

# SEMIPARAMETRIC MIXTURE OF BINOMIAL REGRESSION WITH A DEGENERATE COMPONENT

J. Cao and W. Yao

*Simon Fraser University and Kansas State University*

*Abstract:* Many historical datasets contain a large number of zeros, and cannot be modeled directly using a single distribution. Motivated by rain data from a global climate model, we study a semiparametric mixture of binomial regression, in which both the component proportions and the success probabilities depend on the predictors nonparametrically. An EM algorithm is proposed to estimate this semiparametric mixture model by maximizing the local likelihood function. We also consider a special case in which the component proportions are constant while the component success probabilities still depend on the predictors nonparametrically. This model is estimated by a one-step backfitting procedure, and the estimates are shown to achieve the optimal convergence rates. The asymptotic properties of the estimates for both models are established. The proposed procedures are demonstrated by modelling rain data from a global climate model and historical rain data from Edmonton, Canada. Simulation studies show that satisfactory estimates are obtained for the proposed models for finite samples.

*Key words and phrases:* climate change, EM algorithm, weather data

## 1 Introduction

It is of great interest to study the evidence and impacts of climate change from weather data over periods of time that range from decades to millions of years (Parmesan and Yohe 2003, IPCC 2007, Tebaldi and Sansó 2009, and Smith et al. 2009). While historical weather data are often limited, massive amounts of data for future weather can be generated from a global climate model.

Global climate models are mathematical models of the general circulation of a planetary atmosphere or ocean. There are about 25 versions of global climate models developed in different research centers across the world. Global climate models are commonly used for simulating the atmosphere or ocean of the earth

using complex computer programs. They are widely used for weather forecasting, understanding the climate and projecting climate changes. The Geophysical Fluid Dynamics Laboratory (GFDL) in the U.S.A. developed one global climate model and implemented some computer simulations (Delworth et al. 2006, Gnanadesikan et al. 2006, Wittenberg et al. 2006 and Stouffer et al. 2006).

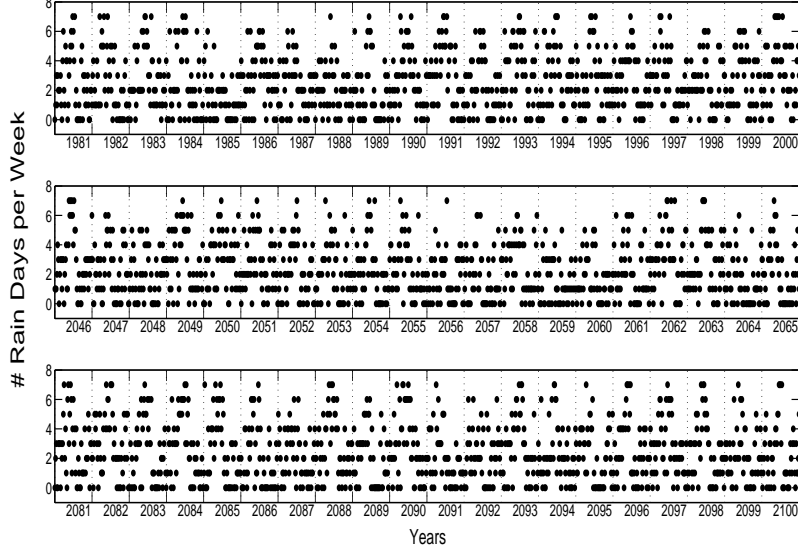


Figure 1: The number of rain days per week in the Edmonton area, Canada in three periods 1981-2000, 2046-2065, 2081-2100 based on the outputs of the GFDL's global climate model.

Figure 1 displays the number of rain days per week in the Edmonton area, Canada in three periods 1981-2000, 2046-2065, 2081-2100 based on the outputs of the GFDL's global climate model. Although such count data might typically be modelled using a binomial distribution, preliminary analysis indicates that this variable does not follow a binomial distribution since too many weeks have no rain days.

Motivated by the above example, we propose a semiparametric mixture of binomial regression model

$$f(X(t) \mid \pi_1(t), p(t)) = \pi_1(t)\text{Bin}(X(t); N, 0) + \pi_2(t)\text{Bin}(X(t); N, p(t)), \quad (1)$$

where  $\text{Bin}(X; N, p)$  is the probability mass function of a binomial distribution for the variable  $X$  based on  $N$  trials with success probability  $p$ . The first component is a degenerate distribution with mass 1 on 0. To make the model (1) identifiable, we assume  $N \geq 2$ ; see Teicher (1961) and Lindsay (1995). Two nonparametric functions,  $\pi_1(t)$  and  $\pi_2(t)$ , are the proportions of zero component and binomial component, respectively, under the constraint  $\pi_1(t) + \pi_2(t) = 1$ . The nonparametric function,  $p(t)$ , is the success probability in the binomial component. The semiparametric mixture model (1) can be used to model many historical data with extra number of zeros. For example, it may be used to model the number of days per week having a forest fire, which are observed with many zeros in winter seasons when no forest fires happen because of snow.

Mixtures of binomial distributions

$$\pi_1 \text{Bin}(x, p_1) + \dots + \pi_m \text{Bin}(x, p_m) \quad (2)$$

were first used by Pearson (1915) to model yeast cell count data. Since then, mixtures of binomial distributions have been used in many fields, such as medicine (Farewell and Sprott, 1988), biology (Brooks et al., 1997) and veterinary science (Böhning et al., 1998). If the component specific probabilities  $p_j$ 's depend on some predictors parametrically, then the model (2) is called "mixture of binomial regression". Mixtures of binomial regression models have wide applications such as in medicine (Zhu and Zhang, 2004), in biology (Follmann and Lambert, 1989), in marketing research (Wedel and DeSarbo, 1933 and De Soete and DeSarbo, 1991), in genetics (Zhang and Merikangas, 2000), in medical research (Lwin and Martin, 1989), in the economics of labor markets (Geweke and Keane, 1999), and in agriculture (Wang and Puterman, 1998). However, the conventional mixtures of binomial regression models require strong parametric assumptions about the  $p_j$ s and cannot account for the dependence of  $\pi_j$ s on the predictors.

The semiparametric mixture model (1) extends mixtures of binomial regression models by removing the parametric assumptions about the component proportions and success probabilities. The two functions  $\pi_1(t)$  and  $p(t)$  in the semiparametric mixture model are estimated using nonparametric smoothing methods such as kernel regression. We propose an EM algorithm (Dempster, Laird, and Rubin, 1977) to maximize the local likelihood function, and prove that the EM algorithm monotonically increases the local likelihood function. The conver-

gence rate of the consistent estimates and their asymptotic normality are also established.

In some cases, one may have some prior knowledge that the component proportions are constant over  $t$ . Therefore, we also consider the following model

$$f(X(t) | \pi_1, p(t)) = \pi_1 \text{Bin}(X(t); N, 0) + \pi_2 \text{Bin}(X(t); N, p(t)). \quad (3)$$

The above semiparametric mixture model (3) requires one to estimate  $p(t)$  locally and  $\pi_1$  globally. Therefore, it is difficult to estimate both  $p(t)$  and  $\pi_1$  efficiently. We propose a one-step backfitting algorithm, in which  $\pi_1$  is first estimated globally given an initial consistent estimate of  $p(t)$  and then  $p(t)$  is updated given the global estimate of  $\pi_1$ . We show that the one-step backfitting estimates for both  $p(t)$  and  $\pi_1$  achieve the optimal convergence rates, and the computation is much more efficient than performing multiple iterations.

The rest of the paper is organized as follows. In Section 2, we introduce the estimation procedure for both proposed semiparametric mixture of binomial regression models. The asymptotic properties of the proposed procedures are established in Section 3. The bandwidth selection is discussed in Section 4. In Section 5, we demonstrate the proposed procedures by modeling the rain data from GFDL's global climate model and the historical rain data from Edmonton, Canada. In Section 6, we use simulations to compare the finite sample performance of the proposed two semiparametric mixture models with a non-mixture model.

