The Annals of Statistics
2020, Vol. 48, No. 4, 2132–2154
https://doi.org/10.1214/19-AOS1880
© Institute of Mathematical Statistics, 2020

DOUBLE-SLICING ASSISTED SUFFICIENT DIMENSION REDUCTION FOR HIGH-DIMENSIONAL CENSORED DATA

BY SHANSHAN DING[1,*], WEI QIAN[1,**] AND LAN WANG[2]

[1] *Department of Applied Economics and Statistics, University of Delaware,* [*] *sding@udel.edu;* [**] *weiqian@udel.edu*
[2] *Department of Management Science, University of Miami,* *lanwang@mbs.miami.edu*

This paper provides a unified framework and an efficient algorithm for analyzing high-dimensional survival data under weak modeling assumptions. In particular, it imposes neither parametric distributional assumption nor linear regression assumption. It only assumes that the survival time T depends on a high-dimensional covariate vector $\mathbf{X}$ through low-dimensional linear combinations of covariates $\Gamma^T\mathbf{X}$. The censoring time is allowed to be conditionally independent of the survival time given the covariates. This general framework includes many popular parametric and semiparametric survival regression models as special cases. The proposed algorithm produces a number of practically useful outputs with theoretical guarantees, including a consistent estimate of the sufficient dimension reduction subspace of $T \mid \mathbf{X}$, a uniformly consistent Kaplan–Meier-type estimator of the conditional distribution function of T and a consistent estimator of the conditional quantile survival time. Our asymptotic results significantly extend the classical theory of sufficient dimension reduction for censored data (particularly that of Li, Wang and Chen in *Ann. Statist.* **27** (1999) 1–23) and the celebrated nonparametric Kaplan–Meier estimator to the setting where the number of covariates p diverges exponentially fast with the sample size n. We demonstrate the promising performance of the proposed new estimators through simulations and a real data example.

1. Introduction.

Literature on high-dimensional data analysis has experienced an explosion recently. However, there still exists relatively little work, particularly with theoretical guarantees, for analyzing high-dimensional data with censored responses where new technical challenges arise. We are interested in studying the relationship between an event time T and a p-dimensional vector of predictors $\mathbf{X} = (X_1, \dots, X_p)^T$. The event time T may not be observed due to right censoring, such as patients dropout. Let $Y = \min(T, C)$ be the observed event time where C denotes the censoring variable, and let $\delta = I(T \le C)$ be the censoring indicator. The observed data consist of $(\mathbf{X}_i, Y_i, \delta_i)$, $i = 1, \dots, n$. In this paper, we develop a general theory for analyzing such censored data in the setting where the number of covariates p can be much larger than the sample size n. A distinguishable feature of our proposal is that we only impose a very general model framework instead of a specific model. In particular, we impose neither parametric distributional assumption nor linear regression assumption.

The basic modeling assumption we adopt is that T depends on $\mathbf{X}$ only through a few linear combinations of covariates $\Gamma^T\mathbf{X}$, where Γ is a $p \times d$ $(d \le p)$ matrix with d usually much smaller than p. Alternatively, we write

$$T \perp\!\!\!\perp \mathbf{X} \mid \Gamma^T\mathbf{X}, \tag{1.1}$$

where $\perp\!\!\!\perp$ stands for independence. The matrix Γ itself is not identifiable as for any $d \times d$ nonsingular matrix A, $A^T\Gamma^T\mathbf{X}$ also satisfies (1.1). Instead, we aim to estimate the smallest

Received February 2018; revised June 2019.
MSC2020 subject classifications. Primary 62H20, 62J07; secondary 62N02.
Key words and phrases. Central subspace, sufficient dimension reduction, variable selection, censored data, ultrahigh dimension, nonparametric estimation.

linear space spanned by the columns of Γ, denoted by $\mathcal{S}_{T|\mathbf{X}}$. In the dimension reduction literature, such a space is referred to as the central subspace, and is known to exist and be unique under mild conditions (Cook (1998)).

Note that (1.1) is equivalent to the statement $F(T\,|\,\mathbf{X}) = F(T\,|\,\Gamma^T\mathbf{X})$, where $F(T\,|\,\mathbf{X})$ and $F(T\,|\,\Gamma^T\mathbf{X})$ are conditional distribution functions of T given $\mathbf{X}$ and $\Gamma^T\mathbf{X}$, respectively. To appreciate the flexibility of this general formulation, we observe that (1.1) encompasses many popular survival analysis model assumptions as special cases regarding $T\,|\,\mathbf{X}$:

- *Proportional hazards or Cox model*: $h(t\,|\,\mathbf{X}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X})$, where the hazard function $h(t\,|\,\mathbf{X}) = -\frac{d}{dt}\log Q(t\,|\,\mathbf{X})$ with $Q(t\,|\,\mathbf{X}) = 1 - F(t\,|\,\mathbf{X})$ being the conditional survival function. This model is equivalent to $F(t\,|\,\mathbf{X}) = 1 - \exp\{-H_0(t)e^{\boldsymbol{\beta}^T\mathbf{X}}\}$, where $H_0(t) = \int_0^t h_0(s)ds$. Hence, (1.1) is satisfied with $\Gamma = \boldsymbol{\beta}$.
- *Accelerated failure time (AFT) regression model*: $\log(T) = \boldsymbol{\beta}^T\mathbf{X} + \varepsilon$, where ε is the random error. Then it is easy to see that (1.1) is satisfied with $\Gamma = \boldsymbol{\beta}$.
- Various semiparametric variants of Cox or AFT model, for example, $\log(T) = \boldsymbol{\beta}_1^T\mathbf{X} + g_1(\boldsymbol{\beta}_2^T\mathbf{X}) + g_2(\boldsymbol{\beta}_3^T\mathbf{x})\varepsilon$, where g_2 is a nonnegative function. The vectors $\boldsymbol{\beta}_i$, $i = 1, \dots, 3$, are unknown; and g_1 and g_2 are also possibly unknown. For this semiparametric regression model, (1.1) is satisfied with Γ whose columns are $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$.

The purpose of this paper is twofold. First, we will estimate the central subspace $\mathcal{S}_{T|\mathbf{X}}$ for high-dimensional censored data, an important problem that has not been explored much in the literature due to technical challenges. The central subspace is known to be a powerful tool for data reduction and visualization. However, it does not directly provide prediction, which is often a main objective for real data analysis. Our second goal is therefore to estimate the conditional distribution function $F(T\,|\,\mathbf{X})$ in the high-dimensional setting while accounting for censoring. This would help answer important practical questions such as what is the probability that a patient can survive more than 6 months given his/her clinical conditions and genetic profile.

2. Related work and contribution.

There exists a rich literature on central subspace estimation for dimension reduction; see Chen, Cook and Zou (2015), Cook and Ni (2005), Hsing and Ren (2009), Kong and Xia (2014), Li (1991), Li (2007), Li and Dong (2009), Li and Wang (2007), Ma and Zhu (2012), Ye and Weiss (2003), Yin and Li (2011), Zhu, Miao and Peng (2006), Bura, Duarte and Forzani (2016), among many important others. More recently, progress has been made in high-dimensional setting ($p > n$) (e.g., Cook, Forzani and Rothman (2012), Li and Yin (2008), Lin, Zhao and Liu (2018), Qian, Ding and Cook (2018), Tan et al. (2018), Wang et al. (2018), Yin and Hilafu (2015), Yu, Dong and Shao (2016), Yu et al. (2013)). However, these methods were mainly designed for complete data and do not apply to data with censored outcomes. When the response variable is censored, several authors have made significant progresses for the classical setting (fixed p or p diverging but $p < n$); see, for example, Li, Wang and Chen (1999), Cook (2003), Li and Li (2004), Li (2005), Xia, Zhang and Xu (2010), Lu and Li (2011), Nadkarni, Zhao and Kosorok (2011), Lopez (2011), Sun et al. (2019) and Zhao, Ma and Lu (2017).

Our first main contribution is to significantly extend the existing theory on sufficient dimension reduction (SDR) for censored data. Based on our proposed new methodology, we establish both the central subspace estimation consistency and variable selection consistency in the ultrahigh-dimensional setting without stringent parametric distributional assumptions. Furthermore, these consistency properties are achieved under relatively mild conditions on the censoring mechanism: we assume the conditional independence condition $T \perp\!\!\!\perp C\,|\,\mathbf{X}$, as opposed to the more restrictive complete independence condition $T \perp\!\!\!\perp C$ or the strong marginally conditional independence condition $T \perp\!\!\!\perp C \mid X_j$, for all j, required by many screening methods.

To the best of our knowledge, our proposal is the first to extend SDR method to the analysis of high-dimensional censored data with theoretical guarantees, where p is allowed to increase at an exponential rate of n. It is worth noting that our approach is very different from existing model-based (mostly based on Cox model) variable selection methods (e.g., Bradic, Fan and Jiang (2011), Du, Ma and Liang (2010), Fan and Li (2002), Fang, Ning and Liu (2017), Huang et al. (2013), Johnson (2009), Tibshirani (1997), Zhang and Lu (2007), Chai et al. (2019)). Moreover, the proposed methods apply to a wide class of survival data models where the proportional hazards assumption is violated.

We also establish uniform convergence of a local Kaplan–Meier estimator in the high-dimensional setting with assistance of the estimated central subspace. The celebrated nonparametric Kaplan–Meier estimator for the conditional distribution of the event time plays a central role in survival analysis but requires the strong independent censoring assumption. On the other hand, important extensions of Kaplan–Meier estimator to covariate-dependent censoring, such as Beran (1981), only works with a few covariates in practical data analysis due to the curse of dimensionality. Our result much extends the practical use of Kaplan–Meier estimator to high-dimensional censored data while permitting covariate-dependent censoring.

In addition, we equip our method with an efficient algorithm for computation. It adopts an iterative strategy to solve the objective functions without inverting any large covariance matrix. Furthermore, we propose a specialized cross-validation method to achieve automatic tuning parameter and structural dimension selection. Its practical effectiveness is demonstrated through extensive numerical and real data evaluations.

The remainder of the paper is organized as follows. Section 3 proposes the new double-slicing assisted SDR method for studying high-dimensional censored data. Section 4 establishes consistency results in both central subspace estimation and variable selection under the ultrahigh-dimensional setting, as well as estimation consistency for conditional survival function. Computational aspects are presented in Section 5. Simulation studies and real data analysis are given in Sections 6 and 7, and concluding remarks are given in Section 8. Proofs, related technical details, and additional computational and numerical results are left to the supplement.

3. Methodology.

In this section, we introduce the *d*ouble-slicing *a*ssisted *S*DR method in *h*igh dimension (abbreviated as DASH) to estimate the central subspace $\mathcal{S}_{T\mid\mathbf{X}}$ and to simultaneously select important covariates for $T \mid \mathbf{X}$ in the ultrahigh-dimensional settings. DASH has two main steps. First, it estimates a cruder augmented central subspace $\mathcal{S}_{(T,C)\mid\mathbf{X}}$ for the conditional distribution of $(T, C) \mid \mathbf{X}$, inspired by the double slicing approach in Li, Wang and Chen (1999) for the classical setting $p < n$. It then provides a refined estimate of the targeted central subspace $\mathcal{S}_{T\mid\mathbf{X}}$ adjusting for censoring in the second step. Based on the central subspace estimation from DASH, we then obtain a nonparametric estimator of the conditional survival function, which also yields an estimator for the conditional quantile function as a byproduct.

3.1. *The DASH method.*

To motivate DASH, we provide some basic intuition for the ideal situation where we fully observe T and there are only fixed number of covariates. In this case, sliced inverse regression (SIR; Li (1991)) is a simple way to estimate the central subspace. Let $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$ and $\Sigma = \mathrm{cov}(\mathbf{X})$ be the mean and covariance matrix of $\mathbf{X}$, respectively. The key observation behind SIR is that $\mathrm{E}(\mathbf{X} - \boldsymbol{\mu} \mid T)$ is in the space $\Sigma\mathcal{S}_{T\mid\mathbf{X}}$ under the linearity condition in Li (1991), which holds for the elliptical distribution family and also holds to a good approximation in high dimension (Hall and Li (1993)). SIR reverses the relation between the response variable and the covariates. Let $\mathcal{I}_t = [a_t, a_{t+1})$, $t = 1, \dots, b$, be b nonoverlapping intervals of a partition of the range of T, where $0 = a_1 < \cdots < a_b < \infty = a_{b+1}$. For a given t, define $\boldsymbol{\xi}_t = \Sigma^{-1}\mathrm{E}(\mathbf{X} - \boldsymbol{\mu} \mid T \in \mathcal{I}_t)$ for $t = 1, \dots, b$. With $M = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_b) \in \mathbb{R}^{p\times b}$ and $\mathcal{S}_M = \mathrm{Span}(M)$, where $\mathrm{Span}(M)$ is the subspace spanned by the columns of M, we have $\mathcal{S}_M \subseteq \mathcal{S}_{T|\mathbf{X}}$. Furthermore, we assume the coverage condition $\mathcal{S}_M \supseteq \mathcal{S}_{T|\mathbf{X}}$, which is usually reasonable with large enough b. With $d = \dim(\mathcal{S}_M)$, the dimension of $\mathcal{S}_M$, Cook and Ni (2005) proposed to estimate $\mathcal{S}_M$ with a d-dimensional subspace that is closest to the columns of M, where the closeness is measured by a quadratic discrepancy function. This general approach subsumes SIR as a special case. However, this quadratic distance approach is not applicable to our setting due to the challenges of random censoring and high-dimensional covariates.

The DASH method we introduce below corrects the bias due to random censoring by employing inverse probability weighting. Such weighting intrinsically also depends on the high-dimensional covariates. We extend the double-slicing approach of Li, Wang and Chen (1999) to obtain estimates of such weights with theoretical guarantee in high dimension. Furthermore, DASH incorporates high-dimensional covariates in estimating $\mathcal{S}_{T|\mathbf{X}}$ by extending the quadratic discrepancy function with appropriate penalization and devising an algorithm that avoids inverting the high-dimensional covariance matrix.

