# Nonparametric Mixture and Quantile Regression with Applications

Thesis Proposal

Alex L. Rojas

**Abstract**

Conditional density estimation and quantile regression are techniques that allow for a better understanding of the relationship between a response variable and a set of covariates in comparison with usual regression methods. Therefore, these techniques are of great importance in many scientific fields where knowledge about conditional means, obtained by regression methods, is not enough to draw valuable conclusions of the problem at hand. Unfortunately, these techniques are not immune to measurement error, which is a rather common problem. Moreover, the available conditional density estimators lack of an ease of interpretability. In this proposal, I first present a conditional density estimator based on finite mixture models and local likelihood estimation, which has the advantage of being easily interpretable. An error-in-variables conditional density estimator based on Fan, Yao and Tong's double kernel approach is presented as well. Second, I develop nonparametric error-in-variables quantile regression based on existing quantile regression methods available for the error-free case.

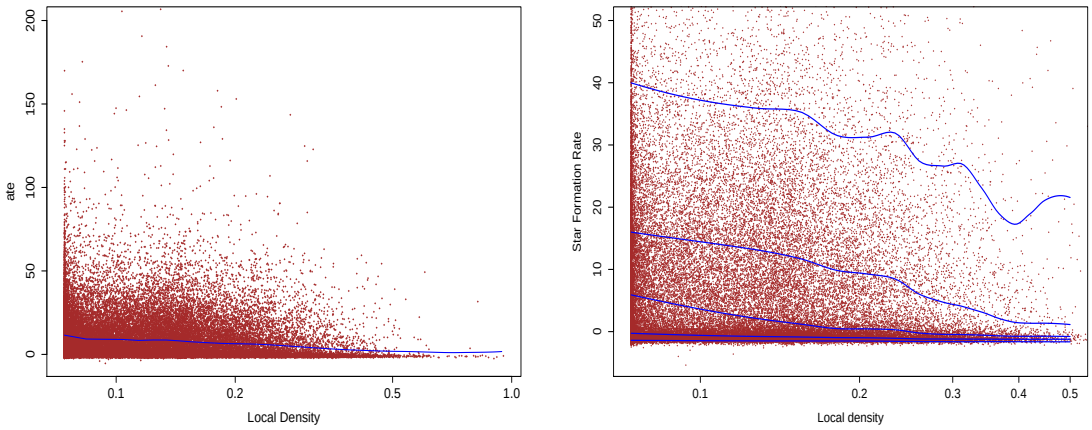
## Scientific Motivation

Through the years, astronomers have had varying ideas about the evolution of the universe. Many of these ideas come from different large-scale patterns observed in small, low-quality astronomical data sets, which point to different theories of how the universe evolved. Nowadays, modern computer technology allows astronomers to collect high-quality data from billions of objects, thus giving an opportunity for testing current theories. As expected, these rich data sets require statistical techniques able to describe complicated, high-dimensional and nonlinear processes with a precise assessment of uncertainty.

Among the hundreds of open questions that astronomers want to answer with new available astronomical surveys is to what degree galaxy evolution is influenced by its local environment. To study the role environment plays in the process of galaxy evolution, cosmologists often use a galaxy's "local density" ($D_l$) as an environmental measure and a galaxy's star formation rate as an evolutionary measure (e.g., Heavens et al., 2004; Gomez et al., 2003; Balogh et al., 2004) . An indicator of the recent star-formation in any given galaxy is the $H_\alpha$ "emission line" (visible in the galaxy spectra). The emission at this specific wavelength is from the process of hot, bright, young stars ionizing the cool, neutral hydrogen that permeates the intergalactic medium. The greater the flux in this line, the greater the amount of star-formation. When no star-formation is present in the galaxy, light at this same wavelength is often seen as an "absorption line," as electrons in the

hydrogen atoms can get excited into a higher energy level. Local density is obtained by estimating the density field on the point-like spatial galaxy distribution using a kernel density estimator.

Figure 1(a) shows the scatterplot between local density and $H_\alpha$ equivalent width (EW($H_\alpha$)) for 47592 galaxies, where $-$EW($H_\alpha$) is a measure of the intensity of an absorption line. A nonparametric estimate of the conditional mean is also included. As can be seen in this figure, galaxies located in very dense regions have a low recent star formation rate, while star-forming galaxies are found in less dense regions. However, no further information is obtained from the conditional mean estimate; thus, we need to use another statistical technique that provide us with more information.




(a) (b)

Figure 1: (a) Scatterplot between $D_l$ and EW($H_\alpha$), and the local likelihood estimate of the conditional mean; (b) smoothed 5th, 25th, 50th, 75th, 95th quantile curves using double-kernel smoothing

One of the available statistical techniques to study the underlying relationship between EW($H_\alpha$) and $D_l$ is quantile regression (Koenker and Bassett, 1978; Yu and Jones, 1998). Figure 1(b) displays a zoomed version of the data plotted in Figure 1(a) along with the 5th, 25th, 50th, 75th and 95th estimated conditional quantiles. We can observe how the conditional density functions have a long right tail and how the 5th and 25th conditional quantiles functions seem to be constant. However, it is impossible to determine how galaxy populations (star-forming and non-star-forming galaxies) interact as galactic systems become denser and thus get a better insight on the underlying cosmology. This situation may be due to the fact that the conditional mean and conditional quantiles do not include all information about the relationship between EW($H_\alpha$) and $D_l$. For this reason, it is needed to go further and estimate the conditional density, which completely describes the conditional relationship among two random variables. On the other hand, we have not accounted for errors in the measurement of local density resulting from the fact that local densities were obtained by means of a kernel density estimator, which may cause attenuation biases (Carroll et al., 1995). These two remarks have inspired us to pursue the following three research aims which can be widely applicable in other scientific contexts as well.

