

# Density Deconvolution by Weighted Kernel Estimators

Martin Hazelton  
Massey University

[m.hazelton@massey.ac.nz](mailto:m.hazelton@massey.ac.nz)

Joint work with Berwin Turlach (University of Western Australia).

# What is Density Deconvolution?

**Model:**  $Y = X + Z$

- $Y$  is ‘contaminated’ observation; density  $g$ .
- $X$  is uncontaminated latent variable; density  $f$ .
- $Z$  is measurement error; known density  $\eta$ .

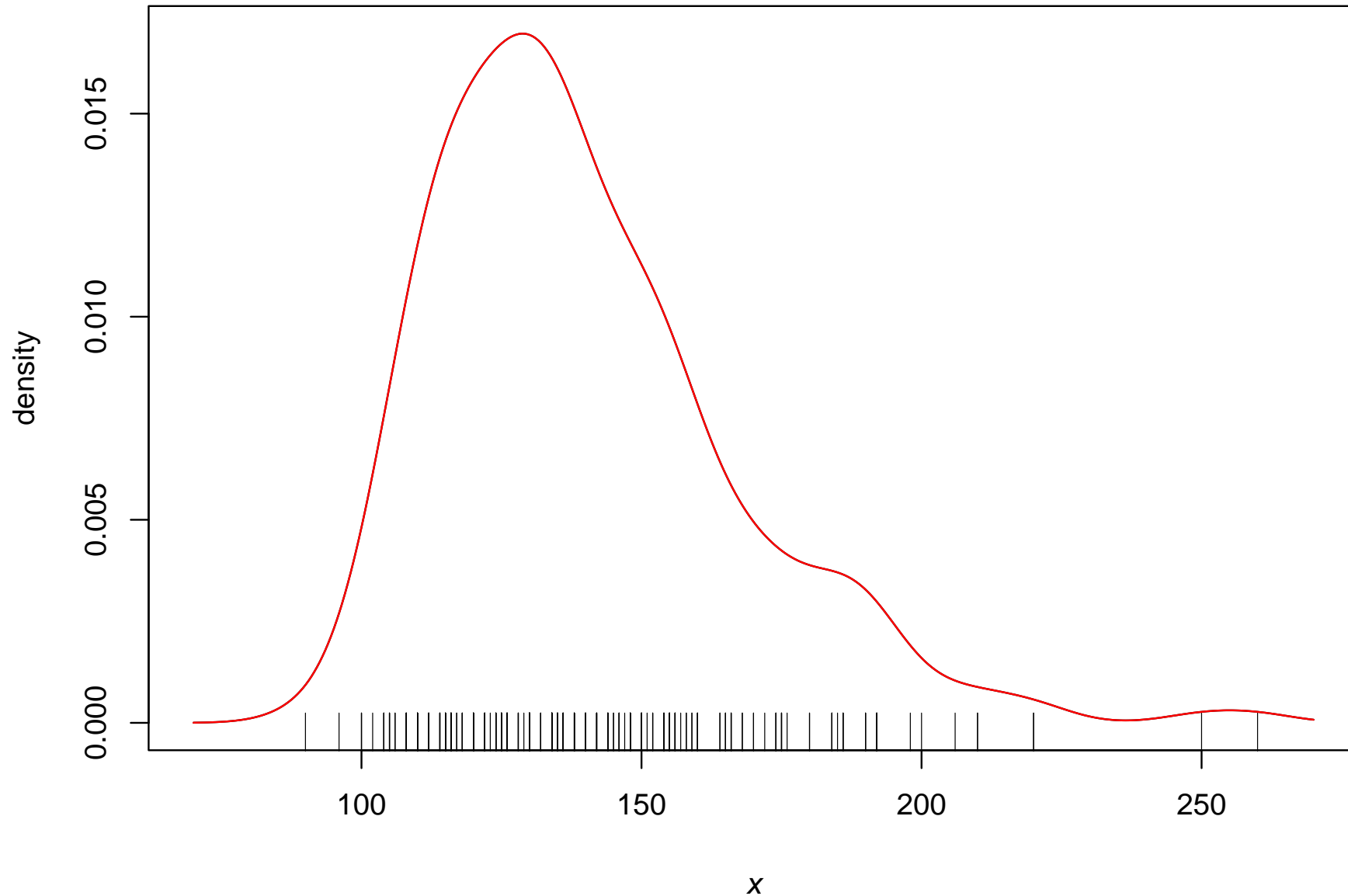
**Data:** Observe random sample  $Y_1, \dots, Y_n$ .

**Aim:** **Nonparametric** estimation of  $f$ .

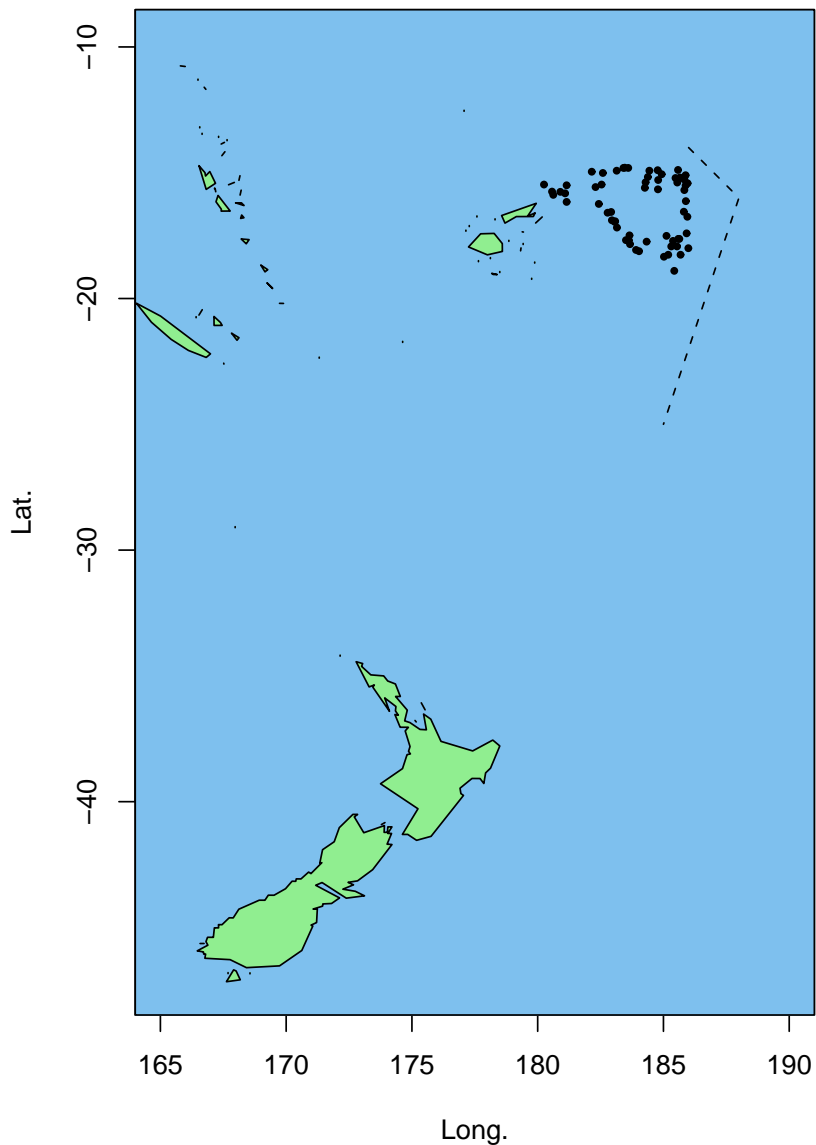
# Example 1: Blood Pressure Measurements

- Data on  $n = 285$  male subjects aged 56 years and over (from Framingham health study).
- Any subject's blood pressure measurement will vary due to
  - Error in measurement;
  - Physiological variations.
- Model  $Y = X + Z$  with
  - $Y$  is measured systolic blood pressure (in mmHg);
  - $X$  is 'long term mean' blood pressure;
  - $Z$  is 'measurement error'; model  $Z \sim N(0, 9.0^2)$ .

# Systolic BP: Convoluted Density Estimate



# Example 2: Tonga Trench Earthquakes



- Points are locations of earthquakes between the Tonga Trench and Fiji.
- Data collected 1960–1965.
- Locations subject to measurement error; std. dev. approximately 0.1 degrees ( $\approx 10\text{km}$ ).
- Will ignore edge effects for simplicity.

# Example 3: Gratuitous World Cup Photo

There are doubtless a number of examples from sport, too . . .



# Structure of Talk

- Density deconvolution in context
- Classical kernel denconvolution
- Weighted kernel estimators
- Choosing the weights
- Theoretical and numerical results
- Examples

# Deconvolution and Inverse Problems

- Density deconvolution is an example of a **positive linear inverse problem**:

$$g(y) = \int \tau(x, y) f(x) dx$$

where  $f$ ,  $g$ ,  $\tau$  are non-negative functions.

