Biometrika (2024), **xx**, x, *pp*. 1–18
Printed in Great Britain

Nonparametric tests of independence for high-dimensional survival data

BY JINHONG LI

Department of Statistics, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, China
jinhongli0106@gmail.com

JICAI LIU

School of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance, 995 Shangchuan Road, Shanghai 201209, China
liujicai1234@126.com

SHUJIE MA

Department of Statistics, University of California, Riverside, USA
shujie.ma@ucr.edu

AND RIQUAN ZHANG

School of Statistics and Information, Shanghai University of International Business and Economics, 1900 Wenxiang Road, Shanghai 201620, China
zhangriquan@163.com

SUMMARY

In survival analysis, testing the independence between survival time and high-dimensional covariates is crucial. However, traditional test methods often prove insufficient for high-dimensional scenarios. This paper introduces a novel framework leveraging the counting process technique, which transforms the originally intractable problem caused by censoring into a test of conditional moment restrictions for complete observations. Within this framework, we construct a rich class of dependency metrics, including the recent kernel logrank test statistic as a special case. We further investigate the asymptotic behavior of these metrics in high dimensions. It reveals that as dimensionality grows, they are approximated by the sum of squared marginal covariances. However, these covariances with complex forms do not clearly delineate the relationship between survival time and covariates. Interestingly, we identify a general class of conditions where the survival time and covariates are dependent, yet the covariances equal zero. This finding suggests that the kernel logrank test in high dimensions can only capture a specific kind of linear dependence for censored data. To overcome this limitation, we propose a new test and rigorously establish its asymptotic normality under both the null and local alternative hypotheses. Extensive numerical studies demonstrate the superior power of our proposed test compared to existing methods.

Some key words: Nonparametric independence test; Conditional moment restrictions; High dimension; Counting process; Survival analysis.

1. INTRODUCTION

In survival data analysis, the primary outcome of interest is the time until an event occurs, such as the time until death in cancer patients or the time until tumor recurrence. A unique feature of survival data is that survival times are usually subject to censoring due to various reasons, such as loss to follow-up, the end of study, or other competing risks. And a central goal is to determine the relationship between survival time and explanatory covariates. Over the last few decades, various regression models have been proposed, including the proportional hazards (PH) model (Cox, 1972), and the accelerated failure time (AFT) model (Kalbfleisch & Prentice, 2002). Unlike these regression approaches, this article focuses on testing the hypothesis:

$$H_0 : T \text{ and } \mathbf{Z} \text{ are independent} \quad \text{versus} \quad H_1 : \text{otherwise}, \tag{1}$$

where T is a survival time and $\mathbf{Z}$ is a covariate vector taking values in $\mathcal{Z} \subseteq \mathbb{R}^p$. Note that in practice, T may not be fully observed due to censoring.

The study of testing (1) can be traced back to the well-known log-rank test for equality of distributions proposed by Mantel (1966). More recently, Edelmann et al. (2021) proposed a survival distance covariance to quantify the dependence between the survival time and covariates. Rindt et al. (2021) transformed censored data into uncensored data and then utilized the HSIC to test for independence. Further, Fernández et al. (2023) introduced a kernel log-rank test by incorporating the supremum over a set of weighted log-rank tests.

The methods mentioned were developed in a low-dimensional setting. However, there is an urgent need to perform tests of independence for high-dimensional data across various fields, such as genomics, medicine and engineering. Recently, some efforts have been directed towards high-dimensional uncensored data. Székely & Rizzo (2013) introduced a bias-corrected sample distance correlation to test for independence in high dimensions. Zhu et al. (2020) proved that the sample distance correlation and HSIC can only capture linear dependencies in the high-dimensional, low-sample-size (HDLSS) setting. Gao et al. (2021) established the limiting distribution of the bias-corrected sample distance correlation both the dimension and sample size increase. In summary, high-dimensional independence testing is a challenging and ongoing area of research. In this article, we are concerned with the topic of high-dimensional survival data, i.e., testing problem (1) when p is large.

In this article, we present a unique viewpoint on problem (1) through employing the counting process technique (Fleming & Harrington, 1991). This approach simplifies problem (1) to a test of conditional moment restrictions with respect to a fully observable stochastic process, as seen in Lemma 1. A significant advantage of this technique is its ability to convert the challenging issue of censoring in problem (1) into testing of conditional moment restrictions (CMR) for complete observations. Testing CMR is a foundational topic in econometrics and finance, with a vast amount of literature. Notable works include Bierens (1982); Newey (1985); Hong & White (1995); Stute (1997); Stinchcombe & White (1998); Escanciano (2006a), and Shao & Zhang (2014). This well-established field provides a wealth of tools that can be applied to our framework. Under this framework, we can construct a rich class of measures, called survival integrated conditional moment (SICM) metrics, to quantify the dependence between survival time and covariates. The kernel logrank test statistic turns out to be a special case in this class.

Another advantage of the novel framework is that it facilitates the investigation of the asymptotic behavior of the proposed measures, including the kernel logrank test. It reveals that the kernel logrank test statistic can be approximated by the sum of squared marginal covariances, as detailed in Theorem 2. However, the presence of censoring complicates these covariances, giving them a complex form. To shed light on this complexity, we provide a general class of conditions under which T and $\mathbf{Z}$ exhibit dependence, yet the covariances equal zero, as described in Theorem 3. This indicates that the kernel logrank test in high dimensions can only capture a special kind of linear dependence for censored data. To overcome this shortcoming, we introduce a new test employing the marginal aggregation of the proposed SICM measures. Extensive simulation studies confirm that our proposed marginal methods outperform existing tests in terms of statistical power.

In summary, the contributions of this article are as follows: (1) We have developed a novel framework that offers a feasible approach to studying the asymptotic behavior of the kernel logrank test within high-dimensional contexts. (2) We demonstrate that the kernel logrank test can be approximated by the sum of squared componentwise covariances, a finding that highlights its inherent limitations. (3) We identify a general class of conditions under which the survival time and the covariates are nonlinearly dependent but uncorrelated, revealing a loss of power for both the kernel logrank test and survival distance covariance methods. (4) To address these limitations, we propose a new marginal test and investigate its asymptotic properties under both the null and local alternative hypotheses within high-dimensional scenarios.

The rest of the paper is organized as follows. In Section 2, we introduce the novel CMR-based framework and propose the SICM metrics. Section 3 establish the asymptotic behavior of the SICM measures in high dimensions. In Section 4, we propose a new SICM-based test utilizing marginal aggregation. In Sections 5 and 6, we assess the finite sample performance of the proposed tests through Monte Carlo simulations and real data applications. Finally, we provide a brief discussion in Section 7. All the technical proofs are presented in the Appendix.

2. CMR-BASED FRAMEWORK AND SICM METRICS

Let T be the continuous survival time and C the censoring time. We denote the observed event time by $X = \min\{T, C\}$ and the censoring indicator by $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Throughout the paper, we assume that C is independent of T conditional on the covariates $\mathbf{Z}$. We define the failure counting process as $N(t) = I(X \leq t, \Delta = 1)$ and the at-risk process $Y(t) = I(t \leq X)$.

Note that problem (1) could be unidentifiable due to censoring. For instance, consider a simple example where $\sup\{t : S_C(t) > 0\} = c_0$ and $\sup\{t : S_T(t) > 0\} = c_0 + 1$, for some positive c_0. If $TI(0 \leq T \leq c_0)$ is independent of $\mathbf{X}$, but $TI(c_0 < T \leq c_0 + 1)$ is dependent on $\mathbf{Z}$, then whether T is dependent on $\mathbf{Z}$ or not cannot be discerned from the observed data, as we only observe T up to c_0. To ensure identifiability, we assume in the article that $TI(T > \tau)$ is independent of $\mathbf{Z}$, where $\tau = \sup\{t : S_X(t) > 0\}$, known as the "identifiable condition". This condition is less restrictive than Assumption 3.1 in Fernández et al. (2023), which states that if $S_{C|\mathbf{Z}=\mathbf{z}}(t) = 0$, then $S_{T|\mathbf{Z}=\mathbf{z}}(t) = 0$ for almost all $\mathbf{z} \in \mathbb{R}^p$, implying that the support set of $S_{C|\mathbf{Z}=\mathbf{z}}(t)$ is at least as large as that of $S_{T|\mathbf{Z}=\mathbf{z}}(t)$. Under the identifiable condition and using the counting process technique, we can show the following result.

LEMMA 1. *Let $dN(t) = N((t+dt)^-) - N(t^-)$ be an increment of $N(\cdot)$ over $[t, t+dt)$. Assume that the identifiable condition holds and $\mathbb{E}\{Y(\tau)|\mathbf{Z}\}$ is bounded away from zero. Then, we have that*

$$T \perp\!\!\!\perp \mathbf{Z} \Longleftrightarrow \mathbb{E}\{Y(t)d\widetilde{N}(t)|\mathbf{Z}\} = 0, \text{ as } dt \to 0^+, \tag{2}$$

where $d\widetilde{N}(t) = dN(t) - \mathbb{E}\{dN(t)|Y(t)\}$ for almost all $t \in (0, \tau]$. Here, " $\Leftrightarrow$ " stands for "equivalent to".

The assumption that $\mathbb{E}\{Y(\tau)|\mathbf{Z}\}$ is bounded away from zero, is standard in survival models. This implies that at least some subjects do not fail at the end time τ. Lemma 1 demonstrates that problem (1) is equivalent to testing the CMR problem $\mathbb{E}\{Y(t)d\widetilde{N}(t)|\mathbf{Z}\} = 0$ with respect to the process $Y(t)d\widetilde{N}(t)$ and $\mathbf{Z}$. Since $N(t)$ and $Y(t)$ are fully observed, the test of the CMR is essentially for complete data. This is a foundational issue in economics and finance with a long history, as outlined in the Introduction. This new insight connects the independence testing in censored data with the CMR testing for complete data. Consequently, the approaches from CMR testing can be utilized to tackle problems involving censored data.