**AIM 1.** Produce conditional density estimators using finite mixture models.

To get a better understanding of the relationship between EW($H_\alpha$) and $D_l$, we use conditional density (CD) estimation. The current approaches to estimate conditional densities are based on kernel methods, but these estimates are highly dependent on the (fixed) bandwidths being used. A factor that contributes to this problem is that the conditional densities of EW($H_\alpha$) given $D_l$ seem to have a long tail; therefore, it is desirable to have a "small" bandwidth in regions of high density and a "large" bandwidth in the tail. In addition, it is hard to get a clear picture of the relationship between EW($H_\alpha$) and $D_l$, since we are not sure which features of the conditional estimates are real. I propose to develop a new set of conditional density estimators using finite mixture models. The main idea is to model each of the mixture parameters as a function of the covariates using local likelihood estimation. By modeling these parameter functions, this approach is expected to provide a better insight on the underlying relationships between random variables than current conditional density estimators.

**AIM 2.** Develop nonparametric error-in-variables quantile regression.

There exist many statistical methods to overcome the presence of measurement error in linear and nonlinear models (Fuller, 1987; Carroll et al., 1995), but few results currently exist for quantile regression functions (see Chesher, 2001; He and Liang, 2002). Given that covariate measurement errors cause changes in conditional quantile regression functions by altering their shape, orientation and location (Chesher, 2001), it is of major interest to develop nonparametric errors-in-variables quantile regression. I undertake this problem in two ways. First, by approaching the quantile regression problem as an approximate M-estimation problem and solving it using corrected score methods (Nakamura, 1990; Novick and Stefanski, 2002). Second, by using a "double-kernel" approach to quantile regression (Yu and Jones, 1998) along with the "deconvolution" kernel proposed by Fan and Truong (1993).

**AIM 3.** Produce a bandwidth selection method for error-in-variable conditional density estimation.

Even though double kernel conditional density estimation (Fan et al., 1996; Hyndman and Yao, 2002) has some drawbacks, it is easily extended to handle covariate measurement error using a deconvolution kernel (Fan and Truong, 1993). However, we still have to investigate if the current approaches for bandwidth selection for double kernel estimators are equally useful when covariate measurement error is present.

This proposal is organized in two sections. The first section corresponds to our first and third aims and the second section to the second aim. In each section a review of relevant literature is presented. Then, I present my approach and specific steps to accomplish the aims of this research. I also provide some preliminary results.

# Conditional Density Estimation Using Finite Mixture Models

The problem of estimating the conditional density of $Y$ given $\boldsymbol{X}$, where $Y \in \mathbb{R}^d$ and $\boldsymbol{X} \in \mathbb{R}^d$ is considered. Addressing this problem is important because the conditional density of $Y$ given $\boldsymbol{X}$ provides a complete description of the stochastic behavior of the response variable given any specific value of its covariates. Therefore, CD estimation generalizes the usual regression model, where the main focus is the conditional mean (i.e., the "center" of the conditional distribution) and quantile regression, which aims to model any conditional quantile. This generalization is most relevant when the conditional mean itself does not reveal the underlying relationship between $\boldsymbol{X}$ and $Y$.

## Current approaches to conditional density estimation

The conditional density function of a variable $Y$ given $X$ is defined as:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

where $f_X(x) \neq 0$. A "natural" estimator of the conditional density is

$$\hat{f}_{Y|X}(y|x) = \frac{\hat{f}_{X,Y}(x,y)}{\hat{f}_X(x)}, \tag{1}$$

where $\hat{f}_{X,Y}(x,y)$ is a kernel estimator of $f_{X,Y}(x,y)$ and $\hat{f}_X(x)$ is a kernel estimator of $f_X(x)$. Hyndman et al. (1996) studied asymptotic properties of this estimator and found two optimal bandwidths, one for the numerator and one for the denominator, with respect to integrated mean-square error.

Conditional density estimation can also be regarded as a nonparametric regression problem (Fan et al., 1996), by noticing that as $\tilde{h} \to 0$

$$\begin{aligned} E\{K_{\tilde{h}}(Y-y)|X=x\} &= \int_{\mathbb{R}} K_{\tilde{h}}(Y-y) f_{Y|X}(y|x)dy \\ &= \int_{\mathbb{R}} K(u) f(u\tilde{h}+y|x)du \\ &\approx f_{Y|X}(y|x), \end{aligned} \tag{2}$$

where $K$ is a symmetric density function on $\mathbb{R}$ and $K_h(t) = h^{-1}K(t/h)$. Therefore, we can estimate $f_{Y|X}(y|x)$ by regressing $K_{\tilde{h}}(Y-y)$ on $X$. This regression problem can be solved using local polynomial regression (see e.g. Fan and Gijbels, 1996) or local likelihood estimation (Tibshirani and Hastie, 1987; Loader, 1999).

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from $f_{X,Y}(x,y)$. Applying the local polynomial technique to the constructed data $(X_1, K_{\tilde{h}}(Y_1-y)), \ldots, (X_n, K_{\tilde{h}}(Y_n-y))$ reduces the estimation of the conditional density, $f_{Y|X}(y|x)$, to find $\hat{\beta}(x,y) = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$ such that

$$\hat{\beta}(x,y) = \arg\min_{\beta} \sum_{i=1}^{n} \left\{K_{\tilde{h}}(Y_i - y) - A(X_i - x, \beta)\right\}^2 W_h(X_i - x) \tag{3}$$

where

$$A(X_i - x, \beta) = \sum_{j=0}^{p} \beta_j \frac{(X_i - x)^j}{j!} \tag{4}$$

and $W$ is a symmetric density function on $\mathbb{R}$ and $W_h(t) = h^{-1}W(t/h)$. The local polynomial estimator of $f_p^{(j)}(y|X = x)$ is $\hat{\beta}_j$, in particular

$$\hat{f}_p(y|x) = A(0, \hat{\beta}(x, y)) = \hat{\beta}_0. \tag{5}$$

From now on, we refer to this class of estimators as 'double-kernel' estimators. Note that if $p = 0$, the double-kernel estimator reduces to

$$\begin{aligned} \hat{f}_0(y|x) &= \sum_{i=1}^{n} W_h(X_i - x) K_{\tilde{h}}(Y_i - y) \Big/ \sum_{i=1}^{n} W_h(X_i - x) \\ &= \frac{1}{nh} \sum_{i=1}^{n} W_h(X_i - x) K_{\tilde{h}}(Y_i - y) / \hat{f}_n(x) \end{aligned} \tag{6}$$

