# Depth Regression for Functional Data

Joydeep Chowdhury and Probal Chaudhuri

Indian Statistical Institute, Kolkata

## Quantitative Methods for Drug Discovery and Development

## Institute of Mathematical Sciences
## National University of Singapore

July 12, 2017

# Why Depth Regression for Functional Data?

## Tecator Data

- This dataset is available in the *R*-package *caret*.
- It contains the percentage values of moisture, fat and protein contents and the spectrum of absorbances for 215 meat samples.
- The moisture, the fat and the protein contents are measured by analytical chemistry.
- The absorbance spectrum of a sample was measured by a spectrometer.

# Why Depth Regression for Functional Data? (contd.)

- Obtaining the spectrum of a sample is cost-efficient, but getting the nutritional values is expensive.
- It is economically important to be able to predict the fat and the protein contents from the absorbance spectrum of a sample.
- We consider the spectrum as a functional covariate (random element in an $L_2$ space) and the fat and the protein contents as response variables. The response may be viewed as real-valued (if we analyze the fat or the protein content individually), or considered as bivariate (if the fat and the protein contents are analyzed simultaneously).

# Why Depth Regression for Functional Data? (contd.)

- We construct local boxplots for both the protein content and the fat content, taking the curve of absorbance spectrum as the covariate. The radius of the neighborhoods of each covariate curve is fixed at 0.25.

# Why Depth Regression for Functional Data? (contd.)

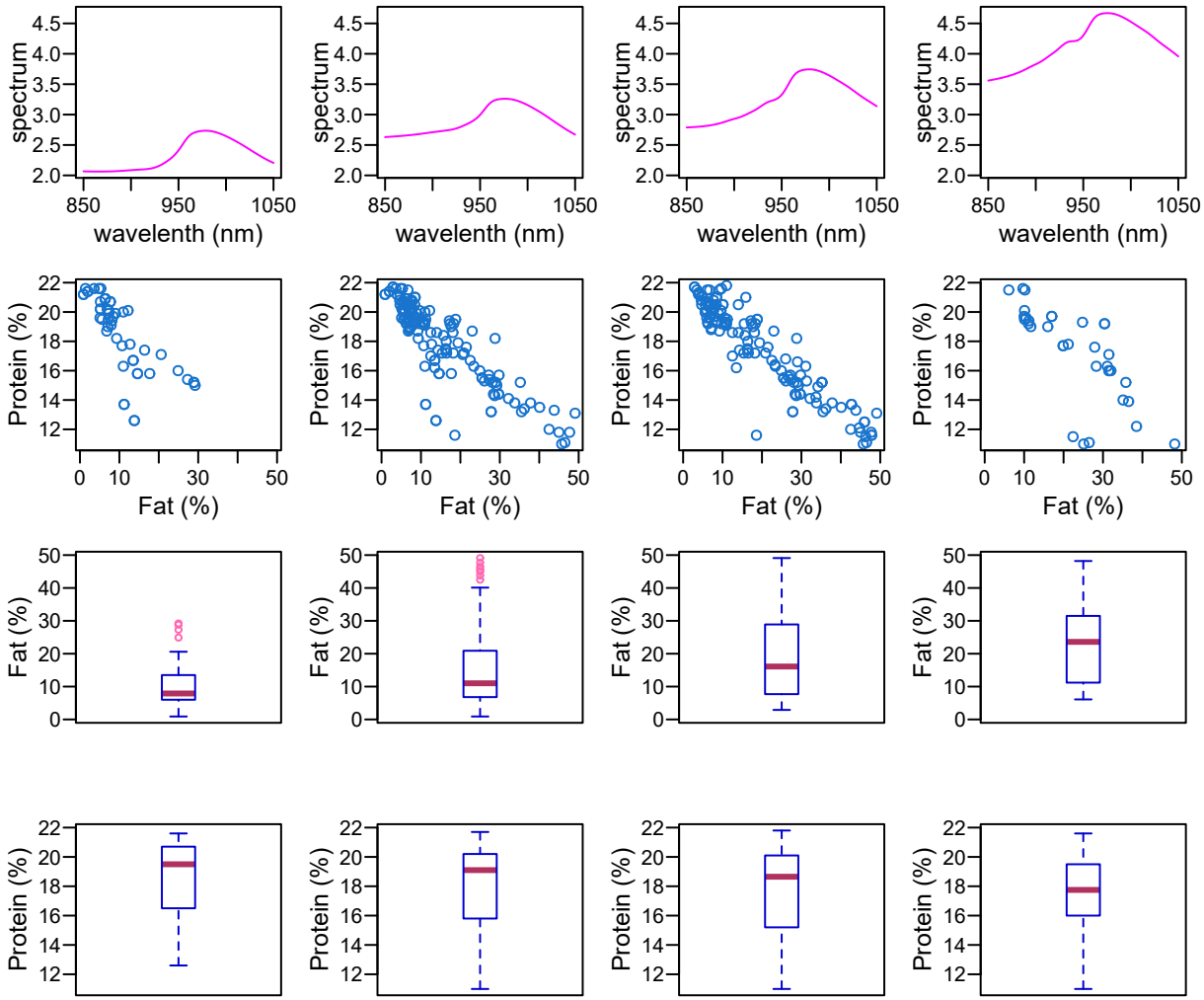



Figure: Local boxplots for the fat and the protein contents of the Tecator data.