For censored data, T may not be completely observed. Instead of slicing T, we consider $\mathcal{H}_y = [t_y, t_{y+1})$, $y = 1, \dots, b$, which form b nonoverlapping intervals of a partition of the observed survival time Y with $0 = t_1 < \cdots < t_b < \infty = t_{b+1}$. Let $S(t \mid \mathbf{x}) = P(C > t \mid \mathbf{X} = \mathbf{x})$ be the conditional survival function of the censoring time C given $\mathbf{X}$. Let $p_y = P(Y \in \mathcal{H}_y)$ and $\tilde{p}_y = P(T \in \mathcal{H}_y)$. It is not hard to verify (see Supplement I.1) that

$$\mathrm{E}(\mathbf{X} - \boldsymbol{\mu} \mid T \in \mathcal{H}_y) = \frac{p_y}{\tilde{p}_y}\mathrm{E}\left\{\frac{\delta(\mathbf{X} - \boldsymbol{\mu})}{S(Y \mid \mathbf{X})} \,\Big|\, Y \in \mathcal{H}_y\right\}. \tag{3.1}$$

Hence, one can correct the censoring bias by inverse probability weighting. Variations of inverse probability weighting for censored data were also used in Cheng, Wei and Ying (1995), Fan and Gijbels (1994), Lu and Li (2011), Nadkarni, Zhao and Kosorok (2011), among others, for classical fixed or small p cases. When the censoring distribution depends on the covariates, the aforementioned literature either imposed relatively strong assumptions such as complete independence or employed a nonparametric estimator for $S(Y \mid \mathbf{X})$ which may suffer from curse of dimensionality in practice. We instead propose a convenient double-slicing assisted procedure to first perform SDR for the joint conditional distribution of $(T, C) \mid \mathbf{X}$ and obtain a uniformly consistent estimator $S(t \mid \mathbf{X})$ nonparametrically under reduced predictor dimension. Let $\hat{S}(t \mid \mathbf{X})$ denote this estimator of $S(t \mid \mathbf{X})$. For now, we assume $\hat{S}(t \mid \mathbf{X})$ has been obtained. The details of estimating $\hat{S}(t \mid \mathbf{X})$ will be described in Section 3.3.

Let $\mathbf{m}_y = \mathrm{E}\{\delta(\mathbf{X} - \boldsymbol{\mu})/S(Y \mid \mathbf{X}) \mid Y \in \mathcal{H}_y\}$, $y = 1, \dots, b$. And define $M_c = (\mathbf{m}_1, \dots, \mathbf{m}_b) \in \mathbb{R}^{p\times b}$. Then (3.1) implies that $\mathcal{S}_{M_c} \subseteq \Sigma\mathcal{S}_{T|\mathbf{X}}$. Unless stated otherwise, we assume throughout the article that $\mathcal{S}_{M_c} \equiv \Sigma\mathcal{S}_{T|\mathbf{X}}$. Let $\mathbf{g} = (\sqrt{p_1}, \dots, \sqrt{p_b})$ and $D_{\mathbf{g}} = \mathrm{diag}(\mathbf{g})$, where $\mathrm{diag}(\mathbf{g})$ denotes the diagonal matrix with diagonal components to be the elements of $\mathbf{g}$. Construct a modified matrix $U_c = M_c D_{\mathbf{g}}$. Then we have $\mathcal{S}_{U_c} = \mathcal{S}_{M_c} = \Sigma\mathcal{S}_{T|\mathbf{X}}$ and U_c is called a kernel matrix. Let $J_y = \{1 \le i \le n : Y_i \in \mathcal{H}_y\}$ be the index set corresponding to slice $\mathcal{H}_y$ and let $N_y = |J_y|$, $y = 1, \dots, b$, where $|\cdot|$ denotes the set cardinality. Then the sample estimates of $\mathbf{m}_y$ and p_y can be formulated by $\hat{\mathbf{m}}_y = N_y^{-1}\sum_{i\in J_y}\delta_i(\mathbf{X}_i - \bar{\mathbf{X}})/\hat{S}(Y_i \mid \mathbf{X}_i)$ and $\hat{p}_y = N_y/n$, where $\bar{\mathbf{X}}$ is the sample mean of $\mathbf{X}$. Correspondingly, we can obtain sample estimates $\hat{M}_c$ and $\hat{\mathbf{g}}$ of M_c and $\mathbf{g}$, and set $\hat{U}_c = \hat{M}_c D_{\hat{\mathbf{g}}}$.

Let Γ_0 be a basis matrix of $\mathcal{S}_{T|\mathbf{X}}$ and let $\boldsymbol{\gamma}_{0j} \in \mathbb{R}^d$ be the jth row of Γ_0. With $\|\cdot\|_2$ being the L_2 norm, let $\mathcal{A}_0 \equiv \{1 \le j \le p : \|\boldsymbol{\gamma}_{0j}\|_2 > 0\}$. Then it is natural to define $\mathbf{X}_{\mathcal{A}_0}$ to be the active variables of $\mathcal{S}_{T|\mathbf{X}}$, where $\mathbf{X}_{\mathcal{A}_0}$ is the sub-vector of $\mathbf{X}$ corresponding to the index set $\mathcal{A}_0$. To achieve SDR and to simultaneously identify the active variables, we estimate Γ_0 by considering the penalized sample objective function

$$F_n(\Gamma, \Phi) = \operatorname{tr}\{(\hat{U}_c - \hat{\Sigma}_n\Gamma\Phi)^T\hat{\Omega}_n(\hat{U}_c - \hat{\Sigma}_n\Gamma\Phi)\} + \lambda\sum_{j=1}^{p} w_j\|\boldsymbol{\gamma}_j\|_2, \tag{3.2}$$

subject to $\Phi\Phi^T = I_d$, where $\operatorname{tr}(\cdot)$ is the trace operator, $\hat{\Sigma}_n$ is the sample covariance matrix of $\mathbf{X}$, $\hat{\Omega}_n$ denotes a sample estimate of Σ^{-1}, and $\boldsymbol{\gamma}_j$ is the jth row vector of Γ. It is important to stress that we avoid finding an explicit form of $\hat{\Omega}_n$ by setting $\hat{\Sigma}_n\hat{\Omega}_n = I_p$ in the sample objective function in our algorithm (Section 5). In (3.2), λ is a tuning parameter and w_j's are penalty weights. Note that each column vector in Γ_0 produces a linear combination of the p covariates, and the above penalty hence encourages sparse linear combinations of all covariates.

REMARK 3.1. We assume throughout the paper that central subspaces $\mathcal{S}_{T|\mathbf{X}}$ and $\mathcal{S}_{(T,C)|\mathbf{X}}$ exist and are unique. This assumption has been shown to hold under mild conditions (when, e.g., the covariates have density with convex support; Cook (1998), Chapter 6; Li (2018), Chapter 2.2. Then, by Proposition A.1 in the Appendix, the set of active (or relevant) variables associated with the central subspace is also unique.

REMARK 3.2. The loss function in (3.2) is motivated by the quadratic discrepancy approach discussed in Cook and Ni (2005) as we can rewrite the population version of the loss function as

$$F(\Gamma, \Phi) = \bigl(\operatorname{vec}(\tilde{U}_c) - \operatorname{vec}(\Gamma\Phi)\bigr)^T W\bigl(\operatorname{vec}(\tilde{U}_c) - \operatorname{vec}(\Gamma\Phi)\bigr), \tag{3.3}$$

where $\operatorname{vec}(\cdot)$ denotes the operator that constructs a vector from a matrix by stacking its columns, $\otimes$ denotes Kronecker product, $\tilde{U}_c = \Sigma^{-1}U_c$ satisfies $\mathcal{S}_{\tilde{U}_c} \equiv \mathcal{S}_{T|\mathbf{X}}$, and $W = I_p \otimes \Sigma$ is positive-definite. See also Proposition A.2 in the Appendix for the motivation. In the population version, (A.1) does not involve Σ^{-1} after expansion, and we thus avoid inverting a large covariance matrix in high dimension. For identifiability, instead of imposing constraints on basis Γ like many SDR methods, we use an alternative constraint $\Phi\Phi^T = I_d$ to overcome computational challenges in the ultrahigh- dimensional setting. As discussed in Section 5, the new constraint will lead to iterative optimization steps that involves singular value decomposition (SVD) of relatively small matrix, which is efficient to compute.

REMARK 3.3. Here, we take $\hat{\Sigma}_n$ to be the sample covariance matrix of $\mathbf{X}$. Alternatively, we can use other estimators such as a thresholded covariance matrix (Bickel and Levina (2008)). Unless stated otherwise, we simply assume sample covariance is used, but we also use the thresholded covariance for numerical studies in Section 6.

REMARK 3.4. The derivation of the penalty weights w_j follows a variant of the adaptive group lasso (Zou (2006), Yuan and Lin (2006)): we first obtain an initial estimator $(\tilde{\Gamma}_0, \tilde{\Phi}_0)$ of (3.2) by using equal weights $w_1 = \cdots = w_p = 1$ with tuning parameter $\tilde{\lambda}$; then we set weights by $w_j = \|\tilde{\boldsymbol{\gamma}}_{0j}\|_2^{-\rho}$ to find the estimator $(\hat{\Gamma}_0, \hat{\Phi}_0)$, where $\tilde{\boldsymbol{\gamma}}_{0j}$ is the jth row of $\tilde{\Gamma}_0$ and ρ is some pre-specified constant. We can also define $w_j = +\infty$ if $\|\tilde{\gamma}_{0j}\|_2 = 0$. The estimator for $\mathcal{S}_{T|\mathbf{X}}$ is then $\mathcal{S}_{\hat{\Gamma}_0}$, and the estimated set of active variables is $\hat{\mathcal{A}}_0 = \{1 \le j \le p : \|\hat{\boldsymbol{\gamma}}_{0j}\|_2 > 0\}$, where $\hat{\Gamma}_0 = (\hat{\boldsymbol{\gamma}}_{01}, \dots, \hat{\boldsymbol{\gamma}}_{0p})^T$. The detailed computational algorithm and tuning parameter selection related to (3.2) are presented in Section 5.

3.2. *Nonparametric estimation of conditional survival function.* We now describe how we can construct a nonparametric estimator of the conditional survival function $Q(T\,|\,\mathbf{X})$

based on the central subspace obtained in Section 3.1. Estimating the conditional survival function is often of independent interest.

The estimated central subspace allows us to estimate $Q(t \mid \mathbf{x}) = P(T > t \mid \Gamma_0^T \mathbf{X} = \Gamma_0^T \mathbf{x})$ as Γ_0 is a basis matrix of $\mathcal{S}_{T|\mathbf{X}}$. The dimension of $\Gamma_0^T \mathbf{X}$ is usually low and we assume it is upper bounded. In addition, as shown in Proposition A.3 in the Appendix, conditioning on $\Gamma_0^T \mathbf{X}$, the conditional independence assumption remains to hold. We generalize the local Kaplan–Meier (KM) estimator (e.g., Beran (1981), Gonzalez-Manteiga and Cadarso-Suarez (1994), Zeng (2004)) to estimate $Q(t \mid \mathbf{x})$ by

$$\hat{Q}(t \mid \mathbf{x}) = \prod_{i=1}^{n} \left\{ 1 - \frac{I(Y_i \le t, \delta_i = 1) B_{ni}(\hat{\Gamma}_0^T \mathbf{x})}{\sum_{k=1}^{n} I(Y_k \ge Y_i) B_{nk}(\hat{\Gamma}_0^T \mathbf{x})} \right\}, \tag{3.4}$$

where $B_{ni}(\hat{\Gamma}_0^T \mathbf{x})$ are nonparametric local weights, and $\hat{\Gamma}_0$ is the estimator for Γ_0 obtained in Section 3.1 with transformation to be a semiorthogonal matrix (e.g., we can simply perform a SVD on $\hat{\Gamma}_0 \hat{\Phi}_0$ to obtain this matrix; this step is only for convenience of technical analysis and, as shown in Proposition A.4 in the Appendix, ensures the existence of a basis $\Gamma_0 \in \mathcal{S}_{T|\mathbf{X}}$ that is close enough to the transformed estimator). Two popular choices of weights are the histogram weight and the Nadaraya–Watson weight. To use the histogram weight, we partition the support $\mathcal{Z} \subset \mathbb{R}^d$ of $\hat{\Gamma}_0^T \mathbf{X}$ into small subdomains; given $\mathbf{z} \in \mathcal{Z}$, let $\mathcal{M}(\mathbf{z})$ be the subdomain containing $\mathbf{z}$ and let $N_n(\mathbf{z})$ be the number of observations $\mathbf{Z}_i = \hat{\Gamma}_0^T \mathbf{X}_i$ $(1 \le i \le n)$ that are contained in $\mathcal{M}(\mathbf{z})$. Then $B_{ni}(\mathbf{z}) = 1/N_n(\mathbf{z})$ if $\mathcal{M}(\mathbf{z}) = \mathcal{M}(\mathbf{Z}_i)$, and $B_{ni}(\mathbf{z}) = 0$ otherwise. To use the Nadaraya–Watson weight, we set $B_{ni}(\mathbf{z}) = K(\frac{\mathbf{Z}_i - \mathbf{z}}{h_n}) / \sum_{k=1}^{n} K(\frac{\mathbf{Z}_k - \mathbf{z}}{h_n})$, where $K(\cdot)$ is a kernel function and h_n is the bandwidth. In Section 4, we establish the uniform convergence for the local KM estimator in (3.4). The proposed estimator much broadens the practical use of Kaplan–Meier estimator in high dimension.

