Electronic Journal of Statistics
Vol. 11 (2017) 1928–1948
ISSN: 1935-7524
DOI: 10.1214/17-EJS1274

Nonparametric distribution estimation in the presence of familial correlation and censoring

Kun Xu

Novartis Pharmaceuticals Corporation,
East Hanover, NJ, USA
e-mail: kun.xu@novartis.com

Yanyuan Ma

Department of Statistics, Penn State University,
University Park, PA, USA
e-mail: yzm63@psu.edu

and

Yuanjia Wang

Department of Biostatistics, Mailman School of Public Health,
Columbia University,
New York, NY, USA
e-mail: yuanjia.wang@columbia.edu

Abstract: We propose methods to estimate the distribution functions for multiple populations from mixture data that are only known to belong to a specific population with certain probabilities. The problem is motivated from kin-cohort studies collecting phenotype data in families for various diseases such as the Huntington's disease (HD) or breast cancer. Relatives in these studies are not genotyped hence only their probabilities of carrying a known causal mutation (e.g., BRCA1 gene mutation or HD gene mutation) can be derived. In addition, phenotype observations from the same family may be correlated due to shared life style or other genes associated with disease, and the observations are subject to censoring. Our estimator does not assume any parametric form of the distributions, and does not require modeling of the correlation structure. It estimates the distributions through using the optimal base estimators and then optimally combine them. The optimality implies both estimation consistency and minimum estimation variance. Simulations and real data analysis on an HD study are performed to illustrate the improved efficiency of the proposed methods.

MSC 2010 subject classifications: Primary 62G08; secondary 62N01.
Keywords and phrases: Bootstrap, efficiency, familial correlation, Huntington's disease, mixed samples, quadratic inference function.

Received May 2016.

Contents

1. Introduction

This work is motivated by research goals arise from studies such as the Cooperative Huntington's Observational Research Trial (COHORT, Dorsey and the Huntington Study Group COHORT Investigators, 2012 [3]). Huntington's disease (HD) is a fatal neurodegenerative disease caused by expanded C-A-G repeats in the huntingtin gene (Huntington Study Group, 1993 [6]). Subjects with expanded CAG repeats at the huntingtin gene will be affected by HD and the distribution of age-at-onset of HD in individuals with mutation is characterized in Langbehn et al. (2004) [8] and Ma and Wang (2012) [10]. However, less discussed in the literature is the effect of CAG expansion status on patient survival. Here, we aim to estimate the age-at-death distribution function for the HD gene expanded individuals. In COHORT, the phenotype information (age-at-death or age at study baseline) in relatives from the same families was collected. One challenge is that the genotype information in a relative is not collected and thus the gene expansion status in some relatives are unknown. Nevertheless, a relative's probability of carrying the huntingtin gene mutation can be obtained from external information (see for example, Ma and Wang 2012 [10]).

Other challenges for the COHORT as well as other family studies include handling of the correlation among the relatives in the same family and the right censoring. Using the probability of carrying a mutation associated with each relative and assuming that the observations are independent given the mutation status, the distribution of any trait of interest for both the mutation carrier and non-carrier populations can be estimated efficiently (Ma and Wang, 2012 [10]) in the absence of censoring. When data are also subject to censoring, Ma and Wang

(2014) [11] further developed effective methods to perform the estimation and inference through a weighted-least-squares formulation. However, both methods assume that relatives in the same family are independent given their genotypes (i.e., no residual familial correlation), and hence it is unclear how to properly handle the potential correlation due to shared life style or causal genes at other loci to improve estimation efficiency.

In this work, we address the within-family correlation for censored mixture data where the subjects' population group identifiers (i.e., mutation carriers or non-carriers) are only known up to the probabilities. Since the familial correlation can be a result of similar life style or shared biological markers other than the gene under study, it may be challenging to choose a satisfactory parametric model for the distribution of the shared latent effects, and therefore we do not make such attempts. We provide two different modeling approaches, and subsequently two estimation procedures. In the first approach, we leave the distribution of the unobserved latent familial effects and their correlation structure completely unspecified. We first eliminate the need to handle the familial correlation by using only one member per family to form a base estimator, and then construct an optimal new estimator that takes advantage of multiple members in a family by resampling and minimizing the variance of the combined estimator. When forming the base estimator, we use the approach by Ma and Wang (2014) [11], which is simple and practically as effective as the efficient estimator. This first proposed method can handle arbitrary distribution functions and arbitrary correlation structures without imposing parametric assumptions or modeling the correlation. In addition, it is easy to compute and flexible. In the second approach, we assume exchangeable correlation structure between family members to improve estimation accuracy but leave the distribution function of the phenotype unspecified to protect against misspecification. In this case, we proceed with a modified weighted least square estimator that takes full advantage of the assumed correlation structure.

The rest of the paper is organized as follows. We present the methods, describe implementation, and demonstrate their optimality property in Section 2. Simulations are carried out in Section 3 to illustrate the performance of the estimators in both simple and complex settings. Finally, we analyze the COHORT data which motivated this work in Section 4 and conclude the paper with some discussions in Section 5. All the technical derivations are in an Appendix.

2. Methodologies

We first define some notations. Suppose there are N families in the study, and the ith family has n_i members, $i = 1, \ldots, N$. The random event time for the jth member of the ith family is S_{ij}. Further, the event is subject to random censoring at time C_{ij}. Let $Y_{ij} = \min(S_{ij}, C_{ij})$ and the censoring indicator $\Delta_{ij} = I(S_{ij} \leq C_{ij})$. Furthermore, we assume there are p different populations, and their event times have cumulative distribution functions $F_1(t), F_2(t), \ldots, F_p(t)$ respectively. Write $\mathbf{F}(t) = \{F_1(t), F_2(t), \ldots, F_p(t)\}^{\mathrm{T}}$. We assume the p event

processes associated with the p populations are independent of the censoring process. Assume for all $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, S_{ij} is a random sample from one of the p populations, but the exact population identifier is not known. We use q_{ijk} to denote the probability of S_{ij} belonging to the kth population, for $k = 1, \ldots, p$. Let $\mathbf{q}_{ij} = (q_{ij1}, q_{ij2}, \ldots, q_{ijp})^{\mathrm{T}}$. Obviously, $\sum_{k=1}^{p} q_{ijk} = 1$. Using these notations, the observed data can be written as $\mathbf{O} = \{(\mathbf{q}_{ij}, Y_{ij}, \Delta_{ij}),\ i = 1, \ldots, N,\ j = 1, \ldots, n_i\}$.

The above data structure typically arises from kin-cohort and quantitative trait locus (QTL) studies, where $\mathbf{q}_{ij}$ only has finitely many, say m, $m < \infty$, possible values, denoted as $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$. We write the frequencies of the occurrences of $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ as $d_1, d_2, \ldots, d_m$. Obviously, $\sum_{i=1}^{m} d_i = \sum_{i=1}^{N} n_i$. For $l = 1, \ldots, m$, we write the m mixture distributions as $H_l(t) = \mathbf{u}_l^{\mathrm{T}} \mathbf{F}(t)$, and let $\mathbf{H}(t) = \{H_1(t), H_2(t), \ldots, H_m(t)\}^{\mathrm{T}}$.

2.1. Independent case

We consider a special case where the observations $\mathbf{O} = \{(\mathbf{q}_{ij}, Y_{ij}, \Delta_{ij}),\ i = 1, \ldots, N,\ j = 1, \ldots, n_i\}$ are independent of each other. Obviously, this happens when each family has only one observation, i.e. $n_i = 1$ for $i = 1, \ldots, N$. This also happens when there is no within-family correlation. In this case, following Ma and Wang (2014) [11], we use the relation

$$\mathbf{F}(t) = \left(\sum_{l=1}^{m} d_l \mathbf{u}_l \mathbf{u}_l^{\mathrm{T}} \right)^{-1} \left\{ \sum_{l=1}^{m} d_l \mathbf{u}_l H_l(t) \right\}, \tag{2.1}$$

and estimate $\mathbf{F}(t)$ through

$$\widehat{\mathbf{F}}(t) = \left(\sum_{l=1}^{m} d_l \mathbf{u}_l \mathbf{u}_l^{\mathrm{T}} \right)^{-1} \left\{ \sum_{l=1}^{m} d_l \mathbf{u}_l \widehat{H}_l(t) \right\}. \tag{2.2}$$

Here $\widehat{H}_l(t)$ is the Kaplan-Meier (KM) estimate (Kaplan and Meier (1958) [7]) for $H_l(t)$. Kaplan and Meier (1958) [7] has established the consistency for the KM estimator, while Breslow and Crowley (1974) [2] has shown that it converges weakly to a Gaussian process. Since $\widehat{\mathbf{F}}(t)$ is a linear transformation of $\widehat{\mathbf{H}}(t)$, it is also consistent and converges weakly to a Gaussian process when $N \to \infty$. These observations will be used in our following derivation when within family correlation exists.

