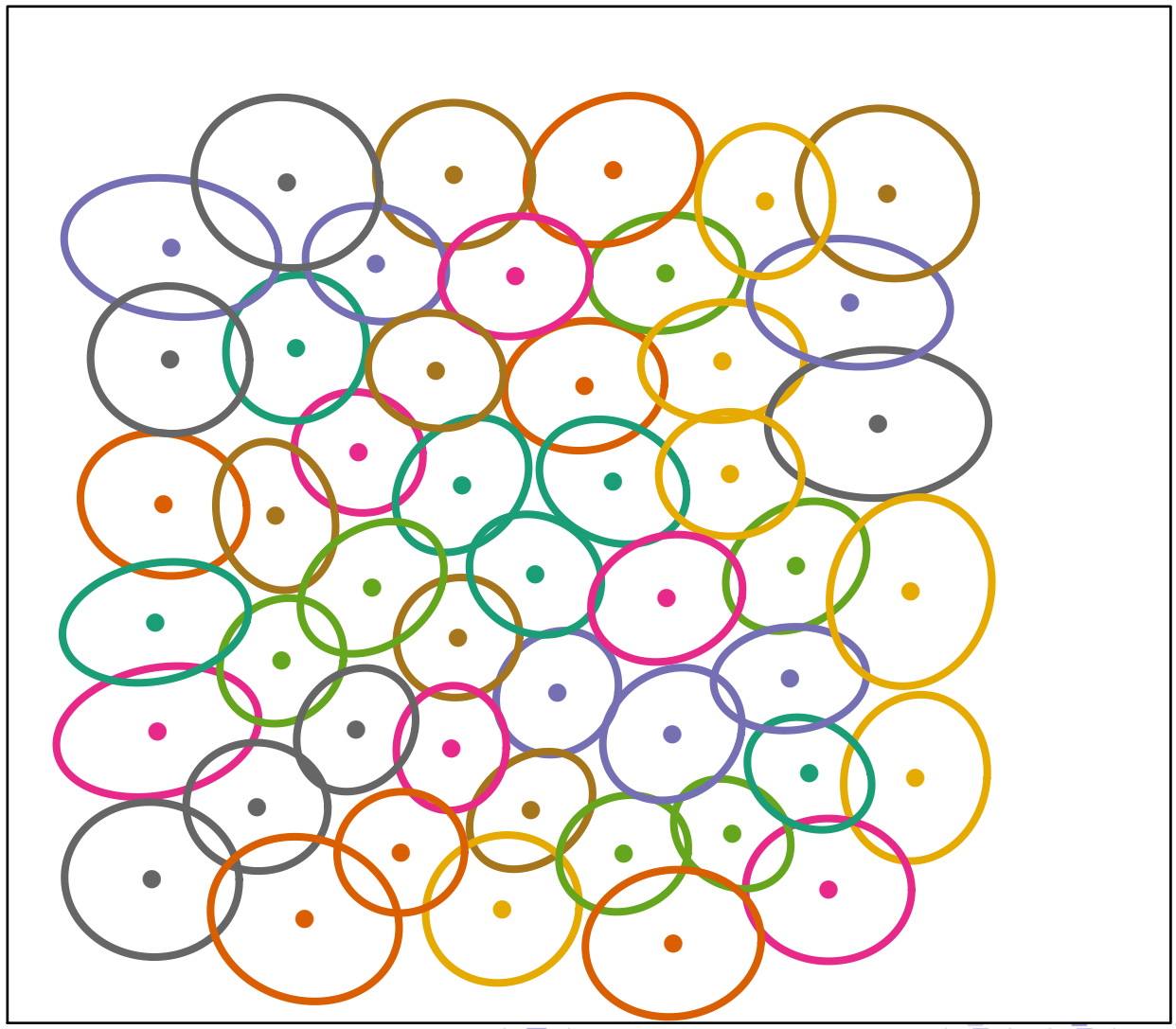


Example Run With $\sigma = 0.50$

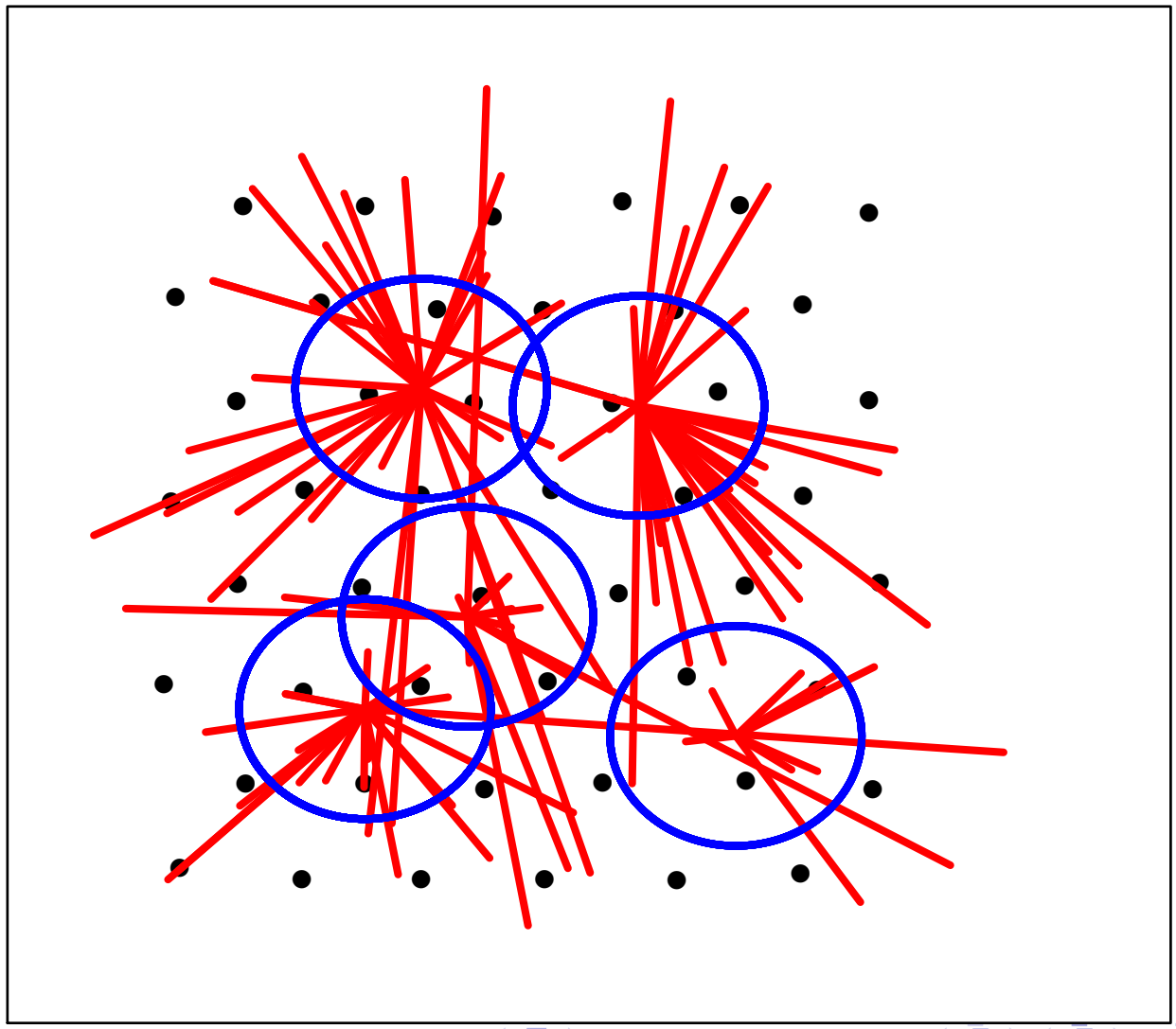
Hexagonal 8x6 Cluster Pattern $n_k = 100$