Following the commonly used integrated conditional moment approach (Bierens, 1982; Stute, 1997; Bierens & Ploberger, 1997; Escanciano, 2006a), we can convert the constraint on the conditional expectation into an infinite number of unconditional moment restrictions. Let $\mathcal{H}$ be a set of measurable functions on $\mathcal{Z}$. Then, we have

$$\mathbb{E}\{Y(t)d\widetilde{N}(t)|\mathbf{Z}\} = 0 \Longleftrightarrow \mathbb{E}\{Y(t)d\widetilde{N}(t)w(\mathbf{Z}, \mathbf{u})\} = 0, \forall\, w(\cdot,\cdot) \in \mathcal{H}, \mathbf{u} \in \mathcal{U}, \tag{3}$$

where $\mathcal{U}$ is some suitable space. The choice of the family $\mathcal{H}$ is crucial to satisfy (3). See Bierens & Ploberger (1997); Escanciano (2006b) for sufficient conditions on such $\mathcal{H}$. Examples of the family $\mathcal{H}$ from the literature include the indicator functions $\{I(\mathbf{Z} \leq \mathbf{u}), \mathbf{u} \in \mathbb{R}^p\}$ (Stute, 1997), the complex exponential functions $\{\exp(i\mathbf{u}^T\mathbf{Z}), \mathbf{u} \in \mathbb{R}^p, i = \sqrt{-1}\}$ (Bierens, 1982), the real exponential functions $\{\exp(\mathbf{u}^T\mathbf{Z}), \mathbf{u} \in \mathbb{R}^p\}$ (Bierens, 1990). More recently, Muandet et al. (2020) have defined $\mathcal{H}$ as the unit ball in a reproducing kernel Hilbert space (RKHS), equipped with an integrally strictly positive definite (ISPD) kernel function.

For simplicity, we hereafter only consider $\mathcal{H} = \{\exp(i\mathbf{u}^T\mathbf{Z}), \mathbf{u} \in \mathbb{R}^p\}$. Note that

$$\mathbb{E}\{w(\mathbf{Z}, \mathbf{u})Y(t)d\widetilde{N}(t)\} = 0 \Longleftrightarrow \mathbb{E}\Big\{\int_0^t w(\mathbf{Z}, \mathbf{u})Y(s)d\widetilde{N}(s)\Big\} = 0, \ \forall t \in [0, \tau].$$

These, along with (2) and (3), motivate us to define the following metric to quantify the dependence between T and $\mathbf{Z}$.

DEFINITION 1. *The Survival Integrated Conditional Moment (SICM) metric between the survival time T and covariates $\mathbf{Z}$ is defined as*

$$\mathrm{SICM}(t; a, \nu) = \int_{\mathbb{R}^p} \Big\|\mathbb{E}\Big\{\int_0^t a(s)\exp(i\boldsymbol{u}^T\mathbf{Z})Y(s)d\widetilde{N}(s)\Big\}\Big\|^2 d\nu(\boldsymbol{u}),$$

where $a(t)$ and $\nu(\boldsymbol{u})$ are two given positive weight functions for which the integral above exists.

It is crucial to select the weight functions $a(t)$ and $\nu(\mathbf{u})$ appropriately in the above definition. The choice of $a(t)$ is discussed in Section 4. For $\nu(\mathbf{u})$, we here consider the following two types of weight functions:

(i) $d\nu(\mathbf{u}) = c_p^{-1}\|\mathbf{u}\|^{-(1+p)}d\mathbf{u}$ with $c_p = \pi^{(p+1)/2}/\Gamma((1+p)/2)$;
(ii) ν is any finite nonnegative Borel measure on $\mathbb{R}^p$.

The two types of weight functions are commonly used in the literature and correspond to two classes of measures: non-integrable and integrable. Specifically, the first type of weight function, proposed by Székely et al. (2007), is a special case of non-integrable weighting measures. In contrast, the second type represents a class of integrable weighting measures, as described by Sriperumbudur et al. (2010). These two types of weight functions have the following appealing merits:

LEMMA 2. (i) *(Székely et al., 2007, Lemma 1) For any $\mathbf{z} \in \mathbb{R}^p$, we have that*

$$\int_{\mathbb{R}^p} \frac{1 - \cos\left(\boldsymbol{u}^T\mathbf{z}\right)}{c_p\|\boldsymbol{u}\|^{1+p}}d\boldsymbol{u} = \|\mathbf{z}\|.$$

(ii) *(Bochners Theorem (Wendland, 2004, Theorem 6.6)) A continuous function $\Psi : \mathbb{R}^p \to \mathbb{C}$ is positive semi-definite if and only if it is the Fourier transform of a finite nonnegative Borel measure ν on $\mathbb{R}^p$, that is, for any $\mathbf{z} \in \mathbb{R}^p$,*

$$\Psi(\mathbf{z}) = \int_{\mathbb{R}^p} e^{-i\boldsymbol{u}^T\mathbf{z}}d\nu(\boldsymbol{u}).$$

Let $K(\mathbf{z}_1, \mathbf{z}_2) = \Psi(\mathbf{z}_1 - \mathbf{z}_2)$ be a shift-invariant kernel for some positive definite function Ψ. According to Bochner's Theorem, all continuous shift-invariant kernels on $\mathbb{R}^p$ can characterize the second set of weight functions mentioned. Popular examples of such kernels include Gaussian kernel $K(\mathbf{z}_1, \mathbf{z}_2) = \exp(-\|\mathbf{z}_1 - \mathbf{z}_2\|^2/\gamma)$, Laplacian kernel $K(\mathbf{z}_1, \mathbf{z}_2) = \exp\left(-\|\mathbf{z}_1 - \mathbf{z}_2\|/\gamma\right)$, and inverse multiquadric kernel $K(\mathbf{z}_1, \mathbf{z}_2) = \{c + \|\mathbf{z}_1 - \mathbf{z}_2\|^2\}^{-\gamma}$, where $\gamma, c > 0$.

According to Lemma 2, the two classes of weight functions mentioned yield two distinct types of the SICM metrics. For clarity, we henceforth refer to these metrics as $\mathrm{SICM}_{L_2}(t; a)$ and $\mathrm{SICM}_K(t; a)$. Then, we have the following important properties of these two metrics.

THEOREM 1. *Let* $\{(X_i, \Delta_i, \mathbf{Z}_i)\}_{i=1}^4$ *be independent and identically distributed (i.i.d.) copies of* $(X, \Delta, \mathbf{Z})$. *Denote* $\widetilde{T}_i = (X_i, \Delta_i)$ *and*

$$\psi_t(\widetilde{T}_i, \widetilde{T}_k) = \int_0^t \frac{a(s)}{\mathbb{E}\{Y(s)\}} \Big[I(X_k \geq s) dN_i(s) - I(X_i \geq s) dN_k(s) \Big].$$

(i) *If* $d\nu(\boldsymbol{u}) = c_p^{-1} \|\boldsymbol{u}\|^{-(1+p)} d\boldsymbol{u}$ *and* $\mathbb{E}\{\|\mathbf{Z}\|^2\} < \infty$, *we have that*

$$\mathrm{SICM}_{L_2}(t; a) = -\mathrm{E}\{\|\mathbf{Z}_1 - \mathbf{Z}_2\| \psi_t(\widetilde{T}_1, \widetilde{T}_3) \psi_t(\widetilde{T}_2, \widetilde{T}_4)\}.$$

(ii) *If* ν *is a finite nonnegative Borel measure on* $\mathbb{R}^p$, *we have that*

$$\mathrm{SICM}_K(t; a) = \mathrm{E}\{K(\mathbf{Z}_1, \mathbf{Z}_2) \psi_t(\widetilde{T}_1, \widetilde{T}_3) \psi_t(\widetilde{T}_2, \widetilde{T}_4)\},$$

where $K(\cdot, \cdot)$ *is a shift-invariant kernel.*

(iii) $\mathrm{SICM}(t; a, \nu) \geq 0$. *Moreover,* $T \perp\!\!\!\perp \mathbf{Z}$ *if and only if* $\mathrm{SICM}(t; a, \nu) = 0$, *for any* $t \in [0, \tau]$.

Theorem 1 states that $\mathrm{SICM}_{L_2}(t; a)$ and $\mathrm{SICM}_K(t; a)$ have closed forms and are thus easily estimated from the data. Further remarks on the theorem are as follows:

(1) In Theorem 1(i), we only focus on a particular case of non-integrable weight functions. In fact, by employing Lévy measures (Böttcher et al., 2018), we can extend the results beyond Theorem 1(i). The Lévy measures constitute a more general class of weight functions, including $c_p^{-1} \|\mathbf{u}\|^{-(1+p)} d\mathbf{u}$.
(2) Theorem 1(ii) restricts our results to shift-invariant kernels, which is a consequence of using finite nonnegative Borel measures. However, by leveraging the conditional moment embedding approach within an RKHS, as proposed by Muandet et al. (2020), we can extend these results to ISPD kernels, which encompass shift-invariant kernels.
(3) An interesting finding is that $\mathrm{SICM}_K(t; a)$ is precisely the population version of the kernel logrank test statistic. In fact, taking $L(t, t') = a(t)a(t')$, $\mathrm{SICM}_K(t; a)$ corresponds to the population version of Ψ_n with a kernel $\mathfrak{K}((t, \mathbf{z}), (t', \mathbf{z}')) = L(t, t')K(\mathbf{z}, \mathbf{z}')$, as presented in Theorem 3.2 by Fernández et al. (2023).
(4) From a methodological perspective, our approach significantly differs from the kernel logrank test, which is derived from the supremum of weights in a weighted log-rank test across an RKHS unit ball. In contrast, we employ the CMR framework. Furthermore, within this framework, we derive $\mathrm{SICM}_{L_2}(t; a)$, a special case of negative type semimetric-based measures (Sejdinovic et al., 2013).

