Conditional Density Estimation using Finite Mixture Models with an Application to Astrophysics

Alex L. Rojas

Advisors:
Christopher R. Genovese, Christopher J. Miller, Robert Nichol
and Larry Wasserman

Center of Automatic Learning and Discovery and Department of Statistics
Carnegie Mellon University

July 26, 2005

Abstract

Conditional density estimation (CDE) is a statistical technique that allows for a better understanding of the relationship between a response variable and a set of covariates, in comparison with usual regression methods. Therefore, this technique is of great importance to many scientific fields where knowledge about conditional means, obtained by regression methods, is not enough to draw valuable conclusions about the problem at hand. There are a variety of conditional density estimators, but most of them lack generality or ease of interpretation. We present a solution to this problem by modeling the conditional density of Y given X using finite mixture models and estimating the parameter functions using local likelihood estimation. We use the proposed estimator to analyze the relationship between galaxy evolution and local density.

Keywords: Mixture Models, EM algorithm, kernel density estimation, local likelihood regression.

1 Introduction

We consider the problem of estimating the conditional density of Y given $\boldsymbol{X}$, where $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}$. 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, CDE 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. For example, consider the following model:

$$Y|X = x \sim \sum_{i=1}^{3} \pi_{i,x} \mathcal{N}(\mu_{i,x}, 9^2), \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with parameters μ and σ^2,

$$\pi_{x,i} = \begin{cases} \dfrac{1}{3} & i = 1 \\ \dfrac{\exp\left(x/20 - 2.3\right)}{0.9} & i = 2 \\ \dfrac{2}{3} - \pi_{x,2} & i = 3 \end{cases} \quad \text{and} \quad \mu_{x,i} = \begin{cases} 20 \cdot \mu_{x,2}(1 - 3 \cdot \pi_{x,2}) - 25 & i = 1 \\ \dfrac{(x-5)^2 + 40}{2} & i = 2 \\ -\mu_{x,2} & i = 3 \end{cases} \tag{2}$$

