Electronic Journal of Statistics
Vol. 6 (2012) 710–737
ISSN: 1935-7524
DOI: 10.1214/12-EJS690

# Efficient distribution estimation for data with unobserved sub-population identifiers

**Yanyuan Ma**

*Department of Statistics*
*Texas A&M University*
*College Station, TX 77845*
*e-mail:* ma@stat.tamu.edu

**and**

**Yuanjia Wang**

*Department of Biostatistics*
*Mailman School of Public Health*
*Columbia University*
*722 West 168th Street*
*New York, NY 10032*
*e-mail:* yuanjia.wang@columbia.edu

**Abstract:** We study efficient nonparametric estimation of distribution functions of several scientifically meaningful sub-populations from data consisting of mixed samples where the sub-population identifiers are missing. Only probabilities of each observation belonging to a sub-population are available. The problem arises from several biomedical studies such as quantitative trait locus (QTL) analysis and genetic studies with ungenotyped relatives where the scientific interest lies in estimating the cumulative distribution function of a trait given a specific genotype. However, in these studies subjects' genotypes may not be directly observed. The distribution of the trait outcome is therefore a mixture of several genotype-specific distributions. We characterize the complete class of consistent estimators which includes members such as one type of nonparametric maximum likelihood estimator (NPMLE) and least squares or weighted least squares estimators. We identify the efficient estimator in the class that reaches the semiparametric efficiency bound, and we implement it using a simple procedure that remains consistent even if several components of the estimator are mis-specified. In addition, our close inspections on two commonly used NPMLEs in these problems show the surprising results that the NPMLE in one form is highly inefficient, while in the other form is inconsistent. We provide simulation procedures to illustrate the theoretical results and demonstrate the proposed methods through two real data examples.

**AMS 2000 subject classifications:** Primary 62G05, 62G20; secondary 62G99.

**Keywords and phrases:** Finite mixed samples, robustness, semiparametric efficiency, nonparametric maximum likelihood estimator (NPMLE).

Received November 2011.

## Contents

## 1. Introduction

In many scientific studies, data arise from a mixture of scientifically meaningful distributions. For example, in a quantitative trait locus (QTL) study, the goal is to identify, map and estimate effect of a QTL predisposing the trait. However, the genomic location of the QTL is unknown, therefore subjects' genotypes at the QTL are not observed. Mixture models are widely used to map QTLs using location-known molecular markers such as single nucleotide polymorphisms (SNPs) or microsatellite markers, see Lander and Botstein (1989) and Wu et al. (2007).

Another example where mixture model is useful is genetic studies where genotypes in relatives of an initial sample (probands) are not collected (Marder et al., 2003; Wang et al., 2008). In these studies, of scientific interest is to estimate the conditional distribution of a trait given a genotype (or penetrance, Khoury et al., 1993). Genotype information in the initial sample of probands are collected. However, it is common that due to high cost of administering in-person interviews in relatives, their genotype information is not collected. For example, in Wacholder et al. (1998) and Wang et al. (2007, 2008), only the probands are genotyped, but none of the first-degree relatives of the probands was genotyped. Distribution of possible genotypes of a relative, however, can easily be obtained given the relationship between the relative and the proband and the genotype in the proband. The relatives' disease history or trait information is usually obtained by administering a systematic and reliable phone-interview (Marder et al., 2003). Distribution of the trait in a relative is then a mixture of conditional distribution of the trait given the relative's genotype and these relatives form the main analysis sample.

A concrete example of such genetic studies is an investigation of association between the APOE gene and the LDL concentrations in young children (Shea et al., 1999). There are three common alleles at the APOE locus ($\varepsilon 2, \varepsilon 3, \varepsilon 4$). The APOE $\varepsilon 3$ is the most prevalent allele in the general population, with frequency 75% to 80%. Previous studies have suggested that the APOE $\varepsilon 4$ allele may be associated with higher LDL cholesterol levels in adults (Davignon et al., 1988). Of interest is the association between APOE $\varepsilon 4$ allele and LDL cholesterol distribution in children.

Subjects included in the study were recruited from a cross-sectional biomarker study of children conducted from 1994 to 1998 (Shea et al., 1999). Proband children were recruited from lists of cardiac patients generated through the Presbyterian Hospital Clinical Information System, private cardiology practices, lipid clinics and pediatric practices. Families with at least one healthy child 4 to 25 years of age were eligible for participation. Siblings of proband children were recruited to the study. The availability of the APOE genotype information of the probands and the sibling relationship enables the calculation of each sibling's probability of carrying the $\varepsilon 4$ allele. The cumulative distribution function of LDL concentration for carriers of $\varepsilon 4$ allele (carrying one or two copies of $\varepsilon 4$) and for the non-carriers (carrying zero copy of $\varepsilon 4$) are of primary interest in this study.

Traditional statistical analysis of mixture data specifies a parametric form of conditional distribution of an outcome given group membership (e.g., Gaussian mixture model, Wu et al., 2007) and estimates mixture probabilities and parameters in the conditional distribution by maximum likelihood through an EM algorithm (McLachlan and Peel, 2000). In this work, we provide nonparametric estimation in the sense that we do not make any distributional assumption on the conditional distributions. One common feature of the two examples introduced before is that the mixture probabilities are easily calculated without using the outcome data or are known, and the mixture populations are scientifically meaningful (e.g., subjects carrying a certain genotype). Treating these mixture probabilities as random variables, each observation in the data consists a vector of mixture probabilities and a continuous outcome, and the observations are assumed to be independent and identically distributed (i.i.d.).

To fix idea, let $Q$ denote a $p$-dimensional vector of random mixture probabilities, and let $p_Q$ denote the probability mass function of $Q$, which has a finite support $u_1, \ldots, u_m$. Let $S$ denote a random outcome, let $L$ denote the unobserved group membership (or genotype), and let $f(s)$ denote the $p$-dimensional conditional density of $S$ given $L$. For simplicity, we assume that $f(s)$ is supported on a compact interval, say $[T_1, T_2]$. For the $i$th subject, $i = 1, \dots, n$, we observe $(q_i, s_i)$, where the joint density of $Q, S$ at $Q = q_i$ and $S = s_i$ is

$$g(q_i, s_i) = p_Q(q_i) q_i^T f(s_i). \tag{1}$$

Here $f(s)$ is a length $p$ vector, where the $j$th component $f_j(s)$ represents the conditional probability density function (PDF) of $s$ given that it belongs to the $j$th genotype group, $j = 1, \dots, p$. Each component of $f(s)$, $f_j(s)$, is the PDF of a trait at time $t$ given the gene mutation status being the $j$th kind in a relative (for example, $j = 1$ denotes carriers and $j = 2$ denotes non-carriers), or the PDF of a quantitative trait given the QTL genotype being the $j$th kind. Let $F(\cdot)$ denote the corresponding $p$ dimensional cumulative distribution function (CDF) of $f(\cdot)$. Our interest is in estimating $F$ at any fixed time $t$. The vector $q_i$ represents probabilities that a relative carries a certain genotype given the proband's genotype, or a vector of probabilities of a subject having a certain QTL genotype given the flanking markers. Obviously $\sum_{j=1}^{p} q_{ij} = 1$. The distribution of $q_i$ (i.e., $p_Q$) depends on study design and can be easily estimated consistently from the empirical distribution of $q_i$. For example, for a backcross QTL experiment, $q_i$ takes four different values depending on the marker genotype frequencies (e.g., Table 10.3 of Wu et al., 2007). The vector of density functions $f$ is completely unspecified, thus $f$ is an infinite-dimensional nuisance parameter with length $p$.

Here, we characterize the complete class of consistent estimators which includes Fine et al. (2004) and Chatterjee and Wacholder (2001). We show that any weighted least squares estimator is a member of this estimation class hence yields a consistent estimator. In addition, we construct a special subclass which obtains the minimum estimation variance and reaches the semiparametric efficiency bound. We inspect two types of widely used NPMLEs and report a surprising finding that they are either inefficient or even inconsistent. Although commonly applied in clinical studies (Sigurdson et al., 2004; Hauptmann et al., 2003; Webb et al., 2006a,b; Hartge et al., 2002), the inconsistency of the second type of NPMLE has not been discovered in the literature before.

The remaining of the paper is organized as follows. In Section 2, weighted least squares estimators are introduced and a complete class of consistent estimators encompassing the least squares is defined. The optimal member of the class is identified and shown to reach the semiparametric efficiency bound. In Section 3, an algorithm to implement the efficient estimator is developed and asymptotic properties of the estimator are proved. In Section 4, two types of commonly used NPMLE estimators are investigated and one type is found to be inefficient while the other is inconsistent. In Section 5, simulation experiments are conducted to investigate the finite sample performance of the developed methods, and several estimators including the efficient estimator, the least squares estimators and the NPMLEs are compared. In Section 6, the proposed methods are implemented to analyze two data examples, one from a genetic linkage study of rice plant height and the other from a study of association between plasma low-density lipoprotein (LDL) cholesterol level and the apolipoprotein-E (APOE) gene. In Section 7, possible extensions of the proposed methods are discussed.

## 2. Estimation procedures

### 2.1. A class of weighted least squares estimators

Although the traditional approach to estimating $F(t)$ is maximum likelihood estimator for a parametric model or NPMLE for a nonparametric model, a very simple weighted estimator can be used if we formulate the same problem from a different angle. Observe that the model in (1) implies $q^T F(t) = E\{I(S \leq t)|q\}$, where $I(\cdot)$ denotes an indicator function. Therefore, viewing the $q_i$'s as covariates and $I(S_i \leq t)$ as response variables, the covariates and the responses are linked by $F(t)$ via a familiar linear regression model

$$Y_i \equiv I(S_i \leq t) = q_i^T F(t) + e_i,$$

where $E(e_i|q_i) = 0$, $i = 1, \ldots, n$. It is straightforward that the $e_i$'s are independent conditional on $q_i$'s, and have the variances $v_i = q_i^T F(t)\{1 - q_i^T F(t)\}$. Thus, weighted least squares based method can be used to estimate $F(t)$. Denote by $M$ an arbitrary $n \times n$ diagonal matrix. Let $A = (q_1, \ldots q_n)^T \in R^{n \times p}$, $Y = (y_1, \ldots y_n)^T \in R^n$, and $e = (e_1, \ldots, e_n)^T \in R^n$. Then we obtain the general WLS estimator

$$\widehat{F}(t) = (A^T M A)^{-1} A^T M Y.$$

The simplest estimator is the OLS where we set $M = I_n$, also derived in Fine et al. (2004) using a different formulation, while the most efficient WLS estimator is obtained when we assign $M$ to be a diagonal matrix with the $i$th diagonal entry equals $v_i^{-1}$. Standard iteratively re-weighted estimation procedure can be used to obtain this optimal WLS (OWLS) estimator. The presence of the matrix $M$ also allows the flexibility to derive other WLS estimators to achieve desired properties such as robustness.