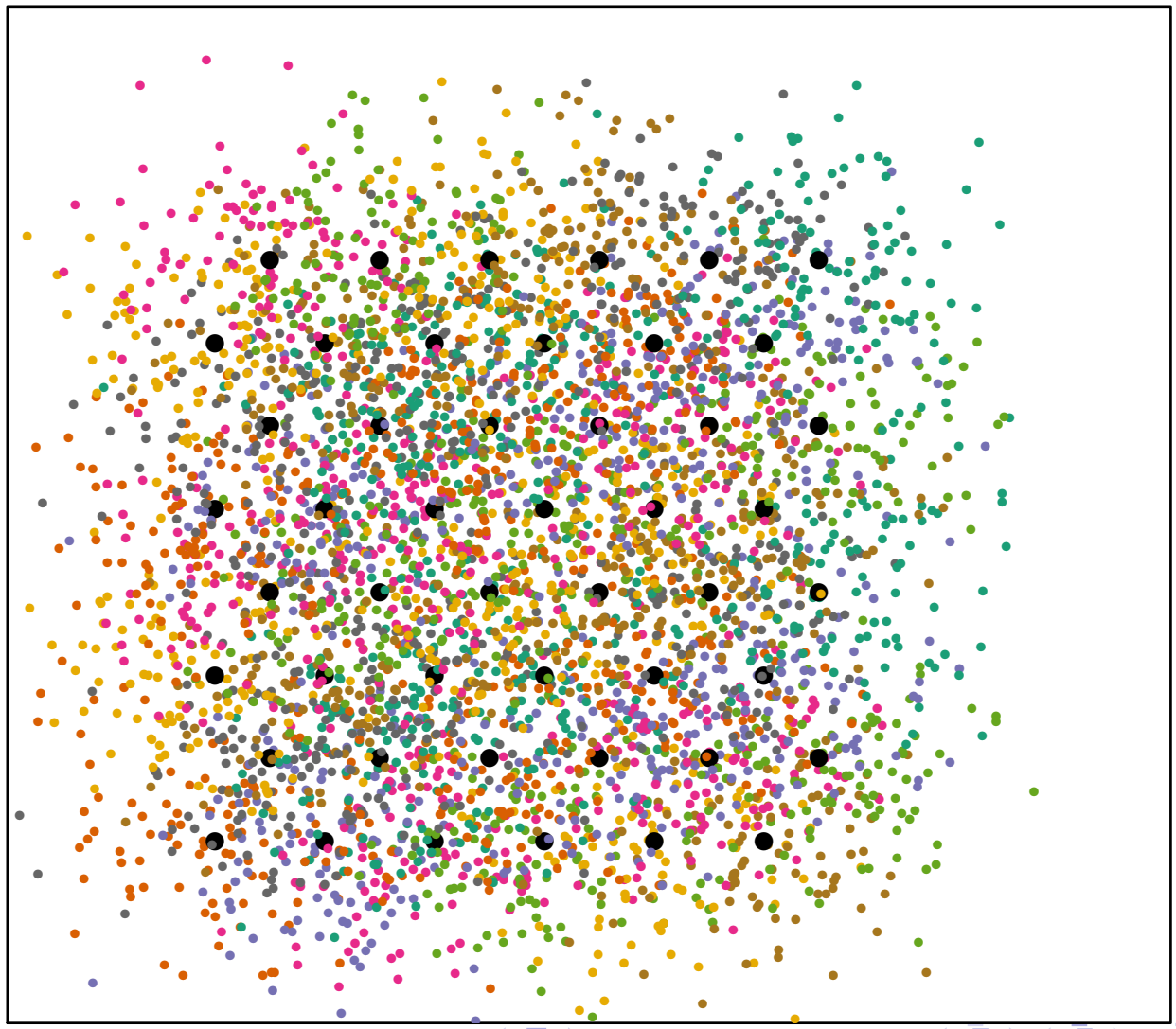
Example Run With $\sigma = 0.50$



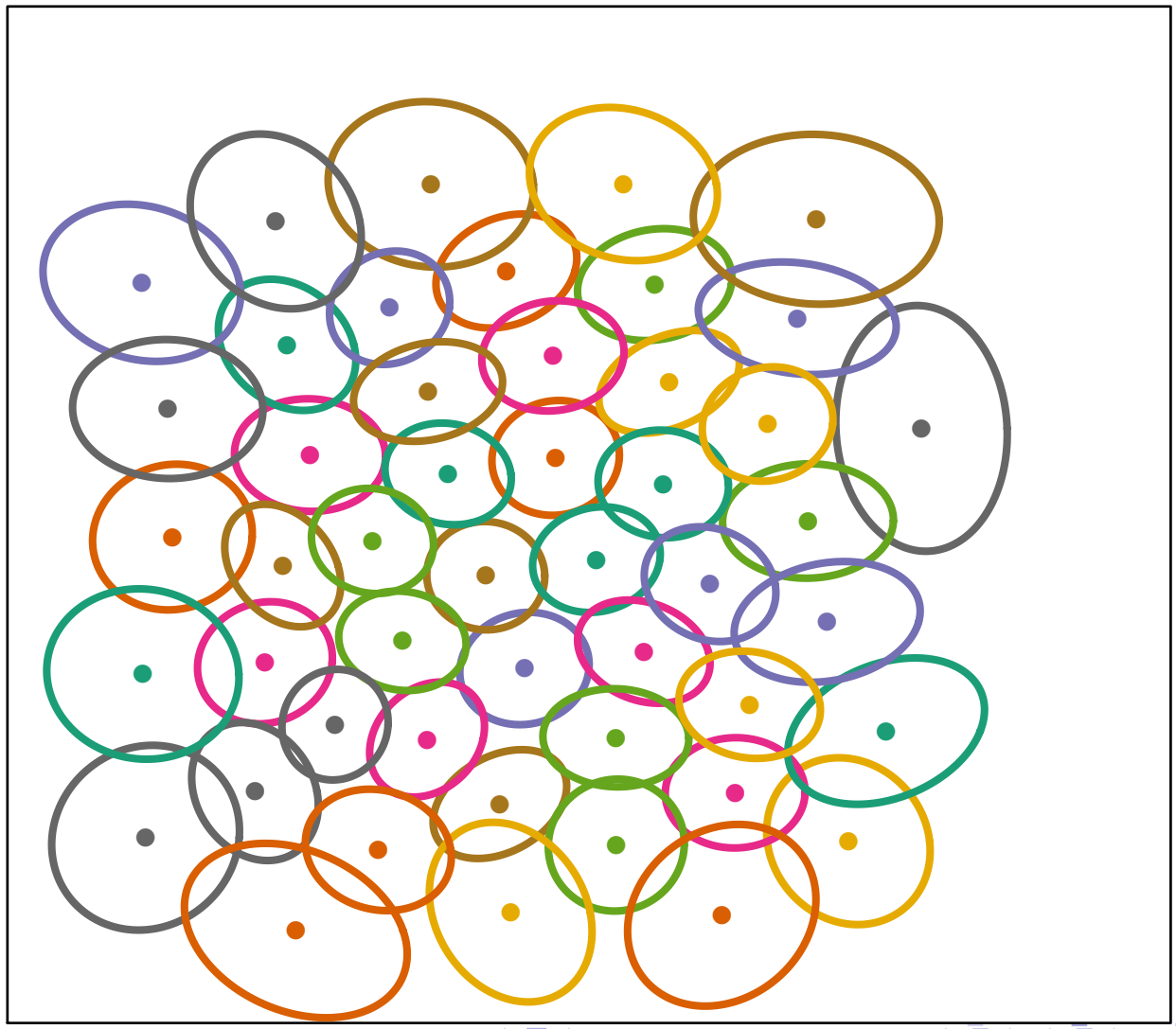
