Biometrika (2015), **102**, 3, *pp.* 515–532
Printed in Great Britain

doi: 10.1093/biomet/asv030
Advance Access publication 14 July 2015

Efficient estimation of nonparametric genetic risk function with censored data

BY YUANJIA WANG

Department of Biostatistics, Mailman School of Public Health, 722 West 168th Street,
New York, New York 10032, U.S.A.
yw2016@columbia.edu

BAOSHENG LIANG, XINGWEI TONG

School of Mathematical Sciences, Beijing Normal University, Beijing 100875, China
liangbs@mail.bnu.edu.cn xweitong@bnu.edu.cn

KAREN MARDER

Department of Neurology and Psychiatry, College of Physicians and Surgeons,
Columbia University, New York, New York 10032, U.S.A.
ksm1@columbia.edu

SUSAN BRESSMAN

The Alan and Barbara Mirken Department of Neurology, Beth Israel Medical Center,
New York, New York 10003, U.S.A.
sbressma@chpnet.org

AVI ORR-URTREGER, NIR GILADI

Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
aviorr@tasmc.health.gov.il nirg@tasmc.health.gov.il

AND DONGLIN ZENG

Department of Biostatistics, CB #7420, University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina 27599, U.S.A.
dzeng@bios.unc.edu

SUMMARY

With the discovery of an increasing number of causal genes for complex human disorders, it is crucial to assess the genetic risk of disease onset for individuals who are carriers of these causal mutations and to compare the distribution of the age-at-onset for such individuals with the distribution for noncarriers. In many genetic epidemiological studies that aim to estimate causal gene effect on disease, the age-at-onset of disease is subject to censoring. In addition, the mutation carrier or noncarrier status of some individuals may be unknown, due to the high cost of in-person ascertainment by collecting DNA samples or because of the death of older individuals. Instead, the probability of such individuals' mutation status can be obtained from various other sources. When mutation status is missing, the available data take the form of censored mixture data. Recently, various methods have been proposed for risk estimation using such data, but none is efficient for estimating a nonparametric distribution. We propose a fully efficient sieve maximum likelihood estimation method, in which we estimate the logarithm of the hazard ratio between genetic mutation groups using B-splines, while applying nonparametric maximum likelihood estimation to the reference baseline hazard function. Our estimator can be calculated via an expectation-maximization algorithm which is much faster than existing methods. We show that our estimator is consistent and semiparametrically efficient and establish its asymptotic distribution. Simulation studies demonstrate the superior performance of the proposed method, which is used to estimate the distribution of the age-at-onset of Parkinson's disease for carriers of mutations in the leucine-rich repeat kinase 2, LRRK2, gene.

Some key words: Empirical process; Mixture distribution; Parkinson's disease; Semiparametric efficiency; Sieve maximum likelihood estimation.

1. INTRODUCTION

The identification of causal genes for many genetic disorders has made personalized risk assessment and prediction of disease onset a real possibility. However, although interest lies in estimating the cumulative risk distributions of disease onset for individuals who are carriers of deleterious mutations or who have a certain haplotype, investigators may encounter missing genotypes or phase information of the haplotypes in a large proportion of individuals. For instance, genotypes of family members may be missing because of the high cost of collecting blood samples from relatives, the death of a relative (Wacholder et al., 1998; Marder et al., 2003; Zhang et al., 2010; Wang et al., 2012; Qin et al., 2014), or limitations in the technology to separate two homologous chromosomes in genotyping. Furthermore, disease onset information is subject to censoring caused by loss to follow-up or death.

When some genotype information is missing, the statistical framework for estimating disease risk distributions associated with genetic mutations is essentially the framework for analysis of censored mixture data. There is a large body of literature on inference for mixture models; see, for example, Titterington et al. (1985) and Mclachlan & Basford (1988) for parametric models, and Hall & Zhou (2003) for nonparametric models. Most of these papers address noncensored outcomes. Many genetic epidemiological studies of disease risk distributions have two features that distinguish them from other censored mixture models. First, each subgroup in the mixture model is biologically meaningful and corresponds to either carriers or noncarriers of a mutation; second, the mixing probability is usually known to investigators or can be inferred from family pedigrees and other external sources. For example, in a case-control genetic study with valid family history information on relatives (Marder et al., 2003), the probability of a relative having a certain genotype is obtained through the relationship between relatives and probands under Mendelian assumptions (Wacholder et al., 1998; Zhang et al., 2010; Wang et al., 2012; Qin et al., 2014). In haplotype studies, the probability of a certain haplotype can be inferred from unphased genotypes under the Hardy–Weinberg equilibrium (Zeng et al., 2006), based on information from external sources such as the HapMap project or from sequencing data (Yang et al., 2013).

In this paper, one application of our proposed method is to a recent study on age-specific risk of Parkinson's disease associated with mutations in the leucine-rich repeat kinase 2, LRRK2, gene (Paisán-Ruíz et al., 2004; Healy et al., 2008). Although Parkinson's disease is traditionally considered a nongenetic disorder, recent studies have identified genetic risk factors for Parkinson's disease, especially in more genetically homogeneous subpopulations such as Ashkenazi Jews (Trinh & Farrer, 2013). The goal of the current study is to estimate age-specific risk of Parkinson's disease in Ashkenazi Jews for LRRK2 gene mutation carriers and compare the distribution with that for noncarriers. Since LRRK2 mutations have low prevalence, it is not efficient to randomly sample individuals from the Ashkenazi Jewish population. Instead, the study used the kin-cohort design (Wacholder et al., 1998), which was initially implemented to study genetic risks of breast cancer. In our study, an initial sample of individuals with Parkinson's disease, i.e., the probands, were sequenced for the LRRK2 mutations and provided age-at-onset information for their first-degree relatives. Most of the relatives were not genotyped due to limited resources and therefore had unknown LRRK2 mutation status. In addition, in the case of older relatives who were deceased, it was not possible to collect blood samples.

Several existing works consider the estimation of distribution functions for such mixture data in a parametric or semiparametric framework (e.g., Diao & Lin, 2005; Zhang et al., 2010). When concerns over model misspecification arise in practice (e.g., Langbehn et al., 2004), use of a nonparametric model and inference through nonparametric maximum likelihood estimation are natural. However, although the Kaplan–Meier estimator is nonparametrically efficient for censored data, in general nonparametric maximum likelihood estimators are either inconsistent or inefficient for mixture data (Wang et al., 2012). To account for censoring and the mixture nature of the problem, while ensuring monotonicity of the estimated distribution function over the entire support, Qin et al. (2014) proposed methods based on a binomial likelihood and a sequence of nonparametric estimates performed by reducing censored data to current status data and implementing the expectation-maximization algorithm (Laird & Ware, 1982) along with the pooled-adjacent-violators algorithm. However, this approach is not guaranteed to be efficient and can be computationally intensive. Other works involving a nonparametric model based on estimating equations and weighting of Kaplan–Meier survival curves include Wacholder et al. (1998) and Fine et al. (2004).

In this paper, we propose a sieve maximum likelihood estimation method to estimate disease risk associated with genetic mutations in censored mixture models. Specifically, we utilize sieve estimation based on B-splines to estimate the log hazard ratios between carriers and noncarriers, while the nonparametric maximum likelihood estimator is used to estimate the reference baseline hazard function.

The estimators we derive for the disease risk distributions are asymptotically efficient. Furthermore, the calculation of the sieve maximum likelihood estimators can be easily implemented via an expectation-maximization algorithm which converges much faster than existing algorithms, due to the existence of closed-form solutions in the M-step. We tackle the theoretical challenge that arises when one functional parameter is estimated using a nonparametric maximum likelihood estimator while the other is estimated using a sieve estimator.

By simulation, we demonstrate substantial efficiency gains of the proposed method. Finally, as a real-data application, we use the method to estimate the age-at-onset of Parkinson's disease for individuals with deleterious LRRK2 mutations (Goldwurm et al., 2011).

2. Method and inference procedure

2·1. *Data and likelihood function*

Let T_i be the age-at-onset of a disease which is subject to random censoring. Let B_i denote the potentially missing mutation status, with 1 indicating the carrier group where each individual has at least one copy of the mutation and 2 indicating the noncarrier group. As in the Parkinson's disease study described in § 1, the probability of being a carrier takes a finite number of values. For example, a child of a heterozygote carrier parent has a probability of 0·5 of carrying this mutation under the Mendelian assumption, so if the mutation prevalence in the general population is f, we have $\text{pr}(B = 1) = 0{\cdot}5(1 + f)$ for this child. For individuals with observed carrier status, $\text{pr}(B = 1)$ equals 1 for carriers and 0 for noncarriers. We denote the finite set of values for the probability $\text{pr}(B = 1)$ by $\{p_1, \ldots, p_m\}$. Our goal is to estimate the risk distribution of the age-at-onset in the mutation group and in the no-mutation group, that is, $F_1(t) = \text{pr}(T \leqslant t \mid B = 1)$ and $F_2(t) = \text{pr}(T \leqslant t \mid B = 2)$, respectively.

Due to right censoring, the observations from n individuals consist of $\{Y_i = T_i \wedge C_i,\ \Delta_i = I(T_i \leqslant C_i),\ \text{pr}(B_i = 1)\}_{i=1,\ldots,n}$, where C_i denotes the censoring time, assumed to be independent of T_i. We introduce an indicator variable G_i to represent m distinct mixing probabilities, so $G_i = g$ indicates $\text{pr}(B_i = 1) = p_g$ $(g = 1, \ldots, m)$. After grouping individuals with the same p_g value together, the likelihood function can be written as

$$\prod_{i=1}^{n}\prod_{g=1}^{m}\left[\{p_g f_1(Y_i) + (1 - p_g) f_2(Y_i)\}^{\Delta_i}\{1 - p_g F_1(Y_i) - (1 - p_g) F_2(Y_i)\}^{1-\Delta_i}\right]^{I(G_i=g)},$$

where f_k is the density function corresponding to F_k $(k = 1, 2)$. Our goal is to estimate F_k.