### 2.2. The complete class of consistent estimators

Although simple to derive and easy to implement, it is unclear whether the class of WLS is complete and whether OWLS is the optimal estimator among all consistent estimators of $F(t)$. To answer these questions and to provide easy variance estimation for any consistent estimator, we perform a formal semiparametric analysis to characterize the complete class of consistent estimators. We derive in Appendix A.1 that the family of all influence functions is

$$\begin{aligned} S_{IF} \quad = \quad & \Big\{ \phi(q,s) : \phi(q,s) = b(q,s) - F(t) - C1_p, \\ & \quad \text{where} \int b(q,s) q^T p_Q(q) d\mu(q) = I(s \leq t) I_p + C \Big\}, \end{aligned} \tag{2}$$

where $I_p$ is a p-dimensional identity matrix, $C$ is an arbitrary $p \times p$ constant matrix, and $1_p$ is a $p$-dimensional vector with all elements being one.

For any qualified $b$-function as described in $S_{IF}$, an estimator for $F(t)$ is

$$\widehat{F}(t) = n^{-1}\sum_{i=1}^{n} b(q_i, s_i) - C_b 1_p, \tag{3}$$

where we use $C_b = \int b(q,s)q^T p_Q(q)d\mu(q) - I(s \le t)I_p$ to denote the constant matrix corresponding to this $b$-function. For example, a convenient choice of $b(q,s)$ is

$$\begin{aligned} b(q,s) \quad &= \quad I(s \le t)\left\{\int h_1(q,s)q^T p_Q(q)d\mu(q)\right\}^{-1} h_1(q,s) \\ &\quad +B\left\{\int h_2(q,s)q^T p_Q(q)d\mu(q)\right\}^{-1} h_2(q,s) + h_3(q), \end{aligned} \tag{4}$$

where $h_1(q,s), h_2(q,s)$, and $h_3(q)$ can be arbitrary functions in $R^p$ such that $\int h_1(q,s)q^T p_Q(q)d\mu(q)$ and $\int h_2(q,s)q^T p_Q(q)d\mu(q)$ are invertible, and $B$ is an arbitrary constant matrix. This characterization provides a simple construction of a very rich class of estimators.

Since $S_{IF}$ contains all the influence functions, any regular asymptotic linear (RAL, Newey, 1990) estimator can be written in the form of (3). For example, we show in Appendix A.2 that the influence function of any WLS estimator is

$$\phi_{WLS} = \{E(WQQ^T)\}^{-1}wq\left\{I(s \le t) - q^T F(t)\right\}.$$

Here, $w$ is a weight variable. For the $i$th individual, $w = w_i$ is the $i$th diagonal entry of $M$. We use $W$ to denote the weight variable when it is considered as a random variable. It is easy to see that this corresponds to choosing $h_1 = wq$, $h_2 = 0$, and $h_3 = -\{E(WQQ^T)\}^{-1}wqq^T F(t) + F(t)$, hence any WLS is indeed a member of $S_{IF}$. In addition, comparing the form of $\phi_{WLS}$ and $S_{IF}$ indicates that the WLS estimators are only a subset of consistent estimators that can be constructed. To further study whether the optimal WLS estimator is the most efficient among all the consistent estimators for $F(t)$, we need to derive the efficient influence function.

### *2.3. The semiparametric efficient estimator*

Projecting an arbitrary influence function $\phi$ onto the tangent space $\Lambda_{\mathcal{T}}$ yields an efficient influence function (Newey, 1990). In Appendix A.3, we derive the form of $\Lambda_{\mathcal{T}}$ and its orthogonal complement, which enables us to derive the following theorem.

**Theorem 1.** *The efficient influence function is*

$$\phi_{eff} = \frac{\{I(s \le t)I_p - K\}A^{-1}(s)q}{q^T f(s)},$$

*where*

$$A(s) = \int \frac{qq^T p_Q(q)}{q^T f(s)} d\mu(q),$$

*and*

$$K = \int_{T_1}^{T_2} I(s \le t) A^{-1}(s) ds \left\{ \int_{T_1}^{T_2} A^{-1}(s) ds \right\}^{-1}.$$

The proof of the Theorem 1 is in Appendix A.4.

It is straightforward to see that the construction of the efficient estimator requires correct specification of the nuisance parameter $f(s)$, which is not always easy to obtain. If we unknowingly mis-specify $f(s)$ as $f^*(s)$ and follow the same construction in Theorem 1 to obtain $\phi^*_{eff}$, then the result is no longer a valid influence function. To see this, note that $\check{\phi} = \frac{\{I(s \le t) - K^*\} A^{*-1}(s) q}{q^T f^*(s)}$, where $A^*(s) = \int \frac{qq^T p_Q(q)}{q^T f^*(s)} d\mu(q)$, and $K^* = \int I(s \le t) A^{*-1}(s) ds \left\{ \int A^{*-1}(s) ds \right\}^{-1}$. We can then easily verify that $E(\check{\phi}) = F(t) - K^* 1_p$, which is not necessarily zero. We thus robustify the influence function by constructing

$$\phi = \frac{\{I(s \le t) - K^*\} A^{*-1}(s) q}{q^T f^*(s)} - F(t) + K^* 1_p. \tag{5}$$

Regardless of the form of $f^*$, (5) always yields a valid influence function. In addition, $\phi = \phi_{eff}$ when $f^*(s) = f_0(s)$ and $\phi$ can be used to estimate $F(t)$ via

$$\widehat{F}(t) = n^{-1} \sum_{i=1}^{n} \frac{\{I(s_i \le t) - K^*\} A^{*-1}(s_i) q_i}{q_i^T f^*(s_i)} + K^* 1_p. \tag{6}$$

**Remark 1.** In (6), we can replace $K^*$ by an arbitrary constant matrix. The resulting estimator remains consistent, and the corresponding $\phi$ is still a valid influence function. However, since different $K^*$ corresponds to different influence function, the estimators have different variances.

In practice, since $f(s)$ is usually either proposed or estimated so that it may be different from $f_0(t)$, it is always a safer choice to use (6) to obtain $\widehat{F}(t)$. We will show in Section 3 that as long as $f(s)$ is consistently estimated, the estimator (6) is guaranteed to provide an efficient estimator for $F(t)$.

### *2.4. Analytic comparison between OWLS and the efficient estimator*

We are now ready to assess whether the OWLS is efficient. Comparing $\phi_{eff}$ with $\phi_{OWLS}$ obtained in Appendix A.2, we find that although the OWLS is optimal among the WLS family, it does not reach the semiparametric efficiency bound. We prove this claim by contradiction. Suppose that the OWLS is efficient, then we would have $\phi_{eff} = \phi_{OWLS} + o_p(1)$, which would imply that for all $(q, s)$ pairs,

$$\frac{\{I(s \leq t) - K\}A^{-1}(s)q}{q^T f(s)} = \left[E\frac{QQ^T}{Q^T F(t)\{1 - Q^T F(t)\}}\right]^{-1} \frac{q\{I(s \leq t) - q^T F(t)\}}{q^T F(t)\{1 - q^T F(t)\}}.$$

Denote $B = E\left(QQ^T/[Q^T F(t)\{1 - Q^T F(t)\}]\right)$, we then have

$$\frac{A^{-1}(s)q}{q^T f(s)} = \frac{B^{-1}q}{q^T F(t)\{1 - q^T F(t)\}} \quad \text{and} \quad \frac{KA^{-1}(s)q}{q^T f(s)} = \frac{B^{-1}q}{1 - q^T F(t)},$$

which leads to $q^T F(t)A^{-1}(s)q = KA^{-1}(s)q$. The left hand-side is a quadratic function of $q$, while the right hand-side is linear, so the above equality will never hold since $q$ cannot be a constant vector of zero.

## 3. Efficient estimator and its asymptotic properties

As we have pointed out, the efficient influence function derived in Theorem 1 involves unknown nuisance parameters $f(s)$ and therefore cannot be directly used to construct an efficient estimator for $F(t)$. Using (6) will provide a robust and locally efficient estimator, in the sense that if $f^*(s) = f_0(s)$, the estimator is indeed efficient, otherwise, the estimator is still guaranteed to be consistent. We now propose a method to construct an estimator that is always efficient. This method avoids estimating the $p$-dimensional PDF $f(s)$ directly, and is simple to implement.

### *3.1. Algorithm for implementing the efficient estimator*

We propose to use the following procedure to construct the efficient estimator.

1. Randomly split the data into two sets. The second set has size $n_2 = n^{5/6}$, and the first set has size $n_1 = n - n_2$. Assume that the first set contains $(q_1, s_1), \ldots, (q_{n_1}, s_{n_1})$ and the second set $(q_{n_1+1}, s_{n_1+1}), \ldots, (q_n, s_n)$.
2. Obtain the empirical estimator of $q^T f(s)$, $\widehat{q^T f(s)}$ from the second set of sample with size $n_2$. Recall that the random vector $Q$ can take $m$ different vector values $u_1, \ldots, u_m$, so for each $k = 1, \ldots, m$, we can calculate a kernel estimate for $u_k^T f(s)$ as
$$\widehat{u_k^T f(s)} = \frac{\sum_{i=n_1+1}^{n} I(q_i = u_k)K_h(s_i - s)}{\sum_{i=n_1+1}^{n} I(q_i = u_k)}.$$
Here $K_h$ is any kernel function with bandwidth $h$ satisfying $(n_2 h)^{-1} = o(1), n_2 h^5 \leq O(1)$ as $n_2 \to \infty$, and $K_h(\cdot) = h^{-1}K(\cdot/h)$.
3. Calculate
$$A(s; \widehat{q^T f}) = \int \frac{qq^T p_Q(q)}{\widehat{q^T f(s)}} d\mu(q) = E_Q\left\{\frac{QQ^T}{\widehat{Q^T f(s)}}\right\} = \sum_{k=1}^{m} \frac{u_k u_k^T p_Q(u_k)}{\widehat{u_k^T f(s)}},$$

where $E_Q$ stands for expectation with respect to $Q$. We construct

$$K_1(\widehat{q^Tf}) = \int_{T_1}^{T_2} I(s \le t) A^{-1}(s;\widehat{q^Tf})ds, \qquad K_2(\widehat{q^Tf}) = \int_{T_1}^{T_2} A^{-1}(s;\widehat{q^Tf})ds$$

using numerical integration, and form $K(\widehat{q^Tf}) = K_1(\widehat{q^Tf})K_2^{-1}(\widehat{q^Tf})$.

