Biometrika (2020), **107**, 2, *pp.* 433–448
Printed in Great Britain
doi: 10.1093/biomet/asaa012
Advance Access publication 15 April 2020

Ensemble estimation and variable selection with semiparametric regression models

BY SUNYOUNG SHIN

Department of Mathematical Sciences, University of Texas at Dallas, 800 W. Campbell Rd., Richardson, Texas 75080, U.S.A.

sunyoung.shin@utdallas.edu

YUFENG LIU

Department of Statistics and Operations Research, CB# 3260, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.

yfliu@email.unc.edu

STEPHEN R. COLE

Department of Epidemiology, CB# 7435, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.

cole@unc.edu

AND JASON P. FINE

Department of Biostatistics, CB# 7420, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.

jfine@email.unc.edu

SUMMARY

We consider scenarios in which the likelihood function for a semiparametric regression model factors into separate components, with an efficient estimator of the regression parameter available for each component. An optimal weighted combination of the component estimators, named an ensemble estimator, may be employed as an overall estimate of the regression parameter, and may be fully efficient under uncorrelatedness conditions. This approach is useful when the full likelihood function may be difficult to maximize, but the components are easy to maximize. It covers settings where the nuisance parameter may be estimated at different rates in the component likelihoods. As a motivating example we consider proportional hazards regression with prospective doubly censored data, in which the likelihood factors into a current status data likelihood and a left-truncated right-censored data likelihood. Variable selection is important in such regression modelling, but the applicability of existing techniques is unclear in the ensemble approach. We propose ensemble variable selection using the least squares approximation technique on the unpenalized ensemble estimator, followed by ensemble re-estimation under the selected model. The resulting estimator has the oracle property such that the set of nonzero parameters is successfully recovered and the semiparametric efficiency bound is achieved for this parameter set. Simulations show that the proposed method performs well relative to alternative approaches. Analysis of an AIDS cohort study illustrates the practical utility of the method.

Some Keywords: Likelihood factorization; Penalized estimation; Prospective cohort study; Semiparametric efficiency; Uncorrelatedness.

1. Introduction

Efficient estimation via the averaging of estimators has been suggested for many statistical models. Keller & Olkin (2004) and references therein studied combining estimators of the mean of a normal distribution from different sources. In meta-analysis, estimators from multiple studies are commonly aggregated to construct an efficient estimator (Borenstein et al., 2009). Lin & Zeng (2010), Liu et al. (2015) and Slud et al. (2018) proposed a method of combining estimators of a common parameter across independent studies, where the joint likelihood is decomposed into independent component likelihoods. Cox (2001) considered a parametric model based on a factorizable likelihood, a simple example of which is a statistical model for two or more independent studies. As a general approach for combining information, he suggested a generalized least squares estimator as an overall estimator, which optimally combines the component estimators with weights calculated from the inverse of their estimated covariance matrices. This estimator is especially useful when component likelihood maximization is straightforward while full likelihood maximization is computationally demanding.

We investigate combining efficient component estimators of the regression parameter in semiparametric regression models, where the full likelihood is factorizable, similarly to Cox (2001). In our so-called ensemble estimation procedure, we optimally combine the estimators of the finite-dimensional regression parameter to obtain an overall estimator which is semiparametric efficient under uncorrelatedness conditions. These results are valuable when estimation of the infinite-dimensional nuisance parameter, which may not be achievable at the usual parametric rate, is challenging, complicating the full likelihood analysis.

The motivation for this work arises from survival analysis of prospective cohort studies using the age scale. Many subjects may have already experienced the event at the time of study enrolment, while those who did not may not experience the event during the course of follow-up. Such doubly censored data, including both left- and right-censored times, may be analysed using a full likelihood analysis of the proportional hazards model (Cox, 1972). Kim et al. (2013) and Su & Wang (2016) developed approximate expectation-maximization algorithms for the Cox model with doubly censored data by considering the left-censored data as missing. While the theoretical properties of the procedures were established, its practical usage is hampered by computational inefficiency and instability. An alternative analysis is to only use data on those subjects that are event-free at enrolment, left-truncating using the age at enrolment. However, data on subjects who had the event prior to enrolment is not utilized, resulting in a loss of information. Most prospective cohort studies collect information related to participants' age of enrolment. We make novel use of the age at enrolment as a left truncation time, which differs from the standard doubly censored data. We refer to this set-up as prospective doubly censored data. It is shown in § 3 that the full likelihood for such data may be decomposed into a current status data likelihood based on event status at enrolment and a left-truncated right-censored data likelihood based on subjects who did not have the event at enrolment. Both component likelihoods have been well studied, with theoretical and computational issues addressed rigorously. The ensemble estimation may be performed based on the component estimators, simplifying both computation and inference. We utilize the likelihood equivalence between prospective doubly censored data and doubly censored data, which Su & Wang (2016) also recognized, but did not utilize.

Variable selection in semiparametric regression is an important practical issue, e.g., when identifying risk factors in cohort studies for HIV infection. Penalization is a popular variable

selection technique, originating in the seminal work of Tibshirani (1996) and Fan & Li (2001) for parametric regression models. Penalization techniques have been adapted to semiparametric regression models. A well-known example is penalized partial likelihood estimation of the regression parameters in the proportional hazards model with right-censored data, yielding results similar to those for parametric models (Tibshirani, 1997; Fan & Li, 2002). When simultaneous estimation of both regression and nuisance parameters is needed, model-specific penalization approaches based on modified likelihoods have been proposed (Cai et al., 2005; Du et al., 2010; Liu & Zeng, 2013). Such penalized estimation is complicated, owing to theoretical and computational difficulties. For semiparametric factorizable likelihoods we suggest ensemble variable selection, in which the approximation technique in Wang & Leng (2007) is employed to construct sparse estimators, with some regression coefficients being exactly zero (Lu et al., 2012). The main idea is to approximate the profile likelihood of the regression parameter by a least squares criterion centred at the unpenalized ensemble estimator.

The proposed ensemble estimation and variable selection procedure offers a general methodology for regression parameter estimation in semiparametric regression models with factorizable likelihoods. The main requirement is the existence of efficient regression estimators for each component. The nuisance parameter may be estimated at different and potentially slower than parametric rates in the component likelihoods, as happens with prospective doubly censored data. In § 2, theoretical properties are established under uncorrelatedness conditions on the component likelihoods with weak conditions on the component estimators. For variable selection with a fixed-dimensional regression parameter, it is shown that the penalized estimator is sparse and correctly selects the true nonzero parameters as the sample size increases. The resulting estimator has the oracle property: its limiting distribution is normal and its asymptotic covariance matrix is the same as that of the efficient estimator, with the true nonzero parameters known a priori.

2. Ensemble framework

2.1. *General methodology*

Consider a semiparametric regression model with a finite-dimensional regression parameter, θ, and an infinite-dimensional nuisance parameter, Λ. Denote (θ_0, Λ_0) as the true parameter value. The regression parameter is in a fixed p-dimensional parameter space, $\Theta \subset \mathbb{R}^p$. Without loss of generality, denote $\theta_0 = (\theta_{10}^{\mathrm{T}}, \theta_{20}^{\mathrm{T}})^{\mathrm{T}}$, where θ_{10} is an s-dimensional nonzero regression coefficient and $\theta_{20} = 0$. We consider a sparse regression parameter where s is a fixed number less than p. The goals are efficient estimation of θ_0, identification of θ_{20}, and oracle estimation of θ_{10}.

The data consist of n independent and identically distributed observations, $(z_1, \ldots, z_n)$, from $P_{\theta_0,\Lambda_0} \in \mathcal{P}$, where $\mathcal{P}$ is a set of probability measures on the sample space $(\Omega, \mathcal{F})$. Suppose the loglikelihood for the semiparametric model based on these data is denoted by $l_n(\theta, \Lambda) = \sum_{i=1}^n l_{\theta,\Lambda}(z_i)$, where $l_{\theta,\Lambda}(z_i)$ is the loglikelihood of the ith observation. It is assumed that the loglikelihood separates into K component likelihoods, $l_n^1(\theta, \Lambda), \ldots, l_n^K(\theta, \Lambda)$, where K is a fixed number and θ is a common parameter to all component likelihoods. That is, the loglikelihood is the summation, $l_n(\theta, \Lambda) = \sum_{k=1}^K l_n^k(\theta, \Lambda)$. Such likelihoods are referred to as factorizable. The same true parameter value applies to all component likelihoods as they are stemmed from $l_n(\theta, \Lambda)$.