In survival analysis, it is usually more convenient to rewrite the observed likelihood function using hazard functions instead of distribution functions. Let $\lambda_k(t)$ be the hazard function for T in the group with $B = k$, and let $\Lambda_k(t)$ be the corresponding cumulative hazard function. Then the likelihood function can be re-expressed as

$$\prod_{i=1}^{n}\prod_{g=1}^{m}\Big\{\big[p_g\lambda_1(Y_i)\exp\{-\Lambda_1(Y_i)\} + (1 - p_g)\lambda_2(Y_i)\exp\{-\Lambda_2(Y_i)\}\big]^{\Delta_i}$$
$$\times\big(1 - p_g[1 - \exp\{-\Lambda_1(Y_i)\}] - (1 - p_g)[1 - \exp\{-\Lambda_2(Y_i)\}]\big)^{1-\Delta_i}\Big\}^{I(G_i=g)}. \quad (1)$$

The goal is to maximize the likelihood function (1) in order to estimate $\Lambda_1(t)$ and $\Lambda_2(t)$ nonparametrically and hence obtain the age-at-onset distributions, $F_1(t)$ and $F_2(t)$. In the likelihood function (1), p_g equals 1 or 0 if an individual is observed to be a carrier or noncarrier, respectively.

2·2. *Sieve maximum likelihood estimation*

At first glance, to estimate Λ_1 or Λ_2 in (1), one could consider a nonparametric maximum likelihood estimator (Zeng & Lin, 2010), where Λ_1 or Λ_2 is treated as a step function with jumps at the observed event times. However, because the support points are ambiguous for event times when the mutation group membership is not observed, the nonparametric maximum likelihood estimator may not be consistent; in fact, bias was observed in simulations of it even for very large samples (Ma & Wang, 2012; Wang et al., 2012). We therefore propose a hybrid approach involving a nonparametric estimator and sieve maximum likelihood estimators that leads to consistent and semiparametrically efficient estimation.

Define $\beta(t) = \log\{\lambda_1(t)/\lambda_2(t)\}$, so $\Lambda_1(t) = \int_0^t \exp\{\beta(s)\}\, \mathrm{d}\Lambda_2(s)$. The likelihood in (1) can be re-expressed as

$$\prod_{i=1}^{n}\prod_{g=1}^{m}\Biggl\{\lambda_2(Y_i)^{\Delta_i}\left(p_g \exp\{\beta(Y_i)\}\exp\left[-\int_0^{Y_i}\exp\{\beta(t)\}\,\mathrm{d}\Lambda_2(t)\right] + (1-p_g)\exp\{-\Lambda_2(Y_i)\}\right)^{\Delta_i} \\ \times\left(p_g \exp\left[-\int_0^{Y_i}\exp\{\beta(t)\}\,\mathrm{d}\Lambda_2(t)\right] + (1-p_g)\exp\{-\Lambda_2(Y_i)\}\right)^{1-\Delta_i}\Biggr\}^{I(G_i=g)}. \quad (2)$$

To maximize (2), we consider using a nonparametric maximum likelihood estimator to estimate the cumulative hazard function in the baseline group, say $\Lambda_2(t)$, but adopting a sieve approximation to estimate $\beta(t)$. Specifically, we assume that Λ_2 jumps at the observed Y_i values with $\Delta_i = 1$, and we use a sieve approximation for the log hazard ratio $\beta(t)$, letting $\beta(t) = \sum_{j=1}^{K_n} \alpha_j \phi_j(t)$ where $\phi_1, \ldots, \phi_{K_n}$ are basis functions for the sieve approximation. The resulting estimator maximizes a partially smoothed likelihood, where the smoothing is performed on the hazard ratio function. The use of a smoothed approximation enables one to borrow information to estimate $\Lambda_1(t)$ and thus avoid specifying its ambiguous support points as required for the nonparametric maximum likelihood estimator. In our implementation, we choose B-splines as the basis functions: we let the spline knots be $0 = t_1 = \cdots = t_l < t_{l+1} < \cdots < \tau = t_{m_n+l} = t_{m_n+l+1} = \cdots = t_{m_n+2l}$, where τ is the study duration, m_n is an integer to be chosen in a data-driven fashion, and l is the order of the B-splines. There are $K_n = m_n + l$ B-spline basis functions in total, denoted by $\{\phi_j : j = 1, \ldots, K_n\}$.

Using the nonparametric maximum likelihood estimator for Λ_2 and the sieve estimate for $\beta(t)$, we aim to maximize (2) or its logarithm over all the parameters, including the jumps of Λ_2 and the spline coefficients $\alpha_1, \ldots, \alpha_{K_n}$. Direct maximization is computationally intensive and inefficient, since the loglikelihood is not convex and the parameters include the potentially large number of jumps of Λ_2. However, using the expectation-maximization algorithm with $B_1, \ldots, B_n$, the mutation statuses of all individuals, treated as missing data, fast numerical convergence can be obtained by virtue of various closed-form solutions in the M-step.

Assuming that the B_i were observed, the complete-data loglikelihood function for $(Y_i, \Delta_i, B_i, G_i)$ $(i = 1, \ldots, n)$ is

$$\sum_{i=1}^{n} I(B_i = 1)\left[\Delta_i \log \delta\Lambda_2(Y_i) + \Delta_i \sum_{j=1}^{K_n}\alpha_j\phi_j(Y_i) - \sum_{Y_k \leqslant Y_i}\delta\Lambda_2(Y_k)\exp\left\{\sum_{j=1}^{K_n}\alpha_j\phi_j(Y_k)\right\}\right] \\ + \sum_{i=1}^{n} I(B_i = 2)\bigl\{\Delta_i \log\delta\Lambda_2(Y_i) - \Lambda_2(Y_i)\bigr\} + \sum_{i=1}^{n}\sum_{g=1}^{m} I(G_i = g, B_i = 1)\log p_g \\ + \sum_{i=1}^{n}\sum_{g=1}^{m} I(G_i = g, B_i = 2)\log(1-p_g),$$

where $\delta\Lambda_2(y)$ denotes the jump of Λ_2 at y. Therefore, the expectation-maximization algorithm consists of the following E- and M-steps. In the E-step, we evaluate the conditional probability of $B_i = 1$ given the data (G_i, Y_i, Δ_i),

$$\frac{p_{G_i} \exp\big[\Delta_i \sum_{j=1}^{K_n} \alpha_j \phi_j(Y_i) - \int_0^{Y_i} \exp\{\sum_{j=1}^{K_n} \alpha_j \phi_j(t)\}\, \mathrm{d}\Lambda_2(t)\big]}{p_{G_i} \exp\big[\Delta_i \sum_{j=1}^{n} \alpha_j \phi_j(Y_i) - \int_0^{Y_i} \exp\{\sum_{j=1}^{K_n} \alpha_j \phi_j(t)\}\, \mathrm{d}\Lambda_2(t)\big] + (1 - p_{G_i}) \exp\{-\Lambda_2(Y_i)\}},$$

which we denote by q_i. Then, in the M-step, we maximize

$$\sum_{i=1}^{n} q_i \left[\Delta_i \log \delta\Lambda_2(Y_i) + \Delta_i \sum_{j=1}^{K_n} \alpha_j \phi_j(Y_i) - \sum_{Y_k \leqslant Y_i} \delta\Lambda_2(Y_k) \exp\left\{\sum_{j=1}^{K_n} \alpha_j \phi_j(Y_k)\right\}\right] + \sum_{i=1}^{n} (1 - q_i)\big\{\Delta_i \log \delta\Lambda_2(Y_i) - \Lambda_2(Y_i)\big\}. \qquad (3)$$

By differentiating (3) with respect to the jumps of Λ_2, we obtain a closed-form solution

$$\delta\Lambda_2(Y_i) = \Delta_i \Bigg/ \sum_{k=1}^{n} I(Y_k \geqslant Y_i) \left[q_k \exp\left\{\sum_{j=1}^{K_n} \alpha_j \phi_j(Y_i)\right\} + (1 - q_k)\right]. \qquad (4)$$

Upon substituting (4) into (3) and differentiating with respect to the α_j, we obtain values of α_j that satisfy the estimating equation

$$\sum_{i=1}^{n} \Delta_i \left(q_i - \frac{\sum_{k=1}^{n} I(Y_k \geqslant Y_i) q_k \exp\{\sum_{j=1}^{K_n} \alpha_j \phi_j(Y_i)\}}{\sum_{k=1}^{n} I(Y_k \geqslant Y_i)\big[q_k \exp\{\sum_{j=1}^{K_n} \alpha_j \phi_j(Y_i)\} + (1 - q_k)\big]}\right) \begin{pmatrix} \phi_1(Y_i) \\ \vdots \\ \phi_{K_n}(Y_i) \end{pmatrix} = 0, \qquad (5)$$

which is easily solved using the Newton–Raphson method. With the updated α values, we use (4) to update the jumps of $\Lambda_2(\cdot)$. We iterate between the E- and M-steps until convergence. The final estimators are denoted by $\hat{\Lambda}_{2n}(t)$ and $\hat{\beta}_n(t) = \sum_{j=1}^{K_n} \hat{\alpha}_j \phi_j(t)$. Although we have chosen Λ_2 to be the baseline group for the nonparametric maximum likelihood estimation and used sieve estimation to obtain a time-dependent log hazard ratio of the first group to the second group, the procedure can also be done in reverse by treating Λ_1 as the baseline group. In the subsequent arguments, for ease of theoretical justification, we will employ this reversely estimated $\hat{\Lambda}_{1n}(t)$ as an estimator of $\Lambda_1(t)$, instead of using $\int_0^t \exp\{\hat{\beta}_n(s)\}\, \mathrm{d}\hat{\Lambda}_{2n}(s)$. Empirically, we find that these two estimators of Λ_1 are almost identical.

Our theoretical results show that $\hat{\Lambda}_{kn}(t)$ $(k = 1, 2)$ converges in distribution to a Gaussian process after normalization. To estimate its asymptotic variance, following Zeng & Lin (2010), one approach is to compute the observed information matrix for the jump sizes of $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ and then use the inverse of this matrix to estimate the asymptotic covariance of $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$. However, this approach may be numerically unstable as it involves the inversion of a potentially high-dimensional information matrix. Alternatively, bootstrapping can be used to estimate the asymptotic covariance of $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$. From our numerical experience, 100 bootstrap samples are usually sufficient. In our algorithm, Λ_2 is updated using the closed form in (4) and the α_j are obtained via the one-step Newton–Raphson solution to (5). Therefore, the computational burden is much less than for existing methods.

