Semiparametric Mixtures of Regressions

Penn State Dept. of Statistics Technical Report #11-02

David R. Hunter and Derek S. Young
Department of Statistics, Penn State University

February 9, 2011

Abstract

We present an algorithm for estimating parameters in a mixture-of-regressions model in which the errors are assumed to be independent and identically distributed but no other assumption is made. This model is introduced as one of several recent generalizations of the standard fully parametric mixture of linear regressions in the literature. A sufficient condition for the identifiability of the parameters is stated and proved. Several different versions of the algorithm, including one that has a provable ascent property, are introduced. Numerical tests indicate the effectiveness of some of these algorithms.

1 Introduction

A finite mixture of regressions model is appropriate when regression data are believed to belong to two or more distinct categories, yet the categories themselves are unobserved (as distinct from the so-called analysis of covariance, or ANCOVA, model in which the categorical variable is observed). This situation could arise when a different regression relationship between predictor and response is believed to exist in each category; yet there are also special cases, such as the case in which each category has the same regression relationship but the errors are distributed differently in the different categories (e.g., when a small proportion of the errors might be considered outliers).

The basic mixture-of-regressions model is

$$y_i = \begin{cases} f_1(\mathbf{x}_i) + \epsilon_{i1} & \text{with probability } \lambda_1 \\ \quad\quad\vdots & \\ f_m(\mathbf{x}_i) + \epsilon_{im} & \text{with probability } \lambda_m. \end{cases} \tag{1}$$

As usual, y_i is the response value corresponding to the predictor vector $\mathbf{x}_i$ and ϵ_{ij} is the associated error conditional on the event that the ith observation comes from the jth component (an event with probability λ_j, where we assume λ_j to be positive).

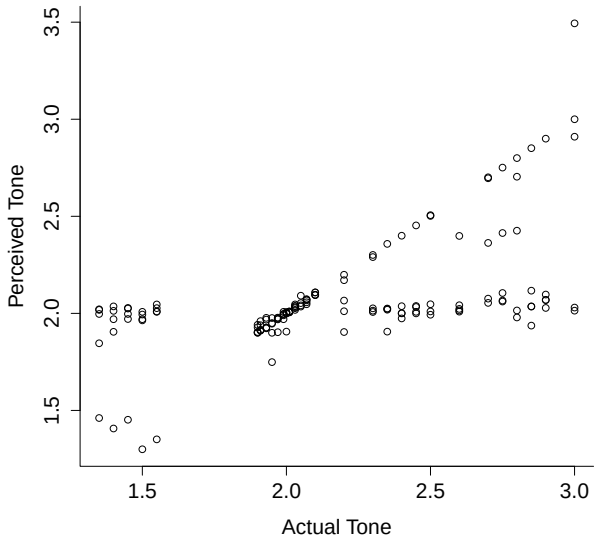

Figure 1: Tone dataset of Cohen (1980).

As a motivating example, the scatterplot of Figure 1 depicts data comparing perceived tone and actual tone for a trained musician, as reported by Cohen (1980). The subject was presented with a fundamental tone plus a series of overtones which are stretched or compressed logarithmically. The subject was asked to tune an adjustable

tone to one octave above the fundamental tone. Two theories of musical perception explored by this study are that the subject would either tune the tones to the nominal octave at a ratio of 2:1 to the fundamental tone (called the *interval memory hypothesis*) or use the overtone to tune the tone to the stretching ratio (called the *partial matching hypothesis*). Since no grouping variable definitively indicates which hypothesis is applicable to the musician for a given tone, modeling these data using ANCOVA is not a possibility. Instead, both DeVeaux (1989) and Viele and Tong (2002) analyzed this dataset by assuming a linear form for the $f_j(x_i)$ functions of Equation (1); i.e., $f_j(x_i) = \beta_0^{(j)} + \beta_1^{(j)} x_i$, $j = 1, 2$.

This article describes the standard parametric mixture-of-linear-regressions model and discusses three recent generalizations, each of which weakens one of the parametric assumptions of the original. The third of these generalizations, which is novel to this article, is discussed in more detail in Section 3. It builds on work on location mixtures of an unspecified symmetric distribution, as introduced by Bordes et al. (2006) and Hunter et al. (2007). We introduce an algorithm for calculating estimates in this model in Section 4 and provide numerical tests of the algorithm in Section 6.

2 Three generalizations

To develop the basic parametric linear mixture of regressions, let us assume that the scalar Y_i is to be regressed on the p-vector $\mathbf{X}_i$ for $1 \le i \le n$, where each $\mathbf{X}_i$ is random with some density, say, $h(\mathbf{x})$; however, as is often the case in regression scenarios, we will largely ignore h and consider only the conditional distribution of Y_i given $\mathbf{X}_i$. We denote by $\delta_{\boldsymbol{\beta}_j}$ the point mass distribution concentrated on the point $\boldsymbol{\beta}_j \in \mathbb{R}^p$.

For parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$, $\lambda_1, \ldots, \lambda_m$, and σ^2, where $\boldsymbol{\beta}_j \in \mathbb{R}^p$, $\sum_j \lambda_j = 1$, $\lambda_j \ge 0$, and $\sigma^2 > 0$, let us assume that

$$\mathbf{B}_i \sim \sum_{j=1}^{m} \lambda_j \delta_{\boldsymbol{\beta}_j}, \tag{2}$$

$$\epsilon_i \sim N(0, \sigma^2), \tag{3}$$

and $\mathbf{X}_i$, $\mathbf{B}_i$, and ϵ_i are jointly independent for each i. Then the basic parametric linear mixture of regressions model may be written as

$$Y_i = \mathbf{X}_i^\top \mathbf{B}_i + \epsilon_i. \tag{4}$$

To estimate the parameters in this model, standard procedures may be applied; for instance, searching for a maximum likelihood estimator is straightforward using a standard EM algorithm for finite mixture models (McLachlan and Peel, 2000). Alternatively, Bayesian methods may be applied, though more care must be exercised when using these methods because of the difficulties presented by label-switching (for instance, see Hurn et al., 2003). The `mixtools` package (Young et al., 2010) for R (R Development Core Team, 2010) includes functions for maximum likelihood estimation and Bayesian estimation for this standard model.

We now present three generalizations of this model, each of which weakens one parametric assumption. The first two of these generalizations are introduced and explored elsewhere, while the third is the subject of the remainder of this article.

1. **Covariate-dependent mixing proportions:**

   Here, we assume that $\mathbf{X}_i$ and $\mathbf{B}_i$ are no longer independent, and in fact each $\lambda_j \equiv \lambda_j(\mathbf{x})$ is some function of the predictor variables. Thus, Equation (2) is replaced by

   $$\mathbf{B}_i | X_i \sim \sum_{j=1}^{m} \lambda_j(\mathbf{X}_i) \delta_{\boldsymbol{\beta}_j}. \tag{5}$$

   Otherwise, Equations (3) and (4) remain unchanged. If $\lambda_j(\mathbf{x})$ is a particular parametric function, namely

   $$\lambda_j(\mathbf{x}) = \frac{\exp\{\mathbf{x}^\top \boldsymbol{\tau}_j\}}{\sum_{\ell=1}^{m} \exp\{\mathbf{x}^\top \boldsymbol{\tau}_\ell\}},$$

   where $\boldsymbol{\tau}_j \in \mathbb{R}^p$ is an unknown *gating* parameter vector, then we get the hierarchical mixtures of experts (HME) model of machine learning; both likelihood-based

(Jordan and Xu, 1995) and Bayesian (Jacobs et al., 1997) estimation methods have been proposed for this model. However, one may alternatively use kernel methods to estimate $\lambda_j(\mathbf{x})$ nonparametrically as in Young and Hunter (2010).