## 2 Methods

In this section, we introduce the estimation procedures and algorithms for the two proposed semiparametric mixture models (1) and (3). In (1), both the mixing proportion  $\pi_1(t)$  and the success probability  $p(t)$  depend on the predictor  $t$  nonparametrically. We propose to use a local likelihood criterion to estimate  $\pi_1(t)$  and  $p(t)$ . An EM algorithm is also proposed to maximize the local likelihood. In (3), the component proportion  $\pi_1$  is a constant parameter while the success probability  $p(t)$  depends on the predictor  $t$  nonparametrically. Therefore, we need to estimate  $\pi_1$  using global data but estimate  $p(t)$  using local data.

We propose a one-step backfitting procedure to achieve the optimal convergence rates for both of the estimates of  $\pi_1$  and  $p(t)$ .

## 2.1 Semiparametric mixture model with time-varying proportions

The semiparametric mixture model (1) has two nonparametric functions to estimate: the proportion of zero component  $\pi_1(t)$  and the success probability  $p(t)$ . Kernel regression is applied to estimate these nonparametric functions. One might also use other nonparametric smoothing methods such as local polynomial, spline smoothing, and wavelets.

The two nonparametric functions, at any point  $t_0$ , are estimated by maximizing the following local log-likelihood (Tibshirani and Hastie 1987 and Fan and Gijbels 1996)

$$\begin{aligned} \ell(\boldsymbol{\theta}(t_0)) = & \frac{1}{n} \sum_{i=1}^n K_h(t_i - t_0) \log \left[ \pi_1(t_0) I(x_i = 0) \right. \\ & \left. + \{1 - \pi_1(t_0)\} \binom{N}{x_i} p(t_0)^{x_i} \{1 - p(t_0)\}^{N-x_i} \right], \end{aligned} \quad (4)$$

where  $\boldsymbol{\theta}(t_0) = \{\pi_1(t_0), p(t_0)\}^T$ ,  $x_i$  is the measurement or observation for  $X(t)$  at  $t_i$ ,  $i = 1, \dots, n$ , and  $K_h(\cdot) = h^{-1}K(t/h)$  is a rescaling of the kernel function  $K(\cdot)$  with the bandwidth  $h$ . In this article, we use the Gaussian kernel for  $K(\cdot)$  for the real data analysis in Section 5 and our simulation study in Section 6. The choice of bandwidth  $h$  will be discussed in Section 4.

Note that there is no explicit solution to the maximization of (4). We propose an EM algorithm to maximize (4). Define a vector of component indicator  $\mathbf{z}_i = (z_{i1}, z_{i2})^T$ , where

$$z_{ij} = \begin{cases} 1, & \text{if } (x_i, t_i) \text{ is from the } j\text{-th component;} \\ 0, & \text{otherwise.} \end{cases}$$

Then the complete local log-likelihood function for the complete data  $\{(x_i, \mathbf{z}_i), i = 1, \dots, n\}$ , after omitting some irrelevant constants, is

$$\begin{aligned} l_c(\boldsymbol{\theta}(t_0)) = & \sum_{i=1}^n K_h(t_i - t_0) [I(x_i = 0) z_{i1} \log \pi_1(t_0) \\ & + z_{i2} \{\log(1 - \pi_1(t_0)) + x_i \log p(t_0) + (N - x_i) \log(1 - p(t_0))\}]. \end{aligned}$$

Let  $\mathbf{y} = \{(x_1, t_1), \dots, (x_n, t_n)\}$  and  $\boldsymbol{\theta}^{(k)}(t_0)$  be the value of  $\boldsymbol{\theta}(t_0)$  after the  $k$ -th EM iteration. The E-step at the  $(k+1)$ -th iteration requires one to find  $E_{\boldsymbol{\theta}^{(k)}(t_0)}(l_c(\boldsymbol{\theta}(t_0)) \mid \mathbf{y})$ . Since  $l_c(\boldsymbol{\theta}(t_0))$  is a linear function of  $z_{ij}$ , the E-step is equivalent to finding the classification probabilities  $p_{ij}^{(k+1)} = E_{\boldsymbol{\theta}^{(k)}(t_0)}(Z_{ij} \mid \mathbf{y})$ , where  $Z_{ij}$  is the random variable corresponding to  $z_{ij}$ . The M step at the  $(k+1)$ -th iteration requires one to maximize  $E_{\boldsymbol{\theta}^{(k)}(t_0)}(l_c(\boldsymbol{\theta}(t_0)) \mid \mathbf{y})$  with respect to  $\boldsymbol{\theta}(t_0)$ , which has explicit solutions.

Let  $\pi_1^{(k)}(t_0)$  and  $p^{(k)}(t_0)$  be the value of  $\pi_1(t_0)$  and  $p(t_0)$  at the  $k$ -th iteration. The EM algorithm to maximize (4) at the  $(k+1)$ -th step, for any given  $t_0$ , is as follows:

**E step** : find the classification probabilities given the current estimate

$$\begin{aligned} p_{i1}^{(k+1)} &= \frac{\pi_1^{(k)}(t_0) \text{Bin}(x_i; N, 0)}{\pi_1^{(k)}(t_0) \text{Bin}(x_i; N, 0) + \{1 - \pi_1^{(k)}(t_0)\} \text{Bin}(x_i; N, p^{(k)}(t_0))} \\ p_{i2}^{(k+1)} &= 1 - p_{i1}^{(k+1)}, i = 1, \dots, n. \end{aligned}$$

**M step** : update  $\{\pi_1(t_0), p(t_0)\}$  by

$$\begin{aligned} \pi_j^{(k+1)}(t_0) &= \frac{\sum_{i=1}^n K_h(t_i - t_0) p_{ij}^{(k+1)}}{\sum_{i=1}^n \sum_{j=1}^2 K_h(t_i - t_0) p_{ij}^{(k+1)}}, j = 1, 2. \\ p^{(k+1)}(t_0) &= \frac{\sum_{i=1}^n K_h(t_i - t_0) p_{i2}^{(k+1)} x_i}{N \sum_{i=1}^n K_h(t_i - t_0) p_{i2}^{(k+1)}}. \end{aligned} \quad (5)$$

The above EM algorithm monotonically increases the local log-likelihood (4) after each iteration, which is shown in the following theorem.

**Theorem 2.1.** *Each iteration of the above E and M steps will monotonically increase the local log-likelihood (4), i.e.,*

$$\ell(\boldsymbol{\theta}^{(k+1)}(t_0)) \geq \ell(\boldsymbol{\theta}^{(k)}(t_0)),$$

for all  $k$ , where  $\boldsymbol{\theta}(t_0) = (\pi_1(t_0), p(t_0))$  and  $\ell(\cdot)$  is defined in (4).

The proof of Theorem 2.1 will be given in the supplementary file.

## 2.2 Semiparametric mixture model with constant proportions

For the semiparametric model (3), the success probability  $p(t)$  needs to be estimated locally but the constant proportion  $\pi_1$  can be estimated globally. Therefore, it is not trivial to estimate both  $p(t)$  and  $\pi_1$  efficiently. In this section, we propose a one-step backfitting procedure to estimate the model (3):  $\pi_1$  is first estimated globally given an initial consistent estimate of  $p(t)$ , and we then update the estimate for  $p(t)$  given the root  $n$  consistent estimate of  $\pi_1$ .

