Statistica Sinica Preprint No: SS-2016-0199R2	
Title	Predicting disease Risk by Transformation Models in the Presence of Missing Subgroup Identifiers
Manuscript ID	SS-2016-0199R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0199
Complete List of Authors	Yuanjia Wang Qianqian Wang and Yanyuan Ma
Corresponding Author	Yuanjia Wang
E-mail	yw2016@cumc.columbia.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

Predicting disease Risk by Transformation Models in the Presence of Unspecified Subgroup Membership

Qianqian Wang, Yanyuan Ma and Yuanjia Wang

University of South Carolina, Penn State University and Columbia University

Abstract

Some biomedical studies lead to mixture data. When a subgroup membership is missing for some of the subjects in a study, the distribution of the outcome is a mixture of the subgroup-specific distributions. Taking into account the uncertain distribution of the group membership and the covariates, we model the relation between the disease onset time and the covariates through transformation models in each sub-population, and develop a nonparametric maximum likelihood-based estimation implemented through the EM algorithm along with its inference procedure. We propose methods to identify the covariates that have different effects or common effects in distinct populations, which enables parsimonious modeling and better understanding of the differences across populations. The methods are illustrated through extensive simulation studies and a data example.

Key words: Censored data, EM algorithm, Laplace transformation, mixed populations, uncertain population identifier, semiparametric models, transformation models

1 Introduction

Biomedical studies can lead to mixture data. When a discrete covariate defining subgroup membership is missing for some of the subjects in a study, the distribution of the outcome is a mixture of the subgroup-specific distributions. One example is the kin-cohort study Wacholder et al. (1998) with the goal of estimating the cumulative risk of disease for mutation carriers Khoury et al. (1993). However, mutation status is only collected in the initial

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

sample of participants, referred as probands, not in their relatives. For example, genetic mutation status is not available for deceased relatives or those who have not undergone genetic testing due to resource constraints. The disease phenotype information for such relatives is available from other sources, such as interviewing the proband in a family Marder et al. (2003). For a late-onset disease, such as Parkinson's disease (PD), parents of study participants are often deceased. Therefore even though age-at-onset of PD is provided by a family member, no genotyping can be performed on deceased parents. When estimating the disease risk distribution for mutation carriers and non-carriers using these relatives' disease onset information, the unknown mutation status needs to be accounted for by using the distribution of mutation status in such relatives as estimated from living relatives who provide blood sample Wang et al. (2012), Ma and Wang (2014).

We consider estimating the subgroup-specific distribution for outcomes that are subject to censoring and with missing subgroup identifiers. The nonparametric models in Wacholder et al. (1998), Wang et al. (2012), and Ma and Wang (2014) do not include any covariates other than the mutation status. We consider how to include covariates that can have identical or different effects across subgroups. Popular semiparametric models for censored outcomes, such as the Cox proportional hazards model, accelerated failure time model, and transformation model have been studied extensively in the literature, but less so in a mixture data setting. Recently, Altstein and Li (2013) proposed a latent subgroup analysis for a semiparametric accelerated failure time model in a clinical trials setting. Our work differs from Altstein and Li (2013) in that the distribution of the subgroup identifiers is available in our problem, and we assume a semiparametric transformation model in each subgroup. A transformation model is applied to analyze neurological disorder data (e.g, Huntington's disease [HD] as in our motivating study) due to its useful biological and clinical interpretations; see for example Zhang et al. (2012).

We propose a semiparametric transformation model for mixture data. Compared to parametric transformation model in the literature Zhang et al. (2012), we allow for greater

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

flexibility to account for subgroup heterogeneity. This is achieved in our model through characterizing the outcome in each subpopulation using a different distribution, indexed by both parameters and error distributions. They can also have both as shared covariate effect and/or a subgroup-specific covariate effect. In addition, we assume an unknown transformation to avoid the difficulty of specifying a parametric transformation. When assuming a homogeneous covariate effect, we account for a missing population identifier by taking advantage of the distribution of the mixing proportion and using a weighted least-square type estimator, which greatly simplifies the procedure. When we assume a subgroup-specific covariate effect, the weighted least-square estimator no longer applies, and we use the EM algorithm. We have performed extensive simulation studies to examine performance of the proposed approach and applied it to estimating the survival function for HD mutation carriers in a large genetic epidemiology study Dorsey and The Huntington Study Group COHORT Investigators (2012).

2 Modeling, Estimation, and Asymptotic Properties

Assume there are n observations from p populations. Here p is usually determined by the research purpose. For genetic studies, populations are defined by mutation carrier status. Throughout, we assume p is pre-determined. Denote the data from the ith observation as $\mathbf{O}_i = (\mathbf{q}_i, \mathbf{x}_i, \mathbf{z}_i, y_i, \delta_i)$, where $\mathbf{q}_i$ is a length p vector, with the jth entry q_{ij} being the probability that the ith observation is randomly sampled from the jth population. We also allow a subject's population membership to be known by allowing $\mathbf{q}_i$ to be a vector with 1 in one component and zero in all others. Let t_i be the time to event and c_i be the censoring time, $y_i = \min(t_i, c_i)$, and $\delta_i = I(t_i \leq c_i)$. Let $\mathbf{x}_i$ denote the covariate vector that has a common effect on the event time across different populations, while $\mathbf{z}_i$ denotes the covariate vector that has a different effect in different populations. For simplicity, we sort the data so that $y_i \leq y_k$ for all $i < k$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

2.1 Model

For the jth population, the linear transformation model we propose has the form

$$H(T) = -\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\alpha}_j + \epsilon_j. \tag{1}$$

Here H is an unknown, monotonically increasing function and, without loss of generality, we assume $H(0) = -\infty$. We assume ϵ_j is independent of $\mathbf{X}$, $\mathbf{Z}$, and has a known population-specific distribution $f_j(\epsilon_j)$. Here, in each population, this is a classical linear transformation model, in which the baseline population distribution can be heterogeneous due to the different choices of f_j. Selection of f_j for each population can be based on scientific or biological knowledge of a particular population. The covariate effect is also allowed to vary, reflected in the population-specific $\boldsymbol{\alpha}_j$. By including the term $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, we also allow the possibility that some covariates have a homogeneous effect across populations. We develop a test to assess whether a covariate exhibits evidence of deviation from a homogeneous effect model.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_p^{\mathrm{T}})^{\mathrm{T}}$, $\Phi(t) = \exp\{H(t)\}$, and $\phi(t) = \exp\{H(t)\}h(t)$. The conditional distribution function of the ith relative from (1) is then

$$\begin{aligned} & f(y_i, \delta_i \mid \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta}, \Phi, \phi) \\ = & \left[h(y_i)\sum_{j=1}^{p} q_{ij} f_j\{H(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j\}\right]^{\delta_i} \left[1 - \sum_{j=1}^{p} q_{ij} F_j\{H(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j\}\right]^{1-\delta_i} \\ = & \phi(y_i)^{\delta_i}\Psi(O_i; \boldsymbol{\theta}, \Phi), \end{aligned}$$

where Φ is a function that depends only on $\boldsymbol{\theta}$ and Φ, but not on ϕ. The model can not be viewed as a transformation model, hence existing estimation procedures do not apply. To ensure identifiability, we require that the $\mathbf{q}_i$ variable takes m different vector values, denoted $\mathbf{u}_1, \ldots, \mathbf{u}_m$, so that the matrix $(\mathbf{u}_1, \ldots, \mathbf{u}_m)$ has rank p. We point out that the identifiability here excludes any permutation. This identifiability is stronger than that up to a permutation in most classical mixture models Holzmann et al. (2006). We can achieve the stronger form of identifiability because the mixture probabilities, while different for different observations, are known.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

2.2 Estimation

We propose a nonparametric maximum likelihood estimator (NPMLE) to estimate $\boldsymbol{\theta}$ and $\Phi(\cdot)$. Specifically, we obtain $\widehat{\boldsymbol{\theta}}$ and $\widehat{H} = \log(\widehat{\Phi})$ through maximizing

$$l(\boldsymbol{\theta}, \Phi) = \sum_{i=1}^{n} \delta_i \log\{\phi(y_i)\} + \sum_{i=1}^{n} \log\{\Psi(O_i; \boldsymbol{\theta}, \Phi)\}$$

with respect to $\boldsymbol{\theta}$ and Φ, where we restrict Φ, hence H, to be a piecewise constant non-decreasing function with non-negative jumps only at the observed event times. Following existing literature Wacholder et al. (1998), Wang et al. (2012), We exclude the probands from the analysis sample and the likelihood to protect against potential ascertainment bias from unknown sources that may be difficult to adjust (e.g., convenience sample of patients visiting a clinic). Given the mutation carrier status, we also assume the relatives' phenotypes are conditionally independent of probands' phenotypes, which is an assumption satisfied by a monogenic disorder with a known genetic cause controlled in the model (e.g., HD in our application).