3. ASYMPTOTIC BEHAVIOR IN HIGH DIMENSIONS

In this section, we study the asymptotic behavior of $\mathrm{SICM}_{L_2}(t; a)$ and $\mathrm{SICM}_K(t; a)$ as $p \to \infty$. To the end, we need the following assumption to control the approximated error of $\|\mathbf{Z}_1 - \mathbf{Z}_2\|$.

Assumption 1. Suppose that

$$\mathbf{Z} = \mathbf{A}\mathbf{U} + \boldsymbol{\mu},$$

where $\mathbf{A} \in \mathbb{R}^{p \times r}$, $\mathbf{U} \in \mathbb{R}^{r \times 1}$ is a r-dimensional random vector with $\mathbb{E}\{\mathbf{U}\} = 0$ and $\mathrm{Var}\{\mathbf{U}\} = I_r$. Here, r is at least as large as p. We also assume that the entries of $\mathbf{U}$ are independent and satisfy $\max_{1 \leqslant k \leqslant r} \mathbb{E}\{U_k^4\} \leq \infty$.

Assumption 1 is common in the random matrix theory literature, and it has been used in the analysis of the spectrum of the kernel random matrix (El Karoui (2010)), and high-dimensional two-sample tests (Bai & Saranadasa (1996); Chen & Qin (2010)). It ensures that $\mathrm{Var}\{\|\mathbf{Z}_1 - \mathbf{Z}_2\|^2\} = O(p)$ and thus the following second-order Taylor expansion holds:

$$\frac{\|\mathbf{Z}_1 - \mathbf{Z}_2\|}{\tau_z} = 1 + \frac{1}{2}\Big(\frac{\|\mathbf{Z}_1 - \mathbf{Z}_2\|^2}{\tau_z^2} - 1 \Big) + O_{_P}(p^{-1}), \tag{4}$$

where $\tau_z^2 = \mathbb{E}\{\|\mathbf{Z}_1 - \mathbf{Z}_2\|^2\}$. Using (4), we can show the following results.

THEOREM 2. *Suppose that Assumption* 1 *holds. Then, as* $p \to \infty$*, we have that*

$$\mathrm{SICM}_{L_2}(t;a) = \frac{1}{\tau_z}\sum_{j=1}^{p}\mathrm{Cov}^2\Big\{Z_j, \int_0^t a(t)Y(t)d\widetilde{N}(t)\Big\} + o_{_P}(1),$$

$$\mathrm{SICM}_{K}(t;a) = \frac{1}{\tau_z\gamma_z}\Psi'\Big(\frac{\tau_z}{\gamma_z}\Big)\sum_{j=1}^{p}\mathrm{Cov}^2\Big\{Z_j, \int_0^t a(s)Y(s)d\widetilde{N}(s)\Big\} + o_{_P}(1),$$

where $\gamma_z = \mathrm{median}\,\{|\mathbf{Z}_s - \mathbf{Z}_t|, s \neq t\}$ *and* $\Psi'(\cdot)$ *is the first derivative of* $\Psi(\cdot)$*, with* $K(\mathbf{z}_1, \mathbf{z}_2) = \Psi(\|\mathbf{z}_1 - \mathbf{z}_2\|/\gamma_z)$.

Theorem 2 indicates that $\mathrm{SICM}_{L_2}(t;a)$ and $\mathrm{SICM}_{K}(t;a)$ are approximated by the sum of squared componentwise covariances up to an asymptotically constant factor. These results extend those of Zhu et al. (2020) from complete to censored data scenarios. However, the involved covariance $\mathrm{Cov}\{Z_j, \int_0^t a(s)Y(s)d\widetilde{N}(s)\}$ is not as intuitive as a regular covariance, as it does not directly measure the relationship between Z_j and T. To the best of our knowledge, such covariance has not been considered in the literature of survival analysis.

We next present some interpretation on $\mathrm{Cov}\{Z_j, \int_0^t a(s)Y(s)d\widetilde{N}(s)\} = 0$, and establish a connection between it and $T \perp\!\!\!\perp \mathbf{Z}$. Note that, for any $t \in [0, \tau]$,

$$\mathrm{Cov}\Big\{Z_j, \int_0^t a(s)Y(s)d\widetilde{N}(s)\Big\} = 0 \Longleftrightarrow \mathbb{E}\{Z_jY(t)d\widetilde{N}(t)\} = 0.$$

This, together with Lemma 1 and (3), motivates us to compare the following two relationships:

$$T \perp\!\!\!\perp \mathbf{Z}, \text{ or equivalently } \mathbb{E}\{w(\mathbf{Z},\mathbf{u})Y(t)d\widetilde{N}(t)\} = 0, \forall w(\cdot,\cdot) \in \mathcal{H}, \tag{5}$$

$$\mathrm{Cov}\Big\{\mathbf{Z}, \int_0^t a(s)Y(s)d\widetilde{N}(s)\Big\} = 0, \text{ or equivalently } \mathbb{E}\{\mathbf{Z}Y(t)d\widetilde{N}(t)\} = 0. \tag{6}$$

We can see that (5) implies (6), but not vice versa. However, since the covariance in (6) takes on a complex form, it appears challenging to find the scenarios where (6) holds while (5) is violated. Interestingly, we identify a sufficient condition that guarantees the existence of the scenarios in the following theorem.

THEOREM 3. *Let* $f_{\mathbf{Z}}(\mathbf{z})$ *be the density function of* $\mathbf{Z}$*, and* $\lambda(t|\mathbf{Z})$ *be the hazard function of* T *conditional on* $\mathbf{Z}$*. Assume that* $f_{\mathbf{Z}}(\mathbf{z})$ *and* $\lambda(t|\mathbf{z})$ *are even functions with respect to each component of* $(z_1, \ldots, z_p)$*. Then,* (6) *holds.*

In Theorem 3, the conditions on $f_{\mathbf{Z}}(\mathbf{z})$ and $\lambda(t|\mathbf{z})$ being componentwise even are easily satisfied in most commonly used survival models. For example, consider the following models:

(1) PH model: $\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp\{g(\mathbf{Z})\}$;
(2) Additive hazards model (Lin & Ying, 1994): $\lambda(t|\mathbf{Z}) = \lambda_0(t) + g(\mathbf{Z})$;
(3) Accelerated hazards model (Chen & Wang, 2000): $\lambda(t|\mathbf{Z}) = \lambda_0(t\exp(g(\mathbf{Z})))$;
(4) AFT model: $\log(T) = -g(\mathbf{Z}) + \varepsilon$, which is equivalent to $\lambda(t|\mathbf{Z}) = \lambda_0(t\exp(g(\mathbf{Z})))\exp\{g(\mathbf{Z})\}$;
(5) Transformed hazards model (Zeng et al., 2005): $\lambda(t|\mathbf{Z}) = G(\lambda_0(t) + g(\mathbf{Z}))$, where $G(\cdot)$ is a known and increasing transformation function.

In the above models, if the regression functions $g(\mathbf{Z})$ and the density function $f_{\mathbf{Z}}(\mathbf{z})$ are componentwise even, the conditions in Theorem 3 hold. This implies that when $g(\mathbf{Z}) \neq 0$, (6) holds but (5) is violated, suggesting that T and $\mathbf{Z}$ are nonlinearly dependent but uncorrelated in the sense of the relationship defined in (6). Thus, we conclude that the kernel logrank test can only capture the linear dependence mentioned in high dimensions, which will be further demonstrated by the simulation studies.

4. MARGINAL SICM-BASED TEST IN HIGH DIMENSIONS

As analyzed above, in high-dimensional cases, directly using SICM to test (1) will become a marginal covariance test. In this section, we propose a SICM-based marginal test, which is equivalent to test

$$H_0' : T \perp\!\!\!\perp Z_s, \forall s \in \{1, \ldots, p\}, \quad \text{versus} \quad H_1' : \text{otherwise}, \tag{7}$$

where Z_s is the s-th component of $\mathbf{Z}$. It is easy to see that H_0 in (1) implies H_0', but not vice versa. Although such componentwise dependence in (7) cannot fully capture the dependence between T and $\mathbf{Z}$, it remains informative in testing H_0, as shown by extensive high dimensional literature, including studies by Chen & Qin (2010); Zhong & Chen (2011) and Zhang et al. (2018). Thus, this insight motivates us to develop a novel nonparametric test based on the SICM for H_0'.

To test problem (7), we can utilize the sum of all the marginal SICM metrics between T and Z_s. Theorem 1(iii) indicates that the summation is nonnegative and equals zero for any $t \in [0, \tau]$ if and only if $T \perp\!\!\!\perp Z_s$ for all $s = 1, ..., p$. Thus, the ideal is proper. Additionally, note that $\psi_t(\widetilde{T}_i, \widetilde{T}_k)$ involves the denominator $\mathbb{E}\{Y(t)\}$. For ease of constructing unbiased test statistics, we select $a(t) = \mathbb{E}\{Y(t)\}$, which can eliminate the denominator. Meanwhile, we take $t = \tau$ for comparison with the kernel logrank test. Therefore, we ultimately propose the following metrics to test the hypothesis (7):

$$\begin{aligned}
\mathrm{SICM}_{L_2}(T|\mathbf{Z}) &= -\sum_{s=1}^{p} \mathbb{E}\{\|Z_{1s} - Z_{2s}\|\psi(\widetilde{T}_1, \widetilde{T}_3)\psi(\widetilde{T}_2, \widetilde{T}_4)\}, \\
\mathrm{SICM}_{K}(T|\mathbf{Z}) &= \sum_{s=1}^{p} \mathbb{E}\{K(Z_{1s}, Z_{2s})\psi(\widetilde{T}_1, \widetilde{T}_3)\psi(\widetilde{T}_2, \widetilde{T}_4)\},
\end{aligned}$$

where

$$\psi(\widetilde{T}_i, \widetilde{T}_k) = \Delta_i I(X_k \geq X_i) - \Delta_k I(X_i \geq X_k).$$

In the following context, we propose new test statistics based on unbiased estimators of the two metrics.