We first introduce how to estimate  $\pi_1$  globally given an initial consistent estimate of  $p(t)$ . Let  $\hat{p}(t)$  and  $\hat{\pi}_1$  denote the initial consistent estimate of  $p(t)$  and  $\pi_1$ , respectively, which can be obtained by maximizing the local log-likelihood (4). Since  $\hat{\pi}_1$  is a local estimator, it does not have root  $n$  convergence rate. To improve the efficiency,  $\pi_1$  can be estimated globally by maximizing the log-likelihood (6) using the EM algorithm after replacing  $p(t)$  in (3) by the consistent estimate  $\hat{p}(t)$ :

$$\ell_1(\pi_1) = \frac{1}{n} \sum_{i=1}^n \log \left[ \pi_1 I(x_i = 0) + \pi_2 \binom{N}{x_i} \hat{p}(t_i)^{x_i} \{1 - \hat{p}(t_i)\}^{N-x_i} \right] \quad (6)$$

Denote  $\tilde{\pi}_1$  as the maximizer of (6). We will prove the root  $n$  consistency of  $\tilde{\pi}_1$  in Section 3.2.

The EM algorithm to maximize (6) at the  $(k+1)$ th step is as follows:

**E step** : find the classification probability given the current estimate

$$\begin{aligned} p_{i1}^{(k+1)} &= \frac{\pi_1^{(k)} \text{Bin}(x_i; N, 0)}{\pi_1^{(k)} \text{Bin}(x_i; N, 0) + \pi_2^{(k)} \text{Bin}(x_i; N, \hat{p}(t_i))} \\ p_{i2}^{(k+1)} &= 1 - p_{i1}^{(k+1)}, i = 1, \dots, n. \end{aligned}$$

**M step** : update  $(\pi_1, \pi_2)$  by

$$\pi_j^{(k+1)} = \frac{\sum_{i=1}^n p_{ij}^{(k+1)}}{n}, j = 1, 2.$$

Next, we can further improve the efficiency of the estimator for  $p(t)$  given the estimate  $\tilde{\pi}_1$ . Replacing  $\pi_1$  in (3) by  $\tilde{\pi}_1$ , we propose to estimate  $p(t_0)$ , for any

given  $t_0$ , by maximizing the local likelihood function

$$\begin{aligned} \ell_2(p(t_0)) = & \frac{1}{n} \sum_{i=1}^n K_h(t_i - t_0) \log \left[ \tilde{\pi}_1 I(x_i = 0) \right. \\ & \left. + (1 - \tilde{\pi}_1) \binom{N}{x_i} p(t_0)^{x_i} (1 - p(t_0))^{N-x_i} \right]. \end{aligned} \quad (7)$$

Denote by  $\tilde{p}(t_0)$  the resulting estimate of  $p(t_0)$  by maximizing (7). Since  $\tilde{\pi}_1$  is a root  $n$  consistent estimate of  $\pi_1$ , the  $\tilde{p}(t_0)$  has the same efficiency as if  $\pi_1$  were known. Therefore,  $\tilde{p}(t_0)$  will be more efficient than  $\hat{p}(t_0)$ , which needs to account for the uncertainty of  $\hat{\pi}_1$  since  $\hat{p}(t_0)$  and  $\hat{\pi}_1$  are estimated locally simultaneously in (4). See Theorem 3.3 and 3.4 for more details.

An EM algorithm is proposed to maximize (7). The estimate for  $p(t_0)$  is updated at the  $(k+1)$ th step as follows:

**E step** : find the classification probability given the current estimate

$$\begin{aligned} p_{i1}^{(k+1)} &= \frac{\tilde{\pi}_1 \text{Bin}(x_i; N, 0)}{\tilde{\pi}_1 \text{Bin}(x_i; N, 0) + (1 - \tilde{\pi}_1) \text{Bin}(x_i; N, p^{(k)}(t_0))} \\ p_{i2}^{(k+1)} &= 1 - p_{i1}^{(k+1)}, i = 1, \dots, n. \end{aligned}$$

**M step** : update  $p(t_0)$  by

$$p^{(k+1)}(t_0) = \frac{\sum_{i=1}^n K_h(t_i - t_0) p_{i2}^{(k+1)} x_i}{N \sum_{i=1}^n K_h(t_i - t_0) p_{i2}^{(k+1)}}.$$

The ascending property of the above EM algorithm can be established along the lines of Theorem 2.1, and is omitted here.

One may further employ the backfitting procedures with a full iteration between estimating  $\pi_1$  and  $p(t)$  (see, for example, Buja, et al. 1989, Hastie and Tibshirani 1990, and Opsomer and Ruppert 1999) or profile likelihood approach (Severini and Staniswalis 1994) to improve the efficiency. However, we will prove in Section 3.2 that the one-step backfitting procedure achieves the optimal convergence rate, but the computation is much more efficient than performing the full iterations or profile likelihood approach.

The idea of one-step estimate has been used by many authors to simplify the computation procedure but yet still provide optimal convergence rates for



both global parameters and nonparametric regression functions for semiparametric modeling. See, for example, Carroll et al. (1997) for generalized partially linear single-index model, and Li and Liang (2008) for generalized partially linear model.

### 3 Statistical Theory

In this section, we will investigate the asymptotic properties of the proposed estimates for the two semiparametric mixture models (1) and (3).

#### 3.1 Semiparametric mixture model with time-varying proportions

We first give the convergence rate of the consistent estimates  $\hat{\pi}_1(t)$  and  $\hat{p}(t)$  for the semiparametric mixture of binomial regression model (1) in Theorem 3.1. The asymptotic normality results are given in Theorem 3.2.

**Theorem 3.1.** *Under Conditions A–D in the appendix, there exists a consistent maximizer  $\hat{\boldsymbol{\theta}}(t_0) = (\hat{\pi}_1(t_0), \hat{p}(t_0))^T$  for the local log-likelihood function (4), such that*

$$\left\| \hat{\boldsymbol{\theta}}(t_0) - \boldsymbol{\theta}_0(t_0) \right\| = O_p \left\{ (nh)^{-1/2} + h^2 \right\},$$

where  $\| \cdot \|$  is the Euclidian norm, and  $\boldsymbol{\theta}_0(t_0)$  is the true value of  $\boldsymbol{\theta}(t_0) = (\pi_1(t_0), p(t_0))^T$ .

The proof of Theorem 3.1 is given in the appendix.

Before showing the asymptotic normality of the parameter estimates in the next theorem, some notations have to be defined. Let

$$f(x, \boldsymbol{\theta}) = \pi_1 \text{Bin}(x; N, 0) + \pi_2 \text{Bin}(x; N, p),$$

where  $\boldsymbol{\theta} = (\pi_1, p)$ . Define  $l(x, \boldsymbol{\theta}) = \log f(x, \boldsymbol{\theta})$  and

$$l_1(x, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(x, \boldsymbol{\theta}), \quad l_2(x, \boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} l(x, \boldsymbol{\theta});$$

$$G(t) = E\{l_1(X, \boldsymbol{\theta}_0(t_0)) \mid T = t\}, \quad \mathcal{I}(t) = -E\{l_2(X, \boldsymbol{\theta}_0(t_0)) \mid T = t\}. \quad (8)$$

The moments of  $K$  and  $K^2$  are denoted respectively by

$$\mu_j = \int t^j K(t) dt \quad \text{and} \quad \nu_j = \int t^j K^2(t) dt.$$

**Theorem 3.2.** *Under the conditions A–D in the appendix, the estimate  $\hat{\boldsymbol{\theta}}(t_0) = (\hat{\pi}_1(t_0), \hat{p}(t_0))^T$ , which is a local maximizer of the local log-likelihood function (4) and satisfies the condition in Theorem 3.1, has the following asymptotic distribution*

$$\sqrt{nh} \left\{ \hat{\boldsymbol{\theta}}(t_0) - \boldsymbol{\theta}_0(t_0) - b(t_0)h^2 + o_p(h^2) \right\} \xrightarrow{D} N \left\{ 0, g^{-1}(t_0) \mathcal{I}^{-1}(t_0) \nu_0 \right\},$$

where  $g(t)$  is the marginal density of the time random variable  $T$ , and

$$b(t_0) = \mathcal{I}^{-1}(t_0) \left\{ \frac{G'(t_0)g'(t_0)}{g(t_0)} + \frac{1}{2}G''(t_0) \right\} \mu_2, \quad (9)$$

The proof of Theorem 3.2 is given in the appendix.