Example Run With $\sigma = 0.50$



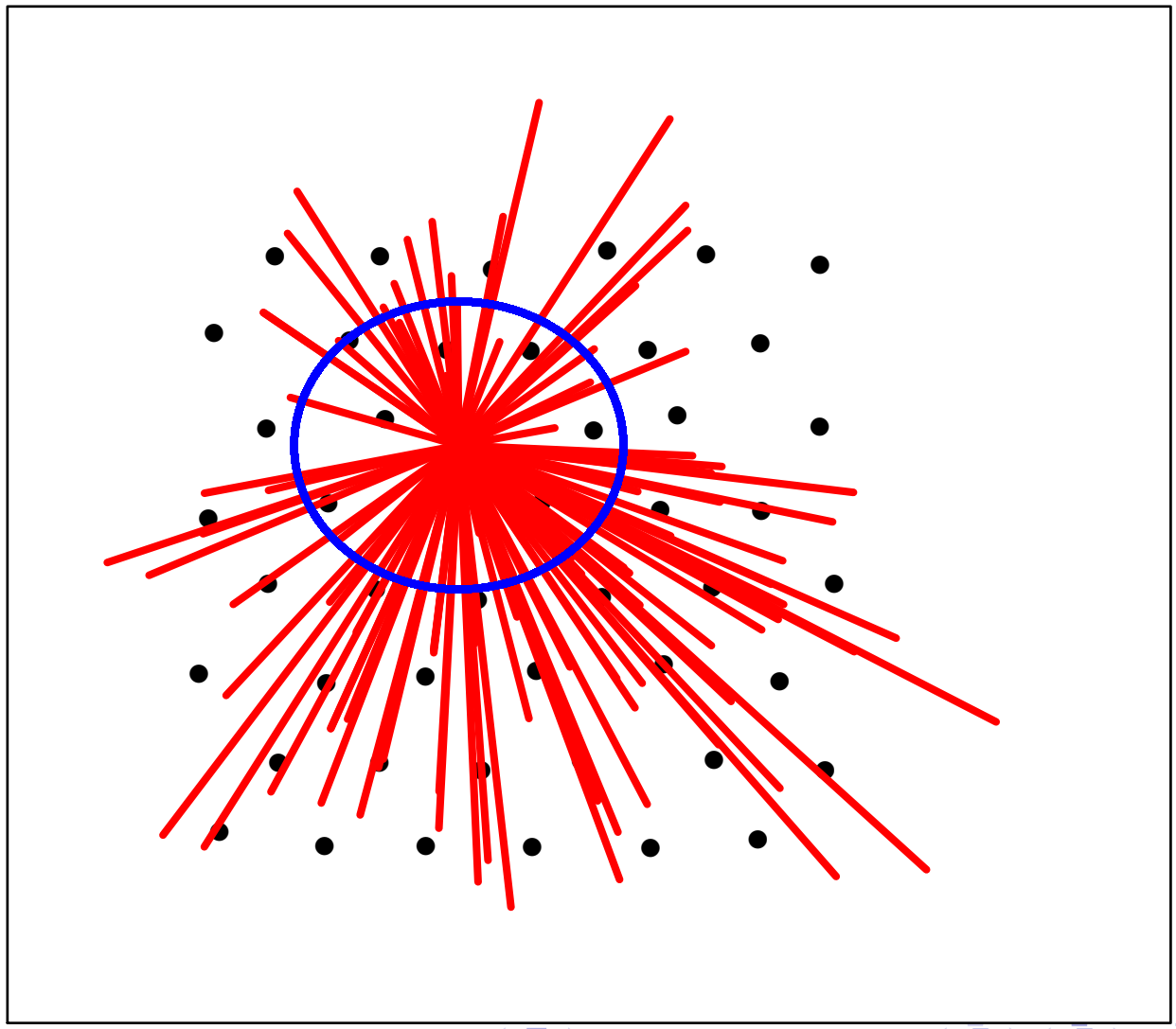
Example Run With $\sigma = 0.75$



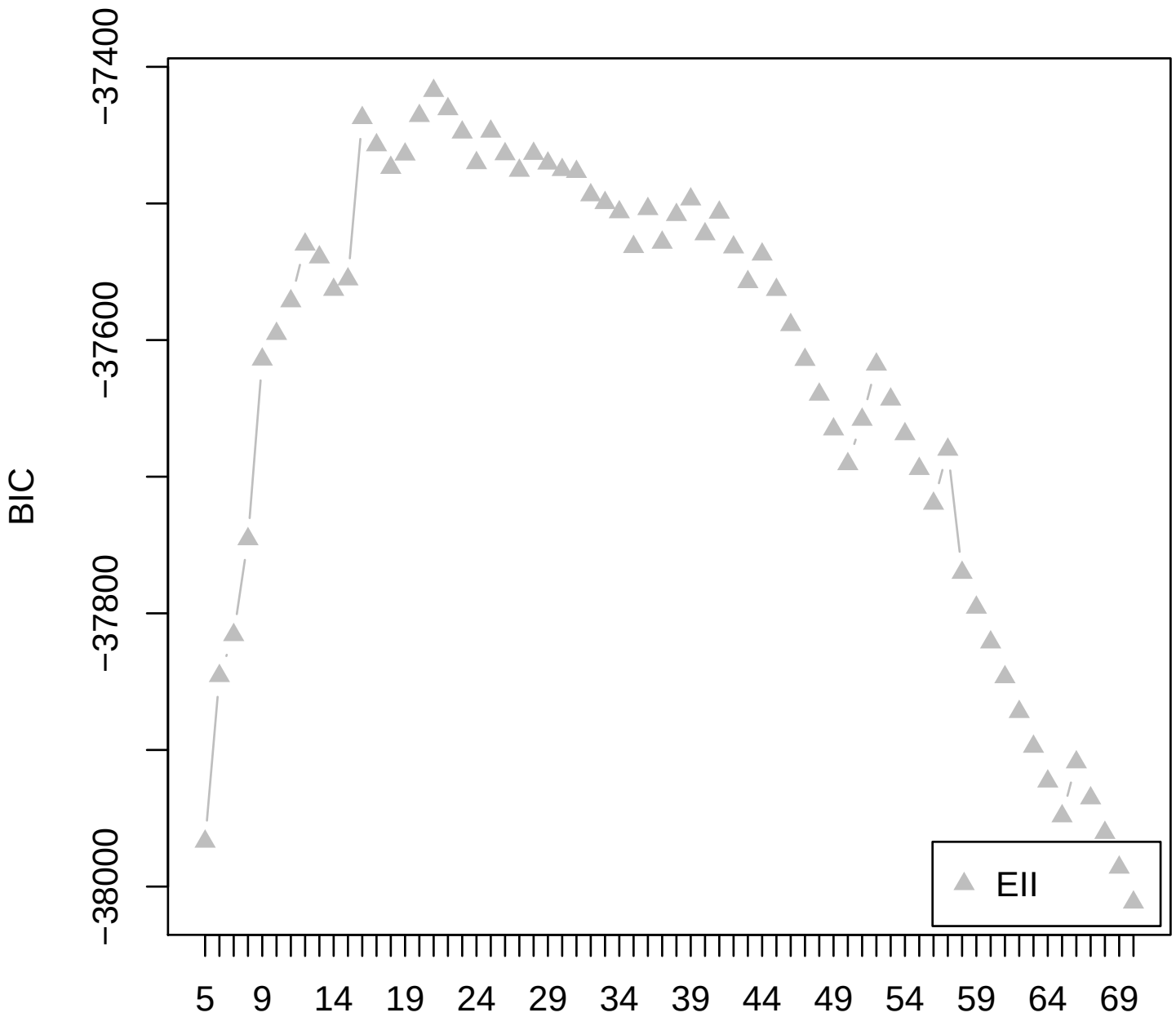
