Essays on Specification Tests for Conditional Hazard and Distribution Models

by

Rui Cui

A dissertation submitted by in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in

Economics

Universidad Carlos III de Madrid

Advisor:

Miguel A. Delgado

May 2019

To my maternal grandparents,

who have given me so much love.

I used to take it for granted,

but now realize how much it means to me.

ACKNOWLEDGEMENTS

There is no doubt that the first thank you goes to my advisor, Prof. Miguel A. Delgado, without whose continuous supervision and guidance the thesis would not have been finished. I could not be more thankful for so much time he has spent on me and advice he has offered me. I could not imagine how my research would end up without him. His unconditional support and genuine care always move me and make me wonder how I am so lucky to have a supervisor like him.

Another person whose appearance has strengthened the feeling of being so lucky is Prof. Winfried Stute. No words can express my gratitude to him. His kindness, academic passion and desire for the truth have and would continue to inspire and support me. His effect on me is perfectly summarized by one interviewer when I was in a job market interview. "It changed your life," the interviewer commented after I had explained how my research begins from his lecture and how it is progressed during the visits to him. I could not agree more with the interviewer's comment.

My special thanks go to Prof. Juan Carlos Escanciano and Prof. Carlos Velasco. They have given me many valuable comments and suggestions for my research. I have learnt a lot from them, not only knowledge but also attitude towards the academic life. They are always my role models. I would also like to thank Prof. Jesús Gonzalo and Prof. Nazarii Salish for their constant advice and support.

Time goes linearly, at least for us who do not have a TARDIS. I want to thank everyone who has been here with me over the last six years in Spain. To my colleague Yuhao, who as also my flatmate probably is the one I have spent the most time with, and to my other econometrics colleague Junji. Thanks to Yunrong, Xiaojun and Mian, with whom I have spent my happiest year in Spain. I would remember our trips and those stupid jokes. Thanks to Meng and Weifeng for the companionship

in the following years. I would miss these days we have dinner and chat. Thanks to Minghai and Julius in the econometrics group. Thanks to my officemate Bea and Ismael, to Cris and all the other colleagues during these years. I would miss you a lot.

Finally, I would like to express my love to my parents, my grandparents and my other family members, the most liberal family. Especially my mother, as an excellent teacher, however, has never tutored me a single lesson and never taught me by command or blame. Her education has successfully kept my curiosity, which is essential and leads me all the way here. Usually, girls are growing more and more like their mother, whether they want to admit it or not. It is such a wonderful thing to expect me to become her.

Contents

0 Introduction **2**

1 Model Checks for Marginal Effects in Proportional Hazard Models **13**

1.1 Introduction 14

1.2 The Cox Hazard Regression Model 18

1.2.1 The Counting Process Approach 19

1.2.2 Estimation Approaches of the Cox Model 20

1.2.3 Test the Covariate Effect in the Cox Model 21

1.3 Principal Component Decomposition 23

1.4 Discrete Approximation of the Covariance Kernel 32

1.5 Numerical Approximation 36

1.6 Simulation 38

1.7 Concluding Remarks 41

1.8 Appendix 41

2 Goodness-of-Fit Tests for the Cox Proportional Hazard Model **56**

2.1 Introduction 57

2.2 Omnibus Test for the Cox Proportional Hazard Model 58

2.2.1 The Cox Model 58

2.2.2 Other Important Models in Duration Analysis 61

2.2.3 Omnibus Test 62

2.3 Tests based on Component Processes 63

2.3.1 Conditional Principal Component Decomposition 63
2.3.2 Asymptotic Theory of Component Processes 66
2.3.3 Test Statistics 69
2.3.4 Other Weight Functions 72
2.4 Simulation Study 74
2.5 Conclusion 79
2.6 Appendix: Proofs 79

3 Goodness-of-Fit Tests for Conditional Distributions 82
3.1 Introduction 83
3.2 Omnibus Test for Conditional Distributions 84
3.3 Tests based on Component Processes 86
3.3.1 Conditional Principal Component Analysis 86
3.3.2 Asymptotic Theory of Component Processes 89
3.3.3 Test Statistics 91
3.3.4 Other Weight Functions 93
3.4 Simulation Study 95
3.5 Conclusion 99
3.6 Appendix: Proofs 99

References 102

Introduction

In this thesis, we discuss the application of principal component analysis (PCA) in specification testing for statistic models. Model check is essential because inferences from an incorrectly specified model can be very misleading. However, when we apply a model checking technique, say using a test statistic, only knowing whether the null model should be rejected provides little information. We also want to know when the model fails, which particular aspects of the data are responsible for such rejection. This raises the question of whether the information in the test can be partitioned into some pieces, each of which measures some distinctive aspect of the data. If possible, we will also study the significance of each piece. This will be much more informative and give us a detailed picture of the nature of the deviation and may suggest some sort of natural alternative. Such a partition can be obtained through principal component decomposition (PCD), the orthogonal decomposition of the test statistic and the pieces are called its principal components (PCs). These PCs play an important role in testing problem—they serve as "special experts" when detecting certain deviations and in many cases one may expect that the main source of deviations only come from the first few.

This thesis provides two PCD methods aimed at improving the efficiency of specification tests for conditional hazard and distribution models. The two methods are both applicable to a general class of models, e.g., the transformation models, however, we demonstrate them in the hazard and distribution regression models. See a summary in Table 1, each row of which will be detailed discussed later in this intro-

Table 1: Application of PCA in Goodness-of-Fit

H_0	Distribution Function	Hazard Function	Mean Regression Function	Proportional Hazard Regression Function
simple	Durbin & Knott (1972)			
composite	Durbin, Knott & Taylor (1975)	Anh & Stute (2012)	Stute (1997)	Chapter 1
conditional	Chapter 3	Chapter 2		

duction. The first PCD method deals with testing for a composite hypothesis when the null hypothesis depends on unknown parameters and it is introduced in chapter 1 with a particular interest in testing the parametric part of the Cox Proportional Hazard Model. While the second one, which we call conditional PCD method, is applicable to the goodness-of-fit problem of conditional models, for which we consider the conditional hazard model in chapter 2 and conditional distribution model in chapter 3.

Simple Hypothesis

The first application of PCA in goodness-of-fit testing is by Durbin and Knott (1972), where they derived the decomposition of the Cramér-von Mises statistic in the context of distribution function specification testing. Suppose we observe a random sample $Y_1, Y_2, \cdots, Y_n$ of variable Y and we want to test whether it follows from a distribution with specified distribution function $F_0(y)$, i.e., the null hypothesis is

$$H_0 : F(y) = F_0(y).$$

The alternative hypothesis will always be the negation of H_0 unless particular H_1 is given in this thesis. Let $U_i = F_0(Y_i)$, the problem is equivalent to testing whether U_i's come from the uniform distribution. The typical Cramér-von Mises test is based

on the uniform empirical process

$$\alpha_n(u) = n^{1/2}(F_n(u) - u),$$

where $F_n(u) = n^{-1}\sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq u\}}$, and is constructed as the functional of the empirical process

$$W_n^2 = \int_0^1 \alpha_n^2(u)du.$$

This is an omnibus test in the sense that it is consistent against all the possible alternatives. Durbin and Knott (1972) have derived the orthogonal decomposition of W_n^2 as

$$W_n^2 = \sum_{j=1}^{\infty} \frac{z_{nj}^2}{j^2\pi^2}, \tag{0.1}$$

where

$$z_{nj} = (2n)^{1/2}\pi j \int_0^1 (F_n(u) - u)\sin(j\pi u)du, \ j = 1, 2, \cdots \tag{0.2}$$

are the PCs of the process $\alpha_n(u)$ and they are uncorrelated with each other. These PCs are obtained by noticing that $\alpha_n(u)$ has the same covariance kernel with the standard Brownian Bridge and hence has eigenvalues $1/(\pi j)^2$ and eigenfunctions $\sqrt{2}sin(j\pi t), j = 1, 2, \cdots$.

The PCs play an important role in testing goodness-of-fit in several ways. First, the asymptotic distribution of each PC is standard normal and they are asymptotically independent with each other, which provides the possibility for distribution-free tests. Second, it is easy to see from (0.2) that the PCs are actually coefficients obtained in a Fourier series expansion of the empirical process. Thus, we may expect that, when j increases, the latter PCs are sensitive to higher-frequency deviations. Finally, from (0.1), we observe that the contribution of the PCs to W_n^2 decreases rapidly with j. Since W_n^2 highly down-weights the latter PC, it will have low power when testing against high-frequency alternatives. Based on these facts, Durbin and Knott (1972) suggested an examination of individual PC to analyze the departure of the observations from the hypothesis. For example, they have shown in practice that the first PC in the normal distribution case is sensitive to mean shift, while the

second PC is sensitive to variance shift, and similar patterns of the third and fourth PC in skewness shift and kurtosis shift. As already mentioned, each PC serves a purpose. It suggests a special design of the tests based on different PCs according to the alternatives of interest.

In addition to studying the PCs individually, there are another two ways to improve the efficiency of the omnibus test. On one hand, given a particular direction of alternative, it is possible to design an optimal directional test based on PCs. On the other hand, it is also possible to construct tests by combining a few of them, and this gives us tests similar to Neyman's smooth tests, which outperform the omnibus test over a large range of alternatives. See for example Eubank and LaRiccia (1992), Ledwina (1994), Janssen (2000) and Escanciano (2009) for power discussion.

Similarly, PCA can also be applied in goodness-of-fit testing of the hazard function. In duration analysis, the hazard function is often modeled directly since it is more informative and completely characterizes the underlying distribution. The hazard function describes the risk of an event happening as a function of time, conditional on not having happened before. Hence it is a natural candidate to describe the dynamics of time-dependent phenomena. To be explicit, we have a random sample $T_1, T_2, \cdots, T_n$ of a nonnegative continuous duration variable T. Given a specified hazard function $\lambda_0(t)$, we want to test that

$$H_0 : \lambda(t) = \lambda_0(t).$$

The counting process approach defines the counting process and the at-risk process as

$$N_i(t) = \mathbb{1}_{\{T_i \le t\}}, \quad Y_i(t) = \mathbb{1}_{\{T_i \ge t\}}.$$

To test H_0, a common strategy is to compare the observed counting process with its expected value. The corresponding distance for each individual is captured in a martingale process that follows from the Doob-Meyer decomposition

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda_0(s)ds.$$

The Cramér-von Mises test is based on the sum of individual martingales

$$\beta_n(t) = n^{-1/2} \sum_{i=1}^{n} M_i(t),$$

and is constructed as

$$W_n^2 = \int_0^\infty \beta_n^2(t) H(dt),$$

where

$$H(t) = \int_0^t exp\left(-\int_0^s \lambda_0(u)du\right) \lambda_0(s)ds$$

is the quadratic variation process of $M(t)$. The steps to get the decomposition of W_n^2 is similar to those when testing the specification of a distribution. Notice that the process $\beta_n(t)$ is a martingale, obtained from the Doob-Meyer decomposition of $F_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{T_i \leq t\}}$. Hence its covariance kernel can be expressed as a transformation of the standard Brownian Motion covariance, i.e.,

$$\text{Cov}(\beta_n(s), \beta_n(t)) = H(s \wedge t).$$

If we define

$$\mu_j = \frac{4}{\pi^2(2j-1)^2}, \quad \varphi_j(t) = \sqrt{2} sin \frac{(2j-1)\pi t}{2}, \quad j = 1, 2, \cdots$$

as the eigenvalues and eigenfunctions of the standard Brownian Motion with covariance structure $K(s,t) = s \wedge t$, then

$$f_j(t) = \varphi_j(H(t)), j = 1, 2, \cdots$$

are the eigenfunctions of $\beta_n(t)$ with associated eigenvalues $\{\mu_j\}_{j=1}^{\infty}$. It is then standard to have the decomposition of the Cramér-von Mises test

$$W_n^2 = \sum_{j=1}^{\infty} \mu_j z_{nj}^2, \tag{0.3}$$

where

$$z_{nj} = \mu_j^{-1/2} \int_0^\infty \beta_n(t) f_j(t) H(dt)$$

are the PCs of $\beta_n(t)$.

Composite Hypothesis

The hypothesis discussed before are simple and the eigenvalues and eigenfunctions of the underlying empirical processes can be obtained from these of the standard Brownian Motion or Brownian Bridge through a suitable transformation. However, when we consider composite hypothesis, where the null depends on some unknown parameters, the decomposition of the Cramér-von Mises statistic is no longer straightforward due to the estimation effect. To be more precise, suppose we are interested in testing whether the distribution function of Y belongs to a parametric family, i.e.,

$$H_0 : F(y) = F(y, \theta), \text{ for some } \theta \in \Theta. \tag{0.4}$$

Or in the hazard case, whether the hazard function of T belongs to a parametric family, i.e.,

$$H_0 : \lambda(t) = \lambda(t, \theta), \text{ for some } \theta \in \Theta. \tag{0.5}$$

Since the empirical process contains unknown parameters, we need to replace them by their estimators when constructing the test statistic. Let us denote the estimator of θ as $\hat{\theta}$. In the distribution case, the Cramér-von Mises statistic turns out to be based on the empirical process after estimation

$$\hat{\alpha}_n(u) = n^{1/2}(\hat{F}_n(u) - u),$$

where $\hat{F}_n(u) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{F(Y_i,\hat{\theta}) \leq u\}}$, i.e.,

$$\hat{W}_n^2 = \int_0^1 \hat{\alpha}_n^2(u) du.$$

While in the hazard case, we have the martingale process after estimation

$$\hat{\beta}_n(t) = n^{-1/2} \sum_{i=1}^{n} \hat{M}_i(t),$$

where

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda(s, \hat{\theta}) ds.$$

The Cramér-von Mises statistic becomes

$$\hat{W}_n^2 = \int_0^\infty \hat{\beta}_n^2(t)\hat{H}(dt),$$

where

$$\hat{H}(t) = \int_0^t exp\left(-\int_0^s \lambda_0(u,\hat{\theta})du\right)\lambda_0(s,\hat{\theta})ds.$$

Although the estimator of the parameters, most commonly converges to the true value, the process with estimated parameters has a different asymptotic distribution than the distribution with known parameters. The limit distribution of $\hat{\alpha}_n$, or $\hat{\beta}_n$, is complicated, since it dependents on the true value of θ, the parametric form of F or λ, and on the particular estimator $\hat{\theta}$. The estimation of θ modifies the Brownian Bridge covariance structure of α_n and $\hat{\beta}_n$ is no longer a martingale. This leads to the difficulty of applying PCA in composite testing problems. That is, how to obtain the eigenvalues and eigenfunctions of the empirical process after estimation.