4. Form

$$\psi(Q,S;\widehat{q^Tf}) = \frac{\{I(S \le t) - K(\widehat{q^Tf})\}A^{-1}(S;\widehat{q^Tf})Q}{\widehat{Q^Tf(S)}} + K(\widehat{q^Tf})1_p,$$

and let the estimator be

$$\widehat{F}(t) = n_1^{-1}\sum_{i=1}^{n_1}\psi(q_i,s_i;\widehat{q^Tf}). \tag{7}$$

The estimation procedure described above is straightforward to implement. Comparing to many other semiparametric problems where the efficient estimator often involves solving integral equations (Rabinowitz, 2000) and iterative procedures (Tsiatis and Ma, 2004), the estimator here is very simple. In addition, unlike most semiparametric problems where the nonparametric functions have to be estimated at a certain rate, sometimes using an under-smoothed bandwidth (Liang and Wang, 2005; Li and Liang, 2008) to reach optimality, we do not have such estimation constraints. In fact, we will show that any consistent estimation of $f(s)$ will be as good as the true $f(s)$ asymptotically. Since consistency can be obtained with a wide range of bandwidth, typically one does not have to go through the computationally intensive cross validation procedure to choose an optimal bandwidth. Finally, we point out that the splitting of the data is solely to facilitate the later theoretical proof and is not mandatory. In reality, one can certainly use the whole data set to estimate $f(s)$ and to form $\widehat{F}(t)$ in (7).

### 3.2. Asymptotics and inferences

We present the asymptotic property of the proposed efficient estimator in the following theorem:

**Theorem 2.** *The estimator constructed in (7) achieves the semiparametric efficiency bound. Specifically, for $n \to \infty$, $\sqrt{n}\{\widehat{F}(t) - F(t)\} \to N(0,V)$ in distribution, where $V = var(\phi_{eff})$ and can be consistently estimated as*

$$n^{-1}\sum_{i=1}^{n}\{\psi(q_i,s_i;\widehat{q^Tf}) - \widehat{F}(t)\}\{\psi(q_i,s_i;\widehat{q^Tf}) - \widehat{F}(t)\}^T.$$

Intuitively, the reason that (7) can reach the semiparametric efficiency is because it solves the estimating equation formed by summing over the robustified influence functions (5) while replacing the unspecified quantities $K^*$, $q^T f^*(s)$ and $A^*$ by their corresponding optimal choices which are, respectively, the nonparametric estimates of $K$, $q^T f(s)$ and $A(s, q^T f)$. The rigorous proof of Theorem 2 is in Appendix A.5.

Since we are able to construct the optimal estimators and estimate their variances, it is straightforward to make inferences based on these results. For example, we can construct a locally most powerful test for the hypothesis $H_0 : F_1(t) - F_2(t) = \delta_0$ versus $H_1 : F_1(t) - F_2(t) \neq \delta_0$. Because of the explicit form of $\widehat{F}(t)$, the Wald test is an obvious choice. Let $\widehat{D} = \widehat{F}_1(t) - \widehat{F}_2(t) - \delta_0$, then the test statistic is

$$T = n\widehat{D}^2/v, \tag{8}$$

where $v = V_{11} - V_{12} - V_{21} + V_{22}$, and $V_{ij}$ is the $(i, j)$th element of the covariance matrix $V$ stated in Theorem 2. It is straightforward that when $n \to \infty$, $T$ has a chi-square distribution with one degree of freedom under $H_0$. Under the local alternative, say $F_1(t) - F_2(t) = \delta/\sqrt{n}$, $T$ has a noncentral chi-square distribution with one degree of freedom and noncentrality parameter $(\delta - \delta_0)^2/v$.

In some applications, one may be interested in testing whether $F_1(t) - F_2(t) = \delta_t$ at several different $t$ values simultaneously, say at $t_1, \dots, t_J$. Letting $a^T = (1, -1)\{F(t_1), \dots F(t_J)\} - \Delta_0^T$, where $\Delta_0 = (\delta_{t_1}, \dots, \delta_{t_J})^T$. This can be written as a problem of testing $H_0 : a = 0$ versus $H_1 : a \neq 0$, Under $H_0$, $a$ has a multivariate normal random distribution with mean zero and variance-covariance matrix $n^{-1}\Sigma$, where $\Sigma_{jk} = (-1, 1)\text{cov}\{\widehat{F}(t_j), \widehat{F}(t_k)\}(-1, 1)^T$ for $j, k = 1, \dots, J$. Here, $\text{cov}\{\widehat{F}(t_j), \widehat{F}(t_k)\}$ can be estimated using

$$n^{-1}\sum_{i=1}^{n}\{\psi_{eff}(q_i, s_i; t_j, \widehat{q^T f}) - \widehat{F}(t_j)\}\{\psi_{eff}(q_i, s_i; t_k, \widehat{q^T f}) - \widehat{F}(t_k)\}^T,$$

where $\psi_{eff}(q_i, s_i, ; t_j, \widehat{q^T f})$ and $\widehat{F}(t_j)$ denote $\psi_{eff}$ and $\widehat{F}$ evaluated at the $i$th observation and calculated at time $t_j$. Thus, we can construct the test statistic

$$T = na^T\Sigma^{-1}a. \tag{9}$$

When $n \to \infty$, under $H_0$, $T$ has a chi-square distribution with $J$ degrees of freedom. Under a local alternative, say $a = \Delta/\sqrt{n}$ for some length $J$ vector $\Delta$, $T$ has a noncentral chi-square distribution with noncentrality parameter $\Delta^T\Sigma^{-1}\Delta$.

## 4. Understanding the NPMLEs

For many nonparametric models, the NPMLE is a widely used estimation procedure. In the literature, two types of NPMLE have been proposed (Wacholder et al., 1998; Chatterjee and Wacholder, 2001). The first type of NPMLE treats each $u_j^T f(s), j = 1, \dots, m$ as an unknown PDF, while the second type treats $f(s)$ as a $p$-dimensional unknown PDF. To explain these two NPMLEs in detail, group the observations in such a way that the first $r_1$ observations form a first subset where each observation has the same $q$ value that equals to $u_1$, the next $r_2$ observations form a second subset with the same $q$ values $u_2$ and so on. Assume that the last $r_m$ observations form the $m$th subset and have the $q$ values equal to $u_m$. We use $\widetilde{F}(t)$ to denote the type I NPMLE of $F(t)$, and $\check{F}(t)$ the type II NPMLE.

The type I NPMLE maximizes

$$\sum_{i=1}^{n} \log\{q_i^T f(s_i)\} = \sum_{j=1}^{m}\sum_{i=1}^{n} \log\{q_i^T f(s_i)\} I(q_i = u_j)$$

with respect to $q_i^T f(s_i)$ for the $i$th subject in the $j$th subset subject to $q_i^T f(s_i) \geq 0$ and $\sum_{i=1}^n q_i^T f(s_i) I(q_i = u_j) = 1$ for $j = 1, \ldots, m$. This is essentially equivalent to performing an empirical density estimation in each of the $m$ groups, where in each group the $q_i$ values are identical. Obviously, the resulting estimation for $q^T f(s)$ in the $j$th group is an empirical PDF with weights $r_j^{-1}$ at the observed values. The procedure then uses $u_j^T F(t) = r_j^{-1} \sum_{i=1}^n I(s_i \leq t, q_i = u_j)$ for $j = 1, \ldots, m$ to recover $\widetilde{F}(t) = (U^T U)^{-1} U^T G(t)$, where we denote $U = (u_1, \ldots, u_m)^T$, and $G(t)$ is a length $m$ vector with the $j$th component equals $r_j^{-1} \sum_{i=1}^n I(s_i \leq t, q_i = u_j)$. It is not difficult to see that

$$\begin{aligned} U^T U &= \sum_{j=1}^{m} u_j u_j^T = \sum_{i=1}^{n} w_i q_i q_i^T, \\ \text{and } U^T G(t) &= \sum_{j=1}^{m} u_j r_j^{-1} \sum_{i=1}^{n} I(s_i \leq t, q_i = u_j) = \sum_{i=1}^{n} w_i q_i I(s_i \leq t), \end{aligned}$$

where $w_i = r_j^{-1}$ if $q_i = u_j$. Thus, the type I NPMLE belongs to the family of WLS estimators (therefore a member of class (2)), where the weights are taken to be $r_j^{-1}$, the inverse of the number of observations in the $j$th group with the same $q_i$ value. However, the weights of this WLS estimator are obviously non-optimal. In addition, intuitively such choice of weights is not reasonable, because it down-weights the contributions from a larger subset. In fact, one would rather downweight the contribution from the observations with less estimation precision, while the quality of the estimation of $F(t)$ from each observation has no definitive link with its subset size.

The type II NPMLE maximizes the same log likelihood, but with respect to $f(s_i)$, subject to $\sum_{i=1}^n f(s_i) = 1_p$ and $f(s_i) \geq 0$ component-wise. It is easy to see that the maximum is obtained when the $r_j$ values of $f(s_i)$ corresponding to the same $u_j$ are the same. We denote this common $f(s_i)$ value by $h_j$, for $j = 1, \ldots, m$. We thus maximize

$$\sum_{j=1}^{m} r_j \log(u_j^T h_j)$$

with respect to $h_j$'s subject to $\sum_{j=1}^m r_j h_j = 1_p$ and $h_j \geq 0$ component-wise. In general, no closed form solution exists for the $h_j$'s, and the EM algorithm is often used to solve this optimization problem and to obtain the $h_j$'s. The NPMLE then proceeds to form

$$\check{F}(t) = \sum_{i=1}^n I(s_i \leq t)\widehat{f}(s_i) = \sum_{j=1}^m \sum_{i=1}^n I(s_i \leq t, q_i = u_j) h_j.$$

The type II NPMLE is different from the type I NPMLE in that here, the term "nonparametric" refers to $f(s)$, not to $u_j^T f(s)$. In the literature, the type II estimator is considered as an improvement of the type I NPMLE. However, our careful investigation reveals that the type II NPMLE is not even consistent, which is a rather counter intuitive result. In Appendix A.6, we give a detailed calculation in a concrete case to explicitly illustrate the inconsistency and in Section 5 we demonstrate the bias of the type II NPMLE in a moderately large sample through simulations.