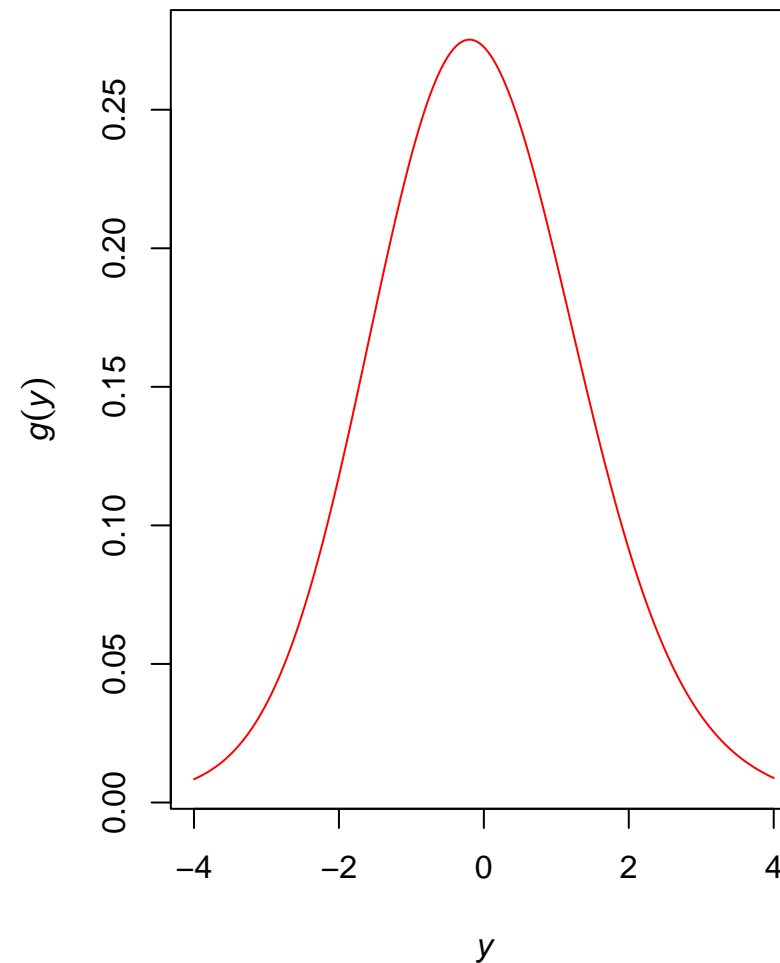
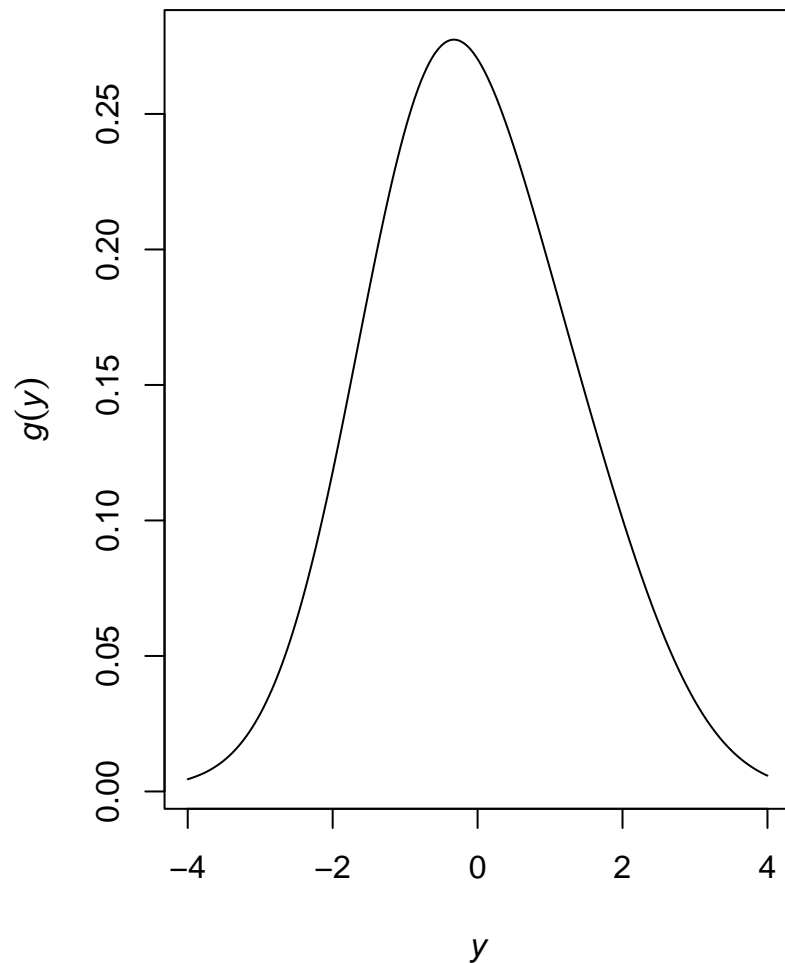
- Many applications, e.g. *image deblurring*, *traffic flow estimation*
- In deconvolution context, densities  $f$ ,  $g$ ,  $\eta$  related by convolution formula:

$$g(y) = \int \eta(y - x) f(x) dx = f * \eta(y)$$

- Such indirect estimation problems can be hard; e.g.  $1/\log(n)$  best rate for deconvolution with Gaussian noise.

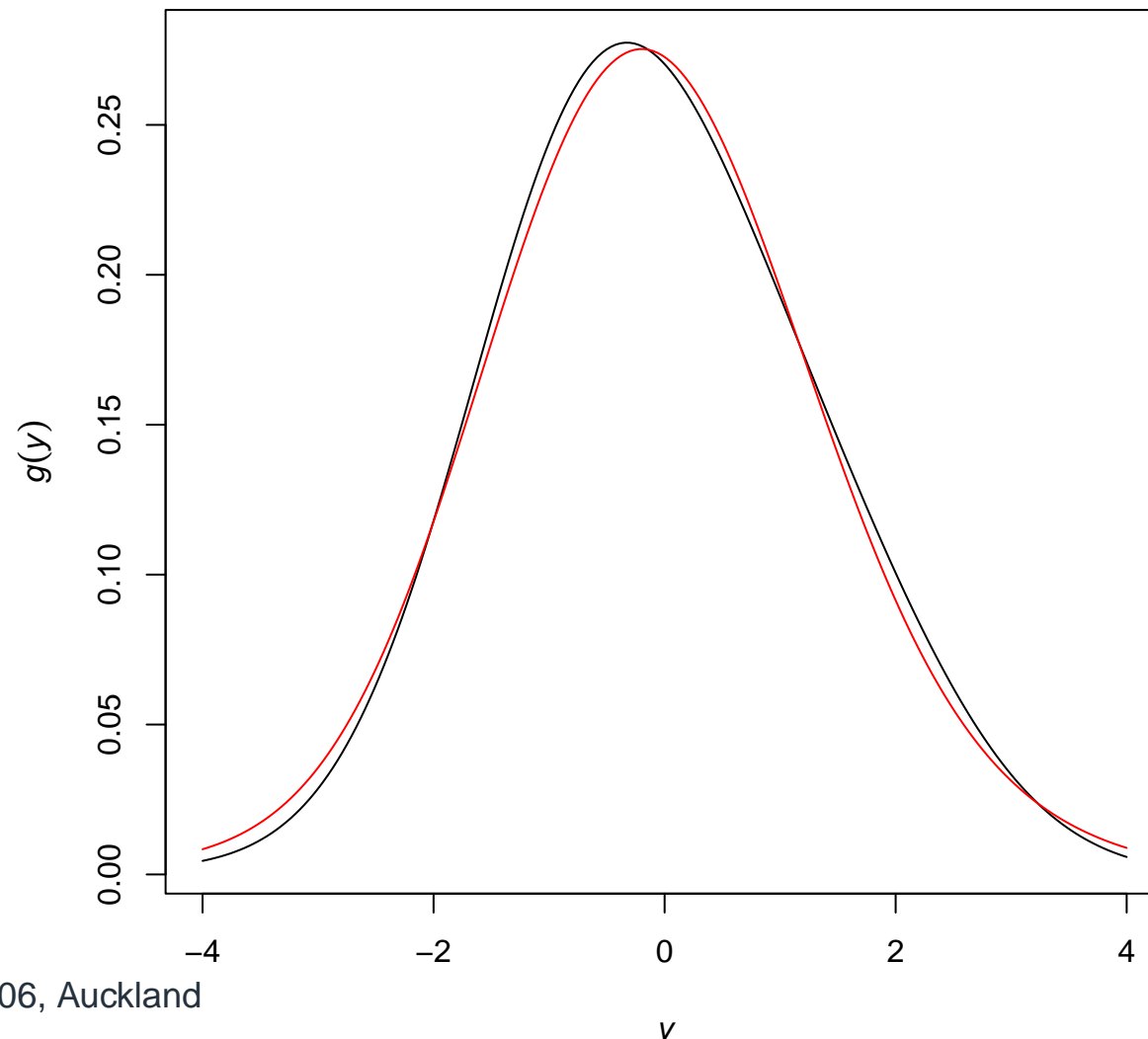
# Spot the Difference # 1

Two convoluted densities ( $g$ ):