A lot of work has been done to deal with the estimation effect and to understand the nature of the process after estimation. Khmaladze (1981) proposed a martingale transformation method, which rules out the estimation effect and makes the resulting test asymptotically distribution free. This method has been applied in various model specification testing problems. See for example Koul and Stute (1999) for time series models, Delgado and Stute (2008) for conditional distributions, and Marzec and Marzec (1997) for conditional hazards. Rather than Khmaladze's idea to remove the estimation effect, Durbin, Knott and Taylor (1975, DKT hereafter) treated the estimation effect in a different way, where they faced up to the problem and investigated the nature of the estimation effect. Following on Durbin and Knott (1972), they developed a PCD method for $\hat{\alpha}_n$, based on a creative idea to construct the eigenfunctions of the process after estimation as linear combinations of the known eigenfunctions before estimation. The components of the Cramér-von Mises statistic, which are standard and asymptotically chi-square distributed, are then used for goodness-of-fit testing. Other papers that have applied DKT's method to deal with

estimation effect are Anh and Stute (2012) and Stute (1997). Anh and Stute (2012) derived the PCD of $\hat{\beta}_n(t)$ in the hazard scenario to test (0.5). Stute (1997) proposed smooth and directional tests for the mean regression function, i.e.,

$$H_0 : m(x) = E(Y \mid X = x) \in \mathcal{M} = \{m(\cdot, \theta) : \theta \in \Theta\}, \tag{0.6}$$

based on the PCD of the marked residual process

$$\hat{\gamma}_n(x) = n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \hat{\varepsilon}_i,$$

where $\hat{\varepsilon}_i = Y_i - m(X_i, \hat{\theta})$ are the residuals from mean regression.

Although DKT's idea is simple and provides a powerful technique for the goodness-of-fit test of a composite hypothesis, it has not drawn much attention in view of such few applications. This might be due to a serious limitation of DKT's method: it only works if the unknown parameters are finite-dimensional and are estimated by the asymptotically efficient estimator. Therefore the existing method is not suitable for nonparametric or semiparametric models, or such models that the efficient estimation is not available.

In chapter 1, motivated from the goodness-of-fit problem of the Cox model, which involves estimation of some finite-dimensional parameters and a nonparametric function, we follow DKT's idea and extend their method to accommodate any root n-consistent estimation of both parametric and nonparametric functions. In particular, we consider the specification test for the parametric part of the Cox model by retaining the proportional hazard assumption, i.e., consider the models specified as

$$\lambda(t \mid X) = \lambda_0(t) g(X), \ a.s.$$

where λ_0 is an unspecified baseline hazard functio and $g(\cdot)$ is a nonnegative function on X. We are testing the Cox specification, i.e.,

$$H_0 : g(X) = exp(\beta^T X), \ a.s. \text{ for some } \beta \in \Theta$$

against its negation. We follow the suggestion of Lin, Wei and Ying (1993) by considering the process

$$\hat{R}_n(x) = n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \hat{M}_i(\infty),$$

where

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) exp(\hat{\beta}^T X_i) d\hat{\Lambda}_0(s)$$

are the martingale residuals under the Cox specification with estimated regression parameter $\hat{\beta}$ and cumulative hazard function $\hat{\Lambda}_0$. Clearly, since the process $\hat{R}_n(x)$ contains an estimation of the finite-dimensional β and the nonparametric function Λ_0, DKT's PCD method does not work for it. Whereas the PCD method we propose has a much larger range of application than DKT's.

Conditional Hypothesis

In economic models, heterogeneity is an important issue and is usually explained through covariate effect. Suppose now rather than the marginal distribution of Y we are interested in the conditional distribution of Y given the covariable X. The X can be multivariate. We want to test whether the conditional distribution belongs to a parametric family, i.e.,

$$H_0 : F(y \mid X) = F(y \mid X, \theta), \ a.s. \text{ for some } \theta \in \Theta. \tag{0.7}$$

Andrews (1997) has proposed a conditional omnibus test based on the empirical process

$$\alpha_n(y, x) = n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \left(\mathbb{1}_{\{Y_i \leq y\}} - F\left(y \mid X_i, \theta\right) \right),$$

which can be viewed as the CUSUM version of $\alpha_n(u)$ w.r.t. the values of X. The difficulty to have PCA in this case lies in the fact that the process is a bivariate one with dependence between y and x. There is no explicit PCD for a multivariate process with possible dependence between its components. However, we notice that, conditional on X_i, PCD of the centered single-event process $\mathbb{1}_{\{Y_i \leq y\}} - F(y \mid X_i, \theta)$

is available because it has the same covariance kernel with a transformed Brownian Bridge. Based on this observation, we propose a conditional PCD method, which consists of two steps: (i) for each i, derive the PCD of the centered single-event process conditional on X_i, (ii) cumulatively sum up the obtained PCs w.r.t. the observed X_i's. It turns out that the summed up PCs form a sequence of new processes on (y, x) and we call them component processes of $\alpha_n(y, x)$. These component processes provide a basis for a class of goodness-of-fit tests. The application of the method in testing (0.7) is discussed in chapter 3, where we construct new goodness-of-fit tests based on the component processes and the tests outperform Andrew's test over a large range of alternatives. Not surprisingly, the conditional PCD method can also be applied in goodness-of-fit testing of conditional hazard models. In chapter 2, we discuss its application in the Cox model, namely, to test

$$H_0 : \lambda(t \mid X) = \lambda_0(t)exp(\beta^T X),\ a.s. \text{ for some } \beta \text{ and nonnegative } \lambda_0(t) \qquad (0.8)$$

against its negation. The conditional PCD is applied on the CUSUM process of martingales

$$R_n(t, x) = n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}} M_i(t),$$

where the martingale $M_i(t)$ takes the form of

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)exp(\beta^T X_i)\lambda_0(s)ds$$

under H_0.

In fact, the two PCD methods, one for composite hypothesis and one for conditional hypothesis, can be used cooperatively. It is convenient to see their relationship through the applications in the Cox model. Roughly speaking, the two PCD methods aim at the decomposition of the bivariate process $R_n(t, x)$ in different directions. The first method deals with a special case of $R_n(t, x)$ by taking $t = \infty$ and carries out the decomposition in x, while the conditional PCD method is for decomposition

of the marks $M_i(t)$ in t conditional on X_i. From another perspective, since x and t play different roles in $R_n(t, x)$, the decomposition methods in x and t should be different. In the Cox case, the two methods complement each other, namely, by obtaining PCD in both directions we are able to examine possible deviations in all aspects.

Applying PCD in goodness-of-fit testing yields more informative pieces of the omnibus test, based on which specially designed tests are available against deviations of interest. However there is a limitation of PCD's application, that is, no explicit PCD is available for a multivariate process with possibly dependent components. Considering the process $R_n(t, x)$ as an example again, its certain dependence structure between x and t prevents an explicit PCD. This is the reason why we choose to do the decomposition separately and our decomposition arguments for $R_n(t, x)$ provide inspiration on how to treat a multivariate process in general. There are other possible ways to avoid multivariate process in goodness-of-fit problem, for instance, in the multivariate regression case Stute, Xu and Zhu (2008) used the univariate residual empirical process instead of the multivariate process of the covariates and PCA on the univariate process was discussed.

In summary, the structure of the thesis is the following. In chapter 1, we introduce the first PCD method, which works for testing composite hypothesis and extends the existing method to have a larger application range. In chapter 2 and 3, we propose a conditional PCD method, which has not been used in any testing problem, and discuss its application in the conditional hazard and conditional distribution models, respectively.

Chapter 1

Model Checks for Marginal Effects in Proportional Hazard Models

1.1 Introduction

The Cox proportional hazard model has been widely used in many fields, including economics, since it was proposed by David Cox in 1972. The model specifies the distribution of the duration time through its hazard rate, which is the best candidate to describe a dynamic time-dependent phenomenon. The Cox model also introduces covariate effects to the hazard rate, which makes regression analysis possible for duration data under censorship. It is a semiparametric model, with finite-dimensional regression parameters and a nonparametric baseline hazard function. Specifically, the conditional hazard rate of the duration variable is assumed to be

$$\lambda(t \mid X) = \lambda_0(t) exp(\beta^T X), \tag{1.1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and the link function takes a linear exponential form. The estimation of the Cox model has been studied by Cox (1972, 1975) through a partial likelihood approach, while its large sample properties have been studied by Tsiatis (1981) and Andersen and Gill (1982) among others.

The Cox specification might fail in two ways. On one hand, the assumption of proportional hazard rates among individuals, i.e., the constant hazard ratio, might fail. On the other hand, the covariate effect might be misspecified. This misspecification might occur with the functional form of the covariables and with the exponential form of the link function. For model checking, various graphical methods and goodness-of-fit tests have been proposed in the literature. Schoenfeld (1980) proposed an omnibus chi-square test similar to the Pearson's test, by comparing the observed and expected numbers of occurrence in cells of a partition of the joint domain of covariables and duration time. However, the resulting test is inconsistent in infinite many departures from the null and the chosen partition favours a particular direction, which is a serious criticism of this approach. He also introduced the Schoenfeld residuals in 1982 and most of the existing tests for checking the proportional hazard assumption are based on the scaled Schoenfeld residuals, see Grambsch

and Therneau (1994). Another commonly used method for model checking consists of using the martingale residuals defined by Barlow and Prentice (1988). The martingale residuals, coming from the Doob-Meyer decomposition of the counting process, provide a basis for goodness-of-fit techniques for general hazard based models, see a comprehensive review in Martinussen and Scheike (2007). For the Cox model, Lin, Wei and Ying (1993) suggested an important class of goodness-of-fit tests based on CUSUM process of the martingale residuals, including an omnibus test that is consistent against any misspecification and several special cases to investigate different features of the Cox model.

In this article, we fix the proportionality assumption and pay attention to the specification test of the covariate effect in the Cox model. The proportional hazard model can be specified as

$$\lambda(t \mid X) = \lambda_0(t)g(X), \ a.s.$$

with an unspecified baseline hazard function λ_0 and a nonnegative function $g(\cdot)$. We want to test the Cox specification, i.e.,

$$H_0 : g(X) = exp(\beta^T X), \ a.s. \text{ for some } \beta \in \Theta, \tag{1.2}$$

against its negation. Lin, Wei and Ying (1993) constructed a Kolmogorov test to check this specification based on a special case of the CUSUM martingale residual process. They used a Monte Carlo simulation technique to approximate the limit distribution of the test. The purpose of this work is to derive principal components of their test. These PCs can be used to design more powerful smooth and directional tests, therefore complement existing proposals. Different PCs serve different purposes to detect particular departures from the null hypothesis.

The functional PCD method has been widely applied in different testing problems to get the decomposition of the test statistic. However, the decomposition requires the eigenvalues and eigenfunctions of the empirical process. Although all stochastic processes in a particular space admit PCDs, not all of them have closed forms for the

eigenvalues and eigenfunctions. Only a few processes with special covariance kernels have an analytical solution to the eigenproblem, such as the Brownian Motion, Brownian Bridge and martingales. Hence in order to be able to apply the PCD method, one needs to seek for these special processes. For example, Durbin and Knott (1972) discussed the specification test for the distribution function by using the standard empirical process. In this case, Donsker's theorem provides a Brownian Bridge limit, which makes PCD possible. Apart from the distribution model, the Brownian Motion and Brownian Bridge, together with their transformations, have been derived as limits in various model specification testing problems, which indicates a wide application of PCD. In the hazard models framework, the Doob-Meyer decomposition, by taking the difference of the counting process and its expected value, yields a martingale. Therefore, test statistics and further decompositions can be obtained based on the PCD of the martingale, e.g., Anh and Stute (2012).

Even if we have a process with available eigensolutions, another problem arises when the hypothesis is composite, in which case the empirical process after estimation differs significantly from the process with the true value of parameters. The limit distribution of the process after estimation becomes complicated since it dependents on the true value of the parameter, the parametric form of the identifier and on the particular estimator. As for the covariance of the limit process, estimation usually causes a shift and destroys the previous special covariance structure. To deal with this problem, Khmaladze (1981) proposed a martingale transformation method, which rules out the estimation effect and makes the resulting test distribution free. This method has been applied in various model specification testing problems. See for example Koul and Stute (1999) to test time series model, Delgado and Stute (2008) to test conditional distribution function, and Marzec and Marzec (1997) to test the Cox model. Rather than the martingale transformation idea to remove the estimation effect, Durbin, Knott and Taylor (1975, DKT) treated the estimation effect in a different way, where they faced up to the problem and inves-

tigated the nature of the estimation effect. In the framework of testing parametric distribution functions, they developed a PCD method for the empirical process after estimation, based on a creative idea to construct the eigenfunctions of the process after estimation, as linear combinations of the known eigenfunctions before estimation. The decomposition of the corresponding Cramér-von Mises test is then standard. Other papers that applied this method to deal with the estimation effect are Stute (1997) and Anh and Stute (2012), for testing parametric mean regression model and parametric hazard model, respectively.

In this paper, for the specification test of covariate effect in the Cox model, we follow the proposal of Lin, Wei and Ying (1993) by considering the CUSUM martingale residual process, which is asymptotically distributed as a transformed Brownian Motion when parameters are known. The challenge consists of providing a PCD for the process after estimation. We could apply DKT's idea, however, their PCD method does not help in any nonparametric or semiparametric setting because of a serious limitation: it only works with a finite-dimensional parametric efficient estimator. Motivated from this, we introduce a different argument to develop DKT's idea, for which we focus on the covariance kernel, rather than the Fourier coefficients in the existing papers. Our argument provides a more general PCD approach to accommodate any root n-consistent estimator of both parametric and nonparametric functions. Hence it is applicable in the Cox model. At the end, as expected, the limit Cramér-von Mises statistic can be decomposed into a weighted sum of independent chi-square components. Different types of tests can be constructed based on these components to improve efficiency.

The structure of this chapter will be the following. A brief introduction of the Cox model, the estimation approaches, and the tests based on CUSUM martingale residual are in section 1.2. The main results, including the PCD of the CUSUM martingale residual process, the construction of smooth tests and the orthogonal decomposition of the omnibus Cramér-von Mises test, are in section 1.3. We intro-

duce a discrete approximation of the covariance kernel, which simplifies the PCD argument in the computation viewpoint, in section 1.4. The numerical approximation is presented in section 1.5. A simulation study illustrating the performance of the test in the finite sample is reported in section 1.6.

1.2 The Cox Hazard Regression Model

In the framework of regression analysis with right-censored duration data, consider a sample $\{Z_i, \Delta_i, X_i\}, i = 1, \cdots, n$ of i.i.d. realizations of $\{Z, \Delta, X\}$. Here Z is the minimum of the non-negative failure and censoring time, which are denoted by T and C, i.e., $Z = min(T, C)$. The indicator $\Delta = \mathbb{1}_{\{T \leq C\}}$ contains the information indicating which of T and C is actually observed, and X is the covariable vector.

The conditional distribution of failure time is usually better described through its hazard functions rather than densities. The conditional cumulative hazard function is given by

$$\Lambda(t \mid X) = \int_0^t \frac{dF(u \mid X)}{1 - F(u- \mid X)}, \tag{1.3}$$

where F is the conditional distribution function of the failure time. If F admits a Lebesgue density f, we have

$$d\Lambda(t \mid X) = \frac{f(t \mid X)}{1 - F(t \mid X)} dt.$$

The function

$$\lambda(t \mid X) = \frac{f(t \mid X)}{1 - F(t \mid X)}$$

is the conditional hazard function. It can also be expressed in terms of a conditional probability as

$$\lambda(t \mid X) = \lim_{h \to 0} h^{-1} P(t \leq T < t + h \mid T \geq t, X). \tag{1.4}$$

It gives us the risk of the event occurring as a function of time, conditional on not having occurred before. In the Cox model, the conditional hazard rate is assumed

to have the multiplicative form as (1.1).