### 3.2 Semiparametric mixture model with constant proportions

Let  $\tilde{\pi}_1$  and  $\tilde{p}(t_0)$  denote the estimates for the semiparametric mixture of binomial regression model (3) using the one-step backfitting procedure introduced in Section 2.2. We first establish the root  $n$  consistency of  $\tilde{\pi}_1$  and give its asymptotic distribution in Theorem 3.3. The asymptotic distribution of  $\tilde{p}(t_0)$  is given in Theorem 3.4.

**Theorem 3.3.** *Under the conditions A – D in the appendix, if  $nh^4 \rightarrow 0$ ,  $nh^2 \log(1/h) \rightarrow \infty$ , then the consistent estimate  $\tilde{\pi}_1$  of  $\pi_1$  has the following asymptotic distribution*

$$\sqrt{n}(\tilde{\pi}_1 - \pi_1) \rightarrow N(0, \mathcal{I}_{\pi_1}^{-2} \Sigma),$$

where

$$\begin{aligned} \mathcal{I}_{\pi_1} &= -E \left\{ \frac{\partial^2 f(x, \pi_1, p(t))}{\partial \pi_1^2} \right\}, \\ \Sigma &= \text{var} \left\{ \frac{\partial f(x, \pi_1, p(t))}{\partial \pi_1} - \mathcal{I}_{\pi_1 p}(t) \psi(t, x) \right\}, \\ \mathcal{I}_{\pi_1 p}(t) &= -E \left\{ \frac{\partial^2 f(X, \pi_1, p(t))}{\partial \pi_1 \partial p} \middle| T = t \right\}, \end{aligned}$$

and  $\psi(t, x)$  is the second entry of  $\mathcal{I}(t)^{-1} l_1(x, \boldsymbol{\theta}(t))$ .

The proof of Theorem 3.3 is given in the appendix. Theorem 3.3 indicates that the bandwidth has to satisfy  $h = o(n^{-1/4})$ , while the optimal bandwidth

for kernel smoothing in the literature usually satisfies  $h = O(n^{-1/5})$ . In other words, undersmoothing is required for  $\tilde{p}(t)$  in order to get the root  $n$  convergence rate for  $\tilde{\pi}_1$ . This condition is consistent with what has been found by Carroll et al. (1997) and Li and Liang (2008).

**Theorem 3.4.** *Under the conditions A – D in the appendix, the consistent estimate  $\tilde{p}(t_0)$  has the following asymptotic distribution*

$$\sqrt{nh} \left\{ \tilde{p}(t_0) - p(t_0) - \tilde{b}(t_0)h^2 \right\} \xrightarrow{D} N \left( 0, g(t_0)^{-1} \mathcal{I}_p(t_0)^{-1} \nu_0 \right),$$

where

$$\begin{aligned} \tilde{b}(t_0) &= \frac{1}{2g(t_0)\mathcal{I}_p(t_0)} \left\{ \Gamma''(t_0)g(t_0) + 2\Gamma'(t_0)g'(t_0) \right\} \mu_2, \\ \Gamma(t) &= E \left\{ \frac{\partial f(x, \pi_1, p(t_0))}{\partial p} \middle| T = t \right\}, \\ \mathcal{I}_p(t) &= \text{var} \left\{ \frac{\partial f(x, \pi_1, p(t_0))}{\partial p} \middle| T = t \right\}. \end{aligned}$$

The proof of the Theorem 3.4 is given in the appendix. Note that  $\mathcal{I}_p(t)$  is the (2, 2) element of  $\mathcal{I}(t)$  and  $\Gamma(t)$  is the second entry of  $G(t)$ , where  $\mathcal{I}(t)$  and  $G(t)$  are defined in (8). Denote by  $\mathcal{I}^{22}(t)$  the (2, 2) element of  $\mathcal{I}^{-1}(t)$ . Note that  $1/\mathcal{I}_p(t_0) \leq \mathcal{I}^{22}(t_0)$ . Comparing the results in Theorem 3.2 and 3.4, we see that the one-step backfitting estimator  $\tilde{p}(t_0)$  has smaller asymptotic bias and variance than the estimator  $\hat{p}(t_0)$ .

## 4 Bandwidth Selection

The nonparametric functions in the two semiparametric mixture of binomial regression models are estimated using the kernel regression with some bandwidth  $h$ . The theoretical optimal bandwidth can be obtained by minimizing the asymptotic weighted mean square error. A practical data-driven bandwidth selector is also introduced based on the Nadaraya-Watson estimator. For simplicity of explanation, we will focus on the semiparametric mixture model (1). The methods are easily adaptive to the semiparametric mixture model (3).

Based on Theorem 3.2, one can see that the asymptotic bias of  $\hat{\theta}$  is  $b(t_0)h^2$  and the asymptotic covariance is  $(nh)^{-1}g^{-1}(t_0)\mathcal{I}^{-1}(t_0)\nu_0$ . A theoretic optimal

bandwidth for estimating  $\boldsymbol{\theta}(t_0)$  can be obtained by minimizing the asymptotic weighted mean squared error (MSE)

$$\begin{aligned} & \mathbb{E} \left[ \left\{ \hat{\boldsymbol{\theta}}(t_0) - \boldsymbol{\theta}(t_0) \right\}^T W \left\{ \hat{\boldsymbol{\theta}}(t_0) - \boldsymbol{\theta}(t_0) \right\} \right] \\ &= b(t_0)^T W b(t_0) h^4 + \frac{\nu_0}{g(t_0) n h} \text{tr} \{ \mathcal{I}^{-1}(t_0) W \} + o_p(a_n), \end{aligned}$$

where  $a_n = \{h^4 + (nh)^{-1}\}$ ,  $\text{tr}(A)$  is the trace of  $A$ , and  $W$  is a weight function. Therefore, the theoretic optimal local bandwidth is

$$\hat{h}_{opt}(t_0) = \left[ \frac{\text{tr} \{ \mathcal{I}^{-1}(t_0) W \} \nu_0}{4b(t_0)^T W b(t_0) g(t_0)} \right]^{1/5} n^{-1/5}, \quad (10)$$

where  $b(t_0)$  is given in (9). If our main interest is  $p(t_0)$ , the weight function  $W$  can be  $\text{diag}\{0, 1\}$ . If we are interested in both  $\pi_1(t_0)$  and  $p(t_0)$ , we can take  $W = \mathcal{I}(t_0)$ , which is proportional to the inverse of the asymptotic variance of  $\hat{\boldsymbol{\theta}}(t_0)$ .

Based on the asymptotic bias and variance, we can also derive a theoretic global bandwidth selector by minimizing the asymptotic integrated weighted mean squared error (MSE)

$$\int \mathbb{E} \left[ \left\{ \hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) \right\}^T W \left\{ \hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) \right\} \right] w(t) dt,$$

where  $w(t)$  is any weight function, such as  $g(t)$  or 1. Therefore the theoretic optimal global bandwidth is

$$\hat{h}_{opt} = \left[ \frac{\nu_0 \int \text{tr} \{ \mathcal{I}^{-1}(t) W \} g^{-1}(t) w(t) dt}{4 \int b(t)^T W b(t) w(t) dt} \right]^{1/5} n^{-1/5}. \quad (11)$$

Note that there are some unknown quantities in the formula (10) and (11). Therefore, they are not ready to use in practice. One of the commonly-used methods is the Plug-In idea (see, for example, Ruppert, Sheather, and Wand 1995), i.e., to replace the unknown quantities in the formula (10) or (11) by some estimates. In addition, one can also use a cross-validation criterion to choose the bandwidth, with a little more computation.