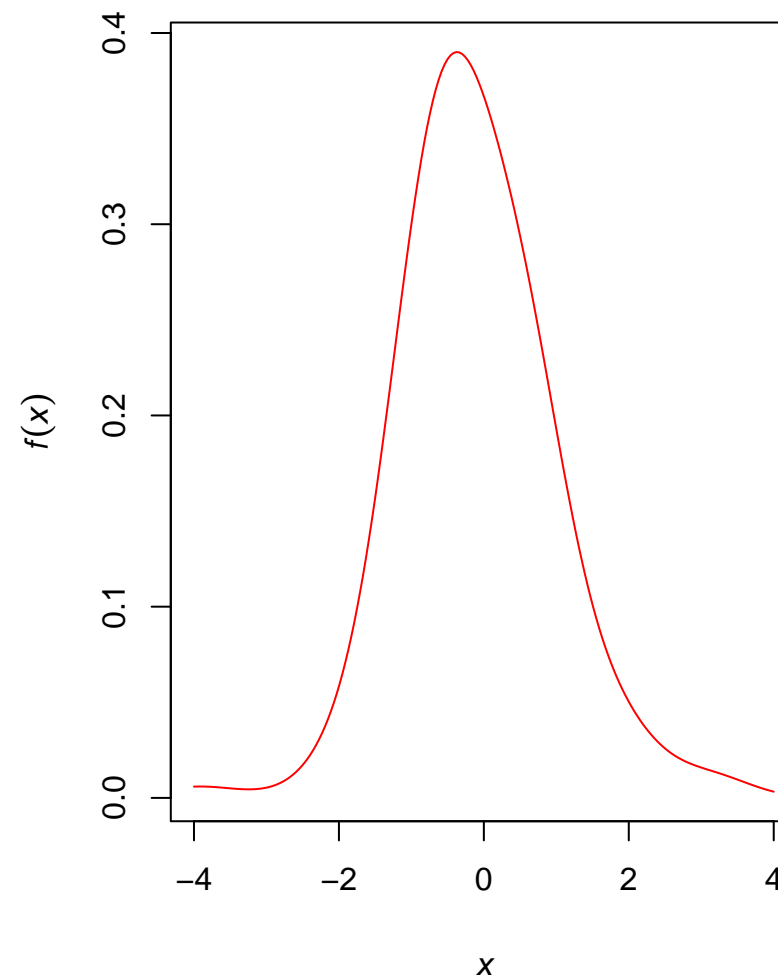
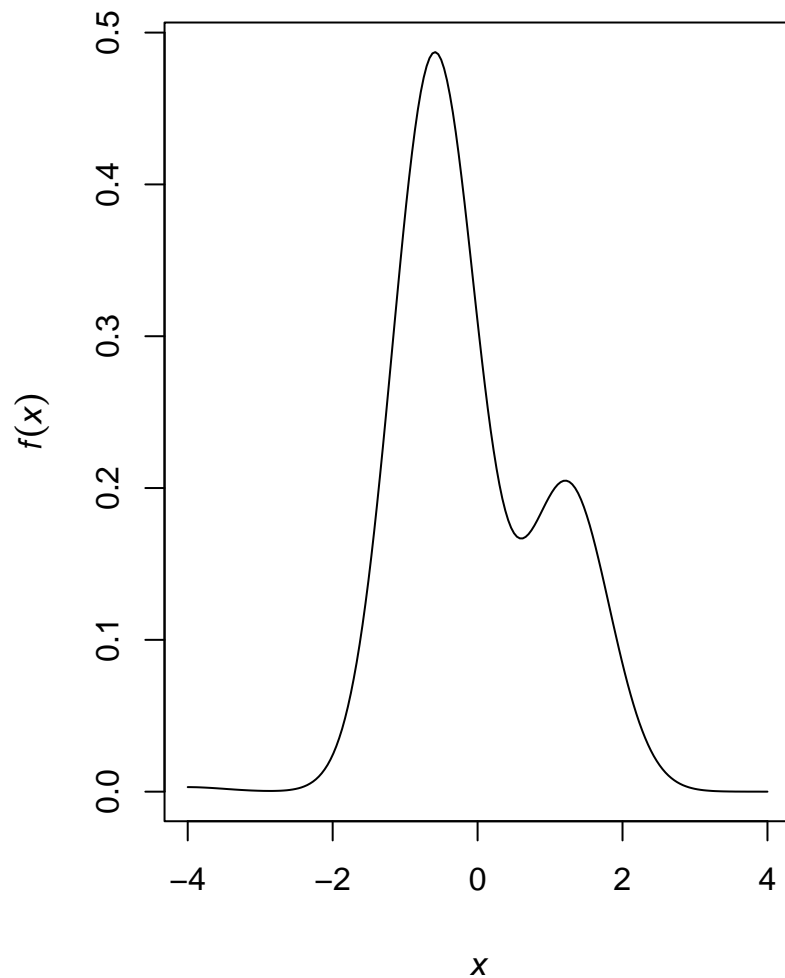
# Spot the Difference # 1

Two convoluted densities ( $g$ ):



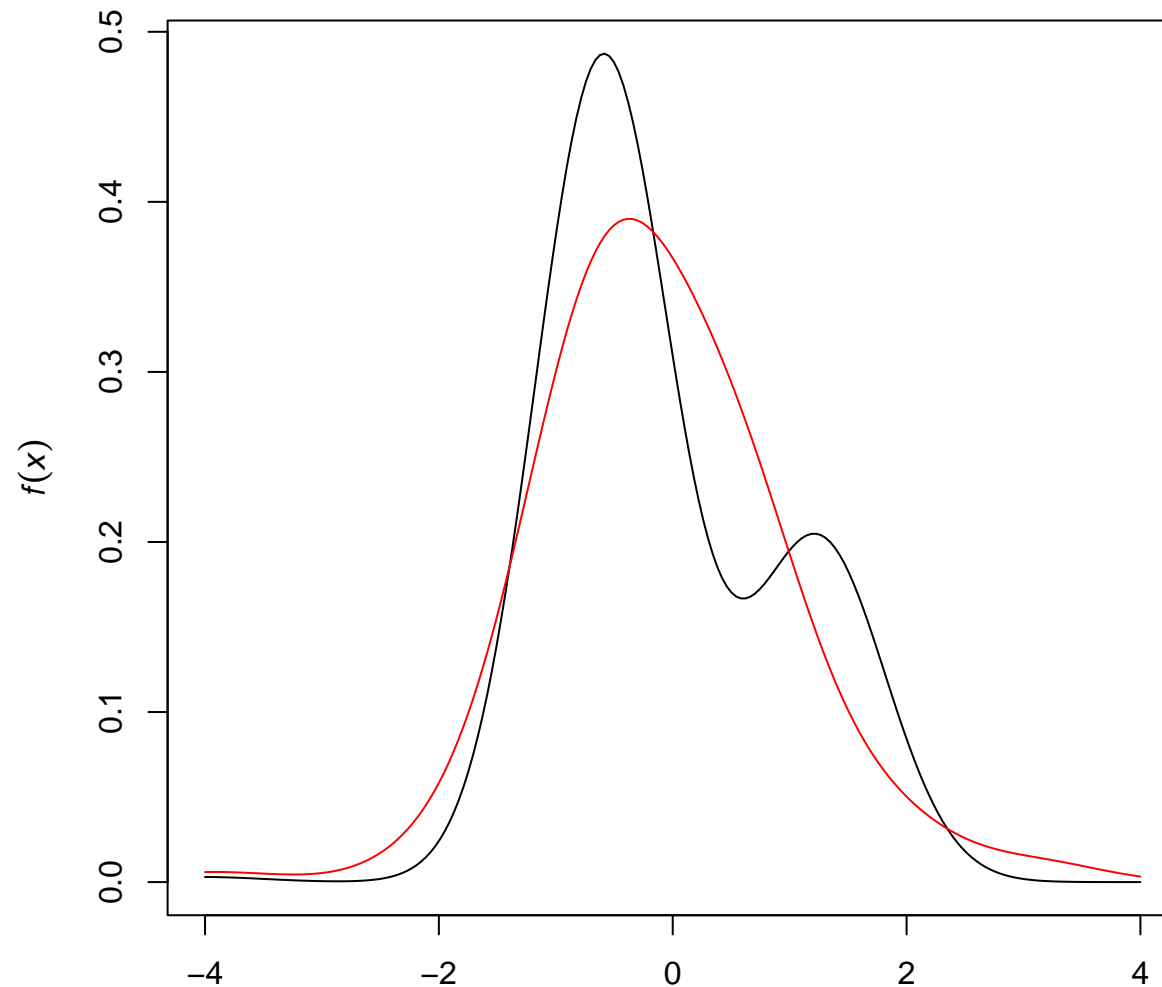
# Spot the Difference # 2

Two deconvoluted densities ( $f$ ):



# Spot the Difference # 2

Two deconvoluted densities ( $f$ ):