Finally, using the proposed nonparametric estimators for $F_k(t) \equiv 1 - \exp\{-\Lambda_k(t)\}$, i.e., $\hat{F}_{kn}(t) = 1 - \exp\{-\hat{\Lambda}_{kn}(t)\}$, we can construct a variety of test statistics to compare the carrier group and the noncarrier group. One test statistic is based on the Kolmogorov–Smirnov test $\mathcal{T}_n =$

$\sup_{t\in[0,\tau]} |\hat{F}_{1n}(t) - \hat{F}_{2n}(t)|$. When $\mathcal{T}_n > c_\alpha$, we reject the null hypothesis that there is no difference between the disease risk distributions of the two groups. Here, α is the significance level and c_α is the $(1-\alpha)$-quantile of the sampling distribution of $\mathcal{T}_n$ under permutations where the variables G_i are permuted. Other possible test statistics include $\mathcal{T}_n = \int_0^\tau \omega(t)|\hat{F}_{1n}(t) - \hat{F}_{2n}(t)|\,\mathrm{d}t$, where $\omega(t)$ is a user-defined weight function that may focus on a specific time range.

2·3. *Generalization to cure rate survival data*

The proposed method can be generalized to analyse cure rate survival data, in which some individuals are considered to be immune to the disease of interest. To this end, we introduce a binary indicator Z to represent cure status. We assume that $\mathrm{pr}(Z=1 \mid B=k)=r_k$ and that the disease risk function among the noncured population is $\mathrm{pr}(T \leqslant t \mid Z=0, B=k) = \tilde{F}_k(t) = 1-\exp\{-\tilde{\Lambda}_k(t)\}$ $(k=1,2)$. The observed data consist of $(Y_i, \Delta_i, G_i, \Delta_i Z_i)$ $(i=1,\ldots,n)$, where Δ_i indicates whether the individual is diseased or cured. In other words, for noncensored individuals, we observe some individuals, usually those who have not experienced disease after a certain age, to be cured; however, the cure status for the censored individuals is unknown. Hence, if we define $\Lambda_k(t) = -\log[r_k + (1-r_k)\exp\{-\tilde{\Lambda}_k(t)\}]$, then the observed likelihood function becomes

$$\prod_{i=1}^{n}\prod_{g=1}^{m}\biggl(\bigl[\lambda_1(Y_i)\exp\{-\Lambda_1(Y_i)\}p_g + \lambda_2(Y_i)\exp\{-\Lambda_2(Y_i)\}(1-p_g)\bigr]^{\Delta_i(1-Z_i)}$$

$$\times\bigl[\exp\{-\Lambda_1(Y_i)\}p_g + \exp\{-\Lambda_2(Y_i)\}(1-p_g)\bigr]^{1-\Delta_i}\biggr)^{I(G_i=g)}$$

$$\times\prod_{i=1}^{n}\prod_{g=1}^{m}\Bigl[\bigl\{(1-r_1)p_g + (1-r_2)(1-p_g)\bigr\}^{\Delta_i(1-Z_i)}\bigl\{r_1 p_g + r_2(1-p_g)\bigr\}^{\Delta_i Z_i}\Bigr]^{I(G_i=g)}. \tag{6}$$

We can estimate the cure rates r_k by maximizing the last part of expression (6), and we estimate $\Lambda_k(t)$ by maximizing the first part using the sieve method proposed in § 2·2. Finally, we estimate $\tilde{F}_k(t)$ via $\tilde{F}_k(t) = [1-\exp\{-\Lambda_k(t)\}]/(1-r_k)$ $(k=1,2)$.

3. Asymptotic results

Let λ_{k0} and Λ_{k0} be the true hazard rates and the cumulative hazard functions for group k $(k=1,2)$ in the setting of §§ 2·1 and 2·2. Then the true log hazard ratio is $\beta_0(t) = \log\{\lambda_{10}(t)/\lambda_{20}(t)\}$. We need the following conditions.

Condition 1. Both $\lambda_{10}(t)$ and $\lambda_{20}(t)$ are r times continuously differentiable on $[0,\tau]$, where $r \geqslant 2$. In addition, there exist g_1 and g_2 such that $p_{g_1}/p_{g_2} \mp (1-p_{g_1})/(1-p_{g_2})$.

Condition 2. The density of C has bounded and continuous rth derivative in $[0,\tau]$, and C is independent of T conditional on G.

Condition 3. The number of interior knots m_n satisfies $m_n^{3/2}/n^{1/2} = O(1)$ and $n^{1/2}/m_n^{2r} \to 0$ as $n \to \infty$.

Conditions 1 and 2 are regularity conditions for the underlying density functions of T in both groups. The second part of Condition 1 ensures that the data contain at least two distinct kinds of p_g to guarantee identifiability of the underlying distributions. In Condition 3, one particular choice for the number of interior knots is $m_n = n^v$, where $1/(4r) < v \leqslant 1/3$. Under these conditions, our first theorem gives the uniform consistency of $\hat{\Lambda}_{1n}$ and $\hat{\Lambda}_{2n}$ in $[0, \tau]$.

THEOREM 1. *Under Conditions* 1–3 *and in the setting of* § § 2·1 *and* 2·2,

$$\sup_{t\in[0,\tau]} \left|\hat{\Lambda}_{1n}(t) - \Lambda_{10}(t)\right| + \sup_{t\in[0,\tau]} \left|\hat{\Lambda}_{2n}(t) - \Lambda_{20}(t)\right| = o_{\mathrm{p}}(1), \quad n \to \infty.$$

To describe the asymptotic distributions of $\hat{\Lambda}_{1n}$ and $\hat{\Lambda}_{2n}$, we first introduce the sets

$$\mathcal{F}_{\mathrm{BV}} = \big\{f(t) : f(t) \text{ has total variance bounded by 1 in } [0, \tau]\big\},$$
$$\mathcal{F}_{\beta} = \big\{g(t) : g(t) \text{ has } r\text{th derivative bounded by 1 in } [0, \tau]\big\}.$$

We then treat both $\hat{\Lambda}_{1n}$ and $\hat{\Lambda}_{2n}$ as bounded stochastic processes in $\mathcal{F}_{\mathrm{BV}}$ by defining $\hat{\Lambda}_{kn}(f) = \int_0^\tau f(s)\,\mathrm{d}\hat{\Lambda}_{kn}(s)$ $(k = 1, 2)$ for $f \in \mathcal{F}_{\mathrm{BV}}$. Similarly, we treat $\hat{\beta}_n(t)$ as a stochastic process on $\mathcal{F}_\beta$ by letting $\hat{\beta}_n(g) = \int_0^\tau g(s)\hat{\beta}_n(s)\,\mathrm{d}s$ for $g \in \mathcal{F}_\beta$. The following theorem shows the weak convergence of these stochastic processes.

THEOREM 2. *Consider* $\{\hat{\Lambda}_{1n}(t) - \Lambda_{10}(t),\ \hat{\Lambda}_{2n}(t) - \Lambda_{20}(t)\}$ *as a stochastic process in* $l^\infty(\mathcal{F}_{\mathrm{BV}} \times \mathcal{F}_{\mathrm{BV}})$. *Then, under Conditions* 1–3 *and in the setting of* § § 2·1 *and* 2·2, $n^{1/2}\{\hat{\Lambda}_{1n}(t) - \Lambda_{10}(t),\ \hat{\Lambda}_{2n}(t) - \Lambda_{20}(t)\}$ *converges in distribution to a zero-mean Gaussian process in* $l^\infty(\mathcal{F}_{\mathrm{BV}} \times \mathcal{F}_{\mathrm{BV}})$ *as* $n \to \infty$. *Furthermore,* $\hat{\Lambda}_{1n}$ *and* $\hat{\Lambda}_{2n}$ *are semiparametrically efficient in the sense of the definition in Bickel et al. (1993). In addition, as a stochastic process in* $l^\infty(\mathcal{F}_\beta)$, $n^{1/2}(\hat{\beta}_n - \beta_0)$ *converges in distribution to a zero-mean Gaussian process as* $n \to \infty$.

Remark 1. Theorem 2 establishes that $n^{1/2}\{\hat{\Lambda}_{1n}(t) - \Lambda_{10}(t)\}$ and $n^{1/2}\{\hat{\Lambda}_{2n}(t) - \Lambda_{20}(t)\}$ converge in distribution to some Gaussian process in $l^\infty([0, \tau])$. By the delta method, this also holds for the corresponding distribution function estimators, $\hat{F}_{1n}(t) = 1 - \exp\{-\hat{\Lambda}_{1n}(t)\}$ and $\hat{F}_{2n}(t) = 1 - \exp\{-\hat{\Lambda}_{2n}(t)\}$. Thus the sieve nonparametric maximum likelihood estimators $\hat{F}_{1n}$ and $\hat{F}_{2n}$ attain the semiparametric efficiency bound and are optimal for the censored mixture data.

Here, semiparametric efficiency is defined in the sense of Bickel et al. (1993, ch. 6). Theorem 2 shows that $\hat{F}_k$, as a function estimator in BV$[0, \tau]$, the set of bounded-variation functions on $[0, \tau]$, is semiparametrically efficient, which means that any bounded linear functional of $\hat{F}_k$ achieves its efficiency bound asymptotically. The weak convergence in Theorem 2 ensures that we can construct a valid confidence band based on these estimators. The proofs of Theorems 1 and 2 are given in the Appendix. The main technical challenge is to handle the mixed convergence rates of the infinite-dimensional parameter estimators, since $\hat{\Lambda}_{kn}$ has a $n^{1/2}$-convergence rate while $\hat{\beta}_n(t)$ has a slower convergence rate. In the proof of Theorem 2, with the derived rates for $\hat{\Lambda}_{kn}$ and $\hat{\beta}_n(t)$ under some suitable norms, the master Z-theorem in van der Vaart & Wellner (1996, § 3.3) is used to derive the asymptotic distributions of the estimators. Theorems 1 and 2 hold for the estimators using the cure rate survival data, because of the similar likelihood function used in the estimation. Although the proposed method is fully efficient based on the assumption of independent T_i given the mutation status B_i, it can easily be generalized to correlated family data by maximizing

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i}\prod_{g=1}^{m}\Big[\big\{p_g f_1(Y_{ij}) + (1 - p_g) f_2(Y_{ij})\big\}^{\Delta_{ij}}\big\{1 - p_g F_1(Y_{ij}) - (1 - p_g) F_2(Y_{ij})\big\}^{1-\Delta_{ij}}\Big]^{I(G_{ij}=g)},$$

where i indicates the family and j indicates an individual in the family. In this case, the proposed inference procedure including the expectation-maximization algorithm and bootstrap over independent families is still valid, and Theorems 1 and 2 hold except that the estimators may not achieve the semiparametric efficiency bound due to the maximization of a marginal likelihood.