2.2. Arbitrary correlation 1: Best Linear Resampled Estimator (BLRE)

In the general situation when members from a same family may be correlated, we propose a two stage procedure that utilizes the results described in Section 2.1. In the first stage, we randomly sample one member from each family, regardless

of the family size, and then use (2.2) to obtain a crude estimation of $\mathbf{F}(t)$. Repeat this process multiple, say R, times to collect multiple estimators for $\mathbf{F}(t)$, denoting these estimators $\widehat{\mathbf{F}}^1(t), \ldots, \widehat{\mathbf{F}}^R(t)$. In the second stage, we aim to combine the multiple estimators from the first stage in an optimal way.

Since each $\widehat{\mathbf{F}}^r(t), r = 1, \ldots, R$ is a consistent estimator of $\mathbf{F}(t)$, it is natural to use a weighted average of these estimators to form an estimator that is not only consistent but also more efficient. In general, we write the combined estimator

$$\widehat{\mathbf{F}}(t) = \mathbf{A}\widehat{\mathbf{F}}_L(t), \tag{2.3}$$

where

$$\widehat{\mathbf{F}}_L(t) = \left[\{\widehat{\mathbf{F}}^1(t)\}^{\mathrm{T}}, \{\widehat{\mathbf{F}}^2(t)\}^{\mathrm{T}}, \ldots, \{\widehat{\mathbf{F}}^R(t)\}^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{2.4}$$

and $\mathbf{A}$ is a $p \times pR$ weight matrix. The consistent requirement mandates $\mathbf{AJ} = \mathbf{I}_p$, where $\mathbf{I}_p$ is the size p identity matrix, and $\mathbf{J}$ is a $pR \times p$ matrix formed by $\mathbf{I}_p$'s, i.e. $\mathbf{J} = (\mathbf{I}_p, \ldots, \mathbf{I}_p)^{\mathrm{T}}$. In the Appendix, we further show that the optimal choice of $\mathbf{A}$ in terms of minimizing the variance of $\widehat{\mathbf{F}}(t)$ is $(\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}$, where $\mathbf{U}$ is the asymptotic variance-covariance matrix of $\sqrt{N}\widehat{\mathbf{F}}_L(t)$. We summarize the above results in Theorem 1 and give the proof in the Appendix.

Theorem 1. *Let $\widehat{\mathbf{F}}(t)$ be given in (2.3). Then as long as $\mathbf{AJ} = \mathbf{I}_p$, $\widehat{\mathbf{F}}(t)$ is a consistent estimator of $\mathbf{F}(t)$ and is asymptoticlly normally distributed. In addition, var$\{\widehat{\mathbf{F}}(t)\}$ is minimized when*

$$\mathbf{A}_{\mathrm{opt}} = (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}. \tag{2.5}$$

The resulting optimal variance of $\sqrt{N}\widehat{\mathbf{F}}(t)$ is

$$\mathbf{V}_1^{\mathrm{opt}} = (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}.$$

To take advantage of the result in Theorem 1, we still need to obtain $\mathbf{U}$. Because of our construction of $\widehat{\mathbf{F}}(t)$, $\mathbf{U}$ is naturally an $R \times R$ block matrix with each block size $p \times p$. Although the diagonal blocks of $\mathbf{U}$ can be approximated using results in Section 2.1, the analysis of the off-diagonal blocks is intractable due to the unspecified correlation structure among family members and the potentially complex pattern resulting from the sampling procedure. Thus, we resort to a bootstrap procedure (Efron (1981) [4] and Akritas (1986) [1]) to assess $\mathbf{U}$. Here, caution needs to be taken in performing the bootstrap procedure. In particular, although our interest is to repeatedly draw family members to form estimators, we need to bootstrap families, not members of the families. Specifically, for $b = 1, \ldots, B$, we randomly draw N families with replacement and with equal probability, and denote the bootstrap sample $\mathbf{O}_b^*$. We then repeat the estimation procedure described above on $\mathbf{O}_b^*$ to obtain $\widehat{\mathbf{F}}_L^{*b}(t)$. The sample variance of $\widehat{\mathbf{F}}_L^{*1}(t), \widehat{\mathbf{F}}_L^{*2}(t), \ldots, \widehat{\mathbf{F}}_L^{*B}(t)$ is then used to estimate $\mathbf{U}$.

The complete procedure of our BLRE is the following.

Algorithm 1.

Step 1. Randomly draw one member from each family, assume the resulting sample contains m different $\mathbf{q}$ values, written as $\mathbf{u}_1, \ldots, \mathbf{u}_m$, with frequency $d_1, \ldots, d_m$. Form

$$\widehat{\mathbf{F}}^r(t) = \left(\sum_{l=1}^{m} d_l \mathbf{u}_l \mathbf{u}_l^{\mathrm{T}} \right)^{-1} \left\{ \sum_{l=1}^{m} d_l \mathbf{u}_l \widehat{H}_l^r(t) \right\}.$$

Step 2. Repeat Step 1 R times ($r = 1, \ldots, R$), and form $\widehat{\mathbf{F}}_L(t)$ using (2.4).

Step 3. Randomly sample N families with replacement from the original families. Perform Steps 1 and 2 on the sampled data, obtain the corresponding $\widehat{\mathbf{F}}_L^{*b}(t)$.

Step 4. Repeated Step 3 B times ($b = 1, \ldots, B$) to obtain $\widehat{\mathbf{F}}_L^{*1}(t), \ldots, \widehat{\mathbf{F}}_L^{*B}(t)$.

Step 5. Calculate the sample variance $\widehat{\mathbf{U}}$ of $\widehat{\mathbf{F}}_L^{*1}(t), \ldots, \widehat{\mathbf{F}}_L^{*B}(t)$. Form the estimator $\widehat{\mathbf{F}}(t) \equiv (\mathbf{J}^{\mathrm{T}} \widehat{\mathbf{U}}^{-1} \mathbf{J})^{-1} \mathbf{J}^{\mathrm{T}} \widehat{\mathbf{U}}^{-1} \widehat{\mathbf{F}}_L(t)$.

2.3. Arbitrary correlation 2: Best Quadratic Inference Function Estimator (BQIF)

We now investigate the issue of familial correlation from a different perspective. The basic idea is that every time when we collect one member from each family, we can construct an estimating equation using Ma and Wang (2014) [11] because the observations are now iid. If we repeat this procedure multiple times, we then have multiple estimating equations. It is natural to consider combining these estimating equations optimally to obtain a final estimator. Specifically, when there is only one member per family, we rewrite the relation in (2.1) as

$$\sum_{i=1}^{N} \left\{ \mathbf{q}_{i1} H_i(t) - \mathbf{q}_{i1} \mathbf{q}_{i1}^{\mathrm{T}} \mathbf{F}(t) \right\} = \mathbf{0}, \tag{2.6}$$

and view $\widehat{\mathbf{F}}(t)$ as the root that solves

$$\sum_{i=1}^{N} \left\{ \mathbf{q}_{i1} \widehat{H}_i(t) - \mathbf{q}_{i1} \mathbf{q}_{i1}^{\mathrm{T}} \mathbf{F}(t) \right\} = \mathbf{0}, \tag{2.7}$$

where $\widehat{H}_i(t)$ is the same KM estimator as before.