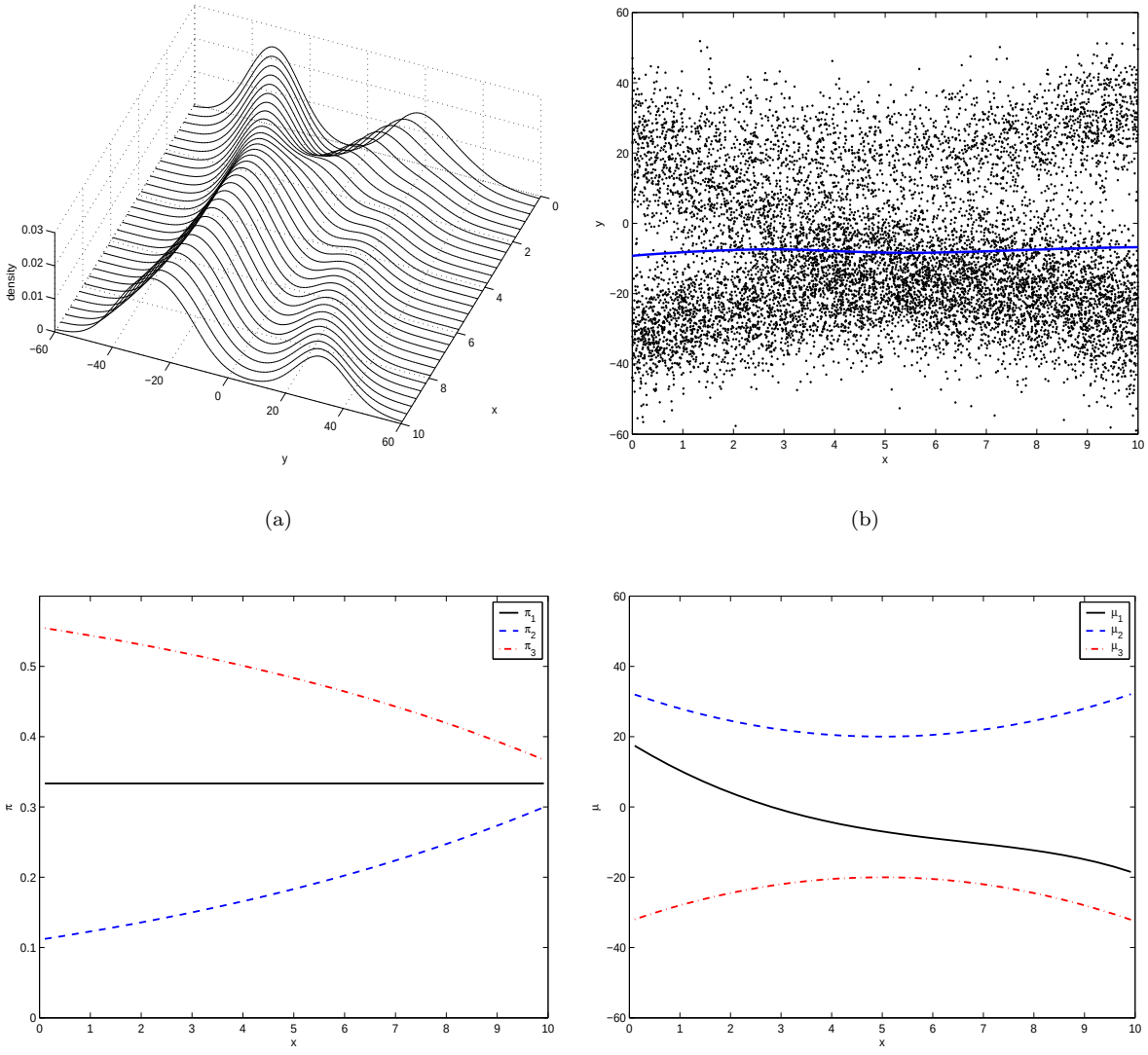


Figure 1: (a) Conditional density of $Y|X = x$, $f_{Y|X=x}(y|x)$, in Eq. (1), (b) a random sample of size 10000 drawn from $f_{Y|X=x}(y|x)$ and estimated conditional mean using local linear regression, (c) mixture proportions $\pi_{x,i}, i = 1, 2, 3$ in Eq. (2), and (d) mixture means $\mu_{x,i}, i = 1, 2, 3$ in Eq. (2).

Figure 1(a) shows the conditional density of Y given X for some values of X. Figure 1(c) and Figure 1(d) display the means and proportions in Eq. (2) as functions of x, respectively. In this model, $E(Y|X = x) = -25/3$, and this is what regression methods will estimate (see Figure 1(b)). It is clear that the conditional mean does not give us important information about the relationship between X and Y. The situation just illustrated does not only appear in "toy" examples, but in "real-world" applications as well. For example, in astrophysics, scientists are interested in studying the relationship between galaxy evolution and its local environment; however, the conditional mean does not give any insight into the cosmology behind this relationship and another statistical tool is needed to study this problem.

Current approaches for CDE are presented in Section 2 and our approach is introduced in Section 3. An application of our CDE approach to the study of the relationship between galaxy evolution and its local environment appears in Section 4. Finally, Section 5 presents our conclusions and future extensions of our work.

2 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 $f_X(x) \neq 0$. Therefore, a "natural" estimator of the conditional density of $Y|X$ can be found if we had an estimator of the joint density $f_{X,Y}(x,y)$ and an estimator of the marginal density of X, $f_X(x)$. The best well-known density estimator is the kernel density estimator. In the multivariate case, it is defined as follows:

$$\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{n|H|}\sum_{i=1}^{n} \mathbf{K}_d\left(H^{-1}(\boldsymbol{x} - \boldsymbol{x}_i)\right) \tag{3}$$

where H is a $d \times d$ nonsingular matrix and $\mathbf{K}_d : \mathbb{R}^d \to \mathbb{R}$ is a d-dimensional kernel, that is, a smooth function such that

$$\int_{\mathbb{R}^d} \mathbf{K}_d(\mathbf{w})d\mathbf{w} = 1, \tag{4}$$

$$\int_{\mathbb{R}^d} \mathbf{w}\mathbf{K}_d(\mathbf{w})d\mathbf{w} = 0. \tag{5}$$

There are various choices for $\mathbf{K}_d$; however, it can be shown theoretically and empirically that the choice of $\mathbf{K}_d$ is not crucial. Given this fact, the kernel of choice is usually the product kernel

$$\boldsymbol{K}_d(\boldsymbol{w}) = \prod_{i=1}^{d} K(w_i), \tag{6}$$

where K is a 1-dimensional kernel, usually a Gaussian kernel, and H is a diagonal matrix with elements $h_j = s_j h$, such that s_j is the standard deviation of the j^{th} variable. The choice of h is the most important problem in kernel density estimation, and it is usually made using plug-in methods or cross-validation (see e.g. Loader, 1999a).

Thus, by means of kernel density estimation, a "natural" estimator of the conditional density is

$$\hat{f}_{Y|X}(y|x) = \frac{\hat{f}_{X,Y}(x,y)}{\hat{f}_X(x)}, \tag{7}$$

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 $\tilde{h} \to 0$

$$\begin{aligned} E\{K_{\tilde{h}}(Y-y)|X=x\} &= \int_{\mathbb{R}} K_{\tilde{h}}(Y-y) f_{Y|X}(y|x)dy \\ &= \int_{\mathbb{R}} K(u) f(u\tilde{h}+y|x)du \\ &\approx f_{Y|X}(y|x), \end{aligned} \tag{8}$$

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, 1999b).

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

$$\hat{\beta}(x,y) = \arg\min_{\beta} \sum_{i=1}^{n} \left\{K_{\tilde{h}}(Y_i-y) - A(X_i-x,\beta)\right\}^2 W_h(X_i-x) \tag{9}$$

where

$$A(X_i-x,\beta) = \sum_{j=0}^{p} \beta_j \frac{(X_i-x)^j}{j!} \tag{10}$$

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

$$\hat{f}_p(y|x) = A(0,\hat{\beta}(x,y)) = \hat{\beta}_0. \tag{11}$$

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

$$\hat{f}_0(y|x) = \sum_{i=1}^{n} w_i(x) K_{\tilde{h}}(Y_i-y) \tag{12}$$

where

$$w_i(x) = \frac{W_h(X_i-x)}{\sum_{l=1}^{n} W_h(X_l-x)}.$$

The estimator in Eq. (12) was first presented by Hyndman et al. (1996) and corresponds to the estimator in Eq. (7) 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. (11) has a smaller bias than the estimator in Eq. (12) (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. (9), by setting $\beta_0 = \ell(\alpha) = \exp(\alpha)$. The second proposal takes