4. Simulation studies

Extensive simulation studies were conducted to compare the small-sample performances of the proposed and existing methods. Our first study used the same distribution functions as in Qin et al. (2014). Specifically, for the carriers, $F_1(t) = \{1 - \exp(-t)\}/\{1 - \exp(-10)\}$, while for the noncarriers, $F_2(t) = \{1 - \exp(-t/2{\cdot}8)\}/\{1 - \exp(-10/2{\cdot}8)\}$ for $0 \leqslant t \leqslant 10$. The mutation probability p_i was randomly chosen from either of the following sets: Case I, $(1, 0{\cdot}6, 0{\cdot}2, 0{\cdot}16)$; Case II, $(0{\cdot}75, 0{\cdot}6, 0{\cdot}5, 0{\cdot}16)$. The censoring time followed a uniform distribution to yield a censoring rate of 20% or 40%. In the second simulation study, we imitated the results from the Parkinson's disease study described in § 5: we generated survival times for carriers and noncarriers using distributions similar to the estimated distributions of the actual data, $F_1 = \text{Wei}(5{\cdot}0, 102)$ and $F_2 = \text{Wei}(5{\cdot}0, 125)$. Furthermore, the sample size was taken to be $n = 2275$ and the mutation probability p_i was taken from $\{0, 0{\cdot}02, 0{\cdot}51, 1\}$, as in the real example. The censoring times were generated from a uniform distribution to achieve a censoring rate of 40% or 80%.

When implementing our method, we used cubic B-spline functions to estimate $\beta(t)$. The number of knots was set at $m_n = \lfloor n^{1/3} \rfloor - 1$ and the location of each interior knot was selected to be evenly distributed at the quantiles of the observed failure times. Some neighbouring knots were combined if the data were found to be too sparse to stably estimate the coefficient of a particular basis function. We also experimented with taking the number of interior knots to be $m_n/2$ or $2m_n$, but the estimates for $\Lambda_1(t)$ and $\Lambda_2(t)$ turned out to vary very little. To avoid local maximization in the expectation-maximization algorithm, we used different initial estimators, including estimates from a published method such as that of Qin et al. (2014); empirically, our algorithm converged to the same results. We used 500 bootstrap samples for variance estimation. Furthermore, we compared our method with the estimator in Qin et al. (2014), which sequentially censors the observed event times to construct a binomial likelihood and applies the pooled-adjacent-violators algorithm for estimation.

The simulation results from 500 replicates for the first scenario are given in Table 1. We present the average estimated values of the cumulative distribution functions F_1 and F_2 at various quartiles. Table 1 suggests that both the sieve estimator and the method of Qin et al. (2014) have small bias, the variance estimate based on the bootstrap agrees adequately with the empirical variability, and the coverage probabilities are close to the nominal level. The sieve estimator is more efficient than the method of Qin et al. (2014) in all simulation settings, and the efficiency gain, which can be as large as 60%, is more evident for the upper quartiles and for the higher censoring rate. A similar advantage of the sieve estimators can be seen in Table 2 for the second simulation scenario. Our method performs well even under 80% censoring; the efficiency gain is up to 15%. In the Supplementary Material, we report root integrated mean squared errors and the average of the pointwise variance for the estimators of Λ_1 and Λ_2. Our estimators for the Λ_i have smaller estimation errors than those of Qin et al. (2014), especially in the case of Λ_1.

We performed two additional simulations, simulations 3 and 4, with crossed distributions; the results are reported in the Supplementary Material. The findings are similar. Finally, we conducted simulation studies to evaluate the permutation test for the Kolmogorov–Smirnov statistic comparing the two distributions. The data generation was similar to that in the second simulation study, except that $F_1 = F_2 = \text{Wei}(5{\cdot}0, 102)$. The empirical Type I error rate was $4{\cdot}6\%$ with

Table 1. *Summary results for the estimated distribution functions in the first simulation scenario* $(\times 10^{-2})$

			Case I							Case II						
			Proposed				EM-PAVA			Proposed				EM-PAVA		
n	$c\%$		Bias	SD	SE	CP	Bias	SD	Ratio	Bias	SD	SE	CP	Bias	SD	Ratio
100	20%	$F_1(Q_{0\cdot25})$	−0·8	7·2	7·1	93	−0·4	7·4	105·5	−1·3	8·9	7·9	90	−0·1	9·5	112·2
		$F_1(Q_{0\cdot50})$	−0·8	9·0	8·9	93	−0·2	9·2	104·6	−2·7	10·9	10·4	92	0·0	11·6	114·0
		$F_1(Q_{0\cdot75})$	−0·8	8·0	8·1	94	−0·1	8·4	108·5	−3·2	10·4	10·1	92	0·6	11·8	130·0
		$F_2(Q_{0\cdot25})$	0·6	7·9	8·1	94	0·4	8·1	104·7	2·8	10·9	10·9	93	0·4	11·0	101·0
		$F_2(Q_{0\cdot50})$	0·4	9·0	8·8	94	0·3	9·2	103·5	3·8	10·6	10·9	94	0·3	12·1	128·8
		$F_2(Q_{0\cdot75})$	−0·5	7·7	7·3	94	−0·3	8·0	108·7	2·8	8·4	7·8	89	1·9	9·5	130·0
	40%	$F_1(Q_{0\cdot25})$	−0·6	7·8	7·2	92	−0·2	7·9	103·1	−1·1	8·2	7·7	92	0·0	8·6	110·6
		$F_1(Q_{0\cdot50})$	−0·7	9·0	9·1	94	−0·1	9·3	106·7	−2·9	10·6	10·3	92	−0·4	11·8	123·4
		$F_1(Q_{0\cdot75})$	−1·1	9·2	8·8	92	−0·2	9·6	109·7	−4·8	11·3	10·4	90	−1·0	13·0	130·8
		$F_2(Q_{0\cdot25})$	0·0	8·4	8·3	92	−0·1	8·6	102·7	2·4	11·2	10·7	91	0·0	12·1	115·4
		$F_2(Q_{0\cdot50})$	0·3	10·0	9·7	94	0·3	10·1	102·1	4·1	12·2	11·7	91	1·3	13·7	125·2
		$F_2(Q_{0\cdot75})$	−2·4	12·0	10·2	88	2·2	14·2	140·8	0·6	11·3	10·7	88	7·0	14·3	160·1
300	20%	$F_1(Q_{0\cdot25})$	−0·5	4·3	4·3	94	−0·4	4·4	102·3	−0·8	5·3	5·3	94	−0·3	5·4	105·7
		$F_1(Q_{0\cdot50})$	−0·1	5·3	5·2	94	0·0	5·3	103·3	−1·7	6·6	6·8	95	−0·7	6·8	108·6
		$F_1(Q_{0\cdot75})$	−0·4	4·8	4·7	95	−0·2	5·0	110·7	−1·9	6·5	6·7	94	−0·2	7·2	122·8
		$F_2(Q_{0\cdot25})$	−0·1	4·7	4·8	95	−0·1	4·7	103·7	1·5	6·9	6·9	95	0·6	6·9	100·2
		$F_2(Q_{0\cdot50})$	0·0	5·0	5·1	96	0·0	5·0	103·6	1·9	7·2	7·0	95	0·4	7·7	113·1
		$F_2(Q_{0\cdot75})$	0·0	4·5	4·4	95	0·0	4·5	103·0	2·4	5·4	5·2	91	1·4	5·7	112·1
	40%	$F_1(Q_{0\cdot25})$	0·1	4·5	4·4	95	0·2	4·5	100·7	−0·6	5·2	5·2	94	0·1	5·4	106·0
		$F_1(Q_{0\cdot50})$	0·3	5·4	5·4	96	0·3	5·4	101·5	−1·6	6·8	7·0	95	−0·5	7·1	108·3
		$F_1(Q_{0\cdot75})$	0·1	5·0	5·1	96	0·1	5·1	107·0	−2·4	6·9	7·1	94	−0·7	7·6	121·7
		$F_2(Q_{0\cdot25})$	−0·4	4·9	5·0	95	−0·4	4·9	101·0	1·7	6·9	7·2	95	0·6	6·9	100·5
		$F_2(Q_{0\cdot50})$	−0·3	5·7	5·8	95	−0·3	5·9	104·1	2·6	7·5	7·7	94	0·7	8·0	114·2
		$F_2(Q_{0\cdot75})$	−0·7	7·5	7·1	92	0·5	8·3	124·6	1·3	8·4	7·8	91	3·0	10·3	150·7

EM-PAVA, the method of Qin et al. (2014); $Q_{0\cdot25}$, $Q_{0\cdot5}$ and $Q_{0\cdot75}$, the first, second and third quartiles, respectively, of F_1 and F_2; $c\%$, censoring rate; Bias, average estimation bias over 500 replications; SD, empirical standard deviation; SE, average of the estimated standard errors from bootstraps; CP, actual coverage probability corresponding to nominal 95% confidence intervals; Ratio, relative efficiency ratio between the proposed method and the method of Qin et al. (2014).

Table 2. *Summary results for the estimated distribution functions in the second simulation scenario* $(\times 10^{-2})$

		Proposed				EM-PAVA		
Censoring		Bias	SD	SE	CP	Bias	SD	Ratio
40%	$F_1(Q_{0\cdot25})$	−0·1	3·1	3·2	95	0·0	3·3	110·3
	$F_1(Q_{0\cdot50})$	−0·1	3·8	4·1	96	0·0	3·8	103·3
	$F_1(Q_{0\cdot75})$	−0·4	3·7	4·1	96	0·0	4·0	115·1
	$F_2(Q_{0\cdot25})$	0·1	1·3	1·3	94	0·1	1·3	103·2
	$F_2(Q_{0\cdot50})$	0·1	1·6	1·5	94	0·0	1·6	100·5
	$F_2(Q_{0\cdot75})$	0·2	1·4	1·3	93	0·0	1·4	103·9
80%	$F_1(Q_{0\cdot25})$	−0·4	4·2	4·1	94	−0·3	4·3	102·5
	$F_1(Q_{0\cdot50})$	−0·7	5·4	5·8	94	−0·4	5·6	104·7
	$F_1(Q_{0\cdot75})$	−1·1	6·0	6·5	95	−0·3	6·4	112·5
	$F_2(Q_{0\cdot25})$	0·1	1·8	1·8	95	0·0	1·8	100·8
	$F_2(Q_{0\cdot50})$	0·2	2·5	2·6	95	0·0	2·5	101·4
	$F_2(Q_{0\cdot75})$	−0·2	4·0	3·7	93	1·0	4·2	107·2

censoring rate 40% and 5·0% with censoring rate 80%. Both are close to the nominal significance level of 5%, so the proposed permutation test appears to be valid.