As one interesting and useful application of the estimated survival function $\hat{Q}(t \mid \mathbf{X})$, we can further estimate the conditional quantile function for $T \mid \mathbf{X}$. That is, given $0 < \tau < 1$, we directly estimate the τth conditional quantile by inverting $\hat{Q}(t \mid \mathbf{X})$. In Section 4, we show that this produces a consistent estimator for $Q_T(\tau \mid \mathbf{X}) = \sup\{t : Q(t \mid \mathbf{X}) < 1 - \tau\}$. See also the numerical illustration in Section 6.

3.3. *Double-slicing assisted preliminary estimator of $S(t \mid \mathbf{X})$ in high dimension.* In Section 3.1, we discussed the need to flexibly estimate $S(t \mid \mathbf{X})$ for inverse probability weighting in (3.1). In this subsection, we provide the details on how such a preliminary estimator can be obtained in high dimension by extending the seminal work of Li, Wang and Chen (1999), which was developed for the small p large n setting. The basic idea is to consider a slightly larger subspace based on the sufficient reduction for $(T, C)|\mathbf{X}$, which also provides a sufficient reduction for $C \mid \mathbf{X}$ and can facilitate the estimation of $S(t \mid \mathbf{X})$. Therefore, we first reduce the dimension of $\mathbf{X}$ through the augmented central subspace $\mathcal{S}_{(T,C)|\mathbf{X}}$ and then estimate $S(t \mid \mathbf{X})$ nonparametrically based on the dimension reduced predictors.

Specifically, because (Y, δ) is a function of (T, C), we have that $\mathcal{S}_{(Y,\delta)|\mathbf{X}} \subseteq \mathcal{S}_{(T,C)|\mathbf{X}}$ (e.g., Theorem 2.3, Li (2018)). In addition, Proposition A.5 in the Appendix shows that the coverage condition $\mathcal{S}_{(Y,\delta)|\mathbf{X}} \supseteq \mathcal{S}_{(T,C)|\mathbf{X}}$ is equivalent to $\mathcal{S}_{T|\mathbf{X}} \subseteq \mathcal{S}_{(Y,\delta)|\mathbf{X}}$ and $\mathcal{S}_{C|\mathbf{X}} \subseteq \mathcal{S}_{(Y,\delta)|\mathbf{X}}$; the latter conditions are often reasonable considering that $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{(Y,\delta)|\mathbf{X}}$ and $P(Y > t \mid \mathbf{X}) = P(T > t \mid \Gamma_0 \mathbf{X}) P(C > t \mid \Gamma_c \mathbf{X})$ from the conditional independence, where Γ_0 and Γ_c are basis of $\mathcal{S}_{T|\mathbf{X}}$ and $\mathcal{S}_{C|\mathbf{X}}$, respectively. Then assuming a full coverage, $\mathcal{S}_{(T,C)|\mathbf{X}}$ can be estimated through $\mathcal{S}_{(Y,\delta)|\mathbf{X}}$ by the sliced inverse regression method with double slicing on Y and δ. This idea eases the central subspace estimation as both Y and δ are observable.

Let $\mathcal{H}_{1,l}$, $l = 1, \dots, b_1$, be the b_1 nonoverlapping intervals of the discretized event time Y without censoring $(\delta = 1)$, and let $\mathcal{H}_{0,l}$, $l = 1, \dots, b_0$, be b_0 nonoverlapping intervals of the discretized Y with censoring ($\delta = 0$). With the double-slicing (DS) practice, similar to our discussion for (3.1), it can be shown that

$$\mathbf{m}_{k,l} := \mathrm{E}(\mathbf{X} \mid Y \in \mathcal{H}_{k,l}, \delta = k) - \boldsymbol{\mu} \in \Sigma\mathcal{S}_{(T,C)|\mathbf{X}}$$

for any $k = 0, 1$, and $l = 1, \ldots, b_k$, under the linearity condition. Define $M_1 = (\mathbf{m}_{1,1}, \ldots, \mathbf{m}_{1,b_1}, \mathbf{m}_{0,1}, \ldots, \mathbf{m}_{0,b_0}) \in \mathbb{R}^{p\times(b_1+b_0)}$, $p_{k,l} = P(Y \in \mathcal{H}_{k,l}, \delta = k)$, and $\mathbf{g}_1 = (\sqrt{p_{1,1}}, \ldots, \sqrt{p_{1,b_1}}, \sqrt{p_{0,1}}, \ldots, \sqrt{p_{0,b_0}})$. By setting $U_1 = M_1 D_{\mathbf{g}_1}$, we have $\mathcal{S}_{U_1} = \mathcal{S}_{M_1} = \Sigma\mathcal{S}_{(T,C)|\mathbf{X}}$. Then U_1 is a kernel matrix for $\mathcal{S}_{(T,C)|\mathbf{X}}$. Let $d_1 = \dim(\mathcal{S}_{(T,C)|\mathbf{X}})$. By replacing U_c with U_1 in (A.1), we have the objective function

$$\begin{aligned} F_1(\Gamma, \Phi) &= \mathrm{tr}\{(U_1 - \Sigma\Gamma\Phi)^T \Sigma^{-1}(U_1 - \Sigma\Gamma\Phi)\} \\ &\text{subject to } \Phi\Phi^T = I_{d_1}, \end{aligned} \tag{3.5}$$

where $\Gamma \in \mathbb{R}^{p\times d_1}$ and $\Phi \in \mathbb{R}^{d\times(b_1+b_0)}$ are parameters, and its minimizer $\Gamma_1 \in \mathbb{R}^{p\times d_1}$ of Γ forms a basis of $\mathcal{S}_{(T,C)|\mathbf{X}}$. Likewise, with $\tilde{U}_1 = \Sigma^{-1}U_1$, this objective function also forms a quadratic discrepancy function resembling (3.3).

Define $J_{k,l} := \{1 \le i \le n : Y_i \in \mathcal{H}_{k,l}$ and $\delta = k\}$ to be the index set corresponding to slice $\mathcal{H}_{k,l}$, and $N_{k,l} = |J_{k,l}|$. Let $\hat{\mathbf{m}}_{k,l} = \bar{\mathbf{X}}_{k,l} - \bar{\mathbf{X}} := \sum_{i\in J_{k,l}} \mathbf{X}_i/N_{k,l} - \bar{\mathbf{X}}$ and $\hat{p}_{k,l} = N_{k,l}/n$, which are used to construct the sample estimators $\hat{M}_1$ and $\hat{\mathbf{g}}_1$ for M_1 and $\mathbf{g}_1$. Set $\hat{U}_1 = \hat{M}_1 D_{\hat{\mathbf{g}}_1}$. Under sparsity, let $\mathcal{A}_1$ be the index set corresponding to active variables of $\mathcal{S}_{(T,C)|\mathbf{X}}$. Assume $q_1 := |\mathcal{A}_1| < p$. We can then use $\hat{U}_1$ to formulate a penalized sample objective function similar to (3.2) to estimate $\mathcal{S}_{(T,C)|\mathbf{X}}$ and $\mathcal{A}_1$ by minimizing

$$F_{1n}(\Gamma, \Phi) = \mathrm{tr}\{(\hat{U}_1 - \hat{\Sigma}_n\Gamma\Phi)^T \hat{\Omega}_n(\hat{U}_1 - \hat{\Sigma}_n\Gamma\Phi)\} + \lambda_1 \sum_{j=1}^{p} w_j \|\boldsymbol{\gamma}_j\|_2, \tag{3.6}$$

subject to $\Phi\Phi^T = I_{d_1}$, where λ_1 is tuning parameter with unequal penalty weights. Here, we perform the same adaptive estimation procedure as described for (3.2), and set $\tilde{\lambda}_1$ to be the tuning parameter corresponding to the initial equal penalty weights. With the estimator $(\hat{\Gamma}_1, \hat{\Phi}_1)$ from (3.6), this DS procedure generates the estimated central subspace $\mathcal{S}_{\hat{\Gamma}_1}$ and variable selection set $\hat{\mathcal{A}}_1 = \{1 \le j \le p : \|\hat{\boldsymbol{\gamma}}_{1j}\|_2 > 0\}$, where $\hat{\Gamma}_1 = (\hat{\boldsymbol{\gamma}}_{11}, \ldots, \hat{\boldsymbol{\gamma}}_{1p})^T$.

Since $\Gamma_1^T\mathbf{X}$ is a sufficient reduction for $C \mid \mathbf{X}$, we have $S(t \mid \mathbf{x}) = P(C > t \mid \Gamma_1^T\mathbf{X} = \Gamma_1^T\mathbf{x})$. Similar to (3.4), we assume d_1 is upper bounded. Thus, with the help of low-dimensional estimator $\hat{\Gamma}_1$, the local KM estimation is applied to estimate $S(t \mid \mathbf{x})$ as

$$\hat{S}(t \mid \mathbf{x}) = \prod_{i=1}^{n} \left\{1 - \frac{I(Y_i \le t, \delta_i = 0)B_{ni}(\hat{\Gamma}_1^T\mathbf{x})}{\sum_{k=1}^{n} I(Y_k \ge Y_i)B_{nk}(\hat{\Gamma}_1^T\mathbf{x})}\right\}, \tag{3.7}$$

where $B_{ni}(\cdot)$'s are the weight functions as described for (3.4).

REMARK 3.5. Since the sufficient reduction for $(T, C) \mid \mathbf{X}$ is also a sufficient reduction for $T \mid \mathbf{X}$, the proposed DS method not only achieves dimension reduction to facilitate the estimation of $S(t \mid \mathbf{X})$, but also provides reduction for the targeted $T \mid \mathbf{X}$. However, as $\mathcal{S}_{(T,C)|\mathbf{X}}$ is usually a larger space than $\mathcal{S}_{T|\mathbf{X}}$, it might still contain redundant information for the ultimate reduction of $T \mid \mathbf{X}$. Nevertheless, the DS method can be efficiently adapted to our high-dimensional framework, and is worthy as an initial assisting step for reduction. Since $\mathcal{S}_{T|\mathbf{X}} \subseteq \mathcal{S}_{(T,C)|\mathbf{X}}$, we have $\mathcal{A}_0 \subseteq \mathcal{A}_1$ and $q_1 \ge q$. When $\mathcal{S}_{(T,C)|\mathbf{X}} = \mathcal{S}_{T|\mathbf{X}}$, $\mathcal{A}_1 = \mathcal{A}_0$, and the DS method itself achieves simultaneous SDR and variable selection for $\mathcal{S}_{T|\mathbf{X}}$. In Section 6, we will evaluate and compare the DS estimator with the DASH estimator under different simulation scenarios.

4. Theoretical results.

In this section, we state the main theoretical results of the proposed methods in ultrahigh dimension that allows p to grow exponentially with n. Under this setting, we first establish consistency properties for the DS method targeting $\mathcal{S}_{(T,C)|\mathbf{X}}$. With these results as preparation, we then investigate consistency properties for DASH targeting $\mathcal{S}_{T|\mathbf{X}}$ and the subsequently estimated conditional survival function $\hat{Q}(t\,|\,\mathbf{x})$ with local KM estimator.

4.1. *DS estimator.* Let $\mu_j = \mathrm{E}(X_j)$, $\mu_{yj} = \mathrm{E}(X_j\,|\,Y \in \mathcal{H}_y)$ and $\mu_{k,l,j} = \mathrm{E}(X_j\,|\,Y \in \mathcal{H}_{k,l}, \delta = k)$ for $1 \le j \le p$, $1 \le y \le b$, $k = 0, 1$ and $1 \le l \le b_k$. Define $\bar{b} = \max(b_1, b_0)$ and $\sigma_{ij} = (\Sigma)_{ij}$. Let $\lambda_{\min}(\cdot)$ denote the minimum eigenvalues of a square matrix. Let $\mathbf{e}_j$ be the vector with the ith element to be one and the rest to be zeros. Consider the following regularity conditions:

(A1) Let G be a basis matrix of $\mathcal{S} = \mathcal{S}_{(T,C)|\mathbf{X}}$ or $\mathcal{S}_{T|\mathbf{X}}$. Assume $\mathbf{X}$ satisfies $\mathrm{E}(\mathbf{X} - \boldsymbol{\mu}\,|\,G^T\mathbf{X}) = AG^T(\mathbf{X} - \boldsymbol{\mu})$ for some matrix $A \in \mathbb{R}^{p\times \dim(\mathcal{S})}$.

(A2) There is a positive constant c_H such that $P(Y \in \mathcal{H}_{k,l}, \delta = k) \ge c_H \bar{b}^{-1}$ for all $k = 0, 1$ and $1 \le l \le b_k$.

(A3) For all $\epsilon > 0$, there is a constant $c_1 > 0$ such that for all $1 \le j \le p$, $P(|X_j - \mu_j| > \epsilon) \le 2\exp(-c_1\epsilon^2)$. In addition, for all $\epsilon > 0$, $1 \le j \le p$, $1 \le y \le b$, $k = 0, 1$ and $1 \le l \le b_k$, there are constants $c_2, \tilde{c}_2 > 0$ such that $P(|X_j - \mu_{yj}| > \epsilon\,|\,Y \in \mathcal{H}_y) \le \tilde{c}_2\exp(-c_2\epsilon^2)$ and $P(|X_j - \mu_{k,l,j}| > \epsilon\,|\,Y \in \mathcal{H}_{k,l}, \delta = k) \le \tilde{c}_2\exp(-c_2\epsilon^2)$.

(A4) Assume that $\sigma_{ij} < \tilde{\sigma}$ $(1 \le i, j \le p)$ and $\lambda_{\min}(\Sigma) > \sigma_*$, where σ_{ij} is the (i, j)-element of Σ, and σ_* and $\tilde{\sigma}$ are some positive constants.

(A5) Assume the nonzero singular values of $\tilde{U}_1$ are bounded away from 0.