Example Run With $\sigma = 0.75$



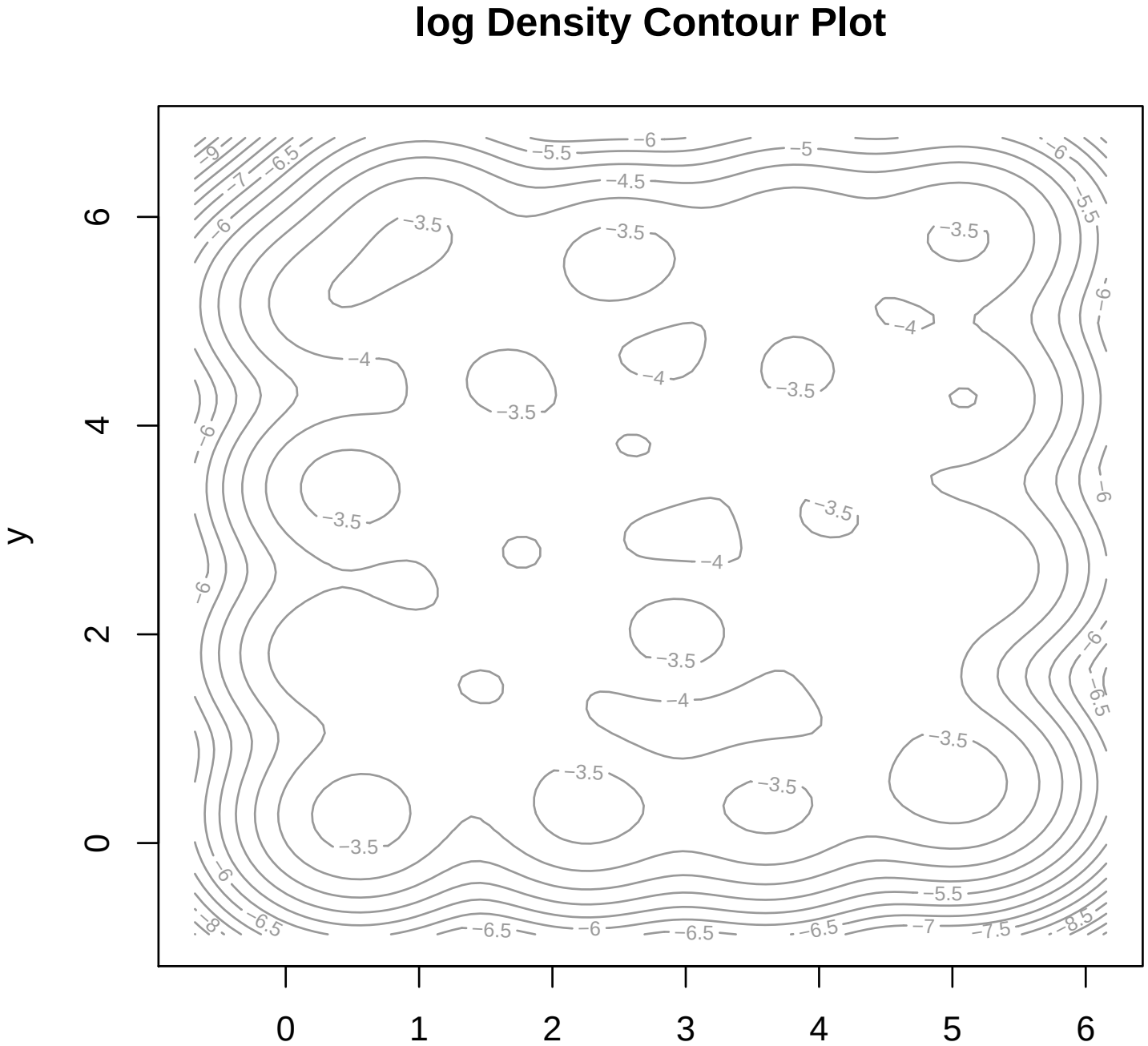
Example Run With $\sigma = 0.75$