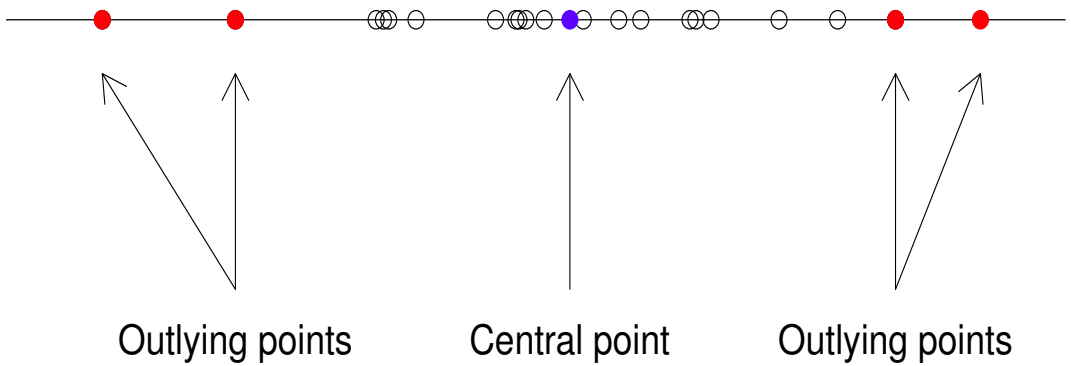
# Why Depth Regression for Functional Data? (contd.)

- Though the local median regression detects the change in the center of the conditional distribution with the change in the covariate, it misses some other important features of the conditional distributions like the variation in the conditional spread of the response.
- The upper and the lower boundaries of the boxplots, which are the local first and the third quartiles respectively, provide an idea about the changes in the conditional spread of the response with the change in the covariate.

# Why Depth Regression for Functional Data? (contd.)

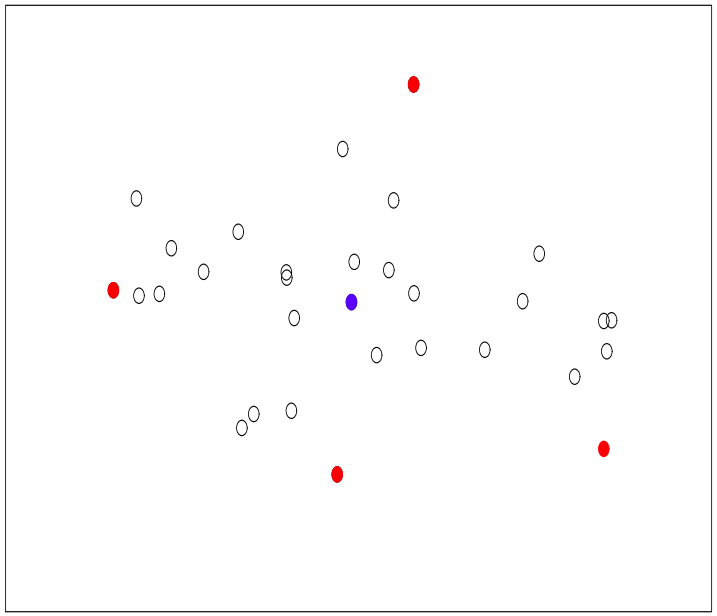
- The two variables are correlated, but the boxplots of each individual variable cannot capture the dependence.

# Data depth: Introduction



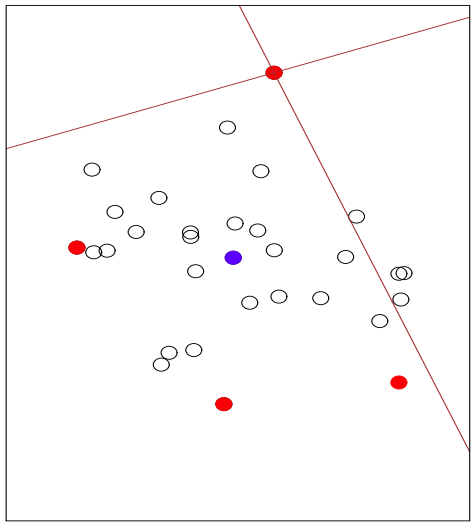

- Data depth gives a center-outward ordering of the points relative to the data cloud.
- The blue point has almost equal number of points on its either sides, while the red points have almost all the observations only on one side.
- Depth of a point $u = \min\{F_n(u), 1 - F_n(u)\}$, where $F_n(\cdot)$ is the proportion of data points that are on the left of $u$ (the empirical distribution function).

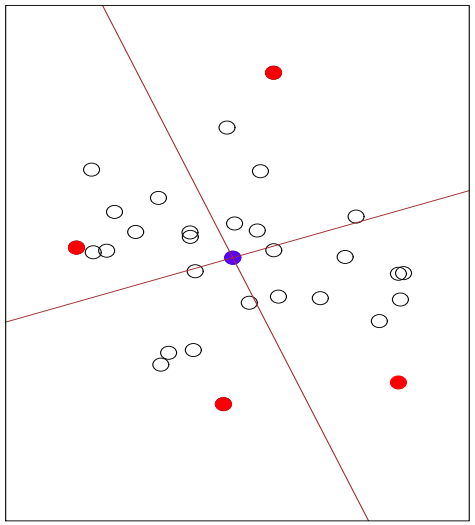
# Data depth: Introduction (contd.)



- How do we extend these ideas in $\mathbb{R}^d$ for $d \geq 2$ ?
- One approach is to consider lines through $\mathbf{u}$ (planes for $d = 3$ and hyper-planes for $d \geq 3$) and look at the proportion of data points lying on the two sides (half-spaces) of the line.