We now give a more general demonstration to show why the type II NPMLE is inconsistent. Suppose the solution to the constrained maximization problem is $h_1, \ldots, h_m$, then the type II NPMLE is

$$\check{F}(t) = \sum_{j=1}^m \left\{ \sum_{i=1}^n I(q_i = u_j) I(s_i \leq t) \right\} h_j = \sum_{j=1}^m r_j G_j(t) h_j = HG(t) = HU\widetilde{F}(t),$$

where $H = (r_1 h_1, \ldots r_m h_m)$, and $U, G(t), \widetilde{F}(t)$ are the same as defined before. We already know that $\widetilde{F}(t)$ is a consistent estimator of $F(t)$. If $\check{F}(t)$ is also consistent, then we would have $HU \rightarrow I_p$ when $n \rightarrow \infty$. This is a much stronger condition than the original constraints of the maximization problem and is in general not satisfied. In fact, this condition means that the type II NPMLE is asymptotically equivalent to the type I NPMLE, which contradicts the original goal of developing a type II estimator. In other words, as a distinct estimator from the type I NPMLE, the type II NPMLE is inconsistent.

## 5. Simulations

To study the finite sample performance of the proposed estimators, we conducted several simulation studies. In all the simulations, the dimension of $F(t)$ is $p = 2$, and the number of simulation iterations is 1000.

### *5.1. Three simulated examples*

In the first simulation experiment, we investigate the performance of the various estimators studied in Sections 2, 3 and 4. Here, $q_i$'s can take six different values, i.e. $m = 6$, while the group sizes $r_j$, $j = 1, \ldots, m$, are randomly generated. The six different $q_i$ values are respectively $(0.3, 0.7)^T$, $(0, 1)^T$, $(0.7, 0.3)^T$, $(0.8, 0.2)^T$,

Table 1

*Bias, empirical standard error (emp se), average estimated standard error (est se), 95% coverage (95% cov) of Simulation 1, sample size $n = 300$, 1000 simulations*

| | $F_1(t) = 0.9295$ | | | | $F_2(t) = 0.3199$ | | | |
|---|---|---|---|---|---|---|---|---|
| Estimator | bias | emp se$^\dagger$ | est se$^*$ | 95% cov | bias | emp se$^\dagger$ | est se$^*$ | 95% cov |
| ORACLE | 0.0003 | 0.529 | 0.538 | 94.9% | −0.0013 | 0.704 | 0.684 | 93.8% |
| EFF | 0.0007 | 0.540 | 0.549 | 94.3% | −0.0017 | 0.726 | 0.704 | 92.5% |
| ROB1 | 0.0001 | 0.545 | 0.555 | 94.5% | −0.0013 | 0.732 | 0.710 | 92.6% |
| ROB2 | 0.0001 | 0.545 | 0.555 | 94.5% | −0.0013 | 0.732 | 0.710 | 92.6% |
| OWLS | −0.0004 | 0.537 | 0.553 | 94.6% | −0.0006 | 0.727 | 0.712 | 93.6% |
| OLS | 0.0001 | 0.545 | 0.559 | 94.6% | −0.0013 | 0.732 | 0.716 | 93.0% |
| NPMLE1 | 0.0001 | 0.570 | 0.581 | 95.0% | −0.0010 | 0.753 | 0.738 | 93.4% |
| NPMLE2 | −0.179 | 0.323 | – | – | 0.2148 | 0.425 | – | – |

$^\dagger$: Empirical standard error × 10

$^*$: Estimated standard error × 10

$(0.5, 0.5)^T$, $(0.6, 0.4)^T$. The two components in the true $F(t)$ both have truncated exponential form, since exponential function is a commonly used parametric model in practice. Specifically, $F_1(t) = \{1 - \exp(-t/3)\}/\{1 - \exp(-10/3)\}$ and $F_2(t) = 1 - \{1 - \exp(t/3 - 10/3)\}/\{1 - \exp(-10/3)\}$ on the interval $(0, 10)$.

We studied eight different estimators. The efficient estimator with true $f(s)$ inserted (hence unrealistic) is denoted ORACLE, while with the estimated $f(s)$ inserted is denoted EFF. Thus EFF is the implemented efficient estimator. Two different kinds of robust estimators are considered, where ROB1 had the $f(t)$ mis-specified, and ROB2 not only used a mis-specified $f(t)$, but also had $K = 0$ plugged in. Specifically, in ROB1, we used the true $f_1(t)$ as the proposed model for $f_2(t)$, and used the true $f_2(t)$ as the proposed model for $f_1(t)$. In ROB2, we proposed uniform model for both $f_1(t)$ and $f_2(t)$. These two estimators are expected to be consistent hence reflecting robustness to mis-specification of the PDFs. We also investigated the proposed OWLS estimator. For comparison, we implemented the OLS, NPMLE1 and NPMLE2 estimators that are used in the literature. We implement the estimation procedures at $t = 6.8$. The resulting estimation mean, sample and estimated standard errors and 95% coverage of the confidence intervals are summarized in Table 1.

It can be seen that all the consistent estimators perform well in finite samples, and the estimated variances are very close to the empirical variances. This indicates that the asymptotic results are relevant for a moderate sample size of $n = 300$. It is very clear that the type II NPMLE yields very large bias. We emphasize here that this bias is not a reflection of small sample size because the bias persists when we increase the sample size to 1000.

We can also see that the type I NPMLE and OLS does not make a very good choice of the weights, hence the estimation standard errors are both larger than the OWLS. This is especially prominent for the type I NPMLE, in that it performs even worse than the simple OLS estimator. The two robust estimator (ROB1 and ROB2) perform very similarly, and both have minimal bias, reflecting the desired robustness property with respect to the PDF estimation. Finally, although in theory the efficient estimator (EFF) should outperform the OWLS

Table 2

*Type I error and power of test in Simulation 1, sample size $n = 300$, 1000 simulations*

| | Type I error | | | | Power | | | |
|---|---|---|---|---|---|---|---|---|
| Estimator | 0.01 | 0.05 | 0.1 | 0.2 | 0.01 | 0.05 | 0.1 | 0.2 |
| ORACLE | 0.016 | 0.062 | 0.105 | 0.194 | 0.198 | 0.424 | 0.546 | 0.700 |
| EFF | 0.017 | 0.061 | 0.118 | 0.198 | 0.177 | 0.400 | 0.523 | 0.676 |
| ROB1 | 0.018 | 0.062 | 0.117 | 0.200 | 0.167 | 0.391 | 0.529 | 0.680 |
| ROB2 | 0.018 | 0.062 | 0.117 | 0.200 | 0.167 | 0.391 | 0.529 | 0.680 |
| OWLS | 0.018 | 0.057 | 0.119 | 0.197 | 0.170 | 0.396 | 0.529 | 0.681 |
| OLS | 0.018 | 0.061 | 0.113 | 0.198 | 0.162 | 0.388 | 0.521 | 0.673 |
| NPMLE1 | 0.023 | 0.062 | 0.100 | 0.204 | 0.148 | 0.354 | 0.496 | 0.655 |

estimator, the performance of OWLS is as satisfactory as EFF. This appears to be often the case in our other simulations not shown here. Thus, using either proposed OWLS or EFF in practice is expected to be adequate.

We also studied the type I error and power of the test (8) in this situation, and present the results in Table 2. The overall performance of the proposed tests is satisfactory. From the left panel of Table 2, we see that all estimators maintain correct size. From the right panel of the same table, we see that the OLS and NPMLE1 have lower power compared to other estimators due to their larger estimation variances.

The second simulation experiment is conducted to closely mimic a QTL mapping data analyzed in Section 6.1. We generated the data from a mixture of two distributions. The first one is a uniform distribution on $(3, 10)$, while the second one has CDF $c(1 - e^{-t/2.5})$ on the interval $(0, 10)$. The mixture probability has four different values which are $(0.02, 0.98)^T, (0.2, 0.8)^T, (0.1, 0.9)^T, (0.98, 0.02)^T$, and the sample size is 100. Based on the performance of the various estimators studied in the first simulation, here we used only the two best estimators, the OWLS and the efficient estimator (EFF) to estimate the two CDFs. We also implemented the type II NPMLE for comparison. We plot the true CDFs, the mean of the estimated CDFs and the 95% pointwise confidence band for each method in Figure 1. As expected, both OWLS and EFF give satisfactory results, while NPMLE2 is clearly biased. Again, we emphasize that the bias of NPMLE2 is not caused by the moderate sample size. In fact, when we increased the sample sizes to 1000, the bias became even more prominent.

Similarly, the third simulation is conducted to closely mimic the LDL data analyzed in Section 6.2. The first CDF is $c_1/\{1 + e^{-(t-3)/0.5}\}$ on the interval $(0, 6)$, and the second CDF is $c_2/\{1 + e^{-(t-2.5)/0.2}\}$ on the interval $(0, 7)$. Note that these two CDFs cross. Here, the mixture probability distribution has three different values which are $(0.15, 0.85)^T, (0.6, 0.4)^T, (0.8, 0.2)^T$, and the sample size is 300. Estimations based on OWLS, EFF and NPMLE2 are computed, and the mean of the estimated CDFs, the 95% pointwise confidence band for each method are presented in Figure 2 together with the true CDFs. Similar to the second simulation, both OWLS and EFF perform well, while NPMLE2 shows large bias.

