

Semiparametric Mixtures of Regressions

David Hunter^{1,2}
Didier Chauveau³
Pierre Vandekerkhove⁴
Laurent Bordes⁵
Derek Young²

¹Le Studium, CNRS Orléans

²Penn State University, USA

³Université d'Orléans

⁴Université Paris-Est Marne-la-Vallée

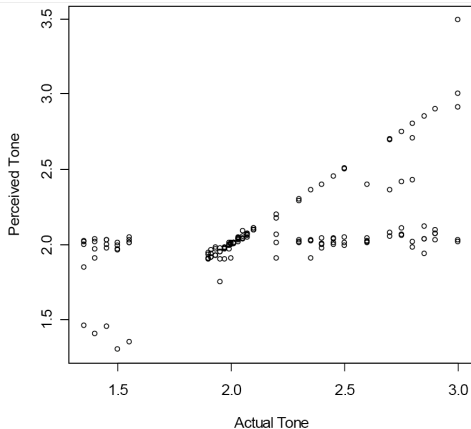
⁵Université de Pau

Pau, June 2008

Mixtures of regressions: A motivating example

Plot taken from PhD dissertation of Derek Young (2007):

- Cohen (1980 PhD dissertation) compared perceived tone and actual tone for a group of musicians.
- Overtones were played as well to try to confuse the musicians.
- Analyzed by DeVeaux (1989) and Viele and Tong (2002) via mixtures of linear regressions.
- Two groups hypothesized, but group variable unobserved.



Outline of talk

- 1 The basic mixture-of-regressions model and some extensions (including the semiparametric model)
- 2 Identifiability of the semiparametric model
- 3 An EM-like algorithm for estimation in the semiparametric model

Next topic. . .

- 1 The basic mixture-of-regressions model and some extensions (including the semiparametric model)
- 2 Identifiability of the semiparametric model
- 3 An EM-like algorithm for estimation in the semiparametric model

A fully parametric mixture of regressions model

Let $(\mathbf{X}_i, \mathbf{B}_i, \epsilon_i)$ be i.i.d., $i = 1, \dots, n$, where

- \mathbf{X}_i and \mathbf{B}_i are p -dimensional; ϵ_i is univariate
- \mathbf{X}_i , \mathbf{B}_i , and ϵ_i are jointly independent
- $\mathbf{B}_i \sim \sum_{j=1}^m \lambda_j \delta_{\beta_j}$ (and δ_{β_j} denotes the distribution concentrated at β_j)
- $\epsilon_i \sim N(0, \sigma^2)$

Then define $Y_i = \mathbf{X}_i^t \mathbf{B}_i + \epsilon_i$.

A fully parametric mixture of regressions model

Let $(\mathbf{X}_i, \mathbf{B}_i, \epsilon_i)$ be i.i.d., $i = 1, \dots, n$, where

- \mathbf{X}_i and \mathbf{B}_i are p -dimensional; ϵ_i is univariate
- \mathbf{X}_i , \mathbf{B}_i , and ϵ_i are jointly independent
- $\mathbf{B}_i \sim \sum_{j=1}^m \lambda_j \delta_{\beta_j}$ (and δ_{β_j} denotes the distribution concentrated at β_j)
- $\epsilon_i \sim N(0, \sigma^2)$

Then define $Y_i = \mathbf{X}_i^t \mathbf{B}_i + \epsilon_i$.

We observe all pairs (Y_i, X_i) and wish to estimate the parameters:

$$\lambda_1, \dots, \lambda_m, \beta_1, \dots, \beta_m, \sigma^2$$

NB: The `mixturetools` package includes functions for maximum likelihood estimation and Bayesian estimation for this model.

Generalization #1: Young and Hunter (2008)

Covariate-dependent mixing proportions

Let $(\mathbf{X}_i, \mathbf{B}_i, \epsilon_i)$ be i.i.d., $i = 1, \dots, n$, where

- \mathbf{X}_i and \mathbf{B}_i are p -dimensional; ϵ_i is univariate
- Do not assume \mathbf{X}_i , \mathbf{B}_i , and ϵ_i are jointly independent
- $\mathbf{B}_i | \mathbf{X}_i \sim \sum_{j=1}^m \lambda_j(\mathbf{X}_i) \delta_{\beta_j}$ (and δ_{β_j} denotes the distribution concentrated at β_j)
- $\epsilon_i \sim N(0, \sigma^2)$

Then define $Y_i = \mathbf{X}_i^t \mathbf{B}_i + \epsilon_i$.