Denote by $L_2(P_{\theta_0,\Lambda_0})$ the space of all functions $g : \Omega \to \mathbb{R}$ with $\int g \, \mathrm{d}P_{\theta_0,\Lambda_0} = 0$ and $\int g^2 \, \mathrm{d}P_{\theta_0,\Lambda_0} < \infty$. Throughout the paper we omit the subscripts (θ_0, Λ_0) denoting the base point. The score function for θ is defined as the derivative of the loglikelihood with respect to θ with fixed Λ_0, i.e., $\dot{l}_\theta(z_i) = \partial l_{\theta,\Lambda_0}(z_i)/(\partial\theta)|_{\theta=\theta_0}$. We consider one-dimensional parametric submodels of Λ,

denoted by $\Lambda(t)$, which approach Λ_0 as $t \to 0$ (van der Vaart, 2000; Kosorok, 2007). The score function for the submodel of Λ is defined as $\partial l_{\theta_0,\Lambda(t)}(z_i)/(\partial t)|_{t=0}$ (Bickel et al., 1993). A collection of score functions for one-dimensional submodels is called a tangent set for Λ. The tangent space for Λ is defined as the closed span of the tangent set, and is denoted as $\dot{\mathcal{P}}_P \subset L_2(P)$. The scores are pointwise Gateau derivatives with respect to a scalar parameter component rather than mean-square derivatives. The efficient score function for θ is defined as the residual of the projection of $\dot{l}_\theta$ onto $\dot{\mathcal{P}}_P$ in $L_2(P)$ (Bickel et al., 1993). Specifically, the efficient score function for θ is written as $\tilde{l}_\theta = \dot{l}_\theta - \prod \dot{l}_\theta$, where $\prod$ is the projection onto $\dot{\mathcal{P}}_P$ in $L_2(P)$. The efficient information matrix is the variance of the efficient score function, i.e., $\tilde{I}_\theta = E\{\tilde{l}_\theta(\tilde{l}_\theta)^{\mathrm{T}}\}$, which is assumed to be positive definite. The semiparametric efficiency bound for θ is $\tilde{I}_\theta$, where $(\tilde{I}_\theta)^{-1}$ is the smallest asymptotic variance among all regular estimators of θ in the semiparametric model. For the kth likelihood component, the score function for θ is $\dot{l}^k_\theta(z_i) = \partial l^k_{\theta,\Lambda_0}(z_i)/(\partial\theta)|_{\theta=\theta_0}$ and the score function for the submodel of Λ is $\partial l^k_{\theta_0,\Lambda(t)}(z_i)/(\partial t)|_{t=0}$. By likelihood factorization, $\dot{l}_\theta(z_i) = \sum_{k=1}^K \dot{l}^k_\theta(z_i)$ and $\partial l_{\theta_0,\Lambda(t)}(z_i)/(\partial t)|_{t=0} = \sum_{k=1}^K \partial l^k_{\theta_0,\Lambda(t)}(z_i)/(\partial t)|_{t=0}$, which are called score function additivities for θ and $\Lambda(t)$. Similarly to the full likelihood, we define a componentwise tangent set and tangent space for Λ. The kth component nuisance-parameter tangent space is denoted as $\dot{\mathcal{P}}^k_P \subset L_2(P)$, the projection onto which is denoted by $\prod_k$. The component efficient score function for θ is $\tilde{l}^k_\theta = \dot{l}^k_\theta - \prod_k \dot{l}^k_\theta$, and the component efficient information matrix is $\tilde{I}^k_\theta = E\{\tilde{l}^k_\theta(\tilde{l}^k_\theta)^{\mathrm{T}}\}$. All the efficient scores and information matrices introduced correspond to single observations. See Bickel et al. (1993), van der Vaart (2000) and Kosorok (2007) for more precise illustrations of the efficient score and information.

The ensemble estimation we propose is an extension of the efficient combination of the component regression estimators, which Cox (2001) suggested for parametric factorizable likelihoods, to semiparametric factorizable likelihoods. Let $\hat{\theta}^k_F$ denote an efficient estimator of θ based on the kth component likelihood, $l^k_n(\theta, \Lambda)$, $k = 1, \ldots, K$. Denote an inverse asymptotic covariance estimator of $\tilde{I}^k_\theta$ by $\hat{I}^k_F$. We suggest an ensemble estimator which minimizes $\sum_{k=1}^K (\theta - \hat{\theta}^k_F)^{\mathrm{T}} \hat{I}^k_F (\theta - \hat{\theta}^k_F)$, and has a closed-form expression, $\hat{\theta}_F = (\sum_{k=1}^K \hat{I}^k_F)^{-1}(\sum_{k=1}^K \hat{I}^k_F \hat{\theta}^k_F)$. Its asymptotic inverse covariance matrix is estimated by $\hat{I}_F = \sum_{k=1}^K \hat{I}^k_F$. The intuition for the procedure is that the log profile likelihood of θ is asymptotically equivalent to the sum of quadratic forms which are centred on the efficient component estimators. The ensemble estimation is extremely useful when an efficient estimator of θ from the full likelihood is computationally very difficult to obtain while the efficient estimators from the component likelihoods can be easily obtained.

Ensemble variable selection applies the least squares approximation approach of Wang & Leng (2007) for efficient variable selection to the ensemble estimator. A least squares approximation replaces the unpenalized objective function based on the preliminary ensemble estimator and is regularized by an adaptive lasso penalty (Zou, 2006). The intermediate estimator with the ensemble variable selection, $\hat{\theta}_{E,\lambda_n}$, is obtained by minimizing

$$Q(\theta) = (\theta - \hat{\theta}_F)^{\mathrm{T}} \hat{I}_F (\theta - \hat{\theta}_F) + \lambda_n \sum_{j=1}^p |\theta_j| / |\hat{\theta}_{Fj}|,$$

where λ_n is a nonnegative tuning parameter. Following Wang & Leng (2007), we select the optimal tuning parameter by minimizing the modified Bayes information criterion: $\mathrm{BIC}_{\lambda_n} = (\hat{\theta}_{E,\lambda_n} - \hat{\theta}_F)^{\mathrm{T}} \hat{I}_F (\hat{\theta}_{E,\lambda_n} - \hat{\theta}_F) + (\log n/n) \sum_{j=1}^p I(\hat{\theta}_{E,\lambda_n,j} \neq 0)$. The selected model is denoted as $\mathcal{A} = \{j : \hat{\theta}_{E,\lambda_n,j} \neq 0\} \subset \{1, \ldots, p\}$. Denote a subspace of Θ supported on the selected model as

$M = \{\theta \in \Theta : \theta_j = 0, \text{ for all } j \in \mathcal{A}^c\}$. The least angle regression algorithm (Efron et al., 2004) can be directly applied, simplifying the implementation of the optimization.

Next, we recalculate each component estimator based on the ensemble variable selected model and compute an overall estimator using the ensemble estimation approach. Denote the refitted estimator for the kth component as $\hat{\theta}^k \in M \subset \mathbb{R}^p$ and its subvector indexed by the ensemble variable selected model as $\hat{\theta}^k_{\mathcal{A}} = (\hat{\theta}^k_j, j \in \mathcal{A}) \in \mathbb{R}^{|\mathcal{A}|}$. We also estimate the asymptotic inverse covariance matrix of $\hat{\theta}^k_{\mathcal{A}}$ as a submatrix of $\hat{I}^k_F$ associated with $\mathcal{A}$, denoted by $\hat{I}^k_{\mathcal{A}}$. Refitting restricted to M can be conveniently performed, similarly to the initial fittings for $\hat{\theta}^k_F$. Ensemble re-estimation provides a closed form of the resulting estimator, $\hat{\theta} \in \mathbb{R}^p$, where $\hat{\theta}_{\mathcal{A}} = (\sum_{k=1}^K \hat{I}^k_{\mathcal{A}})^{-1}(\sum_{k=1}^K \hat{I}^k_{\mathcal{A}}\hat{\theta}^k_{\mathcal{A}}) \in \mathbb{R}^{|\mathcal{A}|}$ and $\hat{\theta}_{\mathcal{A}^c} = 0$. The asymptotic inverse covariance matrix estimator of the resulting estimator restricted to $\mathcal{A}$ is given as $\hat{I}_{\mathcal{A}} = \sum_{k=1}^K \hat{I}^k_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|} \times \mathbb{R}^{|\mathcal{A}|}$.