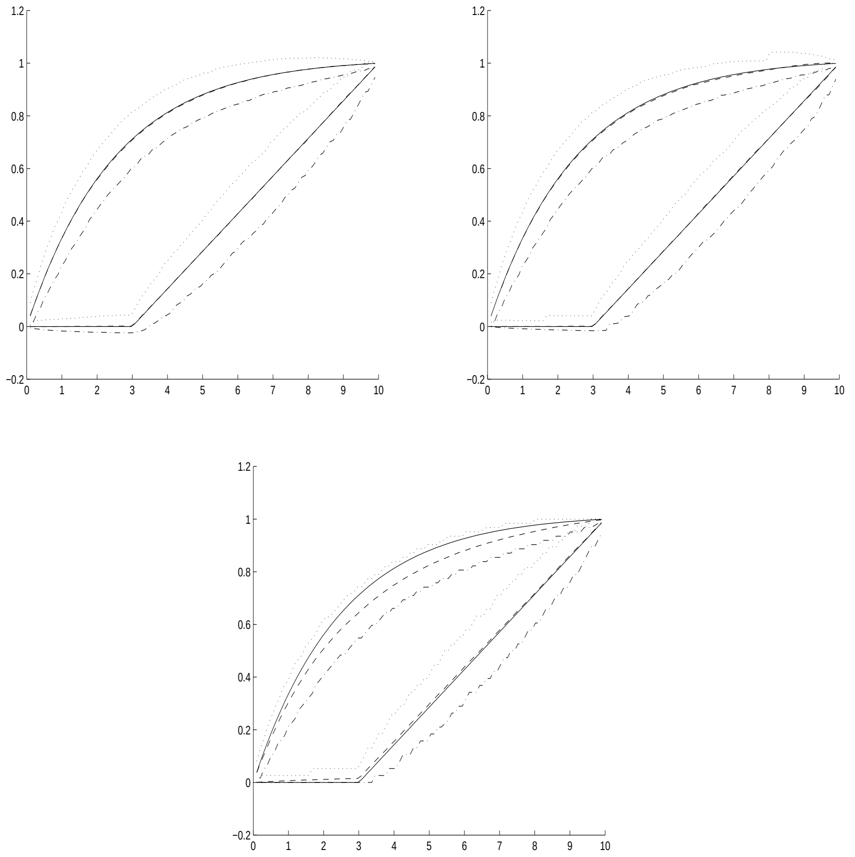

Fig 1. *Simulation 2. True CDF (solid) and the mean (dashed), 95% pointwise confidence band (upper band dotted, lower band dash-dotted) of the estimated CDFs. The OWLS (left), EFF (mid) and NPMLE2 (right) are plotted. The mean and true CDFs are undistinguishable in OWLS and EFF estimators. Sample size is 100, and results are based on 1000 simulations.*

## 6. Real data examples

### *6.1. Estimation from QTL mapping data*

In QTL studies, the trait observations are assumed to be drawn from a mixture of several QTL genotype groups and the mixture probabilities of a subject assuming a certain QTL genotype given flanking markers are calculated based on the study design, the marker genotypes and the recombination fraction between the location-known flanking markers and the putative QTL (Wu et al., 2007). The first example that we use to illustrate our methods is a genetic linkage study used to map QTLs for rice plant height and grain shape. The identified QTL can be used to produce taller rice plants to increase yield. In Huang et al. (1997),

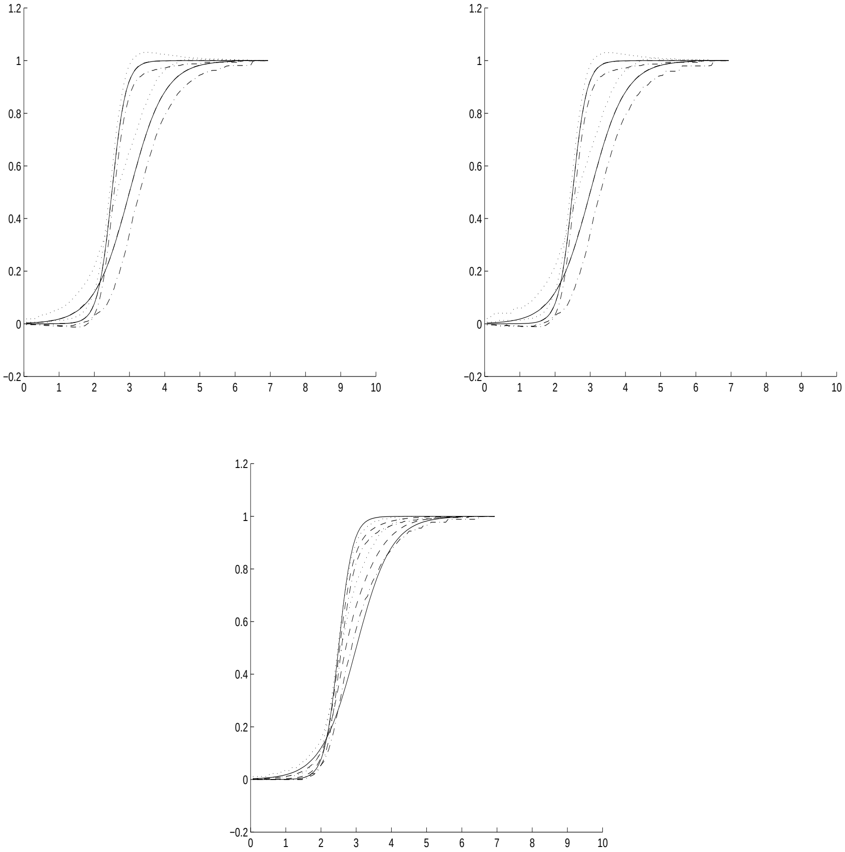

FIG 2. *Simulation 3. True CDF (solid) and the mean (dashed), 95% pointwise confidence band (upper band dotted, lower band dash-dotted) of the estimated CDFs. The OWLS (left), EFF (mid) and NPMLE2 (right) are plotted. The mean and true CDFs are undistinguishable in OWLS and EFF estimators. Sample size is 300, and results are based on 1000 simulations.*

a doubled haploid (DH) population of rice plants was derived from two inbred lines (semi-dwarf IR64 and tall Azucena), creating 123 DH lines each genotyped with 135 RFLP markers and 40 isozyme and RAPD markers. Several traits such as grain shape and plant height were recorded. A DH population is equivalent to a backcross population where the two marker genotypes have an approximately 1:1 distribution ratio. The mixture probabilities $q_i$ of a plant carrying a certain QTL genotype given the flanking markers are computed based on the marker genotypes and the recombination fraction between the marker and the QTL. The details of $q_i$ computation can be found in Table 10.3 of Wu et al. (2007).

Using a Gaussian mixture model, Wu et al. (2007) analyzed the plant height measured at 10 weeks after the rice was transplanted to the field and mapped a QTL for this trait to 199cM on chromosome 1 between the markers RZ730 and RZ801. Here we estimate the cumulative distribution function of the rice

Table 3
*Data example 1. Estimated CDFs of plant height and their standard errors for QTL genotypes bb ($\widehat{F}_1$) and Bb ($\widehat{F}_2$)*

| $t$ | Estimator | $\widehat{F}_1(t)$ | SE($\widehat{F}_1$) | $\widehat{F}_2(t)$ | SE($\widehat{F}_2$) | $p$ value* |
|---|---|---|---|---|---|---|
| 80 | EFF | 0.132 | 0.048 | 0 | 0.006 | 0.011 |
| 80 | OWLS | 0.126 | 0.048 | 0 | 0.001 | 0.011 |
| 110 | EFF | 0.895 | 0.05 | 0.095 | 0.062 | <0.001 |
| 110 | OWLS | 0.927 | 0.043 | 0.098 | 0.062 | <0.001 |
| 140 | EFF | 0.992 | 0.024 | 0.699 | 0.083 | 0.001 |
| 140 | OWLS | 1.000 | 0.006 | 0.684 | 0.082 | 0.000 |

*: $p$ value for testing $H_0 : F_1(t) = F_2(t)$ based on (8)

plant height for each of the two QTL genotypes at the same locus (199cM on chromosome 1) using the model (1).

There were 84 plant height measurements available. Table 3 presents the estimated CDFs and their standard errors for each of the two QTL genotypes at several values of the plant height. We present the efficient estimator (EFF) and the optimal WLS (OWLS). We omitted OLS and the two NPMLEs due to their respective deficiencies. The proposed OWLS and EFF lead to comparable results. The test of $H_0 : F_1(t) = F_2(t)$ based on the test statistic (8) was significant at 5% level for both estimators at three typical values of $t$, indicating a difference in the distribution functions for the two QTL genotypes. In addition, we tested the difference between the two distributions at the three $t$ values simultaneously by the test (9). The null distribution of the test statistic was a chi-square with three degrees of freedom, and the $p$-value was less than 0.01 which indicates a significant difference.

Figure 3 presents the CDFs of rice plant heights for plants carrying each of the two QTL genotypes estimated by the efficient estimator (EFF). It can be seen that there is a large difference in the CDFs across the entire range of the plant height and carrying a risk allele increases the plant height. For example, it was estimated that 90.5% (CI: 78.3%, 100%) of the plants with *Bb* QTL genotype will have plant heights greater than 110, compared to 10.5% (CI: 0.7%, 20.3%) in the *bb* genotype group. This difference is highly significant ($p < 0.001$). These results are consistent with the analysis conducted in Wu et al. (2007).

### 6.2. Estimation from the LDL data

In the LDL example introduced in Section 1, the association between the APOE $\varepsilon 4$ allele and the LDL concentrations in young children is our main research interest. There were 230 subjects included in the data analyses. We show the estimated cumulative distribution function of LDL concentration for carriers of $\varepsilon 4$ allele (carrying one or two copies of $\varepsilon 4$) compared to non-carriers (carrying zero copy of $\varepsilon 4$) at several values of the LDL levels in Table 4. As in data example 1, we present the EFF and the OWLS. Both estimators yielded similar results. The comparison of CDF for carriers versus non-carriers was not significant at 5% level at LDL= 100 or LDL= 260, but was significant at LDL= 180. Similar to the QTL analysis, we tested the difference between two distributions at these

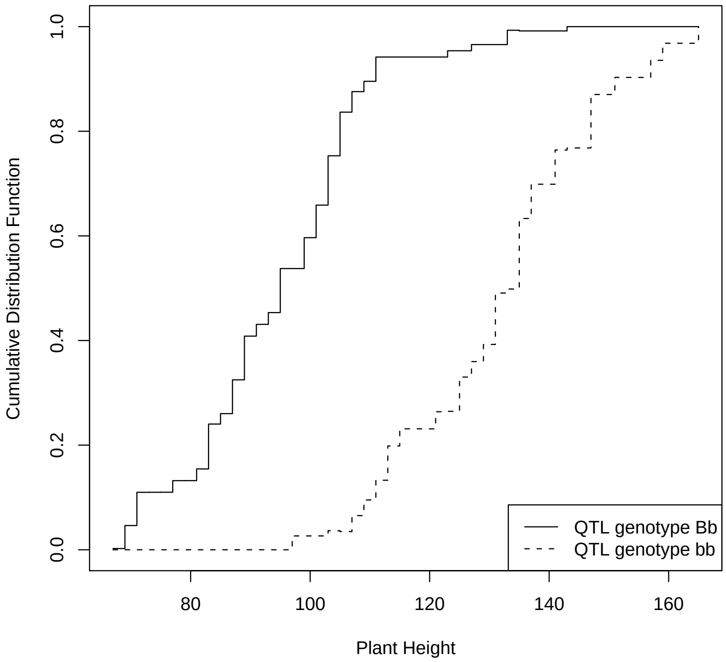

FIG 3. *Data example 1. Estimated cumulative distribution function (CDF) of plant height for QTL genotype Bb (solid) and bb (dashed)*