For simplicity, we only focus on methods based on $\mathrm{SICM}_K(T|\mathbf{Z})$, and methods based on $\mathrm{SICM}_{L_2}(T|\mathbf{Z})$ can be derived similarly. Let $\{(X_i, \Delta_i, \mathbf{Z}_i)\}_{i=1}^n$ be a random sample of $(X, \Delta, \mathbf{Z})$. We propose an unbiased empirical estimate of $\mathrm{SICM}_K(T|\mathbf{Z})$, given by

$$\mathcal{T}_{n,p} = \frac{1}{(n)_4} \sum_{s=1}^{p} \sum_{(i,j,k,l)}^{n} K(Z_{is}, Z_{js})\psi(\widetilde{T}_i, \widetilde{T}_k)\psi(\widetilde{T}_j, \widetilde{T}_l), \tag{8}$$

where $\sum_{(i,j,k,l)}$ denotes the summation over all possible permutations of (i, j, k, l). Throughout the article, we define $(n)_m = n(n-1)\cdots(n-m+1)$.

Note that directly calculating $\mathcal{T}_{n,p}$ using (8) is computationally intensive, with a complexity of order $O(pn^4)$. The following theorem provides a more efficient algorithm to alleviate this computational burden.

THEOREM 4. *$\mathcal{T}_{n,p}$ can be rewritten as*

$$\begin{aligned}
\mathcal{T}_{n,p} = \frac{1}{(n)_4} \sum_{s=1}^{p} \sum_{i=1}^{n} \sum_{j=1}^{n} \Big\{ & K(Z_{is}, Z_{js})[n^2 \bar{\psi}(\widetilde{T}_i)\bar{\psi}(\widetilde{T}_j) + 2n\bar{\psi}(\widetilde{T}_i)\psi(\widetilde{T}_i, \widetilde{T}_j) \\
& - n\bar{\psi}(\widetilde{T}_i, \widetilde{T}_j) - \psi^2(\widetilde{T}_i, \widetilde{T}_j)] + K(Z_{is}, Z_{is})[\psi^2(\widetilde{T}_i, \widetilde{T}_j) - n^2\bar{\psi}^2(\widetilde{T}_i)] \Big\},
\end{aligned}$$

where

$$\begin{aligned}
\bar{\psi}(\widetilde{T}_i, \widetilde{T}_j) &= \frac{1}{n} \sum_{k=1}^{n} \psi(\widetilde{T}_i, \widetilde{T}_k)\psi(\widetilde{T}_j, \widetilde{T}_k), \\
\bar{\psi}(\widetilde{T}_i) &= \frac{1}{n} \sum_{k=1}^{n} \Delta_i I(X_k \geq X_i) - \frac{1}{n} \sum_{k=1}^{n} \Delta_k I(X_i \geq X_k).
\end{aligned} \tag{9}$$

According to Theorem 4, the computation of $\mathcal{T}_{n,p}$ can be significantly optimized, requiring only $O(pn^2)$ computations.

4.1. *Asymptotic null distribution*

In the section, we establish the limiting null distribution of $\mathcal{T}_{n,p}$ in high dimensions by using the theory of martingale central limit theorem (Hall & Heyde, 2014). Before that, we define

$$H(\mathbf{W}_1, \mathbf{W}_2) = \sum_{s=1}^{p} L_1(Z_{1s}, Z_{2s}) L_2(\widetilde{T}_1, \widetilde{T}_2), \quad \mathcal{S}^2 = \mathrm{Var}\{H(\mathbf{W}_1, \mathbf{W}_2)\},$$

where $\mathbf{W}_i = (X_i, \Delta_i, \mathbf{Z}_i)$, and

$$\begin{aligned} L_1(z_{1s}, z_{2s}) &= K(z_{1s}, z_{2s}) - \mathbb{E}\{K(z_{1s}, Z_{1s})\} - \mathbb{E}\{K(z_{2s}, Z_{1s})\} + \mathbb{E}\{K(Z_{1s}, Z_{2s})\}, \\ L_2(\widetilde{t}_1, \widetilde{t}_2) &= \mathbb{E}\{\psi(\widetilde{t}_1, \widetilde{T}_1)\}\mathbb{E}\{\psi(\widetilde{t}_2, \widetilde{T}_1)\}. \end{aligned}$$

To derive the asymptotic null distribution, we need the following assumption to guarantee the conditions in martingale central limit theorem.

Assumption 2. Suppose that

$$\mathbb{E}\{G(\mathbf{W}_1, \mathbf{W}_2)^2\} = o(\mathcal{S}^4), \quad \mathbb{E}\{H^4(\mathbf{W}_1, \mathbf{W}_2)\} = o(n\mathcal{S}^4),$$

where $G(\mathbf{W}_1, \mathbf{W}_2) = E\{H(\mathbf{W}_1, \mathbf{W}_3)H(\mathbf{W}_2, \mathbf{W}_3) \mid \mathbf{W}_1, \mathbf{W}_2\}$.

Assumption 2 is closely connected to the condition (2.1) introduced by Hall (1984). This assumption is broadly applicable. We provide a further analysis of this assumption in Appendix C, specifically under certain dependency structures and Gaussian designs. Under this assumption, we can obtain the following results.

THEOREM 5. *Under H_0' and Assumption* 2*, we have that*

$$\sqrt{\frac{n(n-1)}{2}} \frac{\mathcal{T}_{n,p}}{\mathcal{S}} \xrightarrow{D} N(0,1), \text{ as } n, p \to \infty.$$

In order to formulate our testing procedure based on Theorem 5, the asymptotic variance $\mathcal{S}^2$ needs to be estimated. According to the definition of $\mathcal{S}^2$, we consider the following estimator

$$\mathcal{S}_{n,p}^2 = \{n(n-1)\}^{-1} \sum_{i \neq j}^{n} L_{2,n}^2(\widetilde{T}_i, \widetilde{T}_j) \Big[\sum_{s=1}^{p} L_{1,n}(Z_{is}, Z_{js}) \Big]^2,$$

where $L_{1,n}(Z_{is}, Z_{js})$ is the plug-in estimator of $L_1(Z_{is}, Z_{js})$, given by

$$L_{1,n}(Z_{is}, Z_{js}) K(Z_{is}, Z_{js}) - \frac{1}{n}\sum_{l=1}^{n} K(Z_{is}, Z_{ls}) - \frac{1}{n}\sum_{k=1}^{n} K(Z_{ks}, Z_{js}) + \frac{1}{n^2}\sum_{k,l=1}^{n} K(Z_{ks}, Z_{ls}),$$

and $L_{2,n}(\widetilde{T}_i, \widetilde{T}_j) = \bar{\psi}(\widetilde{T}_i)\bar{\psi}(\widetilde{T}_j)$, with $\bar{\psi}(\widetilde{T}_j)$ defined in (9). The following theorem establishes the consistency of the above estimator.

THEOREM 6. *Suppose that Assumption* 2 *holds. Then, we have that*

$$\frac{\mathcal{S}_{n,p}^2}{\mathcal{S}} \xrightarrow{P} 1, \text{ as } n, p \to \infty.$$

Theorem 6 suggests that $\mathcal{S}_{n,p}^2$ is a consistent estimator of $\mathcal{S}^2$. By Theorems 5 and 6, we can obtain the following result.

COROLLARY 1. *Under H_0' and Assumption* 2*, we have that*

$$Q_{n,p} = \sqrt{\frac{n(n-1)}{2}}\frac{\mathcal{T}_{n,p}}{\mathcal{S}_{n,p}} \xrightarrow{D} N(0,1), \text{ as } n,p \to \infty.$$

By Corollary 1, we reject the null hypothesis H_0' at significant level α if and only if $Q_{n,p} > z_\alpha$, where z_α is the $1-\alpha$ quantile of standard normal.

4.2. *Asymptotic distribution under alternatives*

In the section, we establish the asymptotic distribution under the class of local alternatives H_1', satisfying

$$\begin{aligned}&\mathrm{Var}\left\{\mathbb{E}\{G_1(\mathbf{W}_1,\mathbf{W}_2)|\mathbf{W}_1\}\right\} = o(n^{-1}\mathcal{S}^2), \quad \mathrm{Var}\left\{\mathbb{E}\{H(\mathbf{W}_1,\mathbf{W}_2)|\mathbf{W}_1\}\right\} = o(n^{-1}\mathcal{S}^2),\\ &\mathrm{Var}\left\{G_1(\mathbf{W}_1,\mathbf{W}_2)\right\} = o(\mathcal{S}^2), \quad \mathrm{Var}\left\{G_2(\mathbf{W}_1,\mathbf{W}_2)\right\} = o(\mathcal{S}^2),\end{aligned} \tag{10}$$

where $G_1(\mathbf{W}_1,\mathbf{W}_2) = \mathbb{E}\{\sum_{s=1}^p L_1(Z_{3s},Z_{4s})\psi(\widetilde{T}_3,\widetilde{T}_1)\psi(\widetilde{T}_4,\widetilde{T}_2)|\mathbf{W}_1,\mathbf{W}_2\}$, $G_2(\mathbf{W}_1,\mathbf{W}_2) = \mathbb{E}\{\sum_{s=1}^p L_1(Z_{1s},Z_{3s})\psi(\widetilde{T}_3,\widetilde{T}_4)\psi(\widetilde{T}_1,\widetilde{T}_2)|\mathbf{W}_1,\mathbf{W}_2\}$.