2. **Mixtures of local polynomial regressions:**

   Here, we eliminate the linear regression parameters of Equation (2) and replace $\mathbf{B}_i$ by a random component variable

   $$J_i \sim \sum_{j=1}^{m} \lambda_j \delta_j.$$

   If we were to specify $Y_i = \mathbf{X}_i^\top \boldsymbol{\beta}_{J_i} + \epsilon_i$, then we would obtain the standard mixture of linear regressions. However, we instead merely assume that

   $$Y_i = f_{J_i}(\mathbf{X}_i) + \epsilon_i$$

   for some unspecified functions $f_1, \ldots, f_m$. Issues of identifiability aside, Huang (2009) gives an EM algorithm for estimation of the λ_j and f_j using local likelihood (based on a local polynomial approximation to f_j). In numerical tests, this algorithm appears to perform well. In fact, the algorithm may be extended to the more general case in which the λ_j are assumed to be functions of the predictors $\mathbf{x}_i$, as they are in Equation (5).

3. **Unspecified symmetric error structure:**

   Finally, we assume that Equations (2) and (4) hold, while we replace the parametric assumption (3) by the fully nonparametric

   $$\epsilon_i \sim f, \tag{6}$$

   where f is completely unspecified. Therefore, in this semiparametric model, the conditional distribution of $Y \mid \mathbf{X} = \mathbf{x}$ may be written

   $$g_{\mathbf{x}}(y) = \sum_{j=1}^{m} \lambda_j f(y - \mathbf{x}^\top \boldsymbol{\beta}_j), \tag{7}$$

and the parameters of interest are the λ_j, the $\boldsymbol{\beta}_j$, and f. In the case of regression with an intercept, any location change to f may be absorbed by the intercept parameter, so in this case we may assume without loss of generality that f has median zero.

It is this last of these three semiparametric extensions of the standard mixture-of-regressions model to which we devote the remainder of this article.

3 Nonparametric errors and identifiability

Suppose that $(\mathbf{X}, Y)$ is a multivariate random vector with distribution defined as follows: First, the marginal distribution of $\mathbf{X} \in \mathbb{R}^p$ has (Lebesgue, say) density $h(\mathbf{x})$. Optionally, for regression with an intercept, this distribution guarantees that $X_1 = 1$ and then $\mathbf{X}$ has density $h : \mathbb{R}^{p-1} \to \mathbb{R}$ for its components 2 through p. Second, the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ has density given by Equation (7).

An important question to answer before attempting to estimate parameters in Model (7) is whether the parameters in the model are uniquely identifiable. In this section, we will state and prove a pair of identifiability results. These results make a weak assumption about $h(x)$, namely, that its support contains an open set. However, since the marginal density $h(x)$ may be estimated separately from the parameters in Equation (7), we do not discuss it further, focusing instead on the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$.

To understand why identifiability of the parameters holds in this model, let us temporarily impose a stronger condition on the error density f, namely, that it is symmetric about zero. In a non-regression context, both Bordes et al. (2006) and Hunter et al. (2007) studied univariate mixture models of the form

$$Z \sim \sum_{j=1}^{m} \lambda_j f(z - \mu_j), \tag{8}$$

where f is some symmetric density function. These authors show independently that in the case $m = 2$, the values of λ_j and μ_j and f are uniquely determined, given the distribution of Z, as long as $\lambda_1 \neq 1/2$. Furthermore, Hunter et al. (2007) give sufficient conditions so that when $m = 3$, the parameters are identifiable. These sufficient conditions may be summarized by saying that the values of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ must lie outside a particular subset of $\mathbb{R}^3 \times \mathbb{R}^3$ having Lebesgue measure zero. The conjecture of these authors is that a similar result—identifiability outside of a set having measure zero—holds for general m.

We will argue that in the regression case, even these minor potential impediments to identifiability vanish except in the very particular instance in which two different regression hyperplanes are parallel (i.e., $\boldsymbol{\beta}_j$ is the same as $\boldsymbol{\beta}_k$ for some $j \neq k$ in all but the intercept coordinates). To understand why this is the case, consider Figure 2. In the left-hand plot, we depict two regression lines without intercepts. In this case, we see for one particular value x_0, the distribution of $Y \,|\, X = x_0$ is merely a special case of Model (8). Therefore, if only this conditional distribution were known, it is possible that non-identifiability could occur for parameter values on a set of Lebesgue measure zero. However, any ambiguity in specifying the parameter values is immediately resolved by considering a second value x_1, since only one of the possible sets of parameter values could explain *both* $Y \,|\, X = x_0$ and $Y \,|\, X = x_1$.

The only potential hole in the preceding argument, which will be proven rigorously in Theorem 1, occurs when regression planes are parallel, as shown in the right-hand plot of Figure 2; for in this case, the distribution of $Y \,|\, X$ is independent of X, which means that we are essentially in exactly the situation of Model (8). Thus, we may surmise that identifiability of parameters exists so long as the parameters are identifiable in (8) and there are no two regression planes that are parallel to one another. As these exceptional cases are clearly quite unusual, the use of the estimation algorithm described in Section 4 is certainly justified in practice.

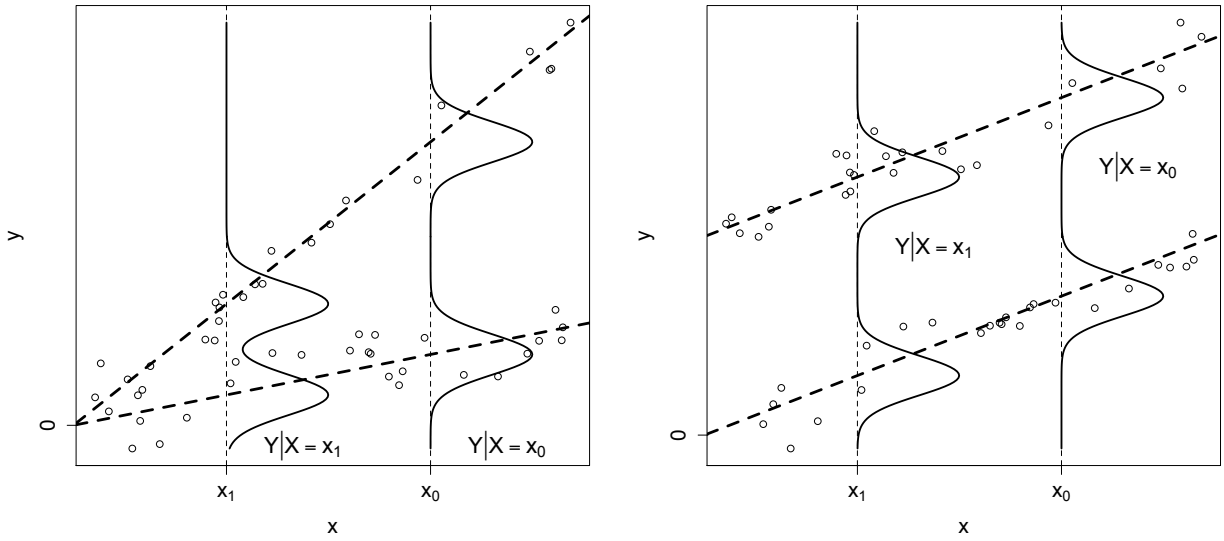

Figure 2: Identifiability follows for the mixture of regressions through the origin (left). However, when an intercept term is added and the lines are parallel (right), the model—and thus the identifiability of parameters—reduces to the non-regression case (8).