Although conceptually simple, the computation of NPMLE is not straightforward because the maximization is with respect to not only $\boldsymbol{\gamma}$, but also $\Phi(\cdot)$ at all the y_i's that are not censored. As sample size increases, the potential number of parameters increases as well, hence the computational problem does not simplify in the asymptotic sense. To overcome the computational difficulty, we use an EM algorithm. To this end, we first use Laplace transformation in each population to obtain

$$1 - F_j(x) = \int_0^\infty \exp(-r_j e^x)\psi_j(r_j)dr_j,$$

where $\psi_j(\cdot)$ is the inverse Laplace transformation of $1-F_j(x)$ as a function of e^x, consequently

$$\begin{aligned} 1 - \sum_{j=1}^{p} q_{ij} F_j\{H(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j\} &= \sum_{j=1}^{p} q_{ij} \int_0^\infty \exp\{-r_{ij} e^{H(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}\psi_j(r_{ij})dr_{ij} \\ &= \sum_{j=1}^{p} q_{ij} \int_0^\infty \exp\{-r_{ij}\Phi(y_i) e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}\psi_j(r_{ij})dr_{ij} \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

and

$$\begin{aligned} & h(y_i)\sum_{j=1}^{p} q_{ij} f_j\{H(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j\} \\ = & \sum_{j=1}^{p} q_{ij}\int_0^{\infty}\exp\{-r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}\phi(y_i)\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j)r_{ij}\psi_j(r_{ij})dr_{ij}. \end{aligned}$$

The ith observation here is $\mathbf{O}_i$, let $\mathbf{D}=(\mathbf{O}_1,\cdots,\mathbf{O}_n)$. Let $0<t_1<\cdots<t_K<\tau$ be the distinct event times, and write the quantities to be estimated as $\boldsymbol{\gamma}=\{\boldsymbol{\theta}^{\mathrm{T}},H(t_1),\ldots H(t_K)\}^{\mathrm{T}}$. The log-likelihood is then $l(\boldsymbol{\gamma};\mathbf{D})=\sum_{i=1}^{n} l_i(\boldsymbol{\gamma};\mathbf{O}_i)$, where

$$l_i(\boldsymbol{\gamma};\mathbf{O}_i)=\log\sum_{j=1}^{p}\int_0^{\infty}\{\phi(y_i)r_{ij}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j)\}^{\delta_i}\exp\{-r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}q_{ij}\psi_j(r_{ij})dr_{ij}.$$

We take advantage of this special data structure and view the population identifiers $\mathbf{G}=(G_1,\ldots,G_n)$ and $\mathbf{r}=(\mathbf{r}_1,\ldots,\mathbf{r}_n)$ as the missing variable, where $G_i=I_j$ represents that the ith observation is a random sample from the jth population, and $\mathbf{r}_i=(r_{i1},\ldots,r_{ip})^{\mathrm{T}}$ is the introduced random effects to facilitate computation. Then the complete data loglikelihood is $l(\boldsymbol{\gamma}\mid\mathbf{D},\mathbf{G},\mathbf{r})=\sum_{i=1}^{n} l_i(\boldsymbol{\gamma}\mid O_i,G_i,\mathbf{r}_i)$, where

$$\begin{aligned} l_i(\boldsymbol{\gamma}\mid O_i,G_i=I_j,r_{ij}) &= \log\left[\{\phi(y_i)r_{ij}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j)\}^{\delta_i}\exp\{-r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}\right] \\ &= \delta_i\log\{\phi(y_i)r_{ij}\}+\delta_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j)-r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}. \end{aligned}$$

This is a Cox model log-likelihood. Thus, in the E-step, we calculate

$$Q(\boldsymbol{\gamma},\boldsymbol{\gamma}^{(u)},\mathbf{D})\equiv E_{\boldsymbol{\gamma}^{(u)}}\{l(\boldsymbol{\gamma}\mid\mathbf{D},\mathbf{G},\mathbf{r})\mid\mathbf{D}\}=\sum_{i=1}^{n}\frac{\int\sum_{j=1}^{p} l_i(\boldsymbol{\gamma}\mid O_i,\mathbf{G}_i=I_j,r_{ij})a_{ij}^{(u)}dr_{ij}}{\int\sum_{j=1}^{p} a_{ij}^{(u)}dr_{ij}},$$

where

$$a_{ij}^{(u)}=\{\phi^{(u)}(y_i)r_{ij}\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)})\}^{\delta_i}\exp\{-r_{ij}\Phi^{(u)}(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}}\}q_{ij}\psi_j(r_{ij}).$$

In the M-step, we maximize $Q(\boldsymbol{\gamma},\boldsymbol{\gamma}^{(u)},\mathbf{D})$ with respect to $\boldsymbol{\gamma}$ subject to the constraints $0<H(t_1)<\cdots<H(t_K)\le 1$ to obtain $\boldsymbol{\gamma}^{(u+1)}$. Specifically, taking derivative with respect to $\boldsymbol{\gamma}$, we obtain estimating equations

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{n}\frac{\int\sum_{j=1}^{p}\{\delta_i\mathbf{x}_i-\mathbf{x}_i r_{ij}\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\}a_{ij}^{(u)}dr_{ij}}{\int\sum_{j=1}^{p} a_{ij}^{(u)}dr_{ij}} \\ &= \sum_{i=1}^{n}\delta_i\mathbf{x}_i-\mathbf{x}_i\Phi(y_i)e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\frac{\sum_{j=1}^{p} e^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}\int r_{ij}a_{ij}^{(u)}dr_{ij}}{\int\sum_{j=1}^{p} a_{ij}^{(u)}dr_{ij}}. \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

For $j = 1, \ldots, p$,

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{n} \frac{\int (\delta_i \mathbf{z}_i - \mathbf{z}_i r_{ij} e^{H(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j}) a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{p} a_{ij}^{(u)} dr_{ij}} \\ &= \sum_{i=1}^{n} \frac{\delta_i \mathbf{z}_i \int a_{ij}^{(u)} dr_{ij} - \mathbf{z}_i \Phi(y_i) e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{p} a_{ij}^{(u)} dr_{ij}}. \end{aligned}$$

For $k = 1, \ldots, K$,

$$\begin{aligned} 0 &= \sum_{y_i \geq t_k} \frac{\int \sum_{j=1}^{p} \left\{ \frac{I(y_i = t_k)}{\phi_k} - r_{ij} e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j} \right\} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{p} a_{ij}^{(u)} dr_{ij}} \\ &= \frac{1}{\phi_k} - \sum_{y_i \geq t_k} \frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \sum_{j=1}^{p} e^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{p} a_{ij}^{(u)} dr_{ij}}. \end{aligned}$$

This yields

$$\phi_k = \left(\sum_{y_i \geq t_k} \frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \sum_{j=1}^{p} e^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{p} a_{ij}^{(u)} dr_{ij}} \right)^{-1},$$

or in general

$$\begin{aligned} \phi(y_k; \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \delta_k \left(\sum_{i=1}^{n} \frac{I(y_i \geq y_k) e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \sum_{j=1}^{p} e^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{p} a_{ij}^{(u)} dr_{ij}} \right)^{-1} \qquad (2) \\ \Phi(y_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{k=1}^{n} I(y_k \leq y_i) \delta_k \left(\sum_{i=1}^{n} \frac{I(y_i \geq y_k) e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \sum_{j=1}^{p} e^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{p} a_{ij}^{(u)} dr_{ij}} \right)^{-1}. \end{aligned}$$

Plugging into the estimating equation for $\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$, we obtain

$$\begin{aligned} \sum_{i=1}^{n} \delta_i \mathbf{x}_i - \mathbf{x}_i \Phi(y_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}} \frac{\sum_{j=1}^{p} e^{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{p} a_{ij}^{(u)} dr_{ij}} &= \mathbf{0} \qquad (3) \\ \sum_{i=1}^{n} \frac{\delta_i \mathbf{z}_i \int a_{ij}^{(u)} dr_{ij} - \mathbf{z}_i \Phi(y_i; \boldsymbol{\beta}, \boldsymbol{\alpha}) e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j} \int r_{ij} a_{ij}^{(u)} dr_{ij}}{\int \sum_{j=1}^{p} a_{ij}^{(u)} dr_{ij}} &= \mathbf{0} \end{aligned}$$

at $j = 1, \ldots, p$.