The conditions in (10) imply a small difference between H_0' and H_1', which can be viewed as local alternatives. Similar interpretations are discussed in Zhong & Chen (2011), Zhang et al. (2018) and Li et al. (2023). The following theorem presents the asymptotic distribution of the proposed test under the local alternatives.

THEOREM 7. *Suppose that Assumption* 2 *holds. Under the local H_1', satisfying* (10)*, we have that*

$$\sqrt{\frac{n(n-1)}{2}}\frac{\mathcal{T}_{n,p} - \mathrm{SICM}_K(T|\mathbf{Z})}{\mathcal{S}} \xrightarrow{D} N(0,1), \text{ as } n,p \to \infty.$$

By Theorem 7 and Corollary 1, the power of the proposed test statistic $Q_{n,p}$ under the local H_1' is asymptotically equal to

$$\Phi\Big(-z_\alpha + \sqrt{\frac{n(n-1)}{2}}\frac{\mathrm{SICM}_K(T|\mathbf{Z})}{\mathcal{S}}\Big),$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$, and z_α denotes the $1-\alpha$ quantile of $N(0,1)$. We can see that the power is determined by

$$\mathrm{SNR} = \sqrt{\frac{n(n-1)}{2}}\frac{\mathrm{SICM}_K(T|\mathbf{Z})}{\mathcal{S}}. \tag{11}$$

Thus, SNR is the signal-to-noise ratio of the test. Our test has non-trivial power under the alternative hypotheses, as long as SNR does not vanish to 0 as $n,p \to \infty$.

5. MONTE CARLO SIMULATIONS

In this section, we conduct simulation studies to assess the finite sample performance of the proposed SICM-based test methods. Our numerical comparison includes our proposed tests based on the metrics $\mathrm{SICM}_{L_2}(T|\mathbf{Z})$ (denoted as SICM_{L_2}) and $\mathrm{SICM}_K(T|\mathbf{Z})$ (denoted as SICM_K), the kernel log-rank test by Fernández et al. (2023) (denoted as KLR), and the inverse probability weighted survival distance covariance by Edelmann et al. (2021) (denoted as IPCW). As suggested by the authors, the critical values of the IPCW and KLR test statistics are computed using permutation and wild bootstrap methods, respectively. For the SICM_K and KLR methods, we consider the Gaussian and Laplace kernels.

Throughout the simulation experiments, the covariates $\mathbf{Z}_i = (Z_{i1},\ldots,Z_{ip})^T$ are generated from a moving average model:

$$Z_{is} = \rho_1\eta_{is} + \rho_2\eta_{i(s+1)} + \ldots + \rho_T\eta_{i(s+T-1)}, s = 1,2,\ldots,p, \tag{12}$$

where $(\eta_{i1}, \ldots, \eta_{i(p+T-1)})$ is $(p+T-1)$-dimensional standard normal vector. Here, we consider that $T = 8$ and $\{\rho_k\}_{k=1}^{T}$ are generated independently from the uniform distribution $U[0,1]$. The censored time C is generated from the exponential distribution $\mathrm{Exp}(\eta^{-1})$. We control the censoring rate to be approximately 20% by selecting η. In each experiment, we perform 1000 replications.

Example 1. In this example, we examine the normal approximation accuracy under the null hypothesis. Towards this goal, we consider the following two scenarios:

(i) $\lambda(t|\mathbf{Z}_i) = \exp(\mathbf{Z}_i^T\boldsymbol{\beta})$, with $\|\boldsymbol{\beta}\| = 0$;
(ii) $\lambda(t|\mathbf{Z}_i) = \lambda_0(t) + \mathbf{Z}_i^T\boldsymbol{\beta}$, with $\|\boldsymbol{\beta}\| = 0$ and $\lambda_0(t) = 0.1$.

Figure 1 displays the kernel density estimates for our test statistic $Q_{n,p}$ in Example 1(i) across various dimensions: $p = (20, 40, 60)$ for $n = 40$, $p = (30, 60, 90)$ for $n = 60$, and $p = (40, 80, 120)$ for $n = 80$. As observed from Figure 1, when n and p increase, the kernel density curves of $Q_{n,p}$ are close to the standard normal distribution. This indicates that the limiting distribution of $Q_{n,p}$ can be well approximated by the standard normal distribution, thus confirming our theoretical results. For the sake of brevity, the analogous results for Example 1(ii) are not presented.

Table 1 further presents the empirical sizes for our proposed tests, as well as those for the IPCW and KLR, for Example 1 at the significance levels $\alpha = 0.05$ and 0.1. The results indicate that the empirical sizes of all tests are acceptably close to the nominal significance levels.

Example 2. In this example, we assess the empirical power of our proposed tests in comparison with the KLR and IPCW methods across diverse scenarios. Our experimental configurations are based on the theoretical framework provided by Theorem 3. Moreover, the regression coefficients, denoted by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, are specified for each scenario as: $\beta_j = \|\boldsymbol{\beta}\|/\sqrt{q}$, for $j = 1, ..., q$, where $q = [\sqrt{p}/2]$, with all other coefficients being zero.

Case 1: Proportional hazards (PH) models

$$\text{(i) } \lambda(t|\mathbf{Z}_i) = \exp((\mathbf{Z}_i^4)^T\boldsymbol{\beta}_1); \qquad \text{(ii) } \lambda(t|\mathbf{Z}_i) = \exp(\log(|\mathbf{Z}_i|)^T\boldsymbol{\beta}_2),$$

with $\|\boldsymbol{\beta}_1\| = 0.04$ and $\|\boldsymbol{\beta}_2\| = 1$.

Case 2: Additive hazards models

$$\text{(i) } \lambda(t|\mathbf{Z}_i) = \lambda_0(t) + (\mathbf{Z}_i^4)^T\boldsymbol{\beta}_1; \qquad \text{(ii) } \lambda(t|\mathbf{Z}_i) = \lambda_0(t) + \log(|\mathbf{Z}_i|)^T\boldsymbol{\beta}_2,$$

with $\|\boldsymbol{\beta}_1\| = \|\boldsymbol{\beta}_2\| = 0.01$ and $\lambda_0(t) = 0.1$.

Case 3: Accelerated failure time (AFT) models

$$\text{(i) } \log(T_i) = (\mathbf{Z}_i^2)^T\boldsymbol{\beta}_1 + \epsilon_i; \qquad \text{(ii) } \log(T_i) = \log(|\mathbf{Z}_i|)^T\boldsymbol{\beta}_2 + \epsilon_i,$$

where $\|\boldsymbol{\beta}_1\| = 0.03$, $\|\boldsymbol{\beta}_2\| = 0.1$, and $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$.

Case 4: Accelerated hazards models

$$\text{(i) } \lambda(t|\mathbf{Z}_i) = \lambda_0(\cos(\mathbf{Z}_i)^T\boldsymbol{\beta}_1); \qquad \text{(ii) } \lambda(t|\mathbf{Z}_i) = \lambda_0((\mathbf{Z}_i^2)^T\boldsymbol{\beta}_2),$$

with $\|\boldsymbol{\beta}_1\| = 4.5$, $\|\boldsymbol{\beta}_2\| = 0.5$, and $\lambda_0(t) = \sqrt{t}$.

Case 5: Transformed hazards models

$$\text{(i) } G(\lambda(t|\mathbf{Z}_i)) = \lambda_0(t) + (\mathbf{Z}_i^2)^T\boldsymbol{\beta}_1; \qquad \text{(ii) } G(\lambda(t|\mathbf{Z}_i)) = \lambda_0(t) + (\log(|\mathbf{Z}_i|))^T\boldsymbol{\beta}_2,$$

where $\|\boldsymbol{\beta}_1\| = \|\boldsymbol{\beta}_2\| = 0.5$, $\lambda_0(t) = t/2$, and $G(x) = (x^s - 1)/s$ with $s = 0.5$.

The empirical powers for Example 2 are summarized in Table 2 and Tables S1-S4 in Appendix D. From these results, we can see that our proposed tests, SICM_{L_2} and SICM_K, outperform the KLR and IPCW tests across all five cases. Moreover, the KLR and IPCW tests suffer a substantial loss of power as the dimension increases, even in the settings of the PH models, which are favored by the KLR.