The estimator in Eq. (6) was first presented by Hyndman et al. (1996) and corresponds to the estimator in Eq. (1) when $f_{X,Y}(x, y)$ is estimated with the product kernel $K_{\tilde{h}} \times W_h$. When $p = 1$, the estimator in Eq. (5) has a smaller bias than the estimator in Eq. (6) (Fan and Gijbels, 1996); however, it is not guaranteed to be non-negative and to integrate to 1, as is the case when $p = 0$ (Hyndman and Yao, 2002). Recognizing this problem, Hyndman and Yao (2002) proposed two new non-negative estimators. The first proposal adds the constraint $\beta_0 > 0$ to the minimization problem in Eq. (3), by setting $\beta_0 = \ell(\alpha) = \exp(\alpha)$. The second proposal takes

$$A(X_i - x, \beta) = \exp\left\{ \sum_{j=0}^{p} \beta_j (X_i - x)^j \right\} \tag{7}$$

and $\hat{f}_p(y|x) = A(0, \tilde{\beta}(x, y)) = \exp\{\tilde{\beta}_0\}$. Hyndman and Yao (2002) noticed that their second proposal is equivalent to using local likelihood estimation for the regression of $K_{\tilde{h}}(Y_i - y)$ against $X_i$ with the Gaussian likelihood and link function $\log(\cdot)$. They also proposed an algorithm for bandwidth selection.

Figure 2 displays the conditional density estimated using the estimator in Eq. (5) with $A$ as in Eq. (7) and $p = 1$ (similar results are obtain with other double-kernel estimators). As can be seen in this figure, kernel conditional density estimation provides a general view of the conditional density function, but we cannot easily unveil the underlying structure of the data. This may be due to the fact that these estimators do not consider local bandwidths to estimate the local structure.

I propose to model the conditional density as a finite mixture model (FMM). In this case, each conditional density has a set of parameters that we model as a function of the conditioning information. Although FMMs involve stronger distributional assumptions than the nonparametric methods previously presented, they require less data and are more easily interpretable. In addition, kernel estimates may be approximated by much smaller mixtures without losing significant information (Scott and Szewczyk, 2001).

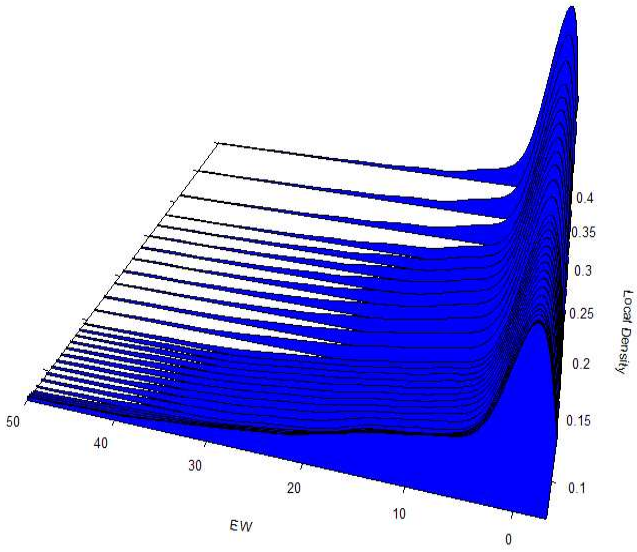



Figure 2: Double-kernel conditional density estimate for the data in Figure 1(a)

## Nonparametric Mixture Regression

Let us assume that the conditional density $f_{Y|X}(y|x)$ of $Y$ given $X$ can be written in the form

$$f_{Y|X}(y|x) = \sum_{i=1}^{k_x} \pi_i(x) g_i(y|\boldsymbol{\theta}_i(x)) \tag{8}$$

where the $g_i(y|\boldsymbol{\theta}_i(x))$, $i = 1, \ldots, k_x$, are densities with a set of parameters $\boldsymbol{\theta}_i(x)$ that depends on $x$, and the $\pi_i(x)$'s are a set of mixing proportions that sums to one for each $x$. Denote $\boldsymbol{\theta}(x) = (\boldsymbol{\theta}_1(x), \ldots, \boldsymbol{\theta}_{k_x}(x))$ and $\pi(x) = (\pi_1(x), \ldots, \pi_{k_x}(x))$. Assuming the model in Eq. (8), I propose to estimate the conditional density by modeling $\pi_i(\cdot)$ and $\boldsymbol{\theta}_i(\cdot)$, $i = 1, \ldots, k_x$, as a function of the conditioning information using local likelihood estimation (Loader, 1999). These functions are referred to as *parameter functions* and the proposed conditional density estimator as *nonparametric mixture regression* (NMR).

Let $\boldsymbol{\eta}(x) = (\pi_1(x), \ldots, \pi_{k_x}(x), \boldsymbol{\theta}_1(x), \ldots, \boldsymbol{\theta}_{k_x}(x))$ and $\ell(y_j, \boldsymbol{\eta}(x_j)) = \log f_{Y|X}(y_i|x_i)$, with $f_{Y|X}$ as in Eq. (8). The local polynomial log-likelihood of a parameter vector

$$\boldsymbol{\eta} = (\pi_1(x_1), \ldots, \pi_{k_{x_1}}(x_1), \boldsymbol{\theta}_1(x_1), \ldots, \boldsymbol{\theta}_{k_{x_1}}(x_1), \ldots, \pi_1(x_n), \ldots, \pi_{k_{x_n}}(x_n), \boldsymbol{\theta}_1(x_n), \ldots, \boldsymbol{\theta}_{k_{x_n}}(x_n))$$

is

$$\mathcal{L}_x(\boldsymbol{\beta}) = \sum_{j=1}^{n} w_j(x) \ell(Y_j, \mathcal{A}(x_j - x, \boldsymbol{\beta})), \tag{9}$$

where

$$\mathcal{A}(t, \boldsymbol{\beta}) = (A_{1,1}(t, \boldsymbol{\beta}_{1,1}), \ldots, A_{1,q_1}(t, \boldsymbol{\beta}_{1,q_1}), \ldots, A_{k_x,q_{k_x}}(t, \boldsymbol{\beta}_{k_x,q_{k_x}}), \ldots, A_{k_x,q_{k_x}}(t, \boldsymbol{\beta}_{k_x,q_{k_x}})),$$

with $A_{l,m}(\cdot,\beta_{l,j})$ as in Eq. (4), $m = 1,\ldots,q_l$, $l = 1,\ldots,k_x$ and $q_l$ the number of parameters of the $l^{\text{th}}$ component. The $\boldsymbol{\beta}$'s are vectors of coefficients and

$$w_j(x) = W\left(\frac{x_j - x}{h(x)}\right), \tag{10}$$

with $W(u)$ a weight function that assigns largest weights to observations close to $x$.