2.2. *Uncorrelatedness conditions*

Cox (2001) showed asymptotic efficiency of ensemble estimation in parametric factorizable likelihoods under a second-order validity condition that makes the component score functions uncorrelated under mild regularity conditions. Along the same lines, we assume the following conditions for semiparametric factorizable likelihoods:

Condition 1. Component score functions for θ are pairwise uncorrelated, i.e., $E\{\dot{l}^k_\theta(\dot{l}^{k'}_\theta)^{\mathrm{T}}\} = 0$, $k \neq k'$.

Condition 2. Component tangent spaces for Λ are pairwise orthogonal, $\dot{\mathcal{P}}^k_P \perp \dot{\mathcal{P}}^{k'}_P$, $k \neq k'$.

Condition 3. Component score functions for θ are orthogonal to all other component nuisance-parameter tangent spaces, i.e., $\dot{l}^k_\theta \perp \dot{\mathcal{P}}^{k'}_P$, $k \neq k'$.

Without loss of generality, we consider a decomposition into two component likelihoods. When $\dot{l}^1_\theta(z_i)$ and $\partial l^1_{\theta_0,\Lambda(t)}(z_i)/(\partial t)|_{t=0}$ are uncorrelated with $\dot{l}^2_\theta(z_i)$ and $\partial l^2_{\theta_0,\Lambda(t)}(z_i)/(\partial t)|_{t=0}$, Conditions 1–3 are met. These facts, along with the score function additivity for θ and for the submodel of Λ, yield that the full efficient score function may be factored into the uncorrelated component efficient score functions, as shown in Proposition 1.

PROPOSITION 1. *Under Conditions 1–3, the full efficient score function for θ is the summation of uncorrelated component efficient score functions, i.e., $\tilde{l}_\theta = \sum_{k=1}^K \tilde{l}^k_\theta$, and $E\{\tilde{l}^k_\theta(\tilde{l}^{k'}_\theta)^{\mathrm{T}}\} = 0$, $k \neq k'$. Consequently, the full efficient information is exclusively divided into component efficient information matrices, i.e., $\tilde{I}_\theta = \sum_{k=1}^K \tilde{I}^k_\theta$.*

The efficient information additivity implies that the ensemble estimation attains the full efficiency bound; see the Supplementary Material.

Consider data consisting of n independent and identically distributed observations of (Z_i, W_i, X_i), where X_i is a covariate. There is a semiparametric regression model with parameters (θ, Λ) which leads to a full loglikelihood, denoted as $\sum_{i=1}^n l_{\theta,\Lambda}(Z_i, W_i \mid X_i)$. The distribution of X_i is independent of the model parameters. The full loglikelihood may be decomposed into the sum of marginal and conditional loglikelihoods, which are $\sum_{i=1}^n l^1_{\theta,\Lambda}(W_i \mid X_i)$ and $\sum_{i=1}^n l^2_{\theta,\Lambda}(Z_i \mid W_i, X_i)$, respectively. Condition 1 is satisfied as $E\{\dot{l}^1_\theta(W_i \mid X_i)\dot{l}^2_\theta(Z_i \mid W_i, X_i)^{\mathrm{T}}\} = 0$ by conditioning and zero expectations of $\dot{l}^2_\theta(Z_i \mid W_i, X_i)$ with respect to the conditional distribution. The same argument can be used to verify Conditions 2 and 3. Thus, Conditions 1–3 are met. The main point is that conditioning and marginalizing are used to establish the decomposition

into the marginal and conditional efficient scores and their uncorrelatedness. A special case of this decomposition is the likelihood for prospective doubly censored data. The full likelihood is decomposed into the conditional likelihood of left-truncated and right-censored data, and the marginal likelihood of current status data. This is discussed in § 3. Similarly, Newey (1990) discussed the uncorrelatedness between a marginal density and a conditional density in semiparametric additive regression models, where the marginal density is ancillary to a parameter of interest. The uncorrelatedness provided theoretical justification for estimation solely based on the conditional density.

2.3. *Theoretical properties*

We now investigate theoretical properties of the full ensemble estimator and the resulting refitted ensemble estimator. Throughout this section we will assume that the uncorrelatedness conditions are satisfied. Key assumptions and results are stated below, with the proofs relegated to the Supplementary Material.

Assumption 1. For every $k = 1, \ldots, K$, $\hat{\theta}_F^k$ is regular, asymptotically linear and semiparametric efficient with respect to the component likelihoods such that $n^{1/2}(\hat{\theta}_F^k - \theta_0) = n^{-1/2} \sum_{i=1}^n (\tilde{I}_\theta^k)^{-1} \tilde{l}_\theta^k(z_i) + o_p(1)$, and $n^{1/2}(\hat{\theta}_F^k - \theta_0)$ converges in distribution to $N\{0, (\tilde{I}_\theta^k)^{-1}\}$.

Assumption 2. A consistent estimator of $\tilde{I}_\theta^k, \hat{I}_F^k$, exists for $k = 1, \ldots, K$.

Assumption 1 indicates that the estimators of the finite-dimensional parameter have the usual $n^{1/2}$ convergence rate with the component likelihoods. Assumptions 1 and 2 may require consistent estimation of the nuisance parameter, potentially converging at different and slower than $n^{1/2}$ rate (Groeneboom & Wellner, 1992; van der Vaart, 2000). The implication is that the exact convergence rates of the nuisance parameter for the component likelihoods are not needed to theoretically justify the ensemble method in the regression parameter estimation. By the uncorrelatedness of the component efficient scores, the component estimators are asymptotically uncorrelated, i.e., $\mathrm{cov}\{n^{1/2}(\hat{\theta}_F^k - \theta_0), n^{1/2}(\hat{\theta}_F^{k'} - \theta_0)\} \to 0$ as $n \to \infty, k \neq k'$.

THEOREM 1 (ASYMPTOTIC EFFICIENCY). *Suppose that Assumptions 1–2 hold. Then* $n^{1/2}(\hat{\theta}_F - \theta_0) = O_p(1)$*, and its asymptotic distribution is* $N\{0, (\tilde{I}_\theta)^{-1}\}$.

Theorem 1 states that the full ensemble estimator is $n^{1/2}$-consistent and asymptotically normal with the semiparametric efficiency achieved. The decomposition of the full efficient score into the uncorrelated component efficient scores is the key to establishing the asymptotic efficiency.

Similarly to θ_0, one may write $\theta = (\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}})^{\mathrm{T}}$, where θ_1 corresponds to the s nonzero components and θ_2 corresponds to the zero components. Define the kth component oracle estimator of θ as a hypothetical estimator based on $l_n^k\{(\theta_1^{\mathrm{T}}, 0^{\mathrm{T}})^{\mathrm{T}}, \Lambda\}$ by using the initial fitting method, denoted by $\check{\theta}_1^k \in \mathbb{R}^s, k = 1, \ldots, K$.

Assumption 3. For every $k = 1, \ldots, K$, $\check{\theta}_1^k$ is regular, asymptotically linear and semiparametric efficient with $n^{1/2}(\check{\theta}_1^k - \theta_{10}) = n^{-1/2} \sum_{i=1}^n (\tilde{I}_{\theta_1}^k)^{-1} \tilde{l}_{\theta_1}^k(z_i) + o_p(1)$, and $n^{1/2}(\check{\theta}_1^k - \theta_{10}) \to N\{0, (\tilde{I}_{\theta_1}^k)^{-1}\}$, where $\tilde{l}_{\theta_1}^k(z_i)$ is a subvector of $\tilde{l}_\theta^k(z_i)$ associated with θ_{10} and $\tilde{I}_{\theta_1}^k$ is a submatrix of $\tilde{I}_\theta^k$ associated with θ_{10}.