(A6) Assume $\min_{j\in\mathcal{A}_1} \mathbf{e}_j^T \tilde{U}_1 \tilde{U}_1^T \mathbf{e}_j > c_{u1} n^{-\tau_1}$ for some $0 \le \tau_1 < 1$.

Condition (A1) is the linearity condition discussed in Section 3. Condition (A2) assumes that all slices of the response have reasonably large marginal probability, which is mild and is often satisfied in practice. The first component in (A3) is the commonly used uniformly sub-Gaussian assumption for the marginal distribution of predictors X_j $(1 \le j \le p)$. The second component in (A3) further assumes uniformly sub-Gaussian condition for X_j given each slice; if b and $\bar{b}$ are upper bounded, these two components are essentially equivalent. Condition (A4) assumes that all elements in Σ are uniformly upper bounded with minimum eigenvalue bounded from zero. Conditions (A5) and (A6) includes mild assumptions on the "kernel" matrix $\tilde{U}_1$: (A5) holds if the nonzero eigenvalues of $\mathrm{Var}(\mathrm{E}(\mathbf{X}\,|\,\tilde{Y}))$ are bounded away from zero, where $\tilde{Y}$ is the double-sliced response; (A6) uses a marginal utility quantity condition (Zhu et al. (2011)) to control signal levels. To obtain consistency results, define $\mathrm{P}_{\mathcal{S}}$ to be the projection matrix onto a given subspace $\mathcal{S}$, and let $\|\cdot\|_F$ be the Frobenius norm. Define $\bar{p}_n = \max(p, n)$. The following condition allows p to grow exponentially with n:

(A7) Assume $\bar{b}^2 q_1 \log \bar{p} = O(n^{1-\zeta_1})$ and $q_1^2 \log \bar{p}_n = O(n^{1-\zeta_1})$ for some constant ζ_1 with $\tau_1 < \zeta_1 < 1$.

THEOREM 4.1. *Assume Conditions* (A1)–(A7) *hold. Suppose that* $\tilde{\lambda}_1 = 2c_{1n}\sqrt{\log \bar{p}_n / n}$, $\lambda_1 = 2^{1-\rho} c_{u1}^{\rho/2} c_{1n} \sqrt{\log \bar{p}_n / n^{1+\rho\tau_1}}$ *and* $\rho(\zeta_1 - \tau_1) \ge 1 - \zeta_1$, *where* $c_{1n} = (C_1\bar{b} + C_2 q_1^{1/2})/2$ *with* C_1, C_2 *being some generic positive constants* (*given in Theorem* I.2 *of Supplement* I.4). *Then the DS estimator of Section* 3.3 *satisfies*

$$\|\mathrm{P}_{\mathcal{S}_{\hat{\Gamma}_1}} - \mathrm{P}_{\mathcal{S}_{(T,C)|\mathbf{X}}}\|_F = O_p\big((\bar{b} + q_1^{1/2})\sqrt{q_1 \log \bar{p}_n / n}\big), \tag{4.1}$$

$$P(\hat{\mathcal{A}}_1 = \mathcal{A}_1) \to 1 \quad \textit{as } n \to \infty. \tag{4.2}$$

Theorem 4.1 shows that the estimation consistency and the variable selection consistency of $\mathcal{S}_{(T,C)|\mathbf{X}}$ can be simultaneously established by our DS method without imposing stringent conditions. In addition to p, the theorem allows q_1 and $\bar{b}$ to diverge with n. If $\bar{b}$ is upper bounded, the convergence rate of the central subspace estimation is $O_p(q_1\sqrt{\log\bar{p}_n/n})$; if q_1 is upper bounded, it is $O_p(\bar{b}\sqrt{\log\bar{p}_n/n})$; with both q_1 and $\bar{b}$ upper bounded, the result is simplified to $O_p(\sqrt{\log\bar{p}_n/n})$.

4.2. *DASH estimator.* The results in Theorem 4.1 target on $\mathcal{S}_{(T,C)|\mathbf{X}}$, which can be a larger subspace than our main interest $\mathcal{S}_{T|\mathbf{X}}$. We next establish theoretical properties for the DASH method targeting $\mathcal{S}_{T|\mathbf{X}}$. Unless stated otherwise, assume that the Nadaraya–Watson kernel-based KM estimator is applied to obtain $\hat{S}(t\,|\,\mathbf{X})$. For technical brevity, we assume uniform kernel function $K(\mathbf{x}) = I(\|\mathbf{x}\|_2 \le 1)$ is used, although other popular kernel choices can apply. Assume data splitting (Cox (1975)) is applied so that $\hat{\Gamma}_1$ and the local KM estimation use different data halves, and $\hat{S}(t\,|\,\mathbf{X}_i)$ is obtained via the leave-one-out technique (Härdle, Hall and Ichimura (1993)), where the product in (3.7) omits the ith sample point. Suppose the domain $\mathcal{X}$ of $\mathbf{X}$ is compact with $\|\mathbf{X}\|_2 \le K$ for some constant $K > 0$. Consider the following regularity conditions:

(C1) Suppose $P(T \ge C \ge t\,|\,\mathbf{X}) \ge \tau_0$ for any $t \in [0, T_0]$ and $\mathbf{X} \in \mathcal{X}$, where τ_0 is some positive constant and T_0 is the maximum follow-up time.

(C2) Given a basis $\Gamma_1 \in \mathcal{S}_{(T,C)|\mathbf{X}}$, $m(t\,|\,\mathbf{z}) := P(T > t\,|\,\Gamma_1^T\mathbf{X} = \mathbf{z})$ and $\tilde{m}(t\,|\,\mathbf{z}) := P(C > t\,|\,\Gamma_1^T\mathbf{X} = \mathbf{z})$, assume that $|m(t\,|\,\mathbf{z}_1) - m(t\,|\,\mathbf{z}_2)| \le c_l\|\mathbf{z}_1 - \mathbf{z}_2\|_2$ and $|\tilde{m}(t\,|\,\mathbf{z}_1) - \tilde{m}(t\,|\,\mathbf{z}_2)| \le \tilde{c}_l\|\mathbf{z}_1 - \mathbf{z}_2\|_2$ hold for any $t \in [0, T_0]$, where c_l and $\tilde{c}_l$ are some positive constants.

(C3) Given a basis $\bar{\Gamma}$ of $\mathcal{S}_{(T,C)|\mathbf{X}}$ or $\mathcal{S}_{T|\mathbf{X}}$, the density of $\bar{\Gamma}^T\mathbf{X}$ is positive and bounded from zero on $\bar{\mathcal{Z}}$, where $\bar{\mathcal{Z}} = \{\bar{\Gamma}^T\mathbf{X} : \mathbf{X} \in \mathcal{X}\}$. In addition, the density of $\Gamma^T\mathbf{X}$ satisfies Lipschitz condition for Γ on a neighborhood of $\bar{\Gamma}$ in $\|\cdot\|_F$-norm.

(C4) For every $1 \le y \le b$, there exists a constant $c_s > 0$ such that $P(Y \in \mathcal{H}_y) \ge c_s b^{-1}$.

(C5) Assume $\min_{j\in\mathcal{A}_0} \mathbf{e}_j^T\tilde{U}_c\tilde{U}_c^T\mathbf{e}_j > c_{u2}n^{-\tau_2}$ for some $0 \le \tau_2 < 1$. The nonzero singular values of $\tilde{U}_c$ are bounded away from zero.

(C6) Assume $bq(\frac{\log\bar{p}_n}{n})^{\frac{2}{2+d_1}} = O(n^{-\zeta_2})$ and $(\bar{b}^2q_1 + q_1^2 + b)bq\log\bar{p}_n = O(n^{1-\zeta_2})$ for some constant ζ_2 with $\tau_2 < \zeta_2 < 1$.

Condition (C1) is commonly assumed in censored data analysis. It implies that for every $t \in [0, T_0)$, $P(C > t\,|\,\mathbf{X}) \ge \tau_0$ and $P(Y > t\,|\,\mathbf{X}) \ge \tau_0^2$. Condition (C2) indicates that conditional survival functions satisfy Lipschitz conditions on $\Gamma_1^T\mathbf{X}$. Condition (C3) is similar to Assumption C1 in Wang et al. (2010) and is used to bound the density of $\Gamma^T\mathbf{X}$ away from zero. Condition (C4) is similar to (A2) and is used to allow reasonable sample size on each slice. Condition (C5) is similar to (A5) and (A6) to regulate the "kernel" matrix. We also allow p to grow exponentially with n in Condition (C6). Theorem 4.2 establishes estimation and selection consistency for DASH in the ultrahigh-dimensional setting.

THEOREM 4.2. *Assume the conditions in Theorem* 4.1 *and* (C1)–(C6) *hold and take* $h_n \asymp (\frac{\log\bar{p}_n}{n})^{\frac{1}{2+d_1}}$. *Define* $\xi_n = b^{1/2}(\log\bar{p}_n/n)^{\frac{1}{2+d_1}} + (\bar{b}q_1^{1/2} + q_1 + b^{1/2})(b\log\bar{p}_n/n)^{1/2}$. *Suppose that* $\tilde{\lambda} = 2c_0\xi_n$, $\lambda = 2^{1-\rho}c_0c_{u2}^{\rho/2}n^{-\rho\tau_2/2}\xi_n$ *and* $\rho(\zeta_2 - \tau_2) \ge 1 - \zeta_2$, *where* c_0 *is some generic positive constant* (*given in Theorem* I.1 *of Appendix* I.3). *Then the DASH estimator satisfies*

$$\begin{aligned}&\|\mathrm{P}_{\mathcal{S}_{\hat{\Gamma}_0}} - \mathrm{P}_{\mathcal{S}_{T|\mathbf{X}}}\|_F\\ &\quad= O_p\biggl((bq)^{1/2}\biggl(\frac{\log\bar{p}_n}{n}\biggr)^{\frac{1}{2+d_1}} + (\bar{b}q_1^{1/2} + q_1 + b^{1/2})\biggl(\frac{bq\log\bar{p}_n}{n}\biggr)^{1/2}\biggr),\end{aligned} \tag{4.3}$$

$$P(\hat{\mathcal{A}}_0 = \mathcal{A}_0) \to 1 \quad \textit{as } n \to \infty. \tag{4.4}$$

Compared to the consistency result (4.1) of the initial DS estimator, the DASH method achieves estimation consistency for $\mathcal{S}_{T|\mathbf{X}}$ with simultaneous variable selection consistency. From (4.3), when b, $\bar{b}$ and q_1 are all upper bounded, the DASH estimator converges at the rate of $O_p((\log \bar{p}_n/n)^{\frac{1}{2+d_1}})$. The somewhat slower convergence rate relative to (4.1) is mainly due to the key inverse probability weighting procedure, but this enables us to target the true central subspace $\mathcal{S}_{T|\mathbf{X}}$, instead of the larger $\mathcal{S}_{(T,C)|\mathbf{X}}$. Interestingly, under classical fixed p scenarios, if $d \ge d_1/2$ (i.e., $\dim(\mathcal{S}_{T|\mathbf{X}})$ is allowed to be as small as half of $\dim(\mathcal{S}_{(T,C)|\mathbf{X}})$), the rate for DASH matches (up to a logarithmic factor) or exceeds the known convergence rate $O_p(n^{-\frac{1}{2(1+d)}})$ in Nadkarni, Zhao and Kosorok (2011).

REMARK 4.1. In the discussion above, we assume that $\mathcal{S}_{(T,C)|\mathbf{X}}$ is truly sparse in the sense that the number of active variables (i.e., $|\mathcal{A}_1|$) is much smaller than p. Interestingly, our results can be extended to a weaker condition to allow an "approximately" sparse scenario where a large number of variables are active for $\mathcal{S}_{(T,C)|\mathbf{X}}$. Specifically, recall that Γ_1 denotes a basis of $\mathcal{S}_{(T,C)|\mathbf{X}}$ that gives a minimizer of (3.5). Given any index set $\mathcal{A} \subset \{1, \dots, p\}$, let $\Gamma_{1,\mathcal{A}}$ be the $|\mathcal{A}| \times d_1$ matrix consisting of the rows of Γ_1 corresponding to $\mathcal{A}$. Then rather than requiring that only a small subset of variables are active, we assume that there exists an index set $\check{\mathcal{A}}_1 \subset \{1, \dots, p\}$ with $q_1 = |\check{\mathcal{A}}_1| < p$ and a (small) parameter $\theta > 0$ such that

$$\|\Gamma_{1,\check{\mathcal{A}}_1^c}\|_{2,1} \le \theta, \tag{4.5}$$

where $\|\cdot\|_{2,1}$ is the $l_{2,1}$ norm (i.e., $\|\Gamma_{1,\check{\mathcal{A}}_1^c}\|_{2,1} = \sum_{j \in \check{\mathcal{A}}_1^c} \|\boldsymbol{\gamma}_{1j}\|_2$ and γ_{1j} is the jth row of Γ_1). By Proposition A.6 in the Appendix, given the objective (3.5), the assumption of (4.5) is well defined as $\|\Gamma_{1,\mathcal{A}}\|_{2,1}$ is unique.

Under the weaker condition of (4.5), on the one hand, $\mathcal{S}_{(T,C)|\mathbf{X}}$ is not strictly sparse according to the definition of active variables in Section 3; on the other hand, as the magnitude of Γ_1 in $\check{\mathcal{A}}_1^c$ is relatively small, it is still possible to approximate the (nonsparse) $\mathcal{S}_{(T,C)|\mathbf{X}}$ by a sparse model. Therefore, the extended setting here, to some extent, shares the flavor of both the abundant variable settings in SDR (e.g., Cook, Forzani and Rothman (2012)) and the approximately sparse conditions in linear models (e.g., Zhang and Huang (2008)). With the weaker form of sparsity, we have the following theorem.