5. Application

Since mutations in the LRRK2 gene were found to be a potential cause of idiopathic Parkinson's disease (Paisán-Ruíz et al., 2004), there has been great interest in estimating the cumulative risk of Parkinson's disease for LRRK2 mutation carriers, especially among Ashkenazi Jews, who have an increased mutation rate (Alcalay et al., 2013). Although such estimates are important for genetic counselling (Goldwurm et al., 2011), results on the risk for LRRK2 carriers in the literature have been inconsistent and estimates vary widely (Goldwurm et al., 2011).

To address these concerns, we aim to estimate the age-specific cumulative risk of Parkinson's disease in LRRK2 carriers and noncarriers. Due to the low prevalence of LRRK2 mutations, a kin-cohort design was used (Marder et al., 2014). To avoid bias in the ascertainment of the initial samples, our units of analysis are the first-degree family members excluding the initial probands (e.g., Wacholder et al., 1998; Wang et al., 2012). Our initial probands were recruited from the Michael J. Fox Foundation Ashkenazi Jewish LRRK2 Consortium; the details of the sample have been reported elsewhere (Alcalay et al., 2013). All probands were screened for G2019S mutations in the LRRK2 gene and common mutations in the glucocerebrosidase gene. To isolate the effect of the LRRK2 mutations on Parkinson's disease risk, we excluded participants with other known genetic risk factors such as glucocerebrosidase mutations. A validated family history instrument (Marder et al., 2003) was applied to the probands or the first-degree relatives themselves if relatives were seen by a neurologist.

The data included information gathered from 2275 first-degree relatives of the probands in the Ashkenazi Jewish LRRK2 Consortium. There were four groups of mutation probabilities: $p_g \in \{0, 0{\cdot}02, 0{\cdot}51, 1\}$, with frequencies 1·6%, 70·9%, 25·4% and 2·1%, respectively. Only 3·7% of relatives had observed genotypes, i.e., their corresponding p_g is either 1 or 0. The first-degree relatives, including parents, siblings and children, of noncarrier probands have $p_g = 0{\cdot}02$ under a 2% population prevalence of LRRK2 in the Ashkenazi Jewish population (Orr-Urtreger et al., 2007) and the Mendelian assumption. Similarly, the first-degree relatives of heterozygote carrier probands have $p_g = 0{\cdot}51$ under the Mendelian assumption. The censoring rate was close to 95%. Due to the high censoring rate, we analysed the data under the cure rate model (6). Individuals who did not develop Parkinson's disease by age 95 were considered immune to the disease, since the greatest documented age-at-onset is 94 years (Driver et al., 2009). In the implementation of the proposed sieve maximum likelihood approach, we used the Bayesian information criterion to choose the number of interior knots and the degree of the B-spline basis. The choices minimizing this criterion were two interior knots and a degree of two. We used bootstrap resampling of families to construct pointwise confidence intervals to ensure valid inference.

In the practice of genetic counselling, it is more useful to provide the population cumulative risks, namely $F_k(t)$ in model (6), regardless of the cure survival status. Therefore we report estimates of $F_k(t)$ in Table 3. These results show that, for carriers, the cumulative risk of Parkinson's disease by age 80 can be as high as 27·4%, with 95% confidence interval 17·6–39·1%, whereas for noncarriers the cumulative risk is 10·4%, with 95% confidence interval 7·8–13·2%. The risk of Parkinson's disease in noncarriers is quite high compared to that in general non-Ashkenazi Jews, whose risk is normally 1%, indicating that this group may have other risk mutations for Parkinson's disease. The estimated lifetime cumulative risk is consistent with some previous findings for LRRK2 mutation carriers in Ashkenazi Jews (Wang et al., 2008), but it contrasts with some other studies, which estimate the risk of Parkinson's disease to be 100% in LRRK2 carriers (Lesage et al., 2005). Methodological issues, including assigning individuals with unobserved

Table 3. *Estimated cumulative risk of Parkinson's disease onset for LRRK2 carriers and noncarriers in the Ashkenazi Jewish LRRK2 Consortium study* ($\times 10^{-2}$)

	Carrier $F_1(\cdot)$			Noncarrier $F_2(\cdot)$		
Age	Risk	SE	95% CI	Risk	SE	95% CI
20	0·0	0·0	(0·0, 0·1)	0·1	0·1	(0·0, 0·3)
30	0·1	0·1	(0·0, 0·3)	0·1	0·1	(0·0, 0·3)
40	0·3	0·4	(0·0, 1·4)	0·2	0·1	(0·0, 0·4)
50	1·8	0·8	(0·5, 3·4)	0·6	0·2	(0·3, 1·1)
60	8·1	1·9	(4·8, 12·5)	2·8	0·6	(1·6, 4·1)
70	18·3	3·9	(11·2, 26·2)	6·8	1·1	(4·9, 9·0)
80	27·4	5·7	(17·6, 39·1)	10·4	1·4	(7·8, 13·2)

CI, confidence interval for the estimated risk value.

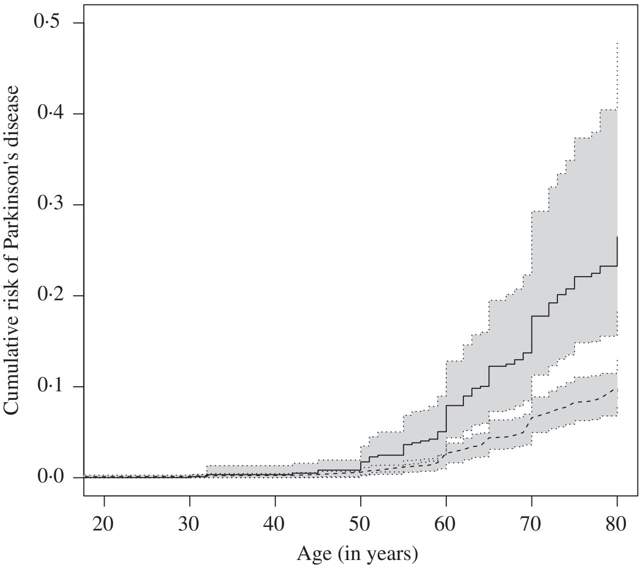

Fig. 1. Estimated cumulative risk functions for Parkinson's disease onset in LRRK2 carriers and noncarriers: the solid curve is the estimated distribution function for carriers and the dashed curve is for noncarriers; the dotted curves are their pointwise 95% confidence intervals, and the shaded regions represent the areas covered by the pointwise confidence intervals.

LRRK2 genotypes to carrier or noncarrier groups based on their Parkinson's disease status, may have contributed to this large discrepancy between the studies. Figure 1 presents the estimated cumulative Parkinson's disease distributions in the two mutation groups, together with their pointwise confidence intervals. The carrier group has a dramatic increase in the risk of developing Parkinson's disease after age 60, as compared to a slower increase in the noncarrier group.

To compare the distributions, we used the Kolmogorov–Smirnov test to examine the maximal difference between the two groups. We computed the p-value for this test based on 1000 permutations, where for each permutation the grouping variable G_i was perturbed. The resulting p-value is less than 0·001. It may be of practical interest to examine some classes of parametric models for the genetic risk functions; for example, within the class of Weibull distributions, we find that the estimated distribution for the carriers is adequately approximated by a Weibull distribution with shape and scale parameters 5 and 102, while the estimated distribution for the noncarriers is close to a Weibull with shape and scale parameters 5 and 125.

The cure rate in carriers was estimated to be 0·3% with 95% confidence interval 0–19·8%, and that for noncarriers was estimated to be 26·6% with 95% confidence interval 17·9–34·6%. There is a significant difference of 26·3% between the two estimated cure rates, with 95% confidence interval 3·6–34·3%. In the noncured population, the cumulative risk of Parkinson's disease for carriers by age 80 was 27·5%; that is, $\tilde{F}_1(t)$ as defined in § 2·3 was 27·5% at age 80, compared with 14·2% for the noncarrier group. The low cure rate in the carrier group suggests a high risk of developing Parkinson's disease if a subject lives long enough. This observation is consistent with what has been reported in the existing clinical literature. For example, Latourelle et al. (2008) reported a high lifetime risk of Parkinson's disease, where the median risk of disease was about 70% and the upper limit of the 95% confidence interval was about 80%.

6. Discussion

One interesting problem is how to handle the different convergence rates of the nonparametric maximum likelihood estimator and the sieve estimators based on B-splines. Alternatively, sieve estimation can also be applied to Λ_2, as was done by Cheng & Wang (2011) for a semiparametric additive transformation model with current status data. One advantage of using the nonparametric maximum likelihood estimator for Λ_2 is that there is no need to determine the number of sieves. Moreover, our nonparametric maximum likelihood estimator has an explicit solution in the M-step of the expectation-maximization algorithm, which leads to computational gain.

Using the reparameterized likelihood function (2), the proposed method can be readily generalized to regression problems where other environmental covariates are included through a proportional hazards model in both groups (Diao & Lin, 2005). Lastly, to efficiently analyse family data, an alternative method using frailty models could be considered to account for within-family dependence through shared frailties.

Acknowledgement

This work was supported by the Michael J. Fox Foundation, the U.S. National Institutes of Health, the National Natural Science Foundation of China and the China Scholarship Council. We thank the editor, an associate editor and referees for comments that have helped to improve the paper.

Supplementary material

Supplementary material available at *Biometrika* online includes a proof of model identifiability, additional tables of results for simulations 1 and 2, and results from additional simulations.

