# ~~Finite Mixtures of Nonparametric Regression~~ Cluster-Weighted Models with Generalized Additive Components

Robert Zimmerman

Department of Statistics and Actuarial Science
University of Waterloo
Supervisor: Dr. Ryan P. Browne

CUMC 2016

# Outline

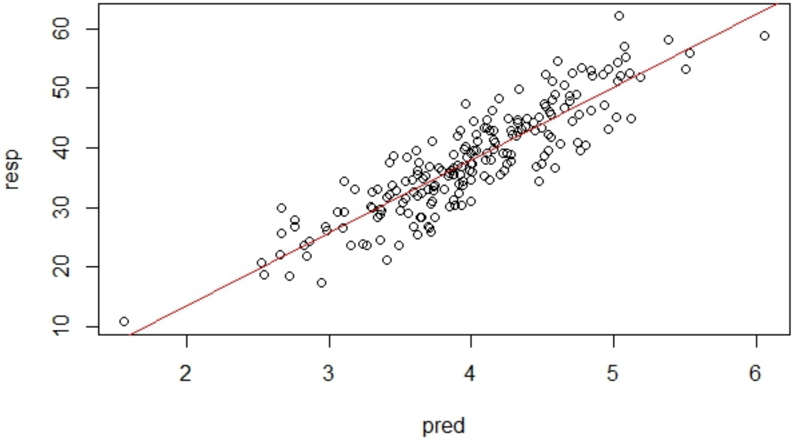
1. The Model Recipe
   - Review: Regression
   - Mixture Models
   - Cluster-Weighted Models
   - Nonparametric Models

# Regression Models

- Suppose we have a set of data $\{(y_i, \mathbf{x_i})\}_{i=1}^n$, where each $y_i \in \mathbb{R}$ is a response (believed to be) based on $\mathbf{x_i} \in \mathbb{R}^p$
- In simple regression, we assume that $Y = f(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}) + \epsilon$, where $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}$, $\mathbb{E}[\epsilon] = 0$, and $f : \mathbb{R} \to \mathbb{R}$ is a deterministic function
- Thus $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = f(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})$
    - When $f(x) = \text{Id}$, this is **linear regression**
    - When $Y$ is in the exponential family and $f(\cdot) = g^{-1}(\cdot)$ is a link function, we get a **generalized linear model**
    - Etc.

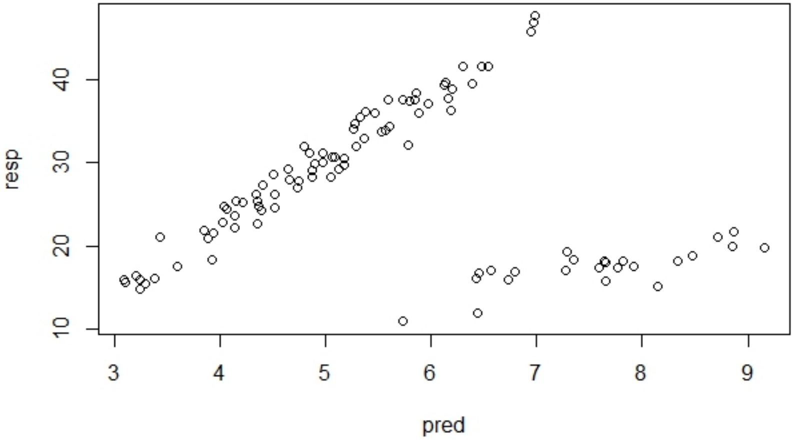
# Regression Models: Example

In linear regression using **ordinary least squares**, we estimate the coefficients $(\beta_{i0}, \boldsymbol{\beta}_i)$ of $\mathbb{E}[Y_i|\boldsymbol{X}_i] = \beta_{i0} + \boldsymbol{\beta}_i^T \mathbf{X_i}$ by minimizing the sum of the (squared) distances between the estimated hyperplane $\hat{y}_i$ and each data point $\mathbf{x_i}$, leading to the estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1} \mathbf{X}^T \mathbf{y}$.