1.2.1 The Counting Process Approach

Another approach to the censored data regression model is based on the analysis of counting process. Define the following two processes

$$N(t) = \mathbb{1}_{\{Z \le t, \Delta = 1\}},$$

$$Y(t) = \mathbb{1}_{\{Z \ge t\}}.$$

Here $N(t)$ is the counting process, and $Y(t)$ is the at-risk process. Applying the Doob-Meyer decomposition, there is a unique predictable process $A(t)$ such that $N(t) - A(t)$ is a martingale and $A(t)$ is called the compensator of $N(t)$. In the counting process approach, instead of modeling conditional hazard rate of T, the compensator process is modeled. Notice that the information contained in $\{Z, \Delta\}$ is equivalent to that contained in $\{N, Y\}$. Actually, these two approaches are equivalent under the conditional independence of T and C on X. To be more specific

$$M(t) = N(t) - \int_0^t Y(u) d\Lambda(u \mid X) \tag{1.5}$$

is a martingale process with the filtration $\mathscr{F}_t = \sigma\{X, N(u), Y(u+) : 0 \le u \le t\}$. Then modeling the compensator $\int_0^t Y(s) d\Lambda(s \mid X)$ is equivalent to modeling the conditional hazard.

The counting process counts the number of occurrence of the event, while the compensator captures its expected value. Thus, the martingale, as the difference of them, plays the same role with the error term in the mean regression model. The Doob-Meyer decomposition serves the same purpose with the projection decomposition, but instead of orthogonality between two random variables, we have martingale process.

Under the counting process framework, if the Cox specification is correct for a given sample, there exists a β and $\lambda_0(t)$, such that

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)exp(\beta^T X_i)\lambda_0(s)ds \quad i = 1, \cdots, n \tag{1.6}$$

are martingales. The corresponding martingale residuals are defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)exp(\hat{\beta}^T X_i)d\hat{\Lambda}_0(s), \tag{1.7}$$

where $\hat{\beta}$ is an estimator of β and $\hat{\Lambda}_0(t)$ is an estimator of the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. These martingale residuals provide a basis for goodness-of-fit test for the Cox model.

1.2.2 Estimation Approaches of the Cox Model

From now on, we restrict the observations of the duration time on $[0, \tau]$, for convenience. This τ is arbitrary, namely, our arguments work for any $\tau < \infty$. It can be easily extended to use all the observations on $[0, \infty)$, which we will discuss in section 1.7. The estimation of the Cox model was suggested by Cox (1972, 1975) using partial likelihood inference. The partial likelihood score function for β is

$$U(\beta, \tau) = \sum_{i=1}^{n} \int_0^\tau \Big(X_i - \bar{X}(\beta, t)\Big) dN_i(t), \tag{1.8}$$

where

$$\bar{X}(\beta, t) = \frac{\sum_{i=1}^n Y_i(t)e^{\beta^T X_i} X_i}{\sum_{i=1}^n Y_i(t)e^{\beta^T X_i}}.$$

The partial likelihood estimator $\hat{\beta}$ is the solution to $U(\beta, \tau) = 0$. Under some mild regularity conditions, $n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution to a centered Gaussian variable with covariance matrix $\Sigma(\beta_0, \tau)^{-1}$. The matrix $\Sigma(\beta, t)$ is defined as

$$\Sigma(\beta, t) = \mathbb{E}\left(\int_0^t (X - \tilde{X}(\beta, s))^2 Y(s)e^{\beta^T X} d\Lambda_0(s)\right),$$

with

$$\tilde{X}(\beta, t) = \frac{\mathbb{E}\left(Y(t)e^{\beta^T X} X\right)}{\mathbb{E}\left(Y(t)e^{\beta^T X}\right)}$$

being the limit of $\bar{X}(\beta, t)$. The cumulative baseline hazard is estimated by the Breslow (1974) estimator

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u) e^{\hat{\beta}^T X_i}}. \tag{1.9}$$

Apart from the likelihood inference, Chen, Jin and Ying (2002) proposed another estimation approach based on moment conditions. They considered a general transformation model with the Cox model as a special case. Motivated by the fact that $M(t)$ is a martingale, the estimation equations for the Cox model are

$$\begin{aligned} &\sum_{i=1}^n \int_0^\tau X_i \left(dN_i(t) - Y_i(t) e^{\beta^T X_i} d\Lambda_0(t) \right) = 0, \\ &\sum_{i=1}^n \left(dN_i(t) - Y_i(t) e^{\beta^T X_i} d\Lambda_0(t) \right) = 0, \ t \geq 0, \end{aligned} \tag{1.10}$$

where $\Lambda_0(t)$ belongs to a collection of nondecreasing step functions with $\Lambda_0(0) = 0$ and with jumps only at the observed duration times. Although with different motivations, the solution of the above equations coincides with the partial likelihood estimator and the Breslow estimator.

1.2.3 Test the Covariate Effect in the Cox Model

To test the specification of the Cox model, i.e., to test

$$H_0 : \lambda(t \mid X) = \lambda_0(t) exp(\beta^T X) \ a.s. \ \text{for some } \beta \text{ and nonnegative } \lambda_0(t),$$

against all the possible alternatives, Lin, Wei and Ying (1993) proposed an omnibus test by considering the CUSUM of martingale residuals

$$\hat{c}(t, x) = n^{-1/2} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \hat{M}_i(t),$$

where $\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) exp(\hat{\beta}^T X_i) d\hat{\Lambda}_0(s), i = 1, \cdots, n$ are martingale residuals with partial likelihood estimator $\hat{\beta}$ and Breslow estimator $\hat{\Lambda}_0(t)$.

To test the covariate effect specification (1.2), they considered a special case of $\hat{c}(t,x)$, by taking $t=\tau$, i.e., the process

$$\hat{c}(x) := n^{-1/2}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\le x\}}\hat{M}_i(\tau).$$

Here we consider any fixed $\tau>0$ but later we will let $\tau\to\infty$. They showed that, under some mild conditions, this process has a large sample behavior

$$\begin{aligned}\hat{c}(x) = n^{-1/2}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\le x\}}M_i(\tau) - \Bigg[& n^{-1/2}\sum_{i=1}^{n}\int_0^\tau l(\beta_0,s,x)dM_i(s) \\ & + \mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x\}}\left(X-\tilde{X}(\beta_0,s)\right)d\Lambda_0(s)\right) \\ & \times \Sigma(\beta_0,\tau)^{-1}n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \left(X_i-\tilde{X}(\beta_0,s)\right)dM_i(s)\Bigg] + o_p(1),\end{aligned} \tag{1.11}$$

with

$$l(\beta,t,x) = \frac{\mathbb{E}\left(Y(t)e^{\beta^T X}\mathbb{1}_{\{X\le x\}}\right)}{\mathbb{E}\left(Y(t)e^{\beta^T X}\right)},$$

and $\tilde{X}(\beta,t)$, $\Sigma(\beta,t)$ defined in section 1.2.2. The term in the square bracket $[\cdots]$ in (1.11) is the estimation effect. It is a sum of two terms, of which the first one is the estimation effect of the nonparametric $\Lambda_0(t)$ and the second one is the estimation effect of parametric β. These two effects are orthogonal in this large sample limit. (1.11) can be rewritten in terms of a sum of i.i.d. martingale integrals

$$\hat{c}(x) = n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \tilde{h}_i(\beta_0,s,x)dM_i(s) + o_p(1), \tag{1.12}$$

with

$$\begin{aligned}\tilde{h}_i(\beta,s,x) = & \mathbb{1}_{\{X_i\le x\}} - l(\beta,s,x) - \mathbb{E}\left(\int_0^\tau Y(s)e^{\beta^T X}\mathbb{1}_{\{X\le x\}}\left(X-\tilde{X}(\beta,s)\right)d\Lambda_0(s)\right) \\ & \times \Sigma(\beta,\tau)^{-1}\left(X_i-\tilde{X}(\beta,s)\right).\end{aligned}$$

Therefore, together with tightness, they showed that the process $\hat{c}(x)$ converges weakly to a centered Gaussian process in the space $D[-\infty,\infty]$, as $n\to\infty$,

$$\hat{c}(x) \xrightarrow{d} \hat{c}_\infty(x). \tag{1.13}$$

The limit Gaussian process has covariance kernel

$$K(t_1, t_2, x_1, x_2) = \mathbb{E}\left[\int_0^\infty \tilde{h}_i(\beta_0, s, x_1)\tilde{h}_i(\beta_0, s, x_2)Y_i(s)e^{\beta_0^T X_i}\lambda_0(s)ds\right].$$

Kolmogorov test based on $\hat{c}(x)$ was constructed, and they proposed a Monte Carlo simulation technique to approximate its limit distribution.

1.3 Principal Component Decomposition

In this section, we develop PCD of the limit Gaussian process $\hat{c}_\infty(x)$ and develop an orthogonal decomposition of the corresponding Cramér-von Mises test statistic. We follow DKT's idea to construct the eigenfunctions of $\hat{c}_\infty(x)$ as linear combinations of the known eigenfunctions before estimation, which are the transformed Brownian Motion eigenfunctions in our case. Although the idea is the same, we provide a more general PCD approach, which allows us to use any root n-consistent estimator, especially for the models where an efficient estimator is not available.

We introduce the following assumptions. The first three are standard assumptions of the Cox model. The fourth one is needed to get the asymptotic distribution of the partial likelihood estimator. The asymptotic results of $\hat{c}(x)$ in section 1.2.3 require assumption (A1)-(A4). In addition, we assume real-valued X's. Extension to the multivariate case is discussed in section 1.7.

(A1). T and C are independent conditional on X.

(A2). $P\{Y(\tau) = 1\} > 0$.

(A3). X is bounded.

(A4). $\Sigma(\beta_0, \tau) = E\left[\int_0^\tau (X - \tilde{X}(\beta_0, s))^2 Y(s)e^{\beta_0 X}\lambda_0(s)ds\right]$ is positive definite.

If the counting process has continuous compensator, which is equivalent to having

continuous $\Lambda_0(t)$, the covariance kernel of $\hat{c}_\infty(x)$ is

$$\begin{aligned}
K^c(x_1,x_2) &:= \mathrm{Cov}(\hat{c}_\infty(x_1),\hat{c}_\infty(x_2)) \\
&= \mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_1\wedge x_2\}}d\Lambda_0(s)\right) \\
&\quad -\int_0^\tau \frac{\mathbb{E}\left(Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_1\}}\right)\mathbb{E}\left(Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_2\}}\right)}{\mathbb{E}\left(Y(s)e^{\beta_0^T X}\right)}d\Lambda_0(s) \\
&\quad -\mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_1\}}\left(X-\tilde{X}(\beta_0,s)\right)d\Lambda_0(s)\right) \\
&\qquad \times \Sigma(\beta_0,\tau)^{-1}\mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_2\}}\left(X-\tilde{X}(\beta_0,s)\right)d\Lambda_0(s)\right).
\end{aligned}$$

To simplify the notation, let us denote functions

$$H_c(x) := \mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x\}}d\Lambda_0(s)\right),$$

$$H_l(x,t) := \mathbb{E}\left(Y(t)e^{\beta_0^T X}\mathbb{1}_{\{X\le x\}}\right),$$

and

$$H_b(x) := \mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x\}}\left(X-\tilde{X}(\beta_0,s)\right)d\Lambda_0(s)\right).$$

Then

$$K^c(x_1,x_2) = H_c(x_1\wedge x_2) - \int_0^\tau \frac{H_l(x_1,s)H_l(x_2,s)}{H_l(\infty,s)}d\Lambda_0(s) - H_b(x_1)\Sigma(\beta_0,\tau)^{-1}H_b(x_2). \tag{1.14}$$

The limiting covariance kernel $K^c(x_1,x_2)$ consists of three terms. The first term $H_c(x_1\wedge x_2)$ is the covariance kernel before estimation, which takes the form of a transformed Brownian Motion covariance. The last two terms are the shifts caused by estimation of the nonparametric $\Lambda_0(t)$ and parametric β, respectively.

We begin with the PCD of the covariance before estimation. Let

$$\mu_k = \frac{4}{\pi^2(2k-1)^2}, \quad \varphi_k(t) = \sqrt{2}sin\frac{(2k-1)\pi t}{2},\ k=1,2,\cdots$$

be the eigenvalues and eigenfunctions of the standard Brownian Motion with covariance kernel $K(s,t)=s\wedge t$. The eigenfunctions of $H_c(x_1\wedge x_2)$ can be obtained by defining the following transformations

$$f_k(x) = \varphi_k(H_c(x)/H_c(\infty)),\ k=1,2,\cdots.$$

Then $\{f_k(x)\}_{k=1}^{\infty}$ forms an orthonormal basis of a subspace of $L^2(\mathbb{R}, H_c(x)/H_c(\infty))$, the Hilbert space of all square integrable functions on $\mathbb{R}$ with the inner product

$$\langle \rho, g \rangle := \int_{-\infty}^{\infty} \rho(x) g(x) \frac{H_c(dx)}{H_c(\infty)},$$

since

$$\begin{aligned} \langle f_k, f_h \rangle &= \int_{-\infty}^{\infty} \varphi_k \left(\frac{H_c(x)}{H_c(\infty)} \right) \varphi_h \left(\frac{H_c(x)}{H_c(\infty)} \right) \frac{H_c(dx)}{H_c(\infty)} \\ &= \int_0^1 \varphi_k(u) \varphi_h(u) du = \begin{cases} 1 & k = h \\ 0 & k \neq h \end{cases}. \end{aligned}$$

Moreover, $\{f_k(x)\}_{k=1}^{\infty}$ are the eigenfunctions of the covariance kernel $H_c(x_1 \wedge x_2)/H_c(\infty)$ with associated eigenvalues $\{\mu_k\}_{k=1}^{\infty}$, i.e.,

$$\int_{-\infty}^{\infty} \frac{H_c(x_1 \wedge x_2)}{H_c(\infty)} f_k(x_1) \frac{H_c(dx_1)}{H_c(\infty)} = \mu_k f_k(x_2).$$

By Mercer's theorem, the covariance kernel $H_c(x_1 \wedge x_2)/H_c(\infty)$ can be decomposed as

$$H_c(x_1 \wedge x_2)/H_c(\infty) = \sum_{k=1}^{\infty} \mu_k f_k(x_1) f_k(x_2). \tag{1.15}$$

It is more convenient to write the decomposition (1.15) into a matrix form. Let us denote $\boldsymbol{f}(x)$ and $\boldsymbol{\mu}$ as the infinite-dimensional vector and matrix of all the eigenfunctions and eigenvalues, i.e.,

$$\boldsymbol{f}(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \end{pmatrix},$$

and

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 & & 0 \\ & \mu_2 & \\ 0 & & \ddots \end{pmatrix}.$$

Then the decomposition (1.15) can be rewritten as

$$H_c(\infty)^{-1} H_c(x_1 \wedge x_2) = \boldsymbol{f}^T(x_1) \boldsymbol{\mu} \boldsymbol{f}(x_2). \tag{1.16}$$