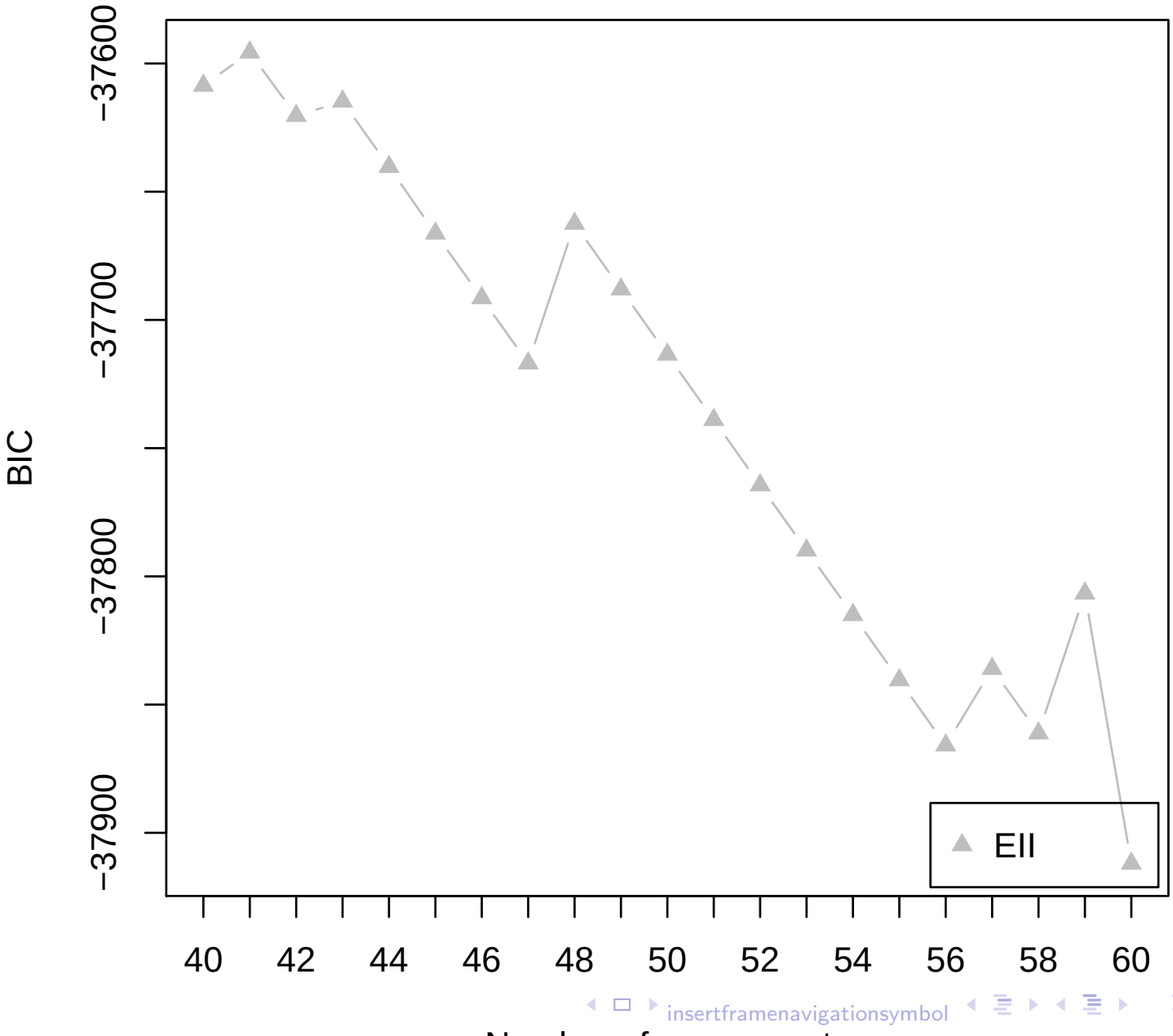
Mclust Run With $\sigma = 0.25$



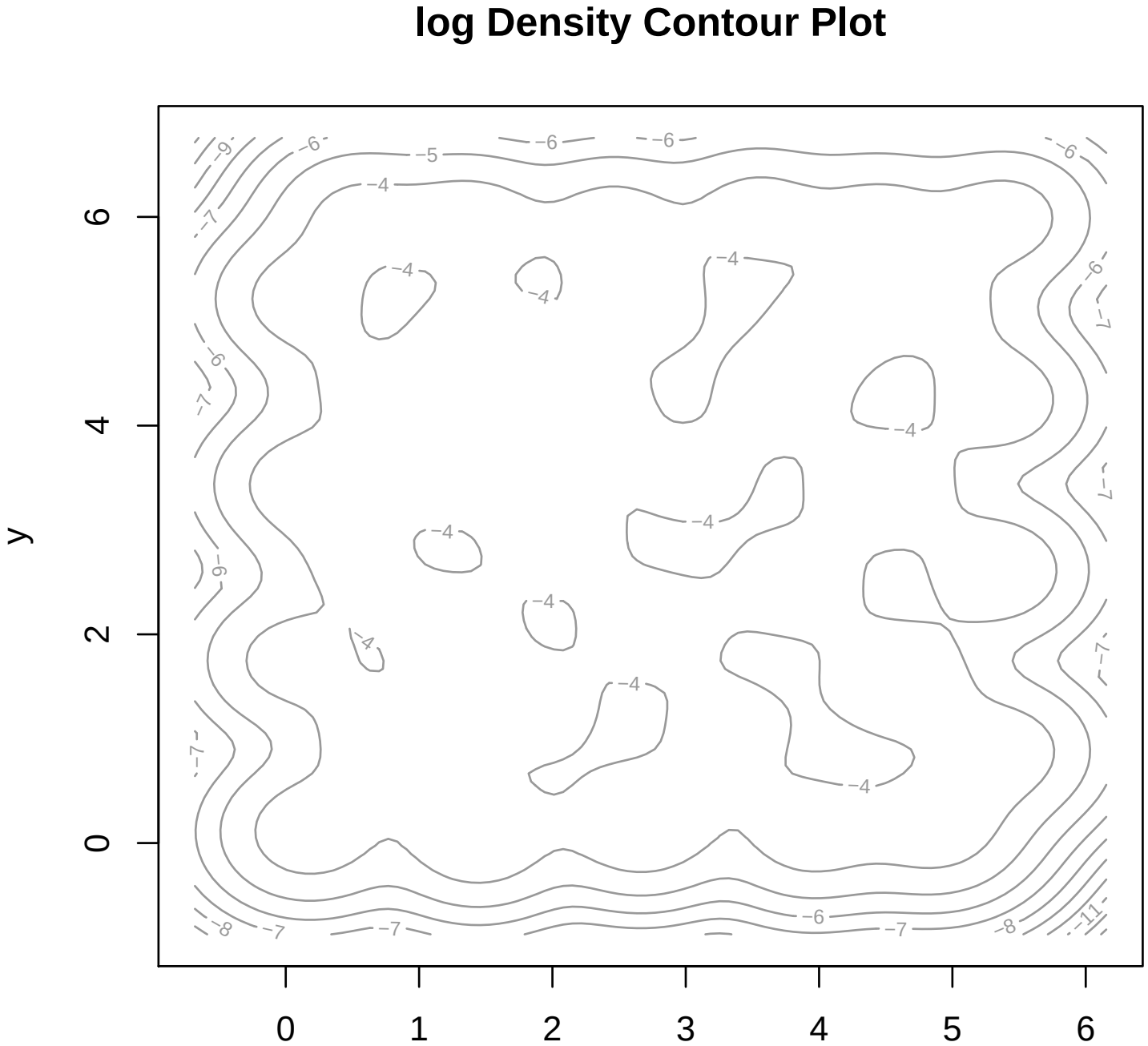
Mclust Run With $\sigma = 0.25$