Noting that (5) is a conventional Nadaraya-Watson estimator if  $p_{i2}$  is either 0 or 1, we can also employ the existing bandwidth selector for Nadaraya-Watson

estimator; see, for example, Rice (1984); Hurvich, Simonoff, and Tsai (1998), based on some initial partition of the data into different components. The initial partition, i.e., the classification probabilities,  $p_{ij}$ , can be estimated by assuming  $\pi_1(t)$  and  $p(t)$  are constant or polynomial functions of  $t$ . This simple initial fit cannot guarantee a consistent estimate but is easy to implement and generally works well. This idea of initial parametric fit has been used by Fan and Gijbels (1996, Sec 4.2). Note that the nonzero observation  $x_i$  must be from the second component. The only uncertainty is the zero observations of  $x$ . Therefore, the impact of misspecification of  $\pi_1(t)$  and  $p(t)$  is very small. One can also iterate the above procedure several times to get a refined bandwidth. We will use this bandwidth selection method in our simulations and real data applications.

## 5 Applications

The semiparametric mixture model is demonstrated using two applications. The first application is modeling past and future rain data generated from GFDL's global climate model. The second one analyzes historical rain data from Edmonton, Canada.

### 5.1 Rain data from GFDL's Global Climate Model

GFDL's computer simulation based on their global climate model generated rain data at 128 grid points in longitude and 64 grid points in latitude over the whole earth in three time periods: 1981-2000, 2046-2065, and 2081-2100. We choose a grid point close to Edmonton, Canada, and analyzed the rain data at this grid point. Figure 1 displays the number of rain days per week during the three time periods. A rain day is defined as a day with more than 1 millimeter of rainfall. The semiparametric mixture of binomial regression model (1) with  $N = 7$  is used to analyze the rain data from each of these three time periods. We will use the bandwidth introduced at the end of Section 4 and the Gaussian kernel for  $K(\cdot)$  for our model (1). Similar choices are used for the other examples.

It is of great interest to estimate the trend of extreme weather when studying the evidence of climate change. For the rain data, extreme weather includes having zero rain days in one week (too dry) or having seven rain days in one

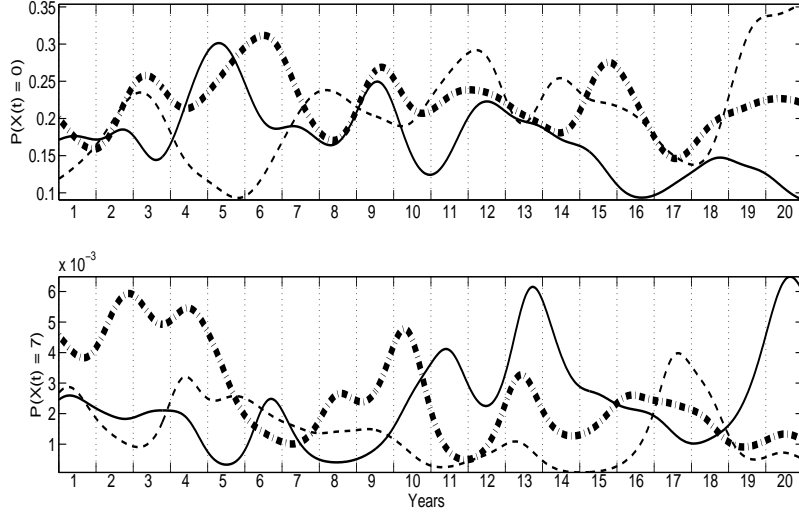


Figure 2: The top and bottom panels show the estimated probabilities of having zero and seven rain days in one week within the three time periods 1981-2000 (solid lines), 2046-2065 (dashed lines) and 2081-2100 (thick dash-dotted lines), respectively.

week (too wet). The probabilities of having zero and seven rain days in one week are calculated as  $P(X(t) = 0) = \hat{\pi}_1(t) + \{1 - \hat{\pi}_1(t)\}\{1 - \hat{p}(t)\}^7$  and  $P(X(t) = 7) = \{1 - \hat{\pi}_1(t)\}\hat{p}(t)^7$ , respectively. Figure 2 displays the estimated  $P(X(t) = 0)$  and  $P(X(t) = 7)$  in the three time periods. The time period 2081-2100 has a high probability of having zero rain days in one week, which is 22.0% on average. The average probability of having zero days in one week in the time period 2081-2100 increases 27.4% and 6.2% from the time periods 1981-2000 and 2046-2065, respectively.

The time period 2081-2100 also has a high probability of having seven rain days in one week, which is 0.26% on average. The time period 2081-2100 has the average probability of having seven rain days in one week increasing 9.4% and 84.0% from the time periods 1981-2000 and 2046-2065, respectively. This indicates that we may have more extreme weather in the time period 2081-2100.

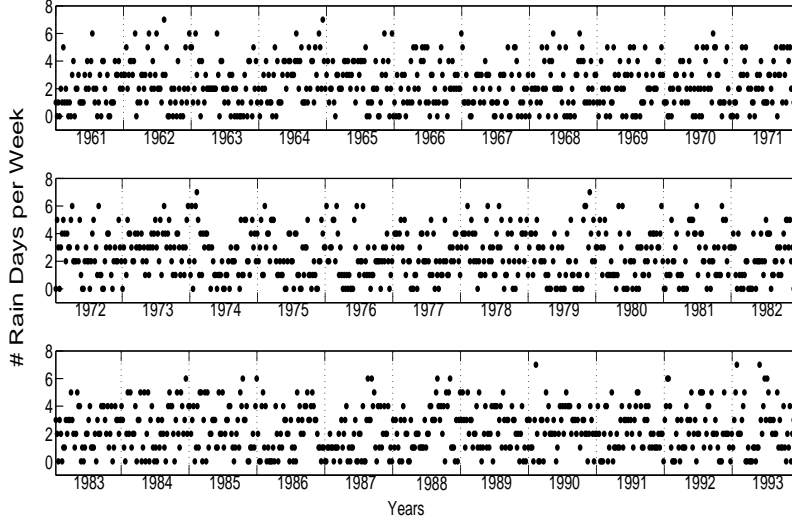


Figure 3: The number of rain days per week in Edmonton, Canada during 1961-1993.

## 5.2 Historical Rain Data in Edmonton

Figure 3 displays the number of rain days per week in Edmonton, Canada during 1961-1993. We fit the rain data with the semiparametric mixture of binomial regression model (1) with  $N = 7$ .

Figure 4 displays the estimates of the probabilities of having zero and seven rain days in one week, which are denoted as  $P(X(t) = 0)$  and  $P(X(t) = 7)$ , respectively. The time period 1986-1987 is extremely dry, having the largest probability of zero rain days in one week (21.1%) and the smallest probability of seven rain days in one week (0.042%). On the other hand, year 1973 has extreme rainfall, with the smallest probability of having zero rain days in one week (7.4%) and the largest probability of having seven rain days in one week (0.200%).

We also fit the same rain data with the non-mixture model:

$$f(X(t) | p(t)) = \text{Bin}(X(t); N, p(t)). \quad (12)$$

The above non-mixture model ignores the degenerate zero component and is equivalent to the semiparametric mixture model (1) when assuming  $\pi_1(t) \equiv 0$ . As a result, it seriously underestimates the probability of having zero and seven

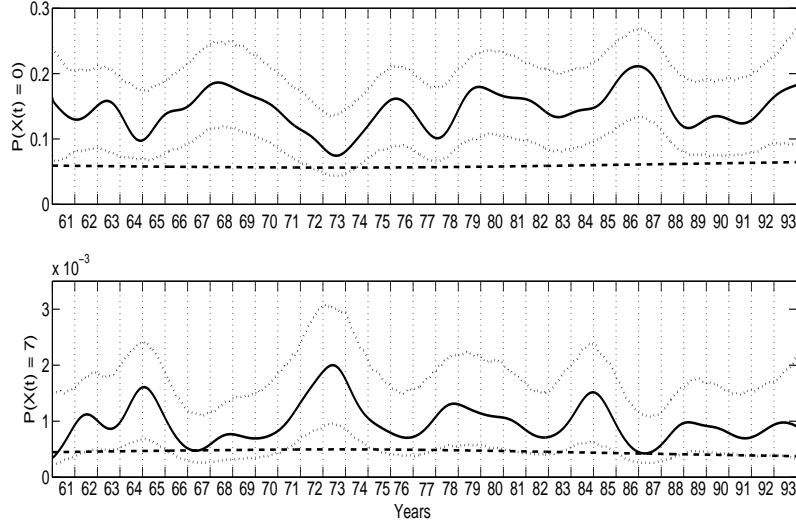


Figure 4: The top and bottom panels display the estimated probabilities of having zero and seven rain days in one week in Edmonton, Canada during 1961-1993, respectively. The solid and dashed lines are the estimates from the semiparametric mixture model (1) and the non-mixture model (12), respectively. The dotted curves are the corresponding 95% confidence intervals for the estimates from the semiparametric mixture model (1).

rain days in one week, which is around 5.9% and 0.046% for the whole time period. In Figure 4, one can also see that the estimates from the non-mixture model are almost flat. This is mainly because that the non-mixture model does not have enough flexibility and cannot capture the variation of the data.