We use the sampling scheme in the first stage of BLRE in section 2.2 to sample R data sets, and write the estimating equation (2.7) based on the rth sampled data $\sum_{i=1}^{N} \mathbf{g}_i^r(t) = \sum_{i=1}^{N} \left\{ \mathbf{q}_i^r \widehat{H}_i^r(t) - \mathbf{q}_i^r (\mathbf{q}_i^r)^{\mathrm{T}} \mathbf{F}(t) \right\} = \mathbf{0}$, $r = 1, \ldots, R$. Here, $\mathbf{q}_i^r$ denotes the $\mathbf{q}$ value of the member from the ith family in the rth sample. Because the number of equations, pR, can be much larger than the number of the parameters p, we resort to the Quadratic Inference Function (QIF) method

(Lindsay and Qu 2003 [9]). Write

$$\sum_{i=1}^{N}\mathbf{g}_i(t)=\sum_{i=1}^{N}\begin{Bmatrix}\mathbf{g}_i^1(t)\\ \mathbf{g}_i^2(t)\\ \vdots\\ \mathbf{g}_i^R(t)\end{Bmatrix}=\sum_{i=1}^{N}\begin{Bmatrix}\mathbf{q}_i^1\widehat{H}_i^1(t)-\mathbf{q}_i^1(\mathbf{q}_i^1)^{\mathrm{T}}\mathbf{F}(t)\\ \mathbf{q}_i^2\widehat{H}_i^2(t)-\mathbf{q}_i^2(\mathbf{q}_i^2)^{\mathrm{T}}\mathbf{F}(t)\\ \vdots\\ \mathbf{q}_i^R\widehat{H}_i^R(t)-\mathbf{q}_i^R(\mathbf{q}_i^R)^{\mathrm{T}}\mathbf{F}(t)\end{Bmatrix}, \tag{2.8}$$

we minimize the quadratic form

$$\left\{\sum_{i=1}^{N}\mathbf{g}_i(t)\right\}^{\mathrm{T}}\mathbf{W}\left\{\sum_{i=1}^{N}\mathbf{g}_i(t)\right\}, \tag{2.9}$$

where $\mathbf{W}$ is a weight matrix. In a typical QIF construction, $\mathbf{g}_i(t)$'s are functions of the ith observation and are hence independent of each other, and the subsequent root-N consistency and asymptotic normality of the resulting estimator have been established in Lindsay and Qu (2003) [9]. However here, it is important to recognize that $\mathbf{g}_i(t)$'s are not independent since they contain $\widehat{H}_i^r(t)$'s, which are estimated based on all the observations from the rth sample for $r=1,\ldots,R$. Nevertheless, in Theorem 2, we show that the resulting estimator still enjoys the usual asymptotic normality property. The proof is in the Appendix.

Theorem 2. *Let* $\widehat{\mathbf{F}}(t)$ *be the minimizer of the quadratic form in (2.9).* $\sqrt{N}\{\widehat{\mathbf{F}}(t)-\mathbf{F}(t)\}\to$ *Normal*$(\mathbf{0},\mathbf{V}_2)$ *in distribution when* $N\to\infty$, *where* $\mathbf{V}_2$ *is a* $p\times p$ *positive-definite matrix. Let* $\mathbf{M}$ *be the asymptotic variance-covariance matrix of* $N^{-\frac{1}{2}}\sum_{i=1}^{N}\mathbf{g}_i(t)$. *When* $\mathbf{W}_{\mathrm{opt}}=\mathbf{M}^{-1}$, $\sqrt{N}\{\widehat{\mathbf{F}}(t)-\mathbf{F}(t)\}$ *achieves the efficiency bound*

$$\mathbf{V}_2^{\mathrm{opt}}=\left[E\left\{\frac{\partial\mathbf{g}_i(t)}{\partial\mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}}\mathbf{M}^{-1}E\left\{\frac{\partial\mathbf{g}_i(t)}{\partial\mathbf{F}^{\mathrm{T}}(t)}\right\}\right]^{-1}. \tag{2.10}$$

Theorem 2 prescribes the choice of the optimal weight matrix. To achieve efficiency, it is essential to estimate $\mathbf{M}$. Because no correlation structure is modeled for members from the same family, we resort to the bootstrap procedure mentioned in Section 2.2 to approximate $\mathbf{M}$. Using the bth bootstrap sample $\mathbf{O}_b^*$, we follow the procedure described above to construct estimation equation $\sum_{i=1}^{N}\mathbf{g}_i^{*b}(t)$. The sample variance of $\sum_{i=1}^{N}\mathbf{g}_i^{*1}(t),\ldots,\sum_{i=1}^{N}\mathbf{g}_i^{*B}(t)$ is then used to estimate $\mathbf{M}$.

The detailed algorithm based on BQIF is the following.

Algorithm 2.

Step 1. Randomly draw one member from each family. Form

$$\sum_{i=1}^{N}\mathbf{g}_i^r(t)=\sum_{i=1}^{N}\left\{\mathbf{q}_i^r\widehat{H}_i^r(t)-\mathbf{q}_i^r(\mathbf{q}_i^r)^{\mathrm{T}}\mathbf{F}(t)\right\}.$$

Step 2. Repeat Step 1 R times ($r = 1, \ldots, R$), and form $\sum_{i=1}^{N} \mathbf{g}_i(t)$ using (2.8).
Step 3. Randomly sample N families with replacement from the original families. Perform Steps 1 and 2 on the sampled data, obtain the corresponding $\sum_{i=1}^{N} \mathbf{g}_i^{*b}(t)$.
Step 4. Repeated Step 3 B times ($b = 1, \ldots, B$) to obtain

$$\sum_{i=1}^{N} \mathbf{g}_i^{*1}(t), \ldots, \sum_{i=1}^{N} \mathbf{g}_i^{*B}(t).$$

Step 5. Calculate the sample variance $\widehat{\mathbf{M}}$ of $\sum_{i=1}^{N} \mathbf{g}_i^{*1}(t), \ldots, \sum_{i=1}^{N} \mathbf{g}_i^{*B}(t)$. Let $\mathbf{W} = \widehat{\mathbf{M}}^{-1}$. Obtain the estimator $\widehat{\mathbf{F}}(t)$ from minimizing (2.9).

In "step 2" of both algorithms (BLRE and BQIF), we need to repeat "step 1" R times. As we repeat, more combinations of family members are formed and included in data analysis. In total there are $\prod_{i=1}^{N} n_i$ ways to form different estimating equations in "step 1" (BQIF). However, in practice, we suggest setting R to be the largest family size initially, and increasing it gradually until no significant improvement can be seen. If the process of increasing R continues and costs substantial computation time, the method described in section 2.5 will be an alternative to consider.

2.4. Equivalence of BLRE and BQIF

To understand the advantages and disadvantages of BLRE and BQIF introduced respectively in Section 2.2 and 2.3, we perform further analysis to compare their relative performance. Given that BLRE is a combination of the estimators from R samples, while BQIF results from solving a combination of estimating equations from the same R samples, it is not surprising that these two procedures are in fact equivalent. In the following, we formally establish that there is a one-to-one mapping between the estimators in the two classes, and in particular, the optimal estimation variances from the two estimators are identical asymptotically.

Because the BLRE is uniquely determined by the weight matrix choice $\mathbf{A}$, while the BQIF is uniquely decided by the weight matrix $\mathbf{W}$, we only need to establish the one-to-one mapping between $\mathbf{A}$ and $\mathbf{W}$ in order to show our results. Define a $pR \times pR$ block diagonal matrix

$$\mathbf{D} = \operatorname{diag}\left[\left\{E\left(\mathbf{q}_{ij}\mathbf{q}_{ij}^{\mathrm{T}}\right)\right\}^{-1}, \ldots, \left\{E\left(\mathbf{q}_{ij}\mathbf{q}_{ij}^{\mathrm{T}}\right)\right\}^{-1}\right].$$

For any weight matrix $\mathbf{W}$ defined in the BQIF estimator, consider

$$\mathbf{A} = (\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1} \tag{2.11}$$

as the weight matrix in BLRE. Obviously, $\mathbf{A}\mathbf{J} = \mathbf{I}_p$. We now investigate the resulting BLRE and BQIF from the corresponding $\mathbf{A}$ and $\mathbf{W}$. Let the BLRE

estimator $\widehat{\mathbf{F}}^{(1)}(t) = \mathbf{A}\widehat{\mathbf{F}}_L(t)$, where $\widehat{\mathbf{F}}_L(t)$ is defined in (2.4). Define $\mathbf{F}_L(t)$ analogously as $\widehat{\mathbf{F}}_L(t)$ and recall the definition of $\mathbf{g}_i(t)$ in (2.8). We can write

$$\sqrt{N}\{\widehat{\mathbf{F}}_L(t) - \mathbf{F}_L(t)\} = N^{-1/2}\mathbf{D}\sum_{i=1}^{N}\mathbf{g}_i(t) + o_p(1), \tag{2.12}$$

which leads to

$$\widehat{\mathbf{F}}^{(1)}(t) = \mathbf{A}\mathbf{D}N^{-1}\sum_{i=1}^{N}\mathbf{g}_i(t) + \mathbf{F}(t) + o_p(N^{-1/2}). \tag{2.13}$$