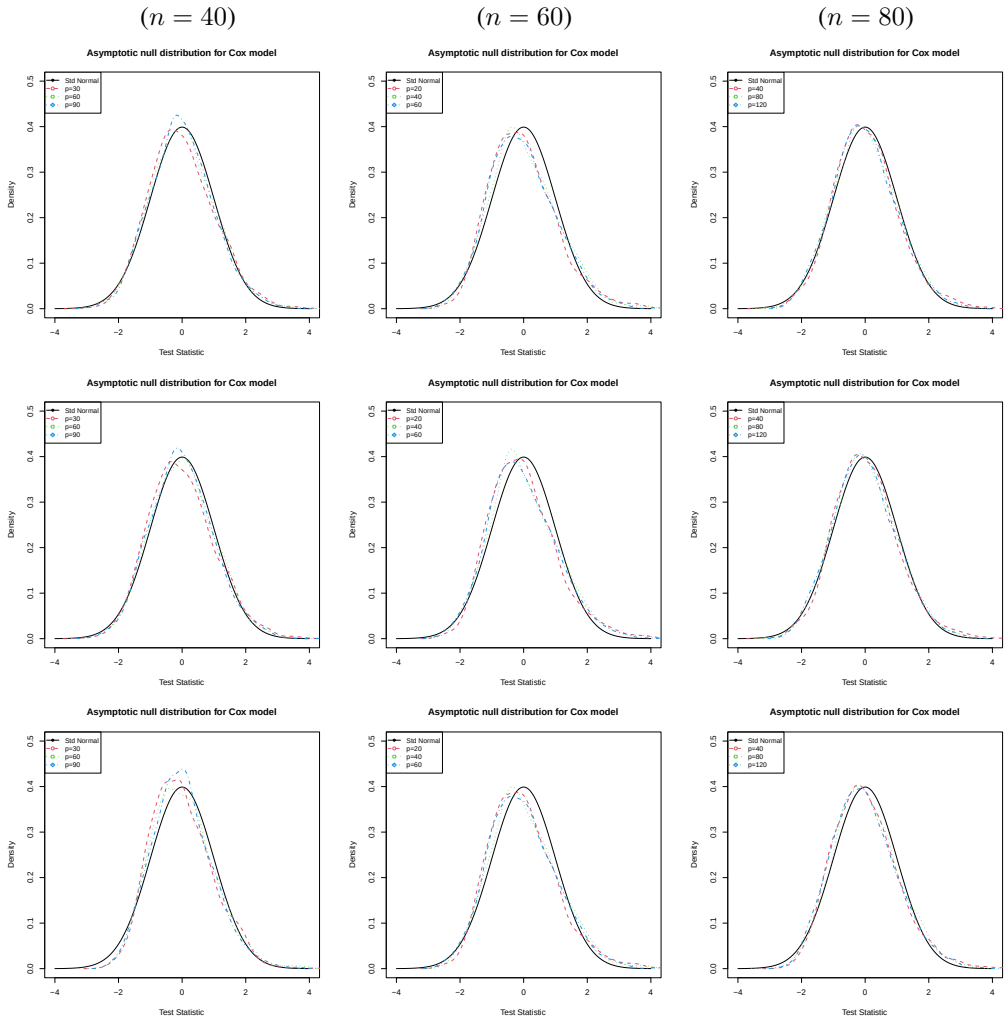

Fig. 1. Density curves of the asymptotic null distribution of $Q_{n,p}$ in Example 1(i). The first and second rows are based on SICM_K with the Gaussian and Laplace kernels, and the third is based on SICM_{L_2}.

The performance observed in Table 2 and Tables S1-S4 in Appendix D are consistent with our theoretical findings. According to Theorem 3, in all five cases, we obtain that $\mathrm{Cov}\{\mathbf{Z}_j, \int_0^t a(s)Y(s)d\widetilde{N}(s)\} = 0$, $j = 1, \ldots, p$, but T and $\mathbf{Z}$ are nonlinearly dependent. As demonstrated in Theorem 2, the KLR test statistic is approximated by the sum of the squares of the aforementioned covariances as $p \to \infty$. As expected, the KLR test fails to capture the nonlinear dependence in the above five cases, Additionally, since the IPCW is a weighted version of the distance covariance (Székely et al., 2007), as shown by Zhu et al. (2020), the IPCW also exhibits a loss of power as the dimension increases.

In the "SNR column of Table 2 and Tables S1-S4 in Appendix D, we also present the estimates of the signal-to-noise ratio as defined in (11). We observe that the empirical powers of of SICM_{L_2} and SICM_K decrease as the SNR decreases, which is consistent with Theorem 7.

Example 3. In this example, we are interested in comparing the performance of our proposed tests against the KLR and IPCW tests in scenarios where the conditions in Theorem 3 are not satisfied. We consider the following two cases:

Table 1. *The empirical sizes for the PH and additive hazards (AH) models for Example* 1 *at the significance level 5% and 10%.*

						Gaussian Kernel		Laplace Kernel	
	n	p	α	IPCW	SICM_{L^2}	KLR	SICM_K	KLR	SICM_K
(PH model)	40	20	0.050	0.030	0.056	0.040	0.058	0.042	0.060
		40	0.050	0.046	0.054	0.038	0.052	0.044	0.052
		80	0.050	0.042	0.052	0.056	0.058	0.054	0.054
		20	0.100	0.084	0.088	0.094	0.108	0.092	0.106
		40	0.100	0.086	0.108	0.098	0.101	0.096	0.102
		80	0.100	0.090	0.105	0.106	0.100	0.106	0.090
	60	30	0.050	0.045	0.052	0.044	0.054	0.044	0.054
		60	0.050	0.046	0.058	0.042	0.056	0.044	0.058
		90	0.050	0.054	0.048	0.050	0.052	0.046	0.046
		30	0.100	0.082	0.096	0.088	0.098	0.098	0.104
		60	0.100	0.110	0.103	0.092	0.106	0.094	0.108
		90	0.100	0.088	0.094	0.094	0.094	0.098	0.094
	80	40	0.050	0.040	0.052	0.054	0.054	0.054	0.054
		80	0.050	0.046	0.058	0.042	0.056	0.044	0.046
		120	0.050	0.054	0.046	0.040	0.052	0.056	0.046
		40	0.100	0.082	0.096	0.098	0.108	0.097	0.104
		80	0.100	0.101	0.103	0.092	0.106	0.094	0.108
		120	0.100	0.098	0.102	0.094	0.094	0.098	0.094
(AH model)	40	20	0.050	0.043	0.046	0.048	0.044	0.048	0.047
		40	0.050	0.044	0.054	0.053	0.056	0.044	0.044
		80	0.050	0.046	0.048	0.046	0.054	0.046	0.048
		20	0.100	0.098	0.106	0.092	0.106	0.090	0.106
		40	0.100	0.097	0.102	0.094	0.106	0.107	0.098
		80	0.100	0.114	0.105	0.104	0.103	0.102	0.104
	60	30	0.050	0.046	0.057	0.056	0.056	0.044	0.056
		60	0.050	0.040	0.056	0.060	0.060	0.048	0.058
		90	0.050	0.054	0.048	0.066	0.046	0.042	0.056
		30	0.100	0.100	0.106	0.100	0.106	0.094	0.104
		60	0.100	0.098	0.102	0.090	0.102	0.096	0.096
		90	0.100	0.106	0.096	0.097	0.108	0.098	0.108
	80	40	0.050	0.040	0.054	0.048	0.066	0.040	0.047
		80	0.050	0.046	0.048	0.060	0.058	0.064	0.048
		120	0.050	0.048	0.058	0.048	0.047	0.038	0.052
		40	0.100	0.098	0.102	0.084	0.106	0.078	0.112
		80	0.100	0.104	0.105	0.102	0.103	0.120	0.104
		120	0.100	0.098	0.104	0.104	0.106	0.090	0.106

(i) $\log(T_i) = \mathbf{Z}_i^T\boldsymbol{\beta} + \epsilon_i$, where $\|\boldsymbol{\beta}\| = 0.03$ and $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$;
(ii) $\lambda(t|\mathbf{Z}_i) = \exp(\mathbf{Z}_i^T\boldsymbol{\beta})$, where $\|\boldsymbol{\beta}\| = 0.04$.

Table 3 summarizes the empirical powers of all tests for Example 3. As illustrated in Table 3, for the AFT model, SICM_{L^2} exhibits superior performance, followed by the IPCW. For the PH model, the KLR demonstrates superior performance, which is unsurprising because of its inheritance of the effectiveness of the classical logrank test in the PH model setting. However, even in this case, SICM_{L^2} is comparable to the KLR. In summary, our proposed methods display a high capability to capture the dependence between T and $\mathbf{Z}$ in scenarios where the conditions in Theorem 3 are not satisfied. Additionally, by comparing

Table 2. *The empirical powers for the PH models in Case 1 for Example* 2 *at the significance level 5% and 10%.*

							Gaussian Kernel		Laplace Kernel	
	n	p	α	SNR	IPCW	SICM_{L^2}	KLR	SICM_K	KLR	SICM_K
(i)	40	20	0.050	2.154	0.040	0.255	0.365	0.605	0.390	0.625
		40	0.050	1.641	0.045	0.200	0.160	0.470	0.170	0.490
		80	0.050	1.225	0.040	0.145	0.055	0.345	0.055	0.345
		20	0.100	2.154	0.085	0.340	0.555	0.750	0.580	0.760
		40	0.100	1.641	0.090	0.305	0.265	0.590	0.315	0.600
		80	0.100	1.225	0.085	0.210	0.135	0.465	0.145	0.485
	60	30	0.050	2.557	0.050	0.315	0.340	0.760	0.325	0.780
		60	0.050	2.031	0.065	0.305	0.200	0.595	0.230	0.610
		90	0.050	1.823	0.045	0.195	0.100	0.530	0.115	0.530
		30	0.100	2.557	0.090	0.490	0.520	0.860	0.520	0.845
		60	0.100	2.031	0.115	0.385	0.330	0.700	0.335	0.695
		90	0.100	1.823	0.085	0.295	0.190	0.650	0.205	0.665
	80	40	0.050	3.415	0.055	0.445	0.375	0.940	0.410	0.940
		80	0.050	2.767	0.055	0.385	0.130	0.825	0.150	0.840
		120	0.050	2.148	0.045	0.250	0.120	0.695	0.125	0.700
		40	0.100	3.415	0.080	0.590	0.580	0.970	0.595	0.975
		80	0.100	2.767	0.115	0.480	0.265	0.895	0.270	0.910
		120	0.100	2.148	0.095	0.375	0.230	0.765	0.235	0.775
(ii)	40	20	0.050	1.477	0.050	0.190	0.160	0.445	0.155	0.450
		40	0.050	1.181	0.060	0.175	0.120	0.345	0.130	0.330
		80	0.050	0.905	0.055	0.115	0.110	0.235	0.125	0.265
		20	0.100	1.477	0.105	0.235	0.340	0.530	0.325	0.520
		40	0.100	1.181	0.120	0.235	0.210	0.435	0.220	0.425
		80	0.100	0.905	0.110	0.175	0.225	0.360	0.220	0.385
	60	30	0.050	1.752	0.075	0.210	0.200	0.530	0.210	0.565
		60	0.050	1.367	0.035	0.185	0.100	0.415	0.110	0.395
		90	0.050	1.325	0.030	0.120	0.090	0.355	0.095	0.375
		30	0.100	1.752	0.130	0.305	0.350	0.650	0.360	0.640
		60	0.100	1.367	0.080	0.245	0.205	0.475	0.210	0.495
		90	0.100	1.325	0.065	0.210	0.170	0.510	0.170	0.505
	80	40	0.050	2.430	0.050	0.285	0.240	0.720	0.255	0.690
		80	0.050	2.079	0.065	0.250	0.125	0.655	0.120	0.640
		120	0.050	1.663	0.040	0.190	0.105	0.495	0.110	0.500
		40	0.100	2.430	0.120	0.390	0.405	0.830	0.425	0.850
		80	0.100	2.079	0.110	0.355	0.255	0.765	0.260	0.765
		120	0.100	1.663	0.080	0.265	0.190	0.630	0.190	0.605