Covariate-dependent mixing proportions

Let $(\mathbf{X}_i, \mathbf{B}_i, \epsilon_i)$ be i.i.d., $i = 1, \dots, n$, where

- \mathbf{X}_i and \mathbf{B}_i are p -dimensional; ϵ_i is univariate
- Do not assume \mathbf{X}_i , \mathbf{B}_i , and ϵ_i are jointly independent
- $\mathbf{B}_i | \mathbf{X}_i \sim \sum_{j=1}^m \lambda_j(\mathbf{X}_i) \delta_{\beta_j}$ (and δ_{β_j} denotes the distribution concentrated at β_j)
- $\epsilon_i \sim N(0, \sigma^2)$

Then define $Y_i = \mathbf{X}_i^t \mathbf{B}_i + \epsilon_i$.

If $\lambda_j(\mathbf{x})$ is a particular parametric function, we get the hierarchical mixtures of experts (HME) model of machine learning.

But one may use kernel methods to estimate $\lambda_j(\mathbf{x})$ nonparametrically.

Mixtures of local polynomial regressions

Let $(\mathbf{X}_i, J_i, \epsilon_i)$ be i.i.d., $i = 1, \dots, n$, where

- \mathbf{X}_i is p -dimensional; ϵ_i is univariate; $J_i \in \{1, \dots, m\}$
- \mathbf{X}_i , J_i , and ϵ_i are jointly independent
- $J_i \sim \sum_{j=1}^m \lambda_j \delta_j$ (and δ_j denotes the distribution concentrated at j)
- $\epsilon_i \sim N(0, \sigma^2)$

Then **define** $Y_i = f_{J_i}(\mathbf{X}_i) + \epsilon_i$, where f_1, \dots, f_m are unknown functions.

Mixtures of local polynomial regressions

Let $(\mathbf{X}_i, J_i, \epsilon_i)$ be i.i.d., $i = 1, \dots, n$, where

- \mathbf{X}_i is p -dimensional; ϵ_i is univariate; $J_i \in \{1, \dots, m\}$
- \mathbf{X}_i , J_i , and ϵ_i are jointly independent
- $J_i \sim \sum_{j=1}^m \lambda_j \delta_j$ (and δ_j denotes the distribution concentrated at j)
- $\epsilon_i \sim N(0, \sigma^2)$

Then **define** $Y_i = f_{J_i}(\mathbf{X}_i) + \epsilon_i$, where f_1, \dots, f_m are unknown functions.

Huang (2008) gives an EM algorithm for estimation of the λ_j and f_j using local likelihood (based on a local polynomial approximation to f_j).

Generalization #3: The semiparametric model

Unspecified error structure

Let $(\mathbf{X}_i, \mathbf{B}_i, \epsilon_i)$ be i.i.d., $i = 1, \dots, n$, where

- \mathbf{X}_i and \mathbf{B}_i are p -dimensional; ϵ_i is univariate
- \mathbf{X}_i , \mathbf{B}_i , and ϵ_i are jointly independent
- $\mathbf{B}_i \sim \sum_{j=1}^m \lambda_j \delta_{\beta_j}$ (and δ_{β_j} denotes the distribution concentrated at β_j)
- **Assume** $\epsilon_i \sim f$ for unknown f .

Then define $Y_i = \mathbf{X}_i^t \mathbf{B}_i + \epsilon_i$.

Generalization #3: The semiparametric model

Unspecified error structure

Let $(\mathbf{X}_i, \mathbf{B}_i, \epsilon_i)$ be i.i.d., $i = 1, \dots, n$, where

- \mathbf{X}_i and \mathbf{B}_i are p -dimensional; ϵ_i is univariate
- \mathbf{X}_i , \mathbf{B}_i , and ϵ_i are jointly independent
- $\mathbf{B}_i \sim \sum_{j=1}^m \lambda_j \delta_{\beta_j}$ (and δ_{β_j} denotes the distribution concentrated at β_j)
- **Assume $\epsilon_i \sim f$ for unknown f .**

Then define $Y_i = \mathbf{X}_i^t \mathbf{B}_i + \epsilon_i$.

This generalization will be the main focus of this talk.