We solve the estimating equations (3) to obtain $\widehat{\boldsymbol{\beta}}^{(u+1)}, \widehat{\boldsymbol{\alpha}}^{(u+1)}, j = 1, \ldots, p$, and then substitute into (2) to obtain $\Phi^{(u+1)}(t)$, and hence also $H^{(u+1)}(t) = \log\{\Phi^{(u+1)}(t)\}$. The procedure iterates between the E-step and the M-step until convergence.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

We point out that, although the functions $\psi_j(r)$'s are left as unknown, we can still calculate $\int a_{ij}^{(u)} dr_{ij}$ and $\int r_{ij} a_{ij}^{(u)} dr_{ij}$ in the M-step. Specifically,

$$\begin{aligned}
\int a_{ij}^{(u)} dr_{ij} &= q_{ij} \left\{1 - F_j(t)\right\}^{1-\delta_i} \left\{h^{(u)}(y_i) f_j(t)\right\}^{\delta_i} \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}}, \\
\int r_{ij} a_{ij}^{(u)} dr_{ij} &= \left\{e^{-t} q_{ij} f_j(t)\right\}^{1-\delta} \left[e^{-t} q_{ij} h^{(u)}(y_i)\{f_j(t) - f_j'(t)\}\right]^{\delta} \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}},
\end{aligned}$$

as shown in Appendix A.1, by taking advantage of the Laplace/inverse Laplace transform relation. In fact, even if an explicit form of $\psi_j(r)$ can be obtained, it is not necessary to go through the calculation because $\psi_j(r)$ itself is not needed. Finally, because ψ_j is defined as the inverse Laplace transform of a bounded function, it always exists for any ϵ distribution.

2.3 Theoretical properties

Although (1) is not a transformation model, under the list of conditions imposed in Appendix A.2, it can be cast into the general framework, Zeng and Lin (2007). To this end, we can verify that our Conditions (a), (b), (c) lead to their conditions (C1), (C2), (C3), respectively. Our Conditions (d) and (e) jointly ensure their conditions (C4) and (C8). Our Condition (f) leads to their condition (C6), and our Condition (g) leads to their conditions (C5), (C7). These are mild conditions mainly imposing identifiability, sufficient smoothness, and boundedness of various functions; They are usually satisfied in practice. Having verified the regularity conditions C1-C7 of Zeng and Lin (2007), we can use their results to obtain the asymptotic properties of the NPMLE in the linear transformation model in the mixture data setting. We state the results in Theorem 1 and provide the proof in Appendix A.3.

Theorem 1. *Let $\boldsymbol{\theta}_0, \Phi_0$ denote the true value of $\boldsymbol{\theta}, \Phi$, and write $\Phi = \{\Phi(t_1), \ldots, \Phi(t_K)\}^{\mathrm{T}}$. Under conditions (a)-(g) of Appendix A.2, $\widehat{\boldsymbol{\theta}}, \widehat{\Phi}$ are consistent, and have the asymptotic property that $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}, \widehat{\Phi} - \Phi)$ converges weakly to a zero mean Gaussian process. Then, for any function $a_1(s)$ with bounded total variation and any vector $\mathbf{a}_2$, $\sqrt{n}\int a_1(s) d\{\widehat{\Phi}(s) - \Phi(s)\} + \sqrt{n}\mathbf{a}_2^{\mathrm{T}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges to a zero mean normal distribution whose variance can be*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

approximated by

$$\mathbf{v}\{a_1(\cdot), \mathbf{a}_2\} \equiv -(\mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}}) \left\{ \frac{\partial^2 l(\widehat{\Phi}, \widehat{\boldsymbol{\theta}})}{\partial(\Phi^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})\partial(\Phi^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}} \right\}^{-1} (\mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}})^{\mathrm{T}},$$

where $\mathbf{a}_1 = \{a_1(t_1), \ldots, a_1(t_K)\}^{\mathrm{T}}$.

2.4 Inference

The main interest is often in the covariate effects described by $\boldsymbol{\theta}$. In such cases, we can perform inference using the results of a profiling procedure: at any $\boldsymbol{\theta}$, we use the same EM algorithm to calculate $\widehat{H}(T, \boldsymbol{\theta})$ except that we hold $\boldsymbol{\theta}$ fixed, and then calculate the information matrix using numerical derivatives. This is a simplification because it bypasses the need to invert a potentially high-dimensional matrix. For example, the $\alpha 100\%$ confidence interval for the jth component of $\boldsymbol{\theta}$, θ_j is

$$\begin{aligned} & \widehat{\theta}_j \pm Z_{(1+\alpha)/2} \left[-\sum_{i=1}^{n} \frac{\partial^2 l_i\{\boldsymbol{\theta}, \widehat{H}(t_1, \boldsymbol{\theta}), \ldots, \widehat{H}(t_K, \boldsymbol{\theta})\}}{\partial \theta_j^2} \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right]^{-1/2} \\ \approx\ & \widehat{\theta}_j \pm Z_{(1+\alpha)/2} \left[\sum_{i=1}^{n} \frac{-l_i\{\widehat{\boldsymbol{\theta}} + b\mathbf{e}_j, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}} + b\mathbf{e}_j), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}} + b\mathbf{e}_j)\}}{b^2} \right. \\ & \left. + \frac{2l_i\{\widehat{\boldsymbol{\theta}}, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}}), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}})\} - l_i\{\widehat{\boldsymbol{\theta}} - b\mathbf{e}_j, \widehat{H}(t_1, \widehat{\boldsymbol{\theta}} - b\mathbf{e}_j), \ldots, \widehat{H}(t_K, \widehat{\boldsymbol{\theta}} - b\mathbf{e}_j)\}}{b^2} \right]^{-1/2}, \end{aligned}$$

where $Z_{(1+\alpha)/2}$ is the $(1+\alpha)/2$ quantile of the standard normal distribution, l_i is the likelihood evaluated at the ith observation, $\mathbf{e}_j$ is the vector with zero components everywhere except the jth component being 1, and b is a small number that facilitates the numerical derivative.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

Likewise, for hypothesis testing of the form $H_0 : \boldsymbol{\theta} = \mathbf{c}$, we can construct the test statistic

$$\begin{aligned}
\mathbf{Z} &= \left[-\sum_{i=1}^{n} \frac{\partial^2 l_i\{\boldsymbol{\theta}, \widehat{H}(t_1,\boldsymbol{\theta}),\ldots,\widehat{H}(t_K,\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right]^{1/2}(\boldsymbol{\theta}-\mathbf{c}) \\
&\approx \left[\left(\sum_{i=1}^{n} \frac{-l_i\{\widehat{\boldsymbol{\theta}}+b\mathbf{e}_j+b\mathbf{e}_k, \widehat{H}(t_1,\widehat{\boldsymbol{\theta}}+b\mathbf{e}_j+b\mathbf{e}_k),\ldots,\widehat{H}(t_K,\widehat{\boldsymbol{\theta}}+b\mathbf{e}_j+b\mathbf{e}_k)\}}{4b^2}\right.\right. \\
&\quad + \frac{l_i\{\widehat{\boldsymbol{\theta}}+b\mathbf{e}_j-b\mathbf{e}_k, \widehat{H}(t_1,\widehat{\boldsymbol{\theta}}+b\mathbf{e}_j-b\mathbf{e}_k),\ldots,\widehat{H}(t_K,\widehat{\boldsymbol{\theta}}+b\mathbf{e}_j-b\mathbf{e}_k)\}}{4b^2} \\
&\quad + \frac{l_i\{\widehat{\boldsymbol{\theta}}-b\mathbf{e}_j+b\mathbf{e}_k, \widehat{H}(t_1,\widehat{\boldsymbol{\theta}}-b\mathbf{e}_j+b\mathbf{e}_k),\ldots,\widehat{H}(t_K,\widehat{\boldsymbol{\theta}}-b\mathbf{e}_j+b\mathbf{e}_k)\}}{4b^2} \\
&\quad \left.\left. - \frac{l_i\{\widehat{\boldsymbol{\theta}}-b\mathbf{e}_j-b\mathbf{e}_k, \widehat{H}(t_1,\widehat{\boldsymbol{\theta}}-b\mathbf{e}_j-b\mathbf{e}_k),\ldots,\widehat{H}(t_K,\widehat{\boldsymbol{\theta}}-b\mathbf{e}_j-b\mathbf{e}_k)\}}{4b^2}\right)_{jk}\right]^{1/2} \\
&\quad \times(\boldsymbol{\theta}-\mathbf{c}),
\end{aligned}$$