$$A(X_i-x,\beta) = \exp\left\{\sum_{j=0}^{p} \beta_j (X_i-x)^j\right\} \tag{13}$$

and $\hat{f}_{Y|X}(y|x) = A(0, \tilde{\beta}(x,y)) = \exp\{\tilde{\beta}_0\}$. 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(a) displays the estimate for a set of conditional densities using the double-kernel estimator (see Eq. (9)) with the function A defined as in Eq. (13) and $p = 1$, for the simulated data in Figure 1(b). As can be seen in this figure, the double-kernel estimator fails to detect the changing behavior that the conditional density presents for $x \in (2, 6)$. This situation is due to the fact that this estimator is not considering information locally in the sample space to estimate the local structure. Another disadvantage of this estimator is that we cannot unveil the underlying structure of the data. We propose an approach to conditional density estimation, such that these two drawbacks are solved. We 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).

3 Conditional Density Estimation using Finite Mixture Models

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

$$f_{Y|X}(y|x) = \sum_{i=1}^{k_x} \pi_i(x) g_i(y; \boldsymbol{\theta}_i(x)) \tag{14}$$

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. (14), we 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. We model these "parameter functions" using local likelihood estimation (Loader, 1999b).

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. (14). 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

$$\mathcal{L}_x(\boldsymbol{\beta}) = \sum_{j=1}^{n} w_j(x) \ell(Y_j, \mathcal{A}(x_j - x, \boldsymbol{\beta})), \tag{15}$$

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. (10), $m = 1, \ldots, q_l$, $l = 1, \ldots, k_x$ and q_l the number of parameters of the l^{th} component. The $\boldsymbol{\beta}$'s are vectors of coefficients and

$$w_j(x) = W\left(\frac{x_j - x}{h(x)}\right), \tag{16}$$

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. (15), that is,

$$\begin{aligned}\hat{\boldsymbol{\beta}}(x) &= \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} w_j(x)\ell(Y_j, \mathcal{A}(x_j - x, \boldsymbol{\beta})) \\ &= \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} w_j(x) \log \sum_{i=1}^{k_x} A_{i,1}(x_j - x, \boldsymbol{\beta}_{i,1}) \cdot g_i(Y_j; A_{i,2}(x_j - x, \boldsymbol{\beta}_{i,2}), \ldots, A_{i,q_i}(x_j - x, \boldsymbol{\beta}_{i,q_i})) \quad (17)\end{aligned}$$

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

We are mainly interested in the parameter functions $\boldsymbol{\theta}_i(x)$ and $\pi_i(x)$, and not in their derivatives; therefore, we set the degree of the local approximation to be zero or one, that is, $A(t, \boldsymbol{\beta}) = \boldsymbol{\beta}_0$ or $A(t, \boldsymbol{\beta}) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 t$. Letting $g_i(y; \boldsymbol{\theta}_i) = \phi(y; \mu_i, \sigma_i^2)$, a density function with parameters μ_i and σ_i^2, the number of parameter functions is three for all the mixture components and Eq. (17) can be written as

$$\hat{\boldsymbol{\beta}}(x) = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} w_j(x) \log \sum_{i=1}^{k_x} \beta_{i,1}^0 \cdot \phi(y_j; \beta_{i,2}^0, \beta_{i,3}^0) \quad (18)$$

for $A(t, \boldsymbol{\beta}) = \boldsymbol{\beta}_0$ and

$$\hat{\boldsymbol{\beta}}(x) = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} w_j(x) \log \sum_{i=1}^{k_x} (\beta_{1\,i}^0 + \beta_{1\,1}^1 x) \cdot \phi(y; \beta_{2\,i}^0 + \beta_{2\,i}^1 x, \beta_{3\,i}^0 + \beta_{i,3}^1 x) \quad (19)$$

for $A(t, \boldsymbol{\beta}) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 t$, with $\beta_{l\,i}^k$ is the l^{th} position of the k^{th} vector of coefficients of the i^{th} component.

Notice that when $A(t, \boldsymbol{\beta}) = \boldsymbol{\beta}_0$, the original problem is reduced to solving a finite mixture problem (see e.g., McLachlan and Peel, 2000; Ripley, 1996); therefore, we can 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. This 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 is critical to modify this algorithm or find other approaches. In this paper, we use 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). Before we describe these three algorithms, we introduced some model selection criteria and a similarity measure for densities.

3.1 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

$$\hat{k}_x = \arg\min_k \{\mathcal{C}(\hat{\boldsymbol{\beta}}(x), k), k = k_{x,min}, \ldots, k_{x,max}\} \quad (20)$$

where $\mathcal{C}(\cdot, k)$ is some model selection criterion. There are many choices for $\mathcal{C}(\cdot, k)$, in this paper we make use of the Bayesian Information Criterion (BIC, Schwarz, 1978; Kass and Raftery, 1995) and the Integrated Squared Error (ISE) criterion.