# Mixture Models: Motivation

What if we find that our data set appears to partition into several distinct groups, or **clusters**?



# Finite Mixture Models

- Suppose $\theta$ is a discrete random variable whose distribution places mass on the elements of $\{1, 2, \ldots, G\}$, and suppose we have $G$ conditional random variables $\{X_g|\theta = g \sim F_g(x)\}_{g=1}^{G}$ which follow their own distinct distributions
- It is easily shown that $F(x) = \sum_{g=1}^{G} F_g(x) \cdot \mathbb{P}(\theta = g)$ defines a distribution function, which we call a **mixture distribution**
- Denoting each **mixing weight** $\pi_g := \mathbb{P}(\theta = g) \in [0, 1]$ and observing that $\sum_{g=1}^{G} \pi_g = 1$, we see that $F(x) = \sum_{g=1}^{G} \pi_g \cdot F_g(x)$ is simply a convex combination of distribution functions

# Finite Mixture Models: Example

Suppose we reach for one of two biased coins, $C_{big}$ and $C_{small}$, such that $\mathbb{P}(C_{big} = H) = 0.75$ and $\mathbb{P}(C_{small} = H) = 0.25$, and then we flip it. Due to their different sizes, we are twice as likely to grab $C_{big}$ as we are $C_{small}$. We can model the distribution of the flipped coin $C$ as a mixture of Bernoulli distributions:

$$
\begin{aligned}
\mathbb{P}(C = H) &= \mathbb{P}(C = C_{big}) \cdot \mathbb{P}(C = H | C = C_{big}) \\
&\qquad + \mathbb{P}(C = C_{small}) \cdot \mathbb{P}(C = H | C = C_{small}) \\
&= \frac{2}{3} \cdot 0.75 + \frac{1}{3} \cdot 0.25 \quad = \frac{7}{12} \\
\mathbb{P}(C = T) &= 1 - \mathbb{P}(C = H) \qquad = \frac{5}{12}
\end{aligned}
$$

This is nothing but the Law of Total Probability.

# Finite Mixture Models: Identifiability

- Suppose that each distribution in a mixture comes from the same family $F$ of distributions, defined on a parameter space $\boldsymbol{\Theta}$ so that $F = \{F_g(x; \boldsymbol{\theta}_g) : \boldsymbol{\theta}_g \in \boldsymbol{\Theta}, g = 1, \ldots, G\}$
- Let $C = \{\sum_{g=1}^{G} \pi_g \cdot F_g(x; \boldsymbol{\theta}_g) : \pi_g > 0, \sum_{g=1}^{G} \pi_g = 1, F_g(x; \boldsymbol{\theta}_g) \in F\}$ be the convex hull of F
- $C$ is **identifiable** all of its members are distinct, up to reordering of summations
- Mixtures that are not identifiable suffer from the **label-switching problem** and are difficult to estimate in general

# Identifiability: Example

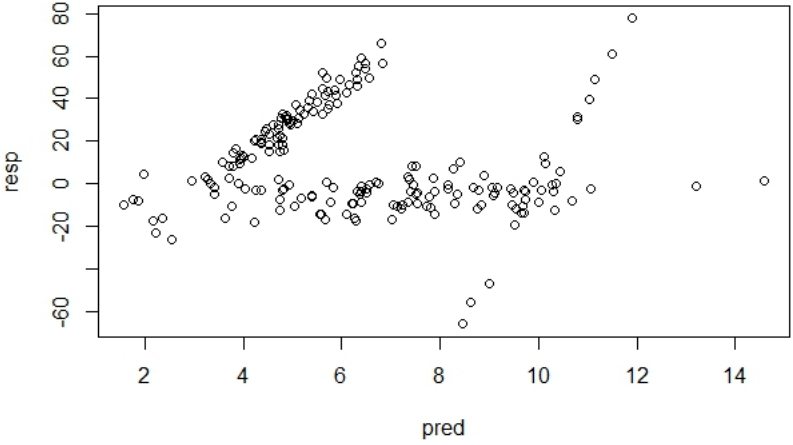
- The mixture of Bernoullis is not identifiable!
- Suppose we did not know $\mathbb{P}(C = C_{big})$ and $\mathbb{P}(C = C_{small})$ beforehand
- $\mathbb{P}(C = H) = \pi \cdot 0.75 + (1 - \pi) \cdot 0.25 = 0.5\pi + 0.25$ and $\mathbb{P}(C = T) = 0.75 - 0.5\pi$ for any $\pi \in (0, 1)$