Let $\hat{\boldsymbol{\beta}}$ be the maximizer of the local likelihood Eq. (9), that is,

$$\begin{aligned}\hat{\boldsymbol{\beta}}(x) &= \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} w_j(x)\ell(Y_j, \mathcal{A}(x_j - x, \boldsymbol{\beta})) \qquad (11)\\ &= \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} w_j(x) \log \sum_{i=1}^{k_x} A_{i,1}(x_j - x, \boldsymbol{\beta}_{i,1}) \cdot g_i(Y_j | A_{i,2}(x_j - x, \boldsymbol{\beta}_{i,2}), \ldots, A_{i,q_i}(x_j - x, \boldsymbol{\beta}_{i,q_i})).\end{aligned}$$

The local likelihood estimate of the set of parameters $\boldsymbol{\eta}(x)$ is then defined as $\hat{\boldsymbol{\eta}}(x) = \mathcal{A}(0, \hat{\boldsymbol{\beta}}(x))$.

Letting $g_i(y|\boldsymbol{\theta}_i) = \phi(y|\mu_i, \sigma_i^2)$, a density function with parameters $\mu_i$ and $\sigma_i^2$, the number of parameter functions is three for all the mixture components. Furthermore, if all parameter function are approximated locally with the same order polynomial, that is, $A_{i,l}(\cdot,\cdot) = A(\cdot,\cdot)$, Eq. (11) can be written as

$$\hat{\boldsymbol{\beta}}(x) = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} w_j(x) \log \sum_{i=1}^{k_x} A(x_j - x, \boldsymbol{\beta}_{i,1}) \cdot \phi(y_j | A(x_j - x, \boldsymbol{\beta}_{i,2}), A(x_j - x, \boldsymbol{\beta}_{i,3})). \tag{12}$$

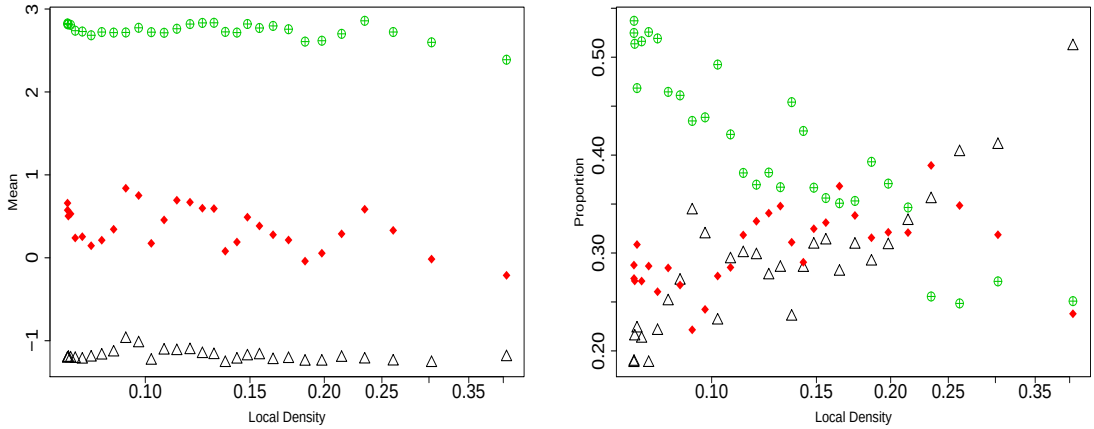
Notice that when $A(t, \boldsymbol{\beta})$ is a constant ($p = 0$), we obtain

$$\hat{\boldsymbol{\beta}}(x) = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} w_j(x) \log \sum_{i=1}^{k_x} \beta_{i,1} \cdot \phi(y_j | \beta_{i,2}, \beta_{i,3}). \tag{13}$$

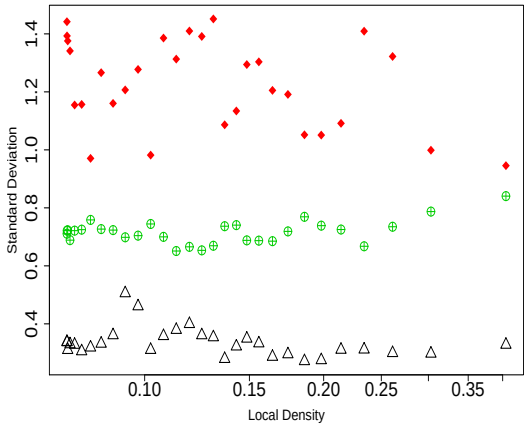
Therefore, the original problem is reduced to solving a finite mixture problem (see e.g., McLachlan and Peel, 2000). We can thus make use of some existing techniques used in mixture models to obtain a conditional density estimate.

The most popular algorithm to estimate the mixture parameters is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), which converges to a maximum likelihood estimate of the mixture parameters. Figures 3(a), 3(b) and 3(c) show the estimated parameter functions using NMR with $k_x = 3$ and the EM algorithm for the data in Figure 1(a). As can be seen in this plot, the predominant effect is observed in the proportion functions, where the fraction of star-forming galaxies decreases with increasing local density. Likewise, the population of non-star-forming galaxies increases with increasing local density. The third component, while noisy, does not undergo a significant change in proportion with density. The means and dispersions of the three populations seems to remain constant as density changes.

Note that the EM algorithm requires the knowledge of $k_x$, plus it is highly dependent on the parameter initialization. These drawbacks may be multiplied in our case, since we need to fit a mixture for each value $x$; therefore, it seems critical to modify this algorithm or find other approaches. I consider two approaches that avoid the drawbacks of the EM algorithm for mixture fitting: (i) the algorithm proposed by Figueiredo and Jain (2002), and (ii) the Iterative pairwise replacement algorithm (IPRA, Scott and Szewczyk, 2001). Notice that the IPRA is not a likelihood-type method, but it is worth to study its behavior given its properties (see below).