TABLE 4

*Data example 2. Estimated CDFs of LDL levels and their standard errors of APOE $\varepsilon 4$ carriers ($\widehat{F}_1$) and non-carriers ($\widehat{F}_2$)*

| $t$ | Estimator | $\widehat{F}_1(t)$ | SE($\widehat{F}_1$) | $\widehat{F}_2(t)$ | SE($\widehat{F}_2$) | $p$ value* |
|---|---|---|---|---|---|---|
| 100 | EFF | 0.719 | 0.108 | 0.619 | 0.054 | 0.496 |
| 100 | OWLS | 0.718 | 0.110 | 0.619 | 0.054 | 0.510 |
| 180 | EFF | 1.000 | 0.014 | 0.921 | 0.024 | 0.037 |
| 180 | OWLS | 1.000 | 0.014 | 0.922 | 0.024 | 0.035 |
| 260 | EFF | 1.000 | 0.006 | 0.984 | 0.011 | 0.364 |
| 260 | OWLS | 1.000 | 0.006 | 0.984 | 0.011 | 0.354 |

*: $p$ value for testing $H_0 : F_1(t) = F_2(t)$ based on (8)

three typical $t$ values simultaneously by (9). The $p$-value was 0.29, indicating a non-significant overall difference of the two distributions at these values.

Figure 4 depicts the CDF of LDL for carriers and non-carriers estimated by the efficient estimator, EFF. It can be seen that there is virtually no difference of the two CDFs in the range from 45 to 130. The CDF for carriers is elevated in the interval (130, 200) compared to non-carriers and the two functions merge again for LDL greater than 200. Previous analyses in the literature focus on the mean LDL concentration. Our analysis shows that the effect of APOE $\varepsilon 4$ on LDL manifests in the range of 130 to 200.

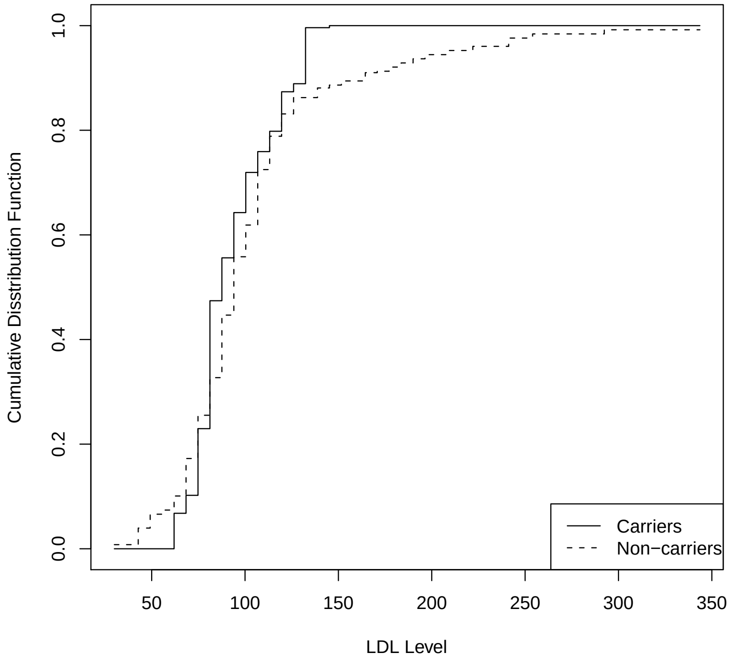

Fig 4. *Data example 2. Estimated CDF of LDL levels for carriers of APOE $\varepsilon 4$ allele (solid) and non-carriers (dashed)*

## 7. Discussion

We have developed nonparametric estimation procedures for mixed samples where the conditional distribution of the outcome given the group membership is completely unspecified and the mixing probabilities are known or can be calculated without using the outcome data. We propose an extremely simple optimal weighted least squares estimator and derive an easy-to-compute efficient estimator which reaches the semiparametric efficiency bound. We illustrate by simulations that the OWLS estimator has good efficiency in many practical situations. We investigate performances of two types of NPMLE and show the surprising results that none of them is efficient and one of them is not even consistent. This is in contrast to many other semiparametric problems where the NPMLE is an efficient estimator.

Although the estimators are constructed for CDFs, it is straightforward to adapt these procedures to estimate a quantile function $F^{-1}(\tau)$. This is because we can then express all the estimators in terms of solving for $F(t)$ from an estimating equation. When we denote $t = F^{-1}(\tau)$, replace $F(t)$ with $\tau$ in these estimating equations, and solve for $t$ from the known $\tau$ value instead of solving for $F(t)$ from the known $t$ value, we can obtain estimators for the quantile functions. For example, the efficient quantile estimator at $\tau$ can be obtained through solving for $t$ from

$$n^{-1}\sum_{i=1}^{n}\frac{\{I(s_i\le t)-K(t,\widehat{q^Tf})\}A^{-1}(s_i;\widehat{q^Tf})q_i}{\widehat{q_i^Tf(s_i)}}+K(t,\widehat{q^Tf})1_p=\tau,$$

where $K$ itself is now a function of $t$ hence we use the notation $K(t,\widehat{q^Tf})$.

The CDFs estimated by the consistent estimators may not be monotone increasing functions of $t$ when the sample size is relatively small. In fact, the type II NPMLE was originally proposed to address this issue, but it unfortunately lead to inconsistency. One way to guarantee the monotonicity is though reparametrization. For example, we could write $f(t)=e^{g(t)}\exp\{-\int_0^t e^{g(u)}du\}$, and treat $g(u)$ as a nuisance parameter, which will guarantee the range of $F(t)=1-\exp\{-\int_0^t e^{g(u)}du\}$ to be monotone and within 0 and 1. However, the additional complexity may not be worth the gain. Instead, we suggest to use a post estimation adjustment, such as a pooled adjacent algorithm (Barlow et al., 1972) to modify the results to achieve monotonicity. For a detailed description, see Wang et al. (2007).

Finally, we point out that one needs to be cautious in interpreting inconsistency of the type II NPMLE. The inconsistency occurs when a pure nonparametric model is used. Parametric models and semiparametric models such as Cox proportional hazards model with a nonparametric baseline or piecewise exponential models are likely to be consistent. An extension of the proposed methods to handle censoring based on full data influence functions discovered here and inverse probability weighting is underway.

## Acknowledgements

The authors wish to thank Dr. Steve Shea and Dr. Rongling Wu for providing data. Ma's research is supported by an NSF grant DMS-0906341, DMS-1206693 and NIH grant NS073671-01. Wang's research is supported by NIH grants AG031113-01A2 and NS073671-01.

## Appendix

### *A.1. Derivation of the complete influence function family*

To perform a formal semiparametric analysis (Bickel et al., 1993; Tsiatis, 2006), we denote by $\theta$ the function that maps the nuisance parameter $f(s)$ to the $p$-dimensional parameter of interest, $F(t)$, i.e., $\theta\{f(s)\}=\int_{T_1}^t f(s)ds$. We denote the infinite dimensional nuisance parameter $f(s)$ as $\eta$, i.e., $\eta=f(x)$.

We now derive a general class of consistent estimators through characterizing the complete influence function set. An influence function $\phi(q,s;\theta,\eta)$ is a mean zero function that satisfies

$$E(\phi S_\gamma^T)=\partial\theta(\gamma_0)/\partial\gamma^T \tag{A.1}$$

for any parametric submodel. A parametric submodel is a model where the original unknown function $f(s)$ is replaced by a parametric PDF model $f(s;\gamma)$, and it satisfies $f(s;\gamma_0) = f_0(s)$. Here $S_\gamma$ is the score function with respect to $\gamma$ evaluated at $\gamma_0$,

$$S_\gamma = \frac{\partial \log\{p_Q(q)q^T f(s;\gamma)\}}{\partial\gamma}\Big|_{\gamma=\gamma_0}$$

and

$$\theta(\gamma) \equiv \theta\{f(s;\gamma)\} = \int_{T_1}^{t} f(s;\gamma)ds.$$

The relation in (A.1) indicates that

$$\int\int_{T_1}^{T_2} \phi q^T \frac{\partial f(s;\gamma_0)}{\partial\gamma^T} ds p_Q(q) d\mu(q) = \int_{T_1}^{t} \frac{\partial f(s;\gamma_0)}{\partial\gamma^T} ds,$$

where $\mu(q)$ is the counting measure of $Q$.