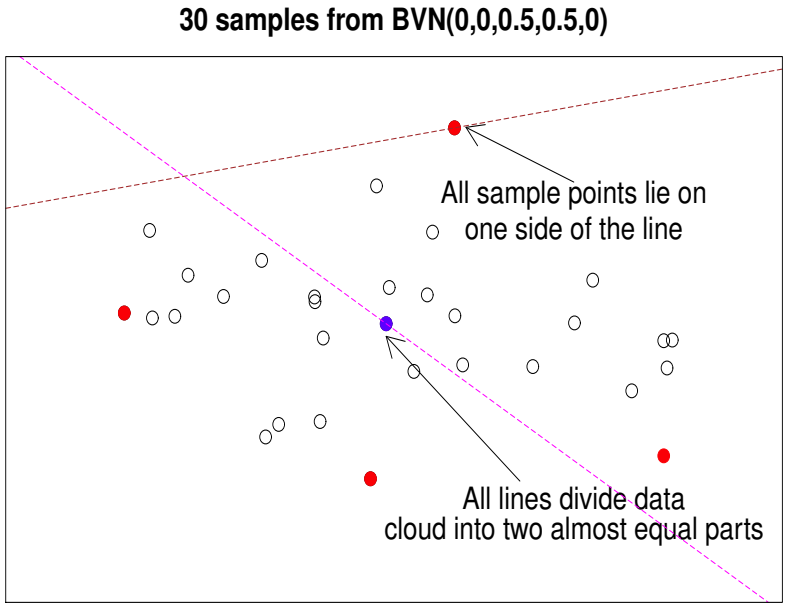
# Data depth: Introduction (contd.)





- We can take the minimum of the proportion of data points in any half-space of the line through $\mathbf{u}$ after considering all the possible lines through $\mathbf{u}$. This is called the Tukey half-space depth.
- All the data points lie on one half-space of one line through the red point, while any line through the blue point has almost equal proportion of data points in both the half-spaces.

# Data depth: Introduction (contd.)




- The half-space depth leads to a center-outward ordering of the points in $\mathbb{R}^d$ with respect to a given data cloud.

# Conditional Depth

- Let the response $\mathbf{Y} \in \mathbb{R}^p$ and the covariate $\mathbf{X} \in \mathcal{C}$, where $(\mathcal{C}, d)$ is a complete separable metric space.
- Let the conditional probability distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{z}$ be denoted as $\mu(\cdot \,|\, \mathbf{z})$, and $\mathbf{x} \in \mathcal{C}$ be a fixed element.
- Conditional Half-space Depth of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is defined as $\rho(\mathbf{y} \,|\, \mathbf{x}) = \inf\{\mu(\{\mathbf{v} \in \mathbb{R}^p \,|\, \mathbf{u}^t\mathbf{v} \geq \mathbf{u}^t\mathbf{y}\} \,|\, \mathbf{x}) \,|\, \mathbf{u} \in \mathbb{R}^p, \|\mathbf{u}\| = 1\}, \mathbf{y} \in \mathbb{R}^p$.

# Conditional Depth (contd.)

- Denote $D(\alpha \,|\, \mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^p \,|\, \rho(\mathbf{y} \,|\, \mathbf{x}) \geq \alpha\}$ for $\alpha \in \mathbb{R}$.
- For $0 \leq r < 1$, let $\alpha(r) = \sup\{\alpha \,|\, \mu(D(\alpha \,|\, \mathbf{x}) \,|\, \mathbf{x}) \geq r\}$.
- The set $D(\alpha(r) \,|\, \mathbf{x})$ is called the conditional $100r\%$ central region of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$.
- The conditional $100r\%$ central region contains $100r\%$ of the conditional probability mass with its elements having higher conditional depth than any point outside this set.

# Kernel estimates

- Let $K(\cdot)$ be a kernel function supported on $[0, 1]$, which is bounded and bounded away from 0, with associated bandwidth $h_n > 0$.
- Denote the sample conditional probability distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ as $\mu_n(\cdot \mid \mathbf{x})$, which puts mass

$$\frac{K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^{n} K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}$$

at the point $Y_i$.

# Kernel estimates (contd.)

- The conditional sample depth function $\rho_n(\cdot \,|\, \mathbf{x})$ is related to $\mu_n(\cdot \,|\, \mathbf{x})$ in the same way as the conditional population depth function $\rho(\cdot \,|\, \mathbf{x})$ is related to $\mu(\cdot \,|\, \mathbf{x})$.
- Denote $D_n(\alpha \,|\, \mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^p \,|\, \rho_n(\mathbf{y} \,|\, \mathbf{x}) \geq \alpha\}$, where $\alpha \in \mathbb{R}$.
- For $0 \leq r < 1$, let $\alpha_n(r) = \sup\{\alpha \,|\, \mu_n(D_n(\alpha \,|\, \mathbf{x}) \,|\, \mathbf{x}) \geq r\}$.
- The conditional sample $100r\%$ central region of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is $D_n(\alpha_n(r) \,|\, \mathbf{x})$.