(a) (b)




(c)

Figure 3: Estimated conditional component (b) mean functions, (c) proportion functions, and (d) standard-deviation functions using nonparametric mixture regression.

As in nonparametric regression, local approximations of higher order are usually prefer due to their ability to reduce bias, particularly at boundary regions. Frequently, local linear polynomial are enough when we are interested in the parameter functions as opposed to their derivatives. However, if the true parameter function are expected to have a substantial curvature, further bias reduction can be obtained by fitting local quadratic polynomials. In any of these cases, Eq. (13) can be found using the EM algorithm for mixture models and the degree of the polynomial is decided by the user.

The following section describe the IPRA and FJME algorithms. We start by introducing some model selection criteria and a similarity measure for densities.

### Model selection

When the number of mixtures $k_x$ is unknown, we could estimate $k_x$ as follows. Use the EM algorithm to obtain a sequence of parameter estimates for a range of values of $k_x$ and estimate $k_x$ as

$$\hat{k}_x = \arg\min_k \{\mathcal{C}(\hat{\boldsymbol{\beta}}(x), k), k = k_{x,min}, \ldots, k_{x,max}\} \tag{14}$$

where $\mathcal{C}(\cdot, k)$ is some model selection criterion. There are multiple choices $\mathcal{C}(\cdot, k)$; however, in this paper we make use of the Bayesian Information Criterion (BIC, Schwarz, 1978) and the Integrated Squared Error (ISE) criterion.

The BIC is defined as

$$BIC(\hat{\boldsymbol{\beta}}(x), k) = -\log \mathcal{L}_x(\hat{\boldsymbol{\beta}}) + \frac{N(k)}{2} \log n_x \tag{15}$$

where $N(k)$ is the total number of estimated parameters and $n_x$ is the sample size. The ISE is defined as

$$ISE(\hat{\boldsymbol{\beta}}(x), k) = \int_{-\infty}^{\infty} (f_k(y|x, \hat{\boldsymbol{\beta}}(x)) - f(y|x))^2 dy \tag{16}$$

where $f_k(y|x, \hat{\boldsymbol{\beta}}(x))$ is a conditional density estimate of $Y$ given $X$ using a mixture of $k$ components.

Another approach, proposed by Figueiredo and Jain (2002), is to consider a mixture of $k_x$ components as a mixture of $k(> k_x)$ components where $k_z(< k)$ components have zero weight. In this case, we need to have a criterion that can select the "best" model in the entire set of available models. As noticed by Figueiredo and Jain (2002), this approach resembles the Minimum Message Length (MML) philosophy (Wallace and Freeman, 1987). MML criteria are based on the idea that statistical inference can be view as data compression. In other words, if we can build a short code for the available data, then we will have a good data generation model (Rissanen, 1989). Figueiredo and Jain (2002) developed the following MML criterion

$$\hat{\boldsymbol{\beta}}(x) = \arg\min_{\boldsymbol{\beta}} -\log \mathcal{L}_x(\hat{\boldsymbol{\beta}}) + N(k) k_{nz} \log n_x + \frac{k_{nz}}{2} \log n_x + \frac{N(k)}{2} \cdot \sum_{j:\pi_j(x)>0} \log \pi_j(x) \tag{17}$$

where $k_{nz}$ is the number of non-zero-probabilities components.

We finish this section by introducing the similarity measure for densities given by Scott and Szewczyk (2001). They defined a similarity measure between two density functions $g_1, g_2$ as

$$\text{sim}(g_1, g_2) = \frac{\int_{-\infty}^{\infty} g_1(t) g_2(t)\, dt}{\left( \int_{-\infty}^{\infty} g_1^2(t)\, dt \int_{-\infty}^{\infty} g_2^2(t)\, dt \right)^{1/2}}, \tag{18}$$

based on the intuition that when $g_1$ and $g_2$ are similar, $\int g_1(x) g_2(x)\, dx$ should be larger than when $g_1$ and $g_2$ are not similar. Scott and Szewczyk (2001) showed that $0 \leq \text{sim}(g_1, g_2) \leq 1$.

### Figueiredo and Jain's algorithm

Figueiredo and Jain (2002) noticed that the MML criterion for mixtures in Eq. (17) is equivalent to an a posteriori density resulting from the use of a Dirichlet-type prior for the $\pi_i(x)$'s and a flat prior for the $\boldsymbol{\theta}_i(x)$'s. Therefore, to minimize Eq. (17), the values $\hat{\pi}_i^{(t+1)}(x)$, $i = 1, \ldots, k$, calculated in the M-step of the traditional EM algorithm for mixtures are changed to

$$\hat{\pi}_i^{(t+1)}(x) = \frac{\max\left\{0, \sum_{j=1}^{n_x} \gamma_{ji}^{(t+1)} - \frac{N}{2}\right\}}{\sum_{m=1}^{k_x} \max\left\{0, \sum_{j=1}^{n_x} \gamma_{jm}^{(t+1)} - \frac{N}{2}\right\}}, \text{ for } i = 1, \ldots, k, \tag{19}$$

where $N$ is the number of parameters that specify each component, and the $\gamma_{ji}^{(t+1)}$'s are the values obtained in the traditional E-step.

Note that the modified M-step eliminates components that are not supported by the data; therefore, the EM algorithm can be initialized with a "large" number of components, which helps to move components across low-likelihood regions and then eliminate all unnecessary components. However, if the algorithm is initialized with an "extremely large" number of components, the first iteration of the modified M-step may eliminate all of them. To avoid this problem, Figueiredo and Jain (2002) used the component-wise EM algorithm (CEM, Celeux et al., 1999). The CEM algorithm differs from the traditional EM algorithm in that it updates the estimation of the $\pi_i(x)$'s and $\boldsymbol{\theta}_i(x)$'s one by one, instead of all together. That is, CEM updates the estimates $\pi_1(x)$ and $\boldsymbol{\theta}_1(x)$ and continues with the E-step, then it updates the estimates of $\pi_2(x)$ and $\boldsymbol{\theta}_2(x)$ and goes to the E-step, and so on.