Now we discuss how to decompose the covariance kernel with estimation shifts. Recall that, the estimation shifts consist of functions in x_1, x_2 and t, which are $H_l(x_1, t)$, $H_l(x_2, t)$, $H_b(x_1)$ and $H_b(x_2)$. For fixed t, they are all univariate functions in x_1 or x_2, thus, they all admit a decomposition on the basis $\boldsymbol{f}(x)$. The objective is to write the estimation shifts as similar "sandwich" forms to (1.16), in which the basis $\boldsymbol{f}(x)$ appear on both sides and the corresponding coefficients appear in the middle. Let us define the coefficients of $H_l(x, t)$, for each t, and the coefficients of $H_b(x)$, on the basis $\boldsymbol{f}(x)$, as infinite-dimensional vectors

$$\boldsymbol{\delta_l}(t) := \begin{pmatrix} \langle H_l(\cdot, t), f_1 \rangle \\ \langle H_l(\cdot, t), f_2 \rangle \\ \vdots \end{pmatrix} \quad \text{and} \quad \boldsymbol{\delta_b} := \begin{pmatrix} \langle H_b, f_1 \rangle \\ \langle H_b, f_2 \rangle \\ \vdots \end{pmatrix}.$$

Then the decompositions of $H_l(x, t)$ and $H_b(x)$ on $\boldsymbol{f}(x)$ are

$$H_l(x, t) = \boldsymbol{f}^T(x)\boldsymbol{\delta_l}(t),$$

for each t, and

$$H_b(x) = \boldsymbol{f}^T(x)\boldsymbol{\delta_b}. \tag{1.17}$$

Plugging these decompositions into the estimation shifts, we have

$$\int_0^\tau \frac{H_l(x_1, s)H_l(x_2, s)}{H_l(\infty, s)} d\Lambda_0(s) = \boldsymbol{f}^T(x_1) \left(\int_0^\tau \boldsymbol{\delta_l}(s) H_l(\infty, s)^{-1} \boldsymbol{\delta_l}^T(s) d\Lambda_0(s) \right) \boldsymbol{f}(x_2),$$

and

$$H_b(x_1)\Sigma(\beta_0, \tau)^{-1} H_b(x_2) = \boldsymbol{f}^T(x_1) \left(\boldsymbol{\delta_b} \Sigma(\beta_0, \tau)^{-1} \boldsymbol{\delta_b}^T \right) \boldsymbol{f}(x_2). \tag{1.18}$$

These are the "sandwich" form that we expect. Together with (1.16), we have

$$\begin{aligned} & K^c(x_1, x_2) \\ &= \boldsymbol{f}^T(x_1) \left(H_c(\infty)\boldsymbol{\mu} \right) \boldsymbol{f}(x_2) - \boldsymbol{f}^T(x_1) \left(\int_0^\tau \boldsymbol{\delta_l}(s) H_l(\infty, s)^{-1} \boldsymbol{\delta_l}^T(s) d\Lambda_0(s) \right) \boldsymbol{f}(x_2) \\ &\qquad - \boldsymbol{f}^T(x_1) \left(\boldsymbol{\delta_b} \Sigma(\beta_0, \tau)^{-1} \boldsymbol{\delta_b}^T \right) \boldsymbol{f}(x_2) \\ &= \boldsymbol{f}^T(x_1) \left(H_c(\infty)\boldsymbol{\mu} - \int_0^\tau \boldsymbol{\delta_l}(s) H_l(\infty, s)^{-1} \boldsymbol{\delta_l}^T(s) d\Lambda_0(s) - \boldsymbol{\delta_b} \Sigma(\beta_0, \tau)^{-1} \boldsymbol{\delta_b}^T \right) \boldsymbol{f}(x_2). \end{aligned} \tag{1.19}$$

If we define an infinite-dimensional matrix

$$\boldsymbol{M} := H_c(\infty)\boldsymbol{\mu} - \int_0^\tau \boldsymbol{\delta_l}(s)H_l(\infty, s)^{-1}\boldsymbol{\delta_l}^T(s)d\Lambda_0(s) - \boldsymbol{\delta_b}\Sigma(\beta_0, \tau)^{-1}\boldsymbol{\delta_b}^T,$$

then

$$K^c(x_1, x_2) = \boldsymbol{f}^T(x_1)\boldsymbol{M}\boldsymbol{f}(x_2). \tag{1.20}$$

The last step to obtain the eigenvalues and eigenfunctions of $K^c(x_1, x_2)$ consists of diagonalizing the matrix $\boldsymbol{M}$. There exist an orthonormal matrix $\boldsymbol{N}$ and a diagonal matrix

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \end{pmatrix}$$

with $\lambda_1 > \lambda_2 > \cdots > 0$, which are actually the matrixes of eigenvectors and eigenvalues of $\boldsymbol{M}$, such that

$$\boldsymbol{M} = \boldsymbol{N}\boldsymbol{\lambda}\boldsymbol{N}^T. \tag{1.21}$$

See Theorem 2.3 in Stute (1997). Thus,

$$K^c(x_1, x_2) = \boldsymbol{f}^T(x_1)\boldsymbol{N}\boldsymbol{\lambda}\boldsymbol{N}^T\boldsymbol{f}(x_2). \tag{1.22}$$

This expression is the PCD of $K^c(x_1, x_2)$, and gives us the eigenvalues and eigenfunctions of $K^c(x_1, x_2)$, which are, in the matrix form,

$$\boldsymbol{\lambda} \text{ and } \boldsymbol{N}^T\boldsymbol{f}(x).$$

Clearly, the eigenfunctions after estimation are linear combinations of the f_k's, and the weights of the linear combinations are contained in the matrix $\boldsymbol{N}$. Let $\boldsymbol{N}_k$ be the k-th column vector of $\boldsymbol{N}$, then $\boldsymbol{N}_k^T\boldsymbol{f}(x)$ is the k-th eigenfunction of $K^c(x_1, x_2)$ with eigenvalue λ_k. Now we are able to study the properties of the Fourier coefficients of the process $\hat{c}_\infty(x)$, which are also called principal components of $\hat{c}_\infty(x)$. By Kac-Siegert, we have the joint distribution of its principal components summarized in Theorem 1.

Theorem 1 *Under the null hypothesis and assumptions* (A1)-(A4)*, the vector of all the principal components of* $\hat{c}_\infty(x)$ *have distribution*

$$\langle \boldsymbol{N}^T \boldsymbol{f}, \hat{c}_\infty \rangle = \boldsymbol{N}^T \begin{pmatrix} \langle f_1, \hat{c}_\infty \rangle \\ \langle f_2, \hat{c}_\infty \rangle \\ \vdots \end{pmatrix} \sim \mathcal{N}_\infty(\mathbf{0}, \boldsymbol{\lambda}). \tag{1.23}$$

Here $\langle f_k, \hat{c}_\infty \rangle = H_c(\infty)^{-1} \int_{-\infty}^{\infty} f_k(x)\hat{c}_\infty(x)H_c(dx)$ *is the* k*-th coefficient of* $\hat{c}_\infty(x)$ *on the basis* $\{f_k(x)\}_{k=1}^{\infty}$.

Theorem 1 says that, the k-th principal component of $\hat{c}_\infty(x)$, which is $\langle \boldsymbol{N}_k^T \boldsymbol{f}, \hat{c}_\infty \rangle$, is distributed as a normal random variable with mean zero and variance λ_k. Besides, the principal components are independent of each other. Suppose we have uniform consistent estimators $\hat{H}_0$, $\hat{f}_k$ and consistent $\hat{\lambda}_k$, $\hat{\boldsymbol{N}}_k$ of the corresponding unknown terms in (1.23), the normalized empirical coefficients of $\hat{c}(x)$ will be

$$\zeta_k := \hat{\lambda}_k^{-1/2} \hat{H}_c(\infty)^{-1} \int_{-\infty}^{\infty} \hat{\boldsymbol{N}}_k^T \hat{\boldsymbol{f}}(x)\hat{c}(x)\hat{H}_c(dx).$$

We immediately have its following convergence result.

Corollary 2 *Under the null hypothesis and assumptions* (A1)-(A4)*, we have the* k*-th normalized empirical coefficient of* $\hat{c}(x)$ *convergences in distribution to a standard normal variable, as* $n \to \infty$*, i.e.,*

$$\zeta_k \xrightarrow{d} \mathcal{N}(0,1), \ k = 1, 2, \cdots. \tag{1.24}$$

The limit normal variables are independent among k*'s.*

From the distribution-free result in Corollary 2, we might construct distribution free smooth tests, which are based on a sum of a few squared coefficients, and even optimal directional tests. We will only discuss the smooth test in this work, for which we have the convergence of the test based on the first r squared coefficients as

$$S_{nr}^2 := \sum_{k=1}^{r} \zeta_k^2 \xrightarrow{d} \chi_r^2. \tag{1.25}$$

How to estimate the unknown terms and approximate $\hat{\lambda}_k$ and $\hat{\boldsymbol{N}}_k$ in practice will be discussed in section 1.5.

In addition to the tests based on a few components, we also have a Cramér-von Mises type omnibus test, which is a functional of the CUSUM process and has a representation as a weighted sum of all squared components. This can be obtained from the quadratic form of (1.23)

$$\langle \boldsymbol{N}^T \boldsymbol{f}, \hat{c}_\infty \rangle^T \langle \boldsymbol{N}^T \boldsymbol{f}, \hat{c}_\infty \rangle = \sum_{k=1}^{\infty} \lambda_k Z_k^2, \tag{1.26}$$

where the Z_k's are independent standard normal variables. The left-hand side of equation (1.26) actually equals to the limit of the corresponding Cramér-von Mises statistics, which is defined as

$$CvM_\infty := H_c(\infty)^{-1} \int_{-\infty}^{\infty} \left(\hat{c}_\infty(x)\right)^2 H_c(dx),$$

and it is the large sample limit of

$$CvM_n := \hat{H}_c(\infty)^{-1} \int_{-\infty}^{\infty} \left(\hat{c}(x)\right)^2 \hat{H}_c(dx).$$

This result is summarized in Corollary 3 and the proof is in Appendix A.

Corollary 3 *Under the null hypothesis and assumptions* (A1)-(A4)*, the limit of the Cramér-von Mises statistic based on* $\hat{c}(x)$ *can be decomposed as a weighted sum of independent chi-square variables with degree of freedom one, i.e.,*

$$CvM_n \xrightarrow{d} CvM_\infty = \sum_{k=1}^{\infty} \lambda_k Z_k^2, \tag{1.27}$$

where the weights λ_k*'s are the eigenvalues of the limit Gaussian process* $\hat{c}_\infty(x)$*.*

Corollary 3 provides the decomposition of the limit omnibus Cramér-von Mises test. To actually compute the weights λ_k's and the eigenfunctions, we need to diagonalize the matrix $\boldsymbol{M}$, and this requires to replace the unknown $K^c(x_1, x_2)$ by its consistent estimator. Since $K^c(x_1, x_2)$ depends on the true value of β and $\Lambda_0(t)$, we need to plug in the partial likelihood and Breslow estimators. Note that the

Breslow estimator of $\Lambda_0(t)$ is discrete and only jumps at the observed uncensored durations. As a consequence, plugging in the discrete estimated $\Lambda_0(t)$ will lead to a structural change of $K^c(x_1, x_2)$, namely, the integrals in $K^c(x_1, x_2)$ with respect to $\Lambda_0(t)$ can now be written as discrete summations. More importantly, the estimation shifts caused by β and $\Lambda_0(t)$, which are separated in $K^c(x_1, x_2)$, can now be combined in a more compact form, which will simplify the PCD argument. In next section, we introduce a discrete approximation of $K^c(x_1, x_2)$, which discretizes the integrals with respect to $\Lambda_0(t)$ into summations and combines the estimation shifts in a compact matrix form. This approximation is helpful from a computation viewpoint.

We finish this section by adding an important remark. We have mentioned that DKT's idea works for any consistent estimator, not only for the efficient one, however, this fact has not been clearly presented in existing papers, where the PCD arguments only work for the efficient estimators. For example, in DKT's original paper, a key step for later PCD is the representation (4.14) of the Fourier coefficients as a projection of a standard normal vector, for which they took advantage of the efficient MLE. Stute (1997) developed the PCD of the marked residual process based on a parallel projection result, which follows from the efficient LSE, see his proof of Theorem 2.1 that relies on the property of efficient estimators. Moreover, these projection results not only rely on the efficiency of the estimators, but also are restricted to single equation estimation problem. Take the Cox model as an example: although the partial likelihood and Breslow estimators are asymptotically efficient, we do not have a similar projection result because they cannot be regarded as the solution of a single equation problem, see Appendix B, where we use an optimal instrumental variable estimation based on the moment conditions of multivariate martingale increments.

While our PCD argument above, by focusing on the covariance kernel, provide a general PCD approach, which accommodates different estimations and also estimation of nuisance parameters in semiparametric models. In general, if the plugged in

estimator is inefficient, the estimation shift will have two additional terms, see (C.4) in Appendix C for example. However, we still have the "sandwich" form for these additional terms, thus, we can sum up the middle coefficient terms as what we do for the two shifts in (1.19). As long as the sum is symmetric, which is guaranteed by the symmetry of the covariance kernel, DKT's idea would work.

To see the limitation of the existing methods more clearly, we may look at the Fourier coefficients in (1.23), for which we have

$$\langle \boldsymbol{f}, \hat{c}_\infty \rangle = \begin{pmatrix} \langle f_1, \hat{c}_\infty \rangle \\ \langle f_2, \hat{c}_\infty \rangle \\ \vdots \end{pmatrix} \sim \mathcal{N}_\infty \left(\mathbf{0}, \boldsymbol{N}\boldsymbol{\lambda}\boldsymbol{N}^T \right) = \mathcal{N}_\infty \left(\mathbf{0}, \boldsymbol{M} \right).$$

If we normalize the coefficients, then

$$\left(H_c(\infty)\boldsymbol{\mu} \right)^{-1/2} \langle \boldsymbol{f}, \hat{c}_\infty \rangle \sim \mathcal{N}_\infty \left(\mathbf{0}, \bar{\boldsymbol{M}} \right), \tag{1.28}$$

where the variance and covariance matrix $\bar{\boldsymbol{M}}$ equals to

$$\begin{aligned} \bar{\boldsymbol{M}} &= \left(H_c(\infty)\boldsymbol{\mu} \right)^{-1/2} \boldsymbol{M} \left(H_c(\infty)\boldsymbol{\mu} \right)^{-1/2} \\ &= \boldsymbol{I}_\infty - \left(H_c(\infty)\boldsymbol{\mu} \right)^{-1/2} \Bigg(\int_0^\tau \boldsymbol{\delta_l}(s) H_l(\infty, s)^{-1} \boldsymbol{\delta_l}^T(s) d\Lambda_0(s) \\ &\qquad + \boldsymbol{\delta_b} \Sigma(\beta_0, \tau)^{-1} \boldsymbol{\delta_b}^T \Bigg) \left(H_c(\infty)\boldsymbol{\mu} \right)^{-1/2}. \end{aligned} \tag{1.29}$$

If the matrix $\bar{\boldsymbol{M}}$ is a projection, which is idempotent, namely $\bar{\boldsymbol{M}} = \bar{\boldsymbol{M}}\bar{\boldsymbol{M}}$, then (1.28) gives us a representation of the normalized coefficients as a projection of independent standard normal variables, i.e.,

$$\left(H_c(\infty)\boldsymbol{\mu} \right)^{-1/2} \langle \boldsymbol{f}, \hat{c}_\infty \rangle = \bar{\boldsymbol{M}}\boldsymbol{Z}, \tag{1.30}$$

where the $\boldsymbol{Z}$ is the multivariate standard normal vector. This is a similar result to (4.14) in DKT and Theorem 2.2 in Stute (1997), where $\bar{\boldsymbol{M}} = \boldsymbol{I} - \boldsymbol{\Delta} \left(\boldsymbol{\Delta}^T \boldsymbol{\Delta} \right)^{-1} \boldsymbol{\Delta}^T$, for some $\boldsymbol{\Delta}$. However, this $\bar{\boldsymbol{M}}$ will be a projection only when the estimator is efficient and comes from a single equation problem. For example, in the Cox model case, the $\bar{\boldsymbol{M}}$ given in (1.29) is not a projection matrix, thus, (1.30) is not true. This explains

why the existing methods fail with inefficient estimators, because the key steps in the existing methods, which are parallel to (1.30), do not hold under inefficient estimators.