and note that $\mathbf{Z}$ is approximately a standard multivariate normal random variable under H_0. Here, we use the notation (A_{jk}) to denote the square matrix $\mathbf{A}$ with size the length of $\boldsymbol{\theta}$ and (j,k) entry A_{jk}.

3 Homogeneous and no covariate effect model

When either $\boldsymbol{\beta}$ or $\boldsymbol{\alpha}_j$ does not appear in (1), the model is more restrictive and the computation simplifies. If $\boldsymbol{\beta}$ does not appear, then there is no homogeneous covariate effect in the transformation model. In terms of estimation, the procedures follows the same line with some minor simplifications. However, if $\boldsymbol{\alpha}_j$ does not appear, (1) greatly simplifies and can be treated quite differently, as we now explain.

The common-effect covariate effect model for the jth population is

$$H(T) = -\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_j,$$

where all the components in the model retain the same interpretation as in (1). The implication of the model is that the heterogeneity between subpopulations is due to the different variability of measurement errors, but not the heterogeneous effect of covariates. The con-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

ditional distribution is then simplified to

$$\begin{aligned} f(Y,\Delta \mid \mathbf{X}) &= \left[h(y)\sum_{j=1}^{p} q_j f_j\{H(y)+\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}\right]^{\delta}\left[1-\sum_{j=1}^{p} q_j F_j\{H(y)+\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}\right]^{1-\delta} \\ &= \left[h(y)\mathbf{q}^{\mathrm{T}}\mathbf{f}\{H(y)+\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\}\right]^{\delta}\left[1-\mathbf{q}^{\mathrm{T}}\mathbf{F}\left\{H(y)+\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right\}\right]^{1-\delta}, \end{aligned}$$

where $\mathbf{f}=(f_1,\ldots,f_p)^{\mathrm{T}}$, $\mathbf{F}=(F_1,\ldots,F_p)^{\mathrm{T}}$, and $h(y)\equiv H'(y)$, because the same transformation H and the same parameter $\boldsymbol{\beta}$ are assumed across all p populations. The population difference is only reflected in the distribution of ϵ_j, which is assumed to be f_j. We can however still use the different f_j's of the model to account for unexplained residual population heterogeneity, for example, different variances.

As before, estimating the distribution in each population is equivalent to estimating H and $\boldsymbol{\beta}$. As the $\mathbf{q}_i$'s have $m\geq p$ different vector values $\mathbf{u}_1,\ldots,\mathbf{u}_m$, assign the n observations to these m groups according to their $\mathbf{q}$ values. Assume there are, respectively, $r_1,\ldots,r_m$ observations in each group. In group k, we can view the model as a transformation model with the same transformation H, the same parameter $\boldsymbol{\beta}$, but a new distribution for ϵ, which has the mixture form $\mathbf{u}_k^{\mathrm{T}}\mathbf{f}(\epsilon)$. Thus, we can use the existing estimation method for transformation models to obtain the estimators of H and $\boldsymbol{\beta}$, using exclusively the kth group data. Denote the resulting estimators as $\widehat{H}_k$ and $\widehat{\boldsymbol{\beta}}_k$. We can then take the weighted average to obtain the final estimator $\widehat{H}(t)=\sum_{k=1}^{m} w_k(t)\widehat{H}_k(t)$ and $\widehat{\boldsymbol{\beta}}=\sum_{k=1}^{m}\mathbf{w}_k\widehat{\boldsymbol{\beta}}_k$. To be consistent with the estimation in the general model (1), we use the NPMLE proposed by Zeng and Lin (2006). Thus, we obtain $\widehat{\boldsymbol{\beta}}_k$, $\widehat{H}_k$ via maximizing

$$\begin{aligned} l_k(H,\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n} I(\mathbf{q}_i=\mathbf{u}_k)\left(\delta_i \log\left[h(y_i)\mathbf{u}_k^{\mathrm{T}}\mathbf{f}\{H(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}\right]\right. \\ &\quad \left.+(1-\delta_i)\log\left[1-\mathbf{u}_k^{\mathrm{T}}\mathbf{F}\left\{H(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right\}\right]\right) \end{aligned}$$

with respect to $\boldsymbol{\beta}$ and H. Here, we restrict $H(y)$ to be a piecewise constant nondecreasing function with nonnegative jumps only at the y_i's where $\mathbf{q}_i=\mathbf{u}_k$ and $\delta_i=1$. We write these jump points $t_1,\ldots,t_K$, and write $\mathbf{H}_k=\{H(t_1),\ldots,H(t_K)\}^{\mathrm{T}}$. Zeng and Lin (2006) showed that the resulting $\widehat{\boldsymbol{\beta}}_k,\widehat{H}_k$ are consistent, and that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta},\widehat{H}_k-H)$ converges weakly to

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

a zero mean Gaussian process. Thus, for any function $a_1(s)$ with bounded total variation and any vector $\mathbf{a}_2$, $\sqrt{n}\int a_1(s)d\{\widehat{H}_k(s) - H(s)\} + \sqrt{n}\mathbf{a}_2^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta})$ converges to a zero mean normal distribution whose variance can be approximated by

$$\mathbf{v}_k\{a_1(\cdot), \mathbf{a}_2\} \equiv -(\mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}}) \left\{ \frac{\partial^2 l_k(\widehat{H}_k, \widehat{\boldsymbol{\beta}}_k)}{\partial(\mathbf{H}_k^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})\partial(\mathbf{H}_k^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}} \right\}^{-1} (\mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}})^{\mathrm{T}},$$

where $\mathbf{a}_1 = \{a_1(t_1), \ldots, a_1(t_K)\}^{\mathrm{T}}$.

It remains to determine the choice of weights $\mathbf{w}_k$. Because the estimation in different group is based on different subjects, they are independent. Hence the optimal weights are proportional to the inverse of the variance of the estimators. The optimal weights for $\widehat{H}(t)$ are then $w_k(t) = \mathbf{v}_k\{I(s \le t), \mathbf{0}\}^{-1}/[\sum_{k=1}^m \mathbf{v}_k\{I(s \le t), \mathbf{0}\}^{-1}]$ and $\mathbf{w}_k$ is a diagonal matrix with the jth diagonal element $w_{kj} = \mathbf{v}_k(0, \mathbf{e}_j)^{-1}/\{\sum_{k=1}^m \mathbf{v}_k(0, \mathbf{e}_j)^{-1}\}$. In practice, this may not work well since it relies on asymptotic results. Based on prior work in Ma and Wang (2014), a simple choice of $w_k(t) = \mathbf{w}_k = r_k^{-1}$ has satisfactory performance.

Because the within group NPMLE already guarantees the monotonicity of each $\widehat{H}_k$, the final weighted average estimator for $\widehat{H}$ is monotone. The asymptotic property of $\widehat{H}$ and $\boldsymbol{\beta}$ is standard: $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \widehat{H} - H)$ converges weakly to a zero mean Gaussian process. Then, for any function $a_1(t)$ with bounded total variation and any vector $\mathbf{a}_2$, $\sqrt{n}\int a_1(s)d\{\widehat{H}(s) - H(s)\} + \sqrt{n}\mathbf{a}_2^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a zero mean normal distribution whose variance can be approximated with

$$\mathbf{v}\{a_1(\cdot), \mathbf{a}_2\} \equiv \sum_{k=1}^m \mathbf{v}_k\{a_1(\cdot)w_k(\cdot), \mathbf{w}_k\mathbf{a}_2\}$$

where $t_1, \ldots, t_K$ are the observed event times.