Appendix

Before proving Theorems 1 and 2, we first show that the information operator for Λ_2 and β is invertible. For $G = g$, we define

$$A_1(\Delta, G, Y) = \Xi_1 \left((1-p_g)\exp\{-\Lambda_2(Y)\} + p_g \exp\{\beta(Y)\} \exp\left[-\int_0^Y \exp\{\beta(t)\}\, \mathrm{d}\Lambda_2(t) \right] \right),$$

$$A_2(\Delta, G, Y) = (1-p_g)\exp\{-\Lambda_2(Y)\}(\Xi_1 + \Xi_2),$$

$$A_3(\Delta, G, Y) = p_g \left[\exp\{\beta(Y)\}\Xi_1 + \Xi_2\right] \exp\left[-\int_0^Y \exp\{\beta(t)\}\, \mathrm{d}\Lambda_2(t) \right],$$

$$B_1(\Delta, G, Y) = p_g \exp\{\beta(Y)\} \exp\left[-\int_0^Y \exp\{\beta(t)\}\, \mathrm{d}\Lambda_2(t) \right] \Xi_1,$$

where

$$\Xi_1 = \Delta \left(p_g \exp\{\beta(Y)\} \exp\left[-\int_0^Y \exp\{\beta(t)\}\, \mathrm{d}\Lambda_2(t) \right] + (1-p_g)\exp\{-\Lambda_2(Y)\} \right)^{-1},$$

$$\Xi_2 = (1-\Delta) \left(p_g \exp\left[-\int_0^Y \exp\{\beta(t)\}\, \mathrm{d}\Lambda_2(t) \right] + (1-p_g)\exp\{-\Lambda_2(Y)\} \right)^{-1}.$$

The loglikelihood function for a single subject is

$$\begin{aligned} l(\Lambda_2,\beta) = \sum_{g=1}^{m} I(G=g) \Bigg\{ \Delta \log \Bigg(& p_g \lambda_2(Y) \exp\{\beta(Y)\} \exp\left[-\int_0^Y \exp\{\beta(t)\}\, \mathrm{d}\Lambda_2(t) \right] \\ & + (1-p_g)\lambda_2(Y)\exp\{-\Lambda_2(Y)\} \Bigg) \\ & + (1-\Delta)\log\left(p_g \exp\left[-\int_0^Y \exp\{\beta(t)\}\, \mathrm{d}\Lambda_2(t) \right] + (1-p_g)\exp\{-\Lambda_2(Y)\} \right) \Bigg\}. \end{aligned}$$

By differentiating $l(\Lambda_2,\beta)$ with respect to Λ_2 and β along submodels $\mathrm{d}\Lambda_2(1+\varepsilon h_1)$ and $\beta+\varepsilon h_2$, respectively, we obtain the score operators $l_{\Lambda_2}(\Lambda_2,\beta)(h_1) = A_1 h_1(Y) - \int_0^Y h_1(t)[A_2 + A_3\exp\{\beta(t)\}]\,\mathrm{d}\Lambda_2(t)$ and $l_\beta(\Lambda_2,\beta)(h_2) = B_1 h_2(Y) - A_3\int_0^Y h_2(t)\exp\{\beta(t)\}\,\mathrm{d}\Lambda_2(t)$. Thus, if we define $\langle f_1, f_2\rangle = E(f_1 f_2)$, then for any $L_2(P)$-integrable functions $\{w_1(\Delta,G,Y), w_2(\Delta,G,Y)\}$ we have

$$\begin{aligned} \left\langle \begin{pmatrix} l_{\Lambda_2}(h_1) \\ l_\beta(h_2) \end{pmatrix}, \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right\rangle &= \left\langle \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \begin{pmatrix} E(A_1 w_1 \mid Y) \\ E(B_1 w_2 \mid Y) \end{pmatrix} \right\rangle \\ &\quad + \int_0^Y \left\langle \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \begin{pmatrix} E\{(A_2+A_3)w_1 \mid Y=t\}\exp\{\beta(t)\} \\ E\{A_3 w_2 \mid Y=t\}\exp\{\beta(t)\} \end{pmatrix} \right\rangle \mathrm{d}\Lambda_2(t). \end{aligned}$$

Therefore,

$$\begin{aligned} l^*_{\Lambda_2} l_{\Lambda_2}(h_1) &= E(A_1^2 \mid Y)h_1(Y) - \int_0^Y E\big(A_1[A_2 + A_3\exp\{\beta(t)\}] \mid Y=t\big)h_1(t)\,\mathrm{d}\Lambda_2(t) \\ &\quad - \int_0^Y E\big\{(A_2+A_3)A_1 \mid Y=t\big\}h_1(t)\,\mathrm{d}\Lambda_2(t) \\ &\quad + \int_0^Y\int_0^t E\big(I(G=g)(A_2+A_3)[A_2+A_3\exp\{\beta(t)\}] \mid Y=s\big)\,\mathrm{d}\Lambda_2(s)h_1(t)\,\mathrm{d}\Lambda_2(t), \\ l^*_\beta l_\beta(h_2) &= E(B_1^2 \mid Y)h_2(Y) - \int_0^Y E\big[B_1A_3\exp\{\beta(t)\} \mid Y=t\big]h_2(t)\,\mathrm{d}\Lambda_2(t) \\ &\quad - \int_0^Y E\left\{ A_3 \sum_{g=1}^{m} I(G=g)B_1 \,\middle|\, Y=t \right\} h_2(t)\,\mathrm{d}\Lambda_2(t) \\ &\quad + \int_0^Y \left(\int_0^t E\left[A_3A_3\exp\{\beta(t)\} \,\middle|\, Y=s \right] \mathrm{d}\Lambda_2(s) \right) h_2(t)\,\mathrm{d}\Lambda_2(t), \end{aligned}$$

where $(l^*_{\Lambda_2}, l^*_\beta)$ is the dual operator of (l_{Λ_2}, l_β). Thus, the information operator $\mathcal{I}(\Lambda_2,\beta) = (l_{\Lambda_2}, l_\beta)^*(l_{\Lambda_2}, l_\beta)$ can be expressed as a Fredholm operator of the first kind, which is the summation of an invertible operator and an integral operator when $\Lambda_2 = \Lambda_{20}$ and $\beta = \beta_0$. Consequently, to show that $\mathcal{I}(\Lambda_{20},\beta_0)$ is invertible, following Rudin (1973) it suffices to show that $\mathcal{I}(\Lambda_{20},\beta_0)$ is one-to-one; that is, we need to prove that for any h_1 and h_2, if $\mathcal{I}(\Lambda_{20},\beta_0)(h_1,h_2) = 0$, which is equivalent to $l_{\Lambda_{20}}(h_1) + l_{\beta_0}(h_2) = 0$, then $h_1 \equiv 0$ and $h_2 \equiv 0$. Suppose that $l_{\Lambda_{20}}(h_1) + l_{\beta_0}(h_2) = 0$; let $\Delta = 1$ and $G = g$, and integrate Y from 0 to any $t \in [0,\tau]$. Then we obtain $\int_0^t [p_g\{h_1(s) + h_2(s)\}\exp\{\beta_0(s)\} +$

$(1-p_g)h_2(s)]\,\mathrm{d}\Lambda_{20}(s)=0$. Hence $p_g\{h_1(t)+h_2(t)\}\exp\{\beta_0(t)\}+(1-p_g)h_2(t)=0$. By Condition 1, we immediately conclude that $h_1=h_2\equiv 0$. Therefore $\mathcal{I}(\Lambda_{20},\beta_0)(h_1,h_2)$ is continuously invertible.

Next, we consider a different Banach space $\mathcal{H}^*=\{(h_1,h_2): h_1\in L_2[0,\tau],\ h_2\in L_2[0,\tau]\}$. Then the above arguments still hold. Hence, the invertibility of $\mathcal{I}(\Lambda_{20},\beta_0)$ implies that $\|\mathcal{I}(\Lambda_{20},\beta_0)\|_{L_2}^2\geqslant c(\|h_1\|_{L_2}^2+\|h_2\|_{L_2}^2)$, where c is a constant. Furthermore, if $\|\Lambda_2-\Lambda_{20}\|_\infty+\|\beta-\beta_0\|_\infty<\varepsilon_0$ for a small ε_0, the continuity of $\mathcal{I}$ in this space gives

$$\big\|\mathcal{I}(\Lambda_2,\beta)(h_1,h_2)\big\|_{L_2}^2\geqslant\frac{c}{2}\big(\|h_1\|_{L_2}^2+\|h_2\|_{L_2}^2\big).$$

We will use this fact in the following consistency proof.

Proof of Theorem 1. We define a sieve space

$$S_n=\bigg\{(\Lambda_2,\beta):\ \Lambda_2\text{ is a step function with jumps at the observed events,}\\ \beta(t)=\sum_{j=1}^{K_n}\alpha_j\phi_j(t)\text{ where the }\phi_j\text{ are B-spline bases}\bigg\}.$$

First, we show that there exists a local maximum of the observed-data likelihood function over S_n such that the proposed estimators $(\hat\Lambda_2,\hat\beta)$ converge to the true parameters in probability under the norm $\|\cdot\|_\infty$.

By Schumaker (2007) and Condition 1, there exists a function $\hat\beta_0(t)=\sum_{j=1}^{K_n}\alpha_{j0}\phi_j(t)$ such that $\|\hat\beta_0-\beta_0\|_\infty=O(m_n^{-r})$. We consider the neighbourhood of $\hat\beta_0$ in the sieve space $\mathcal{N}_{\epsilon_n}=\{\beta:\beta(t)=\sum_{j=1}^{K_n}\alpha_j\phi_j(t)$ with $(\sum_{j=1}^{K_n}|\alpha_j-\alpha_{j0}|^2)^{1/2}\leqslant\epsilon_n\}$, where ϵ_n is to be chosen later. For each $\beta\in\mathcal{N}_{\epsilon_n}$, we define $\hat\Lambda_{2,\beta}=\arg\max_{\Lambda_2}P_n l(\Lambda_2,\beta)$, where Λ_2 is a step function with jumps at the observed failure events. If we choose ϵ_n such that $m_n^{3/2}\epsilon_n\to 0$, then for $\beta\in\mathcal{N}_{\epsilon_n}$ we have

$$\|\beta-\hat\beta_0\|_{\mathrm{BV}}\leqslant\sum_{j=1}^{K_n}|\alpha_j-\alpha_{j0}|\,\|\phi_j'\|_\infty\leqslant O(m_n)\{\epsilon_n^2(m_n+l)\}^{1/2}\to 0.$$