On the other hand, the BQIF, denoted $\widehat{\mathbf{F}}^{(2)}(t)$, is obtained from minimizing (2.9), thus standard Taylor expansion leads to

$$\begin{aligned}
&\widehat{\mathbf{F}}^{(2)}(t)\\
=\ & -\left[E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}}\mathbf{W}E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}\right]^{-1}E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}}\mathbf{W}N^{-1}\sum_{i=1}^{N}\mathbf{g}_i(t)\\
&+\mathbf{F}(t) + o_p(N^{-1/2})\\
=\ & (\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{W}N^{-1}\sum_{i=1}^{N}\mathbf{g}_i(t) + \mathbf{F}(t) + o_p(N^{-1/2})
\end{aligned} \tag{2.14}$$

where the last equality follows from the relation

$$E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\} = -1_R \otimes E\left\{\mathbf{q}_{ij}(\mathbf{q}_{ij})^{\mathrm{T}}\right\} = -\mathbf{D}^{-1}\mathbf{J}.$$

Further using the connection between $\mathbf{A}$ and $\mathbf{W}$ in (2.11), we immediately have $\widehat{\mathbf{F}}^{(1)}(t) = \widehat{\mathbf{F}}^{(2)}(t) + o_p(N^{-1/2})$. Conversely, if $\widehat{\mathbf{F}}^{(1)}(t) = \widehat{\mathbf{F}}^{(2)}(t) + o_p(N^{-1/2})$, subtraction of (2.14) from (2.13) yields (2.11).

Having established the one-to-one mapping between BLRE and BQIF via (2.11), it is not surprising to expect that the optimal weight matrix choices in the two estimator classes, $\mathbf{A}_{\mathrm{opt}}$ and $\mathbf{W}_{\mathrm{opt}}$, also satisfy (2.11). This can be easily verified through using the equality $\mathbf{U} = \mathbf{D}\mathbf{M}\mathbf{D}$, which follows from (2.12). Furthermore, we can explicitly verify that the two optimal asymptotic estimation variances are identical, i.e.

$$\begin{aligned}
\mathbf{V}_1^{\mathrm{opt}} &= (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1} = (\mathbf{J}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{M}^{-1}\mathbf{D}^{-1}\mathbf{J})^{-1}\\
&= \left[E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}}\mathbf{M}^{-1}E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}\right]^{-1} = \mathbf{V}_2^{\mathrm{opt}}.
\end{aligned}$$

2.5. Structured correlation

The estimators proposed in Section 2.2 rely on resampling because we do not impose any assumption on the familial correlation structure. Here, we consider

a special case where the correlation between family members are identical. The assumption is natural when the data do not contain information that allows us to differentiate various levels of correlation among relatives such as those due to shared life style when living together. We denote the common correlation as ρ.

Recall that in Section 2.1, we demonstrate that in the independent case, the estimator of Ma and Wang (2014) [11] is based on (2.2) where $\widehat{H}_l(t)$ is a KM estimator for $H_l(t)$. We can view this as $\widehat{H}_l(t) = \mathbf{u}_l^{\mathrm{T}}\mathbf{F}(t) + \epsilon_l$, where $\epsilon_l \sim N(0, V_l/d_l)$ for $l = 1, \ldots, m$. Ma and Wang (2014) [11] advocates to replace the V_l's with a common value V and use a weighted least square with m observations to recover $\mathbf{F}(t)$.

However, when familial correlation is ρ instead of zero, when estimating $H_l(t)$, we effectively have fewer than d_l observations. The effective number of observations is

$$\widetilde{d}_l = \frac{d_l^2}{d_l + 2\left\{\binom{d_{l1}}{2} + \cdots + \binom{d_{lN}}{2}\right\}\rho},$$

based on the relation

$$\frac{V}{\widetilde{d}_l} = \frac{d_l V + 2\left\{\binom{d_{l1}}{2} + \cdots + \binom{d_{lN}}{2}\right\}\rho V}{d_l^2}.$$

Here d_{li} is defined as the number of members in family i that belong to the same group l, $i = 1, \ldots, N$. Note that some d_{li}'s may be zero by its definition. Further taking into account the correlations between the m groups in a similar way, we propose to estimate $\mathbf{F}(t)$ through

$$\widehat{\mathbf{F}}(t) = (\mathbf{U}\mathbf{W}\mathbf{U}^{\mathrm{T}})^{-1}\{\mathbf{U}\mathbf{W}\widehat{H}(t)\}, \tag{2.15}$$

where $\mathbf{W}^{-1}$ has diagonal elements $\widetilde{d}_1^{-1}, \ldots, \widetilde{d}_m^{-1}$, and the (l, l') entry is

$$\frac{(d_{l1}d_{l'1} + \cdots + d_{lN}d_{l'N})\rho}{d_l d_{l'}}.$$

In matrix $\mathbf{W}$, the only unknown quantity is ρ. We estimate ρ through using the large sample properties of $\widehat{H}_l(t)$ under correlated data as described in the following.

We first sort the observed time $Y_{(1)} \le Y_{(2)} \le \ldots \le Y_{(d_l)}$ in the l-th group and obtain ordered data $[\{Y_{(1)}, \Delta_{(1)}, \mathbf{q}_{(1)} = \mathbf{u}_l\}, \ldots, \{Y_{(d_l)}, \Delta_{(d_l)}, \mathbf{q}_{(d_l)} = \mathbf{u}_l\}]$. Note that they are from the same group, and hence have the same $\mathbf{q}$ value. Following Breslow & Crowley (1974) [2], for correlated observations, KM estimator has the property

$$\sqrt{d_l}\{\widehat{H}_l(t) - H_l(t)\} = \frac{1}{\sqrt{d_l}}\sum_{i=1}^{d_l} a_i^l + O_p(\frac{1}{\sqrt{d_l}}),$$

where

$$\begin{aligned} a_i^l &= H_l(t)\Big[-\int_0^t \frac{I\{Y_{(i)} < u\}}{\{1-G^l(u)\}H_l(u)} dH_l(u) - \int_0^t G^l(u)\{1-G^l(u)\}^{-2} d\widetilde{G}^l(u) \\ &\quad - \int_0^t I\{Y_{(i)} < u, \Delta_{(i)} = 1\}\{1-G^l(u)\}^{-2} dG^l(u) \\ &\quad + \int_0^t \widetilde{G}^l(u)\{1-G^l(u)\}^{-2} dG^l(u) \\ &\quad + I\{Y_{(i)} < t, \Delta_{(i)} = 1\}\{1-G^l(t)\}^{-1} - \widetilde{G}^l(t)\{1-G^l(t)\}^{-1}\Big] \end{aligned}$$

is a function of $\{Y_{(i)}, \Delta_{(i)}, \mathbf{q}_{(i)} = \mathbf{u}_l\}$. Here $G^l(t)$ is the cumulative distribution functions of $Y_{(i)}$ and $\widetilde{G}^l(t) = \Pr\{Y_{(i)} \le t, \Delta_{(i)} = 1\}$.