Testing whether population heterogeneity in the covariate effects is present in (1) is equivalent to testing $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \cdots = \boldsymbol{\alpha}_p$. This can be written as testing $\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$, $\mathbf{A}$ a $(p-1)d_z \times (d_x + pd_z)$ block matrix in which the (j, j) block is $\mathbf{I}$ and the $(2, j)$ block is $-\mathbf{I}$ for $j = 3, \ldots, p+1$. All other blocks are zero. Based on the asymptotic results in Section 2, we can conveniently use a Wald test: under Φ_0, $n(\mathbf{A}\boldsymbol{\theta})^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{A}\boldsymbol{\theta}$ has χ^2 distribution with

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

$(p-1)d_z$ degrees of freedom, where

$$\mathbf{V} = -(\mathbf{0}_{(p-1)d_z \times K}, \mathbf{A}) \left\{ \frac{\partial^2 l(\widehat{\Phi}, \widehat{\boldsymbol{\theta}})}{\partial(\Phi^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})\partial(\Phi^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}} \right\}^{-1} (\mathbf{0}_{(p-1)d_z \times K}, \mathbf{A})^{\mathrm{T}}.$$

When no covariate is included in the model, $\boldsymbol{\beta}$ does not appear. The procedure can then be directly applied with the simplification of deleting all the steps concerning estimating $\boldsymbol{\beta}$: we estimate $H(\cdot)$ from each of the m groups, then combine the results via a weighted average. This is similar to the approaches in Wacholder et al. (1998) and in Ma and Wang (2014), except that the estimation of $H(\cdot)$ in each group is carried out via MLE instead of least squares, and the weight selection is different from that in Wacholder et al. (1998).

4 Simulation Studies

We performed six sets of simulation studies to demonstrate the performance of the proposed method for the transformation model in the mixture data context. We present three of the simulation studies here and relegate the remaining three to Appendix A.4. Our first set of simulations contain homogeneous covariate effects. We generated data using $p = 2$, without $\boldsymbol{\alpha}_j$, and $\mathbf{X}$ a bivariate random vector. The first component of $\mathbf{X}$ was a binary variable, taking values 1 or 0 each with probability 0.5, the second component was uniform on -1 to 1. The transformation H was a logarithm function. We set f_1 to be the extreme value distribution, f_2 to be the logistic distribution. The censoring distribution was exponential, resulting in an overall censoring rate about 25%. The results are in the first block of Table 1 and upper-left plot of Figure 1. For comparison, we also did the estimation treating the homogeneous effect as heterigeneous, and estimated β_1, β_2 as $\alpha_{11}, \alpha_{21}, \alpha_{12}, \alpha_{22}$ instead. The results are in the second block of Table 1 and upper-right plot of Figure 1. These estimations are still consistent, yet the variability roughly doubled.

The second set of simulations studied heterogeneous covariate effects. It included $\boldsymbol{\alpha}_j$, but not $\boldsymbol{\beta}$. We generated data using $p = 2$. $\mathbf{Z}$ was of the same structure as $\mathbf{X}$ in the first simulation for the first two terms and an intercept term for the third term. We kept H

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

the same as in the first simulation. Usually, in transformation models, the intercept term is not identifiable. In our case, the difference of the intercepts in different populations is identifiable, and hence was estimated. Here we set f_1 to be standard normal and f_2 to be a t distribution with 5 degrees of freedom. The censoring distribution was still exponential to achieve a 20% overall censoring rate. Results are in the second block of Table 1 and lower-left plot of Figure 1.

Our third simulation included both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_j$. We generated data using $p = 2$. $\mathbf{X}$ is bivariate with the first component either 1 or 0 with equal probability, and the second component a standard normal. Z was a uniform covariate on [-1 1] and a constant 1 to capture the intercept. The true H was still the log transformation. We took both f_1, f_2 to be normal with mean zero, but the second population had four times the variance as the first. The censoring distribution was exponential yielding a 20% overall censoring rate. The results are in the third block of Table 1 and the lower-right plot of Figure 1.

The simulation studies suggest that the proposed method has satisfactory finite sample performance: the parameter estimation yields small biases in all three simulations, measured by the mean and median of the 1000 estimates; Inference results are precise, in that the sample standard deviation from the 1000 simulations are closely matched by the average and the median of the 1000 estimated standard deviations calculated from the asymptotic results. The overall distribution of the estimated parameters are close to normal, as indicated by the empirical coverage of the 95% confidence intervals, which are close to their nominal levels. The estimation of the transformation function H, as shown in Figure 1, is within expectations. Overall, the average of the curve estimation approximately overlays the true H curve, while the 95% confidence bands have better performance than the typical nonparametric curve estimation. This is because H is estimated as the root-n rate, instead of the usual nonparametric rate. We also tired different transformations than H, with the overall performance similar. The details of these simulations are in Appendix A.4.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

5 Application to Huntington's Disease Study

HD is the most prevalent monogenic neurodegenerative disorder caused by expansion of C-A-G repeats at the HD gene on chromosome 4 MacDonald et al. (1993). Typically neurological, cognitive, and physical symptoms begin to exhibit around 30-50 years of age for affected individuals, and eventually death is from pneumonia, heart failure, or other complications 15-20 years after the diagnosis Foroud et al. (1999). The subjects analyzed here were recruited in the Cooperative Huntington's Observational Research Trial (COHORT, Dorsey and The Huntington Study Group COHORT Investigators 2012), an epidemiological study of the natural history of HD. The probands were recruited primarily at academic research centers from 50 sites in the United States, Canada, and Australia. Probands were either clinically diagnosed with HD or the individuals who pursued HD genetic testing and carried a mutation but who were not clinically diagnosed. The initial probands underwent clinical examination and genotyping for HD mutation, and reported family history information on their first-degree relatives. The relatives were not genotyped because there was no resource for in-person collection of blood samples. Thus the relatives' HD mutation status was unknown, while the distribution of their mutation status could be estimated from the pedigree structure and the probands' carrier status. The full details of the COHORT study design are described in Dorsey and The Huntington Study Group COHORT Investigators (2012) and in Wang et al. (2012).

There were 4105 subjects included in the COHORT analysis, and they were either mutation carriers or not, hence $p = 2$. The heterogeneous covariate effect model (1) was used to study the effect of several covariates on mortality in HD mutation carriers where, for carriers, f_1 was normal with mean zero standard deviation 0.2, and for non-carriers, f_2 was $0.2T_5$, with T_5 a student t with 5 degrees-of-freedom. The main research interest is to predict age at death based on CAG repeats length, adjusting for gender, proband's HD clinical diagnosis status and a relative's relationship to the probands. We assumed all covariates to have differential effect in each mutation group to allow for maximal flexibility.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

The covariates included in the model were: CAG repeats length at the HD gene, gender, and proband's HD diagnosis status.

The results are reported in Table 2. As expected, the effects of CAG repeats length has a significant effect on age-at-death with an estimated effect of -0.76 (SE: 0.09, p-value<0.001). The results suggest that if all covariates are the same, the subjects with one unit CAG longer repeat are expected to have a 2.38 years shorter lifespan. Here 2.38 is calculated as the average of $\widehat{H}^{-1}(U) - \widehat{H}^{-1}(U - 0.76)$ for a random U, where $\widehat{H}$ is the estimated transformation function and is close to a linear function (See Figure 2). This finding is consistent with the clinical literature which indicates an inverse association between CAG repeats length and HD age at diagnosis and death, Foroud et al. (1999), Langbehn et al. (2004). Proband's HD diagnosis also has a significant effect after adjusting for CAG repeats and other covariates: having a positive HD diagnosis in a family member is associated with an earlier mean age-at-death in carrier, potentially due to other shared familial risk factors.

The estimated transformation $H(\cdot)$ and its bootstrap confidence interval are presented in Figure 2. The nonparametric function suggests that a linear transformation may fit the data adequately and, under a parametric approximation, predictions formula for the age-at-death in a mutation carrier subject can be obtained. The approximated linear function is $\widehat{H}(t) = -24.35 + 0.32t$, see Figure 3.