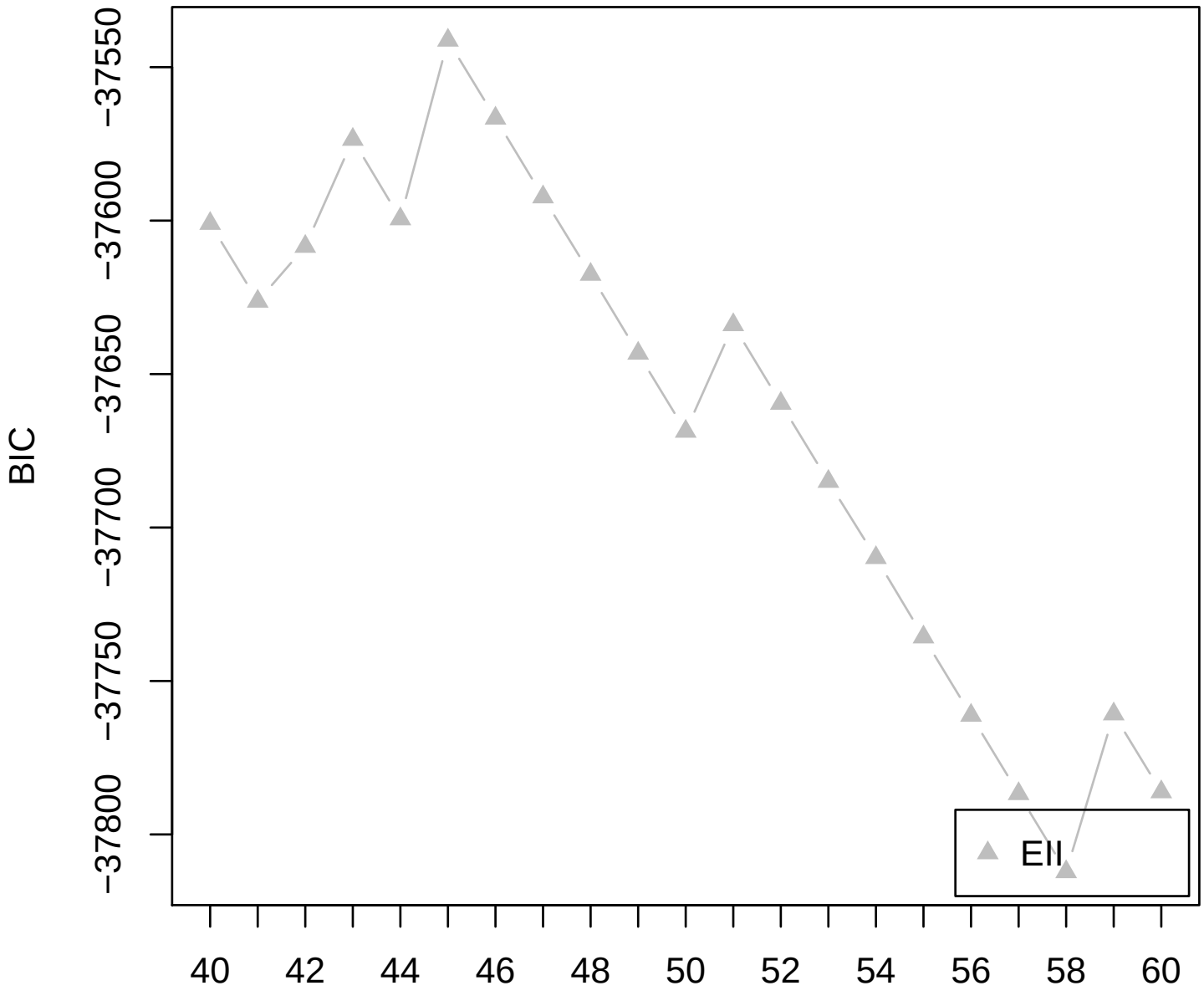
Mclust Run With $\sigma = 0.25$



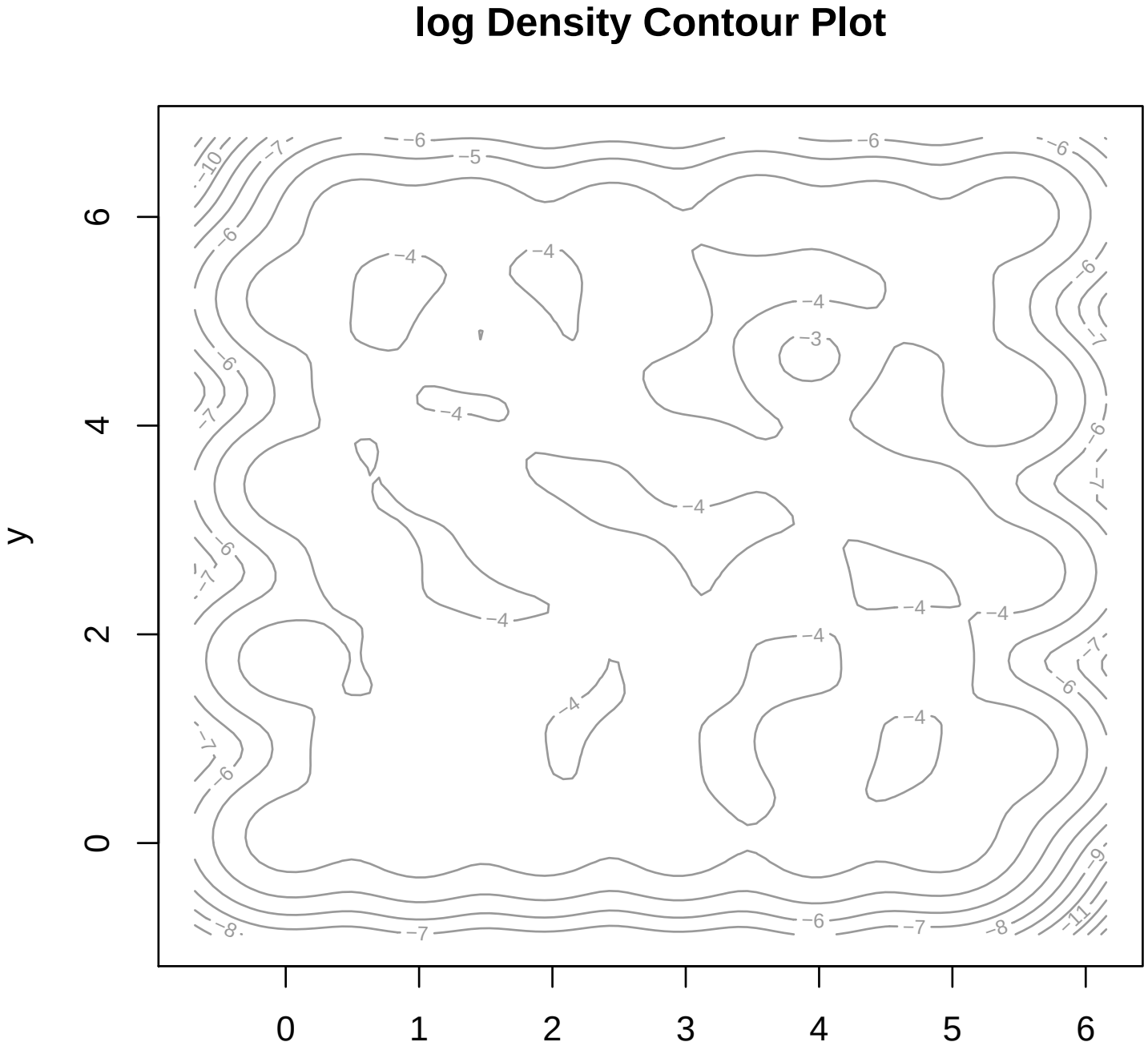