**Theorem (Yakowitz, Spragins (1968))**

*$C$ is identifiable if and only if $F$ is linearly independent over $\mathbb{R}$.*

- With some mild constraints imposed, mixtures of linear regression models are identifiable

When clusters of data are far away from each other, fitting a finite mixture model is relatively straightforward. But this is not always the case:

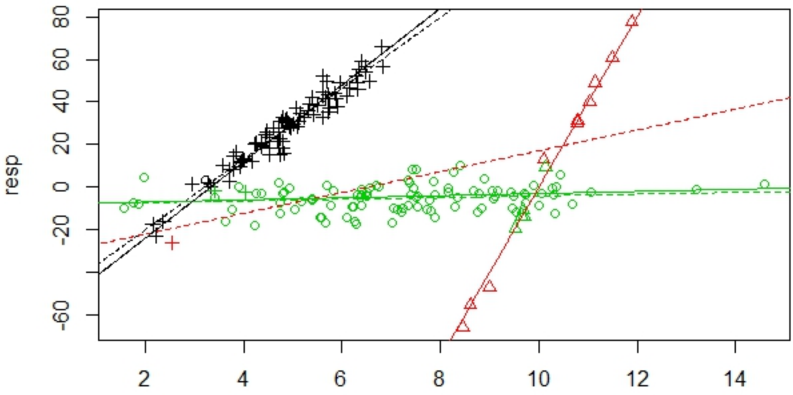


# Cluster Weighted Models: Definition

- Suppose that $\mathbf{y} \in \mathbb{R}^d$ is a *multivariate* response, $\mathbf{x} \in \mathbb{R}^p$ is a vector of explanatory covariates, and $\theta$ and $\pi_g$ are as defined previously
- A **cluster-weighted model** is a specific finite mixture model where $f(\mathbf{x}, \mathbf{y}) = \sum_{g=1}^{G} f_{\mathbf{Y}|\mathbf{X},\theta=g}(\mathbf{y}|\mathbf{x}, g) \cdot f_{\mathbf{X}|\theta=g}(\mathbf{x}|g) \cdot \pi_g$ is given as the *joint density* of $(\mathbf{X}, \mathbf{Y})$
- Here, each conditional density of $\mathbf{y}$ is weighted by both a mixing weight $\pi_g$ as well as a local density of $\mathbf{x}$ within group $g$ (which is usually assumed to be Gaussian)
- Cluster-weighted models allow for modelling data whose clusters may not appear to be distinct

# Cluster Weighted Models: Example

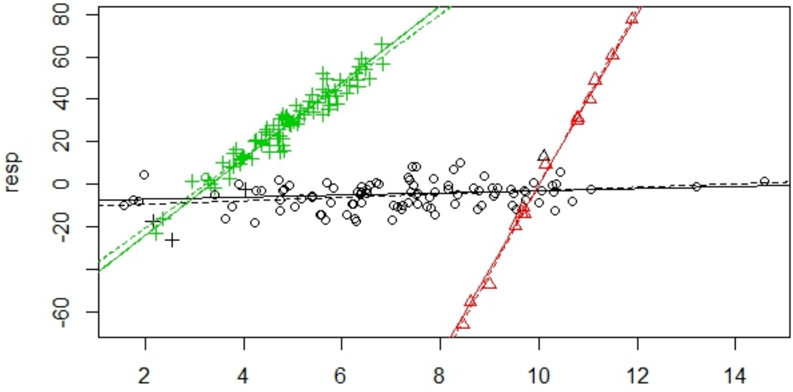
A finite mixture of regressions model was fit using the EM algorithm:




The algorithm classified many points, but failed to correctly classify the cluster which spanned a small portion of the feature space

# Cluster Weighted Models: Example

A cluster-weighted model was fit to the same data:




The algorithm correctly classified *all but five* points, and determined the actual lines that were used to generate the data almost perfectly

# Nonparametric Models

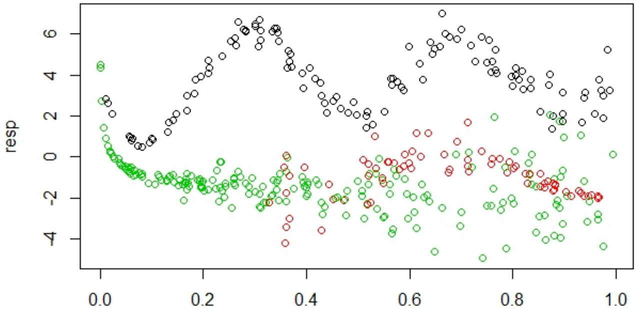
- Traditionally, finite mixture models are **fully parametric**; that is, each probability distribution $F_g(x)$ can be fully specified by a vector of fixed parameters $\boldsymbol{\theta}_g \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ is a finite-dimensional parameter space
  - For example, a mixture of Gaussian distributions of the form $F(x) = \sum_{g=1}^{G} \pi_g \cdot \mathcal{N}(\mu_g, \sigma_g^2)$ is fully parametric, with $\boldsymbol{\theta}_g = (\pi_g, \mu_g, \sigma_g^2)$
- In **nonparametric models**, the components of the distributions are not assumed to be constant, but are instead taken to be unknown functions of the predictors $\{\mathbf{x}_i\}$ themselves
- These functions require estimation

# Nonparametric Models: Example

$$\pi_1(x) = \frac{e^{\sqrt{x}}}{5(1+e^{\sqrt{x}})}, \quad \mu_1(x) = 4 - \frac{3}{2}x^{-\frac{1}{3}}\sin(5\pi x), \quad \sigma_1(x) = x^{\frac{4}{5}}$$

$$\pi_2(x) = \frac{x^2}{2}, \quad \mu_2(x) = -1 + \cos(3\pi x), \quad \sigma_2(x) = \frac{5}{2} - 3\sin(x)$$

$$\pi_3(x) = 1 - \pi_1 - \pi_2, \quad \mu_3(x) = \frac{1}{x^{\frac{3}{10}}} - 3, \quad \sigma_3(x) = 2x$$



# Benefits and Drawbacks

- Nonparametric models allow for much more freedom than parametric models, but there is a drawback
- In parametric models, parameters can be estimated from the data using straightforward approaches based on maximum likelihood estimation
    - In least squares regression, the ordinary least squares estimate is *the* MLE
    - In generalized linear models, quasi-Newton methods like the Fisher scoring algorithm numerically finds a root of the score equation
    - In mixtures of (parametric) Gaussian models, the EM algorithm uses a modified log-likelihood approach to estimate the parameters of the distributions as well as the mixing weights
- Likelihood estimation often fails for nonparametric models!

# Kernel Smoothing

- In nonparametric models, component functions are usually estimated using **kernels**
- If we fix one data point $x_0$, then a kernel $K_{x_0,\lambda}(x)$ assigns a weight $W_{\lambda,j}(x) = \frac{K_{x_0,\lambda}(x - x_j)}{\sum_{i=1}^{n} K_{x_0,\lambda}(x - x_i)}$ to each $x_j \in B_\lambda(x_0)$ based on its distance from $x_0$
- In one dimension, a kernel $K : \mathbb{R} \to \mathbb{R}$ is continuous, bounded, symmetric about 0, and satisfies $\int_{-\infty}^{\infty} K(x)dx = 1$
- In a regression setting, $\hat{f}(x) = \sum_{i=1}^{n} W_{\lambda,i}(x) \cdot y_i$ is a **kernel smoother** that provides a smooth nonparametric estimate of the true function $f(x)$, where $Y = f(X) + \epsilon$