A limitation of our analysis is that probands data were not included to protect against potential bias resulting from unknown sources in the COHORT study that did not use a population-based ascertainment scheme for probands. When the proband ascertainment is population-based, for example, probands are randomly selected from diseased population (case-family design), their data may be included through a retrospective likelihood. It would be interesting to replicate our analysis in an independent study using such a design, including probands data in the analyses.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

6 Discussion

A potentially interesting extension of our method is to further parametrize the mixing distributions and estimate the parameters from data. If the q_{ij}'s are modeled parametrically, semiparametrically, or nonparametrically and estimated as $\widehat{q}_{ij}$, it would be interesting to develop methods to account for the discrepancy between $\widehat{q}_{ij}$ and q_{ij} and to deliver appropriate estimation of the survival function and covariate effect using the $\widehat{q}_{ij}$.

Our method has the flexibility to account for cross-population heterogeneity by characterizing the outcome in each population using different distributions specified by covariate parameters and error distributions (e.g., distinct scale or shape parameter; population-specific covariate effect), while simultaneously allow for common components across populations (e.g., shared covariate effect). Whether or not to adopt population-specific effects or shared effects is often determined by the purpose of the analysis and prior knowledge. In many cases, covariates whose effects are of particular research interest might be assumed to be population-specific as a precaution, while covariates that are not of interest be modeled across population.

We have assumed that the relative observations are independent, and excluded probands from the analyses. In proband-relative studies, multiple relatives from the same family may be collected and thus could have residual familial correlation. Our current approach is still consistent if the probands are representative samples of the probands population, but the inferences developed would no longer be valid. When probands are not representative and there is residual familial aggregation, ascertainment schemes may need to be modeled and probands and relative data analyzed jointly. How to best accommodate familial correlation and adjust for probands ascertainment schemes is highly challenging, and interesting.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

Acknowledgements

This work is supported by NIH grants NS073671, NS082062 and NSF grant DMS1608540. The authors wish to thank CHDI Foundation and COHORT study investigators. Samples and/or data from the COHORT study, which received support from HP Therapeutics, Inc., were used in this study. The authors also thank an associate editor and the reviewers for their constructive comments that have helped improving the quality of the paper.

Table 1: Simulation results based on 1000 repetitions.

	true	mean	median	sd	mean($\widehat{\text{sd}}$)	median($\widehat{\text{sd}}$)	95% CI
				simulation 1.1			
β_1	1.0000	0.9834	0.9703	0.4384	0.4474	0.4472	0.9570
β_2	2.0000	1.9734	1.9626	0.3845	0.3958	0.3954	0.9570
				simulation 1.2			
α_{11}	1.0000	0.9958	0.9992	1.0400	0.9623	0.9414	0.9410
α_{12}	2.0000	2.0420	2.0456	0.8916	0.8539	0.8199	0.9310
α_{21}	1.0000	0.9915	1.0140	0.8581	0.8395	0.8378	0.9420
α_{22}	2.0000	1.9684	1.9879	0.7328	0.7436	0.7350	0.9530
				simulation 2			
α_{11}	1.0000	1.0644	1.0584	1.1017	1.1758	1.1264	0.9530
α_{12}	2.0000	2.0767	2.0493	1.2519	1.3178	1.2870	0.9620
α_{21}	1.5000	1.4353	1.4306	0.7582	0.8072	0.7918	0.9640
α_{22}	3.0000	2.9344	2.9167	0.8787	0.9039	0.8852	0.9490
				simulation 3			
β_1	1.0000	0.9895	0.9915	0.3944	0.3976	0.3974	0.9520
β_2	1.5000	1.4974	1.4894	0.1983	0.2083	0.2079	0.9560
α_1	2.0000	1.9007	1.9443	1.1372	1.1737	1.1683	0.9600
α_2	3.0000	3.0040	2.9988	0.5071	0.5071	0.5028	0.9420

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

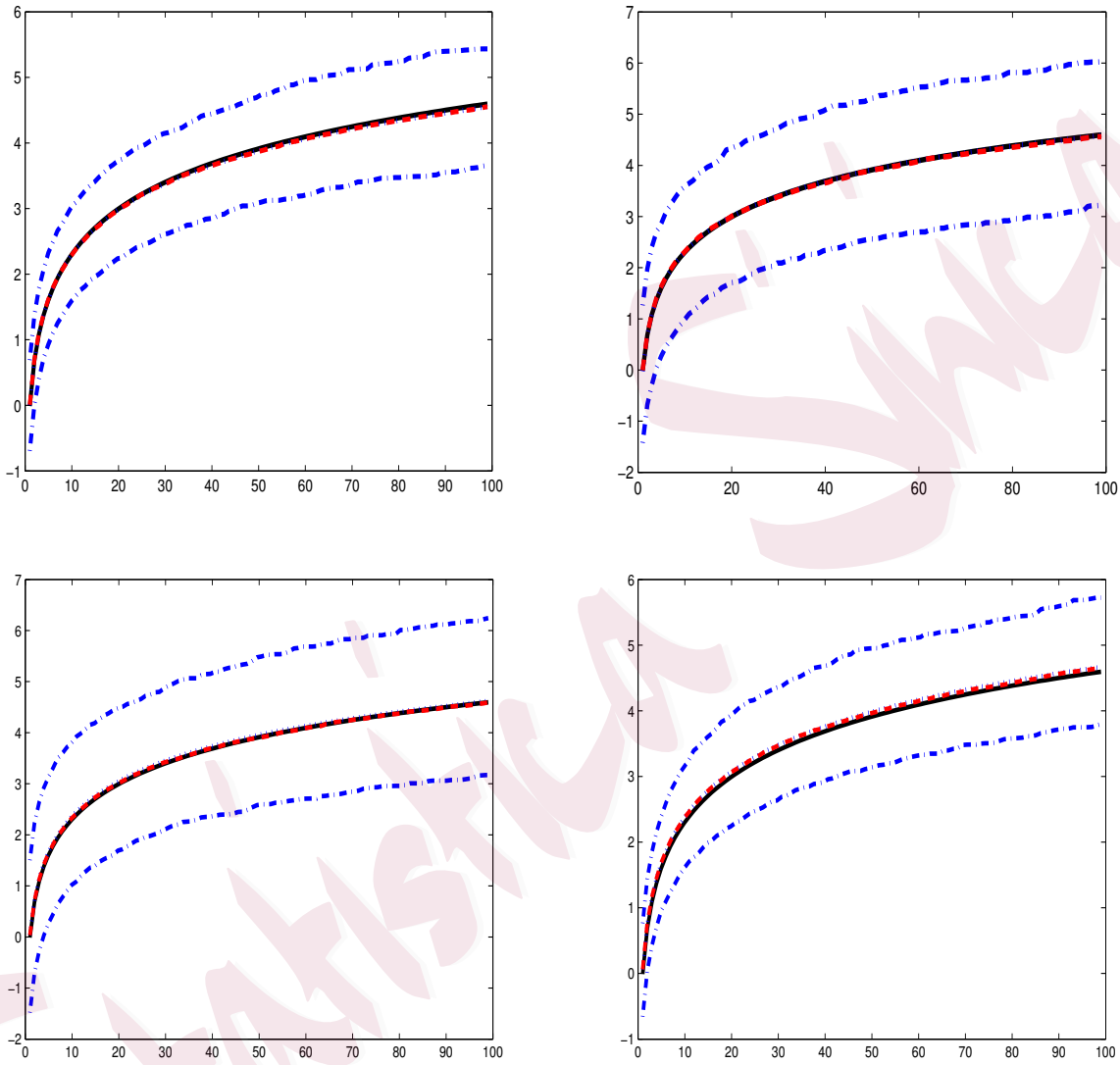

Figure 1: True function (solid line), median estimation (dashed line), mean estimation (dotted line) and 95% confidence band (dash-dotted line) of $H(T)$ in simulations 1.1 (upper-left), 1.2 (upper-right), 2 (lower-left), and 3 (lower-right) .