Parametric bootstrap is applied to obtain the 95% confidence intervals for the probabilities of having zero and seven rain days in one week. It is implemented as follows. The simulated data are generated from the semiparametric mixture model (1) with  $N = 7$  where the true  $\pi_1(t)$  and  $p(t)$  are set as the estimates from the real rain data. Then the semiparametric mixture model is estimated from the simulated data in 1000 simulation replicates. The probabilities of having zero and seven rain days in one week are calculated from the 1000 estimates of the semiparametric mixture model. We then obtain the 95% confidence interval by calculating the 2.5% and 97.5% quantiles of the 1000 probabilities of having zero and seven rain days in one week. Figure 4 displays the 95% confidence



intervals for the probabilities of having zero and seven rain days in one week. The probability of having zero rain days in one week estimated from the non-mixture model is under the lower confidence bound, which indicates that the estimate of  $P(X(t) = 0)$  from the mixture model is significantly higher than that from the non-mixture model.

## 6 Simulations

Simulation studies are implemented to evaluate the finite sample performance of the estimation for our proposed two semiparametric mixture of binomial regression models (1) and (3), and compare them with the non-mixture model (12).

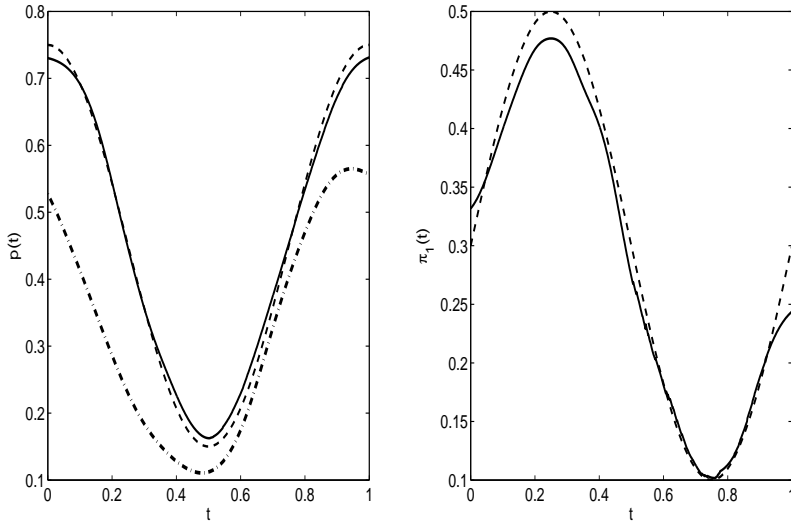


Figure 5: The mean of the estimates for  $p(t)$  and  $\pi_1(t)$  in the semiparametric mixture model (1), which are both plotted in solid lines. The dashed lines are the true functions  $p(t)$  and  $\pi_1(t)$ . The dash-dotted line is the mean of the estimates for  $p(t)$  in the non-mixture model (12).

The simulated data are generated in two scenarios, the first based on the semiparametric mixture model (1) and the second based on the model (3). Both scenarios set the true success probability function  $p(t) = 0.3(1.5 + \cos(2\pi t))$ . In

the first scenario, the true  $\pi_1(t)$  is a time-varying function:  $\pi_1(t) = 0.2(1.5 + \sin(2\pi t))$ . In the second scenario, the true  $\pi_1 = 0.4$ . We evaluate the finite sample performance by varying the sample size as small ( $n = 50$ ), medium ( $n = 100$ ), and large ( $n = 200$ ). The times are generated at  $n$  equally-spaced grid points in  $[0,1]$ . The models (1) and (12) are estimated from the simulated data in the first scenario, and the models (3) and (12) are estimated in the second scenario. The simulation is done with 100 replicates.

Figure 5 displays the mean of the estimates for  $\pi_1(t)$  and  $p(t)$  in the semiparametric mixture model (1) in the first scenario. For comparison, we also add the mean of the estimates for  $p(t)$  in the non-mixture model (12). From Figure 5, one can see that the mean estimates of both  $\pi_1(t)$  and  $p(t)$  in our proposed semiparametric model (1) are very close to the true functions, while the mean estimate of  $p(t)$  in the non-mixture model is smaller than the true  $p(t)$ , which becomes more serious where the true  $\pi_1(t)$  is large.

When the data are simulated based on the first scenario, the estimates for the semiparametric mixture model (1) and the non-mixture model (12) are summarized in Table 1. The average absolute values of biases of  $\hat{p}(t)$  using the semiparametric mixture model are only 12%, 10%, and 7% of those using the non-mixture model when the sample size is 50, 100, and 200, respectively. The estimates using the semiparametric mixture model have slightly smaller average standard deviations for  $p(t)$  than those using the non-mixture model. The semiparametric mixture model also reduces the average RMSE of  $\hat{p}(t)$  by 53%, 60%, and 68% than the non-mixture model when the sample size is 50, 100, and 200, respectively.

When the data are simulated based on the second scenario, the mixture model (3) is estimated using our one-step estimator and the traditional full iterative backfitting algorithm. Table 2 displays the summary of the estimates. Both algorithms have almost the same quality of estimates for  $p(t)$ . The backfitting algorithm has a slightly smaller RMSE for  $\hat{\pi}_1$  than the one-step method, but the one-step method takes less than half the time than the backfitting algorithm. The non-mixture model (12) is also fitted to the same simulated data. The mixture model (3) reduces the average RMSE of  $\hat{p}(t)$  by 62%, 70%, and 77% than the non-mixture model when the sample size is 50, 100, and 200, respectively.

Table 1: The summary of the estimates for the semiparametric mixture model (1) and the non-mixture model (12) when the simulated data are generated based on the model (1). The true success probability function  $p(t) = 0.3(1.5 + \cos(2\pi t))$ , and the true  $\pi_1 = 0.2(1.5 + \sin(2\pi t))$ . The last three columns are the absolute values of bias, standard deviation (SD) and root mean squared error (RMSE) of the estimates for the models (1) and (12), averaged over  $n$  equally spaced points in  $[0,1]$ . The non-mixture model (12) is estimated with the Penalized Iteratively Reweighted Least Squares (P-IRLS) method (see e.g. Wood 2000) using the “mgcv” package in R (R Development Core Team 2010).

| $n$ | Model       |            | BIAS  | SD    | RMSE  |
|-----|-------------|------------|-------|-------|-------|
| 50  | Mixture     | $\pi_1(t)$ | 0.013 | 0.164 | 0.165 |
|     |             | $p(t)$     | 0.017 | 0.080 | 0.082 |
|     | Non-Mixture | $p(t)$     | 0.137 | 0.100 | 0.173 |
| 100 | Mixture     | $\pi_1(t)$ | 0.014 | 0.133 | 0.134 |
|     |             | $p(t)$     | 0.013 | 0.059 | 0.061 |
|     | Non-Mixture | $p(t)$     | 0.132 | 0.069 | 0.152 |
| 200 | Mixture     | $\pi_1(t)$ | 0.013 | 0.099 | 0.100 |
|     |             | $p(t)$     | 0.009 | 0.044 | 0.045 |
|     | Non-Mixture | $p(t)$     | 0.132 | 0.047 | 0.142 |