# Asymptotic properties

- The conditional sample $100r\%$ maximal depth contour of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is defined as $\delta_n(\alpha_n(r) \,|\, \mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^p \,|\, \rho_n(\mathbf{y} \,|\, \mathbf{x}) = \alpha_n(r)\}$, where $0 < r < 1$.
- Under appropriate assumptions,
  - $D_n(\alpha_n(r) \,|\, \mathbf{x}) \to D(\alpha(r) \,|\, \mathbf{x})$ *almost surely* as $n \to \infty$ for any $0 < r < 1$.
  - Given any $\epsilon > 0$, $\delta_n(\alpha_n(r) \,|\, \mathbf{x}) \subseteq \{\mathbf{y} \in \mathbb{R}^p \,|\, \alpha(r) - \epsilon \leq \rho(\mathbf{y} \,|\, \mathbf{x}) < \alpha(r) + \epsilon\}$ *almost surely* for all sufficiently large $n$.
- The shapes of the sample central regions are good approximations of their population counter-parts for large sample sizes as the contours determine the shapes of the central regions.

# Center of the Conditional Distribution

- A point $m(\mathbf{x})$ with $\rho(m(\mathbf{x}) \,|\, \mathbf{x}) \geq \rho(\mathbf{y} \,|\, \mathbf{x})$ for every $\mathbf{y}$ is called a conditional median of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ with respect to the conditional depth $\rho(\cdot \,|\, \mathbf{x})$.
- For the the half-space depth, $m(\mathbf{x})$ becomes the usual conditional median for a univariate response, and $D(\alpha(r) \,|\, \mathbf{x})$ becomes the conditional interquartile interval for $r = 0.5$.

# Center of the Conditional Distribution (contd.)

- $m(\mathbf{x})$ along with the set $D(\alpha(r)\,|\,\mathbf{x})$ can be viewed as a generalization of the box-plot corresponding to the conditional distribution of a real valued response.
- A sample conditional median $m_n(\mathbf{x})$ satisfies $\rho_n(m_n(\mathbf{x})\,|\,\mathbf{x}) \geq \rho_n(\mathbf{y}\,|\,\mathbf{x})$ for every $\mathbf{y}$.

# Center of the Conditional Distribution (contd.)

- The conditional $100r\%$ trimmed mean $m(r \mid \mathbf{x})$ of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is defined as
  $m(r \mid \mathbf{x}) = [\int \mathbf{y} I(\mathbf{y} \in D(\alpha(1-r) \mid \mathbf{x}))\mu(d\mathbf{y} \mid \mathbf{x})]/\mu(D(\alpha(1-r) \mid \mathbf{x}) \mid \mathbf{x})$.
- For a real valued response, the above definition of the conditional trimmed mean coincides with the usual real valued conditional trimmed mean.
- The sample conditional $100r\%$ trimmed mean $m_n(r \mid \mathbf{x})$ is defined by
  $m_n(r \mid \mathbf{x})$
  $= [\int \mathbf{y} I(\mathbf{y} \in D_n(\alpha_n(1-r) \mid \mathbf{x}))\mu_n(d\mathbf{y} \mid \mathbf{x})]/\mu_n(D_n(\alpha_n(1-r) \mid \mathbf{x}) \mid \mathbf{x})$.

# Asymptotic Properties (contd.)

- Under appropriate assumptions,
  - Any sequence of sample deepest points $m_n(\mathbf{x})$ converges to a population deepest point $m(\mathbf{x})$ *almost surely* as $n \to \infty$.
  - $m_n(r \,|\, \mathbf{x}) \to m(r \,|\, \mathbf{x})$ *almost surely* as $n \to \infty$ for any $0 < r < 1$.

# Demonstration: Tecator Data

- We consider the pair of fat and protein contents as the bivariate response, and the absorbance spectra as the covariate.
- The response is bivariate and the covariate is functional.
- The kernel function $K(\cdot)$ used for the estimation is $K(u) = I(0 \leq u \leq 1)$. The bandwidth is computed by the leave-one-out cross validation method, minimizing the mean square error for the conditional median corresponding to the half-space depth.

# Demonstration: Tecator Data (contd.)

- We compute the 50% conditional central regions corresponding to four selected covariate curves ordered by their $L_2$-norms.
- The conditional median and the conditional trimmed mean are plotted as circles inside the central regions.