The BIC is defined as

$$BIC(\hat{\boldsymbol{\beta}}(x), k) = -\log \mathcal{L}_x(\hat{\boldsymbol{\beta}}) + \frac{N(k)}{2} \log n_x \tag{21}$$

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

$$ISE(\hat{\boldsymbol{\beta}}(x), k) = \int_{-\infty}^{\infty} (f_k(y|x; \hat{\boldsymbol{\beta}}(x)) - f(y|x))^2 dy \tag{22}$$

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 viewed 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

$$\hat{\boldsymbol{\beta}}(x) = \arg\min_{\boldsymbol{\beta}} -\log \mathcal{L}_x(\hat{\boldsymbol{\beta}}) + N(k) k_{nz} \log n_x + \frac{k_{nz}}{2} \log n_x + \frac{N(k)}{2} \cdot \sum_{j:\pi_j(x)>0} \log \pi_j(x) \tag{23}$$

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

$$\text{sim}(g_1, g_2) = \frac{\int_{-\infty}^{\infty} g_1(t) g_2(t)\, dt}{\left(\int_{-\infty}^{\infty} g_1^2(t)\, dt \int_{-\infty}^{\infty} g_2^2(t)\, dt \right)^{1/2}}, \tag{24}$$

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 $0 \leq \text{sim}(g_1, g_2) \leq 1$.

3.2 Figueiredo and Jain's algorithm

Figueiredo and Jain (2002) noticed that the MML criterion for mixtures in Eq. (23) 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. (23), 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

$$\hat{\pi}_i^{(t+1)}(x) = \frac{\max\left\{0, \sum_{j=1}^{n_x} \gamma_j^{(t)} - \frac{N}{2}\right\}}{\sum_{m=1}^{k_x} \max\left\{0, \sum_{j=1}^{n_x} \gamma_j^{(t)} - \frac{N}{2}\right\}}, \text{ for } i = 1, \ldots, k, \tag{25}$$

where N is the number of parameters that specify each component, and the $\gamma_j^{(t)}$'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.

3.3 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. (24), are combined using the method of moments (MoM); that is, given two components with parameters (w_1, μ_1, σ_1^2) and (w_2, μ_2, σ_2^2), respectively, the new component will have parameters

$$(w_i + w_{i+1}, w_i'\mu_i + w_{i+1}'\mu_{i+1}, w_i'\sigma_i^2 + w_{i+1}'\sigma_{i+1}^2 + w_i'w_{i+1}'(\mu_i - \mu_{i+1})^2) \tag{26}$$

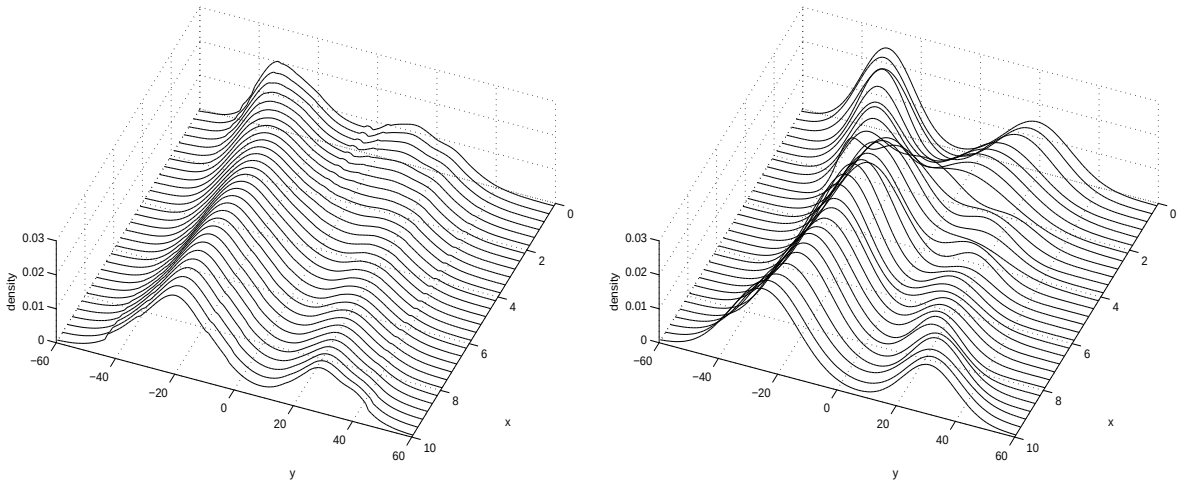
where $w_i' = w_i/(w_i + w_{i+1})$ and $w_{i+1}' = 1 - w_i'$. At this point, we end up with a mixture of $\tilde{k}$ components with the set of parameters $\{(w_1, \mu_1, \sigma_1^2), \ldots, (w_{\tilde{k}}, \mu_{\tilde{k}}, \sigma_{\tilde{k}}^2)\}$, such that $\mu_1 < \ldots < \mu_{\tilde{k}}$. Third, the similarity function in Eq. (24) 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. (26). 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