Because some $\{Y_{(i)}, \Delta_{(i)}, \mathbf{q}_{(i)} = \mathbf{u}_l\}$'s are from a same family, there may exist an in-group correlation ρ between a_i^l and a_j^l in the l-th group. Furthermore, the calculation of $\mathrm{cov}\{\widehat{H}_l(t), \widehat{H}_{l'}(t)\}$ may also involve a between-group correlation between a_i^l and $a_j^{l'}$ if the observations are also from a same family. Since a common correlation ρ is assumed among family members, the between-group correlation is also ρ. We match ρ with the sample correlation from the in-group and between-groups pairs to obtain $\widehat{\rho}$, the detailed procedure contains three steps described in the following.

i. Approximate the integrals in a_i^l and re-organize the summation. For notational simplicity, we still write the result as a_i^l. This gives

$$a_i^l = \widehat{H}_l(t)(B_{(i)}^l - C^l - D_{(i)}^l + E_{(i)}^l),$$

where

$$\begin{aligned} B_{(i)}^l &= d_l \sum_{Y_{(j)} \le t} I\{Y_{(i)} \le Y_{(j)}\} \frac{\Delta_{(j)}}{(d_l - j)(d_l - j + 1)} \\ C^l &= d_l \sum_{Y_{(j)} \le t} \frac{\Delta_{(j)}}{(d_l - j)(d_l - j + 1)} \\ D_{(i)}^l &= d_l \sum_{Y_{(j)} \le t} I\{Y_{(i)} \le Y_{(j)}, \Delta_{(i)} = 1\} \frac{1}{(d_l - j)(d_l - j + 1)} \\ E_{(i)}^l &= I(Y_{(i)} < t, \Delta_{(i)} = 1) \left[1 - \frac{1}{d_l}\sum_{j=1}^{d_l} I\{Y_{(j)} < t\}\right]^{-1}. \end{aligned}$$

ii. Form pair $(a_i^l, a_j^{l'})$ if data $\{Y_{(i)}, \Delta_{(i)}, \mathbf{q}_{(i)} = \mathbf{u}_l\}$ and $\{Y_{(j)}, \Delta_{(j)}, \mathbf{q}_{(j)} = \mathbf{u}_{l'}\}$ are from a same family for $i = 1, \ldots, d_l$, $j = 1, \ldots, d_{l'}$ and $l, l' = 1, \ldots, m$. Denote the total number of pairs by v. We know that $v = \sum_{i=1}^N \binom{n_i}{2}$. Stack all the pairs to create a $v \times 2$ matrix $\mathbf{Z} = (Z_{ij})$.

iii. Calculate

$$\widehat{\rho} = \frac{\sum_{i=1}^{v}(Z_{i1} - \widehat{\mathbf{Z}}_{\cdot 1})(Z_{i2} - \widehat{\mathbf{Z}}_{\cdot 2})}{\sqrt{\sum_{i=1}^{v}(Z_{i1} - \widehat{\mathbf{Z}}_{\cdot 1})^2 \sum_{i=1}^{v}(Z_{i2} - \widehat{\mathbf{Z}}_{\cdot 2})^2}},$$

where $\widehat{\mathbf{Z}}_{\cdot j} = (\sum_{i=1}^{v} Z_{ij})/v$ for $j = 1, 2$.

Once $\mathbf{W}$ is obtained, we form estimator $\widehat{\mathbf{F}}(t)$ from (2.15). Standard calculation shows that it has a Gaussian limiting distribution with variance $(\mathbf{UWU}^{\mathrm{T}})^{-1}$. We denote this estimator as the correlated weighted least square (CWLS) estimator.

3. Simulation studies

We now demonstrate the finite sample performance of the BLRE, BQIF and CWLS methods via two simulation studies. The first simulation is a relatively simple one used to illustrate the effectiveness of the theoretical properties derived in Section 2. In the second simulation, we generated a more complex data, where we increased the number of event distributions, considered larger families and varied family sizes. For both simulations, we generated 1000 data sets.

In the first simulation, we set the sample size $N = 1000$, $p = 2$, $m = 5$ and $n_i = 4$ for $i = 1, \ldots, N$. The two ($p = 2$) true functions $F_1(t)$ and $F_2(t)$ are respectively the distribution functions of two truncated exponential densities, with rate 3 and 5 respectively. The support is $[0, 10]$. To generate correlated survival times for members from a same family, we implement the following procedure. For the ith family, we construct a multivariate distribution, that is, we generate a random vector $(S_{i1}^1, \ldots, S_{i4}^1, S_{i1}^2, \ldots, S_{i4}^2)$ from a Clayton copula with parameter 10. It provides the jth member of the ith family with possible survival time S_{ij}^1 or S_{ij}^2, corresponding to the two functions $F_1(t)$ and $F_2(t)$. We select $S_{ij} = S_{ij}^1$ or S_{ij}^2 with probabilities in the $\mathbf{q}_{ij}$ vector, where $\mathbf{q}_{ij}$ is assigned to five ($m = 5$) different vector values $(0.9, 0.1)^{\mathrm{T}}$,$(0.6, 0.4)^{\mathrm{T}}$,$(0.4, 0.6)^{\mathrm{T}}$, $(0.2, 0.8)^{\mathrm{T}}$ and $(0.15, 0.85)^{\mathrm{T}}$, with probabilities 0.3, 0.3, 0.2, 0.1 and 0.1 respectively. Lastly, we generate the censoring time from a uniform distribution on $(0, 5.4)$, resulting in a censoring rate of 50% approximately. We then create $Y_{ij} = \min(S_{ij}, C_{ij})$ and $\Delta_{ij} = I(S_{ij} \leq C_{ij})$.

We implement CWLS method in section 2.5, and use Algorithms 1 and 2 to carry out the BLRE and BQIF methods, and consider $R = 6$. We implement $B = 500$ bootstrap repetitions to estimate the variance-covariance matrices $\mathbf{U}$ in BLRE and $\mathbf{M}$ in BQIF respectively. We summarize the results of our analysis at $t = 2.5$ in Table 1. As a comparison, we also implemented the method from Ma and Wang (2014) [11], which we named "MW" method. From Table 1, it is clear that Algorithms 1 and 2 produce very similar results. This fact concurs with our theoretical results on the asymptotic equivalence of the two methods in Section 2.4. For all methods considered here, the mean of the 1000 estimates are very close to the true function values, the sample standard deviations of

TABLE 1

Simulation 1 summary statistics of the survival mixture distribution estimators at $t = 2.5$ in the MW (Assuming independence as in Ma and Wang (2014) [11]), CWLS, BLRE and BQIF models ($n = 1000$, $B = 500$). 'emp se' is the empirical standard error and 'est se' is the average of the estimated standard error.

	est mean	emp se	mse	est se	95% cov
		$F_1(t)$ (true=0.5863)			
MW	0.5860	0.0274	0.0015	0.0277	95.2%
CWLS	0.5863	0.0236	0.0011	0.0232	94.8%
BLRE	0.5863	0.0241	0.0011	0.0234	94.2%
BQIF	0.5863	0.0241	0.0011	0.0234	94.2%
		$F_2(t)$(true=0.4551)			
MW	0.4562	0.0339	0.0023	0.0337	94.0%
CWLS	0.4559	0.0281	0.0016	0.0276	93.5%
BLRE	0.4558	0.0289	0.0016	0.0281	93.4%
BQIF	0.4557	0.0289	0.0016	0.0280	93.7%

the 1000 estimates are very close to the average of the estimated standard deviations, and coverage rate of the 95% confidence interval is indeed close to the nominal value. BLRE, BQIF and CWLS reduce the mean squared error (MSE) by 27% for $F_1(t)$ and 30% for $F_2(t)$, indicating large improvement in estimation efficiency. The estimation result of the entire functions $F_1(t)$ and $F_2(t)$ is given in Figure 1, where the mean estimated curves almost overlap with the true curves. The lower and upper bound of the 95% confidence bands of $F_1(t)$ and $F_2(t)$ can be separated in our proposed methods, while they overlap with each other in the Ma and Wang (2014) [11] method.