# Demonstration: Tecator Data (contd.)

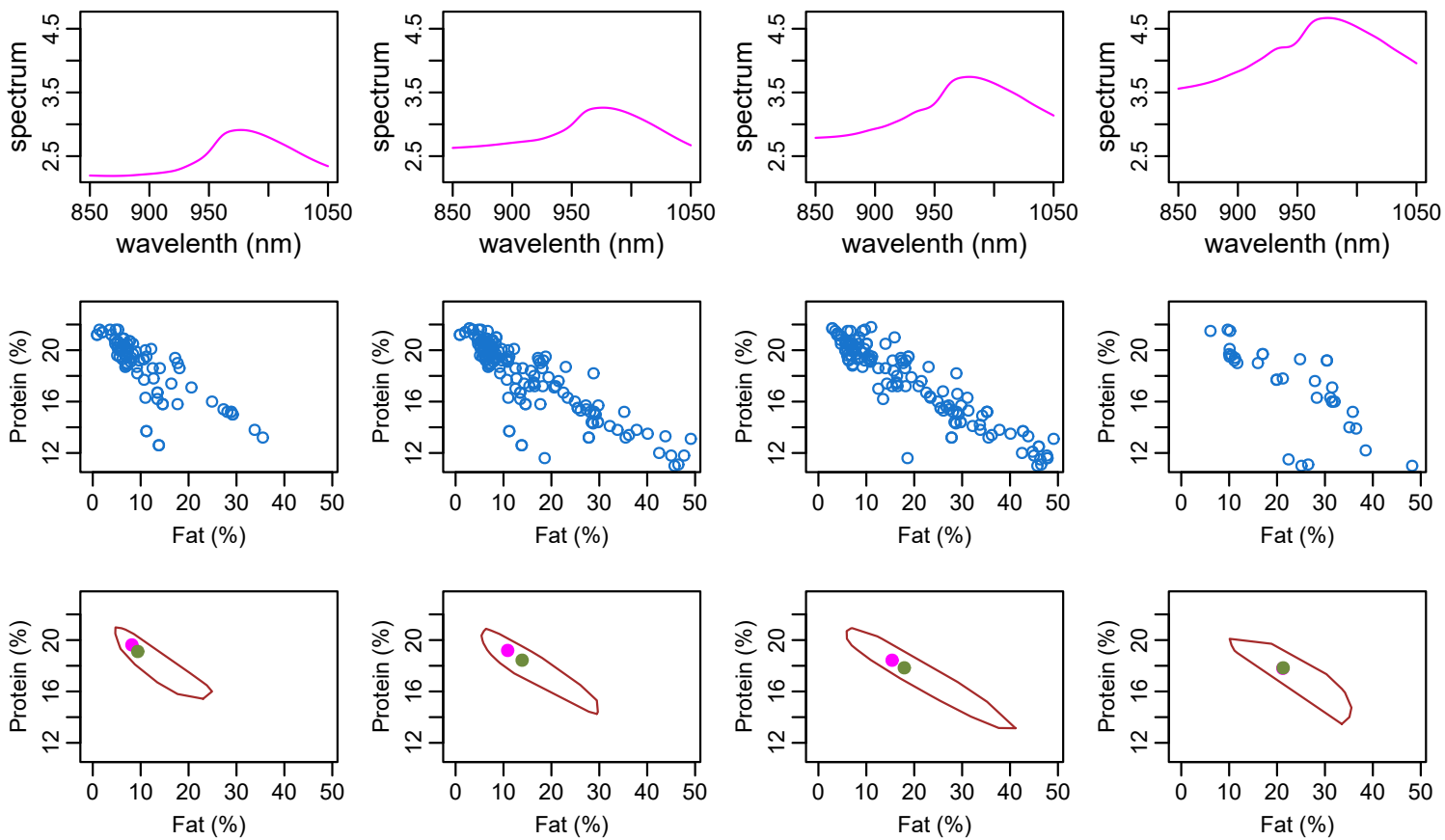



Figure: The selected covariate curves (1st row), scatter plots of the local response values (2nd row) and the conditional central regions for $r = 0.50$ (3rd row) for the Tecator Data with bivariate response.

# Conditional Spread

- A larger conditional central region indicates a higher spread of the conditional distribution of the response.
- The measure $\Delta(r \mid \mathbf{x})$ of conditional spread of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is defined as the diameter of the set $D(\alpha(r) \mid \mathbf{x})$.

# Conditional Spread (contd.)

- For a real valued response and for the half-space depth, $\Delta(r \mid \mathbf{x})$ coincides with a conditional inter-quantile range.
- In particular, $\Delta(0.5 \mid \mathbf{x})$ coincides with the conditional interquartile range of the real valued response.

# Conditional Spread (contd.)

- The estimate of $\Delta(r \mid \mathbf{x})$, denoted as $\Delta_n(r \mid \mathbf{x})$, is defined as $\Delta_n(r \mid \mathbf{x}) = \sup\{\|\mathbf{y}_1 - \mathbf{y}_2\| \mid \mathbf{y}_1, \mathbf{y}_2 \in D_n(\alpha_n(r) \mid \mathbf{x})\}$.
- Under appropriate conditions, $\Delta_n(r \mid \mathbf{x}) \to \Delta(r \mid \mathbf{x})$ *almost surely* as $n \to \infty$ for any $0 < r < 1$.