# Kernel Density Estimation

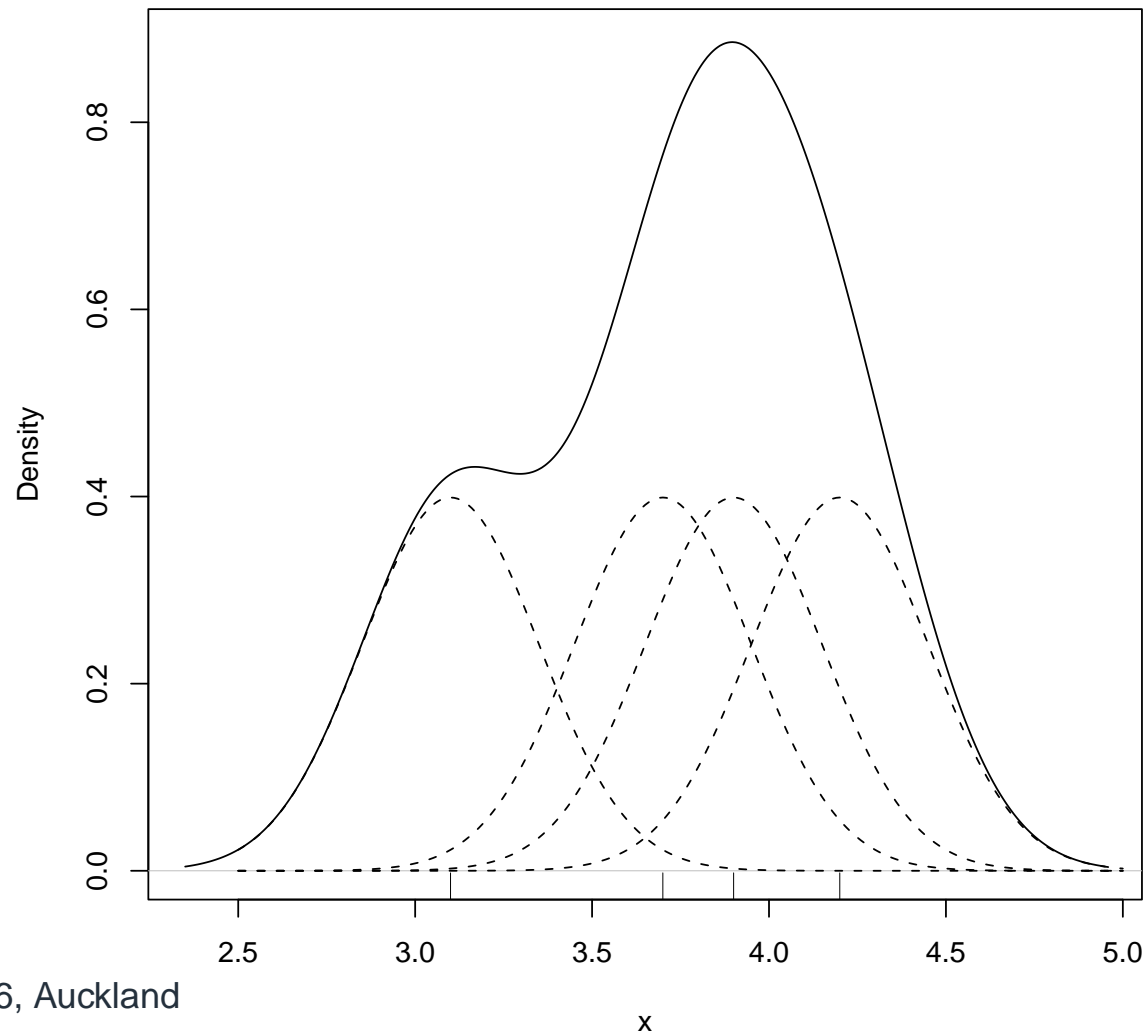
- Popular approach in nonparametric density estimation.
- Standard kernel density estimate constructed from data  $X_1, \dots, X_n$  is

$$\check{f}(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$$

- $K(\cdot)$  is **kernel function**; assumed to be a p.d.f. with symmetry  $K(x) = K(-x)$ ;
- $K_h(x) = h^{-1}K(x/h)$  is scaled kernel;
- $h$  is the **bandwidth**, which controls smoothness of estimate.

# Kernel Density Estimation Graphic

Estimate  $\hat{f}$  is aggregate of 'bumps' centred at data points.



# Classical Kernel Deconvolution

- In deconvolution must modify density estimate constructed from contaminated data  $Y_1, \dots, Y_n$  to account for measurement error.
- Classical approach adapts kernel function. Viz:

$$\check{g}(y) = \hat{f} * \eta(y) \Rightarrow \phi_{\hat{f}}(t) = \phi_{\check{g}}(t) / \phi_{\eta}(t)$$

where  $\phi_f$  is characteristic function of  $f$  etc. Hence

$$\check{g}(y) = n^{-1} \sum_{i=1}^n K_h(y - Y_i) \Rightarrow \hat{f}(y) = n^{-1} \sum_{i=1}^n K_h^Z(y - Y_i; h)$$

where  $K_h^Z(\cdot; h)$  is kernel with characteristic function  $\phi_K / \phi_{\eta}$ .

# Comments on Classical Method

- Deconvolution kernel  $K_h^Z(\cdot; h)$  will take negative values, and so generally will  $\hat{f}$ .
- Evaluation of deconvolution kernel is computationally expensive for multivariate data with complex measurement error structure.
- Most (good) kernel density estimators have form

$$\hat{f}(y) = \int K_h(y - x) d\hat{F}(x)$$

where  $\hat{F}$  consistent for distribution function of  $X$ , and  $K \geq 0$ .

- Classical deconvolution estimator does not have this form.

# Weighted Kernel Estimation

Weighted kernel estimator is

$$\hat{f}_{\mathbf{w}}(y) = n^{-1} \sum_{i=1}^n w_i K_h(y - Y_i)$$

- $\mathbf{w} = (w_1, \dots, w_n)^T$  vector of non-negative weights;
- to make  $\int \hat{f}_{\mathbf{w}}(x) dx = 1$ , impose constraint  $\bar{w} = n^{-1} \sum_{i=1}^n w_i = 1$ .
- Helpful to think in terms of weighting function  $\alpha$ , with  $w_i = \alpha(Y_i)$ .
- Note 'good' form: if  $\hat{F}(A) = n^{-1} \sum_{i=1}^n w_i \mathbf{1}_A(Y_i)$  then

$$\hat{f}_{\mathbf{w}}(y) = \int K_h(y - x) d\hat{F}_{\mathbf{w}}(x)$$

# Motivation for Weighted Estimators

- Define  $\alpha_0(x) = f(x)/g(x)$ , and corresponding weights  $w_{0i} = \alpha_0(Y_i)$  ( $i = 1, \dots, n$ ).
- Then
$$\begin{aligned}\mathbb{E}[\hat{f}_{w_0}(y)] &= \mathbb{E}[w_{0n}K_h(y - Y_n)] \\ &= \int \alpha_0(x)K_h(y - x)g(x) dx \\ &= \int K_h(y - x)f(x) dx \\ &= \mathbb{E}[K_h(y - X_n)]\end{aligned}$$
- Use of weight function  $\alpha_0(x)$  compensates for measurement error (at expense of modest increase in estimator's variance).

# Choosing the Weights

- Intuitively, want  $\hat{f}_w * \eta(y) \approx \check{g}(y)$  for all  $y$ .
- Hence choose weights by so that  $\hat{f}_w * \eta$  and  $\check{g}(y)$  match as closely as possible:

$$\hat{w} = \operatorname{argmin}_w D \left( \hat{f}_w * \eta, \check{g} \right)$$

where  $D$  measures discrepancy between densities.

- Methodology then depends on:
  1. how we measure discrepancy between  $\hat{f}_w * \eta$  and  $\check{g}$ ;
  2. whether we reduce the dimensionality of optimization problem by placing constraints on  $w$ .

# Some Discrepancy Measures

Integrated squared difference (ISE):

$$\hat{w} = \operatorname{argmin}_{w: \bar{w}=1} \int \{\hat{f}_w * \eta(y) - \check{g}(y)\}^2 dy$$

where

- $\hat{f}_w(y) = n^{-1} \sum_{i=1}^n w_i K_h(y - Y_i);$
- $\check{g}(y) = n^{-1} \sum_{i=1}^n K_h(y - Y_i).$

# Discrepancy Measures (cont.)

## Kullback Leibler approach (KL)

$$\hat{w} = \operatorname{argmax}_{w: \bar{w}=1} \int \log\{\bar{f}_w * \eta(y)\} d\hat{G}(y)$$

where

- $\bar{f}_w(x) = n^{-1} \sum_{i=1}^n w_i K_0(y - Y_i)$ , so that

$$\bar{f}_w * \eta(y) = n^{-1} \sum_{i=1}^n w_i \eta(y - Y_i)$$

- $\hat{G}(A) = n^{-1} \sum_{i=1}^n \mathbf{1}_A(Y_i)$  is standard empirical distribution function.

# Discrepancies Compared

- ISE works on natural scale for density estimation.
- ISE requires prior choice of bandwidth, KL does not.
- Theoretical analysis easier for KL.
- Constraint  $\bar{w} = 1$  easier to apply with KL.
- Both methods produce highly encouraging numerical results.