Table 2: The summary of the estimates for the semiparametric mixture model (3), and the non-mixture model (12) when the simulated data are generated based on the model (1). The true success probability function  $p(t) = 0.3(1.5 + \cos(2\pi t))$ , and the true  $\pi_1 = 0.4$ . “|BIAS|”, “SD”, “RMSE” are the absolute values of bias, standard deviation (SD) and root mean squared error (RMSE) of the estimates for the models (3) and (12), averaged over  $n$  equally spaced points in  $[0,1]$ . The non-mixture model (12) is estimated with the Penalized Iteratively Reweighted Least Squares (P-IRLS) method (see e.g. Wood 2000) using the “mgcv” package in R (R Development Core Team 2010). The last column is the computing time in seconds for 100 Simulations.

| $n$ | Model       |         | Method      | BIAS   | SD    | RMSE  | Time   |
|-----|-------------|---------|-------------|--------|-------|-------|--------|
| 50  | Mixture     | $\pi_1$ | One-Step    | 0.032  | 0.074 | 0.081 | 1.656  |
|     |             |         | Backfitting | 0.011  | 0.075 | 0.076 | 3.890  |
|     |             | $p(t)$  | One-Step    | 0.010  | 0.077 | 0.078 | 1.656  |
|     |             |         | Backfitting | 0.009  | 0.079 | 0.079 | 3.890  |
|     | Non-Mixture | $p(t)$  | P-IRLS      | 0.178  | 0.104 | 0.207 | 7.407  |
| 100 | Mixture     | $\pi_1$ | One-Step    | 0.032  | 0.055 | 0.064 | 3.374  |
|     |             |         | Backfitting | 1.9e-5 | 0.060 | 0.060 | 7.412  |
|     |             | $p(t)$  | One-Step    | 0.011  | 0.057 | 0.059 | 3.374  |
|     |             |         | Backfitting | 0.009  | 0.058 | 0.059 | 7.412  |
|     | Non-Mixture | $p(t)$  | P-IRLS      | 0.182  | 0.073 | 0.197 | 8.625  |
| 200 | Mixture     | $\pi_1$ | One-Step    | 0.026  | 0.037 | 0.045 | 7.710  |
|     |             |         | Backfitting | 0.001  | 0.038 | 0.038 | 15.637 |
|     |             | $p(t)$  | One-Step    | 0.009  | 0.042 | 0.044 | 7.710  |
|     |             |         | Backfitting | 0.008  | 0.043 | 0.044 | 15.637 |
|     | Non-Mixture | $p(t)$  | P-IRLS      | 0.181  | 0.053 | 0.190 | 11.625 |

## Appendix

In this section, we will provide a sketch of the proof of Theorem 2.1 and 3.3. The proof of Theorem 3.1, 3.2, and 3.4 are very standard and are omitted here. Please refer to the supplementary file for more detail.

Let  $g(t)$  be the density function for  $t$ . The following technical conditions are imposed in this section. They are not the weakest possible conditions, but they are imposed to facilitate the proofs.

### Technical Conditions:

A  $\pi_1(t)$  and  $p(t)$  has continuous second derivative at  $t_0$  and  $0 < \pi_1(t_0) < 1$  and  $0 < p(t_0) < 1$ . (For the constant proportion semiparametric mixture model (3), we use the same assumption for  $p(t)$  and assume  $0 < \pi_1 < 1$ .)

B  $g(t)$  has continuous second derivative at the point  $t_0$  and  $g(t_0) > 0$ .

C  $K(\cdot)$  is a symmetric (about 0) kernel density with compact support  $[-1, 1]$ .

D The bandwidth  $h$  tends to zero such that  $nh \rightarrow \infty$ .

*Proof of Theorem 2.1.*

Note that

$$\ell(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n K_h(t_i - t_0) \log f(x_i, \boldsymbol{\theta}).$$

Hence,

$$\begin{aligned} \ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^n \log \left\{ \frac{\pi_1^{(k)} B(x_i, N, 0)}{f(x_i, \boldsymbol{\theta}^{(k)})} \frac{\pi_1^{(k+1)} B(x_i, N, 0)}{\pi_1^{(k)} B(x_i, N, 0)} \right. \\ &\quad \left. + \frac{\pi_2^{(k)} B(x_i, N, p^{(k)})}{f(x_i, \boldsymbol{\theta}^{(k)})} \frac{\pi_2^{(k+1)} B(x_i, N, p^{(k+1)})}{\pi_2^{(k)} B(x_i, N, p^{(k)})} \right\} K_h(x_i - x_0) \\ &= \sum_{i=1}^n \log \left\{ r_{i1}^{(k+1)} \frac{\pi_1^{(k+1)} B(x_i, N, 0)}{\pi_1^{(k)} B(x_i, N, 0)} + r_{i2}^{(k+1)} \frac{\pi_2^{(k+1)} B(x_i, N, p^{(k+1)})}{\pi_2^{(k)} B(x_i, N, p^{(k)})} \right\} K_h(x_i - x_0) \end{aligned}$$

Based on the Jensen's inequality, we have

$$\begin{aligned} \ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)}) &\geq \sum_{i=1}^n \left[ r_{i1}^{(k+1)} \log \left\{ \frac{\pi_1^{(k+1)} B(x_i, N, 0)}{\pi_1^{(k)} B(x_i, N, 0)} \right\} K_h(x_i - x_0) \right. \\ &\quad \left. + r_{i2}^{(k+1)} \log \left\{ \frac{\pi_2^{(k+1)} B(x_i, N, p^{(k+1)})}{\pi_2^{(k)} B(x_i, N, p^{(k)})} \right\} K_h(x_i - x_0) \right] \end{aligned}$$

Based on the property of M-step of (5), we have  $\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)}) \geq 0$ .

*Proof of Theorem 3.3.*

Let

$$f(x_i, \pi_1, \hat{p}(t_i)) = \log \left[ \pi_1 I(x_i = 0) + \pi_2 \binom{N}{x_i} \hat{p}(t_i)^{x_i} (1 - \hat{p}(t_i))^{N-x_i} \right].$$

Based on a Taylor expansion of (6), we have that

$$\sqrt{n}(\tilde{\pi}_1 - \pi_1) = B_n^{-1} A_n + o_p(1).$$

where

$$A_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial f(x_i, \pi_1, \hat{p}(t_i))}{\partial \pi_1} \text{ and } B_n = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 f(x_i, \pi_1, \hat{p}(t_i))}{\partial \pi_1^2}$$

It can be shown that

$$B_n = -E \left\{ \frac{\partial^2 f(x_i, \pi_1, p(t_i))}{\partial \pi_1^2} \right\} + o_p(1) = \mathcal{I}_{\pi_1} + o_p(1).$$

It can be shown that

$$\begin{aligned} A_n &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial f(x_i, \pi_1, p(t_i))}{\partial \pi_1} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial^2 f(x_i, \pi_1, p(t_i))}{\partial \pi_1 \partial p} \{\hat{p}(t_i) - p(t_i)\} + O_p(d_{1n}) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial f(x_i, \pi_1, p(t_i))}{\partial \pi_1} + S_{n1} + O_p(d_{1n}). \end{aligned}$$

where  $d_{1n} = n^{-1/2} \|\tilde{\pi}_1 - \pi_1\|_\infty^2 = o_p(1)$ . Based on the proof of Theorem 3.2 (see supplementary file), we have

$$\hat{\boldsymbol{\theta}}(t_i) - \boldsymbol{\theta}(t_i) = \frac{1}{n} g(t_i)^{-1} \mathcal{I}(t_i)^{-1} \sum_{j=1}^n K_h(t_j - t_i) l_1(x_j, \boldsymbol{\theta}(t_i)) + O_p(d_{n2}),$$