We now establish two results that summarize the preceding discussion. The proof of Theorem 1 is given in the Appendix. Denote the joint density by

$$\psi(\mathbf{x}, y) = h(\mathbf{x}) g_{\mathbf{x}}(y) = h(\mathbf{x}) \sum_{j=1}^{m} \lambda_j f(y - \mathbf{x}^\top \boldsymbol{\beta}_j), \tag{9}$$

where $h(\cdot)$ is the marginal density of $\mathbf{X}$ and $g_{\mathbf{x}}(\cdot)$ is the conditional density of $Y|\mathbf{X} = \mathbf{x}$.

Theorem 1 *(Regression without an intercept) If the support of* $\mathbf{X}$ *contains an open set in* $\mathbb{R}^p$*, then all parameters are identifiable; i.e., the left side of Equation (9) uniquely determines the right side.*

Corollary 1 *(Regression with an intercept) If the support of* $\mathbf{X}$ *contains an open subset in* $1 \times \mathbb{R}^{p-1}$ *and* f *is assumed to have median zero, then all parameters in model (9) are identifiable as long as no two of the regression surfaces* $y = \mathbf{x}^\top \boldsymbol{\beta}_j$ *are parallel.*

In other words, identifiability follows as long as no two vectors $(\beta_{j2}, \ldots, \beta_{jp}) \in \mathbb{R}^{p-1}$, $1 \leq j \leq m$, *are equal.*

Remark: The stipulation that the support of $\mathbf{X}$ contains an open set is not necessary for identifiability; it is merely an easy-to-state sufficient condition. For instance, in the left plot of Figure 2, we see that only two distinct support points are required in the case of univariate regression through the origin when $m = 2$. In general, it appears that the minimum number of support points sufficient for identifiability will depend on m and the predictors in some complicated way, so we avoid this question by simply requiring infinitely many support points as implied by the existence of an open set.

4 A semiparametric EM-like algorithm

Assume that we observe data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. As is typical in finite-mixture-model settings, we define the Z_{ij}, $1 \leq i \leq n$ and $1 \leq j \leq m$, to be the indicator that the ith observation comes from the jth mixture component. We do not observe the $\mathbf{Z}_i$, though conceptually we may consider the complete data (in the sense of an EM algorithm) to be $(\mathbf{X}_1, Y_1, \mathbf{Z}_n), \ldots, (\mathbf{X}_n, Y_n, \mathbf{Z}_n)$.

The algorithm we introduce here uses the same intuition as those studied by Benaglia et al. (2009) and Benaglia et al. (2011), though those algorithms were all tailored toward a particular (non-regression) multivariate finite mixture model. Because all of these algorithms bear a strong resemblance to standard EM algorithms for the case of a parametric finite mixture model, we consider them "EM-like" and we retain the so-called "E-step" and "M-step" characteristic of a true EM algorithm.

In the rest of this section, we let $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_m, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, f)$ denote the vector of parameters and t denote the iteration number. Thus, we denote tth-iteration parameters as $\boldsymbol{\theta}^t$, λ_j^t, $\boldsymbol{\beta}_j^t$, and f^t.

- **The E-step:** The "E-step" at the tth iteration consists of finding the so-called

"posterior" probabilities

$$\begin{aligned} p_{ij}^t &\stackrel{\text{def}}{=} P(Z_{ij} = 1 | \text{data}, \boldsymbol{\theta}^t) \\ &= \frac{\lambda_j^t f^t(y_i - \mathbf{x}_i^\top \beta_j^t)}{\sum_{\ell=1}^m \lambda_\ell^t f^t(y_i - \mathbf{x}_i^\top \beta_\ell^t)}. \end{aligned} \tag{10}$$

Because p_{ij}^t depends on all of the other parameters, it is often easiest in practice to skip the first E-step, instead initializing the algorithm by requiring the values of p_{ij}^0 to be given by the user and proceeding to update the parameters in the M-step and the density estimation step. Note that this choice is reflected in the notation, as the tth iteration p_{ij} values depend on the tth iteration parameter values. In other words, our algorithm is actually more of an "ME" algorithm than an "EM" algorithm in practice, in the sense that the E-step is actually the *last* update made during each iteration.

- **The M-step:** In the M-step, the Euclidean parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ are updated. As usual in a finite mixture EM algorithm, each λ_j is the mean of the corresponding posteriors p_{ij}:

$$\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^n p_{ij}^t. \tag{11}$$

However, the update of $\boldsymbol{\beta}_j$ is not as straightforward. In a typical EM algorithm, the updates depend on maximization of an expected conditional log-likelihood function. Here, however, due to the absence of a parametric assumption about the errors, there is no obvious function to maximize. One possibility is to do the best we can in maximizing a nonparametric version of the log-likelihood by setting

$$\boldsymbol{\beta}_j^{t+1} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^n p_{ij}^t f^t(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}). \tag{12}$$

Other possibilities are using least-squares or minimum-L_1 estimators despite the

lack of a likelihood:

$$\boldsymbol{\beta}_j^{t+1} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} p_{ij}^t (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \tag{13}$$

$$\boldsymbol{\beta}_j^{t+1} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} p_{ij}^t |y_i - \mathbf{x}_i^\top \boldsymbol{\beta}|. \tag{14}$$

In our numerical examples of Section 6, we use Equation (13) because it is the most straightforward computationally—it merely involves weighted least squares. We will argue in the discussion section that Equation (12) also has merit based on the smoothed likelihood ideas of Levine et al. (2010) and Chauveau et al. (2010). However, Equation (12) has two drawbacks: First, it requires a numerical optimization, which can be difficult; and second, it depends on the parameter f^t, unlike either (13) or (14), which means that the iterative algorithm cannot be initialized using *only* the p_{ij}^0 values as discussed below Equation (10).

- **The density estimation step:**

  We now employ a third step - a density estimation step. Technically, this step could be considered part of the M-step, though we separate it here due to the fact that this density estimation does not actually maximize an objective function. (However, see Section 5.)

  The density update is done using a form of a weighted kernel density estimate. For a given bandwidth h and kernel density $K(\cdot)$, we take

$$f^{t+1}(u) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^t K\left(\frac{u - y_i + \mathbf{x}_i^\top \boldsymbol{\beta}_j^t}{h}\right). \tag{15}$$

  It is possible to update $f(\cdot)$ while enforcing certain constraints, if this is desired. For instance, the assumption of a symmetric error density may be implemented by defining

$$f^{t+1}(u) = \frac{1}{2nh} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^t \left\{ K\left(\frac{u - y_i + \mathbf{x}_i^\top \boldsymbol{\beta}_j^t}{h}\right) + K\left(\frac{u + y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j^t}{h}\right) \right\}. \tag{16}$$

Alternatively, the common assumption that $E(\epsilon_i) = 0$ may be enforced by defining

$$\mu^{t+1} = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^t \left(y_i - \mathbf{x}_i^\top \beta_j^t \right) \tag{17}$$

and then taking

$$f^{t+1}(u) = \frac{1}{2nh} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^t K \left(\frac{u - \mu^{t+1} + y_i + \mathbf{x}_i^\top \boldsymbol{\beta}_j^t}{h} \right). \tag{18}$$

A similar modification to ensure that the median of $f(\cdot)$ is zero may be implemented by redefining μ^{t+1} to be the weighted median of the residuals $y_i - \mathbf{x}_i^\top \beta_j^t$. Each of these methods of calculating μ^{t+1}—both the weighted mean (17) and the weighted median—implicitly assumes that the kernel function $K(\cdot)$ is symmetric about zero, which is usually a reasonable assumption.