Therefore β has bounded total variation. Define

$$\hat\Lambda_{20}(t)=\sum_{j=1}^n\int_0^t\frac{I(Y_j\geqslant s)}{\sum_{k=1}^n I(Y_k\geqslant s)[q_k\exp\{\beta_0(s)\}+(1-q_k)]}\,\mathrm{d}N_j(s);$$

it is easy to see that $\|\hat\Lambda_{20}-\Lambda_{20}\|_{\mathrm{BV}}=O_{\mathrm{p}}(n^{-1/2})$. Therefore $P_n l(\hat\Lambda_{2,\beta},\beta)\geqslant P_n l(\hat\Lambda_{20},\beta)$, where P_n denotes the empirical measure. Note that $P_n l(\hat\Lambda_{2,\beta},\beta)-P_n l(\hat\Lambda_{20},\beta)$ equals

$$n^{-1}\sum_{i=1}^n\sum_{g=1}^m I(G_i=g)\\ \times\Bigg\{\Delta_i\log\left(\frac{\delta\hat\Lambda_{2,\beta}(Y_i)}{\delta\hat\Lambda_{20}(Y_i)}\,\frac{p_{ig}\exp\{\beta(Y_i)\}\exp\big[-\int_0^{Y_i}\exp\{\beta(t)\}\,\mathrm{d}\hat\Lambda_{2,\beta}(t)\big]+(1-p_{ig})\exp\{-\hat\Lambda_{2,\beta}(Y_i)\}}{p_{ig}\exp\{\beta(Y_i)\}\exp\big[-\int_0^{Y_i}\exp\{\beta(t)\}\,\mathrm{d}\hat\Lambda_{20}(t)\big]+(1-p_{ig})\exp\{-\hat\Lambda_{20}(Y_i)\}}\right)\\ +(1-\Delta_i)\log\left(\frac{p_{ig}\exp\big[-\int_0^{Y_i}\exp\{\beta(t)\}\,\mathrm{d}\hat\Lambda_{2,\beta}(t)\big]+(1-p_{ig})\exp\{-\hat\Lambda_{2,\beta}(Y_i)\}}{p_{ig}\exp\big[-\int_0^{Y_i}\exp\{\beta(t)\}\,\mathrm{d}\hat\Lambda_{20}(t)\big]+(1-p_{ig})\exp\{-\hat\Lambda_{20}(Y_i)\}}\right)\Bigg\}.$$

It is easy to show that $\delta\hat{\Lambda}_{20}(t) = O_{\mathrm{p}}(1/n)$, so we conclude that there exist constants c_1 and c_2 independent of β such that

$$\begin{aligned}
0 &\leqslant P_n l(\hat{\Lambda}_{2,\beta}, \beta) - P_n l(\hat{\Lambda}_{20}, \beta)\\
&\leqslant n^{-1}\sum_{i=1}^{n} c_1 \log\left\{n\delta\hat{\Lambda}_{2,\beta}(Y_i)\right\}\Delta_i\\
&\qquad\qquad + c_2 \log\left(p_i \exp\left[-\int_0^\tau \exp\{\beta(s)\}\,\mathrm{d}\hat{\Lambda}_{2,\beta}(s)\right] + (1-p_i)\exp\{-\hat{\Lambda}_{2,\beta}(\tau)\}\right)\\
&\leqslant n^{-1}\sum_{i=1}^{n} c_1 \log\{n\delta\hat{\Lambda}_{2,\beta}(Y_i)\}\Delta_i + c_2 \log\left[p_i \exp\{-c\hat{\Lambda}_{2,\beta}(\tau)\} + (1-p_i)\exp\{-c\hat{\Lambda}_{2,\beta}(\tau)\}\right]\\
&\leqslant n^{-1}\sum_{i=1}^{n} c_1 \log\{n\delta\hat{\Lambda}_{2,\beta}(Y_i)\}\Delta_i - c_2\hat{\Lambda}_{2,\beta}(\tau) + O_{\mathrm{p}}(1).
\end{aligned}$$

Hence $n^{-1}\sum_{i=1}^{n}\Delta_i \log\{n\delta\hat{\Lambda}_{2,\beta}(Y_i)\} - c_1\hat{\Lambda}_{2,\beta}(\tau)$ is bounded from below in probability. Since $n^{-1}\sum_{i=1}^{n}\Delta_i \log\{n\delta\hat{\Lambda}_{2,\beta}(Y_i)\}$ is less than $\log\{\sum_{i=1}^{n}\Delta_i\delta\hat{\Lambda}_{2,\beta}(Y_i)\} = \log\hat{\Lambda}_2(\tau)$, $\limsup_{n\to\infty}\{\sup_{\beta\in\mathcal{N}_{\epsilon_n}}\hat{\Lambda}_{2,\beta}(\tau)\}$ is finite with probability tending to 1. As a result, $\{\hat{\Lambda}_{2,\beta} : \beta \in \mathcal{N}_{\epsilon_n}\}$ consists of bounded and increasing functions.

From the fact that $P_n l_{\Lambda_2}(\hat{\Lambda}_{2,\beta}, \beta)(h_1) = 0$ we obtain

$$(P_n - P) l_{\Lambda_2}(\hat{\Lambda}_{2,\beta}, \beta)(h_1) = P l_{\Lambda_2}(\hat{\Lambda}_{2,\beta}, \beta)(h_1).$$

The left-hand side of this equation is $O_{\mathrm{p}}(n^{-1/2})$ because l_{Λ_2} is Donsker by virtue of the fact that both $\Lambda_{2,\beta}$ and β belong to BV$[0,\tau]$. Applying the Taylor expansion at the true (Λ_{20}, β_0) to the right-hand side, we get

$$O_{\mathrm{p}}(n^{-1/2}) = -\left\langle \mathcal{I}_1(\Lambda_{20}, \beta_0)(h_1),\ \mathrm{d}\hat{\Lambda}_{2,\beta} - \mathrm{d}\Lambda_{20}\right\rangle_{L_2} + o(\|\hat{\Lambda}_{2,\beta} - \Lambda_{20}\|_{\mathrm{BV}}) + O_{\mathrm{p}}(\|\beta - \beta_0\|_{L_2}),$$

where $\mathcal{I}_1$ is the operator in $\mathcal{I}$ corresponding to Λ_2. Using the invertibility of $\mathcal{I}_1$, we obtain $\|\hat{\Lambda}_{2,\beta} - \Lambda_{20}\|_{\mathrm{BV}} = A_n(n^{-1/2} + \|\beta - \beta_0\|_{L_2})$, where $\sup_{\beta\in\mathcal{N}_{\epsilon_n}}|A_n|$ is a bounded random variable.

We now consider $B_n \equiv P_n\{l(\hat{\Lambda}_{2,\beta}, \beta) - l(\hat{\Lambda}_{20}, \hat{\beta}_0)\}$. Observe that $B_n = (P_n - P)\{l(\hat{\Lambda}_{2,\beta}, \beta) - l(\hat{\Lambda}_{20}, \hat{\beta}_0)\} + P\{l(\hat{\Lambda}_{2,\beta}, \beta) - l(\hat{\Lambda}_{20}, \hat{\beta}_0)\}$. The first term on the right-hand side is equal to $c_n n^{-1/2}$, where $\sup_{\beta\in\mathcal{N}_{\epsilon_n}}|c_n| \to 0$. For the second term, we apply the expansion at the true values and obtain

$$\begin{aligned}
&-\left\langle \mathcal{I}(\Lambda_2^*, \beta^*)(\mathrm{d}\hat{\Lambda}_{2,\beta}/\mathrm{d}\hat{\Lambda}_{20} - \lambda_{20},\ \beta - \beta_0),\ (\mathrm{d}\hat{\Lambda}_{2,\beta}/\mathrm{d}\hat{\Lambda}_{20} - \lambda_{20},\ \beta - \beta_0)\right\rangle_{L_2}\\
&\quad + o(\|\hat{\Lambda}_{20} - \Lambda_{20}\|_\infty^2 + \|\hat{\beta}_0 - \beta_0\|_\infty^2),
\end{aligned}$$

where (Λ_2^*, β^*) is between $(\hat{\Lambda}_{2,\beta}, \beta)$ and (Λ_{20}, β_0). Thus, we have $B_n \leqslant c_n n^{-1/2} - c_1/2\|\beta - \beta_0\|_{L_2}^2 + b_n(n^{-1} + m_n^{-2r})$, where $\sup_{\beta\in\mathcal{N}_{\epsilon_n}}|b_n| \to 0$. Therefore, if $\beta \in \partial\mathcal{N}_{\epsilon_n}$, a result from de Boor (1978) gives $\|\beta - \beta_0\|_{L_2}^2 \geqslant c_2\epsilon_n^2$, so that $B_n \leqslant \{|c_n|n^{-1/2} + b_n(n^{-1} + m_n^{-2r})\} - c_1c_2\epsilon_n^2/2$. Hence, if we choose $\epsilon_n^2 = 4(c_1c_2)^{-1}\{|c_n|n^{-1/2} + b_n(n^{-1} + m_n^{-2r})\}$, then $B_n < 0$, noting that such ϵ_n still satisfies $m_n^{3/2}\epsilon_n \to 0$ due to $r \geqslant 2$ and Condition 3; that is, there exists a local maximum $\hat{\beta}_n$ within this neighbourhood. Consequently, $\|\hat{\beta}_n - \beta_0\|_{\mathrm{BV}} \to 0$ and $\|\hat{\beta}_n - \beta_0\|_{L_2}^2 \leqslant \|\hat{\beta}_n - \hat{\beta}_0\|_{L_2}^2 + O(m_n^{-2r}) \leqslant \epsilon_n^2 + m_n^{-2r} = o_{\mathrm{p}}(n^{-1/2})$. From the result that $\|\hat{\Lambda}_{2,\beta} - \Lambda_{20}\|_{\mathrm{BV}} = A_n(n^{-1/2} + \|\beta - \beta_0\|_{L_2})$, the corresponding $\hat{\Lambda}_{2,\beta}$ satisfies $\|\hat{\Lambda}_{2,\hat{\beta}_n} - \Lambda_{20}\|_{\mathrm{BV}} = O_{\mathrm{p}}(n^{-1/2}) + \|\hat{\beta}_n - \beta_0\|_{L_2} = o_{\mathrm{p}}(n^{-1/4})$. This implies that $\sup_{t\in[0,\tau]}|\hat{\Lambda}_{2n}(t) - \Lambda_{20}(t)| + \sup_{t\in[0,\tau]}|\hat{\beta}_n(t) - \beta_0(t)| = o_{\mathrm{p}}(1)$. By reversing the labels, the same argument gives $\sup_{t\in[0,\tau]}|\hat{\Lambda}_{1n}(t) - \Lambda_{10}(t)| = o_{\mathrm{p}}(1)$. The proof of Theorem 1 is thus complete. □

Proof of Theorem 2. For any $h_1 \in \mathrm{BV}[0,\tau]$ and any h_2 with bounded rth derivative in $[0,\tau]$, we have $P_n l_{\Lambda_2}(\hat\Lambda_{2n},\hat\beta_n)(h_1)=0$ and $P_n l_\beta(\hat\Lambda_{2n},\hat\beta_n)(h_{2n})=0$. Here, h_{2n} is the projection of h_2 onto S_n, and $\|h_{2n}-h_2\|_\infty = O(m_n^{-r})$. This gives

$$G_n\bigl\{l_{\Lambda_2}(\hat\Lambda_{2n},\hat\beta_n)(h_1)+l_\beta(\hat\Lambda_{2n},\hat\beta_n)(h_{2n})\bigr\} = -n^{1/2}P\bigl\{l_{\Lambda_2}(\hat\Lambda_{2n},\hat\beta_n)(h_1)+l_\beta(\hat\Lambda_{2n},\hat\beta_n)(h_{2n})\bigr\}, \qquad \text{(A1)}$$

where $G_n = n^{1/2}(P_n - P)$. It is straightforward to verify that $\{l_\beta(\hat\Lambda_{2n},\hat\beta_n)(h_1)+l_{\Lambda_2}(\hat\Lambda_{2n},\hat\beta_n)(h_2) : \|h_1\|_{\mathrm{BV}}\leqslant 1,\ \|h_2\|_\infty \leqslant 1\}$ is a Donsker class. Hence, the left-hand side of (A1) equals $G_n\{l_{\Lambda_2}(\hat\Lambda_{2n},\hat\beta_n)(h_1)+l_\beta(\Lambda_{20},\beta_0)(h_2)\}+o_{\mathrm{p}}(1)$, where here and in what follows $o_{\mathrm{p}}(1)$ refers to some random element that converges to zero in probability uniformly in (h_1,h_2).