In the second simulation, we set $N = 800$, $m = 5$, $p = 3$ and generate 50, 250,150,100,150,50 and 50 families with sizes $n_i = 8, 10, 12, 14, 15, 16$ and 18 respectively. We set $F_1(t)$, $F_2(t)$ and $F_3(t)$ as truncated exponential densities with rates 5, 9 and 11 on $[0, 10]$. As in simulation 1, we generate correlated survival time for members in a same family from a multivariate distribution. Specifically, for the ith family, we first generate a vector $(S^1_{i1},\ldots,S^1_{in_i},S^2_{i1},\ldots,S^2_{in_i},S^3_{i1},\ldots,S^3_{in_i})$

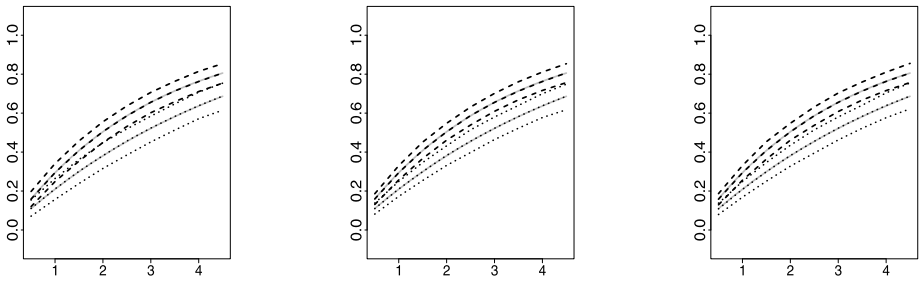

FIG 1. *Simulation study 1 on $F_1(t)$ and $F_2(t)$. True CDFs (solid grey) and mean ($F_1(t)$ dashed, $F_2(t)$ dotted), 95% confidence band ($F_1(t)$ dashed, $F_2(t)$ dotted) of the estimated CDFs. Left: Assuming independence as in Ma and Wang (2014) [11]; Middle: CWLS; Right: BLRE.*

TABLE 2

Simulation 2 summary statistics of the survival mixture distribution estimators at $t = 1.0$ in the MW (Assuming independence as in Ma and Wang (2014) [11]), CWLS, BLRE and BQIF models ($n = 800$, $B = 600$). 'emp se' is the empirical standard error and 'est se' is the average of the estimated standard error.

	est mean	emp se	mse	est se	95% cov
		$F_1(t)$ (true=0.2997)			
MW	0.2989	0.0291	0.001695	0.0291	94.2%
CWLS	0.2984	0.0206	0.000808	0.0196	93.4%
BLRE	0.2978	0.0248	0.001153	0.0231	93.0%
BQIF	0.2977	0.0250	0.001155	0.0229	92.5%
		$F_2(t)$(true=0.2289)			
MW	0.2276	0.0242	0.001337	0.0246	95.2%
CWLS	0.2278	0.0171	0.000733	0.0173	94.9%
BLRE	0.2266	0.0215	0.000888	0.0205	93.0%
BQIF	0.2265	0.0216	0.000891	0.0204	92.4%
		$F_3(t)$(true=0.2135)			
MW	0.2122	0.0228	0.001069	0.0220	93.4%
CWLS	0.2127	0.0166	0.000600	0.0163	93.4%
BLRE	0.2113	0.0204	0.000772	0.0188	92.6%
BQIF	0.2116	0.0203	0.000764	0.0187	92.6%

from a Clayton copula with parameter 20. The survival time S_{ij} of the jth member in the ith family is then assigned to S_{ij}^1, S_{ij}^2 or S_{ij}^3, corresponding to $F_1(t)$, $F_2(t)$ and $F_3(t)$ with probabilities q_{ij1}, q_{ij2} and q_{ij3}. Here $\mathbf{q}_{ij} = (q_{ij1}, q_{ij2}, q_{ij3})^{\mathrm{T}}$ is set to be $(1.00, 0.00, 0.00)^{\mathrm{T}}$, $(0.60, 0.40, 0.00)^{\mathrm{T}}$, $(0.0, 0.20, 0.80)^{\mathrm{T}}$, $(0.20, 0.00, 0.80)^{\mathrm{T}}$ and $(0.30, 0.70, 0.00)^{\mathrm{T}}$. The mixing percentages are 15.92%, 15.92%, 26.37%, 20.90% and 20.90%. We generate censoring time from a uniform distribution on $(0, 5.4)$, resulting in a censoring rate around 37%. Finally we let $Y_{ij} = \min(S_{ij}, C_{ij})$ and $\Delta_{ij} = I(S_{ij} \le C_{ij})$.

We implement Algorithms 1 (BLRE) and 2 (BQIF) with $R = 15$, CWLS method and the method from Ma and Wang (2014) [11]. We use $B = 600$ bootstraps to estimate $\mathbf{U}$ and $\mathbf{M}$. The simulation results at $t = 1.5$ are summarized in Table 2. From the results in Table 2, we find that Algorithms 1 and 2 produce similar results. It again validates our theoretical discovery in Section 2.4, regardless of which distribution and correlation structure we use to generate the data. The mean of the 1000 estimates are fairly close to the true values. The average of the 1000 estimated standard errors are also close to the sample standard errors of the 1000 estimates. The coverage rates of the 95% confidence intervals are close to the nominal level. Compared with the method in Ma and Wang (2014) [11], the CWLS method reduces the MSE by 52.3%, 45.2% and 43.9% for $F_1(t)$, $F_2(t)$ and $F_3(t)$ respectively, while BLRE and BQIF reduce the MSE by 32.0% for $F_1(t)$, 33.6% for $F_2(t)$ and 27.8% for $F_3(t)$, indicating quite large improvement. The estimation results of the entire functions $F_1(t)$, $F_2(t)$ and $F_3(t)$ are given in Figure 2. The mean curves of 1000 simulations and the true distribution curves of $F_1(t)$, $F_2(t)$ and $F_3(t)$ overlay. The 95% confidence bands of $F_1(t)$(red) and $F_2(t)$(blue) in the left panel have an overlapped area, while in middle and right panels, they are better separated. In addition, there

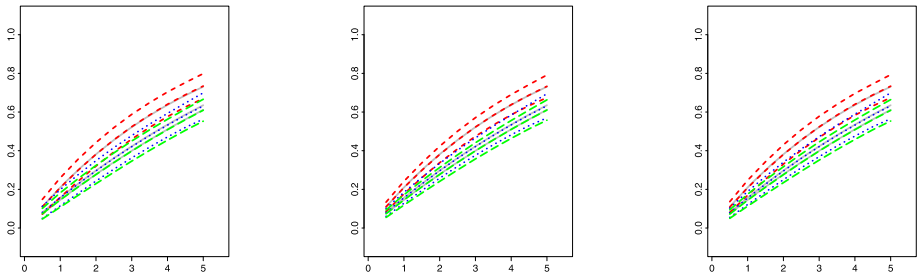

FIG 2. *Simulation study 2. Curves of $F_1(t)$(red), $F_2(t)$(blue) and $F_3(t)$(green) are displayed. True CDFs (solid grey) and mean ($F_1(t)$ dashed, $F_2(t)$ dotted, $F_3(t)$ long-dashed), 95% confidence band ($F_1(t)$ dashed, $F_2(t)$ dotted, $F_3(t)$ long-dashed) of the estimated CDFs. Left: Assuming independence as in Ma and Wang (2014) [11]; Middle: CWLS; Right: BLRE.*

is a gap between 95% confidence bands of $F_1(t)$(red) and $F_3(t)$(green) in the middle and right panels, but not in the left panel.

4. Real data example

In the Cooperative Huntington's Observational Research Trial (COHORT, Dorsey and the Huntington Study Group COHORT Investigators, 2012 [3]), initial participants (probands) were sequenced for huntingtin gene expansion status: subjects with C-A-G repeats length greater than 36 are the HD mutation carriers and will eventually develop HD. The proband participants also provided their relatives' phenotype information such as age-at-death if deceased. However, the relatives were not genotyped due to practical difficulties in collecting blood samples (especially for deceased individuals). It is of interest to estimate the distribution functions using the relatives phenotypes only, while avoid potential ascertainment bias when recruiting proband participants. Comparing survival functions of gene-expanded and non-expanded subjects is essential for understanding the disease risk associated with a causal mutation, for timing intervention in the disease progression course, and for genetic counseling.

The COHORT data includes 771 families with different numbers of first-degree relatives within each family. There are a total of 3661 individuals. The barplot in Figure 3 characterizes the distribution of the family sizes. Using the available relationship (parents, children, siblings) between each family member and his/her proband, we calculated the probability of the family member carrying the huntingtin gene mutation. We obtained three ($m = 3$) different $\mathbf{q}_{ij}$ values in total, $(1.0, 0.0)^{\mathrm{T}}, (0.5, 0.5)^{\mathrm{T}}$ and $(0.0, 1.0)^{\mathrm{T}}$, with frequency 558, 1805 and 1298, respectively. Denote the distribution function of age-at-death in mutation carrier population to be $F_1(t)$ and in non-carrier group to be $F_2(t)$. Our goal is to estimate $\mathbf{F}(t) = \{F_1(t), F_2(t)\}^{\mathrm{T}}$. The COHORT data has approximately 29% censoring, and we assume the censoring time is independent of the event time.