In particular, the parameters of interest to us are

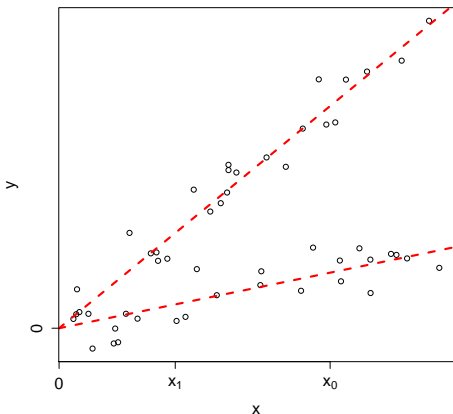
$$\lambda_1, \dots, \lambda_m, \beta_1, \dots, \beta_m, f$$

Next topic. . .

- 1 The basic mixture-of-regressions model and some extensions (including the semiparametric model)
- 2 Identifiability of the semiparametric model**
- 3 An EM-like algorithm for estimation in the semiparametric model

Identifiability: Intuition

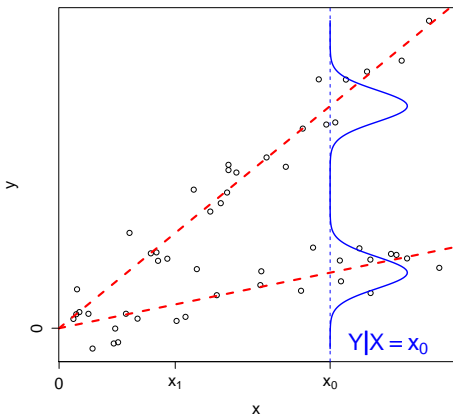
Consider the simplest case: Univariate x_j and $J \in \{1, 2\}$:



- $Y = X\beta_J + \epsilon$ where $\epsilon \sim f$
- Fix $X = x_0$.

Identifiability: Intuition

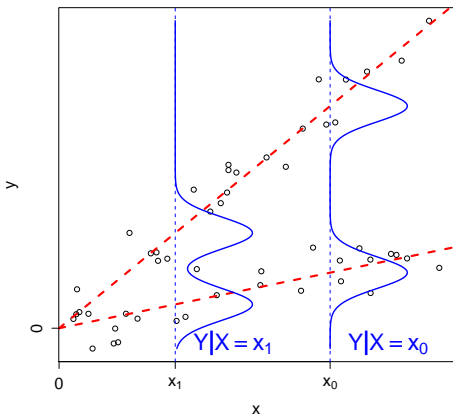
Consider the simplest case: Univariate x_i and $J \in \{1, 2\}$:



- $Y = X\beta_J + \epsilon$ where $\epsilon \sim f$
- Fix $X = x_0$.
- Conditional distribution of Y when $X = x_0$ not identifiable as mixture of shifted versions of f , even if f is assumed (say) symmetric.

Identifiability: Intuition

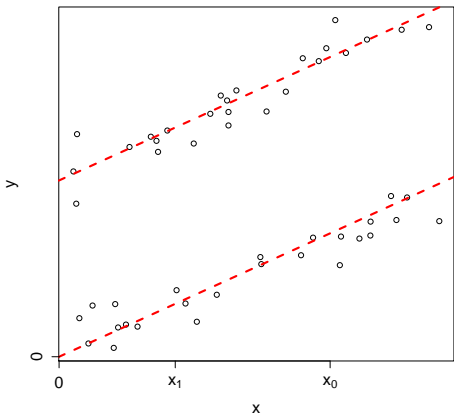
Consider the simplest case: Univariate x_i and $J \in \{1, 2\}$:



- $Y = X\beta_J + \epsilon$ where $\epsilon \sim f$
- Fix $X = x_0$.
- Conditional distribution of Y when $X = x_0$ not identifiable as mixture of shifted versions of f , even if f is assumed (say) symmetric.
- Identifiability depends on using additional X values that change the relative locations of the mixture components.

Mixtures of simple linear regressions

Next allow an intercept: $Y = \beta_{J1} + X\beta_{J2} + \epsilon$, with $\epsilon \sim f$.



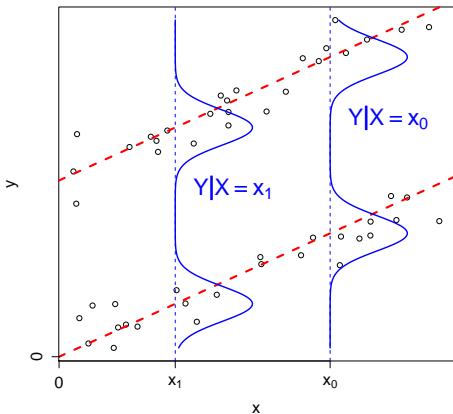
- First, note that for any K ,

$$Y = (\beta_{J1} + K) + X\beta_{J2} + (\epsilon - K)$$

so the model cannot be identifiable.