Remark: It is possible to create a stochastic version of this algorithm, as introduced in a slightly different context by Bordes et al. (2007), by replacing the p_{ij} by randomly generated indicators Z_{ij}^* in Equations (12) through (18). These Z_{ij}^* should be generated at each iteration so that for every i, exactly one Z_{ij}^* equals one, and the rest are zeros, such that $P(Z_{ij}^* = 1) = p_{ij}^t$. Essentially, this approach randomly reassigns each observation to exactly one of the mixture components for the purposes of performing the EM updates. The approach we present here is in some sense splitting each observation among all of the components according to the p_{ij}^t weights.

- **The bandwidth update step (optional):**

  In many kernel density estimation problems, choosing a bandwidth is somewhat tricky and this choice can have a strong impact on the estimates obtained. The usual difficulties are even more pronounced in the case of a finite mixture, since even some standard rules of thumb become impossible to apply in that case. In an article about a similar EM-like algorithm for nonparametric multivariate finite mixtures, Benaglia et al. (2011) address this issue and describe an algorithm that

recalculates the bandwidth at each iteration of the algorithm. This is particularly helpful once the algorithm has begun to identify the mixture structure, since at that stage the mixture information can be exploited in order to apply standard kernel density estimation techniques.

Here, we describe a possible update to the bandwidth that may, if desired, be inserted into each iteration of our algorithm. To wit, we reset the bandwidth h as follows:

$$h^{t+1} = \frac{0.90}{n^{1/5}} \min \left\{ \sigma^{t+1}, \frac{IQR^{t+1}}{1.34} \right\}. \tag{19}$$

Here,

$$\sigma^{t+1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{t} (y_i - \mathbf{x}_i^{\top} \boldsymbol{\beta}_j^{t+1})^2} \tag{20}$$

is an estimate of the standard deviation of the error density based on the residuals at the tth iteration, and IQR^{t+1} is an estimate of the interquartile range that is similarly based on the weighted residuals; see section 3 of Benaglia et al. (2011) for details of the IQR^{t+1} calculation. The update in Equation (19) is an implementation of the rule of thumb advocated by Silverman (1986, p. 46); as an alternative, changing the factor 0.90 to 1.06 gives the rule presented by Scott (1992, Section 6.5). Our estimation software, described in Section 6, allows the user to either set (and fix) the bandwidth or, alternatively, use the iterative update formula (19) with an arbitrary value of the constant factor (the default is 0.90).

5 Maximum smoothed likelihood estimation

The algorithm we present in Section 4 is not a true EM algorithm since there is no likelihood function that may be shown to increase at each iteration. Nonetheless, using recent work of Levine et al. (2010) and Chauveau et al. (2010) as a guide, it is possible

to adapt this algorithm slightly to produce a new algorithm that does increase the value of a smoothed version of the loglikelihood at each iteration.

To this end, we first define the nonlinear smoothing operator

$$\mathcal{N}_h f(x) \;=\; \exp \int \frac{1}{h} K\left(\frac{x-u}{h}\right) \log f(u)\, du.$$

Next, we define a smoothed version of the log-likelihood function of the parameters:

$$\ell_{\text{smoothed}}(\boldsymbol{\lambda}, \boldsymbol{\beta}, f) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{m} \lambda_j \mathcal{N}_h f(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j) \right]. \tag{21}$$

It is now possible to show that a new algorithm, closely resembling that of Section 4, may be defined in such a way that it possesses the desirable *ascent property* enjoyed by all true EM algorithms:

$$\ell_{\text{smoothed}}(\boldsymbol{\lambda}^{t+1}, \boldsymbol{\beta}^{t+1}, f^{t+1}) \geq \ell_{\text{smoothed}}(\boldsymbol{\lambda}^{t}, \boldsymbol{\beta}^{t}, f^{t}), \tag{22}$$

where the subscripted t is the iteration number just as in Section 4. When estimates are obtained from an algorithm that may not be shown to optimize any particular objective function, these estimates are only implicitly defined. With the objective function (21), however, such estimates may be viewed as maximizers of ℓ_{smoothed}. This leads to the possibility that asymptotic results might be possible with further research.

The method of proof of this descent property is introduced by Levine et al. (2010). It is extended by Corollary 1 of Chauveau et al. (2010) to the case of a symmetric error distribution. We do not reprint these proofs here. The algorithm, which very much resembles an EM algorithm, is actually an example of a generalization of EM called a minorization-maximization (MM) algorithm. The class of MM algorithms generalizes the EM algorithms in the sense that in every EM algorithm, the E-step may be shown to be a minorization step. Although a thorough discussion of MM algorithms is beyond the scope of this article, Hunter and Lange (2004) provides an introduction to them and contains citations to many other articles.

The modified algorithm operates according to the following steps for $t = 0, 1, \ldots$:

- **Minorization step:**

$$p_{ij}^t = \frac{\lambda_j^t \mathcal{N}_h f^t(y_i - \mathbf{x}_i^\top \beta_j^t)}{\sum_{\ell=1}^m \lambda_\ell^t \mathcal{N}_h f^t(y_i - \mathbf{x}_i^\top \beta_\ell^t)}. \tag{23}$$

- **Maximization step, part 1:**

$$\lambda_j^{t+1} = \frac{1}{n} \sum_{i=1}^n p_{ij}^t.$$

- **Maximization step, part 2:**

$$f^{t+1}(u) = \frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^m p_{ij}^t K\left(\frac{u - y_i + \mathbf{x}_i^\top \boldsymbol{\beta}_j^t}{h}\right). \tag{24}$$

- **Maximization step, part 3:**

$$\boldsymbol{\beta}_j^{t+1} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^n p_{ij}^t \mathcal{N}_h f^{t+1}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}). \tag{25}$$

The last step, maximization with respect to $\boldsymbol{\beta}$, may present some numerical challenges, though these can be largely overcome if a generic optimizer, such as the `optim` function in R (R Development Core Team, 2010), is used.

Though the proof of the ascent property follows from the arguments in Levine et al. (2010), Chauveau et al. (2010) point out that the presence of part 3 of the maximization step means that technically, the algorithm above is probably best categorized as a minorization-conditional maximization algorithm. Here, this MCM algorithm is a generalization of the expectation-conditional maximization (ECM) paradigm of Meng and Rubin (1993).

It is possible to modify the above algorithm to allow for a symmetric error density f while preserving the important ascent property: To do so, we simply replace Equation (24) by

$$f^{t+1}(u) = \frac{1}{2nh} \sum_{i=1}^n \sum_{j=1}^m p_{ij}^t \left\{ K\left(\frac{u - y_i + \mathbf{x}_i^\top \boldsymbol{\beta}_j^t}{h}\right) + K\left(\frac{u + y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j^t}{h}\right) \right\}. \tag{26}$$

Corollary 1 of Chauveau et al. (2010) proves that when Equation (26) is used in place of (24), the resulting algorithm is still a minorization-maximization algorithm that guarantees the descent property.

6 Numerical examples

6.1 Cohen data

We apply the semiparametric EM algorithm of Section 4 to the Cohen (1980) data. The density estimation step is performed once when assuming zero-symmetric error densities and once without assuming symmetric errors, each time using the least-squares $\boldsymbol{\beta}$ update of Equation (13). The estimates for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are quite similar under each constraint. The left-hand side of Figure 3 shows the fitted regressions when assuming zero-symmetric error densities as well as the standard EM algorithm estimates for the parametric mixture of linear regressions model. The corresponding estimates are also reported in Table 1.