THEOREM 4.3. *Assume Conditions* (A1)–(A5) *and* (A7) *hold. Suppose that* $\theta = o(1)$ *and* $\lambda_1 = 3c_{1n}\sqrt{\log \bar{p}_n/n}$ *with equal weights* $w_j = 1$ *for* $1 \le j \le p$, *where* c_{1n} *is defined as in Theorem* 4.1. *Then the DS estimator satisfies*

$$\|\mathrm{P}_{\mathcal{S}_{\hat{\Gamma}_1}} - \mathrm{P}_{\mathcal{S}_{(T,C)|\mathbf{X}}}\|_F = O_p\bigl(\max\bigl\{(\bar{b} + q_1^{1/2})\sqrt{q_1 \log \bar{p}_n/n}, \theta\bigr\}\bigr). \tag{4.6}$$

In addition, suppose that $\theta = O(\bar{b}\sqrt{q_1 \log \bar{p}_n/n})$ *and* $\theta = O(q_1\sqrt{\log \bar{p}_n/n}))$ *and further assume that the additional conditions of Theorem* 4.2 *hold. Then the DASH estimator satisfies the consistency properties of* (4.3) *and* (4.4).

In particular, if $\theta = 0$, (4.5) reduces back to the strict sparsity conditions for $\mathcal{S}_{(T,C)|\mathbf{X}}$ in Section 3.3, and the convergence result of (4.6) becomes the same as (4.1) in Theorem 4.1. If θ is positive but converges to 0, by approximating $\mathcal{S}_{(T,C)|\mathbf{X}}$ with a parsimonious model from the DS method, we also note that the final DASH estimator may still be consistent for estimation of $\mathcal{S}_{T|\mathbf{X}}$, at some possible cost of the convergence rate.

4.3. *Local KM estimator.* As discussed in Section 3.2, with $\hat{\Gamma}_0$ available, we can subsequently apply popular nonparametric estimation methods to estimate $Q(t \,|\, \mathbf{x})$. In the following, suppose the local KM estimator (3.4) with Nadaraya–Watson kernel weighting scheme is applied. Similar to $\hat{S}(t \,|\, \mathbf{x})$, we assume same techniques with uniform kernel are used, but we can generalize the results to other kernel function choices. The following theorem provides a uniform estimation consistency result for $Q(t \,|\, \mathbf{x})$.

THEOREM 4.4. *Suppose the conditions of Theorem 4.2 are satisfied. Then given any* $\mathbf{x}_0 \in \mathcal{X}$, *the local KM estimator satisfies*

$$
\begin{aligned}
&\sup_{t \in [0, T_0)} \big|\hat{Q}(t \,|\, \mathbf{x}_0) - Q(t \,|\, \mathbf{x}_0)\big| \\
&\quad = O_p\bigg((bq)^{1/2}\bigg(\frac{\log \bar{p}_n}{n}\bigg)^{\frac{1}{2+d_1}} + \big(\bar{b}q_1^{1/2} + q_1 + b^{1/2}\big)\bigg(\frac{bq \log \bar{p}_n}{n}\bigg)^{1/2}\bigg).
\end{aligned}
\tag{4.7}
$$

By employing the Nadaraya–Watson based local KM estimator, the consistency rate upper bound obtained in Theorem 4.4 is the same as that of (4.3). Likewise, when b, $\bar{b}$ and q_1 are all upper bounded, the local KM estimator can converge at the rate of $O_p((\log \bar{p}_n/n)^{\frac{1}{2+d_1}})$. It is worth noting that other nonparametric methods may also be applied to estimate $Q(t \,|\, \mathbf{x})$. For example, histogram-based local KM method may be used and we can provide a similar uniform estimation error bound for $Q(t \,|\, \mathbf{x})$, which is left in Supplement I.5.

COROLLARY 4.1. *Suppose the conditions of Theorem 4.2 are satisfied. Then given any* $\mathbf{x}_0 \in \mathcal{X}$ *and* $0 < \tau < 1$, *the* τ*th conditional quantile estimator* $\hat{Q}_T(\tau \,|\, \mathbf{x}_0) = \sup\{t : \hat{Q}(t \,|\, \mathbf{x}_0) < 1 - \tau\}$ *is a consistent estimator for* $Q_T(\tau \,|\, \mathbf{x}_0)$.

With the estimated survival function, we naturally obtain a consistent estimator for conditional quantiles of the survival time as described in Corollary 4.1. Beyond all these consistency results, post selection inference on SDR (with censored data) may need to directly involve the set of selected variables besides central subspace estimation error; it is an interesting yet challenging problem in its own right, which is left for future studies.

5. Computation.

As we described in Section 3, the algorithm for the DASH method involves a DS estimation step followed by an IPCW based step. We summarize the procedures for DASH in Algorithm 1.

Algorithm 1 DASH method for censored data

```
1. An initial DS estimation step
    (a) Construct the kernel matrix $\hat{U}_1$ and covariance estimation $\hat{\Sigma}_n$.
    (b) Given $d_1$, $\tilde{\lambda}_1$ and $\lambda_1$, find minimizer $(\hat{\Gamma}_1, \hat{\Phi}_1)$ of (3.6).
2. An IPCW-based estimation step
    (a) Given $h_n$, apply the local KM estimation method with the DS estimator to find $\hat{S}(t \,|\, \mathbf{X}_i)$ by (3.7).
    (b) Construct the kernel matrix $\hat{U}_c$.
    (c) Given $d$, $\tilde{\lambda}$ and $\lambda$, find minimizer $(\hat{\Gamma}_0, \hat{\Phi}_0)$ of (3.2).
3. Apply a nonparametric method discussed in Section 3.2 to estimate $Q(t \,|\, \mathbf{X})$.
```

It remains to consider two practically important issues: (1) how to find proper values for tuning parameters and structural dimensions $(d_1, \tilde{\lambda}_1, \lambda_1)$, h_n and $(d, \tilde{\lambda}, \lambda)$; (2) how to solve optimization problems for the objective functions (3.2) in Step 2(c) and (3.6) in Step 1(b). Accordingly, we propose the procedures of tuning parameters and structural dimension determination in Section 5.1. Section 5.2 explains the algorithm used to solve (3.2), which is also applicable to solving Step 1(b).

5.1. *Tuning parameter and structural dimension selection.* We propose to use cross-validation (CV) procedures to determine $(d_1, \tilde{\lambda}_1, \lambda_1)$, h_n and similarly $(d, \tilde{\lambda}, \lambda)$ in their respective steps. As discussed in Yang (2007), CV is often considered as a natural approach for complicated nonparametric procedure/model comparison purposes. As will be seen, the main challenge of applying CV is that the true survival time is not always observable. We propose a new sequential data-driven approach that integrates inverse weighting ideas for prediction error measurements (Gerds and Schumacher (2007)) into the parameter/dimension determination. Since $(d_1, \tilde{\lambda}_1, \lambda_1)$ can be chosen by procedures similar to that of $(d, \tilde{\lambda}, \lambda)$, for brevity, we next assume $(d_1, \tilde{\lambda}_1, \lambda_1)$ has been chosen, and only describe in detail how to find h_n and $(d, \tilde{\lambda}, \lambda)$ sequentially. Suppose the data is partitioned into K folds, and denote the nonoverlapping index set of each fold by $\mathcal{I}_1, \dots, \mathcal{I}_K \subset \{1, \dots, n\}$. Let $\mathbb{Z}_k$ be the validation set in the kth fold and $\mathbb{Z}_{(-k)}$ be the estimation set excluding the kth fold.

5.1.1. *Determining h_n.* We intend to find h_n used in Step 2(a). Given a specified time point τ_m (e.g., $\tau_m = \text{median}(Y_i)$), define $\delta^* = I(C > \tau_m)$. For each candidate h_n, we estimate $\mathrm{E}(\delta^* \mid \mathbf{X}) = S(\tau_m \mid \mathbf{X})$ by the local KM estimation using the estimation set $\mathbb{Z}_{(-k)}$, and denote this estimator by $\hat{S}_{(-k)}(\tau_m \mid \mathbf{X}, h_n)$. Then we define the prediction mean square error $\mathrm{MSE}(h_n) := \mathrm{E}[\delta^* - \hat{S}_{(-k)}(\tau_m \mid \mathbf{X}, h_n)]^2$. With the censored data, δ^* is not always observed. As a solution, $\mathrm{MSE}(h_n)$ can be equivalently written as $\mathrm{E}[\delta^* - \hat{S}_{(-k)}(\tau_m \mid \mathbf{X}, h_n)]^2 W$, where $W = \frac{I(Y > \tau_m)}{Q(\tau_m \mid \mathbf{X})} + \frac{I(Y \le \tau_m)(1-\delta)}{Q(Y \mid \mathbf{X})}$ is a weighting variable. Consequently, given a set $\mathcal{G}$ of candidates h_n's, we can now determine the parameter $h_n = h_{\text{opt}}$ by

$$h_{\text{opt}} := \arg\min_{h \in \mathcal{G}} \widehat{\mathrm{MSE}}(h) = \arg\min_{h \in \mathcal{G}} \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \hat{W}_i \bigl(\delta_i^* - \hat{S}_{(-k)}(\tau_m \mid \mathbf{X}_i, h)\bigr)^2, \tag{5.1}$$

where $\hat{W}_i = \frac{I(Y_i > \tau_m)}{\tilde{Q}(\tau_m \mid \mathbf{X}_i)} + \frac{I(Y_i \le \tau_m)(1-\delta_i)}{\tilde{Q}(Y_i \mid \mathbf{X})}$ and $\tilde{Q}(t \mid \mathbf{X})$ is a local KM estimation of $Q(t \mid \mathbf{X})$ using the reduction $\hat{\Gamma}_1^T \mathbf{X}$. Since our computation is all based on dimension reduced predictors, selection for h_n is computationally fast.

5.1.2. *Determining $(d, \tilde{\lambda}, \lambda)$.* In Step 2(c), to find $(d, \tilde{\lambda}, \lambda)$, we again apply specialized sequential CV and the aforementioned weighting techniques to naturally handle performance evaluation. Specifically, we perform the following steps:

(i) Given a candidate d and a sequence of candidate $\tilde{\lambda}$'s, using data $\mathbb{Z}_{(-k)}$ $(k = 1, \dots, K)$ with $w_j = 1$ $(j = 1, \dots, p)$, compute the solution path of (3.2) and denote the estimators by $(\tilde{\Gamma}_{(-k)}, \tilde{\Phi}_{(-k)})$.

(ii) With $\mathbb{Z}_{(-k)}$ and $\tilde{\Gamma}_{(-k)}$, compute the local KM estimator $\hat{Q}^{\star}_{(-k)}(t \mid \mathbf{X})$ for $Q(t \mid \mathbf{X})$ based on the reduction $\tilde{\Gamma}_{(-k)}^T \mathbf{X}$. The bandwidth for $\hat{Q}^{\star}_{(-k)}(t \mid \mathbf{X})$ is automatically determined by a CV within $\mathbb{Z}_{(-k)}$ using similar procedures as in Section 5.1.1 by switching the roles of T and C.

(iii) With each candidate $(d, \tilde{\lambda})$ and their corresponding $\hat{Q}^{\star}_{(-k)}$'s, evaluate the out-of-sample prediction performance by computing

$$\widehat{\mathrm{MSE}}_2(d, \tilde{\lambda}) := \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \hat{W}_i^{\star} \big(\delta_i^{\star} - \hat{Q}^{\star}_{(-k)}(\tau_m \mid \mathbf{X}_i)\big)^2, \tag{5.2}$$

where $\delta_i^{\star} = I(T > \tau_m)$ and $\hat{W}_i^{\star} = \frac{I(Y_i > \tau_m)}{\hat{S}(\tau_m \mid \mathbf{X}_i)} + \frac{I(Y_i \leq \tau_m)\delta_i}{\hat{S}(Y_i \mid \mathbf{X}_i)}$.

(iv) With $(d, \tilde{\lambda})$ chosen by minimal (5.2), compute (3.2) with the whole data to find $\tilde{\Gamma}_0$. Then set $w_j = \|\tilde{\boldsymbol{\gamma}}_{0j}\|_2^{-\rho}$ and repeat Steps (i)–(iv) with the previously selected d to determine λ.

The determination of $(d_1, \tilde{\lambda}_1, \lambda_1)$ is obtained similarly. But the detailed procedures are slightly different and simpler because DS does not involve the delicate survival function estimation issues. Accordingly, in Step (ii), rather than using the local KM estimator, we apply the tree classification (Breiman et al. (1984)) with pruning to predict the response slice; in Step (iii), rather than using $\widehat{\mathrm{MSE}}_2(d, \tilde{\lambda})$, we compute slice classification errors to evaluate prediction performance.

Since $(d_1, \tilde{\lambda}_1, \lambda_1)$, h_n and $(d, \tilde{\lambda}, \lambda)$ are determined sequentially instead of combinatorially and nonparametric estimation method like local KM estimation and tree classifications are all performed on dimension reduced covariates, these important tuning parameters are thus automatically determined in our algorithm in a computationally efficient way.

5.2. *Objective function optimization.* To solve (3.2) in Algorithm 1, we employ an iterative algorithm, where a Stiefel manifold optimization is directly embedded into a parallelizable coordinate descent to update Γ and Φ iteratively until convergence. This optimization algorithm is motivated by Qian, Ding and Cook (2018) for SDR optimization problems with complete data, and is computationally efficient without inverting any large covariance. For presentation brevity, we leave the detailed algorithm and rationales in Supplement II.1.

6. Simulation studies.

In this section, we evaluate the numerical performance of DASH along with the DS procedure for censored data in high dimension.

6.1. *Performance with different covariance estimation.* Unless stated otherwise, we use the following candidate parameters in our simulation and real data experiments: For $\tilde{\lambda}$, λ, $\tilde{\lambda}_1$ and λ_1, we use a sequence of 50 values between 0.01 and 1 that are evenly spaced in the logarithmic scale. Similarly, for h_n, we use a sequence of 20 values between 0.1 and 1 that are evenly spaced in the logarithmic scale. The candidates for d_1 and d are $\{1, 2, 3\}$. In addition, we simply set $b_1 = b = 5$, $b_0 = 2$, $\tau_m = \text{median}(Y_i)$ and $\rho = 0.5$. Gaussian kernel was used for the estimation of $S(t \mid \mathbf{X})$ and $Q(t \mid \mathbf{X})$.