By Taylor expansion, the right-hand side of (A1) can be written as

$$\begin{aligned}
&-\{1+o_{\mathrm{p}}(1)\}\left\{\int h_1^*\,\mathrm{d}(\hat\Lambda_{2n}-\Lambda_{20})+\int h_2^*(\hat\beta_n-\beta_0)\,\mathrm{d}t\right\}+n^{1/2}O\bigl(\|\hat\Lambda_{2n}-\Lambda_{20}\|_{\mathrm{BV}}^2+\|\hat\beta_n-\beta\|_{L_2}^2\bigr)\\
&\quad = -n^{1/2}\{1+o_{\mathrm{p}}(1)\}\left\{\int h_1^*\,\mathrm{d}(\hat\Lambda_{2n}-\Lambda_{20})+\int h_2^*(\hat\beta_n-\beta_0)\,\mathrm{d}t\right\}+o_{\mathrm{p}}(1),
\end{aligned}$$

where $(h_1^*,h_2^*)=\mathcal{I}(\Lambda_{20},\beta_0)(h_1,h_2)$. This yields

$$G_n\bigl\{l_{\Lambda_2}(\Lambda_{20},\beta_0)(h_1^{**})+l_\beta(\Lambda_{20},\beta_0)(h_2^{**})\bigr\}+o_{\mathrm{p}}(1) = -n^{1/2}\left\{\int h_1\,\mathrm{d}(\hat\Lambda_{2n}-\Lambda_{20})+\int h_2(\hat\beta_n-\beta_0)\,\mathrm{d}t\right\}$$

where $(h_1^{**},h_2^{**})=\mathcal{I}^{-1}(\Lambda_{20},\beta_0)(h_1,h_2)$. In other words, $n^{1/2}\{\hat\Lambda_{2n}(t)-\Lambda_{20}(t),\ \hat\beta_n(t)-\beta_0(t)\}$ converges in distribution to a zero-mean Gaussian process in $l^\infty(\mathcal{F}_{\mathrm{BV}}\times\mathcal{F}_\beta)$. Finally, since $\hat\Lambda_{1n}$ is obtained using the same estimation as for $\hat\Lambda_{2n}$ but reversing the group labels, a similar asymptotically linear expansion can be shown to hold for $n^{1/2}\int h_1\,\mathrm{d}(\hat\Lambda_{1n}-\Lambda_{10})$. Hence, we conclude that $n^{1/2}\{\hat\Lambda_{1n}(t)-\Lambda_{10}(t),\ \hat\Lambda_{2n}(t)-\Lambda_{20}(t)\}$ converges in distribution to a zero-mean Gaussian process in $l^\infty(\mathcal{F}_{\mathrm{BV}}\times\mathcal{F}_{\mathrm{BV}})$.

From the asymptotic linear expansion of $n^{1/2}\{\hat\Lambda_{kn}(t)-\Lambda_{k0}(t)\}$, we note that for any fixed t, the influence function of $\hat\Lambda_{kn}$ is on the tangent space of the score functions. Therefore, the estimators are semiparametrically efficient in the metric space $l^\infty(\mathcal{F}_{\mathrm{BV}}\times\mathcal{F}_{\mathrm{BV}})$ according to Theorem 18.8 in Kosorok (2008). We have thus completed the proof of Theorem 2. □

References

Alcalay, R. N., Mirelman, A., Saunders-Pullman, R., Tang, M., Mejia Santana, H., Raymond, D., Roos, E., Orbe-Reilly, M., Gurevich, T., Bar Shira, A. et al. (2013). Parkinson's disease phenotype in Ashkenazi Jews with and without LRRK2 G2019S mutations. *Mov. Disord.* **28**, 1966–71.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer.

Cheng, G. & Wang, X. (2011). Semiparametric additive transformation model under current status data. *Electron. J. Statist.* **5**, 1735–64.

de Boor, C. (1978). *A Practical Guide to Splines*. Wroclaw: Springer.

Diao, G. & Lin, D. (2005). Semiparametric methods for mapping quantitative trait loci with censored data. *Biometrics* **61**, 789–98.

Driver, J. A., Logroscino, G., Gaziano, J. M. & Kurth, T. (2009). Incidence and remaining lifetime risk of Parkinson disease in advanced age. *Neurology* **72**, 432–8.

Fine, J. P., Zou, F. & Yandell, B. S. (2004). Nonparametric estimation of mixture models, with application to quantitative trait loci. *Biostatistics* **5**, 501–13.

Goldwurm, S., Tunesi, S., Tesei, S., Zini, M., Sironi, F., Primignani, P., Magnani, C. & Pezzoli, G. (2011). Kin-cohort analysis of LRRK2-G2019S penetrance in Parkinson's disease. *Mov. Disord.* **26**, 2144–5.

Hall, P. & Zhou, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31**, 201–24.

Healy, D. G., Falchi, M., O'Sullivan, S. S., Bonifati, V., Durr, A., Bressman, S., Brice, A., Aasly, J., Zabetian, C. P., Goldwurm, S. et al. (2008). Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: A case-control study. *Lancet Neurol.* **7**, 583–90.

Kosorok, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Laird, N. M. & Ware, J. H. (1982). Random-effect models for longitudinal data. *Biometrics* **38**, 963–74.

Langbehn, D. R., Brinkman, R. R., Falush, D., Paulsen, J. S. & Hayden, M. R. (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin. Genet.* **65**, 267–77.

Latourelle, J. C., Sun, M., Lew, M. F., Suchowersky, O., Klein, C., Golbe, L. I., Mark, M. H., Growdon, J. H., Wooten, G. F., Watts, R. L. et al. (2008). The Gly2019Ser mutation in LRRK2 is not fully penetrant in familial Parkinson's disease: The GenePD study. *BMC Med.* **6**, article no. 32.

Lesage, S., Leutenegger, A. L., Ibanez, P., Janin, S., Lohmann, E., Durr, A., Brice, A. & French Parkinson's Disease Genetics Study Group. (2005). LRRK2 haplotype analyses in European and North African families with Parkinson disease: A common founder for the G2019S mutation dating from the 13th century. *Am. J. Hum. Genet.* **77**, 330–2.

Ma, Y. & Wang, Y. (2012). Efficient distribution estimation for data with unobserved sub-population identifiers. *Electron. J. Statist.* **6**, 710–37.

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

Marder, K., Tang, M., Alcalay, R., Mejia-Santana, H., Raymond, D., Mirelman, A., Saunders-Pullman, R., Clark, L., Ozelius, L., Orr-Urtreger, A. et al. (2014). Age specific penetrance of the LRRK2 G2019S mutation in the Michael J Fox Ashkenazi Jewish (AJ) LRRK2 consortium. *Neurology* **82** (10 Suppl.), article no. S17-002.

Mclachlan, G. J. & Basford, K. E. (1988). *Mixture Models, Inference and Applications to Clustering*. New York: Dekker.

Orr-Urtreger, A., Shifrin, C., Rozovski, U., Rosner, S., Bercovich, D., Gurevich, T., Yagev-More, H., Bar-Shira, A. & Giladi, N. (2007). The LRRK2 G2019S mutation in Ashkenazi Jews with Parkinson's disease: Is there a gender effect? *Neurology* **69**, 1595–602.

Paisán-Ruíz, C., Jain, S., Evans, E. W., Gilks, W. P., Simón, J., van der Brug, M., de Munain, A. L., Aparicio, S., Gil, A. M., Khan, N. et al. (2004). Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron* **44**, 595–600.

Qin, J., Garcia, T., Ma, Y., Tang, M., Marder, K. & Wang, Y. (2014). Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint. *Ann. Appl. Statist.* **8**, 1182–208.

Rudin, W. (1973). *Functional Analysis*. New York: McGraw-Hill.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge: Cambridge University Press.

Titterington, D. M., Smith, A. F. M. & Markov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.

Trinh, J. & Farrer, M. (2013). Advances in the genetics of Parkinson disease. *Nature Rev. Neurol.* **9**, 445–54.

van der Vaart, A. & Wellner, J. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

Wacholder, S., Hartge, P., Struewing, J., Pee, D., McAdams, M., Brody, L. & Tucker, M. (1998). The kin-cohort study for estimating penetrance. *Am. J. Epidemiol.* **148**, 623–30.

Wang, Y., Clark, L. N., Louis, E. D., Mejia-Santana, H., Harris, J., Cote, L. J., Waters, C., Andrews, D., Ford, B., Frucht, S. et al. (2008). Risk of Parkinson's disease in carriers of parkin mutations: Estimation using the kin-cohort method. *Arch. Neurol.* **65**, 467–74.

Wang, Y., Garcia, T. & Ma, Y. (2012). Nonparametric estimation for censored mixture data with application to the Cooperative Huntington's Observational Research Trial. *J. Am. Statist. Assoc.* **107**, 1324–38.

Yang, W., Hormozdiari, F., Wang, Z., He, D., Pasaniuc, B. & Eskin, E. (2013). Leveraging reads that span multiple single nucleotide polymorphisms for haplotype inference from sequencing data. *Bioinformatics* **29**, 2245–52.

Zeng, D. & Lin, D. (2010). A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statist. Sinica* **20**, 871–910.

Zeng, D., Lin, D., Avery, C. L. & North, K. E. (2006). Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics* **7**, 486–502.

Zhang, H., Olschwang, S. & Yu, K. (2010). Statistical inference on the penetrances of rare genetic mutations based on a case-family design. *Biostatistics* **11**, 519–32.

[*Received March* 2014. *Revised April* 2015]