# Demonstration: Tecator Data (contd.)

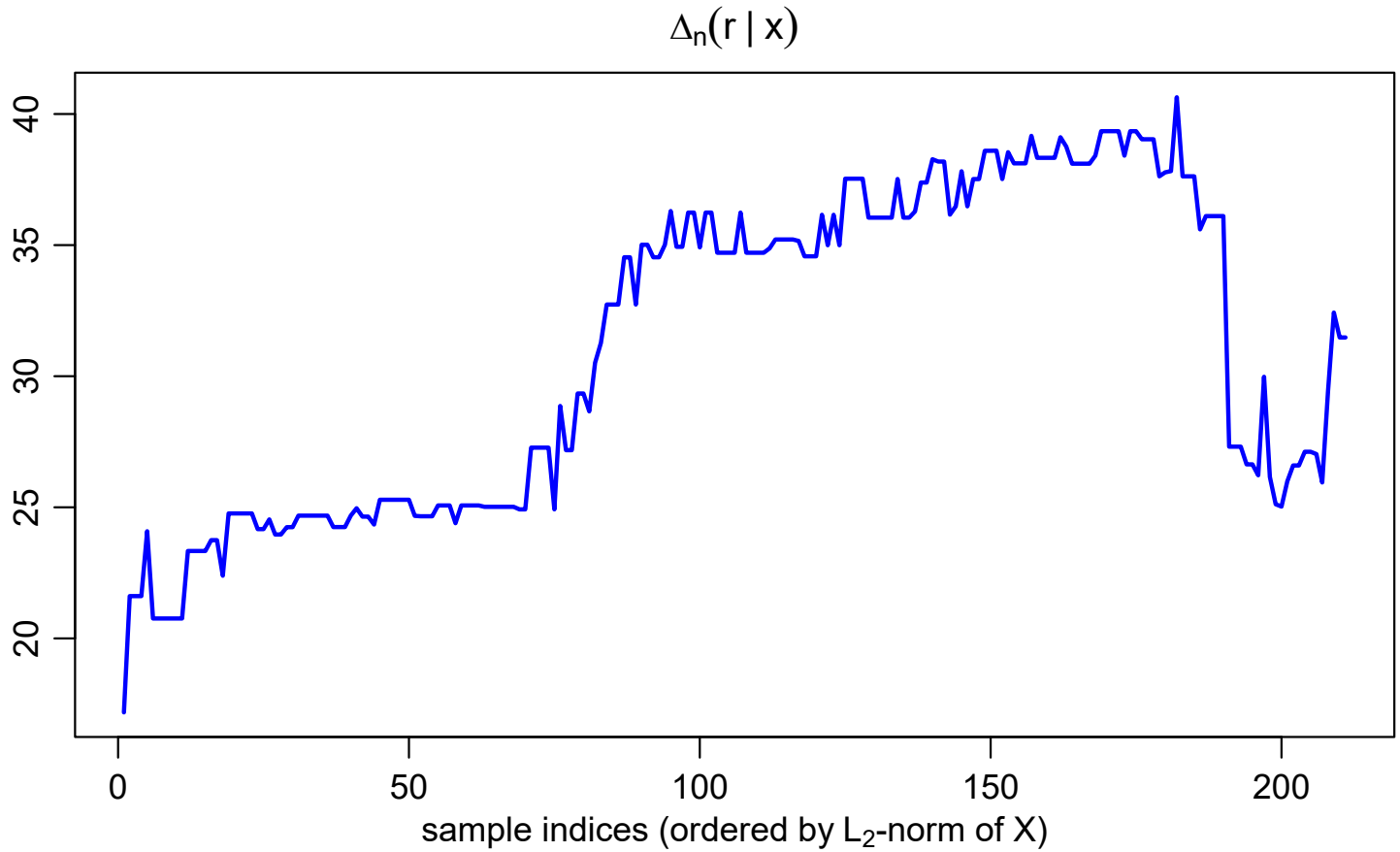



Figure: Plot of $\Delta_n(r \mid \mathbf{x})$ $r = 0.50$ for the Tecator Data with bivariate response (fat and protein contents).

# Test of Heteroscedasticity

- A nonparametric test for heteroscedasticity can be developed based on the conditional central regions.
- Our hypotheses are $H_0 : \Delta(r \,|\, \mathbf{x})$ is constant over $\mathbf{x}$, and $H_A : \Delta(r \,|\, \mathbf{x})$ varies with $\mathbf{x}$.

# Test of Heteroscedasticity (contd.)

- Recall that the estimate $\Delta_n(r \mid \mathbf{x})$ is the diameter of the conditional sample central region $D_n(\alpha_n(r) \mid \mathbf{x})$, and computationally expensive as a result.
- We consider a different estimate of $\Delta(r \mid \mathbf{x})$.

# Test of Heteroscedasticity (contd.)

- Define $\Delta'_n(r \mid \mathbf{x}) = \max\{\|\mathbf{Y}_i - \mathbf{Y}_j\| \mid \mathbf{Y}_i, \mathbf{Y}_j \in D_n(\alpha_n(r) \mid \mathbf{x})\}$.
- Under appropriate conditions, $\Delta'_n(r \mid \mathbf{x}) \to \Delta(r \mid \mathbf{x})$ *almost surely* as $n \to \infty$ for any $0 < r < 1$.

# Test of Heteroscedasticity (contd.)

- Our test-statistic is

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\left[\Delta_n'(r \mid X_i) - \left(\frac{1}{n}\sum_{j=1}^{n}\Delta_n'(r \mid X_j)\right)\right]^2 .$$

Large values of $T_n$ will bear evidence against $H_0$.

# Test of Heteroscedasticity (contd.)

- The p-value for the test is computed based on a permutation procedure.
- Step 1: Consider a permuted sample $(X_1, Y_1^*), ..., (X_n, Y_n^*)$, where $Y_1^*, \cdots, Y_n^*$ is a random permutation of $Y_1, \cdots, Y_n$.
- Step 2: The value of $T_n$ is computed for all such permuted samples, and the empirical distribution of those values is taken as an approximation of the null distribution of $T_n$.
- Step 3: The p-value for the test is computed as the proportion of those values of $T_n$ which are larger than the actually observed value of $T_n$.

# Test of Heteroscedasticity (contd.)

- In the Tecator data, the p-value for the proposed test, based on 500 random permutations, turns out to be 0 for the bivariate response (fat and protein contents) and 0.002 for the trivariate response (moisture, fat and protein contents).

# Power study

- We consider a functional covariate $X \in L_2[0,1]$ given by $X(t) = Be^t$, where $B \sim Uniform[0,1]$ and $t \in [0,1]$.
- Let $\Sigma_p$ denote a $p \times p$ matrix whose $(i,j)$-th element is $\sigma_{ij} = 0.5 + 0.5I(i=j)$
- The conditional distribution of $Y$ given $X$ is $MVN_p(0, (1 + a\|X\|_2)\Sigma_p)$, where $a \geq 0$.