Mixtures of simple linear regressions

Next allow an intercept: $Y = \beta_{J1} + X\beta_{J2} + \epsilon$, with $\epsilon \sim f$.



- First, note that for any K ,

$$Y = (\beta_{J1} + K) + X\beta_{J2} + (\epsilon - K)$$

so the model cannot be identifiable.

- But even if f is assumed (say) symmetric about zero, identifiability can fail:
- Additional X values give no new information if the regression lines are parallel.

Two results

Denote the joint density by

$$\psi(\mathbf{x}, y) = h(\mathbf{x})g_{\mathbf{x}}(y) = h(\mathbf{x}) \sum_{j=1}^m \lambda_j f(y - \mathbf{x}^t \beta_j), \quad (1)$$

where h = marginal of \mathbf{X} , $g_{\mathbf{x}}$ = conditional of $Y|\mathbf{X} = \mathbf{x}$.

Two results

Denote the joint density by

$$\psi(\mathbf{x}, y) = h(\mathbf{x})g_{\mathbf{x}}(y) = h(\mathbf{x}) \sum_{j=1}^m \lambda_j f(y - \mathbf{x}^t \beta_j), \quad (1)$$

where h = marginal of \mathbf{X} , $g_{\mathbf{x}}$ = conditional of $Y|\mathbf{X} = \mathbf{x}$.

Theorem

If the support of \mathbf{X} contains an open set in \mathbb{R}^p , then model (1) is identifiable.

Corollary: Regression with an intercept

If the support of \mathbf{X} contains an open subset in $1 \times \mathbb{R}^{p-1}$, then model (1) is identifiable as long as no two of the regression surfaces $y = \mathbf{x}^t \beta_j$ are parallel; i.e., as long as no two vectors $(\beta_{j2}, \dots, \beta_{jp}) \in \mathbb{R}^{p-1}$ are equal.

Two results

Denote the joint density by

$$\psi(\mathbf{x}, y) = h(\mathbf{x})g_{\mathbf{x}}(y) = h(\mathbf{x}) \sum_{j=1}^m \lambda_j f(y - \mathbf{x}^t \beta_j), \quad (1)$$

where h = marginal of \mathbf{X} , $g_{\mathbf{x}}$ = conditional of $Y|\mathbf{X} = \mathbf{x}$.

Theorem

If the support of \mathbf{X} contains an open set in \mathbb{R}^p , then model (1) is identifiable.

Corollary: Regression with an intercept

If the support of \mathbf{X} contains an open subset in $1 \times \mathbb{R}^{p-1}$, then model (1) is identifiable as long as no two of the regression surfaces $y = \mathbf{x}^t \beta_j$ are parallel; i.e., as long as no two vectors $(\beta_{j2}, \dots, \beta_{jp}) \in \mathbb{R}^{p-1}$ are equal. \Rightarrow “Generic Identifiability”

Next topic. . .

- 1 The basic mixture-of-regressions model and some extensions (including the semiparametric model)
- 2 Identifiability of the semiparametric model
- 3 An EM-like algorithm for estimation in the semiparametric model**

Notation

Recall the model:

$$Y|\mathbf{X} \sim \sum_{j=1}^m \lambda_j f(y - \mathbf{X}^t \beta_j).$$

(Also, $\mathbf{X} \sim h$ but this fact can be ignored for now.)

Data are $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$.

Recall the model:

$$Y|\mathbf{X} \sim \sum_{j=1}^m \lambda_j f(y - \mathbf{X}^t \beta_j).$$

(Also, $\mathbf{X} \sim h$ but this fact can be ignored for now.)

Data are $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$.

Define the **unobserved data** Z_{ij} , $1 \leq i \leq n$ and $1 \leq j \leq m$, to be the indicator that the i th observation comes from the j th mixture component.

That is,

$$Z_{ij} = I\{Y_i|\mathbf{X}_i \sim f(y - \mathbf{X}_i^t \beta_j)\}.$$

The E-step

The mixture-of-regressions EM-like algorithm is a generalization of a similar algorithm studied by Bordes, Chauveau, and Vandekerckhove (2007) and Benaglia, Chauveau, and Hunter (2008).