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{k-1}(x) &= \arg\min_{\boldsymbol{\beta}} \int_{-\infty}^{\infty} (f_{k-1}(y|x; \boldsymbol{\beta}(x)) - f(y|x))^2 \, dy \\
&= \arg\min_{\boldsymbol{\beta}} \int_{-\infty}^{\infty} f_{k-1}(y|x; \boldsymbol{\beta}(x))^2 \, dy - \int_{-\infty}^{\infty} 2 f_{k-1}(y|x; \boldsymbol{\beta}(x)) f(y|x) \, dy + \int_{-\infty}^{\infty} f(y|x)^2 \, dy \\
&= \arg\min_{\boldsymbol{\beta}} \int_{-\infty}^{\infty} f_{k-1}(y|x; \boldsymbol{\beta}(x))^2 \, dy - \int_{-\infty}^{\infty} 2 f_{k-1}(y|x; \boldsymbol{\beta}(x)) f(y|x) \, dy \\
&\approx \arg\min_{\boldsymbol{\beta}} \int_{-\infty}^{\infty} f_{k-1}(y|x; \boldsymbol{\beta}(x))^2 \, dy - 2 \sum_{j=1}^{n_x} w_j(x) f_{k-1}(y_j|x; \boldsymbol{\beta}(x)).
\end{aligned} \tag{27}$$

3.4 Simulated Example

In this section we apply the proposed algorithm using IPRA and Figueiredo and Jain's algorithm (FJEM) to fit the mixture parameters for each given x, to the data shown in Figure 1(b). The BIC was used to select the number of components for each value of X. Figure 2 shows the estimated conditional density of Y given X, while Figure 3 shows the conditional mean for each component as function of x.