In the following, we set $n = 200$ and $p = 1000$. The covariate vector $\mathbf{X}$ was generated from a multivariate normal distribution with mean zero and covariance matrix Σ that has an exponential decay structure, such that $[\Sigma]_{i,j} = 0.5^{|i-j|}$, $i, j = 1, \dots, p$. The survival time T was generated from the linear transformation model $T = \exp(-2.5 + \boldsymbol{\beta}^T \mathbf{X} + 0.25\varepsilon)$, where $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \dots, 0)^T$, and the error ε follows the standard extreme value distribution $\varepsilon = \log[-\log(1 - U)]$ with uniformly distributed U on $[0, 1]$. This corresponds to both an AFT and proportional hazards model. The censoring time C was generated in two scenarios: Case 1: $C = \exp(-2 + \boldsymbol{\beta}^T \mathbf{X} + 0.5\varepsilon_1)$; Case 2: $C = \exp(-1 + \boldsymbol{\beta}_1^T \mathbf{X} + 0.5\varepsilon_1)$, where $\boldsymbol{\beta}_1 = (0, \dots, 0, 1, 1, 1, 1, 1)$, and ε_1 takes the extreme value distribution and is independent of ε. The censoring rates are on average about 27% for Case 1 and 40% for Case 2. It can be seen that in Case 1, $\mathcal{S}_{T \mid \mathbf{X}} = \mathcal{S}_{(T,C) \mid \mathbf{X}} = \text{Span}(\boldsymbol{\beta})$, and thus DS can be directly applied to

TABLE 1
Comparison of different estimations based on 100 *runs*

Cases	Method	Averaged Frobenius-norm loss	Frequency (%)			$\overline{C}_v$	$\overline{IC}_v$
			$d=1$	$d=2$	$d=3$		
Case 1	Oracle	–	100	0	0	5	0
	SC-DS	0.265 (0.010)	100	0	0	5.00	1.00
	SC-DASH	0.323 (0.012)	100	0	0	4.99	0.31
	TC-DS	0.191 (0.010)	100	0	0	5.00	0.42
	TC-DASH	0.220 (0.011)	100	0	0	5.00	0.33
Case 2	Oracle	–	100	0	0	5	0
	SC-DS	0.735 (0.051)	5	94	1	4.99	13.97
	SC-DASH	0.384 (0.035)	85	15	0	4.94	1.80
	TC-DS	0.672 (0.050)	13	87	0	5.00	7.68
	TC-DASH	0.278 (0.030)	86	14	0	4.99	1.10

estimate the target $\mathcal{S}_{T|\mathbf{X}}$, while in Case 2, $\mathcal{S}_{T|\mathbf{X}} = \operatorname{Span}(\boldsymbol{\beta})$ is a proper subspace of $\mathcal{S}_{(T,C)|\mathbf{X}} = \operatorname{Span}(\boldsymbol{\beta}, \boldsymbol{\beta}_1)$ and DASH is required to estimate the true central subspace.

We examined both DS and DASH in the high-dimensional setting with covariance matrix Σ estimated by sample covariance (SC) and thresholded covariance (TC), respectively. We denote them by SC-DS, SC-DASH, TC-DS and TC-DASH. The thresholded covariance is obtained by a univariate lasso-type thresholding rule in Rothman, Levina and Zhu (2009) designed for (approximately) sparse covariance matrix estimation. We recorded the estimated structural dimension d, and used Frobenius norm loss of projection matrix $\|\mathrm{P}_{\hat{\mathcal{S}}_{T|\mathbf{X}}} - \mathrm{P}_{\mathcal{S}_{T|\mathbf{X}}}\|_F$ to evaluate the estimation accuracy of the central subspace. To quantify variable selection performance, we used C_v to denote the number of correctly identified active variables, and IC_v to denote the number of incorrectly identified active variables. The procedure above was repeated 100 times for each model and the results are summarized in Table 1.

The empirical results showed that when $\mathcal{S}_{T|\mathbf{X}} = \mathcal{S}_{(T,C)|\mathbf{X}}$ (Case 1), both methods successfully identified relevant variables with relatively low IC_v rates, and correctly selected the structural dimension. As expected from the equivalence of $\mathcal{S}_{T|\mathbf{X}}$ and $\mathcal{S}_{(T,C)|\mathbf{X}}$ in this example, DS resulted in central subspace estimation similar to or slightly better than that of DASH. However, in Case 2 when $\mathcal{S}_{T|\mathbf{X}} \subset \mathcal{S}_{(T,C)|\mathbf{X}}$, DS tends to estimate the larger space $\mathcal{S}_{(T,C)|\mathbf{X}}$ and thus predominantly suggested larger structural dimensions than the truth of targeted $\mathcal{S}_{T|\mathbf{X}}$. Therefore, DS resulted in much higher estimation error in central subspace estimation and selected larger number of irrelevant variables than that of DASH. In practice, since the relationship between $\mathcal{S}_{T|\mathbf{X}}$ and $\mathcal{S}_{(T,C)|\mathbf{X}}$ is unknown, we suggest using the DASH method rather than simply stopping at double slicing: if the selected dimension of DASH is the same as that of DS, one may adopt results of DS for central subspace estimation and variable selection; on the other hand, if the selected dimension of DASH is smaller, we adopt final results from DASH. The thresholding covariance methods TC-DS and TC-DASH gave better estimation results than their sample covariance counterparts, suggesting potential gain by imposing the extra covariance thresholding step if covariance matrix is sparse.

6.2. *Conditional quantile estimation.* Since the proposed methods have the potential to facilitate nonparametric estimation of conditional survival functions and consequently conditional quantile functions, we further evaluated the numerical performance in estimating conditional quantile functions under nonlinear and heteroscedastic scenarios. Due to space constraint, we leave detailed numerical results in Supplement II.2.1.

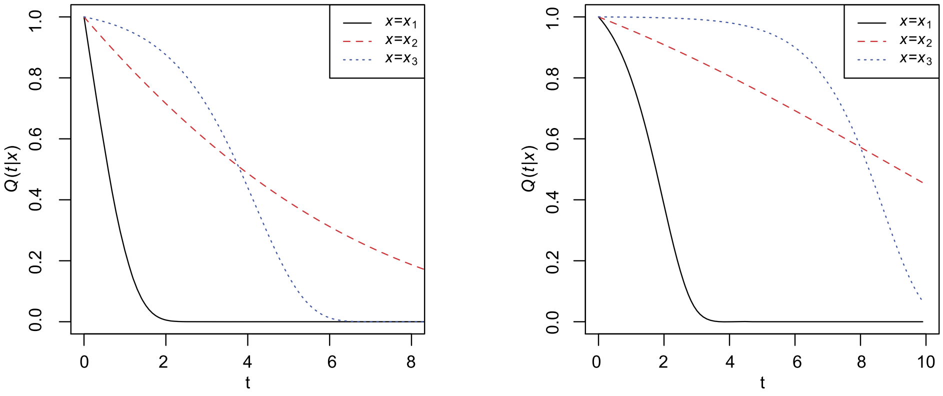

FIG. 1. *True survival functions $Q(t|\mathbf{x})$ for some different covariate values $\mathbf{x}$. Left panel: Case A; Right panel: Case B.*

6.3. *Simulation examples with possible model misspecification.* It is well known that the Cox proportional hazard model requires the relative effect of covariates to be unchanged over time. Next, we provide example case studies where this assumption does not hold. Specifically, with $p = 1000$, consider the following two hazard functions for T: (Case A) $h(t \mid \mathbf{X}) = \exp(-2\boldsymbol{\beta}^T\mathbf{X} + |\boldsymbol{\beta}^T\mathbf{X}|t)$ and (Case B) $h(t \mid \mathbf{X}) = \exp(-(2 + \boldsymbol{\beta}_2^T\mathbf{X})^2 + (\boldsymbol{\beta}_2^T\mathbf{X})^2 t)$, where $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \dots, 0)^T$, $\boldsymbol{\beta}_2 = (1, 1, 1, 0, \dots, 0, 1)^T$ and $\mathbf{X} = (X_1, \dots, X_p)^T$ consists of i.i.d. standard normal variables. Unless stated otherwise, our simulation sample size is set at $n = 200$. For both cases, we have $\dim(\mathcal{S}_{T|\mathbf{X}}) = 1$, but the proportional hazard assumption is not satisfied. Indeed, for each case, its survival functions $Q(t \mid \mathbf{x})$ at three different covariate values (denoted by $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$) are given in Figure 1, where $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ correspond to the first quartile, median and third quantile of $\boldsymbol{\beta}^T\mathbf{X}$ (or $\boldsymbol{\beta}_2^T\mathbf{X}$) from a simulated data set. We observed the "crossing" pattern in both cases.

In the following, we consider various censoring time and covariate generating scenarios under possible model misspecification settings in Sections 6.3.1–6.3.2. Due to space constraint, additional scenarios and related numerical results are left to Supplements II.2.2–II.2.6.

6.3.1. *Large structural dimension d_1.* First, for Case A, let $C = \exp(2 + \boldsymbol{\beta}_3^T\mathbf{X} + \varepsilon)$, where $\boldsymbol{\beta}_3 = (1, 1, 1, 1, 0, \dots, 0)^T$ and ε is the random error; for Case B, let $C = \exp(3 + \boldsymbol{\beta}_4\mathbf{X} + \varepsilon)$, where $\boldsymbol{\beta}_4 = (0, \dots, 0, 1, 1, 1)^T$. We consider two different choices of the random error (Lu and Li (2011)): (PH) the extreme value distribution $\varepsilon = \log[-\log(1 - U)]$ and (PO) the logistic distribution $\varepsilon = \log[U/(1 - U)]$, where U is uniform distribution on $[0, 1]$. Both cases have $d_1 = \dim(\mathcal{S}_{(T,C)|\mathbf{X}}) = 2$, and we denote these censoring time scenarios by Case A_0 and Case B_0, respectively. In contrast, it is also possible to encounter scenarios with relatively large d_1. For example, for Case A, consider $C = \exp(2 + \boldsymbol{\beta}_3^T\mathbf{X} + 0.5X_4X_5 + 0.1X_6^2 + \varepsilon)$; for Case B, consider $C = \exp(3 + \boldsymbol{\beta}_4^T\mathbf{X} + 0.5X_4X_5 + 0.1X_6^2 + \varepsilon)$. Then both cases have $d_1 = 5$, and we denote them by Case A_1 and Case B_1, respectively.

When true d_1 is relatively large, as the local KM method is involved in our estimation for conditional survival function of censoring time, by the curse of dimensionality and Theorem 4.2, it is expected that the use of a smaller (or underestimated) d_1 can often lead to better empirical results. Similar practice has been adopted in classical work of SDR survival models: for example, Xia, Zhang and Xu (2010) introduced a "working dimension" to find reduced space to apply kernel estimation for survival function of Y; Lu and Li (2011) applied a proportional hazard model to implement the estimation of survival function of C.

Accordingly, to illustrate the performance with possibly underestimated d_1, we simply set $d_1 = 2$ for the DS step, which is correct for Cases A_0 and B_0 but is underestimated for Cases

TABLE 2

Averaged simulation results with relatively large structural dimension d_1

		PH			PO		
Case	Method	Frobenius-norm loss	$\overline{C}_v$	$\overline{IC}_v$	Frobenius-norm loss	$\overline{C}_v$	$\overline{IC}_v$
A_0	Oracle	–	5	0	–	5	0
$(d_1 = 2)$	Coxnet	1.31 (0.01)	1.02	5.44	1.30 (0.01)	1.1	6.45
	DS	1.21 (0.01)	4.77	29.90	1.24 (0.01)	4.87	38.88
	DASH	0.63 (0.02)	4.76	8.88	0.65 (0.02)	4.83	11.35
B_0	Oracle	–	4	0	–	4	0
$(d_1 = 2)$	Coxnet	1.33 (0.01)	0.53	3.81	1.35 (0.01)	0.45	3.68
	DS	1.15 (0.01)	3.94	14.74	1.18 (0.01)	3.84	16.27
	DASH	0.52 (0.02)	3.88	3.53	0.47 (0.02)	3.84	3.47
A_1	Oracle	–	5	0	–	5	0
$(d_1 = 5)$	Coxnet	1.32 (0.01)	0.78	5.05	1.29 (0.01)	1.00	4.84
	DS	1.18 (0.01)	4.78	24.55	1.22 (0.01)	4.82	32.33
	DASH	0.59 (0.02)	4.77	8.35	0.65 (0.02)	4.82	10.25
B_1	Oracle	–	4	0	–	4	0
$(d_1 = 5)$	Coxnet	1.34 (0.01)	0.39	3.62	1.38 (0.01)	0.34	2.99
	DS	1.20 (0.01)	3.80	18.39	1.19 (0.01)	3.88	18.19
	DASH	0.55 (0.02)	3.78	3.71	0.47 (0.02)	3.85	3.48

A_1 and B_1. Sample covariance estimation is used for both DS and DASH. The averaged simulation results over 100 runs summarized in Table 2 show that, like in Cases A_0 and B_0, DASH remains to perform reasonably well in both estimation and variable selection for Cases A_1 and B_1 despite our use of underestimated d_1. In practice, a user may adopt the CV procedure of Section 5.1.2 to determine d_1 (DASH gives similar results and the details are thus omitted). On the other hand, the benchmark Coxnet method (Simon et al. (2011)) does not give satisfactory results in these case studies.