1.4 Discrete Approximation of the Covariance Kernel

In this section, we introduce a discrete approximation of $K^c(x_1, x_2)$, which is inspired by an estimation of discrete $\Lambda_0(t)$ and β based on moment conditions of discrete martingale increments. The estimation of discrete $\Lambda_0(t)$ and β is introduced in Appendix B, and the discrete approximation of $K^c(x_1, x_2)$ is developed in Appendix C. Here we only show the result by introducing the following notations.

Let $0 = t_0 < t_1 < t_2 < \cdots < t_m = \tau$ be a finite grid on time. (For example, we can take $t_j = j\tau/m$, $j = 1, \cdots, m$, in which case the partition is of equal length τ/m.) Denote the approximated increments of the compensator $A(t)$ as

$$\Delta A_j := Y(t_j)e^{\beta_0^T X}(\Lambda_0(t_j) - \Lambda_0(t_{j-1})),\ j = 1, 2, \cdots, m,$$

with the true value of β and $\Lambda_0(t)$. Define a covariance kernel

$$K^m(x_1, x_2) := \boldsymbol{v}^T \left(\boldsymbol{H}(x_1 \wedge x_2) - \boldsymbol{G}^T(x_1)\boldsymbol{A}^{-1}\boldsymbol{G}(x_2) \right) \boldsymbol{v}, \tag{1.31}$$

with $\boldsymbol{v} = (1, \cdots, 1)^T$ as the one vector of dimension m,

$$\boldsymbol{H}(x) := \mathbb{E}\left(\mathbb{1}_{\{X \le x\}} \begin{pmatrix} \Delta A_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta A_m \end{pmatrix} \right),$$

$$\boldsymbol{G}(x) := \mathbb{E}\left(\mathbb{1}_{\{X \le x\}} \begin{pmatrix} \Delta A_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta A_m \\ \Delta A_1 X & \cdots & \Delta A_m X \end{pmatrix} \right),$$

and

$$\boldsymbol{A} := \mathbb{E}\begin{pmatrix} \Delta A_1 & \cdots & 0 & \Delta A_1 X \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \Delta A_m & \Delta A_m X \\ \Delta A_1 X & \cdots & \Delta A_m X & \left(\sum_{j=1}^m \Delta A_j\right) X^2 \end{pmatrix}.$$

The covariance kernel $K^m(x_1, x_2)$ consists of two terms, where the first one is a transformed Brownian Motion covariance, and the second one is the estimation shift, which is more concise and compact compared to that in $K^c(x_1, x_2)$. The $K^m(x_1, x_2)$ is our discrete approximation of $K^c(x_1, x_2)$. To understand this fact, it is convenient to first specify the discrete approximation for an arbitrary integral. Let $g(t)$ be an arbitrary function, and we consider its integration with respect to $\Lambda_0(t)$. According to the given grid, the integral can be approximated by the discretized summation, i.e.,

$$\int_0^\tau g(s)d\Lambda_0(s) \approx \sum_{j=1}^m g(t_j)(\Lambda_0(t_j) - \Lambda_0(t_{j-1})). \tag{1.32}$$

The right-hand side term is the discrete approximation of the left-hand side integral. In our specific case, recall that $K^c(x_1, x_2)$ consists of integrals in the left-hand side form, while $K^m(x_1, x_2)$ consists of its corresponding right-hand side summation. Moreover, as in general, the discretized summation converges to the integral, when the grid gets finer and finer, one may expect that $K^m(x_1, x_2)$ converges to $K^c(x_1, x_2)$. This is summarized in Proposition 4. See the proof in Appendix A for details.

Proposition 4 *Let $m \to \infty$, we have pointwise convergence of the covariance kernels*

$$\lim_{m\to\infty} K^m(x_1, x_2) = K^c(x_1, x_2).$$

Therefore, for a given data, if we take the grid t_j's at the observed uncensored durations, where the Breslow estimator jumps, the estimated $K^m(x_1, x_2)$ equals

to the estimated $K^c(x_1, x_2)$. It then suffices to derive the PCD of the estimated $K^m(x_1, x_2)$. Although the PCD argument for $K^m(x_1, x_2)$ is the same with that for $K^c(x_1, x_2)$ in section 1.3, it becomes simpler because we only need to deal with one estimation shift. We discuss briefly the PCD of $K^m(x_1, x_2)$ in the remaining section.

Let us define

$$H(x) := \boldsymbol{v}^T \boldsymbol{H}(x) \boldsymbol{v},$$

then the first part of $K^m(x_1, x_2)$ is $H(x_1 \wedge x_2)$. Its eigenfunctions are

$$h_k(x) := \varphi_k \left(H(x)/H(\infty) \right), \ k = 1, 2, \cdots,$$

with the same associated eigenvalues $\{\mu_k\}_{k=1}^{\infty}$. Then $\{h_k(x)\}_{k=1}^{\infty}$ forms an orthonormal basis of a subspace of $L^2(\mathbb{R}, H(x)/H(\infty))$, the Hilbert space of all square integrable functions on $\mathbb{R}$ with the inner product

$$\langle \rho, g \rangle := \int_{-\infty}^{\infty} \rho(x) g(x) \frac{H(dx)}{H(\infty)}.$$

We write the eigenfunctions in matrix form

$$\boldsymbol{h}(x) = \begin{pmatrix} h_1(x) \\ h_2(x) \\ \vdots \end{pmatrix},$$

then the decomposition of the first part is

$$H(\infty)^{-1} H(x_1 \wedge x_2) = \boldsymbol{h}^T(x_1) \boldsymbol{\mu} \boldsymbol{h}(x_2). \tag{1.33}$$

The next step is to decompose the estimation shift, which is $\boldsymbol{v}^T \boldsymbol{G}^T(x_1) \boldsymbol{A}^{-1} \boldsymbol{G}(x_2) \boldsymbol{v}$ in a compact matrix form, on the basis $\boldsymbol{h}(x)$. Define the coefficients of $\boldsymbol{v}^T \boldsymbol{G}^T(x)$ as a $\infty \times (m+1)$ matrix

$$\boldsymbol{\delta_g} := \begin{pmatrix} \langle \boldsymbol{v}^T \boldsymbol{G}^T, h_1 \rangle \\ \langle \boldsymbol{v}^T \boldsymbol{G}^T, h_2 \rangle \\ \vdots \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{m} \langle G_{1j}, h_1 \rangle & \cdots & \sum_{j=1}^{m} \langle G_{(m+1)j}, h_1 \rangle \\ \sum_{j=1}^{m} \langle G_{1j}, h_2 \rangle & \cdots & \sum_{j=1}^{m} \langle G_{(m+1)j}, h_2 \rangle \\ \vdots & & \vdots \end{pmatrix},$$

where G_{ij} is the ij'th element in the matrix $\boldsymbol{G}$. Then, we have

$$\boldsymbol{v}^T\boldsymbol{G}^T(x) = \boldsymbol{h}^T(x)\boldsymbol{\delta_g}, \tag{1.34}$$

and the "sandwich" form of the estimation shift is

$$\boldsymbol{v}^T\boldsymbol{G}^T(x_1)\boldsymbol{A}^{-1}\boldsymbol{G}(x_2)\boldsymbol{v} = \boldsymbol{h}^T(x_1)\boldsymbol{\delta_g}\boldsymbol{A}^{-1}\boldsymbol{\delta_g}^T\boldsymbol{h}(x_2). \tag{1.35}$$

Together with (1.33), we have

$$\begin{aligned} K^m(x_1,x_2) &= \boldsymbol{h}^T(x_1)\left(H(\infty)\boldsymbol{\mu}\right)\boldsymbol{h}(x_2) - \boldsymbol{h}^T(x_1)\left(\boldsymbol{\delta_g}\boldsymbol{A}^{-1}\boldsymbol{\delta_g}^T\right)\boldsymbol{h}(x_2) \\ &= \boldsymbol{h}^T(x_1)\left(H(\infty)\boldsymbol{\mu} - \boldsymbol{\delta_g}\boldsymbol{A}^{-1}\boldsymbol{\delta_g}^T\right)\boldsymbol{h}(x_2). \end{aligned} \tag{1.36}$$

Now the matrix need to be diagonalized is

$$\boldsymbol{Q} := H(\infty)\boldsymbol{\mu} - \boldsymbol{\delta_g}\boldsymbol{A}^{-1}\boldsymbol{\delta_g}^T.$$

Let an orthonormal matrix $\boldsymbol{P}$ and a diagonal matrix

$$\boldsymbol{\nu} = \begin{pmatrix} \nu_1 & & 0 \\ & \nu_2 & \\ 0 & & \ddots \end{pmatrix}$$

be the matrixes of eigenvectors and eigenvalues of $\boldsymbol{Q}$, such that

$$\boldsymbol{Q} = \boldsymbol{P}\boldsymbol{\nu}\boldsymbol{P}^T. \tag{1.37}$$

Then

$$K^m(x_1,x_2) = \boldsymbol{h}^T(x_1)\boldsymbol{P}\boldsymbol{\nu}\boldsymbol{P}^T\boldsymbol{h}(x_2). \tag{1.38}$$

This is the PCD of $K^m(x_1,x_2)$, and its eigenvalues and eigenfunctions are given by

$$\boldsymbol{\nu} \quad \text{and} \quad \boldsymbol{P}^T\boldsymbol{h}(x).$$

In summary, for a given set of data, we proceed as follows to apply the smooth tests and the Cramér-von Mises test. First, take the grid t_j's at the observed uncensored durations, and obtain the estimated $K^m(x_1,x_2)$ according to this grid.

Second, compute the eigenvalues and eigenfunctions of the estimated $K^m(x_1, x_2)$ through the procedure discussed above in this section. The eigenvalues and eigenfunctions in Corollary 2 and 3 will be approximated by the ones of the estimated $K^m(x_1, x_2)$. How to estimate $K^m(x_1, x_2)$ and diagonalize the infinite-dimensional matrix $\boldsymbol{Q}$ is discussed in next section.

1.5 Numerical Approximation

In order to actually compute $\boldsymbol{P}$ and $\boldsymbol{\nu}$, which are the eigenvectors and eigenvalues of the infinite-dimensional matrix $\boldsymbol{Q}$, we follow the approximation method suggested by DKT, i.e., to approximate the first q components by computing the eigenvectors and eigenvalues of the finite-dimensional matrix $\boldsymbol{Q}_q$ with estimation, which is

$$\hat{\boldsymbol{Q}}_q = \hat{H}(\infty)\boldsymbol{\mu}_q - \hat{\boldsymbol{\delta}}_{\boldsymbol{g}_q}\hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{\delta}}_{\boldsymbol{g}_q}^{\ T}, \tag{1.39}$$

with

$$\boldsymbol{\mu}_q = \begin{pmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_q \end{pmatrix},$$

and

$$\hat{\boldsymbol{\delta}}_{\boldsymbol{g}_q} = \begin{pmatrix} \langle \boldsymbol{v}^T\hat{\boldsymbol{G}}^T, \hat{h}_1 \rangle \\ \vdots \\ \langle \boldsymbol{v}^T\hat{\boldsymbol{G}}^T, \hat{h}_q \rangle \end{pmatrix}.$$

Here $\hat{H}$, $\hat{\boldsymbol{G}}$, $\hat{\boldsymbol{A}}$ and $\hat{h}_k$ are estimators of H, $\boldsymbol{G}$, $\boldsymbol{A}$ and h_k, $k = 1, \cdots, q$. All the terms that need to be estimated, except for β and $\Lambda_0(t)$, are

$$F_X(x),\ \mathbb{E}(\Delta A_j \mid X = x),\ j = 1, \cdots, m.$$

The $F_X(x)$, which is the distribution function of X, can be replaced by the empirical one. For consistent estimation of $\mathbb{E}(\Delta A_j \mid X = x)$, we assume the following additional assumption.

(A5). C is independent of X.

Then

$$\begin{aligned}\mathbb{E}(\Delta A_j \mid X = x) &= \mathbb{E}(Y(t_j) \mid X = x)e^{\beta^T x}(\Lambda_0(t_j) - \Lambda_0(t_{j-1})) \\ &= P(T \geq t_j, C \geq t_j \mid X = x)e^{\beta^T x}(\Lambda_0(t_j) - \Lambda_0(t_{j-1})) \\ &= P(T \geq t_j \mid X = x)P(C \geq t_j \mid X = x)e^{\beta^T x}(\Lambda_0(t_j) - \Lambda_0(t_{j-1})) \\ &= exp(-\Lambda_0(t_j)e^{\beta^T x})P(C \geq t_j)e^{\beta^T x}(\Lambda_0(t_j) - \Lambda_0(t_{j-1})). \qquad (1.40)\end{aligned}$$

The second to the last equation follows from assumption (A1), and the last equation follows from assumption (A5). A natural consistent estimator is

$$exp(-\hat{\Lambda}_0(t_j)e^{\hat{\beta}^T x})(1 - \hat{F}_c(t_j-))e^{\hat{\beta}^T x}(\hat{\Lambda}_0(t_j) - \hat{\Lambda}_0(t_{j-1})),$$

where $\hat{F}_c$ is the Kaplan-Meier estimator of the distribution function of the censoring time C. We will immediately obtain uniform consistent $\hat{H}$, $\hat{\boldsymbol{G}}$, $\hat{h}_k$ and consistent $\hat{\boldsymbol{A}}$ from the above estimators of $F_X(x)$ and $\mathbb{E}(\Delta A_j \mid X = x)$, $j = 1, \cdots, m$. For example,

$$\hat{H}(x) = \sum_{j=1}^{m} \int_{-\infty}^{x} exp(-\hat{\Lambda}_0(t_j)e^{\hat{\beta}^T u})(1 - \hat{F}_c(t_j-))e^{\hat{\beta}^T u}(\hat{\Lambda}_0(t_j) - \hat{\Lambda}_0(t_{j-1}))\hat{F}_X(du),$$

and

$$\hat{h}_k(x) = \varphi_k\left(\hat{H}(x)/\hat{H}(\infty)\right).$$

For $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{A}}$, it is similar but lengthy, thus we omit them here.