Parameter	Parametric EM	SP EM (Zero-Symmetric)	SP EM (No Symmetry)
$\boldsymbol{\beta}_1$	1.916	1.775	1.753
	0.043	0.119	0.130
$\boldsymbol{\beta}_2$	-0.019	0.021	-0.030
	0.992	0.979	1.006
λ_1	0.698	0.676	0.678

Table 1: Estimates for the Cohen (1980) data obtained from the EM algorithm for the parametric mixture of linear regressions approach as well as the semiparametric "EM-like" algorithms.

The kernel-based estimator of the residual density function should have variance $h^2 + \sigma^2_{\text{NP}}$, where σ_{NP} is the expression in Equation (20). Figure 3 compares a mean-zero normal density having this variance to the nonparametrically estimated error

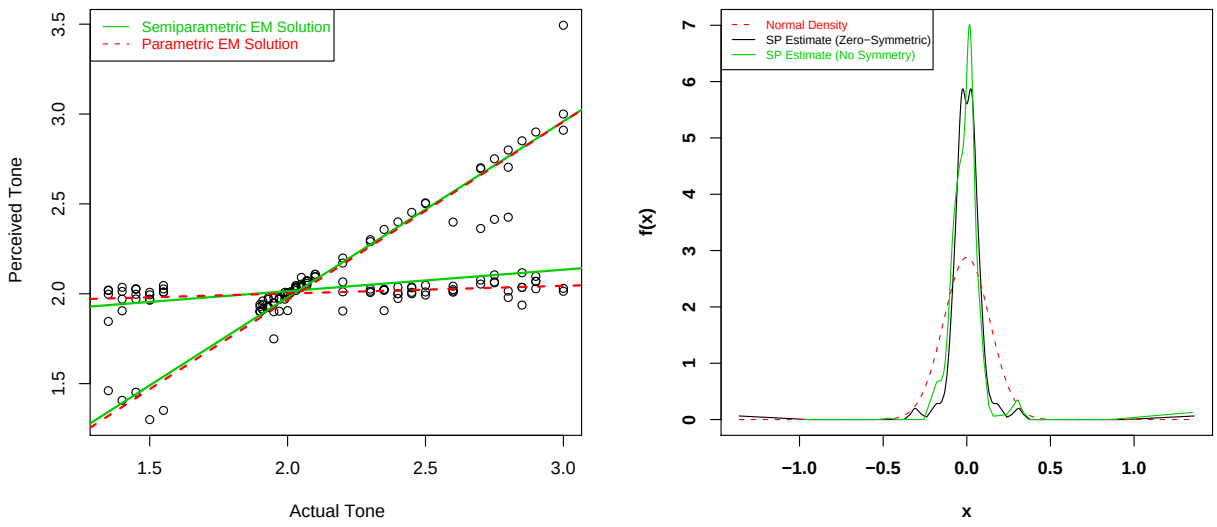

Figure 3: The Cohen (1980) data with the parametric mixtures of linear regressions EM fit and the zero-symmetric semiparametric mixtures of regressions EM fit (left). Kernel density-based estimates of the error density and the normal density with mean 0 and standard deviation $\sqrt{h^2 + \sigma^2_{\text{NP}}}$ (right). For both semiparametric fits, the final bandwidth chosen by the algorithm is $h = 0.021$.

density, both with and without enforcement of the zero-symmetry assumption. The nonparametric estimates have heavier tails than the normal density, which is certainly to be expected in regression situations where outliers may be present, though this difference evidently does not affect the regression parameter estimates strongly.

One aspect of the semiparametric algorithm that does appear to affect the estimation for the Cohen dataset is the choice of $\boldsymbol{\beta}$ update. For these data, when the nonparametric update in Equation (12) is used, we find that many (randomly-generated) choices for the starting parameters lead to a solution with essentially one component, where both regression lines coincide with the more horizontal component shown in Figure 3. This is unsurprising since the residuals that clearly belong to the second component in the figure, though they would be far too large using a normal errors

assumption, are easily accommodated by a fully nonparametric error model. However, in our experience, if the semiparametric algorithm is "directed" toward the more visually obvious two-component solution for a few iterations, it has no trouble identifying the two components, a fact that reveals the particular importance of the starting parameter values when using Equation (12). We find that a hybrid algorithm, which uses a weighted average of Equations (13) and (12) in which the weights begin nearly completely in favor of the least-squares update and evolve to completely in favor of the nonparametric update, works well. However, the tuning of this hybrid algorithm could be a topic for future investigation.

6.2 Heavy-tailed errors

To compare our algorithm with a standard parametric EM algorithm in a situation where the is parametric normal-error assumption is known to be incorrect, we simulated 100 (x, y) pairs from a 2-component mixture-of-regressions model with error terms distributed according to a t_3 distribution. The realizations of the predictor variable, $x_1, \ldots, x_{100}$, were chosen to be an equally-spaced set of values over the closed interval $[0, 10]$. The response values were generated according to

$$y_i = \begin{cases} 1 + 6x_i + \epsilon_i & \text{with probability } 0.25 \\ 8 + 2x_i + \epsilon_i & \text{with probability } 0.75, \end{cases} \tag{27}$$

where the ϵ_i are independent t_3 random variables.

The standard mixtures-of-regressions EM algorithm assuming normal errors with equal component variances and several versions of the semiparametric approach are applied to each of the 1000 simulated datasets. The semiparametric approach may assume errors to be either non-symmetric, as in Equation (15), or symmetric, as in Equation (16); it also uses either the nonparametric $\boldsymbol{\beta}$ update (12) or the least-squares update (13). In order to try to eliminate the effect of the choice of starting value on the algorithms so as to compare the "best-case" results of the various estimation methods, we started the parametric EM algorithm at the true values of $\boldsymbol{\beta}$ and we started the

Algorithm	$\hat{\lambda}_1$	$\hat{\boldsymbol{\beta}}_1$	$\hat{\boldsymbol{\beta}}_2$
Parametric EM assuming normal errors	0.00242*	1.93*	0.17*
Symmetric errors, nonparametric $\boldsymbol{\beta}$ update	0.00231	0.60	0.15
No symmetric errors, nonparametric $\boldsymbol{\beta}$ update	0.00235	0.61	0.14
Symmetric errors, least-squares $\boldsymbol{\beta}$ update	0.00230*	0.66*	0.20*
No symmetric errors, least-squares $\boldsymbol{\beta}$ update	0.00235*	0.65*	0.21*

Table 2: Mean squared distance between estimates and true parameters in 1000 trials of five different algorithms for the t_3-error example. Values marked with asterisks were calculated after dropping one dataset that contained an extreme outlier.

nonparametric algorithms using the correct component assignments for each of the data points.

As seen in Table 2, the lowest mean-squared errors were achieved by the semiparametric algorithm using the nonparametric update (12) of $\boldsymbol{\beta}$. The difference between the symmetric (16) and non-symmetric (15) errors appears to be negligible. The least-squares update (13) used in the fourth and fifth rows of Table 2 produces estimates that sometimes resemble the pure nonparametric estimates of rows 2 and 3 and sometimes are closer to the parametric estimates of row 1. In one of the 1000 datasets, an extreme outlier completely ruined the estimates for the algorithms in rows 1, 4, and 5; in particular, including this dataset raises the mean squared error for $\boldsymbol{\beta}_2$ above 160 for each of these rows.

The algorithms using the fully nonparametric update including Equation (12) are the clear winners in this particular test; however, we offer some mitigating observations based on our experience. For one thing, both the initial programming effort and the computing time per iteration required for the maximization of Equation (12) is much greater than those required for Equation (13). Secondly, we find that the parametric EM algorithm assuming normal errors is surprisingly robust in a number of non-normal error situations. For instance, t-distributed errors with 5 or more degrees of freedom seem to make the parametric EM competitive with the semiparametric algorithms.