### Iterative Pairwise Replacement Algorithm

Scott and Szewczyk (2001) proposed the Iterative Pairwise Replacement Algorithm (IPRA), which is an algorithm for fitting mixture models sequentially. The main idea behind IPRA is that kernel estimates may be approximated by much smaller mixtures. This algorithm starts by first constructing a kernel density estimate, using either the unbiased cross-validation (UCV) bandwidth (Rudemo, 1982; Browman, 1984) or the normal reference rule (Silverman, 1986). Second,

it sequentially eliminates the redundant components in the mixture until $\tilde{k}$ components remain, where $\tilde{k}$ is selected based on the sample size. Each time, the two closest components, in terms of the similarity measure in Eq. (18), are combined using the method of moments (MoM); that is, given two components with parameters $(w_1, \mu_1, \sigma_1^2)$ and $(w_2, \mu_2, \sigma_2^2)$, respectively, the new component will have parameters

$$(w_i + w_{i+1}, w_i'\mu_i + w_{i+1}'\mu_{i+1}, w_i'\sigma_i^2 + w_{i+1}'\sigma_{i+1}^2 + w_i'w_{i+1}'(\mu_i - \mu_{i+1})^2) \tag{20}$$

where $w_i' = w_i/(w_i + w_{i+1})$ and $w_{i+1}' = 1 - w_i'$. At this point, we end up with a mixture of $\tilde{k}$ components with the set of parameters $\{(w_1, \mu_1, \sigma_1^2), \ldots, (w_{\tilde{k}}, \mu_{\tilde{k}}, \sigma_{\tilde{k}}^2)\}$, such that $\mu_1 < \ldots < \mu_{\tilde{k}}$. Third, the similarity function in Eq. (18) is used to compare the current $k^*$-component mixture and the $(k^*-1)$-component mixture obtained by combining each pair of adjacent components using the MoM. The pair that maximizes $\text{sim}(\hat{g}_k, \hat{g}_{k-1})$ are then combined as in Eq. (20). This process continues until a model with only $k_0$ components ($k_0$ is usually less than 30) is obtained. Next, using "$L_2E$ with data," explained below, a pairwise combination is carried out until there is only one component left. At each step the BIC and the $L_2E$ criterion are collected. Finally, an appropriate number of components is chosen based on these criteria.

"$L_2E$ with data" refers to the method of finding the best $(k-1)$-component mixture, $f_{k-1}(y|x, \hat{\boldsymbol{\beta}}(x))$, by keeping all but one component fixed on an initial $k$-component estimate, $f_k(y|x, \hat{\boldsymbol{\beta}}(x))$, in terms of ISE. That is, we need to find the set of parameters such that

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{k-1}(x) &= \arg\min_{\boldsymbol{\beta}} \int_{-\infty}^{\infty} (f_{k-1}(y|x, \boldsymbol{\beta}(x)) - f_{Y|X}(y|x))^2 \, dy \\
&\approx \arg\min_{\boldsymbol{\beta}} \int_{-\infty}^{\infty} f_{k-1}(y|x, \boldsymbol{\beta}(x))^2 \, dy - 2\sum_{j=1}^{n_x} w_j(x) f_{k-1}(y_j|x, \boldsymbol{\beta}(x)).
\end{aligned} \tag{21}$$

## Measurement Error

Let us assume that our variable of interest $X$ cannot be measure precisely or is unobservable and instead $Z = X + e$ is manifest. This will be the error model used from now on. Note that

$$\phi_Z(t) = \phi_X(t)\phi_e(t)$$

where $\phi_U$ is used to denote the characteristic function of a random variable $U$. Using a kernel $W$ with bandwidth $h_n$ and applying the Fourier inversion theorem, Fan and Truong (1993) consider the following estimator of $f_X(x)$:

$$\begin{aligned}
\hat{f}_n(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \phi_W(th_n) \frac{\hat{\phi}_n(t)}{\phi_e(t)} \, dt \\
&= \frac{1}{nh_n} \sum_{i=1}^{n} W_n\left(\frac{x - Z_i}{h_n}\right),
\end{aligned} \tag{22}$$

where $\hat{\phi}_n(t)$ is the empirical characteristic function:

$$\hat{\phi}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \exp\left(itX_i\right),$$

and

$$W_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(-itx\right)\frac{\phi_W(t)}{\phi_e(t/h_n)}\, dt. \tag{23}$$

As in the nonparametric kernel regression (see, Fan and Truong, 1993), this deconvolution kernel can be used to construct a new nonparametric error-in-variables conditional density estimator based on the CD estimator in Eq. (6) and the deconvolution kernel density estimator in Eq. (22), as follows

$$\hat{f}_{0,n}(y|x) = \frac{1}{nh_n}\sum_{i=1}^{n} W_n\left(\frac{x - Z_i}{h_n}\right) K_{\tilde{h}}(y - Y_i)/\hat{f}_n(x). \tag{24}$$

## Future Work

I have presented a new conditional density estimator which has the advantage of being easily interpretable and provides us with information not easily accessible with existent regression methods. Furthermore, I constructed a nonparametric error-in-variables conditional density estimator based on a error-free conditional density estimator. However, there are some open issues regarding these two estimators, which I will study as part of my thesis. First, when applying these estimators, as in all other methods where kernel functions are used, bandwidth selection is a critical issue. In the case of error-in-variables CD estimation, I will build on the bandwidth selection methods proposed by Fan and Truong (1993) and Fan and Yim (2004). For nonparametric mixture regression, one might like to choose a separate bandwidth for each fitting point; however, in practice, doing this in a sensible manner is difficult. Therefore, I will simply study this estimator using a nearest neighbor bandwidth and a constant bandwidth selected by using generalized cross validation. Second, I will compare the performance of the Figueiredo and Jain's algorithm, the classical EM algorithm for mixture models and the IPRA when they are used for nonparametric mixture regression.