# The Curse of Dimensionality

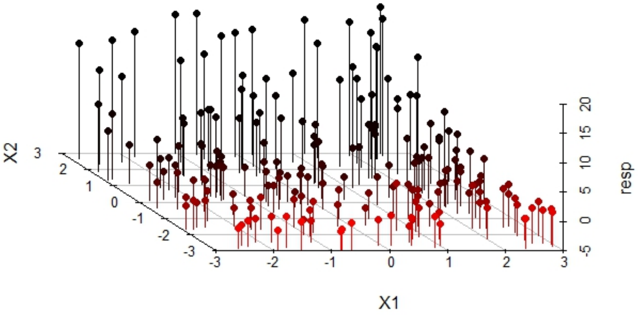
- In local regression, the radius $\lambda = \lambda(x_0)$ is called the **bandwidth**
- Because in general, the data is spread out non-uniformly, **variable bandwidth selection** must be used to determine $\lambda$ at each point
- Typically this is done by the ***k*-nearest neighbours algorithm**, which searches for the $k$ points closest to $x_0$
- In low dimensions, this is straightforward
- However, as the dimension grows, our feature space becomes sparser and we must search a much larger volume for the same $k$ points
- This is an example of the **curse of dimensionality**
- To circumvent this, dimension reduction techniques or feature selection algorithms may be used that restrict the data used

# Generalized Additive Models: Definition

- Recall that in our regression setting, we assumed that $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = f(\beta_0 + \beta_1 x_1 + \cdots \beta_p x_p)$
- This model is useful, but the requirement that the argument of $f(\cdot)$ be linear in the $x_i$'s is often too restrictive
- In a **generalized additive model**, we assume more generally that $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = f(\alpha_0 + \alpha_1(x_1) + \cdots \alpha_p(x_p))$, where each function $a_i : \mathbb{R} \to \mathbb{R}$ is *smooth*
- We can apply kernel smoothing techniques to each $\alpha_i$ individually, and thus avoid the curse of dimensionality
- Smoothing splines are another choice

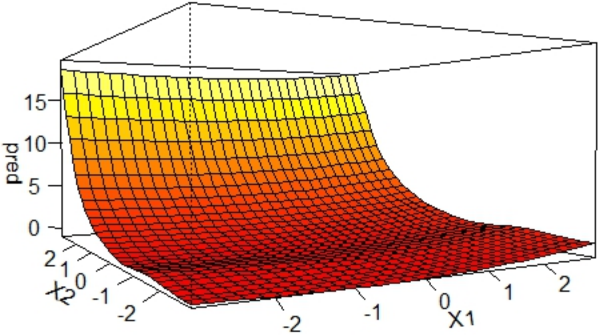
# Generalized Additive Models: Example

$$\alpha_0 = -1, \quad \alpha_1(x_1) = \frac{(x_1+1)^2}{10}, \quad \alpha_2(x_2) = e^{x_2}, \quad f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



# Generalized Additive Models: Continued

A GAM was fit to the above data:



# Cluster Weighted Models with GAMs

- A **cluster-weighted model with generalized additive components** is a finite mixture model where the joint density of $(\mathbf{X}, \mathbf{Y})$ takes the form

$$f(\mathbf{x}, \mathbf{y}) = \sum_{g=1}^{G} f_{\mathbf{Y}|\mathbf{X},\theta=g} \left( \alpha_{g,0} + \sum_{j=1}^{p} \alpha_{g,j}(x_j) | \mathbf{x}, g \right) \cdot f_{\mathbf{X}|\theta=g}(\mathbf{x}|g) \cdot \pi_g$$

where each function $\alpha_{g,h} : \mathbb{R} \to \mathbb{R}$ is smooth

# Summary: Why These are Good

- Finite mixture models are more versatile than "single" models because they allow for clustered data
- Cluster-weighted models are more versatile than finite mixture models because the additional weighting term allows for more accurate identifying of clusters
- Nonparametric models are more versatile than parametric models because they allow the components of distribution functions to vary
- GAMs are more versatile than simple additive models because they allow each covariate to vary in its own (smooth) way