# Normalizing the KL Method

Use unconstrained optimization of following objective function:

$$\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} \int \log\{\bar{f}_{\mathbf{w}} * \eta(y)\} d\hat{G}(y) - n^{-1} \sum_{i=1}^n w_i = Q_{KL}(\mathbf{w})$$

*Rationale:* suppose  $\int \bar{f}_{\mathbf{w}}(x) dx = n^{-1} \sum_{i=1}^n w_i = t$ . Then

$$\begin{aligned} Q_{KL}(\mathbf{w}) &= \int \log\{\bar{f}_{\mathbf{w}} * \eta(y)\} d\hat{G}(y) - t \\ &= \int \log\{t \bar{f}_{\mathbf{w}/t} * \eta(y)\} d\hat{G}(y) - t \\ &= \log(t) + (1 - t) + Q_{KL}(\mathbf{w}/t) \leq Q_{KL}(\mathbf{w}/t) \end{aligned}$$

# $n$ -Dimensional Optimization

- Optimization of ISE or KL objective functions with respect to  $w_1, \dots, w_n$  is problematic:
  - Computational difficulties finding global minimum;
  - Difficulties with theoretical analysis (since infinitely many parameters);
- Simple (or simple-minded?) **pragmatic solution**:
  - Choose  $w = 1$  as initial value;
  - Choose standard quasi-Newton optimization algorithms;
  - Stop when local minimum found.

# Regularizing the Optimization Problem

- Reduce dimensionality of optimization problem by assuming that weight function  $\alpha$  has some given parametric form.
- Example (when  $Y$  is univariate):

$$w_i = \alpha(Y_i) = e^{s(Y_i|\boldsymbol{\theta})}$$

where  $s(\cdot|\boldsymbol{\theta})$  is a cubic spline.

- Some motivation: when  $X$  and  $Z$  are normal, then optimal weight function satisfies

$$\alpha_0(Y_i) = e^{q(Y_i|\boldsymbol{\theta})}$$

where  $q$  is a quadratic polynomial.

# Some Theory for the KL Method

Assume:

- $\alpha$  in parametric class  $\mathcal{A}$ ;
- $f/g = \alpha_0 \in \mathcal{A}$ ;
- Some regularity conditions.

Then

$$\mathbb{E}\{|\hat{\boldsymbol{w}} - \boldsymbol{w}_0|^2\} = o(n^{-4/5})$$

where  $\hat{\boldsymbol{w}}$  is KL optimal weight vector.

# More Theory for the KL Method

- One we have estimated weights  $\hat{w}$ , we can estimate target density by

$$\hat{f}_{\hat{w}}(y) = n^{-1} \sum_{i=1}^n \hat{w}_i K_h(y - Y_i)$$

- In unweighted kernel density estimation, best rate is

$$\mathbb{E}[\{\check{f}(x) - f(x)\}^2] = O(n^{-4/5})$$

- Earlier theory implies  $\hat{f}_{\hat{w}}(x)$  mimics  $\hat{f}_{w_0}(x)$  asymptotically.

**Result** – deconvolution estimator retains usual rate:

$$\mathbb{E}[\{\hat{f}_{\hat{w}}(x) - f(x)\}^2] = O(n^{-4/5})$$

# But I Thought You Said . . .

- Recall that best rate for nonparametric deconvolution with Gaussian error is  $O(1/\log(n))$  (or powers thereof).
- How come weighted estimator so good?
- Poor rates for density deconvolution are driven by some hard to estimate densities e.g. with lots of small wiggles.
- Assumption of parametric form for  $\alpha$  deals with such problems.

# Bandwidth Selection

- Practical performance of  $\hat{f}_{\hat{w}}(x)$  dependent on bandwidth  $h$  (controls amount of smoothing).
- Recall: choice of weights:
  - does not depend on  $h$  for the KL method;
  - does depend on  $h$  for the ISE approach.
- For ISE approach, have tried *ad hoc* solution: apply off-the-shelf bandwidth selection (e.g. Sheather-Jones method).

# Bandwidth Selection for KL Method

- Usual approach is to seek to minimize

$$\begin{aligned}\text{MISE}\{f_{\hat{w}}\} &= \mathbb{E} \int \{f_{\hat{w}}(x; h) - f(x)\}^2 dx \\ &= \mathbb{E} \int \{f_{\hat{w}}(x; h)\}^2 dx - 2\mathbb{E}[f_{\hat{w}}(Y; h)w_0(Y)] + \int f^2\end{aligned}$$

- Can use **leave-one-out (cross-validation) approach**.
- Choose  $h$  to minimize

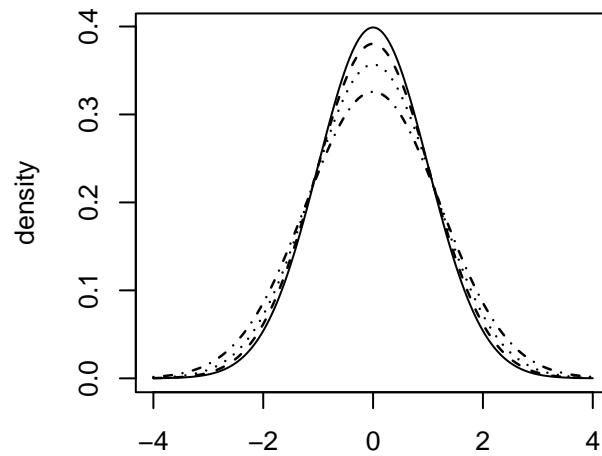
$$\Delta(h) = \int \{f_{\hat{w}}(x; h)\}^2 dx - \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \hat{w}_i \hat{w}_j K_h(Y_i - Y_j).$$

# Simulation Study

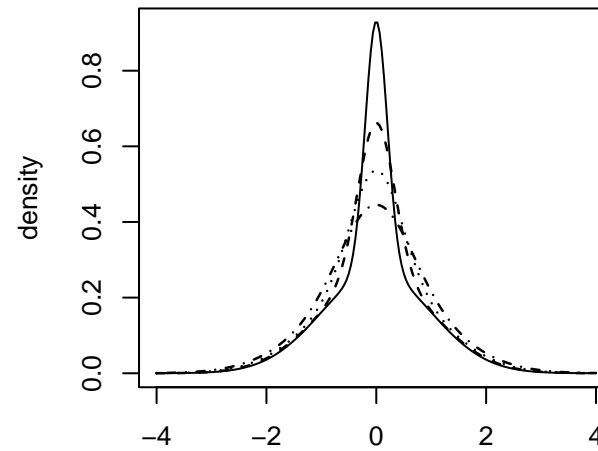
- Four target densities
- Three levels of measurement error (low, medium and high values for standard deviation of  $Z$ )
- Two sample sizes:  $n = 100$ ,  $n = 400$ .
- Two estimation methods:
  - Classical kernel deconvolution, CL;
  - Weighted kernel deconvolution, WT.
- For each combination, 400 data sets generated and integrated squared error (ISE) computed from each density estimate.