# Power study (contd.)

- Each of the simulated level and power is computed based on 500 independent replications of the data.
- The nominal level of the test is fixed at 5%, i.e., we reject the null hypothesis when the computed p-value is less than 0.05.

# Power study (contd.)

Bivariate $Y$ $(p = 2)$

| sample size | $a = 0$ | $a = 5$ | $a = 10$ | $a = 15$ |
|---|---|---|---|---|
| $n = 100$ | 0.048 | 0.354 | 0.474 | 0.48 |
| $n = 200$ | 0.048 | 0.558 | 0.598 | 0.59 |
| $n = 400$ | 0.06 | 0.646 | 0.692 | 0.698 |

Trivariate $Y$ $(p = 3)$

| sample size | $a = 0$ | $a = 5$ | $a = 10$ | $a = 15$ |
|---|---|---|---|---|
| $n = 100$ | 0.044 | 0.378 | 0.476 | 0.516 |
| $n = 200$ | 0.056 | 0.514 | 0.558 | 0.602 |
| $n = 400$ | 0.044 | 0.626 | 0.656 | 0.686 |

# Conditional Skewness

- One can get an idea of the conditional skewness of the response by comparing the conditional $100r\%$ trimmed mean for some $0 < r < 1$ with the conditional median.
- If the conditional trimmed mean coincide with the conditional median, one may conclude that the conditional distribution is symmetric.

# Conditional Skewness: Tecator Data

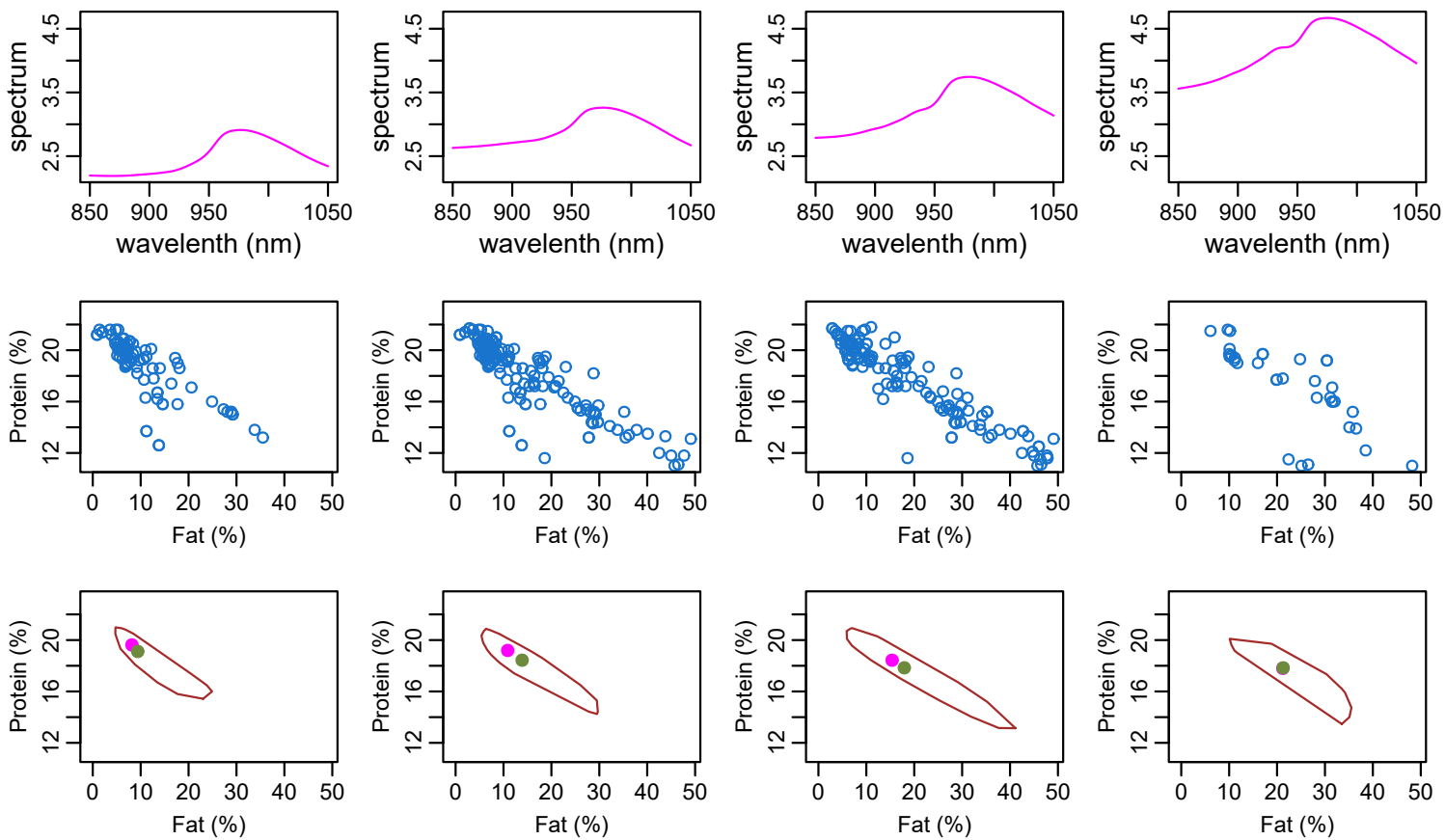



Figure: The selected covariate curves (1st row), scatter plots of the local response values (2nd row) and the conditional central regions for $r = 0.50$ (3rd row) for the Tecator Data with bivariate response.