Semiparametric EM-like

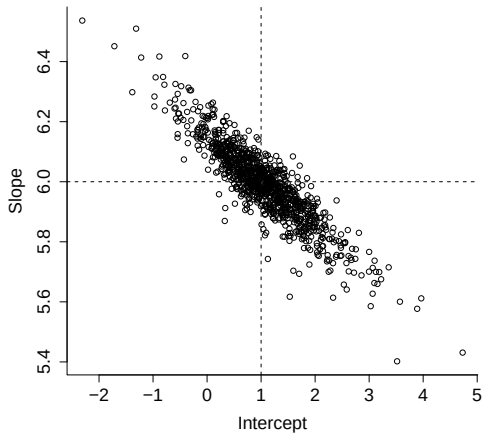

Figure 4: First-component pairs (β_0, β_1) from the Section 4 algorithm using Equations (12) and (15). The dotted lines mark the true parameter values.

Parametric EM

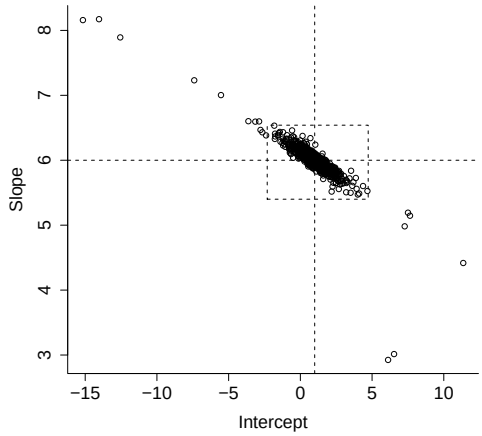

Figure 5: As in Figure 4 but using an EM algorithm assuming normal errors. The dotted rectangle would be just large enough to contain all of the points in Figure 4.

Finally, as we noted in Section 6.1, the fully nonparametric algorithm is so flexible that sometimes, depending on how it is started, it misses the mixture structure entirely and simply classifies some of the signal in the data as noise. The last observation suggests that further work on choosing starting values would be fruitful.

7 Discussion

This article discusses several nonparametric extensions to the standard parametric mixture-of-regressions model. For one of these extensions, which removes all assumptions about the parametric form of the residuals, we provide a proof of identifiability as well as several possible algorithmic approaches to performing estimation. Judging from tests of several of these approaches on actual and simulated data, they appear very effective.

One gap in the identifiability result is the fact that regression hyperplanes are not permitted to be parallel in the case of regression with an intercept. However, as this is the only potential identifiability concern and it only eliminates a subset of the parameter space having Lebesgue measure zero from the set of uniquely identifiable parameters, it is not clear whether this gap has important practical consequences.

Choosing the bandwidth h is a practical challenge in implementing the algorithms described here. Here, we used simplistic bandwidth estimates based on rules of thumb presented in Silverman (1986) and Scott (1992), but these guidelines are tricky to implement in the mixture setting until something is known about which components each observation might belong to. This suggests that an iteratively updated bandwidth is possible. Though we do not discuss this issue here, recent work by Benaglia et al. (2011) and Chauveau et al. (2010) does so in detail for related algorithms in a non-regression setting.

The choice of how to update $\boldsymbol{\beta}$ in the maximization step of the algorithm of Section 4 is not clear. In our experience, the fully nonparametric update (12) gives robust results, yet it has the drawbacks that it is much more difficult than the least-squares update (13) to program and it takes much longer to calculate. If one were to implement the algorithm of Section 5, the programming and computational burden might be even greater. On the other hand, this latter approach, since it may be shown to optimize a nonlinearly smoothed log-likelihood function, holds the promise that such an algorithm might yield theoretical large-sample results such as consistency or a particular rate of convergence. Therefore, this article opens up the possibility of a whole range of algorithms for fitting mixture-of-regression models: On one hand, there is the traditional parametric EM algorithm, which is quick but nonrobust, and on the other are various, more flexible alternatives that possess differing degrees of robustness against unusual error distributions.

References

Benaglia, T., Chauveau, D., and Hunter, D. R. (2009). An EM-Like Algorithm for Semi- and Non-Parametric Estimation in Multivariate Mixtures. *Journal of Computational and Graphical Statistics*, 18:505–526.

Benaglia, T., Chauveau, D., and Hunter, D. R. (2011). Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures. In Hunter, D. R., Richards, D. S. P., and Rosenberger, J. L., editors, *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger*, pages 15–27. World Scientific, Singapore.

Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, 51(11):5429–5443.

Bordes, L., Mottelet, S., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Annals of Statistics*, 34(3):1204–1232.

Chauveau, D., Hunter, D. R., and Levine, M. (2010). Estimation for conditional independence multivariate finite mixture models. Technical Report 10–06, Pennsylvania State University.

Cohen, E. (1980). *Inharmonic Tone Perception*. Phd dissertation.

DeVeaux, R. D. (1989). Mixtures of linear regressions. *Computational Statistics and Data Analysis*, 8(3):227–245.

Huang, M. (2009). *Nonparametric Techniques in Finite Mixture of Regressions Model*. Phd dissertation, The Pennsylvania State University.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58:30–37.

Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.*, 35(1):224–251.

Hurn, M., Justel, A., and Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79.

Jacobs, R. A., Peng, F., and Tanner, M. A. (1997). A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks*, 10(2):231–241.

Jordan, M. I. and Xu, L. (1995). Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8(9):1409–1431.

Levine, M., Hunter, D. R., and Chauveau, D. (2010). Maximum smoothed likelihood for multivariate mixtures. Technical Report 10–04, Pennsylvania State University.

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

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Scott, D. W. (1992). *Multivariate Density Estimation*. John Wiley & Sons Inc., New York.

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

Viele, K. and Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330.

Young, D. S., Benaglia, T., Chauveau, D., Hunter, D. R., Elmore, R. T., Xuan, F., Hettmansperger, T. P., and Thomas, H. (2010). *The mixtools Package: Tools for Mixture Models*. R Package Version 0.4.4.

Young, D. S. and Hunter, D. R. (2010). Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics and Data Analysis*, 54:2253–2266.

A Identifiability

Proof of Theorem 1: Let us consider only the conditional distribution of $Y|\mathbf{X}$ for the moment.

Let

$$\Delta_m^p(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{j=1}^{m} \lambda_j \delta_{\boldsymbol{\beta}_j}$$

denote the discrete distribution putting mass λ_j on $\boldsymbol{\beta}_j \in \mathbb{R}^p$, $1 \leq j \leq m$, as seen in Equation (2). If V and $\mathbf{W}$ are independent random variables such that $V \sim f$ and $\mathbf{W} \sim \Delta_m^p(\boldsymbol{\lambda}, \boldsymbol{\beta})$, then $Y|\mathbf{X}$ is distributed as $V + \mathbf{W}^\top \mathbf{X}$. Let $\Phi_Z(t)$ denote the characteristic function of an arbitrary random Z. Then

$$\phi_{Y|\mathbf{X}}(t) = \phi_V(t)\phi_{\mathbf{W}^\top \mathbf{X}}(t) = \phi_V(t)\phi_{\mathbf{W}}(t\mathbf{X}). \tag{28}$$

Note that $\phi_{\mathbf{W}(\cdot)}$ is a function from $\mathbb{R}^p \to \mathbb{C}$.