# Test Densities

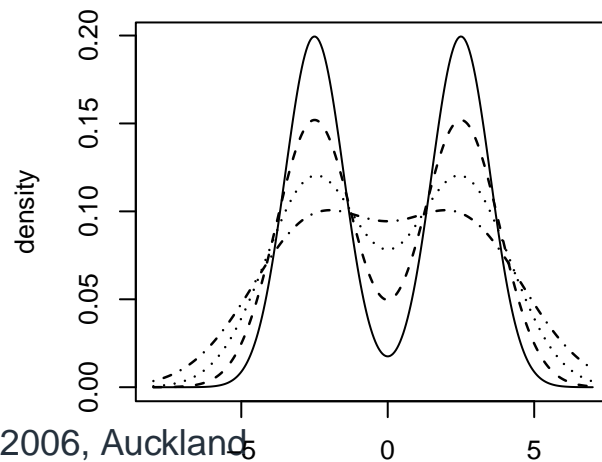
Density 1



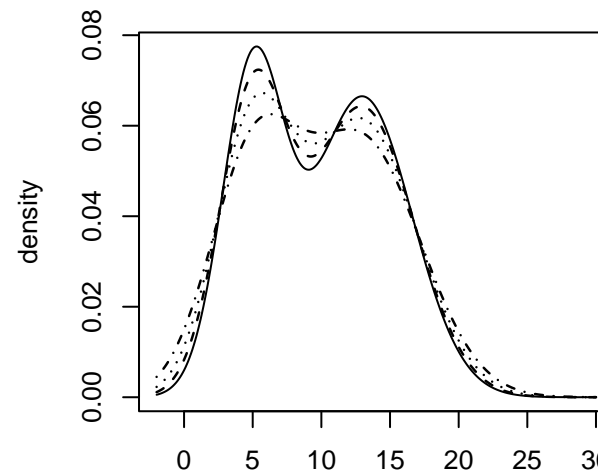
Density 2



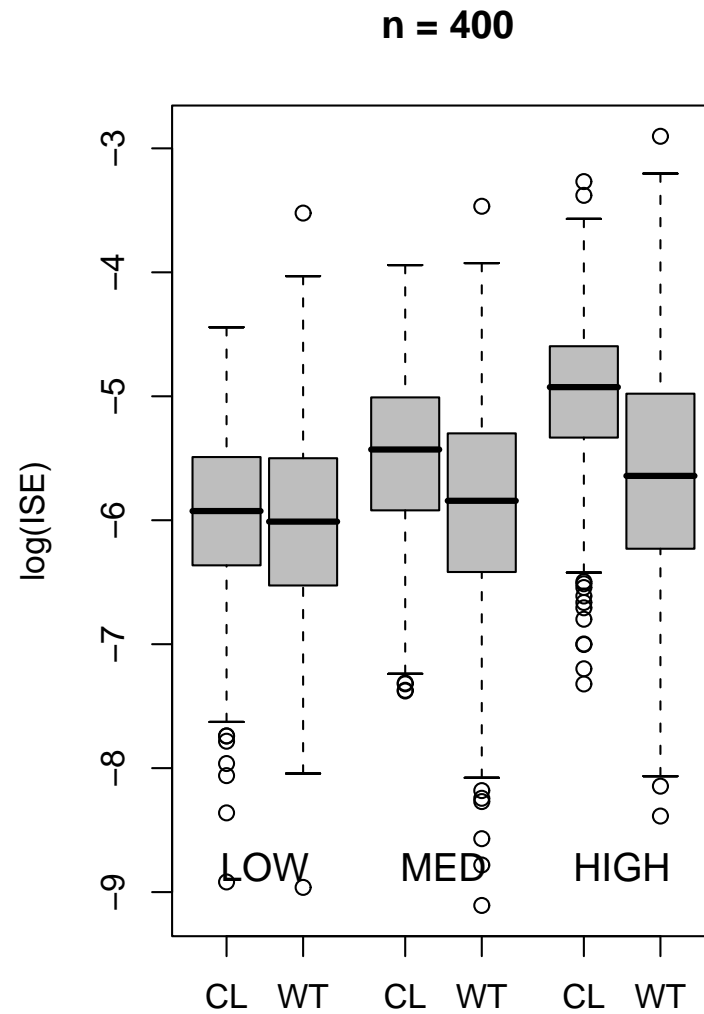
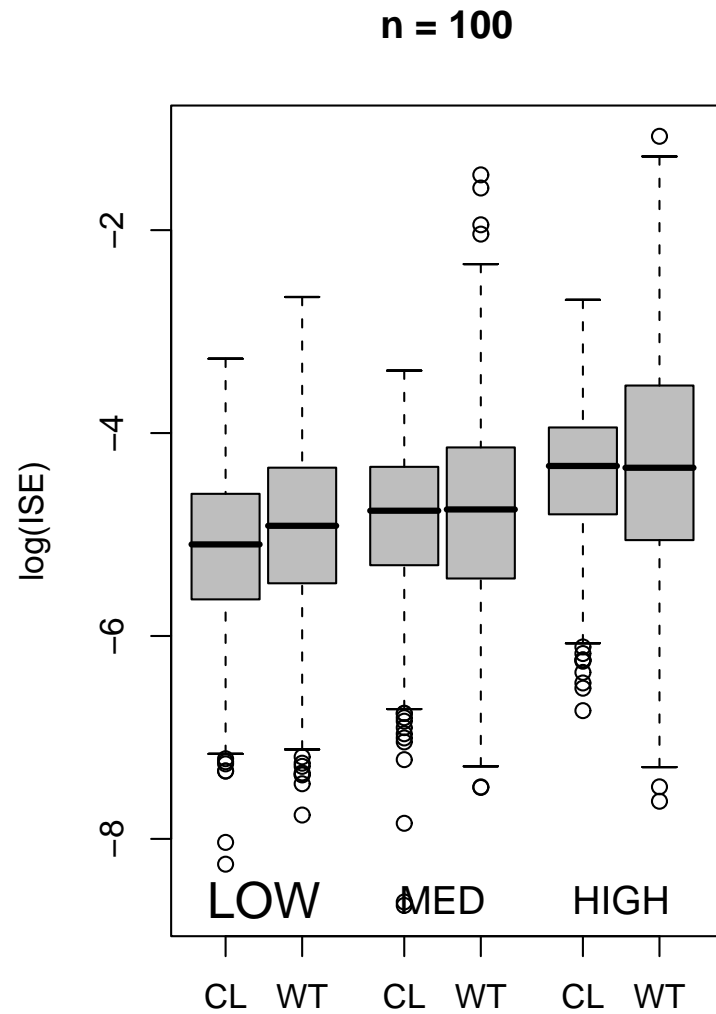
Density 3



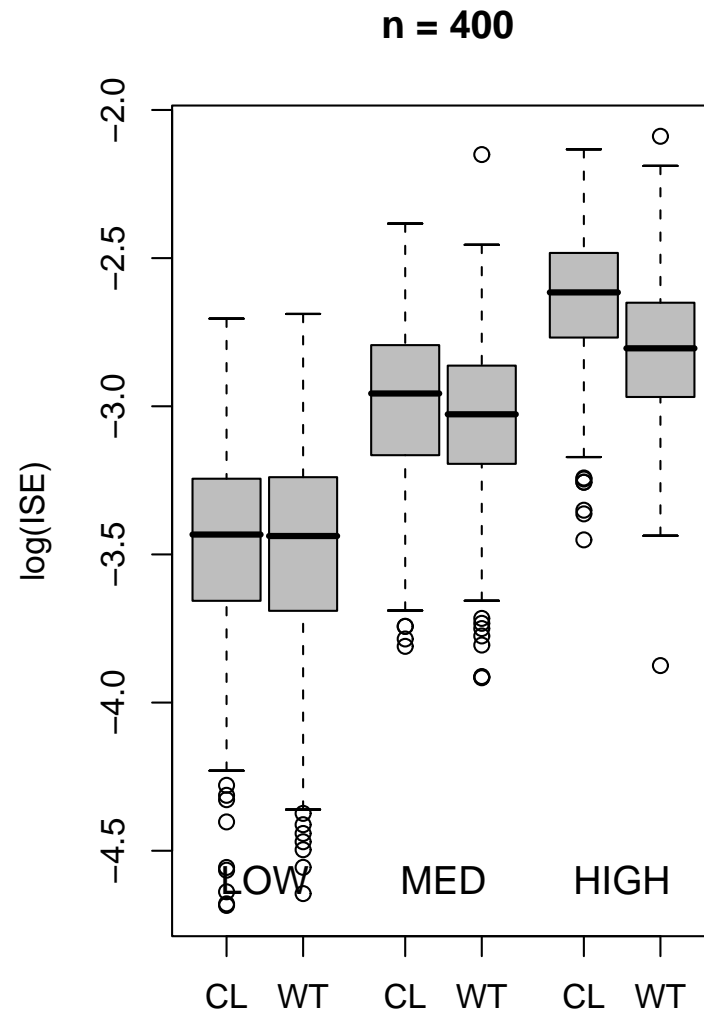
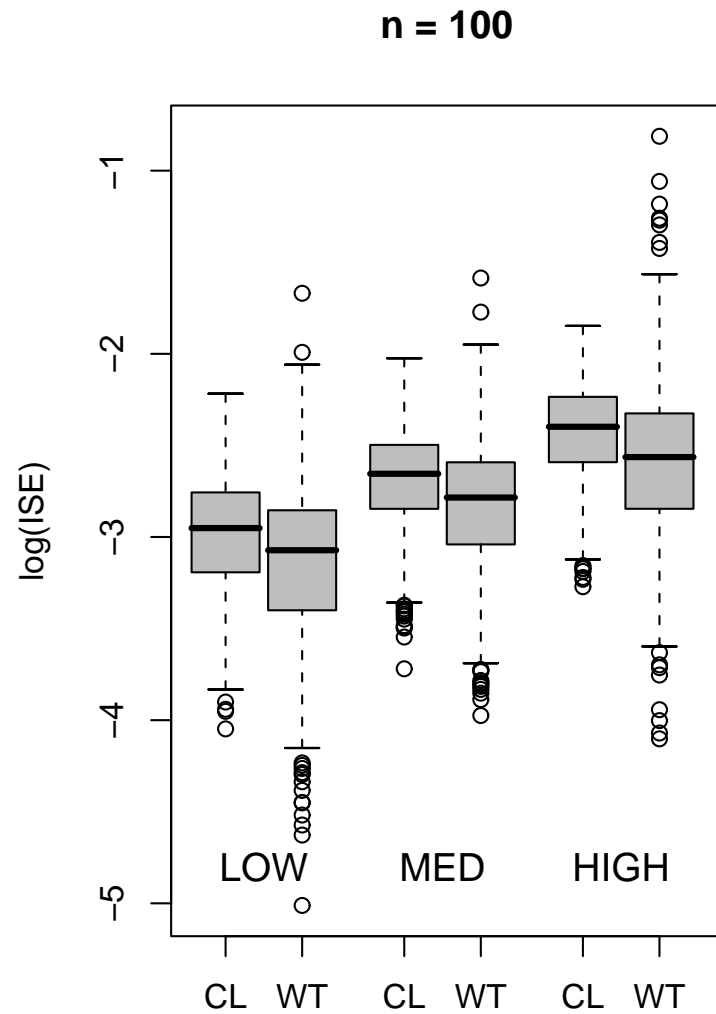
Density 4