- The E-step consists of finding the “posterior” probabilities

$$\begin{aligned} p_{ij} &\stackrel{\text{def}}{=} P(Z_{ij} = 1 | \text{data, starting parameter values}) \\ &= \frac{\lambda_j^0 f^0(y_i - \mathbf{x}_i^t \beta_j^0)}{\sum_{\ell=1}^m \lambda_\ell^0 f^0(y_i - \mathbf{x}_i^t \beta_\ell^0)}. \end{aligned}$$

- NB: The “starting parameter values” are

$$\lambda_1^0, \dots, \lambda_m^0, \beta_1^0, \dots, \beta_m^0, f^0$$

The M-step

In the M-step, the Euclidean parameters are updated.

- As usual, each λ_j is the mean of the corresponding p_{ij} :

$$\lambda_j^{(\text{new})} = \frac{1}{n} \sum_{i=1}^n p_{ij}$$

- For β_j , in an ordinary EM algorithm we maximize a likelihood but in this case, there is not really a likelihood and therefore there is no obvious choice! Possibilities include:

$$\textcircled{1} \quad \beta_j^{(\text{new})} = \arg \min_{\beta} \sum_{i=1}^n p_{ij} (y_i - \mathbf{x}_i^t \beta)^2$$

$$\textcircled{2} \quad \beta_j^{(\text{new})} = \arg \min_{\beta} \sum_{i=1}^n p_{ij} |y_i - \mathbf{x}_i^t \beta|$$

$$\textcircled{3} \quad \beta_j^{(\text{new})} = \arg \max_{\beta} \sum_{i=1}^n p_{ij} f^0(y_i - \mathbf{x}_i^t \beta)$$

The density estimation step

We now employ a third step, a density estimation step.

For some bandwidth h and kernel density $K(\cdot)$, update the estimate of f :

$$f^{(\text{new})}(u) = \frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^m p_{ij} K\left(\frac{u - y_i + \mathbf{x}_i^t \beta_j}{h}\right)$$

The density estimation step cont'd

$$f^{(\text{new})}(u) = \frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^m p_{ij} K\left(\frac{u - y_i + \mathbf{x}_i^t \beta_j}{h}\right)$$

- It is possible to impose some constraints here, such as:
 - f must have zero mean or median
 - f must be symmetric about zero
- Because this step does not maximize a likelihood, the algorithm is not a true EM algorithm

The density estimation step cont'd

$$f^{(\text{new})}(u) = \frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^m z_{ij}^* K\left(\frac{u - y_i + \mathbf{x}_i^t \beta_j}{h}\right)$$

- It is possible to impose some constraints here, such as:
 - f must have zero mean or median
 - f must be symmetric about zero
- Because this step does not maximize a likelihood, the algorithm is not a true EM algorithm
- One may create a stochastic algorithm by replacing p_{ij} by a randomly simulated Bernoulli variable z_{ij}^* .

- How should one choose the bandwidth h ?
 - Should h be adaptively updated?
 - Should h be different for different components?

- How should one choose the bandwidth h ?
 - Should h be adaptively updated?
 - Should h be different for different components?
- How should β be updated?

- How should one choose the bandwidth h ?
 - Should h be adaptively updated?
 - Should h be different for different components?
- How should β be updated?
- In regression with an intercept, how detrimental is nonidentifiability?

Open questions

- How should one choose the bandwidth h ?
 - Should h be adaptively updated?
 - Should h be different for different components?
- How should β be updated?
- In regression with an intercept, how detrimental is nonidentifiability?
- Is there any sort of objective function that is being maximized by this algorithm?

Some references

- Bordes, L., Chauveau, D., and Vandekerckhove, P. (2007), An EM algorithm for a Semiparametric Mixture Model, *Computational Statistics and Data Analysis*, **51**: 5429–5443.
- Cohen, E. (1980), Inharmonic Tone Perception, PhD thesis, Stanford University, unpublished.
- DeVeaux, R. D. (1989), Mixtures of Linear Regressions, *Computational Statistics and Data Analysis*, **8**(3): 227–245.
- Viele, K. and Tong, B. (2002), Modeling with Mixtures of Regressions, *Statistics and Computing*, **12**(4): 315–330.
- Young, D. et al (2008), mixtools: Tools for Analyzing Finite Mixture Models, R package version 0.3.1