# Conditional Skewness (contd.)

- Higher the distance between a conditional trimmed mean and the conditional median, $\|m(r \mid \mathbf{x}) - m(\mathbf{x})\|$, we can say that higher is the conditional skewness of the distribution.
- The distance between a conditional trimmed mean $m(r \mid \mathbf{x})$ and the conditional median $m(\mathbf{x})$ depends on the spread of the conditional distribution.
- We can take the quantity $\|m(r \mid \mathbf{x}) - m(\mathbf{x})\|$ scaled by a measure of conditional spread at $\mathbf{x}$ as a measure of conditional skewness of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$.

# Conditional Skewness (contd.)

- We define the measure $\Psi(r \mid \mathbf{x})$ of conditional skewness of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ as $\Psi(r \mid \mathbf{x}) = \|m(r \mid \mathbf{x}) - m(\mathbf{x})\|/\Delta(r \mid \mathbf{x})$.
- $\Psi(r \mid \mathbf{x})$ is estimated by the sample analogue $\Psi_n(r \mid \mathbf{x}) = \|m_n(r \mid \mathbf{x}) - m_n(\mathbf{x})\|/\Delta_n(r \mid \mathbf{x})$.
- Under appropriate conditions, $\Psi_n(r \mid \mathbf{x}) \to \Psi(r \mid \mathbf{x})$ *almost surely* as $n \to \infty$, for any $0 < r < 1$.

# Demonstration: Tecator Data (contd.)

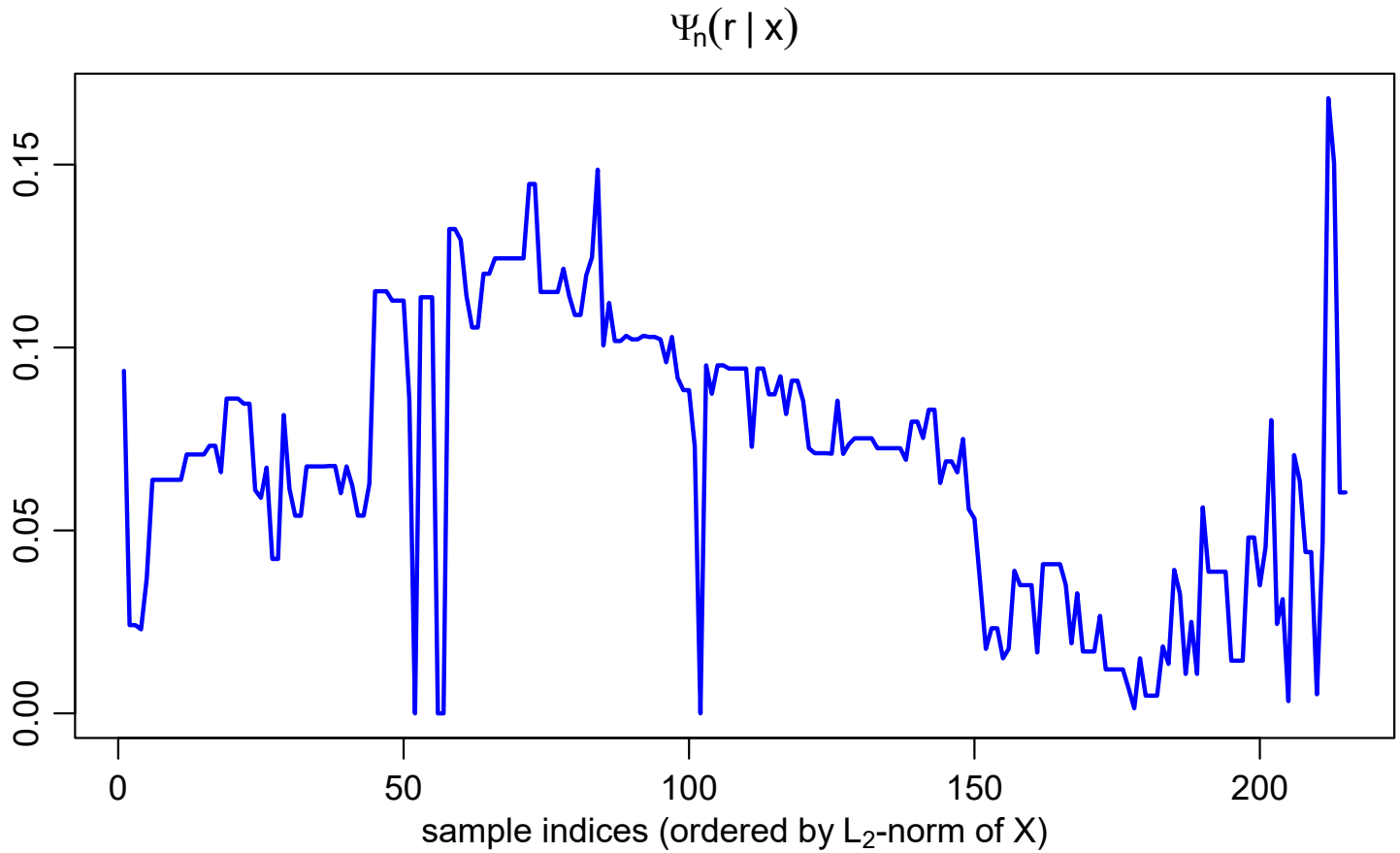



Figure: Plot of $\Psi_n(r \mid \mathbf{x})$ with $r = 0.50$ for the Tecator Data with bivariate response (fat and protein contents).

*Thank You*