Let $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\beta}^{(1)}, f^{(1)}, h^{(1)})$ and $(\boldsymbol{\lambda}^{(2)}, \boldsymbol{\beta}^{(2)}, f^{(2)}, h^{(2)})$ denote two sets of parameters for model (9). In other words, for $i \in \{1, 2\}$, $f^{(i)}(\cdot)$ is arbitrary, $\boldsymbol{\beta}_j^{(i)} \in \mathbb{R}^p$ for $1 \leq j \leq m$, and $\boldsymbol{\lambda}^{(i)} \in \mathbb{R}^m$ has nonnegative entries that sum to one. By definition, the parameters of model (9) are identifiable if

$$h^{(1)}(\mathbf{x}) \sum_{j=1}^{m} \lambda_j^{(1)} f^{(1)}(y - \mathbf{x}^\top \boldsymbol{\beta}_j^{(1)}) = h^{(2)}(\mathbf{x}) \sum_{j=1}^{m} \lambda_j^{(2)} f^{(2)}(y - \mathbf{x}^\top \boldsymbol{\beta}_j^{(2)}) \quad \text{a.e. } (y, \mathbf{x}) \tag{29}$$

implies

$$(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\beta}^{(1)}, f^{(1)}(y), h^{(1)}(\mathbf{x})) = (\boldsymbol{\lambda}^{(2)}, \boldsymbol{\beta}^{(2)}, f^{(2)}(y), h^{(2)}(\mathbf{x})) \quad \text{a.e. } (y, \mathbf{x}).$$

Integrating with respect to y, if equation (29) is true, then we obtain $h^{(1)}(\mathbf{x}) = h^{(2)}(\mathbf{x})$ a.e. $\mathbf{x}$. Therefore, the superscripts are not needed on $h(\mathbf{x})$ and we may ignore the marginal distribution of $\mathbf{X}$ when considering identifiability.

Suppose that $(V_1, \mathbf{W}_1)$ and $(V_2, \mathbf{W}_2)$ are two pairs of independent random variables such that $V_i \sim f^{(i)}$ and $W_i \sim \Delta_m^p(\boldsymbol{\lambda}^{(i)}, \boldsymbol{\beta}^{(i)})$ for $i \in \{1, 2\}$, where we also assume that no two of the $\beta_j^{(1)}$ vectors are the same and no two of the $\beta_j^{(2)}$ vectors are the same. Then Equation (29) implies

$$V_1 + \mathbf{W}_1^\top \mathbf{x} \stackrel{\mathcal{D}}{=} V_2 + \mathbf{W}_2^\top \mathbf{x} \tag{30}$$

for almost all $\mathbf{x}$ in the support of $h(\cdot)$. By Equation (28), there exists $\mathbf{x}_0$ such that

$$\phi_{V_1}(t)\phi_{\mathbf{W}_1}(t\mathbf{x}_0) = \phi_{V_2}(t)\phi_{\mathbf{W}_2}(t\mathbf{x}_0) \quad \text{for all } t \in \mathbb{R}. \tag{31}$$

In particular, since all characteristic functions equal one at $t = 0$, there exists $\epsilon > 0$, depending on $\mathbf{x}_0$, such that each of the four characteristic functions in Equation (31) is nonzero for $t \in (-\epsilon, \epsilon)$. Therefore,

$$\phi_{V_1}(t) = \phi_{V_2}(t)\frac{\phi_{\mathbf{W}_2}(t\mathbf{x}_0)}{\phi_{\mathbf{W}_1}(t\mathbf{x}_0)} \quad \text{for all } -\epsilon < t < \epsilon, \tag{32}$$

so for almost all $\mathbf{x} \neq \mathbf{x}_0$ in the support of $h(\cdot)$, Equations (28) and (32) imply

$$\phi_{\mathbf{W}_2}(t\mathbf{x}_0)\phi_{\mathbf{W}_1}(t\mathbf{x}) = \phi_{\mathbf{W}_1}(t\mathbf{x}_0)\phi_{\mathbf{W}_2}(t\mathbf{x}) \quad \text{for all } -\epsilon < t < \epsilon. \tag{33}$$

Letting i denote $\sqrt{-1}$, Equation (33) may be written explicitly: For all $t \in (-\epsilon, \epsilon)$,

$$\sum_{j=1}^{m}\sum_{k=1}^{m} \lambda_j^{(2)}\lambda_k^{(1)} \exp\{it(\mathbf{x}_0^\top \boldsymbol{\beta}_j^{(2)} + \mathbf{x}^\top \boldsymbol{\beta}_k^{(1)})\} =$$
$$\sum_{j=1}^{m}\sum_{k=1}^{m} \lambda_j^{(1)}\lambda_k^{(2)} \exp\{it(\mathbf{x}_0^\top \boldsymbol{\beta}_j^{(1)} + \mathbf{x}^\top \boldsymbol{\beta}_k^{(2)})\}. \tag{34}$$

To simplify notation, let $\gamma_1, \ldots, \gamma_{m^2}$ denote the m^2 values of $\mathbf{x}_0^\top \boldsymbol{\beta}_j^{(2)} + \mathbf{x}^\top \boldsymbol{\beta}_k^{(1)}$ as j and k range from 1 to m. Similarly, let $\delta_1, \ldots, \delta_{m^2}$ denote the m^2 values of $\mathbf{x}_0^\top \boldsymbol{\beta}_j^{(1)} + \mathbf{x}^\top \boldsymbol{\beta}_k^{(2)}$. Then Equation (33) says that the $2m^2$ functions

$$\{\exp(it\gamma_1), \ldots, \exp(it\gamma_{m^2}), \exp(it\delta_1), \ldots, \exp(it\delta_{m^2})\}$$

are linearly dependent on the interval $-\epsilon < t < \epsilon$. Letting $a_1(t), \ldots, a_{2m^2}(t)$ denote these functions, this implies that the Wronskian function defined by

$$W(t) = \mathrm{Det} \begin{pmatrix} a_1(t) & \cdots & a_{2m^2}(t) \\ a_1'(t) & \cdots & a_{2m^2}'(t) \\ \vdots & & \vdots \\ a_1^{(2m^2-1)}(t) & \cdots & a_{2m^2}^{(2m^2-1)}(t) \end{pmatrix}$$

must be zero for all $-\epsilon < t < \epsilon$. But $W(0)$ is the determinant of the Vandermonde matrix

$$\begin{pmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ i\gamma_1 & \cdots & i\gamma_{m^2} & i\delta_1 & \cdots & i\delta_{m^2} \\ (i\gamma_1)^2 & \cdots & (i\gamma_{m^2})^2 & (i\delta_1)^2 & \cdots & (i\delta_{m^2})^2 \\ \vdots & & & & & \vdots \\ (i\gamma_1)^{2m^2-1} & \cdots & (i\gamma_{m^2})^{2m^2-1} & (i\delta_1)^{2m^2-1} & \cdots & (i\delta_{m^2})^{2m^2-1} \end{pmatrix},$$

which implies that

$$0 = |W(0)| = \prod_{r=1}^{m^2} |\delta_r - \gamma_r| \prod_{1 \le r < s \le m^2} \prod |\gamma_s - \gamma_r||\delta_s - \delta_r||\delta_r - \gamma_s||\delta_s - \gamma_r|. \tag{35}$$

From the definition of $\gamma_1, \ldots, \gamma_{m^2}$, for any $r < s$, there exist distinct ordered pairs (j_1, k_1) and (j_2, k_2) such that

$$\gamma_s - \gamma_r \;=\; \mathbf{x}_0^\top (\boldsymbol{\beta}_{j_1}^{(2)} - \boldsymbol{\beta}_{j_2}^{(2)}) + \mathbf{x}^\top (\boldsymbol{\beta}_{k_1}^{(1)} - \boldsymbol{\beta}_{k_2}^{(1)}); \tag{36}$$

and, reversing the roles of $\mathbf{x}_0$ and $\mathbf{x}$, the same observation holds for $\delta_s - \delta_r$. Thus far, $\mathbf{x}_0$ and $\mathbf{x}$ have been arbitrary elements of the support of $h(\cdot)$ that satisfy Equation (30). This support is assumed to contain an open set, so we can certainly choose $\mathbf{x}_0$ and $\mathbf{x}$

so that none of the $\gamma_s - \gamma_r$ nor $\delta_s - \delta_r$ equals zero since $\boldsymbol{\beta}_{j_1}^{(2)} - \boldsymbol{\beta}_{j_2}^{(2)}$ and $\boldsymbol{\beta}_{k_1}^{(1)} - \boldsymbol{\beta}_{k_2}^{(1)}$ cannot both be zero.