Mclust Run With $\sigma = 0.25$



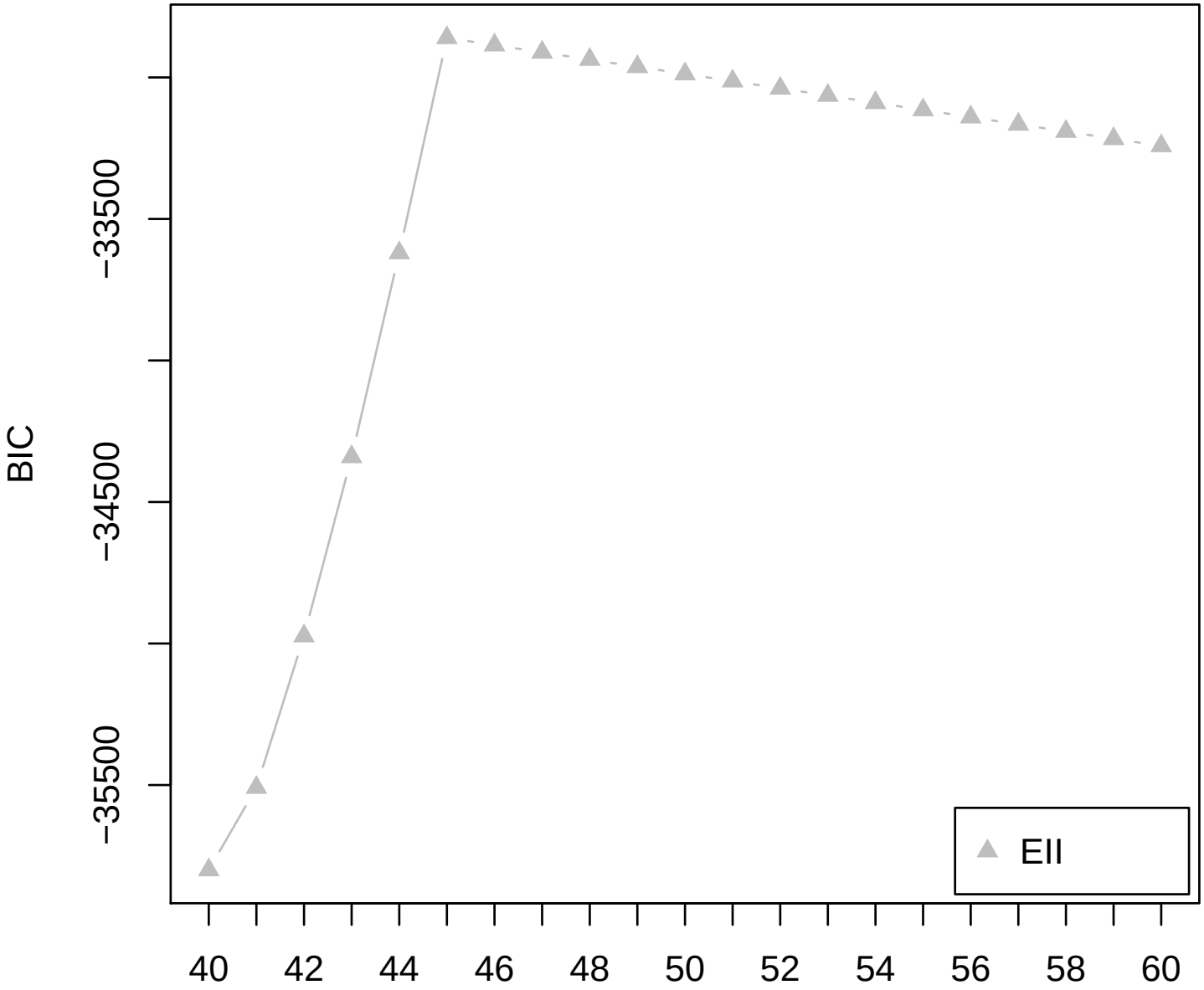
Mclust Run With $\sigma = 0.25$ (Random Start Matters)