Table 3 with Table 2 and Tables S1-S4 in Appendix D, we observe that the L_2-type metrics, SICM_{L^2} and IPCW, perform better under linear dependence as in Example 3 than under the nonlinear dependence as in Example 2. This finding is consistent with research on uncensored data.

Table 3. *The empirical powers for the AFT and PH models in Example* 3 *at the significance level 5% and 10%.*

							Gaussian Kernel		Laplace Kernel	
	n	p	α	SNR	IPCW	SICM_{L^2}	KLR	SICM_K	KLR	SICM_K
(i)	40	20	0.050	1.142	0.260	0.410	0.290	0.335	0.290	0.315
		40	0.050	1.016	0.255	0.380	0.245	0.275	0.245	0.280
		80	0.050	1.098	0.260	0.430	0.320	0.345	0.305	0.335
		20	0.100	1.142	0.435	0.545	0.455	0.425	0.460	0.410
		40	0.100	1.016	0.375	0.460	0.375	0.365	0.390	0.385
		80	0.100	1.098	0.430	0.540	0.445	0.400	0.430	0.420
	60	30	0.050	1.419	0.375	0.475	0.395	0.410	0.390	0.400
		60	0.050	1.533	0.435	0.580	0.430	0.425	0.420	0.415
		90	0.050	1.667	0.420	0.595	0.475	0.495	0.465	0.490
		30	0.100	1.419	0.480	0.575	0.510	0.505	0.510	0.505
		60	0.100	1.533	0.600	0.675	0.595	0.535	0.575	0.535
		90	0.100	1.667	0.575	0.655	0.565	0.585	0.560	0.575
	80	40	0.050	1.932	0.510	0.660	0.505	0.530	0.505	0.525
		80	0.050	2.054	0.595	0.695	0.530	0.590	0.525	0.595
		120	0.050	2.158	0.610	0.740	0.660	0.610	0.665	0.600
		40	0.100	1.932	0.680	0.755	0.615	0.640	0.620	0.635
		80	0.100	2.054	0.720	0.770	0.660	0.670	0.665	0.675
		120	0.100	2.158	0.785	0.820	0.790	0.705	0.780	0.720
(ii)	40	20	0.050	0.619	0.180	0.265	0.265	0.190	0.255	0.190
		40	0.050	0.712	0.160	0.295	0.280	0.260	0.280	0.250
		80	0.050	0.609	0.150	0.245	0.225	0.205	0.230	0.195
		20	0.100	0.619	0.305	0.345	0.385	0.280	0.385	0.300
		40	0.100	0.712	0.290	0.375	0.410	0.295	0.400	0.280
		80	0.100	0.609	0.275	0.325	0.360	0.265	0.355	0.260
	60	30	0.050	0.988	0.225	0.360	0.315	0.290	0.295	0.295
		60	0.050	1.059	0.255	0.345	0.355	0.290	0.360	0.270
		90	0.050	1.119	0.250	0.405	0.370	0.355	0.385	0.360
		30	0.100	0.988	0.385	0.435	0.445	0.375	0.465	0.385
		60	0.100	1.059	0.395	0.455	0.540	0.435	0.525	0.430
		90	0.100	1.119	0.405	0.490	0.545	0.450	0.540	0.440
	80	40	0.050	1.617	0.365	0.560	0.555	0.425	0.545	0.420
		80	0.050	1.585	0.365	0.595	0.610	0.450	0.600	0.465
		120	0.050	1.416	0.400	0.555	0.575	0.430	0.575	0.425
		40	0.100	1.617	0.565	0.640	0.720	0.555	0.735	0.530
		80	0.100	1.585	0.560	0.690	0.740	0.595	0.740	0.585
		120	0.100	1.416	0.600	0.660	0.725	0.520	0.720	0.550

6. REAL DATA ANALYSIS

6.1. *Association test for gene pathways*

We applied the proposed test to assess the association of pathways with survival time applied to 117 advanced stage serous ovarian cancers (ASSOC), with a 42.73% censoring rate, provided by Dressman et al. (2007), is available from the R package "curatedOvarianData".

The gene expression values were obtained for 13,104 genes. Pathway information was obtained from the Gene Ontology (GO) database, using the BioConductor GO package Carlson et al. (2019). Pathways that were considered to be of specific interest were cell cycle (GO:0007049), DNA repair (GO:0006281), angiogenesis (GO:0001525), blood vessel development (GO:0001568) and apoptosis (GO:0006915).

Table 4. *$P-$values of the GO terms for six tests, and their number of genes.*

				Gaussian Kernel		Laplace Kernel	
Pathway	Genes	IPCW	SICM_{L^2}	KLR	SICM_K	KLR	SICM_K
All genes	13104	**	***	***	***	***	***
Cell cycle	1292	**	***	***	***	***	***
DNA repair	416	**	***	***	***	***	***
Angiogenesis	437	**	***	***	***	***	***
Blood vessel development	434	**	***	***	***	***	***
Apoptosis	1479	*	***	***	***	***	***

NOTE: *$P-$value<0.1, **$P-$value<0.05, ***$P-$value<0.01.

The results of all genes and five pathways of primary interest are given in Table 4. In this dataset the expression profile over the set of all genes on the chip is significantly associated with survival. And we can conclude that the interested biological pathways, including those involved in the cell cycle, DNA repair, angiogenesis, blood vessel development, and apoptosis, are clearly associated with survival. These pathways provide crucial insights into how gene expression correlates with patient outcomes, highlighting the broader predictive value of analyzing gene expression profiles in understanding survival.

6.2. *The comparative P value*

Note that small $P-$value does not mean that every gene on the chip is associated with survival. The six methods only test the null hypothesis that the whole gene set or pathway is not associated with survival. This null hypothesis depends only on the observed survival and genes of the pathway itself: the results are absolute, not relative to other pathways. Next, we introduce comparative $P-$value, denoted as $\bar{p}$, proposed by Goeman et al. (2005). The comparative $P-$value fulfills a role different from the $P-$value and should only be used alongside it. It indicates whether the $P-$ value of a group of genes is much lower than the $P-$ value of the genome of the same size in the dataset. The algorithm of $\bar{p}$ can be summarised as follow

Algorithm 1. The comparative $P-$value algorithm.

Require:

Let $t = 1$ and B be the resample times;

The whole gene expression dataset, $(X_i, \Delta_i, \mathbf{Z}_i)_{i=1}^n$ where n is sample size and $\mathbf{Z}$ include all 13,104 genes;

The genes set of GO term, $\mathbf{Z}^{GO} = (Z_1, ..., Z_m)^T$ where m is the number of genes in this GO term;

Ensure:

Comparative $P-$value, $\bar{p}$;

Calculate the $P-$value for GO terms by using $(X_i, \Delta_i, \mathbf{Z}_i^{GO})_{i=1}^n$, denote it as P_{GO};

While $t <= B$, do:

(a) Randomly resample m genes from all 13,104 genes, denote as $\mathbf{Z}^R$;

(b) Calculate the $P-$value by using $(X_i, \Delta_i, \mathbf{Z}_i^R)_{i=1}^n$, denote it as P_t;

(c) t=t+1.

Calculate $\bar{p} = \frac{\sum_{t=1}^B I(P_t <= P_{GO})}{B}$.

return $\bar{p}$;

In table 5, we can see that the association between DNA repair and survival is more relevant than for gene sets of the same size in this dataset. Additionally, other pathways are clearly associated with survival, as can be seen from the $P-$values in Table 4. However, the comparative $P-$ values of the six test methods are not all small. For example, the comparative $P-$value of Cell cycle by KLR with Gaussian kernel is 0.885 which means that more than 88.5% of random gene sets have a lower $P-$value than Cell cycle.