Similarly, for any r and s we may write

$$\delta_s - \gamma_r \quad = \quad \mathbf{x}_0^\top(\boldsymbol{\beta}_{j_1}^{(1)} - \boldsymbol{\beta}_{j_2}^{(2)}) + \mathbf{x}^\top(\boldsymbol{\beta}_{k_1}^{(2)} - \boldsymbol{\beta}_{k_2}^{(1)}) \tag{37}$$

for some j_1, k_1, j_2, k_2. Reasoning as before, $\mathbf{x}_0$ and $\mathbf{x}$ may be chosen so that not only are none of the $\gamma_s - \gamma_r$ or $\delta_s - \delta_r$ equal to zero, but so that $\delta_s - \gamma_r$ cannot be zero either unless there exist some j_1, k_1, j_2, k_2 with

$$\boldsymbol{\beta}_{j_1}^{(1)} - \boldsymbol{\beta}_{j_2}^{(2)} = \boldsymbol{\beta}_{k_1}^{(2)} - \boldsymbol{\beta}_{k_2}^{(1)} = 0.$$

But in light of Equation (35), this must occur. In other words, we must have $\boldsymbol{\beta}_j^{(1)} = \boldsymbol{\beta}_k^{(2)}$ for at least one pair (j, k).

Without loss of generality, we may rearrange subscripts and assume $\boldsymbol{\beta}_1^{(1)} = \boldsymbol{\beta}_1^{(2)}$. Equation (34) may then be rewritten

$$\lambda_1^{(2)}\lambda_1^{(1)} \exp\{it(\mathbf{x}_0^\top + \mathbf{x}^\top)\boldsymbol{\beta}_1^{(1)}\} + \sum_{(j,k)\neq(1,1)}\sum \lambda_j^{(2)}\lambda_k^{(1)} \exp\{it(\mathbf{x}_0^\top\boldsymbol{\beta}_j^{(2)} + \mathbf{x}^\top\boldsymbol{\beta}_k^{(1)})\} =$$
$$\lambda_1^{(2)}\lambda_1^{(1)} \exp\{it(\mathbf{x}_0^\top + \mathbf{x}^\top)\boldsymbol{\beta}_1^{(1)}\} + \sum_{(j,k)\neq(1,1)}\sum \lambda_j^{(1)}\lambda_k^{(2)} \exp\{it(\mathbf{x}_0^\top\boldsymbol{\beta}_j^{(1)} + \mathbf{x}^\top\boldsymbol{\beta}_k^{(2)})\},$$

which is equivalent to

$$\sum_{(j,k)\neq(1,1)}\sum \lambda_j^{(2)}\lambda_k^{(1)} \exp\{it(\mathbf{x}_0^\top\boldsymbol{\beta}_j^{(2)} + \mathbf{x}^\top\boldsymbol{\beta}_k^{(1)})\} =$$
$$\sum_{(j,k)\neq(1,1)}\sum \lambda_j^{(1)}\lambda_k^{(2)} \exp\{it(\mathbf{x}_0^\top\boldsymbol{\beta}_j^{(1)} + \mathbf{x}^\top\boldsymbol{\beta}_k^{(2)})\}. \tag{38}$$

Now, by an argument exactly like the one following equation (34), we conclude that there must exist distinct ordered pairs (j_1', k_1') and (j_2', k_2') such that

$$\boldsymbol{\beta}_{j_1'}^{(1)} - \boldsymbol{\beta}_{j_2'}^{(2)} = \boldsymbol{\beta}_{k_1'}^{(2)} - \boldsymbol{\beta}_{k_2'}^{(1)} = 0.$$

Again, without loss of generality we may therefore assume that $\boldsymbol{\beta}_2^{(1)} = \boldsymbol{\beta}_2^{(2)}$. (NB: We have used the fact here that by assumption, neither $\boldsymbol{\beta}_j^{(1)}$ nor $\boldsymbol{\beta}_j^{(2)}$ may be equal to $\boldsymbol{\beta}_1^{(1)} = \boldsymbol{\beta}_1^{(2)}$ for $j \neq 1$.) Continuing in this way, we finally conclude that

$$\boldsymbol{\beta}_j^{(1)} = \boldsymbol{\beta}_j^{(2)} \quad \text{for all } j. \tag{39}$$

Because of (39), we may eliminate the superscripts on the $\boldsymbol{\beta}$ and simplify equation (34) as follows:

$$\sum_{j<k}\sum \left(\lambda_j^{(2)}\lambda_k^{(1)} - \lambda_j^{(1)}\lambda_k^{(2)}\right) \exp\{it(\mathbf{x}_0^\top \boldsymbol{\beta}_j + \mathbf{x}^\top \boldsymbol{\beta}_k)\} = \sum_{j<k}\sum \left(\lambda_j^{(2)}\lambda_k^{(1)} - \lambda_j^{(1)}\lambda_k^{(2)}\right) \exp\{it(\mathbf{x}^\top \boldsymbol{\beta}_j + \mathbf{x}_0^\top \boldsymbol{\beta}_k)\}. \tag{40}$$

For equation (40), we may use exactly the same argument that followed equation (34); however, this time we have the additional constraints that $j < k$ and thus $\boldsymbol{\beta}_j$ is never equal to $\boldsymbol{\beta}_k$. Thus, we arrive at a contradiction, meaning that there is no nontrivial set of coefficients in equation (40), which is to say that $\lambda_j^{(2)}\lambda_k^{(1)} - \lambda_j^{(1)}\lambda_k^{(2)} = 0$ for all $j < k$. We conclude that $\boldsymbol{\lambda}^{(1)} = \boldsymbol{\lambda}^{(2)}$.

It remains only to prove that $f^{(1)} = f^{(2)}$ almost everywhere. Returning to equation (31), we observe that we have just proven that $\mathbf{W}_1$ and $\mathbf{W}_2$ have the same distribution, which means that $\phi_{V_1}(t)$ must equal $\phi_{V_2}(t)$ whenever $\phi_{\mathbf{W}_1}(t\mathbf{x}_0) = \phi_{\mathbf{W}_1^\top \mathbf{x}_0}(t)$ is not zero. But $\phi_{\mathbf{W}_1^\top \mathbf{x}_0}(t)$, viewed as a function of a complex variable, is analytic on the entire plane and nonconstant, which means that it cannot take the value zero on any interval of the real line. Thus, $\phi_{V_1}(t) = \phi_{V_2}(t)$ except possibly on a subset of $\mathbb{R}$ not containing any intervals, so because these functions must be continuous we conclude that $\phi_{V_1}(t) = \phi_{V_2}(t)$ for all $t \in \mathbb{R}$. ■