Given any parametric submodel of the form $g(q,s;\gamma) = p_Q(q)q^T f(s;\gamma)$, where $\gamma = (\gamma_1,\ldots,\gamma_p)^T$, and $f(s;\gamma) = \{f_1(s;\gamma_1),\ldots,f_p(s;\gamma_p)\}^T$, the parameter of interest is

$$\begin{aligned}
\theta\{f(s;\gamma)\} &= \left\{\int_{T_1}^{t} f_1(s;\gamma_1)ds,\ldots,\int_{T_1}^{t} f_p(s;\gamma_p)ds\right\}^T \\
&= \left\{\int_{T_1}^{T_2} I(s\le t) f_1(s;\gamma_1)ds,\ldots,\int_{T_1}^{T_2} I(s\le t) f_p(s;\gamma_p)ds\right\}^T.
\end{aligned}$$

On one hand, the partial derivative of the parameter of interest with respect to $\gamma$ is a block diagonal matrix of the form

$$\begin{aligned}
&\frac{\partial\theta\{f(s;\gamma)\}}{\partial\gamma^T}\Big|_{\gamma=\gamma_0} \\
= \;& \operatorname{diag}\left\{\int_{T_1}^{T_2} I(s\le t) f'_{1\gamma_1}(s;\gamma_{10})ds,\ldots,\int_{T_1}^{T_2} I(s\le t) f'_{p\gamma_p}(s;\gamma_{p0})ds\right\}.
\end{aligned}$$

On the other hand, the score vector $S_\gamma$ evaluated at the truth is

$$S_\gamma = \left\{\frac{q_1 f'^T_{1\gamma_1}(s;\gamma_{10})}{q^T f(s)},\ldots,\frac{q_p f'^T_{p\gamma_p}(s;\gamma_{p0})}{q^T f(s)}\right\}^T.$$

Recall that (A.1) requires

$$\int_{T_1}^{T_2} I(s\le t) f'_{j\gamma_j}(s;\gamma_{j0})ds = \int\int_{T_1}^{T_2} \phi_j q_j f'_{j\gamma_j}(s;\gamma_{j0}) p_Q(q) ds d\mu(q)$$

for $j = 1, \dots, p$, and

$$\int \int_{T_1}^{T_2} \phi_k q_j f'_{j\gamma_j}(s; \gamma_{j0}) p_Q(q) ds d\mu(q) = 0$$

for $k \neq j$. Here $\phi_j$ is the $j$th component of $\phi$. Because $f(s)$ is completely unspecified, the function $f'_\gamma(s; \gamma_0)$ can be any function that satisfies $\int_{T_1}^{T_2} f'_\gamma(s; \gamma_0) ds = 0$. It then follows almost everywhere that $\int \phi_j q_j p_Q(q) d\mu(q) - I(s \leq t)$ is a constant and $\int \phi_j q_k p_Q(q) d\mu(q)$ is also a constant for $k \neq j$. These requirements can be written concisely as

$$\int \phi(q,s) q^T p_Q(q) d\mu(q) = I(s \leq t) I_p + C. \tag{A.2}$$

Note that a legitimate influence function also needs to have mean zero, hence

$$0 = E(\phi) = \int_{T_1}^{T_2} \int \phi(q,s) q^T p_Q(q) d\mu(q) f(s) ds = F(t) + C 1_p.$$

Thus, we can write $\phi(q,s)$ as $\phi(q,s) = b(q,s) - F(t) - C1_p$, where $b$ satisfies (A.2). This gives the desired family of influence functions described in (2).

### *A.2. Influence function of the WLS*

Denote the $i$th diagonal entry in $M$ as $w_i$ for $i = 1, \dots, n$. When we view the weight $w_i$ as a random variable, we denote it as $W_i$. Since our arguments are general for any $i = 1, \dots, n$, we often omit the subscript $i$, and use $w$ or $W$ for the corresponding quantities. From

$$\widehat{F}(t) = \left( \frac{1}{n} \sum_{i=1}^n w_i q_i q_i^T \right)^{-1} \frac{1}{n} \sum_{i=1}^n w_i q_i I(s_i \leq t),$$

we obtain

$$\begin{aligned}\sqrt{n}\{\widehat{F}(t) - F(t)\} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \left( \frac{1}{n} \sum_{i=1}^n w_i q_i q_i^T \right)^{-1} w_i q_i I(s_i \leq t) \right\} - \sqrt{n} F(t) \\ &= \frac{1}{\sqrt{n}} \left( \frac{1}{n} \sum_{i=1}^n w_i q_i q_i^T \right)^{-1} \sum_{i=1}^n \left\{ w_i q_i I(s_i \leq t) - w_i q_i q_i^T F(t) \right\}.\end{aligned}$$

Note that $E\left\{ W_i Q_i I(S_i \leq t) - W_i Q_i Q_i^T F(t) \right\} = 0$, hence

$$\sqrt{n}\{\widehat{F}(t) - F(t)\} = \frac{1}{\sqrt{n}} \{E(WQQ^T)\}^{-1} \sum_{i=1}^n \{ w_i q_i I(s_i \leq t) - w_i q_i q_i^T F(t) \} + o_p(1).$$

So the influence function of WLS is

$$\phi_{WLS}(q,s) = \{E(WQQ^T)\}^{-1} wq \left\{I(s \le t) - q^T F(t)\right\}.$$

Specifically, for the OLS and the optimal WLS estimators, the influence functions are respectively

$$\begin{aligned}\phi_{OLS}(q,s) &= \{E(QQ^T)\}^{-1} q \left\{I(s \le t) - q^T F(t)\right\},\\ \text{and } \phi_{OWLS}(q,s) &= \left[E\frac{QQ^T}{Q^T F(t)\{1-Q^T F(t)\}}\right]^{-1} \frac{q\left\{I(s \le t) - q^T F(t)\right\}}{q^T F(t)\{1-q^T F(t)\}}.\end{aligned}$$

### *A.3. Derivation of $\Lambda_{\mathcal{T}}$ and $\Lambda_{\mathcal{T}}^{\perp}$*

We denote the collection of mean zero functions orthogonal to all the elements in $\Lambda_{\mathcal{T}}$ as $\Lambda_{\mathcal{T}}^{\perp}$. Consider the space of tangent vectors contributed from the $j$th component $f_j(s)$ only, we obtain

$$\Lambda_j = \left\{\frac{q_j h(s)}{q^T f(s)} : \int h(s)ds = 0, h \in R^p\right\}.$$

Combining the $\Lambda_j$'s for $j = 1, \dots, p$, the nuisance tangent space is therefore

$$\Lambda_{\mathcal{T}} = \left\{\frac{h(s)q}{q^T f(s)} : \int h(s)ds = 0, h \in R^{p\times p}\right\}.$$

Furthermore, it is easy to see that

$$\Lambda_{\mathcal{T}}^{\perp} = \left\{r(q,s) : \int r(q,s)q^T p_Q(q)d\mu(q) = C, C1_p = 0\right\},$$

where $C$ is a constant $p \times p$ matrix.

### *A.4. Proof of Theorem 1*

We only need to verify that $\phi_{eff}$ given in Theorem 1 satisfies $\phi_{eff} = \Pi(\phi|\Lambda_{\mathcal{T}}) = \phi - \Pi(\phi|\Lambda_{\mathcal{T}}^{\perp})$, where $\Pi$ denotes an orthogonal projection.

To show this, we first point out that $K1_p = F(t)$. This is because from the definition of $A(s)$, we have

$$f(s) = A^{-1}(s)\int \frac{qq^T f(s)p_Q(q)}{q^T f(s)} d\mu(q) = A^{-1}(s)\int qp_Q(q)d\mu(q).$$

Integrate the both sides of the above equation from $T_1$ to $T_2$ and from $T_1$ to $t$ respectively, we obtain

$$\begin{aligned}1_p &= \int_{T_1}^{T_2} A^{-1}(s)ds \int qp_Q(q)d\mu(q),\\ F(t) &= \int_{T_1}^{T_2} I(s \le t)A^{-1}(s)ds \int qp_Q(q)d\mu(q),\end{aligned}$$

and the result follows.

Now, letting $h_1(q,s) = h_2(q,s) = A^{-1}(s)q/q^T f(s)$, $h_3(q) = K1_p$ and $B = -K$, we can easily verify that the corresponding $b(q,s)$ in (4) has the form

$$b_{eff} = \{I(s \leq t)I_p - K\}\frac{A^{-1}(s)q}{q^T f(s)} + K1_p.$$

Since

$$\int b_{eff}(q,s)q^T p_Q(q)d\mu(q) = I(s \leq t)I_p - K + K1_p \int q^T p_Q(q)d\mu(q),$$

its corresponding influence function is

$$\begin{aligned} & b_{eff}(q,s) - F(t) - \left\{-K + K1_p \int q^T p_Q(q)d\mu(q)\right\} 1_p \\ = \quad & b_{eff}(q,s) - F(t) + K1_p - K1_p \int q^T 1_p p_Q(q)d\mu(q) \\ = \quad & b_{eff}(q,s) - F(t). \end{aligned}$$

Note that the above expression equals $\phi_{eff}$. Thus, we have shown that $\phi_{eff}$ is a valid influence function hence $\phi_{eff} \in \Lambda_{\mathcal{T}}$.

Now, for any $\phi \in \Lambda_{\mathcal{T}}$, we need to show $\phi - \phi_{eff} \in \Lambda_{\mathcal{T}}^{\perp}$. We have

$$\begin{aligned} & \int (\phi - \phi_{eff})\, q^T p_Q(q)d\mu(q) \\ = \quad & \int \left[\phi - \{I(s \leq t)I_p - K\}\frac{A^{-1}(s)q}{q^T f(s)}\right] q^T p_Q(q)d\mu(q) \\ = \quad & \int \phi q^T p_Q(q)d\mu(q) - \{I(s \leq t)I_p - K\} \\ = \quad & -C - \{F(t) + C1_p\} \int q^T p_Q(q)d\mu(q) + K \end{aligned}$$

is a constant matrix. In the last equality, we used the fact that an influence function $\phi$ can be written as $\phi = b - F(t) - C1_p$, where $\int dq^T p_Q(q)d\mu(q) = I(s \leq t)I_p - C$. From

$$\left[-C - \{F(t) + C1_p\}\int q^T p_Q(q)d\mu(q) + K\right]1_p = -C1_p - \{F(t) + C1_p\} + K1_p = 0$$

and follow the description of $\Lambda_{\mathcal{T}}^{\perp}$, we indeed have $\phi - \phi_{eff} \in \Lambda_{\mathcal{T}}^{\perp}$. □

### *A.5. Proof of Theorem 2*

First, we note that all the approximations are caused by $\widehat{q^T f}$, which is estimated using the second subset of the data. No other estimation or approximation is involved in our construction. From (7) we obtain

$$n_1^{1/2}\left\{\widehat{F}(t)-F(t)\right\} = n_1^{-1/2}\sum_{i=1}^{n_1}\{\psi(q_i,s_i;\widehat{q^Tf})-F(t)\}$$

$$= n_1^{-1/2}\sum_{i=1}^{n_1}\{\psi(q_i,s_i;q^Tf)-F(t)\}+n_1^{-1/2}\sum_{i=1}^{n_1}\{\psi(q_i,s_i;\widehat{q^Tf})-\psi(q_i,s_i;q^Tf)\}.$$

Note that $A(s;q^Tf)=A(s), K(q^Tf)=K$, and $K1_p=F(t)$, hence

$$\psi(q,s;q^Tf)-F(t) \quad = \quad \frac{\{I(s_i\le t)-K\}\,A^{-1}(s)q}{q^Tf(s)}+K1_p-F(t)=\phi_{eff}(q,s).$$