Assumption 3 is an oracle version of Assumption 1, which states that semiparametric efficiency of the component oracle estimators. That is, the asymptotic variances of the component oracle

estimators attain the semiparametric efficiency bound with respect to their component likelihoods. Assumptions 1–3 follow the assumptions in Wang & Leng (2007). Write $\hat{\theta} = (\hat{\theta}_1^{\mathrm{T}}, \hat{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\hat{\theta}_1 \in \mathbb{R}^s$ and $\hat{\theta}_2 \in \mathbb{R}^{p-s}$.

THEOREM 2 (SELECTION CONSISTENCY). *Suppose that Assumptions 1–3 hold. If* $n^{1/2}\lambda_n \to 0$ *and* $n\lambda_n \to \infty$, *then* $n^{1/2}(\hat{\theta} - \theta_0) = O_p(1)$ *and* $\mathrm{pr}(\hat{\theta}_2 = 0) \to 1$.

THEOREM 3 (ORACLE PROPERTY). *Suppose that Assumptions 1–3 hold. If* $n^{1/2}\lambda_n \to 0$ *and* $n\lambda_n \to \infty$, $n^{1/2}(\hat{\theta}_1 - \theta_{10})$ *converges in distribution to* $N\{0, (\tilde{I}_{\theta_1})^{-1}\}$, *where* $\tilde{I}_{\theta_1}$ *is a submatrix of* $\tilde{I}_\theta$ *associated with* θ_{10}.

Theorem 2 demonstrates that the resulting estimator is $n^{1/2}$-consistent and selection consistent. With probability tending to 1, the estimator successfully recovers the true sparse model. Theorem 3 states that the resulting estimator of the nonzero coefficients is asymptotically normal with oracle variance, $(\tilde{I}_{\theta_1})^{-1}$. These theoretical properties are maintained with a different penalty other than the adaptive lasso when the tuning parameter conditions of Wang & Leng (2007) hold.

3. PROSPECTIVE DOUBLY CENSORED DATA

3.1. *Likelihood construction and factorization*

Suppose a prospective cohort study monitors n independent individuals. Each subject has a quadruplet of random variables: the enrolment time, the failure time, the study termination time and the covariate, denoted by (C_i, T_i, R_i, x_i), $i = 1, \ldots, n$. By definition, $C_i \leqslant R_i$. Our interest is the conditional distribution of the failure time given x_i, denoted by $F(T_i = t \mid x_i)$. The observed data are prospective doubly censored. One observes whether subjects experienced the event before enrolment at time C_i. If not, one continues observing whether an event occurred during follow-up to time R_i and the failure time at which such an event occurred. Denote the left censoring status at enrolment and the right censoring status with two indicators, $\delta_i = I(T_i \leqslant C_i)$ and $\nu_i = I(T_i \leqslant R_i)$, respectively. The censoring indicator pair has three possible values: $(0,\ 0)$, $(0,\ 1)$, and $(1,\ 1)$. Assume that T_i and $(C_i,\ R_i)$ are independent given x_i. The joint distribution of $(C_i,\ R_i)$ is assumed free of parameters in the conditional distribution of T_i. Denote the minimum of the failure time and the right censoring time as $Y_i = T_i \wedge R_i$. The observed data of each subject are $(C_i,\ C_i \vee Y_i,\ \delta_i,\ \nu_i, x_i)$. Contrary to conventional doubly censored data, where the left censoring time C_i is only observed on subjects with $T_i \leqslant C_i$, with prospectively doubly censored data, C_i is always observed.

The likelihood function for prospective doubly censored data is

$$\prod_{i=1}^{n} F(C_i \mid x_i)^{\delta_i} f(C_i \vee Y_i \mid x_i)^{\nu_i(1-\delta_i)} \{1 - F(C_i \vee Y_i \mid x_i)\}^{(1-\nu_i)(1-\delta_i)}.$$

The likelihood contains the same information as the likelihood for conventional doubly censored data since $C_i = C_i \vee Y_i$ for subjects who already had the event at enrolment. The use of the enrolment time, C_i, on all subjects facilitates the decomposition of the likelihood as follows:

$$\prod_{i=1}^{n} F(C_i \mid x_i)^{\delta_i} \{1 - F(C_i \mid x_i)\}^{1-\delta_i} \prod_{i=1}^{n} \left\{ \frac{f(Y_i \mid x_i)}{1 - F(C_i \mid x_i)} \right\}^{\nu_i(1-\delta_i)} \left\{ \frac{1 - F(Y_i \mid x_i)}{1 - F(C_i \mid x_i)} \right\}^{(1-\nu_i)(1-\delta_i)}.$$

The first component is the likelihood of current status data at enrolment, in which each individual has (C_i, δ_i, x_i) as its observed triplet. The second component is the likelihood of left-truncated right-censored data for the subjects who did not have the event prior to enrolment, in which each individual has (C_i, Y_i, ν_i, x_i) as its observed quadruplet.

We employ the proportional hazards model for the distribution of T_i given x_i. The conditional hazard rate is assumed to satisfy $h(t \mid x) = h(t)\exp(x\beta)$, where $h(t)$ is the baseline hazard at time t and β is the regression parameter. The parameter space Θ of the regression parameter is a known compact subset of $\mathbb{R}^p$. The baseline hazard rate can be an arbitrary nonnegative function of t. We define the baseline cumulative hazard function by $H(t) = \int_0^t h(u)\,\mathrm{d}u$. The true regression parameter and the true baseline cumulative hazard function are denoted as β_0 and H_0. Without loss of generality, write $\beta_0 = (\beta_{10}^{\mathrm{T}}, \beta_{20}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p$, where β_{10} is an s-dimensional nonzero parameter vector and β_{20} is a $(p-s)$-dimensional zero parameter vector. The corresponding loglikelihood is expressed as:

$$\begin{aligned} l_n^{\mathrm{PDCD}}(\beta, H) &= \sum_{i=1}^{n} \delta_i \log[1 - \exp\{-\exp(x_i^{\mathrm{T}}\beta)H(C_i)\}] - \sum_{i=1}^{n}(1-\delta_i)\exp(x_i^{\mathrm{T}}\beta)H(C_i) \\ &\quad + \sum_{i=1}^{n} \nu_i(1-\delta_i)\{x_i^{\mathrm{T}}\beta + \log h(Y_i) - \exp(x_i^{\mathrm{T}}\beta)H(Y_i) + \exp(x_i^{\mathrm{T}}\beta)H(C_i)\} \\ &\quad + \sum_{i=1}^{n}(1-\nu_i)(1-\delta_i)\{-\exp(x_i^{\mathrm{T}}\beta)H(Y_i) + \exp(x_i^{\mathrm{T}}\beta)H(C_i)\} \qquad (1) \\ &= \sum_{i=1}^{n} l_{\beta,H}^{\mathrm{CS}}(C_i, \delta_i \mid x_i) + \sum_{i=1}^{n} l_{\beta,H}^{\mathrm{LTRC}}(Y_i, \nu_i \mid C_i, \delta_i, x_i) \\ &= l_n^{\mathrm{CS}}(\beta, H) + l_n^{\mathrm{LTRC}}(\beta, H). \qquad (2) \end{aligned}$$

The first two terms of (1) correspond to the loglikelihood for current status data, while the remaining two terms of (1) correspond to the loglikelihood for left-truncated right-censored data. The full efficient score and information for β are denoted as $\tilde{l}_\beta$ and $\tilde{I}_\beta = E\{\tilde{l}_\beta(\tilde{l}_\beta)^{\mathrm{T}}\}$. The efficient score and information for the current status component are $\tilde{l}_\beta^{\mathrm{CS}}$ and $\tilde{I}_\beta^{\mathrm{CS}} = E\{\tilde{l}_\beta^{\mathrm{CS}}(\tilde{l}_\beta^{\mathrm{CS}})^{\mathrm{T}}\}$, respectively. Similarly, denote the efficient score and information for the left-truncated right-censored data component as $\tilde{l}_\beta^{\mathrm{LTRC}}$ and $\tilde{I}_\beta^{\mathrm{LTRC}} = E\{\tilde{l}_\beta^{\mathrm{LTRC}}(\tilde{l}_\beta^{\mathrm{LTRC}})^{\mathrm{T}}\}$.