# Nonparametric Error-in-variables Quantile Regression

Quantile regression (QR) is a statistical technique intended to estimate the full range of quantile functions and capable of providing a more complete insight of the underlying relationships among random variables than usual regression. It is of particular interest in problems where the shape or the dispersion of conditional distributions varies with covariate values. This technique was introduced by Koenker and Bassett (1978), and has been applied to many scientific fields such as biology and ecology (Cade and Noon, 2003, and references therein), finance (Bassett and Chen, 2001) and econometrics (Fitzenberger et al., 2002) to mention a few. Despite the advances in QR function estimation and inference methods, and the many applications of QR methods, covariate measurement error (CME) has received scant attention in the context of QR estimation. QR is not immune to CME as showed by Chesher (2001): CME alters the shape, orientation and location of conditional quantile regression functions. Furthermore, ignoring this error may result in invalid conclusions.

In the following three sections, I introduce two nonparametric approaches to QR and a method to handle CME in regression problems. These existent methodologies will then be used to develop nonparametric error-in-variables quantile regression.

## Nonparametric Quantile Regression

Let $W$ be a symmetric density kernel and denote the distribution function of $W$ by $\Omega$. As in Eq. (2), it can be shown that as $\tilde{h} \to 0$

$$E\left\{ \Omega\left(\frac{y-Y}{\tilde{h}}\right) \middle| X = x \right\} \approx F(y|x), \tag{25}$$

where $F_{Y|X}$ denotes the conditional distribution of $Y$ given $X$.

Yu and Jones (1998) proposed to use this fact to estimate conditional quantiles. More precisely, their estimator uses a further linear approximation:

$$E\left\{ \Omega\left(\frac{y-Y}{\tilde{h}}\right) \middle| X = z \right\} \approx F(y|z) \approx F(y|x) + \dot{F}(y|x)(z-x) \equiv a + b(z-x) \tag{26}$$

where $\dot{F}(y|x) = \partial F(y|x)/\partial x$. Now, define $\tilde{a} = \hat{F}_{h,\tilde{h}}(y|x)$, where

$$(\tilde{a}, \tilde{b}) = \arg\min_{a,b} \sum_{j=1}^{n} \left( \Omega\left(\frac{y-Y_j}{\tilde{h}}\right) - a - b(X_j - x) \right)^2 K\left(\frac{x-X_j}{h}\right). \tag{27}$$

Solving Eq. (27) we obtain

$$\hat{F}_{h,\tilde{h}}(y|x) = \sum_{j=1}^{n} w_j^*(x;h) \Omega\left(\frac{y-Y_j}{\tilde{h}}\right) \tag{28}$$

where $w_i^*(x;h)$ is the weight function associated with local linear fitting,

$$w_j^*(x;h) = w_j(x;h) \Big/ \sum_{l=1}^{n} w_l(x;h)$$

where

$$w_j(x,h) = K\left(\frac{x - X_j}{h}\right)[S_{n,2} - (x - X_j)S_{n,1}],$$

with

$$S_{n,l} = \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)(x - X_j)^l.$$

To obtain a conditional quantile estimator, define $\tilde{q}_p(x)$ to satisfy $\hat{F}_{h,\tilde{h}}(\tilde{q}_p(x)|x) = p$, so that

$$\tilde{q}_p(x) = \hat{F}^{-1}_{h,\tilde{h}}(p|x). \tag{29}$$

Another possibility to estimate quantile functions, proposed by (Jones and Hall, 1990), is to find the solution $\hat{a}$ of the following equation

$$H_p(a) \equiv \sum_{j=1}^{n} W_j(x)\psi_0(Y_j - a) = 0, \tag{30}$$

where

$$\psi_0(z) = \begin{cases} \alpha I_{(0,\infty)}(z) - (1-\alpha)I_{(-\infty,0)}(z) & z \neq 0 \\ 0 & z = 0. \end{cases} \tag{31}$$

with $I_A(z)$ the indicator function of $A$.

## Monte Carlo Corrected Scores

Nakamura (1990) proposed a method for the analysis of data measured with error, called the corrected-score method. Nakamura's method is very attractive due to its generality and theoretical robustness properties but, the identification of the corrected score has to be performed on a model-by-model basis. To overcome this problem, Novick and Stefanski (2002) proposed to use Monte Carlo methods to obtain corrected scores. I follow this approach to construct a conditional quantile function estimator. In general, the corrected-score method does not assume a model for the observed data, but rather assumes that, in the absence of measurement error, the parameter of interest is consistently estimated by an $m$ estimator. It is assumed that the parameter of interested is consistently estimated by $\hat{\theta}$ satisfying

$$\sum_{j=1}^{n} \psi(Y_j, X_j, \hat{\theta}) = 0$$

where the score function $\psi$ is conditionally unbiased, that is,

$$E_\theta\{\psi(Y_j, X_j, \theta)|X_j\} = 0 \qquad j = 1, \ldots, n.$$

A corrected score is a function $\tilde{\psi}$ of the observed data having the property that

$$E\{\tilde{\psi}(Y_j, Z_j, \theta)|(Y_j, X_j)\} = \psi(Y_j, X_j, \theta), \qquad j = 1, \ldots, n.$$

Suppose that $V_{k,j} \sim \mathcal{N}(0,1)$ and consider $\widetilde{Z}_{k,j} = Z_j + i\sqrt{\sigma}V_{k,j}$, where $i = \sqrt{-1}$. Given that a score function $\psi$ can be expanded in an infinite power series in its second argument, and that expectation and summation can be interchanged, Novick and Stefanski (2002) showed that

$$\mathrm{E}[\mathrm{Re}\{\psi(Y_j, \widetilde{Z}_{k,j}, \theta)|(Y_j, X_j)\}] = \psi(Y_j, X_j, \theta),$$

that is, $\mathrm{Re}\{\psi(Y_j, \widetilde{Z}_{k,j}, \theta)\}$ is a corrected score. This corrected score depends on the particular generated random variable $V_{k,j}$; thus, it is desirable to eliminate this dependency by generating $m$ random samples and taking the average corrected score, that is,

$$\tilde{\psi}(Y_j, Z_j, \theta, \sigma) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{Re}\{\psi(Y_j, \widetilde{Z}_{m,j}, \theta)\}. \tag{32}$$