In addition, we considered two other case examples (Case 5 and Case 6) with misspecified (underestimated) d_1, which also showed reasonable numerical performance by DASH. We leave the detailed results including bootstrap confidence intervals in Supplement II.2.5 and Table II.3.

6.3.2. *Approximately sparse central subspace.* As seen from above, different from the penalized Cox model (Coxnet) or other partial likelihood related methods, our proposal does not assume the proportional hazard or require a specific model form, and is therefore intended to be robust to model misspecification in this perspective (which is different from the Coxnet's robustness to model misspecification; e.g., Lu, Goldberg and Fine (2012)). In addition, as is pointed out in Remark 4.1, although we require sparsity in the central subspace, our proposal can be extended to an "approximately" sparse scenarios where a large number of variables are active for $\mathcal{S}_{(T,C)|\mathbf{X}}$ but $\mathcal{S}_{(T,C)|\mathbf{X}}$ can be approximated by a more sparse structure. In the following, we provide numerical illustration on the "approximately" sparse scenarios.

Modifying the settings in Section 6.3.1, under Case A of T, assume $C = \exp(2 + \tilde{\boldsymbol{\beta}}_3^T \mathbf{X} + \varepsilon)$, where $\tilde{\boldsymbol{\beta}}_3 = \boldsymbol{\beta}_3 + \frac{1}{\sqrt{m}} \boldsymbol{\delta}_A$, and $\boldsymbol{\delta}_A \in \mathbb{R}^p$ has its first 100 elements being 1 and the other elements being 0; denote this censoring time scenario by Case A_2. Similarly, under Case B of T, assume $C = \exp(3 + \tilde{\boldsymbol{\beta}}_4^T \mathbf{X} + \varepsilon)$, where $\tilde{\boldsymbol{\beta}}_4 = \boldsymbol{\beta}_4 + \frac{1}{\sqrt{m}} \boldsymbol{\delta}_B$, and $\boldsymbol{\delta}_B \in \mathbb{R}^p$ has its first and last 50 elements being 1 and the other elements being 0; denote this censoring time scenario by Case B_2. Set $m = 20$, 100 or 500. Clearly, larger m indicates better approximation and smaller

TABLE 3
Averaged simulation results with approximately sparse central space

			PH			PO		
Case	m	Method	Frobenius-norm loss	$\overline{C}_v$	$\overline{IC}_v$	Frobenius-norm loss	$\overline{C}_v$	$\overline{IC}_v$
A_2		Oracle	–	5	0	–	5	0
	20	Coxnet	1.38 (0.01)	0.41	3.52	1.37 (0.01)	0.51	4.38
		DS	1.33 (0.01)	4.81	68.54	1.26 (0.01)	4.84	47.85
		DASH	0.83 (0.02)	4.64	13.77	0.76 (0.02)	4.77	13.09
	100	Coxnet	1.35 (0.01)	0.50	3.21	1.35 (0.01)	0.68	4.52
		DS	1.23 (0.01)	4.79	40.27	1.22 (0.01)	4.81	32.49
		DASH	0.70 (0.02)	4.77	11.00	0.66 (0.02)	4.79	8.92
	500	Coxnet	1.34 (0.01)	0.44	2.65	1.30 (0.01)	1.09	6.17
		DS	1.21 (0.01)	4.77	35.61	1.20 (0.01)	4.82	27.58
		DASH	0.67 (0.02)	4.76	11.50	0.62 (0.02)	4.80	9.11
B_2		Oracle	–	4	0	–	4	0
	20	Coxnet	1.37 (0.01)	0.19	2.02	1.39 (0.01)	0.19	2.60
		DS	1.34 (0.02)	3.49	43.80	1.33 (0.02)	3.65	39.82
		DASH	1.02 (0.03)	3.00	10.46	0.83 (0.03)	3.28	6.68
	100	Coxnet	1.35 (0.01)	0.25	2.40	1.38 (0.01)	0.32	3.80
		DS	1.22 (0.01)	3.84	20.77	1.24 (0.02)	3.80	22.83
		DASH	0.64 (0.03)	3.76	4.55	0.55 (0.03)	3.75	3.68
	500	Coxnet	1.36 (0.01)	0.33	4.11	1.36 (0.01)	0.24	2.73
		DS	1.19 (0.01)	3.86	18.33	1.18 (0.01)	3.83	17.81
		DASH	0.58 (0.03)	3.83	4.65	0.45 (0.02)	3.83	2.95

deviation of $\mathcal{S}_{(T,C)|\mathbf{X}}$ in Case A_2 (or B_2) from the original sparsity structure in Case A_0 (or B_0). We kept other simulation settings the same as that of Section 6.3.1, and summarized the results in Table 3.

Compared to Coxnet, the DASH method can still provide reasonable results in both estimation and variable selection even though some relevant variables got missed in estimation of the censoring time probability. Meanwhile, as the DASH estimation is partially influenced by the central subspace estimation errors in the first DS step, in alignment with Theorem 4.3, the DASH estimation errors tend to decrease as we decreased the relatively weak signals by increasing m from 20 to 500. We expect similar phenomenon applies to approximately sparse scenarios for the survival time probability as well. To some extent, these numerical results provide some evidence that the DASH method may often be applicable when the central subspace violates the strict sparsity assumption and some relevant variables (with weak signals) are missed.

7. A real data example. In this section, we describe our analysis on the kidney renal clear cell carcinoma (KIRC) data, which was downloaded directly from the National Cancer Institute's GDC Data Portal platform (https://portal.gdc.cancer.gov/) under project ID TCGA-KIRC. The KIRC data contains 530 patients with clinical information (including survival time and censoring time) as well as their gene expressions based on 57,251 genes from the next-generation sequencing technique known as RNA-Seq. Considering that RNA-Seq contains count data with unique structures including high skewness with many zeros, widely different sequencing depth and over-dispersion, we performed data preprocessing steps, the description of which is left in Supplement II.3. The cleaned data after preprocessing has $p = 2962$ genes and sample size $n = 265$ in both training and testing sets.

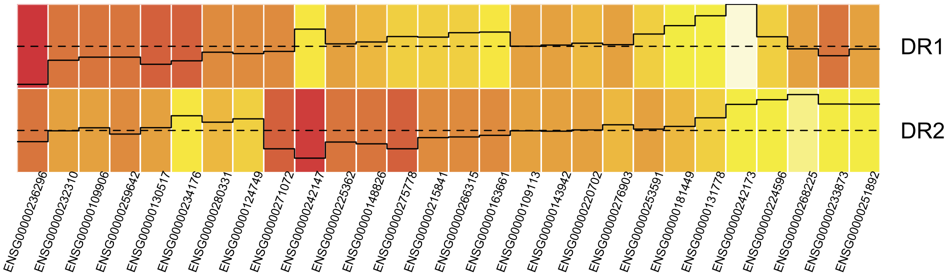

FIG. 2. *Loadings of regression directions for DASH. Column labels are the gene names. The solid lines represent relative magnitude and sign of the loading values, and the dashed lines represent* 0.

We then applied the proposed DASH method along with the DS procedure to the cleaned training set to obtain the estimated basis $\hat{\Gamma}_0$ and $\hat{\Gamma}_1$, respectively. We also applied the Coxnet method as a benchmark, which implicitly assumes $d = 1$ under the parametric Cox model. Both DS and DASH methods chose structural dimension $d = 2$, and the number of selected genes (#Var) are given in Table 3, with DASH having the smallest model size among the three. In particular, we plot loadings of the two directions (denoted by DR1 and DR2) of $\hat{\Gamma}_0$ by a heatmap in Figure 2, with names of the selected genes marked under the corresponding cells. Interestingly, we found that multiple selected genes such as ENSG00000131778 (CHD1L, Cheng, Su and Xu (2013)) and ENSG00000181449 (SOX2, Santini et al. (2014)), among others, have been recently reported as useful prognostic biomarkers.

In the following, we show that the basis estimators of DS and DASH methods indeed performed well in predicting KIRC patient's cancer prognosis. Specifically, with DS and DASH estimators given above, the sufficient predictors had dimension $d = 2$, and we used the training data to build a gradient boosting machine (GBM; Friedman (2001)), where a proportional hazard model $h(t \mid \mathbf{X}) = h_0(t) \exp(R(Z_1, Z_2))$ is assumed with (Z_1, Z_2) being the sufficient predictors, $R(Z_1, Z_2)$ is the general (nonlinear) link function that can be viewed as a patient's risk score, and $h_0(t)$ is a base hazard function. The DASH link function surface in Figure 3 exhibited nonlinear patterns of the two sufficient predictors, as opposed to the linear assumption of Coxnet (the DS showed similar nonlinear patterns and is thus omitted).

Using the models built above, we then computed the risk scores for all patients in the testing set and assigned them into the "high-risk" and "low-risk" groups using the median of the training-set risk scores as the cutoff. In Figure 4(a) and Figure 4(b), we plotted the testing-set KM estimator curves and their confidence interval curves of the two groups generated

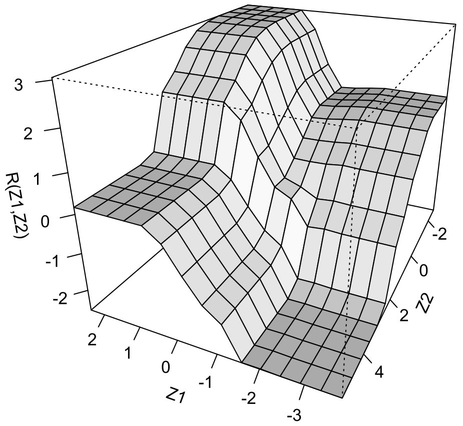

FIG. 3. *Link function $R(Z_1, Z_2)$ from GBM using DASH predictors.*

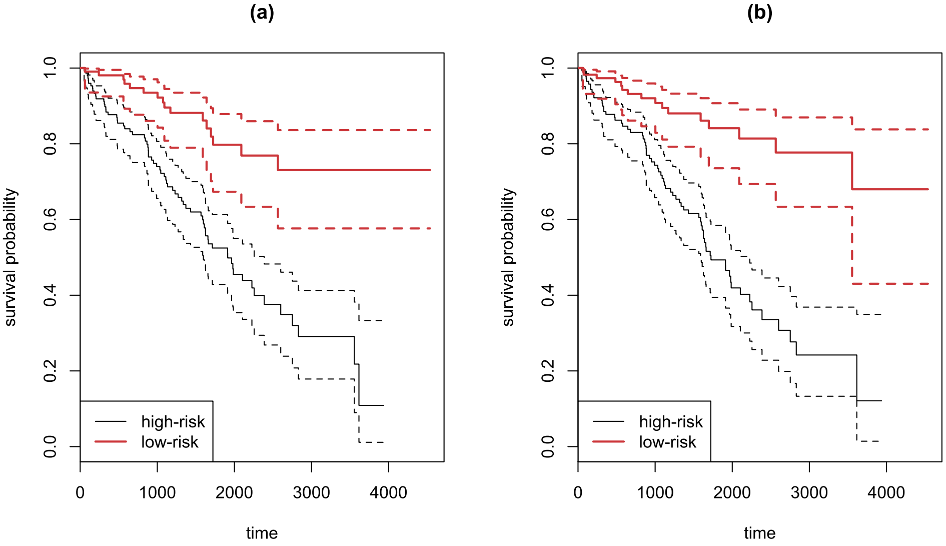

FIG. 4. *The testing set Kaplan–Meier curves of segmented patient groups.* (*a*) *DS method*; (*b*) *DASH method. In both panels, the thin lines represent "high-risk" group and the thick lines represent "low-risk" group. The dashed lines are the confidence interval curves.*

from DASH and DS methods, respectively. We also performed log-rank tests to compare survival functions of the two groups and showed p-values in Table 3. Both the KM curves and p-values obtained from our methods confirmed that the high-risk and low-risk groups segmented by the risk scores indeed had significantly different prognostic patterns.

Furthermore, we evaluated the risk scores computed above by the time-dependent receiver operating characteristic (ROC) curve analysis (Heagerty, Lumley and Pepe (2000)). Using median(Y) of the training set as the cutoff time point, we considered sensitivity-specificity curves constructed using the testing set, and the corresponding areas under curve (AUC) are listed in Table 3. Satisfactorily, DS and DASH performed very competitively compared to the benchmark. It is not surprising that DS can perform well in this example given that it selected the same structural dimension as DASH. In addition, our evaluations on prediction performance showed similar patterns and advantages, the details of which are left in Supplement II.3.

8. Concluding remarks.

We propose a promising model-free double-slicing assisted SDR method for high-dimensional censored data, which is a flexible alternative to existing model-based approaches, such as high-dimensional Cox models. The new development achieves simultaneous dimension reduction and variable selection while preserving full information for the distribution of $T \mid \mathbf{X}$. With new technical tools to handle censored response, we establish both estimation consistency and variable selection consistency that allow p to grow exponentially with n, and obtain uniform convergence for the nonparametric survival function estimation. As evidenced by numerical studies, our model-free proposal can greatly facilitate the practical application of robust nonparametric approaches in the estimation of conditional survival functions with high-dimensional covariates.