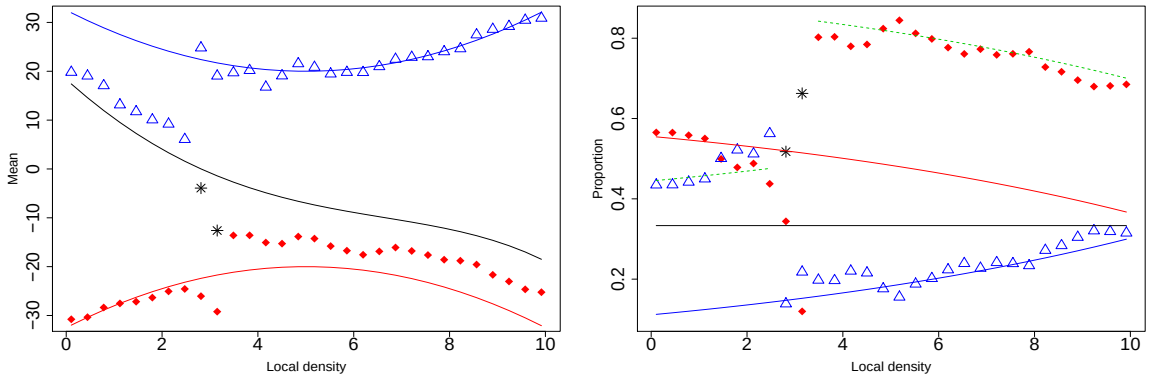
(a) double-kernel with $p = 1$

(b) EM

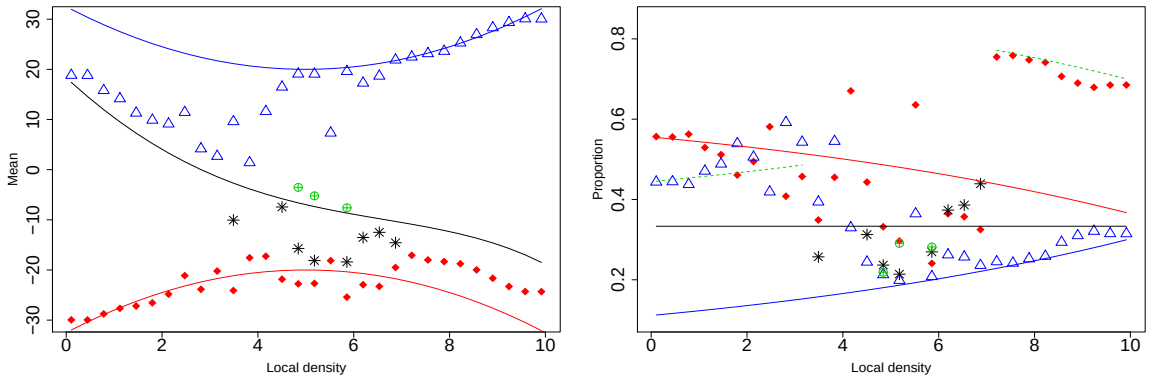
(c) IPRA

(d) FJEM

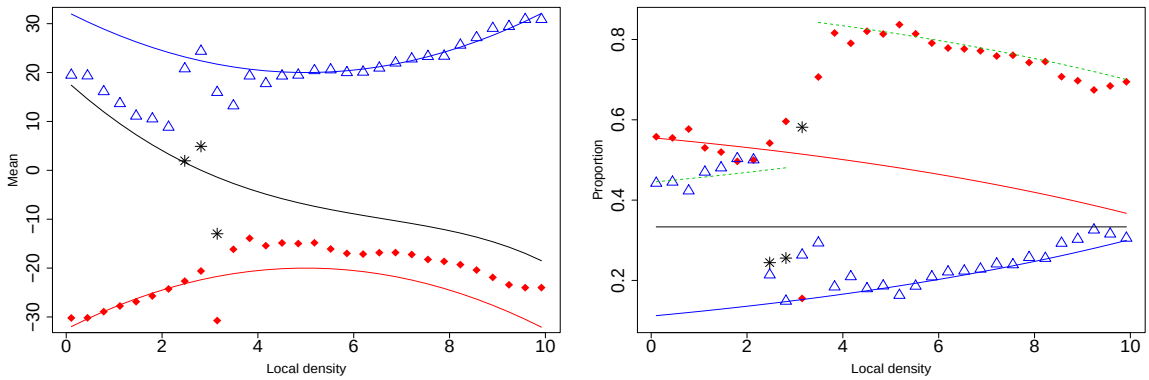
Figure 2: Conditional density estimates for the data in Figure 1(b)



(a) Usual EM



(b) Figueiredo and Jain's algorithm



(c) IPRA

Figure 3: True (solid lines) and estimated mean and proportion functions for the model in Eq. (1). The green (dashed) lines in the second column are the sum of the proportions corresponding to the closes two components.

Notice how the estimates obtained using the EM algorithm, FJEM and IPRA capture the overall structure of the conditional density (see Figure 1(a)), while the double-kernel, as mentioned before, fails. Even though these three estimates capture the general shape of the conditional density, there are some extra bumps. These bumps are due to extra components that these algorithm consider to be important. It is important to mention that for some values of X, the IPRA with three components was very close to the true values; however, the BIC preferred a model with only two components. This situation only happened with IPRA and not with FJEM nor the EM algorithm.

Regarding the parameter functions, the three algorithms determine that there are only two components for most values of x. Notice how these algorithms replace the two closest means by their weighted average for each x. This was somehow expected given that for values of x where two components are very close (less than one standard deviation apart) and the proportion parameter of one of them is small; therefore, the bimodality is not evident (see Figure 1(a)). For values values of x close to three, there is an important change on the structure of the conditional density, which is captured by increasing the number of components to three or even four.

4 Application: galaxy evolution vs. local environment

The galaxy population today is primarily described by two distinct populations. The first are red, elliptically shaped galaxies with little ongoing star-formation. The second are blue, disk-like or morphologically disturbed galaxies with active star-formation. This segregation, while not entirely understood, has been known for a long time.

One of the most fundamental questions in modern astrophysics is how these populations came about. For example, cosmological models indicate that over time, small galaxies will merge into larger and larger systems. While physically, pressure and friction will strip galaxies of the cool gas they need to form new stars. Observationally, it is fairly common to detect highly disrupted galaxy pairs (perhaps merging) undergoing bursts of star-formation. Thus, it has been hypothesized that the old, red population of galaxies formed from less massive star-forming galaxies via this hierarchical structure formation. Alternative theories include top-down approaches, where galaxies fragment from large systems into smaller pieces.

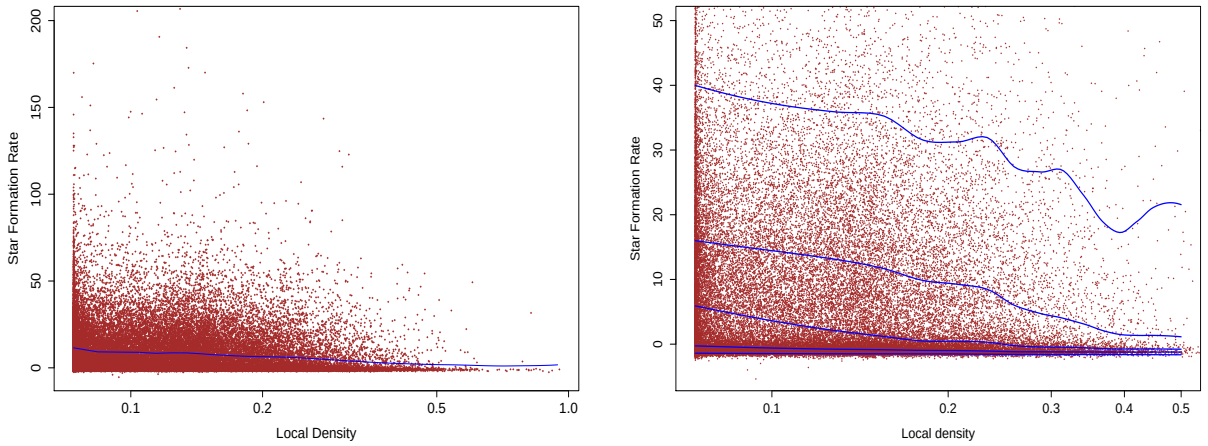
One evolutionary measure that has received recent attention is a galaxy's star formation rate (e.g., Heavens et al., 2004; Gomez et al., 2003; Balogh et al., 2004). The star formation rate (SFR) can be measured by looking back to galaxies with increasing redshifts[1] at different wavelengths. In this work, we use the H_α "emission line" (visible in the galaxy spectra) as an indicator of the recent star-formation in any given galaxy. 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. Thus, the two known populations of galaxies show either H_α emission (star-forming) or H_α absorption (non-star-forming).