## Proposed Estimators

In the same manner that I used the deconvolution kernel in Eq. (23) to account for errors in covariates when estimating conditional densities, the deconvolution kernel can be used for quantile regression as well. More precisely, using the double-kernel quantile regression estimator in Eq. (28) with a local constant approximation, instead of a local linear approximation, I propose the following quantile function estimator involving errors in variables:

$$\tilde{q}_{p,n}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} W_n\left(\frac{x - Z_i}{h_n}\right)\Omega\left(\frac{y - Y_i}{\tilde{h}}\right)/\hat{f}_n(x). \tag{33}$$

A second proposal is to use Monte Carlo corrected scores where the score function is taken to be $\psi_0$ as in Eq. (31). It is possible that my choice of score function does not follow the assumptions of the Monte Carlo corrected score method; however, this method has been shown to perform well even in cases where its theoretical assumptions do not hold (Novick and Stefanski, 2002). Thus, there is some possibility that this estimator can perform better than its naive counterpart, where a naive estimator refers to an estimator that disregards measurement errors.

## Future Work

I will study the performance of the two proposed error-in-variables quantile regression estimators and compare them to their naive estimators. I will also develop a bandwidth selection method based on the bandwidth selection methods for the naive counterparts of the proposed estimators. Finally, the estimator in Eq. (33) uses only a local constant approximation; thus, I will study the possibility of using a deconvolution kernel in the local linear approximation case.

## References

Balogh, M., Eke, V., Miller, C., Lewis, I., Bower, R., Couch, W., Nichol, R., Cannon, R., Cole, S., Colless, M., Collins, C., Cross, N., Dalton, G., De-Propis, R., Driver, S. P., Efstathiou, G., Ellis, R. S., Frenk, C. S., Glazebrook, K., Gomez, P., Gray, A., Hawkins, E., Jackson, C., Lahav, O., Lumsden, S., Maddox, S., Madgwick, D., Peder Norberg, J. A. P., Percival, W., Peterson, B. A., Sutherland, W., and Taylor, K. (2004), "Galaxy ecology: groups and low-density environments in the SDSS and 2dFGRS," *Monthly Notices of the Royal Astronomical Society*, 348, 1355 – 1372.

Bassett, G. W. and Chen, H.-L. (2001), "Portafolio Style: Return-based Attribution using Quantile Regression," *Empirical Economics*, 19, 293 – 305.

Browman, A. W. (1984), "An Alternative Method of Cross-validation for the Smoothing of Density Estimates," *Biometrika*, 71, 353 – 360.

Cade, B. S. and Noon, B. R. (2003), "A Gentle Introduction to Quantile Regression for Ecologists," *Frontiers in Ecology and the Environment*, 1, 412–420.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London, England: Chapman and Hall.

Celeux, G., Chrétien, S., Forbes, F., and Mkhadri, A. (1999), "A component-wise EM algorithm for mixtures," Tech. Rep. 674, INRIA, Rhône-Alpes, France.

Chesher, A. (2001), "Parameter Approximations for Quantile Regressions with Measurement Error," CEMMAP working paper CWP02/01.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society*, 39, 1–38.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall.

Fan, J. and Truong, Y. K. (1993), "Nonparametric Regression with Errors in Variables," *The Annals of Statistics*, 21, 1900 – 1925.

Fan, J., Yao, Q., and Tong, H. (1996), "Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems," *Biometrika*, 83, 189–206.

Fan, J. and Yim, T. (2004), "A crossvalidation method for estimating conditional densities," *Biometrika*, 91, 819–834.

Figueiredo, M. and Jain, A. K. (2002), "Unsupervised learning of finite mixture models," *IEEE Transaction on Pattern Analysis and Machine Intelligence - PAMI*, 24, 381–396.

Fitzenberger, B., Koenker, R., and Machado, J. A. F. (eds.) (2002), *Economic Applications of Quantile Regression*, Physica Verlag.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Gomez, P., Nichol, R., Miller, C., Balogh, M., Goto, T., Zabludoff, A., Romer, K., Bernardi, M., Sheth, R., Hopkins, A., Castander, F., Connolly, A., Schneider, D., Brinkmann, J., Lamb, D., SubbaRao, M., and York, D. (2003), "Galaxy Star-Formation as a Function of Environment in the Early Data Release of the Sloan Digital Sky Survey," *Astrophys.J.*, 584, 210–227.

He, X. and Liang, H. (2002), "Quantile Regression Estimates for a Class of Linear and Partially Linear Errors-in-Variables Models," *Statistica Sinica*, 10, 129 – 140.

Heavens, A., Panter, B., Jimenez, R., and Dunlop, J. (2004), "The star-formation history of the Universe from the stellar populations of nearby galaxies," *Nature*, 428, 625 – 627.

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996), "Estimating and Visualizing Conditional Densities," *Journal of Computational and Graphical Statistics*, 5, 315–336.

Hyndman, R. J. and Yao, Q. (2002), "Nonparametric Estimation and Symmetry Test for Conditional Density Functions," *Journal of Nonparametric Statistics*, 14, 259–278.

Jones, M. C. and Hall, P. (1990), "Mean Square Error Properties of Kernel Estimates of Regression Quantiles," *Statistics and Probability Letters*, 10, 283 – 289.

Koenker, R. and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Loader, C. R. (1999), *Local Regression and Likelihood*, New York: Springer-Verlag.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Nakamura, T. (1990), "Corrected Score Function for Errors-in-variables Models: Methodology and Application to Generalized Linear Models," *Biometrika*, 77, 127 – 137.

Novick, S. J. and Stefanski, L. A. (2002), "Corrected Score Estimation via Complex Variable Simulation Extrapolation," *Journal of the American Statistical Association*, 97, 472 – 481.

Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, Singapore: World Scientific.

Rudemo, M. (1982), "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics*, 9, 65 – 78.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Scott, D. W. and Szewczyk, W. F. (2001), "From Kernels to Mixtures," *Technometrics*, 43, 323 – 335.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Tibshirani, R. and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567.

Wallace, C. and Freeman, O. (1987), "Estimation and inference via compact coding," *Journal of the Royal Statistical Society*, 49, 241–252.

Yu, K. and Jones, M. C. (1998), "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237.