Table 5. *comparative $P-$value of the GO terms for six tests based on $B = 10000$ random gene sets.*

			Gaussian Kernel		Laplace Kernel	
Pathway	IPCW	SICM_{L^2}	KLR	SICM_K	KLR	SICM_K
All genes	-	-	-	-	-	-
Cell cycle	<0.001	0.114	0.885	0.121	<0.001	0.114
DNA repair	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Angiogenesis	<0.001	0.775	<0.001	0.636	<0.001	0.755
Blood vessel development	<0.001	0.901	<0.001	0.821	0.808	0.901
Apoptosis	<0.001	0.126	<0.001	0.137	<0.001	0.126

6.3. *Gene rank in pathway*

To further explore the impact of each gene on the DNA repair, we calculate standardized statistics for each component of $\mathbf{Z}^{DNA}$. And the original proposed test statisitc can be rewritten as

$$\mathcal{T}_{n,p} = \sum_{s=1}^{416} \mathcal{T}_{n,s},$$

where $\mathcal{T}_{n,s}$ is the test statistic of s-th component of $\mathbf{Z}^{DNA}$. Similar to variable scanning, each gene in this pathway has a rank. And we only show the top four genes, namely POLA1 (Polymerase (DNA Directed) α-1), TDP2 (Tyrosyl-DNA Phosphodiesterase 2), SMG1 (Suppressor with Morphological effect on Genitalia 1), and SMC1A (Structural Maintenance of Chromosomes, SMC).

To visualize the four genes selected that were significantly associated with survival. We artificially classified genes with expression above the average into the "high" group and the others into the "low" group. The effects of high or low expression levels of these genes on survival time were comfirmed using KaplanCMeier survival analysis and compared statistically using the log-rank test, with representative genes shown in Figure 2. We can see that the survival curves of individuals grouped by these four genes are significantly different, and the logrank test is also significant, which is consistent with our previous conclusion.

7. DISCUSSION

In this article, we develop a new CMR-based framework and introduce a novel class of SICM metrics to test independence between survival time and covariates. Utilizing this framework, we demonstrate that the kernel logrank test in high dimensions can only detect a specific type of linear dependence in censored data. To address this limitation, we propose a marginal statistic based on the proposed SICM metrics. The usefulness of the proposed methods has been validated through simulation studies and empirical analysis of two real datasets.

We note that our marginal test procedure is constructed using a sum-of-squares-based statistic, which is particularly useful for dense signals. For sparse signals, we can draw upon the ideas presented by Chen et al. (2019) to enhance the power of our tests. These will be interesting topics for future research.

It is crucial to select an appropriate weight function $a(\cdot)$ for $\mathrm{SICM}(t; a, \nu)$ as defined in Definition 1. In Section 4, we choose $a(t) = \mathbb{E}\{Y(t)\}$ to eliminate the denominator of $\psi_t(\widetilde{T}_i, \widetilde{T}_k)$. This choice facilitates the construction of an unbiased test statistic, and its asymptotic distributions under the null and local alternative hypotheses, as presented in Theorems 5 and 7. Nonetheless, identifying the optimal $a(\cdot)$ that maximizes the statistical power remains a challenging task for future research.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes proofs of Theorems **??**-**??** and Lemma **??** as well as additional simulation results and empirical analyses of two real datasets.

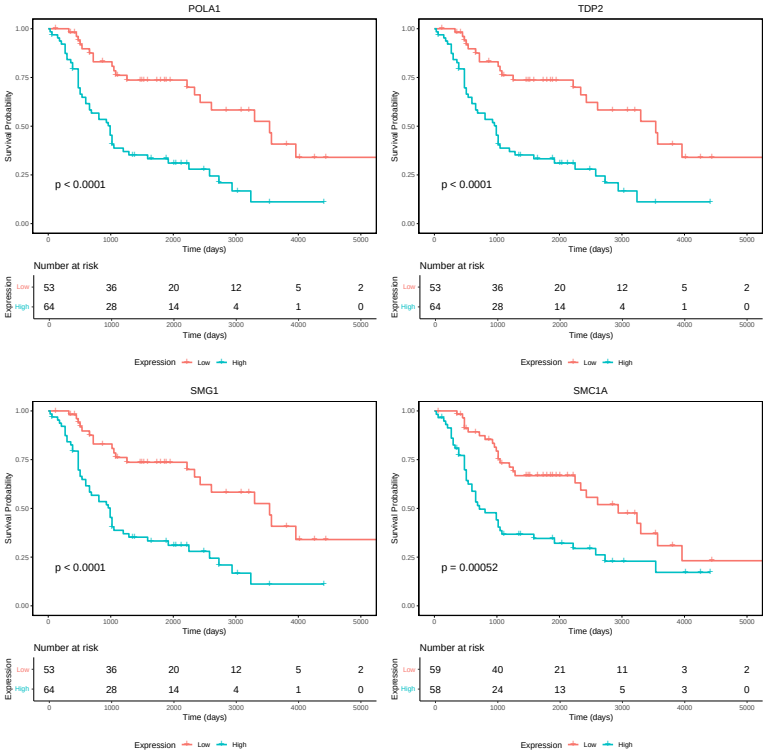

Fig. 2. KaplanCMeier survival curves for four individual genes signiffcantly associated with survival in ASSOC patients.

REFERENCES

BAI, Z. D. & SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica* **6**, 311–329.

BIERENS, H. & PLOBERGER, W. (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica* **65**, 1129–1152.

BIERENS, H. J. (1982). Consistent model specification tests. *Journal of Econometrics* **20**, 105–134.

BIERENS, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica* **58**, 1443–1458.

BÖTTCHER, B., KELLER-RESSEL, M. & SCHILLING, R. L. (2018). Detecting independence of random vectors: generalized distance covariance and gaussian covariance. *Modern Stochastics: Theory and Applications* **5**, 353–383.

CARLSON, M., FALCON, S., PAGES, H. & LI, N. (2019). Go. db: A set of annotation maps describing the entire gene ontology. *R package version 3* .

CHEN, S., LI, J. & ZHONG, P. (2019). Two-sample and ANOVA tests for high dimensional means. *The Annals of Statistics* **47**, 1443–1474.

CHEN, S. X. & QIN, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.

CHEN, Y. & WANG, M. (2000). Analysis of accelerated hazards models. *Journal of the American Statistical Association* **95**, 608–618.

COX, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **34**, 187–202.

DRESSMAN, H. K., BERCHUCK, A., CHAN, G., ZHAI, J., BILD, A., SAYER, R., CRAGUN, J., CLARKE, J., WHITAKER, R. S., LI, L. et al. (2007). An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *Journal of clinical oncology* **25**, 517–525.

EDELMANN, D., WELCHOWSKI, T. & BENNER, A. (2021). A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing. *Biometrics* **78**, 867–879.

EL KAROUI, N. (2010). The spectrum of kernel random matrices. *The Annals of Statistics* .

ESCANCIANO, J. C. (2006a). A consistent diagnostic test for regression models using projections. *Econometric Theory* **22**, 1030–1051.

ESCANCIANO, J. C. (2006b). Goodness-of-fit tests for linear and nonlinear time series models. *Journal of the American Statistical Association* **101**, 531–541.

FERNÁNDEZ, T., GRETTON, A., RINDT, D. & SEJDINOVIC, D. (2023). A kernel log-rank test of independence for right-censored data. *Journal of the American Statistical Association* **118**, 925–936.

FLEMING, T. R. & HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons.

GAO, L., FAN, Y., LV, J. & SHAO, Q. (2021). Asymptotic distributions of high-dimensional distance correlation inference. *The Annals of Statistics* **49**, 1999–2020.

GOEMAN, J. J., OOSTING, J., CLETON-JANSEN, A.-M., ANNINGA, J. K. & VAN HOUWELINGEN, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics* **21**, 1950–1957.

HALL, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis* **14**, 1–16.

HALL, P. & HEYDE, C. C. (2014). *Martingale Limit Theory and Its Application*. Academic Press.

HONG, Y. & WHITE, H. L. (1995). Consistent specification testing via nonparametric series regression. *Econometrica* **63**, 1133–1159.

KALBFLEISCH, J. D. & PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

LI, R., XU, K., ZHOU, Y. & ZHU, L. (2023). Testing the effects of high-dimensional covariates via aggregating cumulative covariances. *Journal of the American Statistical Association* **118**, 2184–2194.

LIN, D. & YING, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163–170.

MUANDET, K., JITKRITTUM, W. & KÜBLER, J. (2020). Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, vol. 124.

NEWEY, W. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica* **53**, 1047–1070.

RINDT, D., SEJDINOVIC, D. & STEINSALTZ, D. (2021). A kernel-and optimal transport-based test of independence between covariates and right-censored lifetimes. *The International Journal of Biostatistics* **17**, 331–348.

SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. & FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* **41**, 2263 – 2291.

SHAO, X. & ZHANG, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association* **109**, 1302–1318.

SRIPERUMBUDUR, B. K., GRETTON, A., FUKUMIZU, K., SCHÖLKOPF, B. & LANCKRIET, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research* **11**, 1517–1561.

STINCHCOMBE, M. B. & WHITE, H. (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* **14**, 295–325.

STUTE, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics* **25**, 613–641.

SZÉKELY, G. J. & RIZZO, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis* **117**, 193–213.

SZÉKELY, G. J., RIZZO, M. L. & BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

WENDLAND, H. (2004). *Scattered data approximation*, vol. 17. Cambridge university press.

ZENG, D., YIN, G. & IBRAHIM, J. G. (2005). Inference for a class of transformed hazards models. *Journal of the American Statistical Association* **100**, 1000–1008.

ZHANG, X., YAO, S. & SHAO, X. (2018). Conditional mean and quantile dependence testing in high dimension. *The Annals of Statistics* **46**, 219–246.

ZHONG, P. & CHEN, S. (2011). Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association* **106**, 260–274.

ZHU, C., ZHANG, X., YAO, S. & SHAO, X. (2020). Distance-based and RKHS-based dependence metrics in high dimension. *The Annals of Statistics* **48**, 3366 – 3394.