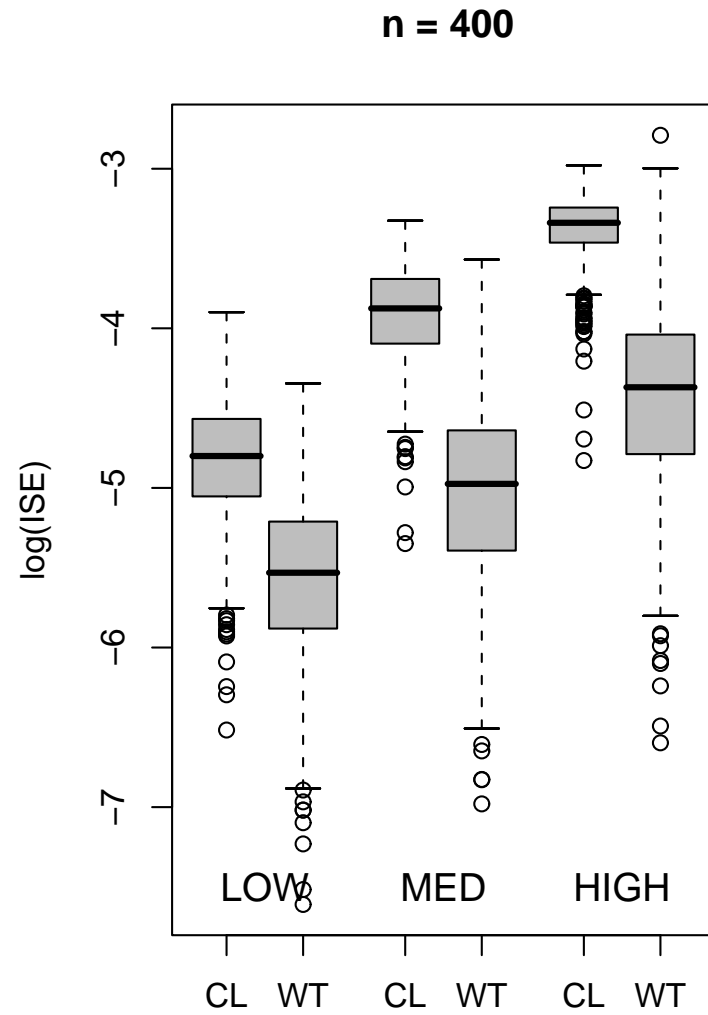
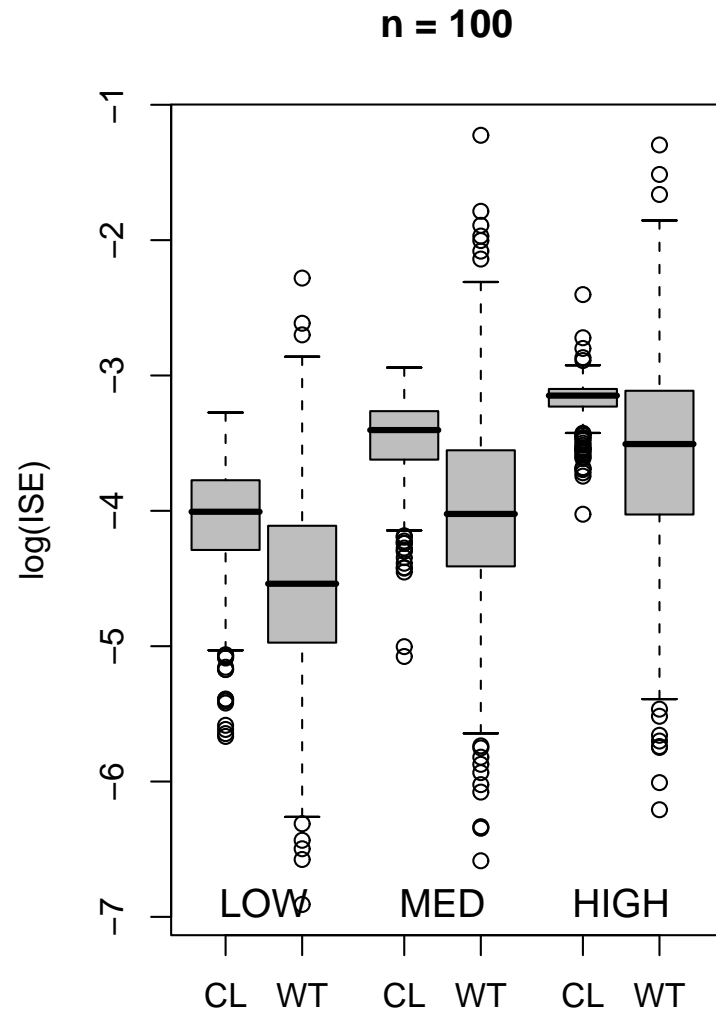
# Results for Density 1



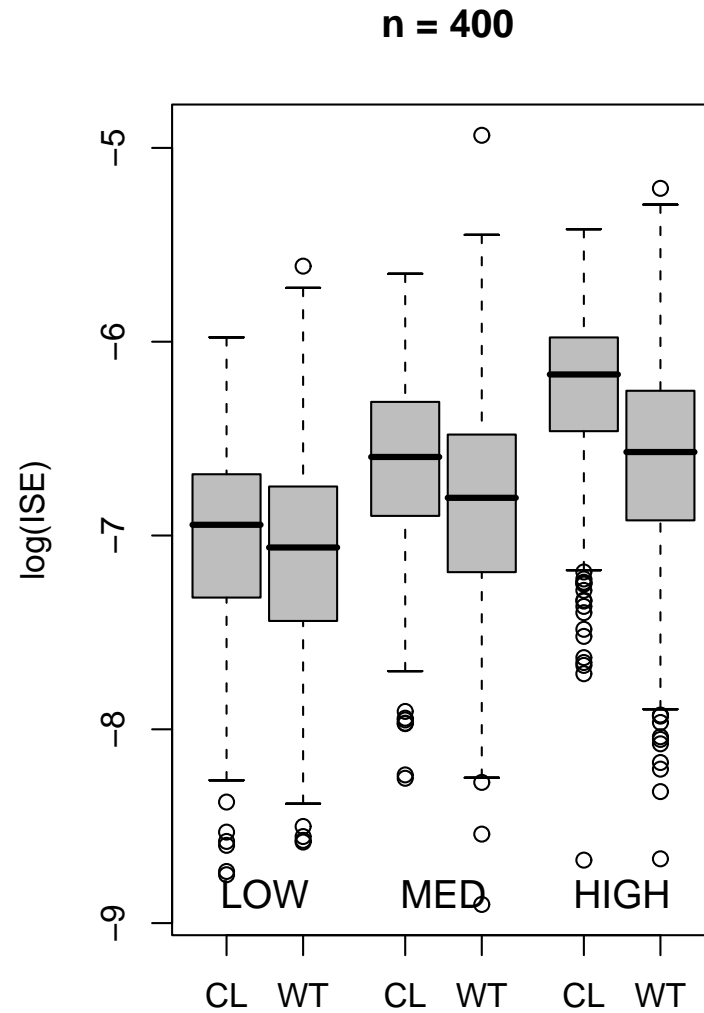
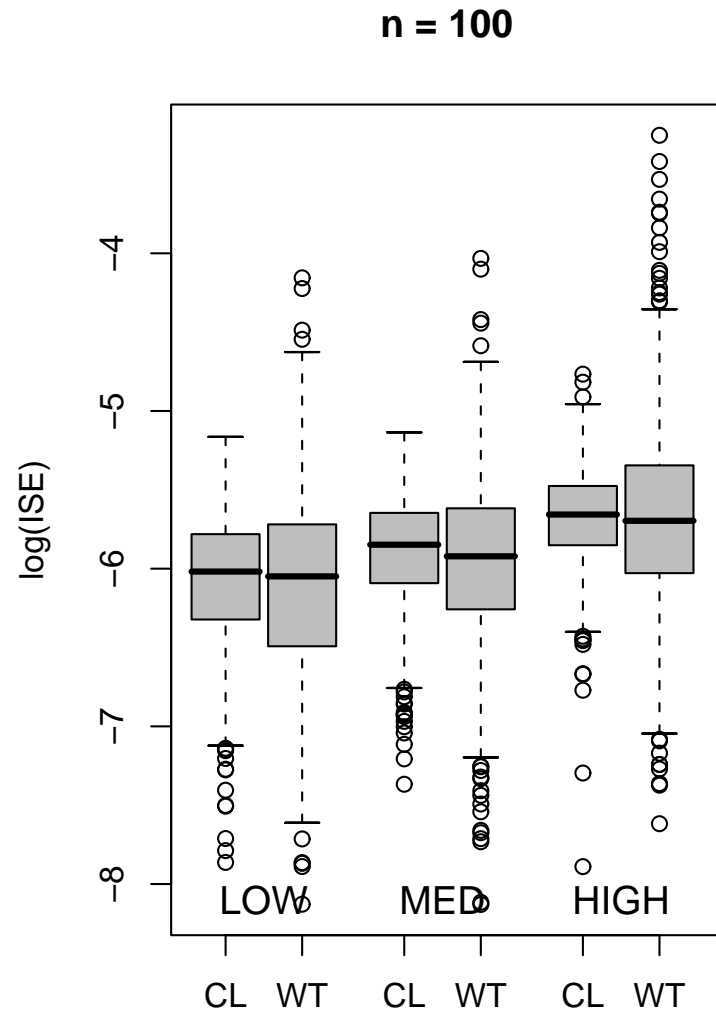
# Results for Density 2