We use 47252 galaxies of the Sloan Digital Sky Survey (SDSS) to study the relationship between galaxy evolution and local environment. The available data consist of the X, Y, Z positions of these galaxies, measured in Megaparsecs (Mpc), and their hydrogen emission line equivalent width at the wavelength 6564 $\mathring{\text{A}}$ ($EW(H_\alpha)$). The X, Y, Z positions have been calculated from the galaxy position in the Right Ascension-Declination (RA-Dec) coordinate system and their redshift positions. These calculations have been carried out assuming the following cosmological model: Hubble parameter (or present-day expansion rate of the

[1] A relevant glossary appears at the end of this paper

Universe) $H_0 = 70$, cosmological constant $\lambda = 0.7$ and the matter density $\Omega_M = 0.3$.

To study the role environment plays in the process of galaxy evolution, we map the density field using a kernel density estimator on the point-like spatial galaxy distribution. The bandwidth selection for this estimator was carried out using least-squares cross-validation. All computations were carried out using the software written by Gray and Moore (2003). We can then analyze the H_α emission as a function of environment. The scatter plot of $EW(H_\alpha)$ versus local density is displayed in Figure 4(a). As can be seen in this figure, galaxies located in very dense regions have a low star formation rate, while star-forming galaxies are found in less dense regions.



(a) Scatter plot and estimated conditional mean

(b) Estimated quantile functions, $\alpha = 0.05, .25, .5, .75, .95$

Figure 4: $EW(H_\alpha)$ versus Local density.

The relationship between SFR and local density could be studied by using available quantile regression techniques (Koenker and Bassett, 1978; Koenker et al., 1994; Yu and Jones, 1998). However, when using quantile regression techniques (see Figure 4(b)), we cannot easily draw meaningful conclusions about the underlying cosmology. Plus, it is impossible to determine how galaxy populations interact as galactic systems become denser. Thus, we estimate the conditional density with the hope of finding more meaningful features in this relationship.

Let X be the estimated local density and $Y = EW(H_\alpha) + \lambda$, where λ is a location parameter. We model the conditional density of Y given X using the following model Eq. (14) with

$$g_1(y, \boldsymbol{\theta}_1(x)) = \mathcal{N}(\mu_{1,x}, \sigma^2_{x,1}) \quad (28)$$

$$g_2(y, \boldsymbol{\theta}_2(x)) = \mathcal{LN}(\mu_{2,x}, \sigma^2_{x,2}) \quad (29)$$

$$g_3(y, \boldsymbol{\theta}_3(x)) = \mathcal{LN}(\mu_{3,x}, \sigma^2_{x,3}) \quad (30)$$

where $\mathcal{LN}(\mu, \sigma^2)$ is the log-normal distribution with parameters μ and σ^2. This model was selected, using the BIC and the L_2E criterion, over the set of possible models including from one to five normal components and from zero to four log-normal components. We have included log-normal components because of the long tail of the conditional densities. The location parameter λ was estimated using its profile likelihood (see Figure 5).

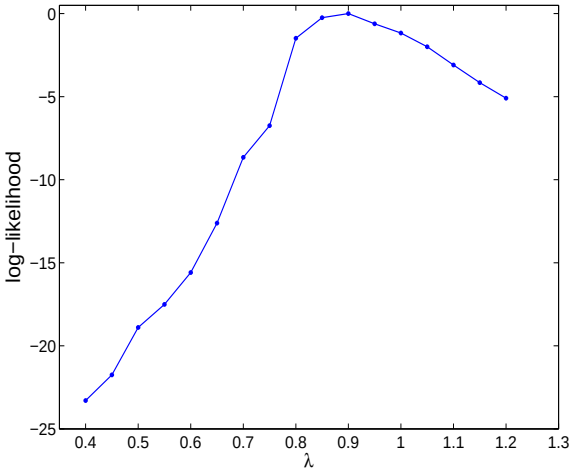


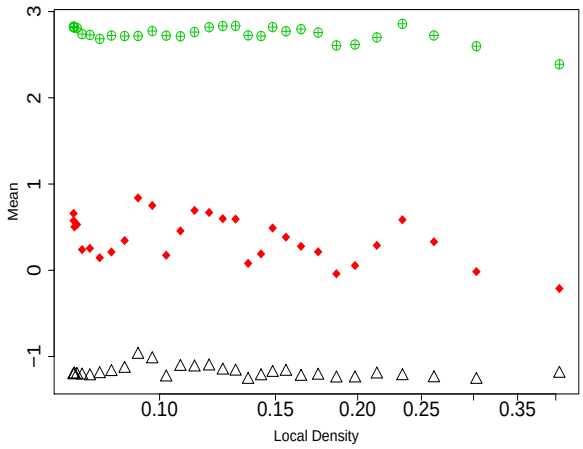
Figure 5: λ's profile likelihood

The estimated parameter functions estimated using the EM algorithm can be found in Figure 6. IPRA and FJEM were not consider because, it is not perfectly clear how to implement these algorithms when mixing normal and log-normal components.