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

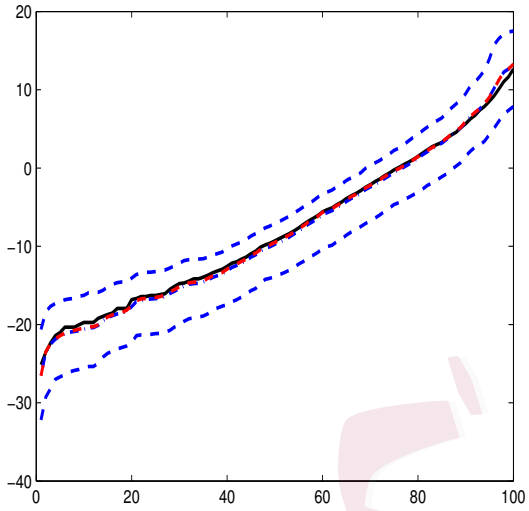

Figure 2: Estimated H function (solid line), median estimation (dashed line), mean estimation (dash-dotted line) and 95% confidence band (dashed line) of $H(T)$ in data analysis. Median, mean and 95% confidence band are based on 1000 bootstrapped samples.

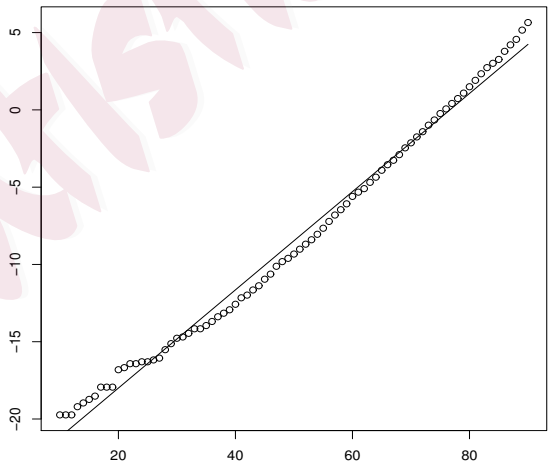

Figure 3: Fitted linear function $\widehat{H}(t)$ versus age t for HD data analysis.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

Table 2: COHORT analysis results: estimated covariate effects (age, gender, proband's diagnosis of HD), their standard errors, and p-values.

	Carriers				Non-carriers			
	$\alpha_{1intercept}$	α_{1Age}	$\alpha_{1Gender}$	$\alpha_{1ProDiag}$	$\alpha_{2intercept}$	α_{2Age}	$\alpha_{2Gender}$	$\alpha_{2ProDiag}$
est	-33.65	0.76	-0.67	1.79	-7.07	0.18	2.82	-2.30
se	4.28	0.09	0.70	1.00	1.25	0.03	0.67	0.84
p-value	< 0.001	< 0.001	0.34	0.07	< 0.001	< 0.001	< 0.001	0.006

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

Appendix

A.1 Derivation of $\int a_{ij}^{(u)} dr_{ij}$ and $\int r_{ij} a_{ij}^{(u)} dr_{ij}$

Here we show the derivation of the relationships

$$\begin{aligned}
\int a_{ij}^{(u)} dr_{ij} &= q_{ij} \left\{1 - F_j(t)\right\}^{1-\delta_i} \left\{h^{(u)}(y_i) f_j(t)\right\}^{\delta_i} \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}} \\
\int r_{ij} a_{ij}^{(u)} dr_{ij} &= e^{-t} q_{ij} f_j(t) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}} \quad \text{if } \delta_i = 0 \\
\int r_{ij} a_{ij}^{(u)} dr_{ij} &= -e^{-t} q_{ij} h^{(u)}(y_i) \{f_j'(t) - f_j(t)\} \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}} \quad \text{if } \delta_i = 1.
\end{aligned}$$

Let $t = H^{(u)}(y_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)} + \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}$. Then

$$a_{ij}^{(u)} = \{h^{(u)}(y_i) r_{ij} \exp(t)\}^{\delta_i} \exp(-r_{ij}e^t) q_{ij} \psi_j(r_{ij}) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}}.$$

When $\delta_i = 0$,

$$\begin{aligned}
\frac{da_{ij}^{(u)}}{dt} &= -r_{ij}e^t \exp(-r_{ij}e^t) q_{ij} \psi_j(r_{ij}) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}} \\
\frac{d^2a_{ij}^{(u)}}{dt^2} &= -r_{ij}e^t \exp(-r_{ij}e^t) q_{ij} \psi_j(r_{ij}) + r_{ij}^2 e^{2t} \exp(-r_{ij}e^t) q_{ij} \psi_j(r_{ij}) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}}.
\end{aligned}$$

When $\delta_i = 1$,

$$\begin{aligned}
\frac{da_{ij}^{(u)}}{dt} &= -h^{(u)}(y_i) r_{ij}^2 e^{2t} \exp(-r_{ij}e^t) q_{ij} \psi_j(r_{ij}) \\
&\quad + h^{(u)}(y_i) r_{ij} e^t \exp(-r_{ij}e^t) q_{ij} \psi_j(r_{ij}) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}}
\end{aligned}$$

Thus, when $\delta_i = 0$,

$$r_{ij} a_{ij}^{(u)} = -e^{-t} \frac{da_{ij}^{(u)}}{dt} \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}, \delta_i=0},$$

and when $\delta_i = 1$,

$$r_{ij} a_{ij}^{(u)} = h^{(u)}(y_i) e^{-t} \left(\frac{d^2 a_{ij}^{(u)}}{dt^2} - \frac{da_{ij}^{(u)}}{dt} \right) \Big|_{t=H^{(u)}(y_i)+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u)}+\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}_j^{(u)}, \delta_i=0}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

A.2 List of regularity conditions

(a) The parameter value $\boldsymbol{\theta}_0$ belongs to the interior of a compact set $\boldsymbol{\Theta} \in \mathbb{R}^{d_\theta}$, and $\phi_0(t) > 0$ for all $t \in [0, \tau]$. (C1).

(b) With probability 1, $\text{pr}(Y_i \geq \tau \mid \mathbf{X}_i, \mathbf{Z}_i) > \delta_0 > 0$ for some constant $\delta_0 > 0$. (C2).

(c) $f_j(s)$ is bounded away from zero and infinity on its support for $j = 1, \ldots, p$. (C3).

(d) $f_j(s)$ is three times continuously differentiable and, the $f_j^{(v)}(s)/\exp(ks), v = 0, \ldots, 3, k = 2, \ldots, 4$, are square integrable on $(-\infty, \log(\tau)]$ for $j = 1, \ldots, p$. (C4), (C8).

(e) The covariates $\mathbf{X}, \mathbf{Z}$ have finite kth moments, $k = 1, \ldots, 6$. (C4), (C8).

(f) The first moment of $\log f_j(s)$ exists for $j = 1, \ldots, p$. (C6).

(g) $m \geq p$ and the matrix $(\mathbf{u}_1, \ldots \mathbf{u}_m)$ has rank p. (C5), (C7).

A.3 Proof of Theorem 1

Because NPMLE for the linear transformation model in the mixture model setting we consider can be cast into the general framework established in Zeng and Lin (2007), we prove Theorem 1 through verifying the conditions (C1)-(C8) required by them.

Condition (a) ensures that the true parameter value is in the interior of a compact set of the parameter space, with Conditions (c) and (d), we further guarantee the differentiability and positivity of the hazard function. This leads to condition (C1) of Zeng and Lin (2007).

Condition (b) is equivalent to their (C2).

Condition (c) guarantees that (C3) of Zeng and Lin (2007) is satisfied.

Condition (C4) of Zeng and Lin (2007) is a type of Lipschitz condition, with respect to both parameter and function; It is guaranteed by the stronger differentiability conditions in our Condition (d) and and the moment conditions in (e).

Our Condition (g) is stated in their (C5).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

Condition (C6) of Zeng and Lin (2007) requires sufficient smoothness and boundedness of the hazard functions and some functions derived from them, as do our Conditions (c), (d) and (f).

Condition (C7) there is an identifiability condition that arises due to the generality of the framework they consider; It is guaranteed to hold under our Condition (g) and the parameterization requiring $H(0) = -\infty$.

Condition (C8) of Zeng and Lin (2007) strengthen their (C4) to hold along each path in a neighborhood of the true parameter value, while our Conditions (d) and (e) are imposed for all the parameter values in a compact set jointly ensuring that this holds.