3.2. *Application of ensemble methodology*

We apply the ensemble estimation and selection procedure for the Cox model with prospective doubly censored data based on the likelihood decomposition into the likelihood for the current status data and the likelihood for left-truncated right-censored data. Maximum likelihood estimation of both the regression parameter and the baseline hazard function has been extensively studied for both components (Huang, 1996; Andersen et al., 1997; Klein & Moeschberger, 2003; Sun, 2007), yielding efficient estimators for β which aid the application of the ensemble methodology.

For the current status data, denote the maximum likelihood estimator of the regression parameter and the baseline cumulative hazard function from $l_n^{\mathrm{CS}}(\beta, H)$ by $\hat{\beta}_F^{\mathrm{CS}}$ and $\hat{H}_F^{\mathrm{CS}}$. We maximize the current status data likelihood using the iterative convex minorant algorithm (Pan, 1999; Murphy & van der Vaart, 2000). A covariance estimator is obtained by bootstrap, denoted by $(\hat{I}_F^{\mathrm{CS}})^{-1} = B^{-1}\sum_{b=1}^{B}\{\hat{\beta}_F^{\mathrm{CS}}(b) - B^{-1}\sum_{b=1}^{B}\hat{\beta}_F^{\mathrm{CS}}(b)\}\{\hat{\beta}_F^{\mathrm{CS}}(b) - B^{-1}\sum_{b=1}^{B}\hat{\beta}_F^{\mathrm{CS}}(b)\}^{\mathrm{T}}$, where $\hat{\beta}_F^{\mathrm{CS}}(b)$

is the bootstrap regression estimate from the bth bootstrap sample (Pan, 1999; Cheng & Huang, 2010).

Since left-censored subjects have no contribution to the likelihood with left-truncated right-censored data, the likelihood for the Cox model may be rewritten with only left-uncensored subjects. Partial likelihood may be used to obtain the maximum likelihood estimator of the regression parameter with left-truncated right-censored data, without simultaneous estimation of H (Klein & Moeschberger, 2003). Denote the estimator of the regression parameter by $\hat{\beta}_F^{\mathrm{LTRC}}$. The negative second derivative of the log partial likelihood is used to estimate the asymptotic inverse covariance matrix for $\hat{\beta}_F^{\mathrm{LTRC}}$. We scale the negative second derivative by the proportion of left-uncensored subjects to address the zero contribution of left-censored subjects. The estimator is denoted by $\hat{I}_F^{\mathrm{LTRC}}$ (Klein & Moeschberger, 2003).

We now illustrate the ensemble estimation and variable selection procedure. First, ensemble estimation is performed with the initial estimators, $\hat{\beta}_F^{\mathrm{CS}}$, $\hat{\beta}_F^{\mathrm{LTRC}}$, $\hat{I}_F^{\mathrm{CS}}$ and $\hat{I}_F^{\mathrm{LTRC}}$. The full ensemble estimator is $\hat{\beta}_F = (\hat{I}_F^{\mathrm{CS}} + \hat{I}_F^{\mathrm{LTRC}})^{-1}(\hat{I}_F^{\mathrm{CS}}\hat{\beta}_F^{\mathrm{CS}} + \hat{I}_F^{\mathrm{LTRC}}\hat{\beta}_F^{\mathrm{LTRC}})$ and its inverse covariance estimator is $\hat{I}_F = \hat{I}_F^{\mathrm{CS}} + \hat{I}_F^{\mathrm{LTRC}}$. The ensemble variable selection is implemented by computing

$$\hat{\beta}_{E,\lambda_n} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}(\beta - \hat{\beta}_F)^{\mathrm{T}}\hat{I}_F(\beta - \hat{\beta}_F) + \lambda_n \sum_{j=1}^{p} |\beta_j|/|\hat{\beta}_{Fj}|,$$

where λ_n is a tuning parameter. The tuning parameter minimizing modified BIC is chosen. The selected model is denoted as $\mathcal{M} = \{j : \hat{\beta}_{E,\lambda_n,j} \neq 0\}$. Then, to obtain the refitted estimators, we maximize the component likelihoods over a parameter subspace, $L = \{\beta \in \mathbb{R}^p, \beta_j = 0, \text{ for all } j \notin \mathcal{M}\} \subset \mathbb{R}^p$. The subvectors of the refitted estimators indexed by $\mathcal{M}$ are denoted as $\hat{\beta}_{\mathcal{M}}^{\mathrm{CS}} = (\hat{\beta}_j^{\mathrm{CS}}, j \in \mathcal{M})$ and $\hat{\beta}_{\mathcal{M}}^{\mathrm{LTRC}} = (\hat{\beta}_j^{\mathrm{LTRC}}, j \in \mathcal{M})$. We estimate their asymptotic covariance matrices in a similar manner to estimating the unpenalized asymptotic covariance matrices, and denote them by $\hat{I}_{\mathcal{M}}^{\mathrm{CS}}$ and $\hat{I}_{\mathcal{M}}^{\mathrm{LTRC}}$, respectively. Alternatively, we may use submatrices of $\hat{I}_F^{\mathrm{CS}}$ and $\hat{I}_F^{\mathrm{LTRC}}$ associated with $\mathcal{M}$. The refit ensemble estimator restricted to L is a weighted combination of the refitted estimators, denoted by $\hat{\beta} \in \mathbb{R}^p$, where $\hat{\beta}_{\mathcal{M}} = (\hat{I}_{\mathcal{M}}^{\mathrm{CS}} + \hat{I}_{\mathcal{M}}^{\mathrm{LTRC}})^{-1}(\hat{I}_{\mathcal{M}}^{\mathrm{CS}}\hat{\beta}_{\mathcal{M}}^{\mathrm{CS}} + \hat{I}_{\mathcal{M}}^{\mathrm{LTRC}}\hat{\beta}_{\mathcal{M}}^{\mathrm{LTRC}})$ and $\hat{\beta}_{\mathcal{M}^C} = 0$. The inverse of its covariance matrix restricted to $\mathcal{M}$ is estimated by $\hat{I}_{\mathcal{M}} = (\hat{I}_{\mathcal{M}}^{\mathrm{CS}} + \hat{I}_{\mathcal{M}}^{\mathrm{LTRC}})$.

3.3. *Theoretical properties*

The likelihood decomposition with prospective doubly censored data for the Cox model is an exemplary case of the marginal and conditional likelihood decomposition introduced in § 2.2. The full data likelihood for the Cox model is the product of the marginal current status data likelihood and the conditional left-truncated right-censored data likelihood for the Cox model as in (2). By defining $Z \equiv (Y, \nu)$ and $W \equiv (C, \delta)$, the uncorrelatedness conditions are met as a special instance of the decomposition into marginal and conditional likelihoods, which is discussed in § 2.2. Hence, under Proposition 1, $\tilde{I}_\beta$ is the sum of the uncorrelated $\tilde{I}_\beta^{\mathrm{CS}}$ and $\tilde{I}_\beta^{\mathrm{LTRC}}$.

Under additional regularity conditions on the component estimation procedures, such that Assumptions 1–2 hold, consistency, asymptotic normality and semiparametric efficiency of the full ensemble estimator are achieved following Theorem 1.

COROLLARY 1. *Under Conditions* 4–10 *of the Supplementary Material,* $n^{1/2}(\hat{\beta}_F - \beta_0)$ *converges in distribution to* $N\{0, (\tilde{I}_\beta)^{-1}\}$.