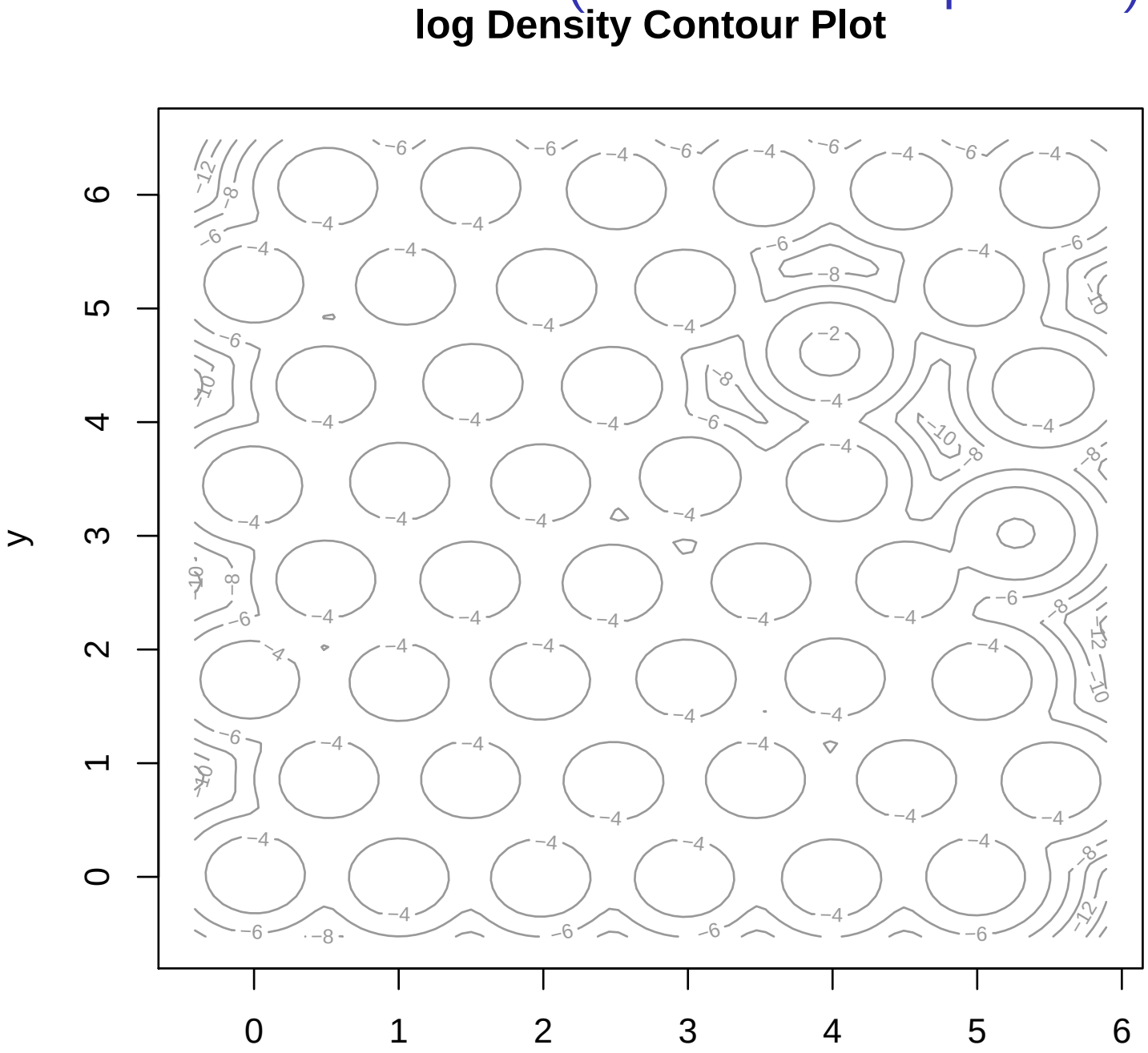
Mclust Run With $\sigma = 0.25$



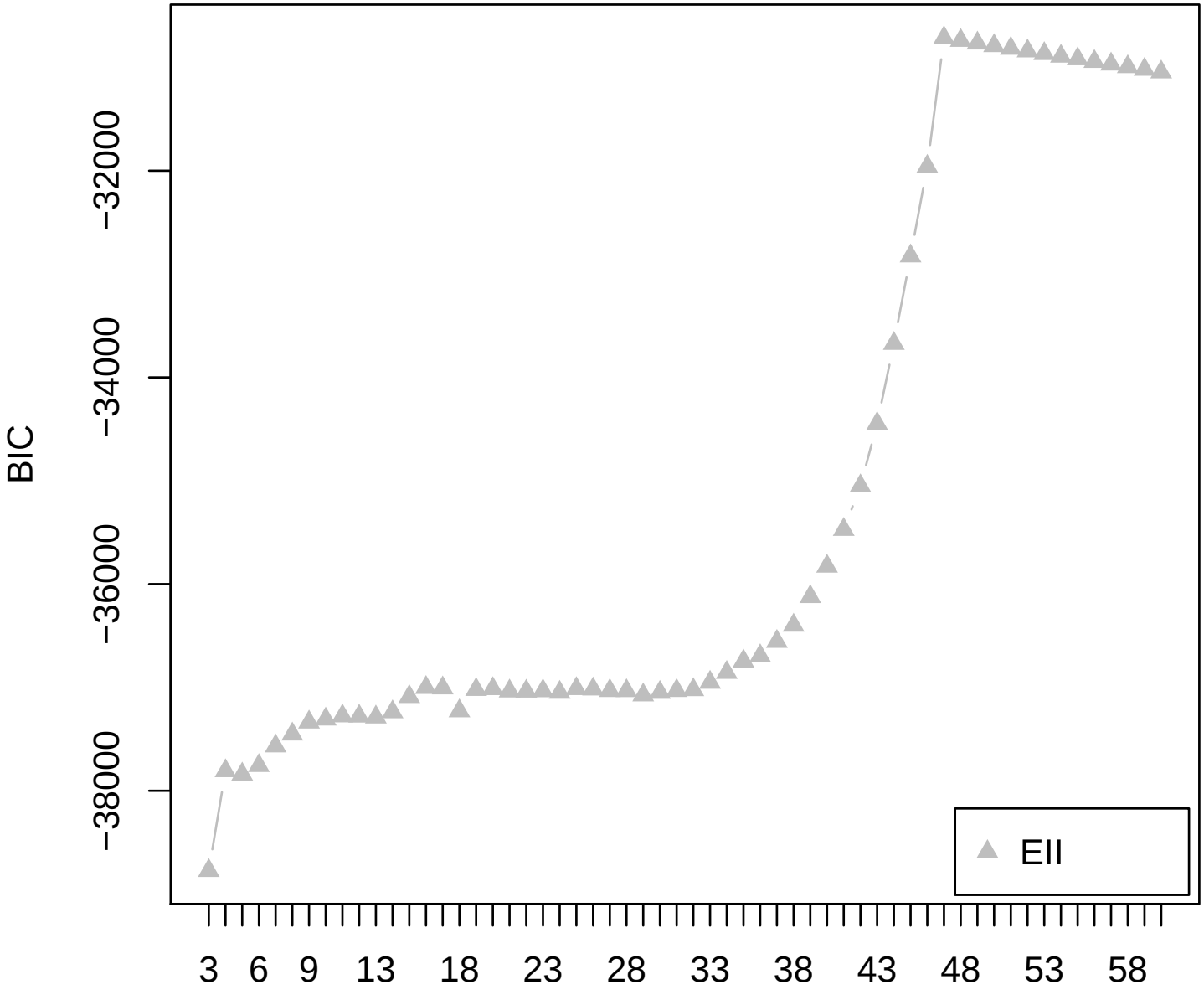
Mclust Run With $\sigma = 0.15$ (easier)



Mclust Run With $\sigma = 0.15$ (easier — not perfect)



Mclust Run With $\sigma = 0.15$ (easier)



Wrapup

- Diagnostics is challenging in higher dimensions, whether there are outliers or not
- Running L2E or other robust procedures in parallel with MLE is quite effective at uncovering anomalies, if any
- The initialization problem for MLE and many robust procedures can be replaced by probing with many random initializations using partial mixture estimation and L2E
- Probing those results can uncover the truth (or much of it)
- We have tried to give a flavor to such analyses in this talk
- The software is available at my web site
  http://www.stat.rice.edu/∼scottdw
- Thank you, especially to the organizers of such an outstanding event

More References

- Aitkin, M. and Wilson, G.T. (1980), *Mixture models, outliers, and the EM algorithm,* Technometrics, **22** (1980), 325 – 331.
- Banfield, J.D. and Raftery, A.E. (1993), *Model-Based Gaussian and Non-Gaussian Clustering,* Biometrics, **49**, 803 – 821.
- Basu, A., Harris, I.R., Hjort, H.L., and Jones, M.C. (1998), *Robust and Efficient Estimation by Minimising a Density Power Divergence.* Biometrika, **85**, 549 – 560.
- Beran, R. (1977), *Robust Location Estimates,* The Annals of Statistics, **5**, 431 – 444.
- Brown, L.D. and Hwang, J.T.G. (1993), *How To Approximate a Histogram By a Normal Density,* The American Statistician, **47**, 251 – 255.
- Hjort, H.L. (1994), *Minimum L2 and Robust Kullback-Leibler Estimation,* Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and

Random Processes, eds. P. Lachout and J.Á. Víšek, Prague Academy of Sciences of the Czech Republic, pp. 102 – 105.


Huber, P.J. (1981), *Robust Statistics.* John Wiley & Sons, New York.


Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.


Scott, D.W. (1999), *Remarks on Fitting and Interpreting Mixture Models,* Computing Science and Statistics, K. Berk and M. Pourahmadi, Eds., **31**, 104 – 109.


Scott, D.W. (2001), *Parametric Statistical Modeling by Minimum Integrated Square Error,* Technometrics, **43**, 274 – 285.


Wang, N. and Raftery, A.E. (2002), *Nearest-neighbor variance estimation: Robust covariance estimation via nearest-neighbor cleaning,* Journal of the American Statistical Association, **97**, 994 – 1019.