□

A.4 Additional simulations

Our fourth simulation is the same as the first, except that the true transformation H is $\log\{t/(1-t)\}$. In this case, the overall censoring rate is about 25%. The results are in Table 3 and Figure 4.

Similarly, the fifth simulationis the same as the second but with $H = \log\{t/(1-t)\}$, and an overall censoring rate of about 20%. The results are in Table 3 and Figure 4.

The sixth simulation is the same as the third except that $H = \log\{t/(1-t)\}$, with an overall censoring rate of about 25%. The results are in Table 3 and Figure 4.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

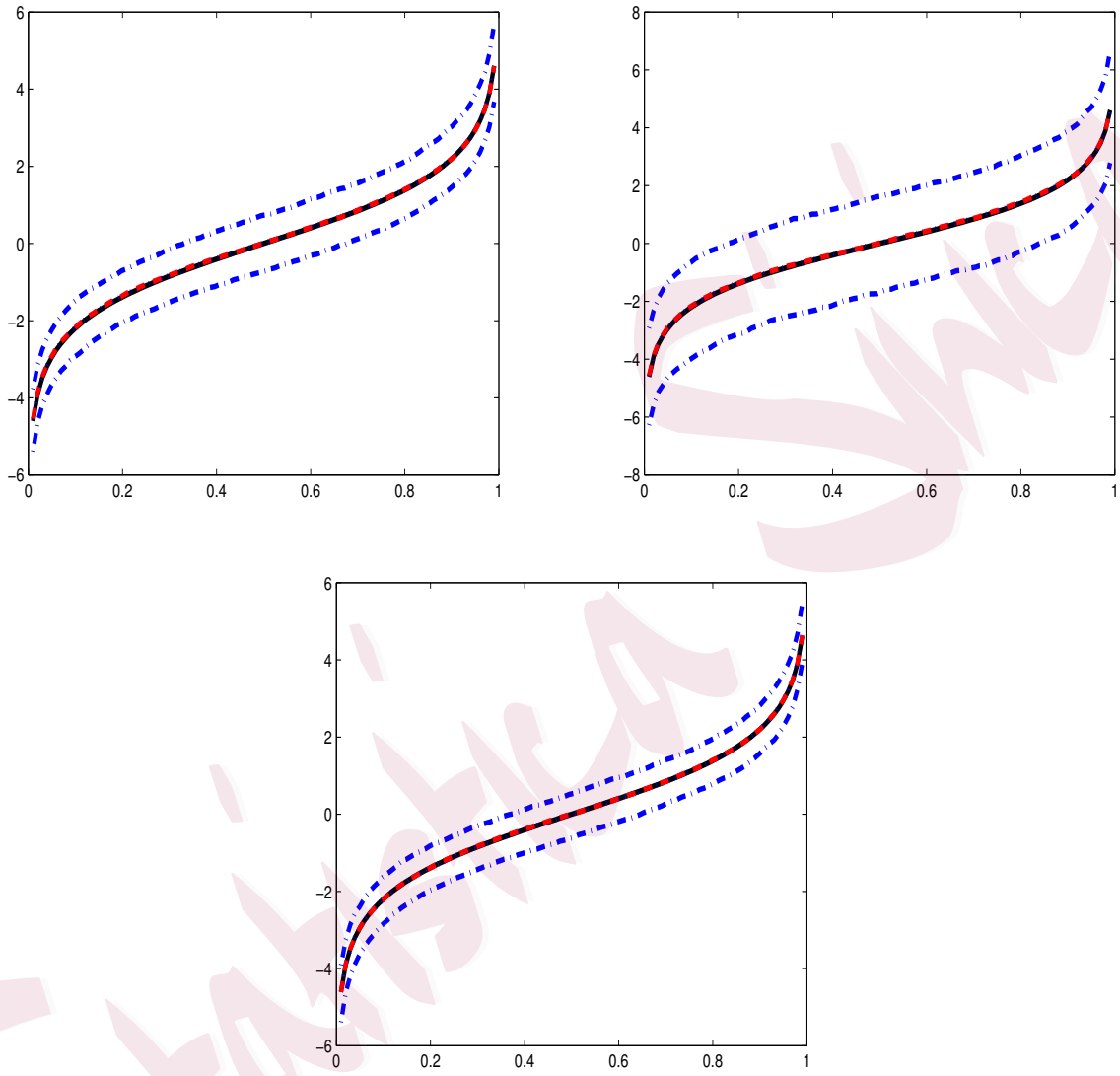

Figure 4: True function (solid line), median estimation (dashed line), mean estimation (dotted line) and 95% confidence band (dash-dotted line) of $H(T)$ in simulations 1 (upper-left), 2 (upper-right), and 3 (lower) .

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

Table 3: Simulation results. Results based on 1000 simulations.

	true	mean	median	sd	mean($\widehat{\text{sd}}$)	median($\widehat{\text{sd}}$)	95% CI
				simulation 4			
β_1	1.0000	0.9809	0.9776	0.4393	0.4605	0.4601	0.9650
β_2	2.0000	1.9693	1.9565	0.3974	0.4088	0.4084	0.9540
				simulation 5			
α_{11}	1.0000	0.9893	0.9986	0.6229	0.6363	0.6351	0.9590
α_{12}	2.0000	1.9895	1.9988	0.5339	0.5552	0.5535	0.9550
α_{21}	1.5000	1.4764	1.4410	1.1660	1.1346	1.1292	0.9530
α_{22}	3.0000	2.9565	2.9681	0.9947	0.9971	0.9933	0.9460
				simulation 6			
β_1	1.0000	0.9973	0.9914	0.2951	0.2982	0.2978	0.9590
β_2	1.5000	1.5038	1.4982	0.1551	0.1569	0.1567	0.9590
α_1	2.0000	1.8943	1.9186	0.7693	0.7955	0.7945	0.9510
α_2	3.0000	3.0311	3.0257	0.3728	0.3609	0.3595	0.9560

References

Altstein, L. and Li, G. (2013), 'Latent subgroup analysis of a randomized clinical trial through a semiparametric accelerated failure time mixture model', *Biometrics* **69**, 52–61.

Dorsey, E. R. 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**, e29522.

Foroud, T., Gary, J., Ivashina, J. and Conneally, P. (1999), 'Differences in duration of huntington's disease based on age at onset', *Journal of Neurology, Neurosurgery and Psychiatry* **66**, 52–56.

Holzmann, H., Munk, A. and Stratmann, B. (2006), 'Identifiability of finite mixtures of elliptical distributions', *Scandinavian Journal of Statistics* **33**, 753–763.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

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

Langbehn, D. R., Brinkman, R. R., Falush, D., Paulsen, J. S. and Hayden, M. R. (2004), 'A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length', *Clinical genetics* **65(4)**, 267–277.

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.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L. et al. (1993), 'A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes', *Cell* **72(6)**, 971–983.

Marder, L., Sulzbach, G. O., Bernardes, A. M. and Ferreira, J. Z. (2003), 'Removal of cadmium and cyanide from aqueous solutions through electrodialysis', *Journal of the Brazilian Chemical Society* **14**, 610–615.

Wacholder, S., Hartge, P., Struewing, J., Pee, D., McAdams, M., Brody, L. and Tucker, M. (1998), 'The kin-cohort study for estimationg penetrance', *American Journal of Epidemiology* **148**, 623630.

Wang, Y., Garcia, T. P. and Ma, Y. (2012), 'Nonparametric estimation for censored mixture data with application to the cooperative huntington's observational research trial', *Journal of the American Statistical Association* **107**, 1324–1338.

Zeng, D. and Lin, D. Y. (2006), 'Efficient estimation of semiparametric transformation models for counting process', *Biometrika* **93**, 627–640.

Zeng, D. and Lin, D. Y. (2007), 'Maximum likelihood estimation in semiparametric regression models with censored data', *Journal of the Royal Statistical Society* **69**, 507–564.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0199

Zhang, Y., Long, J. D., Mills, J. A., Warner, J. H., Lu, W. paulsen, J. S. and the PREDICT-HD Investigators and Coordinators of the Huntington Study Group (2012), 'Indexing disease progression at study entry with individuals at-risk for huntington disease', *American journal of medical genetics: neuropsychiatric genetics* .