Conditions 4–10 in the Supplementary Material are regularity conditions that allow semiparametric efficient estimation of the regression parameter and consistent estimation of the asymptotic

inverse variance for both of the component data likelihoods. For details, see Huang (1996), Andersen et al. (1997), Murphy & van der Vaart (1999), van der Vaart (2000) and Cheng & Huang (2010). Both of the component oracle estimators achieve the semiparametric efficiency bound, which makes Assumption 3 hold, as both loglikelihoods are functions of the linear predictor, $x_i^{\mathrm{T}}\beta$. Similarly to β_0, write $\hat{\beta} = (\hat{\beta}_1^{\mathrm{T}}, \hat{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\hat{\beta}_1 \in \mathbb{R}^p$ and $\hat{\beta}_2 \in \mathbb{R}^{p-s}$.

COROLLARY 2. *Under Conditions* 4–10 *of the Supplementary Material, if* $n^{1/2}\lambda_n \to 0$ *and* $n\lambda_n \to \infty$, *then* $n^{1/2}(\hat{\beta} - \beta_0) = O_p(1)$ *and* $\mathrm{pr}(\hat{\beta}_2 = 0) \to 1$.

COROLLARY 3. *Under Conditions* 4–10 *of the Supplementary Material, if* $n^{1/2}\lambda_n \to 0$ *and* $n\lambda_n \to \infty$, *then* $n^{1/2}(\hat{\beta}_1 - \beta_{10})$ *converges in distribution to* $N\{0, (\tilde{I}_{\beta_1})^{-1}\}$, *where* $\tilde{I}_{\beta_1}$ *is a submatrix of* $\tilde{I}_\beta$ *associated with* β_{10}.

Corollary 2 establishes that the resulting estimator has $n^{1/2}$-consistency and consistency in variable selection. Corollary 3 shows that the resulting estimator restricted to the nonzero parameters is asymptotically normally distributed and achieves a semiparametric efficiency bound.

A practical approach to estimating H may be to use Breslow's estimator with the left-truncated right-censored data, denoted as $\hat{H}_F^{\mathrm{LTRC}}$. The estimator achieves the regular convergence rate of $n^{1/2}$ since the left-truncated right-censored data contain the informative left-uncensored samples observed to experience the event during the study period, as do the prospective doubly censored data (Chang & Yang, 1987; Kim et al., 2010). Since the current status data lack the samples with the exact event time observed, the convergence rate of $\hat{H}_F^{\mathrm{CS}}$ is $n^{1/3}$, which is slower than the regular convergence rate.

4. SIMULATION STUDIES

Extensive simulation experiments were conducted to evaluate the finite sample performance of estimators of the regression parameters from our ensemble procedure. We consider estimators not only based on prospective doubly censored data, but based either on current status data or on left-truncated right-censored data for comparison. The component likelihood-based estimators include the component maximum likelihood estimators, least squares approximation estimators (Wang & Leng, 2007) and refit least squares approximation estimators. Since the expectation-maximization algorithms of Kim et al. (2013) and Su & Wang (2016) are not available in software packages, comparisons of our method with those methods based on the full likelihood were not attempted. We fit the current status data and left-truncated right-censored data using the `intcox` and `survival` R packages (R Development Core Team, 2020). We use 1000 bootstrap replicates to estimate the variance matrix with current status data.

We consider the following exponential hazard model: $h(t \mid x) = \exp(x^{\mathrm{T}}\beta)$, where $\beta = (0.8, 0, 0, 1, 0, 0, 0.6, 0, 0, 0)$. The covariates, x, were generated from a multivariate normal distribution, $N(0, \Sigma)$, where $\Sigma_{ij} = 0.5^{|i-j|}$. The enrolment time follows an exponential distribution and the right censoring time follows an exponential distribution shifted by the corresponding enrolment time. Results are given based on 500 simulated datasets. We consider two settings on left and right censoring rates: (20%, 20%) and (30%, 30%).

Table 1 summarizes the simulation results with sample sizes of 250 and 500, respectively. In component estimation based on current status data or left-truncated right-censored data, both least squares approximation and refit least squares approximation simultaneously perform variable selection and parameter estimation. Each least squares approximation estimate is a benchmark for the variable selection of the refit least squares approximation estimate, thus both share the

Table 1. *Comparison of the ensemble method and the component-based methods*

Censoring rates	(20%, 20%)						(30%, 30%)					
	RMSE	TP	FP	UF	CF	OF	RMSE	TP	FP	UF	CF	OF
						$n = 250$						
						CS						
Oracle	22.80	3.00	0.00	0	100	0	34.69	3.00	0.00	0	100	0
MLE	4.41	3.00	7.00	0	0	100	8.26	3.00	7.00	0	0	100
LSA	5.16	2.40	0.19	36	48	11	12.22	2.81	0.22	12	68	18
Refit LSA	5.78	2.40	0.19	36	48	11	14.78	2.81	0.22	12	68	18
						LTRC						
Oracle	73.45	3.00	0.00	0	100	0	63.46	3.00	0.00	0	100	0
MLE	16.56	3.00	7.00	0	0	100	13.53	3.00	7.00	0	0	100
LSA	48.73	3.00	0.21	0	82	18	35.53	3.00	0.34	0	73	27
Refit LSA	49.31	3.00	0.21	0	82	18	36.29	3.00	0.34	0	73	27
						Ensemble						
Oracle	100	3.00	0.00	0	100	0	100	3.00	0.00	0	100	0
Initial ensemble	22.78	3.00	7.00	0	0	100	23.78	3.00	7.00	0	0	100
EVS	63.04	3.00	0.11	0	90	10	57.37	3.00	0.12	0	89	11
CS refit	20.80	3.00	0.11	0	90	10	30.62	3.00	0.12	0	89	11
LTRC refit	58.59	3.00	0.11	0	90	10	53.13	3.00	0.12	0	89	11
Refit ensemble	75.68	3.00	0.11	0	90	10	78.40	3.00	0.12	0	89	11
						$n = 500$						
						CS						
Oracle	24.27	3.00	0.00	0	100	0	39.80	3.00	0.00	0	100	0
MLE	5.38	3.00	7.00	0	0	100	9.51	3.00	7.00	0	0	100
LSA	9.79	2.98	0.27	2	76	22	16.55	3.00	0.25	0	79	21
Refit LSA	13.08	2.98	0.27	2	76	22	23.72	3.00	0.25	0	79	21
						LTRC						
Oracle	77.02	3.00	0.00	0	100	0	60.45	3.00	0.00	0	100	0
MLE	18.40	3.00	7.00	0	0	100	15.59	3.00	7.00	0	0	100
LSA	54.87	3.00	0.17	0	86	14	42.19	3.00	0.17	0	86	14
Refit LSA	52.54	3.00	0.17	0	86	14	44.02	3.00	0.17	0	86	14
						Ensemble						
Oracle	100	3.00	0.00	0	100	0	100	3.00	0.00	0	100	0
Initial ensemble	24.08	3.00	7.00	0	0	100	24.96	3.00	7.00	0	0	100
EVS	66.67	3.00	0.07	0	93	7	57.45	3.00	0.11	0	90	10
CS refit	21.12	3.00	0.07	0	93	7	32.08	3.00	0.11	0	90	10
LTRC refit	66.31	3.00	0.07	0	93	7	52.60	3.00	0.11	0	90	10
Refit ensemble	77.99	3.00	0.07	0	93	7	73.97	3.00	0.11	0	90	10

CS, analyses based only on the current status data; LTRC, analyses based only on the left-truncated right-censored data; Ensemble: analyses based on the ensemble procedure; Initial ensemble: the full ensemble estimation; EVS, the ensemble variable selection procedure; Refit on LTRC/CS, the refitting based on LTRC/current status data; Refit ensemble: the ensemble re-estimation; RMSE: relative mean squared errors (%) of the estimators to the ensemble oracle estimator; TP/FP, average number of true positives/false positives, respectively; UF, CF and OF, percentages of underfitting, correct fitting and overfitting to the true model. Larger RMSE correspond to higher efficiency. Ensemble oracle estimator has RMSE of 100%.

same average number of true positives and false positives, and proportion of over/underfittings. The refitting increases efficiency in most cases. Estimation based on left-truncated right-censored data has a superior performance over current status data based estimation in terms of efficiency and variable selection, which is attributable to the fact that 75% of left-uncensored samples have the exact failure time observed. The ensemble oracle estimator is an efficient combination of the oracle estimators of both current status data and left-truncated right-censored data, which is a practical proxy to the oracle estimator with prospective doubly censored data. The ensemble variable

selection procedure successfully specifies the correct model in over 90% of the simulations, with reduced selection of true zero covariates than component estimation. The intermediate refitting lowers efficiency in estimation. However, the resulting estimators have resilience, and it is the most efficient among all but the ensemble oracle estimators. As the sample size increases, all procedures have higher accuracy in estimation and variable selection.