Now consider the infinite-dimensional matrix

$$\tilde{\boldsymbol{Q}} = \begin{pmatrix} \hat{\boldsymbol{Q}}_q & 0 \\ 0 & \hat{H}(\infty)\boldsymbol{\mu}_{\infty-q} \end{pmatrix},$$

with

$$\boldsymbol{\mu}_{\infty-q} = \begin{pmatrix} \mu_{q+1} & & 0 \\ & \mu_{q+2} & \\ 0 & & \ddots \end{pmatrix}.$$

Suppose that $\hat{\boldsymbol{Q}}_q$ has eigenvalues $\tilde{\nu}_1, \cdots, \tilde{\nu}_q$ and eigenvectors $\tilde{\boldsymbol{p}}_k = (p_{1k}, \cdots, p_{qk})^T$, $1 \leq k \leq q$. Then $\tilde{\boldsymbol{Q}}$ has eigenvalues $\tilde{\nu}_1, \cdots, \tilde{\nu}_q, \hat{H}(\infty)\mu_{q+1}, \hat{H}(\infty)\mu_{q+2}, \cdots$ and eigenvectors

$$\boldsymbol{p}_{k0} = (\tilde{\boldsymbol{p}}_k, 0, 0, \cdots)^T, \text{ for } 1 \leq k \leq q$$

and

$$\boldsymbol{p}_{k0} = \boldsymbol{e}_k, \text{ which is the } k\text{-th unit vector for } k > q.$$

We use these eigenvalues and eigenvectors of $\tilde{\boldsymbol{Q}}$ to approximate the ones of $\boldsymbol{Q}$.

Finally, the smooth test statistic in (1.25) will be approximated by, for $r < q$,

$$S_{nr}^2 \approx \sum_{k=1}^{r} \left[\tilde{\nu}_k^{-1/2} \hat{H}(\infty)^{-1} \int_{-\infty}^{\infty} \boldsymbol{p}_{k0}^T \hat{\boldsymbol{h}}(x) \hat{c}(x) \hat{H}(dx) \right]^2. \tag{1.41}$$

The distribution of CvM_∞ in (1.27) will be approximated by

$$CvM_{\infty q} := \sum_{k=1}^{q} \tilde{\nu}_k Z_k^2 + c_1 \chi_{c_2}^2, \tag{1.42}$$

where the Z_k's are independent standard normal variables and $\chi_{c_2}^2$ is a suitable chi-square variable with degree of freedom c_2 and c_1 is such that $CvM_{\infty q}$ has the same mean and variance as CvM_∞.

1.6 Simulation

We consider the following DGPs for a simulation study. We take $\Lambda_0(t) = t$, and X from uniform distribution $U(0, 1)$ in all cases. The censoring time variable is drawn from a uniform distribution such that the percentage of censorship is around 30%. We repeat the DGPs 1000 times to find the size and power of the tests.

DGP1: The Cox Model

$$\lambda(t \mid X) = exp(2X).$$

DGP2: Missing Variable

$$\lambda(t \mid X) = exp(2X - 5X^2).$$

DGP3: Wrong Link Function

$$\lambda(t \mid X) = 1 + sin(4X).$$

We run three tests for each case, including the smooth test S_{n1}^2, which is only based on the first component, the omnibus Cramér-von Mises test, and the Schoenfeld chi-square test for comparison. For the Schoenfeld test, we consider three different partitions of the covariable range: "p2" indicates the partition of the covariable range at its median, which is 1/2; "p3" indicates the partition at 1/3 and 2/3; and "p4" indicates the partition at 1/4, 2/4 and 3/4. The partition in the time axis is at the sample median of the duration times, in all "p2", "p3" and "p4". For our omnibus Cramér-von Mises test, we approximate the first q eigenvalues of the infinite-dimensional matrix, and we take $q = 25, 50, 100$ in this simulation. We also run the first component test, because the data-driven technique for the smooth test, e.g., Ledwina (1994), usually suggests using only the first component. Table 1.1 shows the simulation results.

The power of the Schoenfeld test varies a lot with different partitions, while different values of q do not cause too much variation in the power of the first component test and the Cramér-von Mises test. For both alternatives, the Cramér-von Mises test has larger power compared to the Schoenfeld test, meanwhile, the first component test behaves slightly better than the Cramér-von Mises omnibus test.

Table 1.1: Estimated size and power of first component test, omnibus CvM test and Schoenfeld test at 5%

	DGP1: Cox								
	S_{n1}^2			CvM			Sch		
	$q = 25$	50	100	$q = 25$	50	100	p2	p3	p4
$n = 50$	0.042	0.045	0.044	0.036	0.046	0.056	0.045	0.064	0.081
$n = 75$	0.053	0.054	0.055	0.037	0.043	0.051	0.049	0.065	0.066
$n = 100$	0.052	0.050	0.049	0.039	0.042	0.050	0.058	0.043	0.060
$n = 150$	0.049	0.052	0.049	0.033	0.040	0.047	0.053	0.052	0.067
$n = 200$	0.050	0.047	0.049	0.038	0.045	0.046	0.063	0.058	0.070

	DGP2: Missing Variable								
	S_{n1}^2			CvM			Sch		
	$q = 25$	50	100	$q = 25$	50	100	p2	p3	p4
$n = 50$	0.272	0.268	0.263	0.241	0.237	0.232	0.065	0.158	0.177
$n = 75$	0.497	0.493	0.488	0.447	0.444	0.438	0.063	0.225	0.250
$n = 100$	0.637	0.628	0.633	0.599	0.583	0.580	0.060	0.306	0.345
$n = 150$	0.835	0.834	0.833	0.828	0.820	0.813	0.082	0.445	0.525
$n = 200$	0.941	0.939	0.942	0.932	0.928	0.931	0.075	0.591	0.679

	DGP3: Wrong Link Function								
	S_{n1}^2			CvM			Sch		
	$q = 25$	50	100	$q = 25$	50	100	p2	p3	p4
$n = 50$	0.343	0.343	0.347	0.297	0.293	0.281	0.064	0.181	0.215
$n = 75$	0.563	0.567	0.562	0.522	0.521	0.513	0.065	0.261	0.310
$n = 100$	0.687	0.684	0.685	0.678	0.671	0.669	0.056	0.325	0.390
$n = 150$	0.887	0.884	0.888	0.876	0.871	0.869	0.047	0.479	0.566
$n = 200$	0.964	0.962	0.962	0.964	0.961	0.962	0.044	0.637	0.725

1.7 Concluding Remarks

We have restricted the data on $[0, \tau]$ for convenience. In fact, all the results in section 1.3 and the approximate procedure for a given data in section 1.4 still hold if we use all the observations on $[0, \infty)$. We only need to change the τ's to ∞, and change assumption (A2) to

(A2)*. For each $\tau < \infty$, $P\{Y(\tau) = 1\} > 0$.

So far we have assumed real-valued X. However, there is no easy extension to the multivariate case, since no explicit PCD is available for a multivariate process with possibly dependent components. One possibility is to consider the CUSUM process of the martingale residuals on each component of X, and test the specification of each component of X one by one.

To sum up, we develop an orthogonal decomposition of the omnibus Cramér-von Mises test for the specification of the covariate effect in proportional hazard models. From this decomposition, we not only can approximate the limit distribution of the omnibus test numerically, but more importantly, we may feel free to reweight the components to obtain more powerful smooth tests and directional tests. The extension to general semiparametric transformation models is possible, and will be discussed elsewhere.

1.8 Appendix

Appendix A: Proofs

Proof of Theorem 1, Corollary 2 and Corollary 3:

(1.23) follows from Kac-Siegert and the fact that $\hat{c}_\infty$ is a Gaussian process. Since

$\hat{H}_0$, $\hat{f}_k$ are uniform consistent and $\hat{\lambda}_k$, $\hat{\boldsymbol{N}}_k$ are consistent, we have

$$\int_{-\infty}^{\infty} \hat{\boldsymbol{N}}_k^T \hat{\boldsymbol{f}}(x)\hat{c}(x)\hat{H}_c(dx) = \int_{-\infty}^{\infty} \boldsymbol{N}_k^T \boldsymbol{f}(x)\hat{c}(x)H_c(dx) + o_p(1).$$

For a continuous F_X, which leads to a continuous H_c, the weak convergence of the coefficients (1.24) follows from the weak convergence of $\hat{c}$ in (1.13) and the continuous mapping theorem. Since $\{f_k(x)\}_{k=1}^{\infty}$ is an orthonormal basis, we have

$$\hat{c}_\infty(x) = \sum_{k=1}^{\infty} \langle \hat{c}_\infty, f_k \rangle f_k(x). \tag{A.1}$$

Then

$$\begin{aligned} CvM_\infty &= \int_{-\infty}^{\infty} \left(\sum_{k=1}^{\infty} \langle \hat{c}_\infty, f_k \rangle f_k(x) \right)^2 H_c(dx)/H_c(\infty) \\ &= \sum_{k=1}^{\infty} \langle \hat{c}_\infty, f_k \rangle^2 \\ &= \langle \boldsymbol{f}, \hat{c}_\infty \rangle^T \langle \boldsymbol{f}, \hat{c}_\infty \rangle \\ &= \langle \boldsymbol{f}, \hat{c}_\infty \rangle^T \boldsymbol{N}\boldsymbol{N}^T \langle \boldsymbol{f}, \hat{c}_\infty \rangle \\ &= \langle \boldsymbol{N}^T\boldsymbol{f}, \hat{c}_\infty \rangle^T \langle \boldsymbol{N}^T\boldsymbol{f}, \hat{c}_\infty \rangle. \end{aligned}$$

The second to the last equation follows from the fact that $\boldsymbol{N}$ is an orthogonal matrix. By (1.26), we have Corollary 3.

Proof of Proposition 4:

First, we need to get the inverse of $\boldsymbol{A}$. To do this, write $\boldsymbol{A}$ as the following block matrix, by separating the nonparametric part with the parametric part,

$$\boldsymbol{A} = \left(\begin{array}{c|c} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \hline \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{array} \right) = \left(\begin{array}{ccc|c} \mathbb{E}(\Delta A_1) & \cdots & 0 & \mathbb{E}(\Delta A_1 X) \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \mathbb{E}(\Delta A_m) & \mathbb{E}(\Delta A_m X) \\ \hline \mathbb{E}(\Delta A_1 X) & \cdots & \mathbb{E}(\Delta A_m X) & \sum_{j=1}^m \mathbb{E}(\Delta A_j X^2) \end{array} \right). \tag{A.2}$$

Define

$$\Sigma^m := \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} = \sum_{j=1}^{m} \left(\mathbb{E}\left(\Delta A_j X^2\right) - \frac{\mathbb{E}\left(\Delta A_j X\right)^2}{\mathbb{E}\left(\Delta A_j\right)} \right).$$

Then by the inverse formula of block matrix

$$\boldsymbol{A}^{-1} = (\Sigma^m)^{-1}\left(\begin{array}{cccc|c} \frac{\Sigma^m}{\mathbb{E}(\Delta A_1)} + \frac{\mathbb{E}(\Delta A_1 X)^2}{\mathbb{E}(\Delta A_1)^2} & \frac{\mathbb{E}(\Delta A_1 X)\mathbb{E}(\Delta A_2 X)}{\mathbb{E}(\Delta A_1)\mathbb{E}(\Delta A_2)} & \cdots & \frac{\mathbb{E}(\Delta A_1 X)\mathbb{E}(\Delta A_m X)}{\mathbb{E}(\Delta A_1)\mathbb{E}(\Delta A_m)} & -\frac{\mathbb{E}(\Delta A_1 X)}{\mathbb{E}(\Delta A_1)} \\ \frac{\mathbb{E}(\Delta A_1 X)\mathbb{E}(\Delta A_2 X)}{\mathbb{E}(\Delta A_1)\mathbb{E}(\Delta A_2)} & \frac{\Sigma^m}{\mathbb{E}(\Delta A_2)} + \frac{\mathbb{E}(\Delta A_2 X)^2}{\mathbb{E}(\Delta A_2)^2} & \cdots & \frac{\mathbb{E}(\Delta A_2 X)\mathbb{E}(\Delta A_m X)}{\mathbb{E}(\Delta A_2)\mathbb{E}(\Delta A_m)} & -\frac{\mathbb{E}(\Delta A_2 X)}{\mathbb{E}(\Delta A_2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\mathbb{E}(\Delta A_1 X)\mathbb{E}(\Delta A_m X)}{\mathbb{E}(\Delta A_1)\mathbb{E}(\Delta A_m)} & \frac{\mathbb{E}(\Delta A_2 X)\mathbb{E}(\Delta A_m X)}{\mathbb{E}(\Delta A_2)\mathbb{E}(\Delta A_m)} & \cdots & \frac{\Sigma^m}{\mathbb{E}(\Delta A_m)} + \frac{\mathbb{E}(\Delta A_m X)^2}{\mathbb{E}(\Delta A_m)^2} & -\frac{\mathbb{E}(\Delta A_m X)}{\mathbb{E}(\Delta A_m)} \\ \hline -\frac{\mathbb{E}(\Delta A_1 X)}{\mathbb{E}(\Delta A_1)} & -\frac{\mathbb{E}(\Delta A_2 X)}{\mathbb{E}(\Delta A_2)} & \cdots & -\frac{\mathbb{E}(\Delta A_m X)}{\mathbb{E}(\Delta A_m)} & 1 \end{array}\right)$$

$$= \left(\begin{array}{cccc|c} \frac{1}{\mathbb{E}(\Delta A_1)} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{\mathbb{E}(\Delta A_2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\mathbb{E}(\Delta A_m)} & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 \end{array}\right)$$

$$+(\Sigma^m)^{-1}\begin{pmatrix} -\frac{\mathbb{E}(\Delta A_1 X)}{\mathbb{E}(\Delta A_1)} \\ \vdots \\ -\frac{\mathbb{E}(\Delta A_m X)}{\mathbb{E}(\Delta A_m)} \\ 1 \end{pmatrix}\left(-\frac{\mathbb{E}(\Delta A_1 X)}{\mathbb{E}(\Delta A_1)}, \cdots, -\frac{\mathbb{E}(\Delta A_m X)}{\mathbb{E}(\Delta A_m)}, 1 \right).$$

Let us denote the above components of $\boldsymbol{A}^{-1}$ as

$$\boldsymbol{B} := \left(\begin{array}{cccc|c} \frac{1}{\mathbb{E}(\Delta A_1)} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{\mathbb{E}(\Delta A_2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\mathbb{E}(\Delta A_m)} & 0 \\ \hline 0 & 0 & \cdots & 0 & 0 \end{array}\right),$$

and

$$\boldsymbol{C} := \left(-\frac{\mathbb{E}(\Delta A_1 X)}{\mathbb{E}(\Delta A_1)}, \cdots, -\frac{\mathbb{E}(\Delta A_m X)}{\mathbb{E}(\Delta A_m)}, 1 \right).$$

Thus,

$$\boldsymbol{A}^{-1} = \boldsymbol{B} + (\Sigma^m)^{-1}\boldsymbol{C}^T\boldsymbol{C}. \tag{A.3}$$

Since

$$\boldsymbol{v}^T\boldsymbol{G}^T(x) = \left(\mathbb{E}(\mathbb{1}_{\{X\le x\}}\Delta A_1), \cdots, \mathbb{E}(\mathbb{1}_{\{X\le x\}}\Delta A_m), \sum_{j=1}^{m}\mathbb{E}(\mathbb{1}_{\{X\le x\}}\Delta A_j X)\right),$$

by simple computations of matrix, we have

$$\boldsymbol{v}^T\boldsymbol{G}^T(x_1)\boldsymbol{B}\boldsymbol{G}(x_2)\boldsymbol{v} = \sum_{j=1}^{m} \frac{\mathbb{E}(\mathbb{1}_{\{X\le x_1\}}\Delta A_j)\mathbb{E}(\mathbb{1}_{\{X\le x_2\}}\Delta A_j)}{\mathbb{E}(\Delta A_j)},$$

and

$$\boldsymbol{v}^T\boldsymbol{G}^T(x)\boldsymbol{C}^T = \sum_{j=1}^{m}\left(\mathbb{E}(\mathbb{1}_{\{X\le x\}}\Delta A_j X) - \frac{\mathbb{E}(\mathbb{1}_{\{X\le x\}}\Delta A_j)\mathbb{E}(\Delta A_j X)}{\mathbb{E}(\Delta A_j)}\right).$$