Similar to Li and Liang (2008) and Carroll et al. (1997), we can prove that  $n^{1/2}d_{n2} = o_p(1)$  uniformly in  $t_i$ , if  $nh^2/\log(1/h) \rightarrow \infty$ . Let  $\psi(t_j, x_j)$  be the second entry of  $\mathcal{I}(t_j)^{-1}l_1(x_j, \boldsymbol{\theta}(t_j))$ . Since  $p(t_i) - p(t_j) = O(t_i - t_j)$  and  $K(\cdot)$  is symmetric about 0, we have

$$\begin{aligned} S_{n1} &= \frac{1}{n^{-3/2}} \sum_{j=1}^n \sum_{i=1}^n \frac{\partial^2 f(x_i, \pi_1, p(t_i))}{\partial \pi_1 \partial p} g(t_i)^{-1} \psi(t_j, x_j) K_h(t_j - t_i) + O_p(n^{1/2}h^2) \\ &= S_{n2} + O_p(n^{1/2}h^2). \end{aligned}$$

It can be shown, by calculating the second moment, that  $S_{n2} - S_{n3} = o_p(1)$ , where  $S_{n3} = -n^{-1/2} \sum_{j=1}^n \xi(t_j, x_j)$ , with

$$\xi(t_j, x_j) = -E \left\{ \frac{\partial^2 f(x, \pi_1, p(t_j))}{\partial \pi_1 \partial p} \mid t = t_j \right\} \psi(t_j, x_j) = \mathcal{I}_{\pi_1 p}(t_j) \psi(t_j, x_j).$$

By condition  $nh^4 \rightarrow 0$ , we know

$$A_n = n^{-1/2} \sum_{i=1}^n \left\{ \frac{\partial f(x_i, \pi_1, p(t_i))}{\partial \pi_1} - \xi(t_i, x_i) \right\} + o_p(1).$$

We can show that  $E(A_n) = 0$ . Define

$$\Sigma = \text{var}(A_n) = \text{var} \left\{ \frac{\partial f(x, \pi_1, p(t))}{\partial \pi_1} - \xi(t, x) \right\}.$$

Based on the central limit theorem, we can have  $\sqrt{n}(\tilde{\pi}_1 - \pi_1) \rightarrow N(0, \mathcal{I}_{\pi_1}^{-2}\Sigma)$ .

## Acknowledgment

The authors would like to thank Dr. Peter Hall, the associate editor and two referees for their very constructive suggestions and comments, which helps to greatly improve this paper. They also thank Dr. Claudia Tebaldi for providing the rain data from global climate models. JC's research is partly supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

Brooks, S. P., Morgan, B. J. T., Ridout, M. S., and Pack, S. E. (1997). Finite mixture models for proportions. *Biometrics*, 53, 1097-1115.

- Buja, A., Hastie, T. J., and Tibshirani, R. J. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics*, 17, 453-555.
- Böhning, D. (1998). Zero-inflated Poisson models and C.A.MAN: A tutorial collection of evidence. *Biometrical Journal*, 40, 833-843.
- Carroll, R. J. Fan, J., Gijbels, I., and Wang, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92, 477-489.
- De Soete, G. and DeSarbo, W. S. (1991). A latent class probit model for analyzing pick any/N data. *Journal of Classification*, 8, 45-63.
- Delworth et al. (2006). GFDL's CM2 global coupled climate models - Part 1: Formulation and simulation characteristics, *Journal of Climate*, 19, 643-674.
- Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, B39, 1-38.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.
- Farewell, V. T. and Sprott, D. A. (1988). The use of a mixture model in the analysis of count data. *Biometrics*, 44, 1191-1194.
- Follmann, D. A. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84, 295-300.
- Fong, D. Y. T. and Yip, P. S. F. (1993). An EM algorithm for a mixture model of count data. *Statistics & Probability Letters*, 17, 53-60.
- Fong, D. Y. T. and Yip, P. S. F. (1993). A note on informatino loss in analyzing a mixture model for count data. *Communications in Statistics-Theory and Methods*, 24, 3197-3209.
- Geweke, J. and Keane, M. (1999). Mixture of normals probit models. In C. Hsiao, M. H. Pearson, K. L. Lahiri, and L. F. Lee (Eds). *Analysis of Panels*



- and Limited Dependent Variables*, 49-78. Cambridge: Cambridge University Press.
- Gnanadesikan et al. (2006). GFDL's CM2 global coupled climate models - Part 2: The baseline ocean simulation, *Journal of Climate*, 19, 675-697.
- Hastie, T. and Tibshirani, R. J. (1990). *Generalised Additive Models*. London: Chapman and Hall.
- Hurvich, C. M., Simonoff, J. S., and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society*, B60, 271-293.
- IPCC (2007). Climate change 2007: the physical science basis (summary for policy makers). IPCC.
- Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics*, 34, 1-14.
- Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics*, 36, 261-286.
- Lindsay, B. G., (1995), *Mixture Models: Theory, Geometry, and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics v 5, Hayward, CA: Institute of Mathematical Statistics.
- Lwin, T. and Martin, P. J. (1989). Probits of mixtures. *Biometrics*, 45, 721-732.
- Meng, X. L. (1997). The EM algorithm and medical studies: a historical link. *Statistical Methods in Medical Research*, 6, 3-23.
- Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability Applied*, 10, 186-190.
- Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, 8, 715-732.
- Parmesan, C. and Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, 421, 37-42.

- Pearson, K. (1915). On certain types of compound frequency distributions in which the components can be individually described by binomial series. *Biometrika*, 11, 139-144.
- 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, <http://www.R-project.org>.
- Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics*, 12, 1215-1230.
- Smith, R.L., C. Tebaldi, D. Nychka and L. O. Mearns (2009). Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, 104, 97-116.
- Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semi-parametric models. *Journal of the American Statistical Association*, 89, 501-511.
- Stouffer et al. (2006). GFDL's CM2 global coupled climate models - Part 4: Idealized climate response, *Journal of Climate*, 19, 723-740.
- Tebaldi, C. and Sansó, B. (2009). Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. *Journal of the Royal Statistical Society, Series A*, 1, 83-106.
- Tibshirani, R. J. and Hastie, T. J. (1987). Local Likelihood Estimation. *Journal of the American Statistical Association*, 82, 559-567.
- Tierney, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32, 244-248.
- Wang, P. and Puterman, M. L. (1998). Mixed logistic regression models. *Journal of Agricultural, Biological, and Environmental Statistics*, 3, 175-200.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, Ser. A, 26, 359-372.
- Wedel, M. and DeSarbo, W. S. (1933). A latent class binomial logit methodology for the analysis of paired comparison data. *Decision Sciences*, 24, 1157-1170.

- Wittenberg et al. (2006). GFDL's CM2 global coupled climate models - Part 3: Tropical Pacific climate and ENSO, *Journal of Climate*, 19, 698-722.
- Wood, S. N. (2000). Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties, *Journal of the Royal Statistical Society*, B62(2), 413-428.
- Yin, P. (1988). Inference about the mean of a Poisson distribution in the presence of a nuisance parameter. *Australian Journal of Statistics*, 30, 299-306.
- Yin, P. (1991). Conditional inference for a mixture model for the analysis of count data. *Communications in Statistics-Theory and Methods*, 20, 2045-2057.
- Zhu, H.-T. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society*, B66, 3-16.
- Zhang, H. P. and Merikangas, K. (1964). A frailty model of segregation analysis: Understanding the familial transmission of alcoholism. *Biometrics*, 56, 815-823.

J. Cao

Department of Statistics & Actuarial Science, Simon Fraser University  
8888 University Drive, Burnaby, BC, V5A1S6, Canada  
E-mail: jca76@sfu.ca

W. Yao

Department of Statistics, Kansas State University  
101 Dickens Hall, Manhattan, KS, 66506-0802, USA  
E-mail: wxyao@ksu.edu