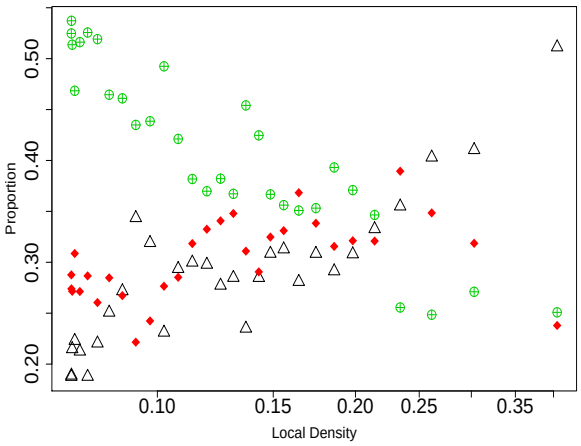
We find that regardless of density, the distribution of H_α is best fit by three components. We detect a high star-formation component and a component corresponding to absorption in H_α. The third component lies in between these two, with a mean (or median) near zero. We can hypothesize that this is either (a) the population of galaxies transitioning from star-forming to non-starforming; (b) a distinct population of another type, for example Active Galaxy Nuclei (AGN) which could be related to (a); or (c) an artifact of our fitting procedure, where the third component may simply indicate that the two main distributions are not well characterized by Gaussians and/or Lognormals.

We examine how the means, dispersions, and proportions of these three populations vary as a function of density. We find that the predominant effect is seen in the proportions, where the fraction of star-forming galaxies decreases with increasing density. Likewise, the population of non-star-forming galaxies increases with increasing 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 stay constant as density changes.

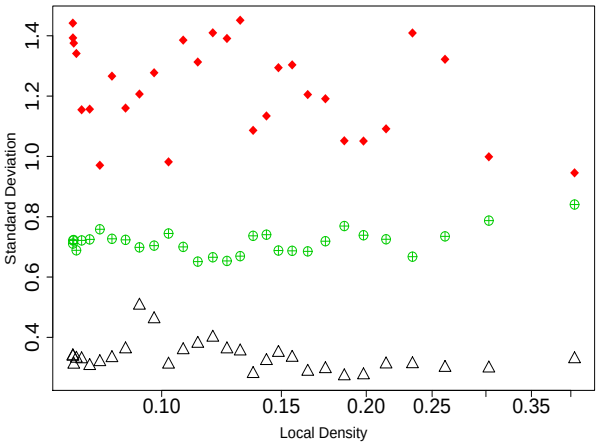
In conclusion, the new, large, and high-quality astronomical datasets (e.g., the SDSS used here), have allowed us to carefully study the distributions of H_α in galaxies as a function of environment. We detect three (not two) components, and we hypothesize on the physical nature of this third component. Likewise, we show that it is the proportions, and not the means or dispersions of the H_α distributions that vary with environment.



(a) Means



(b) Proportions



(c) Standard Deviations

Figure 6: Estimated parameter functions

5 Discussion

We have proposed a new valuable tool to carry out conditional density estimation based on local likelihood estimation and finite mixture models. Our estimator allows us to obtain a better insight of the underlying relationship between random variables than current approaches to conditional density estimation and quantile regression. We have shown this fact with the Astrophysical data presented in Section 4. Even though our estimator produces good results, there is still need for further research. For example, it is not clear which finite mixture model estimator to use. The use of the EM algorithm is natural since it is well understood, easy to implement and we can easily include different family density functions. Plus, it has produced good results in the experiments we have carried out. However, it suffers from some drawbacks that may appear when analyzing other data sets. For this reason we need to further study the other algorithms mentioned in this paper (i.e. IPRA and FJEM) and other such that Dirichlet Process Mixture Models (Escobar and West, 1995).

On future work we shall also concentrate on how to select an "optimal" bandwidth for the weighting function in Eq. (16) and how to estimate the standard deviations of our estimated parameter functions.

Glossary

- One **Parsec** $= 3.085678 \times 10^{16}$m.
- The **RA-Dec coordinate system** is the most natural coordinate system for the stars. Stars are fixed in RA-Dec coordinates, and the coordinate is moving with the sky as time goes by.
- The term **electromagnetic spectrum** refers to the collection of possible wavelengths of electromagnetic radiation.
- A **redshift** is a shift in the frequency of a photon toward lower energy, or longer wavelength. The redshift is defined as the change in the wavelength of the light divided by the rest wavelength of the light.
- The **Cosmological Redshift** is a redshift caused by the expansion of space. The wavelength of light increases as it traverses the expanding universe between its point of emission and its point of detection by the same amount that space has expanded during the crossing time.
- An **emission line** is the name for a portion of the electromagnetic radiation spectrum that is from a unique photonic discharge.

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.

Browman, A. W. (1984), "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, 71, 353 – 360.

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

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.

Escobar, M. D. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577 – 588.

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

Fan, J., Yao, Q., and Tong, H. (1996), "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems," *Biometrika*, 83, 189–206.

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.

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.

Gray, A. and Moore, A. (2003), "Rapid Evaluation of Multiple Density Models," in *Artificial Iintelligence and Statistics*.

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.

Kass, R. E. and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773 – 795.

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

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantiles Smoothing Splines," *Biometrica*, 81, 673–680.

Loader, C. R. (1999a), "Bandwidth selection: classical or plug-in?" *The Annals of Statistics*, 27, 415–438.

— (1999b), *Local Regression and Likelihood*, New York: Springer-Verlag.

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

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge, UK: Cambridge University Press.

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.