Then

$$\begin{aligned}\boldsymbol{v}^T\boldsymbol{G}^T(x_1)\boldsymbol{A}^{-1}\boldsymbol{G}(x_2)\boldsymbol{v} &= \boldsymbol{v}^T\boldsymbol{G}^T(x_1)\boldsymbol{B}\boldsymbol{G}(x_2)\boldsymbol{v} + (\Sigma^m)^{-1}\boldsymbol{v}^T\boldsymbol{G}^T(x_1)\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{G}(x_2)\boldsymbol{v}\\ &= \sum_{j=1}^{m}\frac{\mathbb{E}(\mathbb{1}_{\{X\le x_1\}}\Delta A_j)\mathbb{E}(\mathbb{1}_{\{X\le x_2\}}\Delta A_j)}{\mathbb{E}(\Delta A_j)}\\ &\quad + (\Sigma^m)^{-1}\left(\sum_{j=1}^{m}\left(\mathbb{E}(\mathbb{1}_{\{X\le x_1\}}\Delta A_j X) - \frac{\mathbb{E}(\mathbb{1}_{\{X\le x_1\}}\Delta A_j)\mathbb{E}(\Delta A_j X)}{\mathbb{E}(\Delta A_j)}\right)\right)\\ &\quad \times\left(\sum_{j=1}^{m}\left(\mathbb{E}(\mathbb{1}_{\{X\le x_2\}}\Delta A_j X) - \frac{\mathbb{E}(\mathbb{1}_{\{X\le x_2\}}\Delta A_j)\mathbb{E}(\Delta A_j X)}{\mathbb{E}(\Delta A_j)}\right)\right).\end{aligned} \tag{A.4}$$

Now take $m\to\infty$, we have the following limits.

$$\begin{aligned}&\lim_{m\to\infty}\sum_{j=1}^{m}\frac{\mathbb{E}(\mathbb{1}_{\{X\le x_1\}}\Delta A_j)\mathbb{E}(\mathbb{1}_{\{X\le x_2\}}\Delta A_j)}{\mathbb{E}(\Delta A_j)}\\ =\ &\lim_{m\to\infty}\sum_{j=1}^{m}\frac{\mathbb{E}(Y(t_j)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_1\}})\mathbb{E}(Y(t_j)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_2\}})}{\mathbb{E}(Y(t_j)e^{\beta_0^T X})}(\Lambda_0(t_j)-\Lambda_0(t_{j-1}))\\ =\ &\int_0^\tau\frac{\mathbb{E}(Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_1\}})\mathbb{E}(Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_2\}})}{\mathbb{E}(Y(s)e^{\beta_0^T X})}d\Lambda_0(s).\end{aligned}$$

$$\begin{aligned}&\lim_{m\to\infty}\Sigma^m\\ =\ &\lim_{m\to\infty}\sum_{j=1}^{m}\left(\mathbb{E}\left(Y(t_j)e^{\beta_0^T X}X^2\right)(\Lambda_0(t_j)-\Lambda_0(t_{j-1})) - \frac{\left(\mathbb{E}\left(Y(t_j)e^{\beta_0^T X}X\right)(\Lambda_0(t_j)-\Lambda_0(t_{j-1}))\right)^2}{\mathbb{E}\left(Y(t_j)e^{\beta_0^T X}\right)(\Lambda_0(t_j)-\Lambda_0(t_{j-1}))}\right)\\ =\ &\int_0^\tau\mathbb{E}\left(Y(s)e^{\beta_0^T X}X^2\right)d\Lambda_0(s) - \int_0^\tau\frac{\mathbb{E}\left(Y(t_j)e^{\beta_0^T X}X\right)^2}{\mathbb{E}\left(Y(t_j)e^{\beta_0^T X}\right)}d\Lambda_0(s)\\ =\ &\mathbb{E}\int_0^\tau\left(X-\frac{\mathbb{E}\left(Y(s)e^{\beta_0^T X}X\right)}{\mathbb{E}\left(Y(s)e^{\beta_0^T X}\right)}\right)^2 Y(s)e^{\beta_0^T X}d\Lambda_0(s)\\ =\ &\Sigma(\beta_0,\tau).\end{aligned}$$

$$
\begin{aligned}
&\lim_{m\to\infty}\sum_{j=1}^{m}\left(\mathbb{E}(\mathbb{1}_{\{X\le x\}}\Delta A_j X)-\frac{\mathbb{E}(\mathbb{1}_{\{X\le x\}}\Delta A_j)\mathbb{E}(\Delta A_j X)}{\mathbb{E}(\Delta A_j)}\right)\\
=&\lim_{m\to\infty}\sum_{j=1}^{m}\left(\mathbb{E}(Y(t_j)e^{\beta_0^T X}\mathbb{1}_{\{X\le x\}}X)-\frac{\mathbb{E}(Y(t_j)e^{\beta_0^T X}\mathbb{1}_{\{X\le x\}})\mathbb{E}(Y(t_j)e^{\beta_0^T X}X)}{\mathbb{E}(Y(t_j)e^{\beta_0^T X})}\right)(\Lambda_0(t_j)-\Lambda_0(t_{j-1}))\\
=&\int_0^\tau \mathbb{E}\left(Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x\}}\left(X-\frac{\mathbb{E}(Y(s)e^{\beta_0^T X}X)}{\mathbb{E}(Y(s)e^{\beta_0^T X})}\right)\right)d\Lambda_0(s)\\
=&\mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x\}}\left(X-\frac{\mathbb{E}(Y(s)e^{\beta_0^T X}X)}{\mathbb{E}(Y(s)e^{\beta_0^T X})}\right)d\Lambda_0(s)\right).
\end{aligned}
$$

Combining these limits, we have

$$
\begin{aligned}
&\lim_{m\to\infty}\boldsymbol{v}^T\boldsymbol{G}^T(x_1)\boldsymbol{A}^{-1}\boldsymbol{G}(x_2)\boldsymbol{v}\\
&=\int_0^\tau\frac{\mathbb{E}(Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_1\}})\mathbb{E}(Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_2\}})}{\mathbb{E}(Y(s)e^{\beta_0^T X})}d\Lambda_0(s)\\
&+\mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_1\}}\left(X-\frac{\mathbb{E}(Y(s)e^{\beta_0^T X}X)}{\mathbb{E}(Y(s)e^{\beta_0^T X})}\right)d\Lambda_0(s)\right)\\
&\quad\times\Sigma(\beta_0,\tau)^{-1}\mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_2\}}\left(X-\frac{\mathbb{E}(Y(s)e^{\beta_0^T X}X)}{\mathbb{E}(Y(s)e^{\beta_0^T X})}\right)d\Lambda_0(s)\right). \quad \text{(A.5)}
\end{aligned}
$$

A comment on Σ^m is worth to be made. By Cauchy-Schwarz inequality and by taking the two variables in the inequality as $\sqrt{\Delta A_j}X$ and $\sqrt{\Delta A_j}$, we can show that Σ^m is positive (positive definite if X is a vector). Σ^m is the asymptotic variance of $\hat{\beta}_0$, and it is automatically positive definite. We can conclude that its limit $\Sigma(\beta_0,\tau)$ is nonnegative definite, but not guaranteed to be positive definite. That is the reason why we need assumption (A4) in the continuous case.

Together with

$$
\begin{aligned}
\lim_{m\to\infty}\boldsymbol{v}^T\boldsymbol{H}(x_1\wedge x_2)\boldsymbol{v}&=\lim_{m\to\infty}\sum_{j=1}^{m}\mathbb{E}(\mathbb{1}_{\{X\le x_1\wedge x_2\}}\Delta A_j)\\
&=\lim_{m\to\infty}\sum_{j=1}^{m}\mathbb{E}\left(Y(t_j)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_1\wedge x_2\}}(\Lambda_0(t_j)-\Lambda_0(t_{j-1}))\right)\\
&=\mathbb{E}\left(\int_0^\tau Y(s)e^{\beta_0^T X}\mathbb{1}_{\{X\le x_1\wedge x_2\}}d\Lambda_0(s)\right),
\end{aligned}
$$

we have

$$\lim_{m\to\infty}K^m(x_1,x_2)=K^c(x_1,x_2).$$

It is easy to see that $K^m(x_1, x_2)$ is the discrete approximation of $K^c(x_1, x_2)$ by discretizing the integrals with respect to $\Lambda_0(t)$ into summations.

Appendix B: Another Estimation based on Discrete Martingale Increments

We introduce a different estimation procedure that is based on moment conditions of discrete martingale increments. The difference from the existing approaches is how we deal with the nonparametric $\Lambda_0(t)$. Specifically, our approach consists of two steps: (i) consider $\Lambda_0(t)$ as a simple function that only jumps at finite points in a given grid, and estimate this discrete $\Lambda_0(t)$ and β through moment conditions of martingale increments, (ii) go to the limit by taking the grid of points finer and finer. In another word, we approximate $\Lambda_0(t)$ by simple functions and approximate estimation of $\Lambda_0(t)$ and β by estimations of simple functions and β.

Let $0 = t_0 < t_1 < t_2 < \cdots < t_m = \tau$ be a finite grid on time. To approximate $\Lambda_0(t)$ by simple functions, we introduce the following discrete assumption.

(B1). $\Lambda_0(t)$ is a step function on $[0, \tau]$ and only has jumps on t_j's, $j = 1, \cdots, m$.

Note that assuming (B1) is equivalent to assume discrete distribution function, since in the discrete case, the cumulative hazard function has the same atoms as the distribution function. Under (B1), the unspecified baseline hazard is fully characterized by m jump sizes together with the given grid. Hence, the semiparametric estimation problem of the Cox model turns out to be a parametric one with $m + 1$ parameters, including m jump sizes and the one dimensional β. To estimate these parameters, we construct moment conditions of martingale increments that are generated by the given grid.

The martingale increments, generated by the given grid, are, for $j = 1, 2, \cdots, m$,

$$\begin{aligned} M(t_j) - M(t_{j-1}) &= N(t_j) - N(t_{j-1}) - (A(t_j) - A(t_{j-1})) \\ &= N(t_j) - N(t_{j-1}) - \int_{t_{j-1}}^{t_j} Y(s) e^{\beta^T X} d\Lambda_0(s) \\ &= N(t_j) - N(t_{j-1}) - e^{\beta^T X}(\Lambda_0(t_j \wedge Z) - \Lambda_0(t_{j-1} \wedge Z)). \end{aligned} \tag{B.1}$$

Since Z is random, to characterize these increments, it requires infinitely many parameters that measure the differences of $\Lambda_0(t)$. However, under the discrete assumption (B1), these increments equal to

$$\varepsilon_j := N(t_j) - N(t_{j-1}) - Y(t_j) e^{\beta^T X}(\Lambda_0(t_j) - \Lambda_0(t_{j-1})),\ j = 1, 2, \cdots, m.$$

Now they are defined by finite parameters, including β and the m jump sizes $\Lambda_0(t_j) - \Lambda_0(t_{j-1}),\ j = 1, 2, \cdots, m$. To simplify notation, let us define log-transformation of the jump sizes as

$$\eta_j := \ln(\Lambda_0(t_j) - \Lambda_0(t_{j-1})),\ j = 1, 2, \cdots, m, \tag{B.2}$$

then

$$\varepsilon_j = N(t_j) - N(t_{j-1}) - Y(t_j) e^{\beta^T X + \eta_j},\ j = 1, 2, \cdots, m. \tag{B.3}$$

Let us write the parameters needed to estimate as a vector

$$\boldsymbol{\theta} := (\eta_1, \eta_2, \cdots, \eta_m, \beta)^T,$$

and the martingale increments also as a vector

$$\boldsymbol{\varepsilon} := (\varepsilon_1, \cdots, \varepsilon_m)^T,$$

where we suspends the dependence of m to simplify notation, but should always keep it in mind that they all depend on the grid we take.

We can form estimation equations from the condition

$$\mathbb{E}(\boldsymbol{\varepsilon} \mid X) = 0. \tag{B.4}$$

For example, let $\boldsymbol{W}$ be a valid instrumental variable that is formed by some transformation of X, the parameter $\boldsymbol{\theta}$ can be estimated based on the moment conditions

$$\mathbb{E}(\boldsymbol{W}\boldsymbol{\varepsilon}) = 0. \tag{B.5}$$

One option consists of choosing $\boldsymbol{W}$ as the optimal instrumental variable in the sense that the instrumental variable estimator is asymptotically efficient, and the optimal instrumental variable takes the form of

$$\boldsymbol{W} = \mathbb{E}\left(\frac{\partial \boldsymbol{\varepsilon}^T}{\partial \boldsymbol{\theta}} \mid X\right) \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \mid X)^{-1}. \tag{B.6}$$

To find this $\boldsymbol{W}$, let us first simplify the notation by denoting the increments of the compensator in (B.3) as

$$\Delta A_j := Y(t_j)e^{\beta_0^T X + \eta_j},\ j = 1, 2, \cdots, m,$$

with the true value of β and η_i's. The first derivative is easy to compute, which is

$$\mathbb{E}\left(\frac{\partial \boldsymbol{\varepsilon}^T}{\partial \boldsymbol{\theta}} \mid X\right) = -\begin{pmatrix} \mathbb{E}(\Delta A_1 \mid X) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbb{E}(\Delta A_m \mid X) \\ \mathbb{E}(\Delta A_1 \mid X)X & \cdots & \mathbb{E}(\Delta A_m \mid X)X \end{pmatrix}.$$

For the variance of $\boldsymbol{\varepsilon}$, careful attention should be paid, since we are assuming discrete compensator. The variance of the martingale with discrete compensator has a different form from the one in the continuous case, for which we refer to Fleming and Harrington (1991) in section 2.5 for the continuous case and section 2.6 for the discrete case. In the discrete case, the conditional variances of the martingale increments are

$$\mathbb{E}(\varepsilon_j^2 \mid X) = \mathbb{E}(\Delta A_j(1 - \Delta A_j) \mid X),\ j = 1, 2, \cdots, m,$$

and the conditional variance and covariance matrix is

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \mid X) = \begin{pmatrix} \mathbb{E}(\Delta A_1(1 - \Delta A_1) \mid X) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbb{E}(\Delta A_m(1 - \Delta A_m) \mid X) \end{pmatrix}.$$

Note that the variance and covariance matrix is a diagonal matrix due to the martingale property. However, this discrete form of variance will be simplified, when going to the continuous time limit. Since

$$\mathbb{E}(\Delta A_j(1-\Delta A_j) \mid X) \approx \mathbb{E}(\Delta A_j \mid X),$$

the variance of martingale increment, in the continuous case, is expected to be the expectation of the increment of the compensator, i.e.,

$$\mathbb{E}((M(t_j)-M(t_{j-1}))^2 \mid X) = \mathbb{E}(A(t_j)-A(t_{j-1}) \mid X).$$

Therefore, if we approximate the variance and covariance matrix of $\boldsymbol{\varepsilon}$ by its continuous form, i.e.,

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \mid X) \approx \begin{pmatrix} \mathbb{E}(\Delta A_1 \mid X) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbb{E}(\Delta A_m \mid X) \end{pmatrix}, \tag{B.7}$$

we can take $\boldsymbol{W}$ as the approximative optimal weight for the instrumental variable estimator, which is a $(m+1)\times m$ matrix and is specified as

$$\boldsymbol{W} = -\begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \\ X & \cdots & X \end{pmatrix}. \tag{B.8}$$

For a sample of size n and $i = 1, \cdots, n$, denote the i's individual vector as $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{im})^T$ and the residual as $\hat{\boldsymbol{\varepsilon}}_i = (\hat{\varepsilon}_{i1}, \cdots, \hat{\varepsilon}_{im})^T$, with

$$\hat{\varepsilon}_{ij} = N_i(t_j) - N_i(t_{j-1}) - Y_i(t_j)e^{\hat{\beta}^T X_i + \hat{\eta}_j},\ j = 1, 2, \cdots, m, i = 1, \cdots, n.$$

Then the estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_m, \hat{\beta})^T$ is obtained by solving a $m+1$ simultaneous equations

$$\sum_{i=1}^{n} \boldsymbol{W}_i \hat{\boldsymbol{\varepsilon}}_i = \sum_{i=1}^{n} \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \\ X_i & \cdots & X_i \end{pmatrix} \begin{pmatrix} \hat{\varepsilon}_{i1} \\ \vdots \\ \hat{\varepsilon}_{im} \end{pmatrix} = 0. \tag{B.9}$$

The estimation equations (B.9) are actually the discrete version of the estimation equations (1.10) proposed by Chen, Jin and Ying (2002). However, the difference is that, in (B.9), the estimation equations are combined in a compact matrix form.