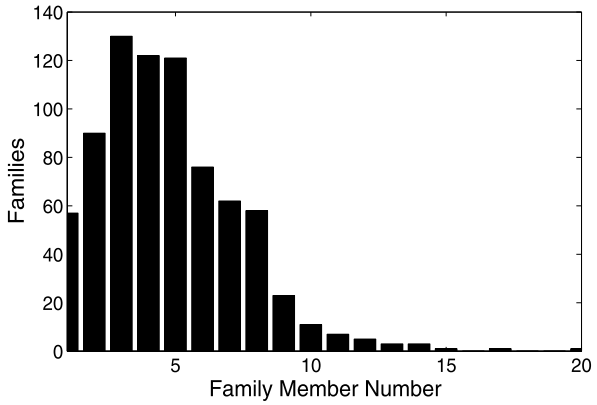

FIG 3. *Barplot of the size of families in COHORT. The highest percentage* 16.86% *happens when* $n_i = 3$. *The largest family has* $n_i = 20$ *members with the smallest percentage* 0.13%.

We implemented Algorithm 1 (BLRE), Algorithm 2 (BQIF), and CWLS developed in section 2. We performed $B = 500$ bootstraps to estimate the variance-covariance $\mathbf{U}$ in BLRE and $\mathbf{M}$ in BQIF. The results corresponding to $R = 16$ are reported. The distribution curve based on four methods at life span $0 - 90$ are depicted in Figure 4. The estimated $F_1(t)$ and $F_2(t)$, and their 95% confidence bands are provided. It is clear that the huntingtin gene mutation carriers have a much lower survival rates than non-carriers, especially in the age range 50 to 90. This indicates that the detrimental effect of the Huntington's disease on survival is most severe in the mid- to old age range. The difference in survival probability starts to be present as early as age 40, which is the same age as the mean of HD disease onset age (Foroud et al. 1999 [5]). The result is also consistent clinical observation that majority of gene carriers die between age 45 and 70 (Foroud et al. 1999 [5]).

As a comparison, we also performed the analysis of Ma and Wang (2014) [11], where the within-family correlation is ignored. We present the estimated curves of $F_1(t)$ and $F_2(t)$, and their 95% pointwise confidence bands in the upper left panel of Figure 4. From these plots, we can see that for estimating the distribution function in the carrier population, the confidence band of MW appears wider than all other three methods. To quantify this observation, we calculated the integrated confidence interval width and obtained the values 0.0577, 0.04, 0.0544 and 0.0545 for MW, CWLS, BLRS and BQIF, respectively. This corresponds to a reduction of 30.6%, 5.6% and 5.5% of the three methods in comparison with MW, indicating improved efficiency of the three proposed methods. Our methods here also provide an assessment of the familial correlation, which is estimated to be smaller than 0.1 across all ages and has a general decreasing trend. Specifically, the correlation is 0.0822 at age $t = 40$, decreases to 0.046 by age 65, and almost diminishes beyond age 70.

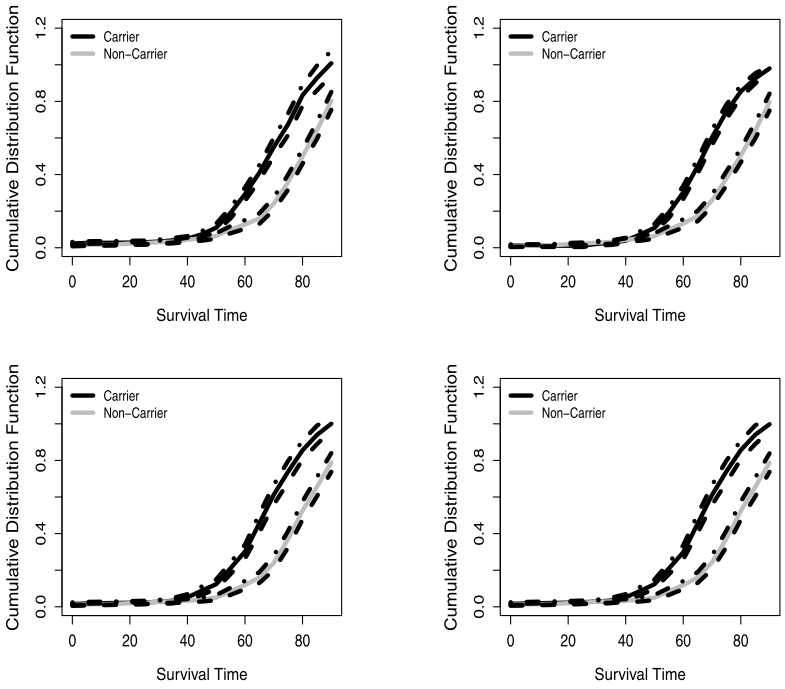

FIG 4. *Distribution of the age-at-death for huntingtin gene mutation carriers and non-carriers in COHORT: estimated CDFs (solid) and the* 95% *confidence band (upper band dot-dashed, lower band dashed). Upper Left: Ignore within-family correlation (MW; Assuming independence as in Ma and Wang (2014) [11]); Upper Right: Treat within-family correlation (CWLS); Lower left: BLRE.*

5. Discussion

In this paper, we propose various methods to account for within-family correlation for mixture data from multiple populations, while the population label is only known up to a probability. Such data arise from kin-cohort studies, such as COHORT study of HD and other disorders, for instance, breast cancer (Wacholder et al. 1998 [12]) or Parkinson's disease (Wang et al. 2008 [13]). The proposed estimators are easy to implement, and assume unstructured correlation or exchangeable correlation. The finite sample efficiency of the estimators under unstructured correlation relies on both the number of resamples and the bootstrap size. Although in theory, large values of both are preferable, in practice, one can always gradually increase these values and stop when the improvement becomes sufficiently small. Comparing with ignoring correlation information, the proposed estimators provide efficiency gain as observed in simulation studies and real data analysis (Section 3 and 4).

The relative performance of the proposed BLRE, BQIF and CWLS methods in terms of variance reduction depends on many factors such as sample size, family sizes, true underlying model and covariate values. The practical ef-

ficiency gain of accounting for the correlation by the optimal estimator could be somewhat limited in a real data example especially when the sample size of the data is large. However, other studies with smaller sample sizes and lower expected number of diseased subjects in carrier group could potentially benefit more from our method by improving estimation efficiency.

Here, p denotes the number of latent distributions and it is typically small in practice. In our application example, the goal is to estimate the survival function in HD-gene mutation carriers, as compared to non-carriers. For HD, which is an autosomal dominant disease, $p = 2$. In general, mode of inheritance is specified as autosomal dominant ($p = 2$), autosomal recessive ($p = 2$), or additive model ($p = 3$), depending on the nature of an inherited disorder. Thus, p is often rather small in these applications. However, when p becomes large in other settings, the proposed model and methods are not suitable because not enough information is available in the data to estimate many latent distributions without stronger parametric assumptions. Parametric or semi-parametric models and methods are needed to enable estimation under reasonable sample size.

Lastly, we point out that one way to relax the exchangeable correlation assumption is to break large extended families into nuclear families and assume a hierarchical model for the correlation structure for the members in the nuclear family and in the extended family.

Appendix

A.1. Proof of Theorem 1

Write $\mathbf{A}$ as $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_R)$. Under the constraint $\mathbf{AJ} = \mathbf{I}_p$, we have

$$\begin{aligned} E\{\widehat{\mathbf{F}}(t)\} &= E\{\mathbf{A}\widehat{\mathbf{F}}_L(t)\} = \sum_{r=1}^{R} \mathbf{A}_r E\{\widehat{\mathbf{F}}^r(t)\} \\ &= \sum_{r=1}^{R} \mathbf{A}_r \mathbf{F}(t) + o_p(1) = \mathbf{AJF}(t) + o_p(1) = \mathbf{F}(t) + o_p(1). \end{aligned}$$

This shows that $\widehat{\mathbf{F}}(t)$ is a consistent estimator. In addition, because $\widehat{\mathbf{F}}_L(t)$ has normal distribution (Ma and Wang 2012 [10]), as linear combination, $\widehat{\mathbf{F}}(t)$ is also normally distributed.