APPENDIX: PROPOSITIONS

PROPOSITION A.1. *Suppose the central subspace $\mathcal{S}_{T|\mathbf{X}}$ exists and is unique. Then the set of active variables $\mathcal{A}_0$ for $\mathcal{S}_{T|\mathbf{X}}$ is also unique.*

PROPOSITION A.2. *Given the objective function*

$$F(\Gamma, \Phi) = \operatorname{tr}\{(U_c - \Sigma\Gamma\Phi)^T \Sigma^{-1}(U_c - \Sigma\Gamma\Phi)\} \quad \text{subject to } \Phi\Phi^T = I_d, \tag{A.1}$$

where $\Gamma \in \mathbb{R}^{p\times d}$ *and* $\Phi \in \mathbb{R}^{d\times b}$ *are parameters. Let* (Γ_0, Φ_0) *be any minimizer of* (A.1). *Then* $\Gamma_0 \in \mathbb{R}^{p\times d}$ *forms a basis matrix of* $\mathcal{S}_{T|\mathbf{X}}$ *and* Φ_0 *is the corresponding coordinate matrix.*

PROPOSITION A.3. *Let* $\mathcal{A}_0$ *be the set of active variables for* $\mathcal{S}_{T|\mathbf{X}}$ *and let* Γ_0 *be a basis matrix of* $\mathcal{S}_{T|\mathbf{X}}$. *Then conditional on* $\mathbf{X}_{\mathcal{A}_0}$ *or* $\Gamma_0^T\mathbf{X}$, T *and* C *are independent.*

PROPOSITION A.4. *Let* $\hat{\Xi} = \hat{\Gamma}_0\hat{\Phi}_0$. *Suppose* $\hat{\Gamma}_{01} \in \mathbb{R}^{p\times d}$ *is the semiorthogonal matrix consisting of left-singular vectors of* $\hat{\Xi}$ *with the nonzero singular values. If* $\|\hat{\Xi} - \tilde{U}_c\|_F \to 0$ *as* $n \to 0$ *and Condition* (C5) *holds, then for large enough* n, *there exists a basis* Γ_0 *of* $\mathcal{S}_{T|\mathbf{X}}$ *such that*

$$\|\hat{\Gamma}_{01} - \Gamma_0\|_F \le \sqrt{2}\|\sin\Theta(\mathcal{S}_{\hat{\Xi}}, \mathcal{S}_{T|\mathbf{X}})\|_F \le c_a\|\hat{\Xi} - \tilde{U}_c\|_F, \tag{A.2}$$

where $c_a > 0$ *is some constant, and* $\Theta(\mathcal{S}_{\hat{\Xi}}, \mathcal{S}_{T|\mathbf{X}})$ *is* $d \times d$ *diagonal matrix in which the* j*th diagonal entry is the* j*th principle angle between* $\mathcal{S}_{\hat{\Xi}}$ *and* $\mathcal{S}_{T|\mathbf{X}}$.

PROPOSITION A.5. *The coverage* $\mathcal{S}_{(Y,\delta)|\mathbf{X}} \supseteq \mathcal{S}_{(T,C)|\mathbf{X}}$ *holds if and only if* $\mathcal{S}_{T|\mathbf{X}} \subseteq \mathcal{S}_{(Y,\delta)|\mathbf{X}}$ *and* $\mathcal{S}_{C|\mathbf{X}} \subseteq \mathcal{S}_{(Y,\delta)|\mathbf{X}}$.

PROPOSITION A.6. *Let* Γ_1 *be any minimizer of* Γ *in* (3.5). *Given any index set* $\mathcal{A} \subset \{1, \dots, p\}$, *let* $\Gamma_{1,\mathcal{A}}$ *be the* $|\mathcal{A}| \times d_1$ *submatrix of* Γ_1 *consisting of the rows corresponding to* $\mathcal{A}$. *Then given objective function* (3.5) *and an index set* $\mathcal{A}$, *we have that* $\|\Gamma_{1,\mathcal{A}}\|_{2,1}$ *is unique.*

Acknowledgment. We would like to thank the Editors, Associate Editor and two anonymous referees for their valuable comments that help to improve this manuscript significantly.

The first and second authors were supported in part by NSF Grant DMS-1916376 and DE-CTR ACCEL/NIH U54 GM104941 SHoRe award.

The third author was supported in part by NSF Grant DMS-1712706.

SUPPLEMENTARY MATERIAL

Supplement to "Double-slicing assisted sufficient dimension reduction for high-dimensional censored data" (DOI: 10.1214/19-AOS1880SUPP; .pdf). The supplemental file (Ding, Qian and Wang (2020)) contains proofs, technical details and numerical results. Supplement I.1 gives proofs of the propositions; Supplement I.2 assembles some useful lemmas; Supplements I.3–I.5 provide proofs for the main theorems; Supplement II contains additional results on computation, simulation and data analysis.

REFERENCES

BERAN, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley, CA.

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600

BRADIC, J., FAN, J. and JIANG, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39** 3092–3120. MR3012402 https://doi.org/10.1214/11-AOS911

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees. Wadsworth Statistics/Probability Series.* Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

BURA, E., DUARTE, S. and FORZANI, L. (2016). Sufficient reductions in regressions with exponential family inverse predictors. *J. Amer. Statist. Assoc.* **111** 1313–1329. MR3561952 https://doi.org/10.1080/01621459.2015.1093944

CHAI, H., ZHANG, Q., HUANG, J. and MA, S. (2019). Inference for low-dimensional covariates in a high-dimensional accelerated failure time model. *Statist. Sinica* **29** 877–894. MR3931392

CHEN, X., COOK, R. D. and ZOU, C. (2015). Diagnostic studies in sufficient dimension reduction. *Biometrika* **102** 545–558. MR3394274 https://doi.org/10.1093/biomet/asv016

CHENG, W., SU, Y. and XU, F. (2013). CHD1L: A novel oncogene. *Molecular Cancer* **12** 170.

CHENG, S. C., WEI, L. J. and YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82** 835–845. MR1380818 https://doi.org/10.1093/biomet/82.4.835

COOK, R. D. (1998). *Regression Graphics*: *Ideas for Studying Regressions Through Graphics*. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. A Wiley-Interscience Publication. MR1645673 https://doi.org/10.1002/9780470316931

COOK, R. D. (2003). Dimension reduction and graphical exploration in regression including survival analysis. *Stat. Med.* **22** 1399–1413.

COOK, R. D., FORZANI, L. and ROTHMAN, A. J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *Ann. Statist.* **40** 353–384. MR3014310 https://doi.org/10.1214/11-AOS962

COOK, R. D. and NI, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100** 410–428. MR2160547 https://doi.org/10.1198/016214504000001501

COX, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189 https://doi.org/10.1093/biomet/62.2.441

DING, S., QIAN, W. and WANG, L. (2020). Supplement to "Double-slicing assisted sufficient dimension reduction for high-dimensional censored data." https://doi.org/10.1214/19-AOS1880SUPP.

DU, P., MA, S. and LIANG, H. (2010). Penalized variable selection procedure for Cox models with semiparametric relative risk. *Ann. Statist.* **38** 2092–2117. MR2676884 https://doi.org/10.1214/09-AOS780

FAN, J. and GIJBELS, I. (1994). Censored regression: Local linear approximations and their applications. *J. Amer. Statist. Assoc.* **89** 560–570. MR1294083

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656 https://doi.org/10.1214/aos/1015362185

FANG, E. X., NING, Y. and LIU, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1415–1437. MR3731669 https://doi.org/10.1111/rssb.12224

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

GERDS, T. A. and SCHUMACHER, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* **63** 1283–1287, 1316. MR2414608 https://doi.org/10.1111/j.1541-0420.2007.00832.x

GONZALEZ-MANTEIGA, W. and CADARSO-SUAREZ, C. (1994). Asymptotic properties of a generalized Kaplan–Meier estimator with some applications. *J. Nonparametr. Stat.* **4** 65–78. MR1366364 https://doi.org/10.1080/10485259408832601

HALL, P. and LI, K.-C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.* **21** 867–889. MR1232523 https://doi.org/10.1214/aos/1176349155

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171 https://doi.org/10.1214/aos/1176349020

HEAGERTY, P. J., LUMLEY, T. and PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56** 337–344.

HSING, T. and REN, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *Ann. Statist.* **37** 726–755. MR2502649 https://doi.org/10.1214/07-AOS589

HUANG, J., SUN, T., YING, Z., YU, Y. and ZHANG, C.-H. (2013). Oracle inequalities for the LASSO in the Cox model. *Ann. Statist.* **41** 1142–1165. MR3113806 https://doi.org/10.1214/13-AOS1098

JOHNSON, B. A. (2009). On lasso for censored data. *Electron. J. Stat.* **3** 485–506. MR2507457 https://doi.org/10.1214/08-EJS322

KONG, E. and XIA, Y. (2014). An adaptive composite quantile approach to dimension reduction. *Ann. Statist.* **42** 1657–1688. MR3262464 https://doi.org/10.1214/14-AOS1242

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, L. (2005). Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics* **22** 466–471.

LI, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011 https://doi.org/10.1093/biomet/asm044

LI, B. (2018). *Sufficient Dimension Reduction*: *Methods and Applications with R*. *Monographs on Statistics and Applied Probability* **161**. CRC Press, Boca Raton, FL. MR3838449 https://doi.org/10.1201/9781315119427

LI, B. and DONG, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Ann*. *Statist*. **37** 1272–1298. MR2509074 https://doi.org/10.1214/08-AOS598

LI, L. and LI, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics* **20** 3406–3412.

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J*. *Amer*. *Statist*. *Assoc*. **102** 997–1008. MR2354409 https://doi.org/10.1198/016214507000000536

LI, K.-C., WANG, J.-L. and CHEN, C.-H. (1999). Dimension reduction for censored regression data. *Ann*. *Statist*. **27** 1–23. MR1701098 https://doi.org/10.1007/3-540-48294-6_14

LI, L. and YIN, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131, 323. MR2422826 https://doi.org/10.1111/j.1541-0420.2007.00836.x

LIN, Q., ZHAO, Z. and LIU, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *Ann*. *Statist*. **46** 580–610. MR3782378 https://doi.org/10.1214/17-AOS1561

LOPEZ, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Comm*. *Statist*. *Theory Methods* **40** 2639–2660. MR2860770 https://doi.org/10.1080/03610926.2010.489175

LU, W., GOLDBERG, Y. and FINE, J. P. (2012). On the robustness of the adaptive lasso to model misspecification. *Biometrika* **99** 717–731. MR2966780 https://doi.org/10.1093/biomet/ass027

LU, W. and LI, L. (2011). Sufficient dimension reduction for censored regression. *Biometrics* **67** 513–523. MR2829020 https://doi.org/10.1111/j.1541-0420.2010.01490.x

MA, Y. and ZHU, L. (2012). A semiparametric approach to dimension reduction. *J*. *Amer*. *Statist*. *Assoc*. **107** 168–179. MR2949349 https://doi.org/10.1080/01621459.2011.646925

NADKARNI, N. V., ZHAO, Y. and KOSOROK, M. R. (2011). Inverse regression estimation for censored data. *J*. *Amer*. *Statist*. *Assoc*. **106** 178–190. MR2816712 https://doi.org/10.1198/jasa.2011.tm08250

QIAN, W., DING, S. and COOK, R. D. (2018). Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *J*. *Amer*. *Statist*. *Assoc*. To appear.

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J*. *Amer*. *Statist*. *Assoc*. **104** 177–186. MR2504372 https://doi.org/10.1198/jasa.2009.0101

SANTINI, R., PIETROBONO, S., PANDOLFI, S., MONTAGNANI, V., D'AMICO, M., PENACHIONI, J., VINCI, M., BORGOGNONI, L. and STECCA, B. (2014). SOX2 regulates self-renewal and tumorigenicity of human melanoma-initiating cells. *Oncogene* **33** 4697–4708.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J*. *Stat*. *Softw*. **39** 1–13.

SUN, Q., ZHU, R., WANG, T. and ZENG, D. (2019). Counting process-based dimension reduction methods for censored outcomes. *Biometrika* **106** 181–196. MR3912390 https://doi.org/10.1093/biomet/asy064

TAN, K. M., WANG, Z., LIU, H. and ZHANG, T. (2018). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated Rayleigh flow. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **80** 1057–1086. MR3874310 https://doi.org/10.1111/rssb.12291

TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat*. *Med*. **16** 385–395.

WANG, J.-L., XUE, L., ZHU, L. and CHONG, Y. S. (2010). Estimation for a partial-linear single-index model. *Ann*. *Statist*. **38** 246–274. MR2589322 https://doi.org/10.1214/09-AOS712

WANG, T., CHEN, M., ZHAO, H. and ZHU, L. (2018). Estimating a sparse reduction for general regression in high dimensions. *Stat*. *Comput*. **28** 33–46. MR3741635 https://doi.org/10.1007/s11222-016-9714-6

XIA, Y., ZHANG, D. and XU, J. (2010). Dimension reduction and semiparametric estimation of survival models. *J*. *Amer*. *Statist*. *Assoc*. **105** 278–290. MR2656052 https://doi.org/10.1198/jasa.2009.tm09372

YANG, Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann*. *Statist*. **35** 2450–2473. MR2382654 https://doi.org/10.1214/009053607000000514

YE, Z. and WEISS, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J*. *Amer*. *Statist*. *Assoc*. **98** 968–979. MR2041485 https://doi.org/10.1198/016214503000000927

YIN, X. and HILAFU, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **77** 879–892. MR3382601 https://doi.org/10.1111/rssb.12093

YIN, X. and LI, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann*. *Statist*. **39** 3392–3416. MR3012413 https://doi.org/10.1214/11-AOS950

YU, Z., DONG, Y. and SHAO, J. (2016). On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *Ann*. *Statist*. **44** 2594–2623. MR3576555 https://doi.org/10.1214/15-AOS1424

YU, Z., ZHU, L., PENG, H. and ZHU, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika* **100** 641–654. MR3094442 https://doi.org/10.1093/biomet/ast005

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x

ZENG, D. (2004). Estimating marginal survival function by adjusting for dependent censoring using many covariates. *Ann. Statist.* **32** 1533–1555. MR2089133 https://doi.org/10.1214/009053604000000508

ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520

ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. MR2410017 https://doi.org/10.1093/biomet/asm037

ZHAO, G., MA, Y. and LU, W. (2017). Efficient estimation for dimension reduction with censored data. arXiv preprint, arXiv:1710.05377.

ZHU, L., MIAO, B. and PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.* **101** 630–643. MR2281245 https://doi.org/10.1198/016214505000001285

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849 https://doi.org/10.1198/jasa.2011.tm10563

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735