From (5), we see that $\psi(q,s;\widehat{q^Tf})-F(t)$ is an influence function. Thus, the difference between $\psi(q,s;\widehat{q^Tf})$ and $\psi(q,s;q^Tf)$ is the difference between a valid influence functions and its projection on $\Lambda_{\mathcal{T}}$, hence is orthogonal to $\Lambda_{\mathcal{T}}$. Specifically, we have

$$\psi(q,s;\widehat{q^Tf})-\psi(q,s;q^Tf)=\{\psi(q,s;\widehat{q^Tf})-F(t)\}-\{\psi(q,s;q^Tf)-F(t)\}\perp\Lambda_{\mathcal{T}}$$

and it has mean zero. Consequently, the estimator $\widehat{F}(t)$ is consistent and has variance

$$\mathrm{var}\left[n_1^{1/2}\left\{\widehat{F}(t)-F(t)\right\}\right]=\mathrm{var}(\phi_{eff})+\mathrm{var}\{\psi(q,s;\widehat{q^Tf})-\psi(q,s;q^Tf)\}.$$

When $n_2\to\infty$, the number of observations that satisfy $q_i=u_k$ also goes to infinity in probability due to the randomness of the data. Thus, the kernel estimator for $\widehat{u_k^Tf(s)}$ satisfies $\widehat{u_k^Tf(s)}-u_k^Tf(s)=o_p(1)$ uniformly on any compact set of $s$ for each $k\in\{1,\ldots,m\}$. Therefore, $\widehat{q^Tf}(s)-q^Tf(s)=o_p(1)$ as $n\to\infty$. Note that $\psi(q,s;q^Tf)$ is a pathwise differentiable function of $q^Tf$, it then follows that $\mathrm{var}\{\psi(q,s;\widehat{q^Tf})-\psi(q,s;q^Tf)\}=o(1)$. This proves that $\widehat{F}(t)$ is indeed an efficient estimator. □

### *A.6. Inconsistency of the type II NPMLE*

Consider a very simple and explicit case where $p=m=2$, $u_2=(1,0)^T$, while $u_1\neq(1,0)^T$ and $u_1\neq(0,1)^T$. This corresponds to the situation where there exists two genotypes, and for the first $r_1$ observations we know that they belong to the first group with probability $u_{11}$ and belong to the second group with probability $u_{12}=1-u_{11}$; while for the last $r_2$ observations, we know that they are from the first group. Under this special case, the NPMLE becomes

$$\max_{h_1,h_2}(u_1^Th_1)^{r_1}(u_2^Th_2)^{r_2}=(u_{11}h_{11}+u_{12}h_{12})^{r_1}h_{21}^{r_2}$$

subject to $r_1h_{11}+r_2h_{21}=1$, $r_1h_{12}+r_2h_{22}=1$, and $h_{ij}\ge 0$ for $i,j=1,2$. Obviously, the maximum is obtained only when $h_{12}=r_1^{-1}$ and $h_{22}=0$.

This can be written as $\widehat{f}_2(s_i) = r_1^{-1} I(q_i = u_1)$ for all $i = 1, \ldots, n$. Hence the NPMLE2 for the PDF $f_2(s)$ puts zero weights on observations that are known to be drawn from the first group, and puts equal weights, $r_1^{-1}$, on other observations. Such result is equivalent to the standard empirical likelihood estimation of a PDF when we are only given observations $s_1, \ldots, s_{r_1}$ drawn as a random sample from this PDF. Hence its corresponding CDF estimation $\widehat{F}_2(t) = \sum_{i=1}^n \widehat{f}_2(s_i) I(s_i \leq t) = r_1^{-1} \sum_{i=1}^{r_1} I(q_i = u_1) I(s_i \leq t)$ is a consistent estimate of the corresponding true CDF. However, $s_1, \ldots, s_{r_1}$ is a random sample from a mixture of two populations, where the mixture probability is $u_{11}$ for being from the first population and is $u_{12}$ for the second population. In other words, the estimator $\widehat{F}_2(t)$ is a consistent estimator of $u_{11}F_1(t) + u_{12}F_2(t)$. Obviously, $u_{11}F_1(t) + u_{12}F_2(t)$ does not equal to $F_2(t)$ unless $u_{11} \equiv 0$. Consequently, the type II NPMLE is not consistent for this simple case.

## References

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions*. New York: John Wiley.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: The Johns Hopkins University Press. MR1245941

Chatterjee, N. and Wacholder, S. (2001). "A Marginal Likelihood Approach for Estimating Penetrance from Kin-cohort Designs". *Biometrics*, **57**, 245-252. MR1833313

Davignon, J., Gregg, R.E. and Sing, C.F. (1988). "Apolipoprotein E Polymorphism and Atherosclerosis". *Arteriosclerosis*, **8**, 1-21.

Fine, J.P., Zou, F. and Yandell, B.S. (2004). Nonparametric estimation of the effects of quantitative trait loci. *Biometrics*, **5**, 501-513.

Hartge, P., Chatterjee, N., Wacholder, S., Brody, L.C., Tucker, M.A., Struewing, J.P. (2002). Breast cancer risk in Ashkenazi BRCA1/2 mutation carriers: effects of reproductive history. *Epidemiology.* **13(3)**, 255-261.

Hauptmann, M., Sigurdson, A.J., Chatterjee, N., Rutter, J.L., Hill, D.A., Doody, M.M., Struewing, J.P. (2003). Re: Population-Based, CaseControl Study of HER2 Genetic Polymorphism and Breast Cancer Risk. *Journal of the National Cancer Institute*, **95**, 1251-1252.

Hixson, J.E. (1991). "Apolipoprotein E Polymorphisms Affect Atherosclerosis in Young Males: Pathobiological Determinants of Atherosclerosis in Youth (PDAY) Research Group". *Arterioscler Thromb*, **11**, 237-244.

Huang, N., Parco, A., Mew, T., Magpantay, G., McCouch, S., Gulderdoni, E., Xu, J., Subudhi, P., Angeles, E. and Khush, G. (1997). "RFLP Mapping of Isozymes, RAPD and QTLs for Grain Shape, Brown Planthopper Resistance in a Doubled Haploid Rice Population". *Molecular Breeding.* **3**, 105-113

Khoury, M., Beaty, H. and Cohen, B. (1993). *Fundamentals of Genetic Epidemiology.* New York: Oxford University Press.

Lander, E.S. and Botstein, D. (1989). “Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps”. *Genetics*, **121** 743-756.

Li, R. and Liang, H. (2008). “Variable selection in semiparametric regression modeling”. *Annals of Statistics*, **36**, 261-286. MR2387971

Liang, H. and Wang, N. (2005). “Large sample theory in a semiparametric partially linear errors-in-variables model”. *Statistica Sinica*, **15**, 99-117 MR2125722

Marder K., Levy, G., Louis, E.D., Mejia-Santana, H., Cote, L., Andrews, H., Harris, J., Waters, C., Ford, B., Frucht, S., Fahn, S. and Ottman, R. (2003). Accuracy of family history data on Parkinson’s disease. *Neurology*, 61, 18-23.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley. MR1789474

Newey, W.K. (1990). “Semiparametric Efficiency Bounds”. *Journal of Applied Econometrics*, **5**, 99-135.

Rabinowitz, D. (2000). “Computing the Efficient Score in Semi-parametric Problems”. *Statistica Sinica*, **10**, 265-280. MR1742112

Sigurdson, A.J., Hauptmann, M., Chatterjee, N., Alexander, B.H., Doody, M.M., Rutter, J.L., Struewing, J.P. (2004). Kin-cohort estimates for familial breast cancer risk in relation to variants in DNA base excision repair, BRCA1 interacting and growth factor genes. *BMC Cancer*, **4**, 9.

Shea, S., Isasi, C.R., Couch, S., Starc, T.J., Tracy, R.P., Deckelbaum, R., Talmud, P., Berglund, L., and Humphries, S.E. (1999). “Relations of Plasma Fibrinogen Level in Children to Measures of Obesity, the (G-455->A) Mutation in the Beta-Fibrinogen Promoter Gene, and Family History of Ischemic Heart Disease: the Columbia University BioMarkers Study”. *American Journal of Epidemiology*, **150**, 737-46.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer. MR2233926

Tsiatis, A.A. and Ma, Y. (2004). “Locally Efficient Semiparametric Estimators for Functional Measurement Error Models”. *Biometrika*, **91**, 835-848. MR2126036

Wacholder, S., Hartge, P., Struewing, J., Pee, D., McAdams, M., Brody, L. and Tucker, M. (1998). “The Kin-cohort Study for Estimating Penetrance”. *American Journal of Epidemiology*, **148**, 623–630.

Wang, Y., Clark, L.N., Marder, K. and Rabinowitz, D. (2007). “Nonparametric Estimation of Genotype-specific Age-at-onset Distributions From Censored Kin-cohort Data”. *Biometrika*, **94**, 403-414. MR2380568

Wang, Y., Clark, L.N., Louis, E.D., Mejia-Santana, H., Harris, J., Cote, L.J., Waters, C., Andrews, D., Ford, B., Frucht, S., Fahn, S., Ottman, R., Rabinowitz, D. and Marder, K. (2008). Risk of Parkinson’s disease in carriers of Parkin mutations: estimation using the kin-cohort method. *Arch Neurol.* 65(4):467-474.PMID: 18413468

WEBB, E.L., RUDD, M.F., AND HOULSTON, R.S. (2006a). Case-control, kin-cohort and meta-analyses provide no support for STK15 F31I as a low penetrance colorectal cancer allele. *British Journal of Cancer*, **95**, 1047-1049.

WEBB, E.L., RUDD, M.F., SELLICK, G.S., GALTA, R., BETHKE, L., WOOD, W., FLETCHER, O., PENEGAR, S., WITHEY, L., QURESHI, M., JOHNSON, N., TOMLINSON, I., GRAY, R., PETO, J., HOULSTON, R.S. (2006b). Search for low penetrance alleles for colorectal cancer through a scan of 1467 non-synonymous SNPs in 2575 cases and 2707 controls with validation by kin-cohort analysis of 14 704 first-degree relatives. *Hum Mol Genet*, **15(21)**, 3263-3271.

WU, R., MA, C., AND CASELLA, G. (2007). *S*tatistical Genetics of Quantitative Traits: Linkage, Maps, and QTL. New York: Springer. MR2344949