The variance of $\widehat{\mathbf{F}}(t) = \mathbf{A}\widehat{\mathbf{F}}_L(t)$ is $\mathbf{AUA}^{\mathrm{T}}$ for a general $\mathbf{A}$ matrix. For any $\mathbf{A}$ that satisfies $\mathbf{AJ} = \mathbf{I}_p$, we have

$$\begin{aligned} & \mathbf{AUA}^{\mathrm{T}} - \mathbf{A}_{\mathrm{opt}}\mathbf{U}\mathbf{A}_{\mathrm{opt}}^{\mathrm{T}} \\ = \; & \mathbf{AUA}^{\mathrm{T}} - (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1} \\ = \; & (\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AUA}^{\mathrm{T}}\mathbf{AJ}(\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ})^{-1} - (\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1} \\ = \; & (\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ})^{-1}\left\{\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AUA}^{\mathrm{T}}\mathbf{AJ} - \mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ}(\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AJ}\right\} \end{aligned}$$

$$
\begin{aligned}
& (\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{J})^{-1} \\
= \; & (\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{U}^{\frac{1}{2}}\left\{I-\mathbf{U}^{-\frac{1}{2}}\mathbf{J}(\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-\frac{1}{2}}\right\} \\
& \mathbf{U}^{\frac{1}{2}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{J}(\mathbf{J}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{J})^{-1}.
\end{aligned}
$$

It is easy to verify that $I-\mathbf{U}^{-\frac{1}{2}}\mathbf{J}(\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-\frac{1}{2}}$ is an idempotent matrix, hence it is semi-positive definite. Therefore, $\mathbf{A}\mathbf{U}\mathbf{A}^{\mathrm{T}}-(\mathbf{J}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{J})^{-1}$ is also semi-positive definite.

A.2. Proof of Theorem 2

Taking derivative of the quadratic form (2.9) with respect to $\mathbf{F}(t)$ and omit the higher order terms, we obtain

$$
E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}}\mathbf{W}\sum_{i=1}^{N}\mathbf{g}_i(t)=\mathbf{0}. \tag{A.1}
$$

In the following, we first investigate $N^{-\frac{1}{2}}\sum_{i=1}^{N}\mathbf{g}_i(t)$.

For $r=1,\ldots,R$, write the observations in the rth sample as $\{O_i^r:\ O_i^r=(\mathbf{q}_i^r,Y_i^r,\Delta_i^r), i=1,\ldots,N\}$. Because there are m possible values for $\mathbf{q}_i^r$'s, we can divide these N observations into m groups $\mathbf{O}_1^r,\ldots\mathbf{O}_m^r$, where

$$
\mathbf{O}_l^r=\left\{O_{l,k}^r:\ O_{l,k}^r=(\mathbf{u}_l,Y_{l,k}^r,\Delta_{l,k}^r),\ k=1,\ldots,d_l^r\right\},
$$

and the Kaplan-Meier estimator in the respective group is denoted $\widehat{H}_l^r(t)$ for $l=1,\ldots,m$.

From Breslow and Crowley (1974) [2], we have the asymptotic expansion

$$
\sqrt{d_l^r}\{\widehat{H}_l^r(t)-H_l(t)\}=(d_l^r)^{-1/2}\sum_{k=1}^{d_l^r}a(O_{l,k}^r)+o_p(1),
$$

where $a(O_{l,k}^r)$ is a function of the kth observation $O_{l,k}^r$ and $E\{a(O_{l,k}^r)\}=0$. Inserting this relation into (2.8), we have

$$
\begin{aligned}
N^{-\frac{1}{2}}\sum_{i=1}^{N}\mathbf{g}_i(t) &= N^{-\frac{1}{2}}\sum_{l=1}^{m}\begin{bmatrix}\sqrt{d_l^1}\mathbf{u}_l\sqrt{d_l^1}\{\widehat{H}_l^1(t)-H_l(t)\}\\ \vdots \\ \sqrt{d_l^R}\mathbf{u}_l\sqrt{d_l^R}\{\widehat{H}_l^R(t)-H_l(t)\}\end{bmatrix} \\
&= N^{-\frac{1}{2}}\sum_{l=1}^{m}\begin{bmatrix}\mathbf{u}_l\sum_{k=1}^{d_l^1}a(O_{l,k}^1)\\ \vdots \\ \mathbf{u}_l\sum_{k=1}^{d_l^R}a(O_{l,k}^R)\end{bmatrix}+o_p(1) \\
&= N^{-\frac{1}{2}}\sum_{i=1}^{N}\begin{bmatrix}\mathbf{q}_i^1 a(O_i^1)\\ \vdots \\ \mathbf{q}_i^R a(O_i^R)\end{bmatrix}+o_p(1),
\end{aligned} \tag{A.2}
$$

where the first equality is obtained through rewriting the summation in (2.8), and the last equality is obtained similarly. Viewing $\mathbf{q}_i^r$'s as random quantities, we have that $N^{-\frac{1}{2}} \sum_{i=1}^{N} \mathbf{g}_i(t)$ is the average of independently identically distributed mean zero random quantities hence it converges to a mean zero normal distribution with variance denoted $\mathbf{M}$.

Standard Taylor expansion of (A.1) then yields

$$\sqrt{N}\{\widehat{\mathbf{F}}(t) - \mathbf{F}(t)\} \to N\{\mathbf{0}, \mathbf{B}^{-1}\mathbf{C}(\mathbf{B}^{-1})^{\mathrm{T}}\}$$

in distribution when $N \to \infty$, where

$$\begin{aligned} \mathbf{B} &= E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}} \mathbf{W} E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}, \\ \mathbf{C} &= \left[E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}} \mathbf{W}\mathbf{M}\mathbf{W} E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}\right]. \end{aligned}$$

Similar derivation as in the proof of Theorem 1 can be used to show that the optimal choice of the weight matrix is $\mathbf{W} = \mathbf{M}^{-1}$, and the resulting variance is

$$\left[E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}^{\mathrm{T}} \mathbf{M}^{-1} E\left\{\frac{\partial \mathbf{g}_i(t)}{\partial \mathbf{F}^{\mathrm{T}}(t)}\right\}\right]^{-1}.$$

Acknowledgements

The project was supported by NSF grant DMS-1608540 and NIH grants NS073671, NS082062. The first author thanks Yongtao Guan for support, encouragement and for providing computational facilities.

References

[1] Akritas, M. G. (1986). Bootstrapping the kaplan-meier estimator. *Journal of the American Statistical Association* **81** 1032–1038. MR0867628

[2] Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics* **2** 437–453. MR0458674

[3] Dorsey, E. and the Huntington Study Group COHORT Investigators (2012). Characterization of a large group of individuals with huntington disease and their relatives enrolled in the cohort study. *PLoS ONE* **7** 429–522.

[4] Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association* **76** 312–319. MR0624333

[5] Foroud, T., Gray, J., Ivashina, J., and Conneally, P. (1999). Differences in duration of huntingtons disease based on age at onse. *Journal of Neurological Neurosurg Psychiatry* **66** 52–56.

[6] Huntington Study Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on huntingtons disease chromosomes. *Cell* **72** 971–983.

[7] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53** 457–481. MR0093867

[8] Langbehn, D.R. and Brinkman, R.R. and Falush, D. and Paulsen, J.S. and Hayden, M.R. and International Huntington's Disease Collaborative Group (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clinical Genetics* **65** 267–77.

[9] Lindsay, B. G. and Qu, A. (2003). Inference functions and quadratic score tests. *Statistical Science* **18** 394–410. MR2061916

[10] Ma, Y. and Wang, Y. (2012). Efficient semiparametric estimation for mixture data. *Electronic Journal of Statistics* **6** 710–737.

[11] Ma, Y. and Wang, Y. (2014). Estimating disease onset distribution functions in mutation carriers with censored mixture data. *Journal of the Royal Statistical Society, Series C* **63** 1–23. MR3148266

[12] Wacholder, S., Hartge, P., Struewing, J. P., Pee, D., McAdams, M., Brody, L., and Tucker, M. (1998). The kin-cohort study for estimating penetrance. *American Journal of Epidemiology* **148** 623–630.

[13] 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 disease in carriers of parkin mutations: estimation using the kin-cohort method. *Archvies of Neurology* **65** 467–474.