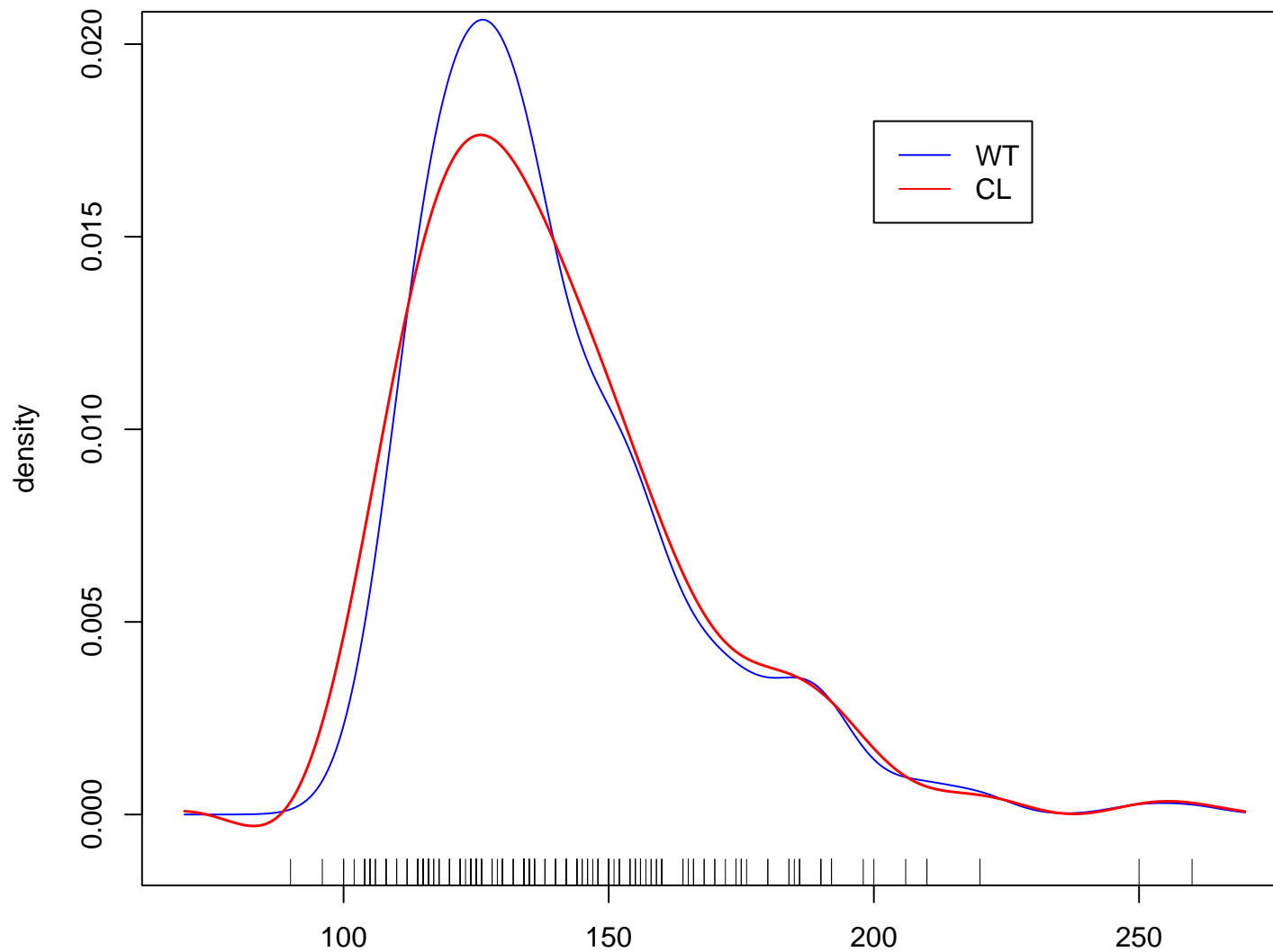
# Results for Density 3



# Results for Density 4

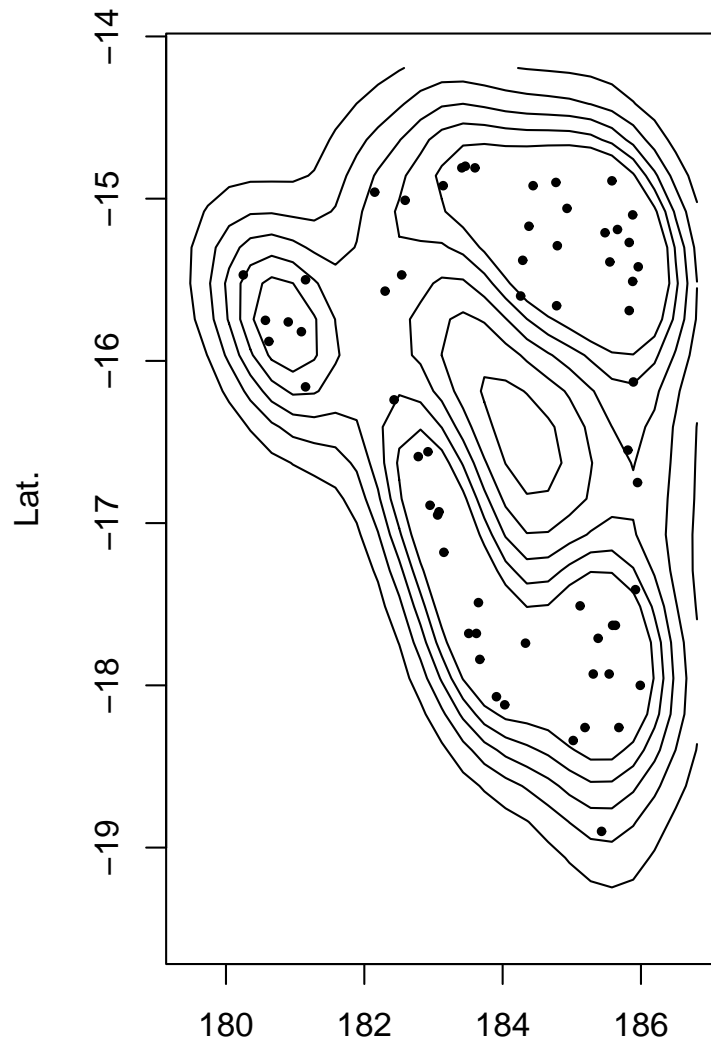


# Deconvolution for Blood Pressure Data

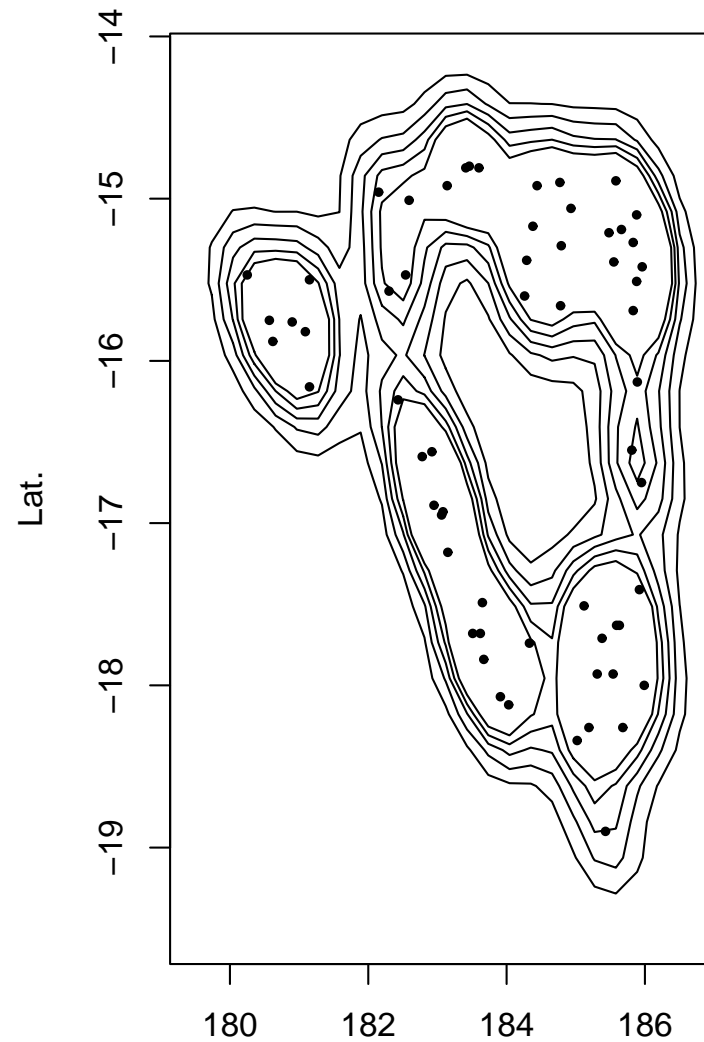


# Deconvolution for Earthquake Data

**Convolved density**



**Deconvolved density**



# Parting Thoughts

- Weighted kernel estimators show considerable promise for density deconvolution.
- Theoretical and numerical analyses are encouraging.
- Issues regarding implementation remain; e.g.
  - Knot selection for spline weight function;
  - Improved bandwidth selection;