We also examine the finite sample performance of the asymptotically valid variance estimators of the regression parameter estimators with comparison to empirical variances. In the Supplementary Material we show that average estimated standard errors of $\hat{\beta}_1$ obtained throughout the procedure for $\beta_1 = 0.8$ are in good agreement with empirical standard errors of $\hat{\beta}_1$ with sample sizes 250 and 500, except for current status data based estimation. The noticeable discrepancy between average estimated standard errors and empirical standard errors of the estimation based on current status data is mainly driven by poor variable selection performance for β_1 of the least squares approximation technique. The average estimated standard errors of the refitted ensemble estimators are within $\pm 6\%$ of the corresponding empirical standard errors. The 95% empirical coverage probabilities of $\hat{\beta}_1$ for β_1 from the ensemble procedure are generally close to the nominal 95% level. Other results are quite similar and are omitted.

Further, we empirically computed covariances between $\hat{\beta}_F^{\mathrm{CS}}$ and $\hat{\beta}_F^{\mathrm{LTRC}}$, which are asymptotically zero matrices. In the Supplementary Material we also report the Frobenius norm, spectral norm and L_1 norm of the empirical covariances from all the scenarios we consider. The norms are rather small and decrease in magnitude as the sample size increases.

5. MULTICENTER AIDS COHORT STUDY

The Multicenter AIDS Cohort Study was initiated to elucidate the natural history of HIV (Kaslow et al., 1987). Nearly 5619 homosexual and bisexual men were enrolled across the United States. Every six months, the participants underwent a physical examination and completed questionnaires and laboratory testing. The studies collected extensive information on participants' demographics, sexual behaviours and medical histories. We considered the information from their first visit as possible risk factors. The seropositivity for HIV type 1 was determined by positive enzyme-linked immunosorbent assays with confirmatory Western blots (Kaslow et al., 1987). We analysed the time to HIV infection on the age scale. Subjects with information missing or record errors were dropped from the analysis. We also excluded subjects whose time gap between the last negative seroconversion visit and the first positive seroconversion visit exceeded 4 years. The analytic dataset included 5102 subjects; 2038 were HIV infected prior to their first visit, 448 became infected during the course of the study, and 2616 were not infected either prior to the study or during follow-up. The risk factors considered are participants' sexual behaviour, medical histories, smoking and drinking behaviour, drug usage, and socioeconomic status.

Table 2 presents the analysis results from the entire ensemble procedure. For comparison, in the Supplementary Material we report the results of the unpenalized maximum likelihood estimation, least squares approximation and refit least squares approximation based either on current status data or on left-truncated right-censored data. While the left-truncated right-censored data analysis concludes that the highest educational risk group is people who have attended college with no degree, both the current status data based estimation and the refitted ensemble estimation conclude that less educated people have higher HIV risk, which agrees with previous findings (Catania et al., 2001; Simard et al., 2012). Further, these analyses select genital warts, cocaine use and Hispanic ethnicity as risk factors, contrary to the left-truncated right-censored data based estimation. These differences may be explained by the fact that roughly 40% of subjects were HIV infected before enrolment and only 9% of participants without HIV at enrolment contracted HIV during the

Table 2. *Results from the ensemble method on the Multicenter AIDS Cohort Study*

Covariates	Initial ensemble (SE)		EVS	LTRC refit (SE)	CS refit (SE)	Refit ensemble (SE)
REC2P	0.58 (0.05)	*	0.58	0.44 (0.12)	0.59 (0.05)	0.57 (0.05)
REC2Y	0.06 (0.05)		—	—	—	—
CON2P	0.07 (0.23)		—	—	—	—
CON2Y	−0.24 (0.21)		—	—	—	—
DIABE	−0.23 (0.29)		—	—	—	—
GONOE	0.53 (0.08)	*	0.50	0.55 (0.10)	0.50 (0.14)	0.62 (0.08)
RADTE	−0.03 (0.21)		—	—	—	—
WARTE	0.37 (0.05)	*	0.33	0.05 (0.11)	0.39 (0.05)	0.34 (0.05)
NDRNK	−0.02 (0.01)		—	—	—	—
PACKS	−0.03 (0.03)		—	—	—	—
NEEDL	0.54 (0.10)	*	0.48	0.84 (0.22)	0.45 (0.12)	0.57 (0.10)
COK2Y	0.49 (0.06)	*	0.52	0.11 (0.10)	0.55 (0.09)	0.41 (0.07)
HAS2Y	0.12 (0.08)	*	—	—	—	—
MSX2Y	0.31 (0.08)	*	0.28	0.56 (0.14)	0.27 (0.09)	0.24 (0.07)
OPI2Y	−0.05 (0.12)		—	—	—	—
UNEMP	0.19 (0.09)		—	—	—	—
BLACK	0.71 (0.08)	*	0.66	0.52 (0.18)	0.69 (0.08)	0.70 (0.08)
HISPA	0.28 (0.10)	*	0.16	0.18 (0.22)	0.32 (0.13)	0.20 (0.11)
OTHER	0.06 (0.20)		—	—	—	—
PRECOL	−0.16 (0.07)	*	—	—	—	—
COL	−0.31 (0.08)	*	−0.13	−0.31 (0.13)	−0.20 (0.06)	−0.24 (0.06)
POSTCOL	−0.44 (0.07)	*	−0.30	−0.10 (0.12)	−0.43 (0.11)	−0.29 (0.07)

REC2P, *REC2Y*, whether participants had anal receptive/insertive sex; *CON2P*, *CON2Y*, whether participants had anal receptive/insertive sex with condom; *DIABE*, diabetes; *GONOE*, gonorrhoea; *RADTE*, radiation therapy/treatment; *WARTE*, genital/anal warts; *NDRNK*, number of drinks per day; *PACKS*, number of cigarette packs smoking per day; *NEEDL*, needle sharing; *COK2Y*, cocaine use; *HAS2Y*, marijuana/hashish use; *MSX2Y*, drugs with sex; *OPI2Y*, heroin/other opiates use; *UNEMP*, current unemployment; *BLACK*, black ethnicity; *HISPA*, hispanic ethnicity; *OTHER*, the other ethnicities; *PRECOL*, college attendance with no degree; *COL*, bachelor's degree; *POSTCOL*, master's degree and above. White ethnicity and high school diploma were used as base categories. Based on the initial ensemble estimator, significant covariates are marked with * at level 0.05.

study period. This suggests that the current status data is more informative than the left-truncated right-censored data, as reflected in the ensemble results. Based on the ensemble procedure, anal receptive sex is strongly associated with HIV infection with a 77% increase in risk. The high risk of HIV infection for anal receptive sex without a condom has been well documented (Ekstrand et al., 1999; Sullivan et al., 2009). Subjects who have had sexual diseases such as gonorrhoea and genital warts are also seen to be at higher risk, by 86% and 40%, respectively. In addition, as expected, needle sharing, cocaine use and drug use with sex increase HIV infection risk, by 77%, 51% and 27%, respectively. African Americans and poorly educated subjects have similar expected increases in HIV risk.

6. Discussion

The ensemble methodology for semiparametric factorizable likelihoods has merits in both reducing computational burden and simplifying asymptotic inference. The theoretical results were established in a paradigm where the sample size increases while the number of covariates is fixed. Certainly, such results are not applicable when the number of covariates grows with

the sample size, as in high-dimensional data applications. As the associate editor pointed out, in the presence of several nonzero parameters whose values are less than the order of $n^{-1/2}$, the ensemble variable selection might not succeed in identifying the correct model, which leads to efficiency loss in the follow-up ensemble estimation. Further work is needed in this area.

For prospective doubly censored data, the likelihood factorization readily accommodates models other than the proportional hazards model where efficient estimators and consistent variance estimators exist. Efficient estimation of the regression parameter and consistent estimation of the asymptotic variance in the accelerated failure time model have previously been studied with left truncation and right censoring by Lai & Ying (1991), and with current status data by Shen (2000). Such methods might be employed to construct a fully efficient ensemble estimator of the regression parameter in this model. As another example, efficient estimation of the regression parameter and consistent estimation of the asymptotic variance in a general class of transformation models, including the proportional hazards model, is, in principle, accomplished with left-truncated right-censored data using the counting process likelihood estimators in Zeng & Lin (2007). Efficient estimation of transformation models with current status data has been rigorously studied in Zhang et al. (2013). The ensemble methodology might be utilized in conjunction with these component estimators for fully efficient estimation of the regression parameter in the transformation model. Implementations of these ensemble procedures are important topics for future research.

The ensemble methodology is similar in spirit to meta-analysis, where one combines information across a fixed number K component likelihoods. Although there is growing interest in meta-analysis using individual-level data, such data may not be available for many of the small biomedical studies reported in the literature. The issue was addressed by leveraging efficient study-specific estimates of a common parameter or its different, but possibly overlapping, elements (Lin & Zeng, 2010; Liu et al., 2015; Slud et al., 2018; Kundu et al., 2019). Similarly, our proposed methods are potentially useful in these applications, if one assumes the number of studies is fixed, with the number of observations in each study tending to infinity. Frequently, the number of studies is small relative to the sample sizes within studies in meta-analysis (Riley et al., 2010). The ensemble procedure yields efficient overall estimates of regression parameters by optimally combining efficient study-specific estimates.

The ensemble procedure provides efficient and oracle estimators of regression parameters with minimal assumptions on estimation of the nuisance parameter. The regularity conditions permit nonstandard and potentially different rates of convergence for the component estimators of the nuisance parameter, as with the Cox model with prospective doubly censored data. A challenging topic which merits further investigation is whether optimal estimation of the nuisance parameter might be achieved by combining component estimators.

Acknowledgement

The authors thank Anthony Davison, the editor, an associate editor and a referee for their constructive comments that substantially improved the paper. This work was supported by the U.S. National Institutes of Health and National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes all proofs of asymptotic properties for the ensemble method and detailed numerical results.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1997). *Statistical Models Based on Counting Processes*. New York: Springer.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester: John Wiley & Sons.

Cai, J., Fan, J., Li, R. & Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303–16.

Catania, J. A., Osmond, D., Stall, R. D., Pollack, L., Paul, J. P., Blower, S., Binson, D., Canchola, J. A., Mills, T. C., Fisher, L., et al. (2001). The continuing HIV epidemic among men who have sex with men. *Am. J. Public Health.* **91**, 907–14.

Chang, M. N. & Yang, G. L. (1987). Strong consistency of a nonparametric estimator of the survival function with doubly censored data. *Ann. Statist.* **15**, 1536–47.

Cheng, G. & Huang, J. (2010). Bootstrap consistency for general semiparametric M-estimation. *Ann. Statist.* **38**, 2884–915.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–200.

Cox, D. R. (2001). Some remarks on likelihood factorization. In *State of the Art in Probability and Statistics: Festschrift for Willem R. van Zwet*, M. de Gunst, C. Klaassen & A. W. van der Vaart, eds. pp. 165–172. Beachwood, OH: IMS Lecture Notes – Monograph Series.

Du, P., Ma, S. & Liang, H. (2010). Penalized variable selection procedure for Cox models with semiparametric relative risk. *Ann. Statist.* **38**, 2092–117.

Efron, B., Hastie, T. J., Johnstone, I. M. & Tibshirani, R. J. (2004). Least angle regression. *Ann. Statist.* **32**, 407–99.

Ekstrand, M. L., Stall, R. D., Paul, J. P., Osmond, D. H. & Coates, T. H. (1999). Gay men report high rates of unprotected anal sex with partners of unknown or discordant HIV status. *AIDS* **13**, 1525–33.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fan, J. & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74–99.

Groeneboom, P. & Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel: Birkhäuser Verlag.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540–68.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F. & Rinaldo, C. R. (1987). The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *Am. J. Epidemiol.* **126**, 310–18.

Keller, T. & Olkin, I. (2004). Combining correlated unbiased estimators of the mean of a normal distribution. In *A Festschrift for Herman Rubin*, A. DasGupta, ed. pp. 218–227. Beachwood, OH: IMS Lecture Notes – Monograph Series.

Kim, Y., Kim, B. & Jang, W. (2010). Asymptotic properties of the maximum likelihood estimator for the proportional hazards model with doubly censored data. *J. Mult. Anal.* **101**, 1339–51.

Kim, Y., Kim, J. & Jang, W. (2013). An EM algorithm for the proportional hazards model with doubly censored data. *Comp. Statist. Data Anal.* **57**, 41–51.

Klein, J. P. & Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.

Kosorok, M. R. (2007). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Kundu, P., Tang, R. & Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106**, 567–85.

Lai, T. L. & Ying, Z. (1991). Rank regression methods for left-truncated and right-censored data. *Ann. Statist.* **19**, 531–56.

Lin, D. Y. & Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97**, 321–32.

Liu, D., Liu, R. Y. & Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *J. Am. Statist. Assoc.* **110**, 326–40.

Liu, X. & Zeng, D. (2013). Variable selection in semiparametric transformation models for right-censored data. *Biometrika* **100**, 859–76.

Lu, W., Goldberg, Y. & Fine, J. P. (2012). On the robustness of the adaptive lasso to model misspecification. *Biometrika* **99**, 717–31.

Murphy, S. A. & van der Vaart, A. W. (1999). Observed information in semiparametric models. *Bernoulli* **5**, 381–412.

Murphy, S. A. & van der Vaart, A. W. (2000). On profile likelihood. *J. Am. Statist. Assoc.* **95**, 449–65.

NEWEY, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Economet.* **5**, 99–135.
PAN, W. (1999). Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. *J. Comp. Graph. Statist.* **8**, 109–20.
R DEVELOPMENT CORE TEAM (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
RILEY, R. D., LAMBERT, P. C. & ABO-ZAID, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *Br. Med. J.* **340**, c221.
SHEN, X. (2000). Linear regression with current status data. *J. Am. Statist. Assoc.* **95**, 842–52.
SIMARD, E. P., FRANSUA, M., NAISHADHAM, D. & JEMAL, A. (2012). The influence of sex, race/ethnicity, and educational attainment on human immunodeficiency virus death rates among adults, 1993–2007. *Arch. Internal Med.* **172**, 1591–8.
SLUD, E., VONTA, I. & KAGAN, A. (2018). Combining estimators of a common parameter across samples. *Statist. Theory Rel. Fields* **2**, 158–71.
SU, Y.-R. & WANG, J.-L. (2016). Semiparametric efficient estimation for shared-frailty models with doubly-censored clustered data. *Ann. Statist.* **44**, 1298–331.
SULLIVAN, P. S., SALAZAR, L., BUCHBINDER, S. & SANCHEZ, T. H. (2009). Estimating the proportion of HIV transmissions from main sex partners among men who have sex with men in five US cities. *AIDS* **23**, 1153–62.
SUN, J. (2007). *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer.
TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.
TIBSHIRANI, R. J. (1997). The lasso method for variable selection in the Cox model. *Statist. Med.* **16**, 385–95.
VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. New York: Cambridge University Press.
WANG, H. & LENG, C. (2007). Unified LASSO estimation by least squares approximation. *J. Am. Statist. Assoc.* **102**, 1039–48.
ZENG, D. & LIN, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). *J. R. Statist. Soc. B* **69**, 507–64.
ZHANG, B., TONG, X., ZHANG, J. & WANG, C. (2013). Efficient estimation for linear transformation models with current status data (with discussion). *Commun. Statist. B* **42**, 3191–203.
ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

[*Received on* 21 *July* 2016. *Editorial decision on* 12 *July* 2019]