The estimator $\hat{\boldsymbol{\theta}}$ from (B.9) yields an estimator of β and $\Lambda_0(t)$ under assumption (B1). Note that the assumption (B1) is based on a given grid, and for different grids it describes different simple functions. Thus, for a given data, the estimation from (B.9) will be different if we take different grids. A special grid is worth to mention. Recall that the Breslow estimator of $\Lambda_0(t)$ is a step function and only jumps at the observed uncensored durations. Hence, given a data set, we could take the grid at the observed uncensored durations. With this special grid, the estimator $\hat{\boldsymbol{\theta}}$ from (B.9) coincides with the partial likelihood estimator of β and the Breslow estimator of $\Lambda_0(t)$.

Without conditioning on the data, we could also achieve the partial likelihood estimator and the Breslow estimator by our estimation procedure. This requires the second step of our approach: go to the limit by taking the grid finer and finer. When going to the continuous time limit, as the simple functions described in (B1) approximate the true $\Lambda_0(t)$, the estimations from (B.9) also approximate the partial likelihood estimation and the Breslow estimation. This result is summarized in Proposition 5.

Proposition 5 *Given a data that restricts the observations of the duration time on* $[0, \tau]$, *for any* $0 < \tau < \infty$. *Let* $m \to \infty$, *the estimator of* β *and* $\Lambda_0(t)$ *from* (B.9) *converges to the partial likelihood estimator of* β *and the Breslow estimator of* $\Lambda_0(t)$, *pointwise.*

Proof. The equations (B.9) are

$$\begin{cases} \sum_{i=1}^n \left(N_i(t_1) - N_i(t_0) - Y_i(t_1)e^{\hat{\beta}^T X_i}(\hat{\Lambda}_0(t_1) - \hat{\Lambda}_0(t_0))\right) = 0, \\ \vdots \\ \sum_{i=1}^n \left(N_i(t_m) - N_i(t_{m-1}) - Y_i(t_m)e^{\hat{\beta}^T X_i}(\hat{\Lambda}_0(t_m) - \hat{\Lambda}_0(t_{m-1}))\right) = 0, \\ \sum_{i=1}^n \sum_{j=1}^m X_i \left(N_i(t_j) - N_i(t_{j-1}) - Y_i(t_j)e^{\hat{\beta}^T X_i}(\hat{\Lambda}_0(t_j) - \hat{\Lambda}_0(t_{j-1}))\right) = 0. \end{cases}$$

It is easy to see that they are the discrete version of equation (1.10) with discrete $\Lambda_0(t)$. By simple computation, (B.9) is equivalent to

$$\begin{cases} \hat{\Lambda}_0(t_1) - \hat{\Lambda}_0(t_0) = \frac{\sum_{i=1}^n (N_i(t_1) - N_i(t_0))}{\sum_{i=1}^n Y_i(t_1)e^{\hat{\beta}^T X_i}}, \\ \vdots \\ \hat{\Lambda}_0(t_m) - \hat{\Lambda}_0(t_{m-1}) = \frac{\sum_{i=1}^n (N_i(t_m) - N_i(t_{m-1}))}{\sum_{i=1}^n Y_i(t_m)e^{\hat{\beta}^T X_i}}, \\ \sum_{i=1}^n \sum_{j=1}^m X_i \left(N_i(t_j) - N_i(t_{j-1}) - \frac{Y_i(t_j)e^{\hat{\beta}^T X_i} \sum_{i=1}^n (N_i(t_j) - N_i(t_{j-1}))}{\sum_{i=1}^n Y_i(t_j)e^{\hat{\beta}^T X_i}}\right) = 0. \end{cases}$$

Special care should be taken in the case that there is no observation at or after t_j. In this case, $Y_i(t_j) = 0$, for all $i = 1, \cdots, n$, the denominator in the above fractions will be zero, thus we take

$$\hat{\Lambda}_0(t_j) - \hat{\Lambda}_0(t_{j-1}) = 0.$$

Now let $m \to \infty$, the first m equations become

$$d\hat{\Lambda}_0(t) = \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n Y_i(t)e^{\hat{\beta}^T X_i}}.$$

The last equation, by changing the order of summation, is

$$\sum_{i=1}^n \sum_{j=1}^m X_i \left(N_i(t_j) - N_i(t_{j-1})\right) - \sum_{i=1}^n \sum_{j=1}^m \frac{\sum_{i=1}^n Y_i(t_j)e^{\hat{\beta}^T X_i} X_i}{\sum_{i=1}^n Y_i(t_j)e^{\hat{\beta}^T X_i}} (N_i(t_j) - N_i(t_{j-1})) = 0,$$

then take $m \to \infty$, it becomes

$$\sum_{i=1}^n \int_0^\tau X_i dN_i(t) - \sum_{i=1}^n \int_0^\tau \frac{\sum_{i=1}^n Y_i(t)e^{\hat{\beta}^T X_i} X_i}{\sum_{i=1}^n Y_i(t)e^{\hat{\beta}^T X_i}} dN_i(t) = 0$$

Thus, in the continuous time limit, (B.9) becomes

$$\begin{cases} d\hat{\Lambda}_0(t) = \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n Y_i(t)e^{\hat{\beta}^T X_i}}, \ 0 \le t \le \tau, \\ \sum_{i=1}^n \int_0^\tau \left(X_i - \frac{\sum_{i=1}^n Y_i(t)e^{\hat{\beta}^T X_i} X_i}{\sum_{i=1}^n Y_i(t)e^{\hat{\beta}^T X_i}}\right) dN_i(t) = 0, \end{cases}$$

which yields the partial likelihood estimator and the Breslow estimator. □

Appendix C: Test Specification of Covariate Effect in Discrete Case

We continue with Appendix B. Consider the same process as Lin, Wei and Ying (1993), i.e., the process $n^{-1/2}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\leq x\}}M_i(\tau)$, but with the discrete estimation obtained in Appendix B, in order to derive a discrete approximation of the covariance kernel $K^c(x_1,x_2)$. Under assumption (B1), the process equals to

$$R(x) := n^{-1/2}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\leq x\}}\boldsymbol{v}^T\boldsymbol{\varepsilon}_i, \tag{C.1}$$

with $\boldsymbol{v}=(1,\cdots,1)^T$. Its covariance kernel is

$$\mathrm{Cov}(R(x_1),R(x_2)) = \boldsymbol{v}^T\mathbb{E}\left(\mathbb{1}_{\{X\leq x_1\wedge x_2\}}\begin{pmatrix}\Delta A_1(1-\Delta A_1) & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \Delta A_m(1-\Delta A_m)\end{pmatrix}\right)\boldsymbol{v}, \tag{C.2}$$

which is a diagonal matrix due to the martingale property, and the variances of the martingale increments take the discrete form. Consider the process $R(x)$ after estimation

$$\hat{R}(x) = n^{-1/2}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\leq x\}}\boldsymbol{v}^T\hat{\boldsymbol{\varepsilon}}_i,$$

we have the following theorem and corollary demonstrate its large sample properties.

Theorem 6 *Under the null hypothesis and assumptions* (A1)(A2)(A3) *and* (B1)*, we have, uniformly in x, as $n\to\infty$,*

$$\hat{R}(x) = R(x) - \boldsymbol{v}^T\boldsymbol{G}^T(x)\boldsymbol{A}^{-1}n^{-1/2}\sum_{i=1}^{n}\boldsymbol{W}_i\boldsymbol{\varepsilon}_i + o_p(1), \tag{C.3}$$

with

$$\boldsymbol{G}(x) := \mathbb{E}\left(\mathbb{1}_{\{X\leq x\}}\begin{pmatrix}\Delta A_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \Delta A_m\\ \Delta A_1 X & \cdots & \Delta A_m X\end{pmatrix}\right),$$

and

$$\boldsymbol{A} := \mathbb{E}\begin{pmatrix} \Delta A_1 & \cdots & 0 & \Delta A_1 X \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \Delta A_m & \Delta A_m X \\ \Delta A_1 X & \cdots & \Delta A_m X & \left(\sum_{j=1}^m \Delta A_j\right) X^2 \end{pmatrix}.$$

Corollary 7 *Under the null hypothesis and assumptions* (A1)(A2)(A3) *and* (B1)*, the process* $\hat{R}(x)$ *converges weakly to a centered Gaussian process in the space* $D(-\infty, \infty)$*, as* $n \to \infty$*,*

$$\hat{R}(x) \xrightarrow{d} \hat{R}_\infty(x).$$

Proof. The Taylor expansion of $\hat{R}_n(x)$ is given by

$$\hat{R}(x) = R(x) - n^{-1}\sum_{i=1}^n \mathbb{1}_{\{X_i \le x\}} \boldsymbol{v}^T \begin{pmatrix} \Delta A^*_{i1} & \cdots & 0 & \Delta A^*_{i1} X_i \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \Delta A^*_{im} & \Delta A^*_{im} X_j \end{pmatrix} \times n^{1/2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right),$$

where $\Delta A^*_{ij} = Y_i(t_j) e^{\beta^{*T} X_i + \eta^*_j}$ with (β^*, η^*_j) being proper value between the estimator $(\hat{\beta}, \hat{\eta}_j)$ and its true value. From the Taylor expansion of (B.9), we have

$$n^{1/2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = n^{-1}\sum_{i=1}^n \begin{pmatrix} \Delta A^{**}_{i1} & \cdots & 0 & \Delta A^{**}_{i1} X_i \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \Delta A^{**}_{im} & \Delta A^{**}_{im} X_i \\ \Delta A^{**}_{i1} X_i & \cdots & \Delta A^{**}_{im} X_i & (\sum_{j=1}^m \Delta A^{**}_{ij}) X_i^2 \end{pmatrix}^{-1} \times n^{-1/2} \sum_{i=1}^n \boldsymbol{W}_i \boldsymbol{\varepsilon}_i,$$

where $\Delta A^{**}_{ij} = Y_i(t_j) e^{\beta^{**T} X_i + \eta^{**}_j}$ with $(\beta^{**}, \eta^{**}_j)$ being proper value between the estimator $(\hat{\beta}, \hat{\eta}_j)$ and its true value.

Under (B1), $\hat{\boldsymbol{\theta}}$ coincide with the partial likelihood estimator of β and the Breslow estimator of $\Lambda_0(t)$, thus it is consistent. It then follows from the uniform SLLN

(Jennrich, 1969) that

$$n^{-1}\sum_{i=1}^{n}\begin{pmatrix} \Delta A_{i1}^{**} & \cdots & 0 & \Delta A_{i1}^{**}X_i \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \Delta A_{im}^{**} & \Delta A_{im}^{**}X_i \\ \Delta A_{i1}^{**}X_i & \cdots & \Delta A_{im}^{**}X_i & (\sum_{j=1}^{m}\Delta A_{ij}^{**})X_i^2 \end{pmatrix}^{-1} \xrightarrow{p} \boldsymbol{A}^{-1},$$

and

$$n^{-1}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\le x\}}\begin{pmatrix} \Delta A_{i1}^{*} & \cdots & 0 & \Delta A_{i1}^{*}X_i \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \Delta A_{im}^{*} & \Delta A_{im}^{*}X_i \end{pmatrix} \xrightarrow{p} \boldsymbol{G}^T(x),$$

uniformly, as $n\to\infty$.

Thus, we have Theorem 4. The tightness result is provided by Lin, Wei and Ying (1993). By tightness and CLT, we have Corollary 5. □

The covariance kernel of the limit Gaussian process is then easy to compute, which is

$$\begin{aligned}
&\mathrm{Cov}(\hat{R}_\infty(x_1),\hat{R}_\infty(x_2)) \\
=&\mathrm{Cov}(R(x_1),R(x_2)) \\
&+\boldsymbol{v}^T\boldsymbol{G}^T(x_1)\boldsymbol{A}^{-1}\mathbb{E}\left(\boldsymbol{W}\begin{pmatrix} \Delta A_1(1-\Delta A_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta A_m(1-\Delta A_m) \end{pmatrix}\boldsymbol{W}^T\right)\boldsymbol{A}^{-1}\boldsymbol{G}(x_2)\boldsymbol{v} \\
&-\boldsymbol{v}^T\boldsymbol{G}^T(x_1)\boldsymbol{A}^{-1}\mathbb{E}\left(\mathbb{1}_{\{X\le x_2\}}\boldsymbol{W}\begin{pmatrix} \Delta A_1(1-\Delta A_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta A_m(1-\Delta A_m) \end{pmatrix}\right)\boldsymbol{v} \\
&-\boldsymbol{v}^T\boldsymbol{G}^T(x_2)\boldsymbol{A}^{-1}\mathbb{E}\left(\mathbb{1}_{\{X\le x_1\}}\boldsymbol{W}\begin{pmatrix} \Delta A_1(1-\Delta A_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta A_m(1-\Delta A_m) \end{pmatrix}\right)\boldsymbol{v}.
\end{aligned} \tag{C.4}$$

This covariance kernel will be dramatically simplified if we apply the approximation (B.7) again. In this case, it reduces to a covariance kernel

$$\mathrm{Cov}(R(x_1), R(x_2)) \approx \boldsymbol{v}^T \left(\mathbb{E}\left(\mathbb{1}_{\{X \leq x_1 \wedge x_2\}} \begin{pmatrix} \Delta A_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Delta A_m \end{pmatrix} \right) - \boldsymbol{G}^T(x_1) \boldsymbol{A}^{-1} \boldsymbol{G}(x_2) \right) \boldsymbol{v}, \tag{C.5}$$

which is the $K^m(x_1, x_2)$ that we introduce in section 1.4. This reduction is no accident, but because of the particular optimal instrumental variable estimation.

Chapter 2

Goodness-of-Fit Tests for the Cox Proportional Hazard Model

2.1 Introduction

As we mentioned in chapter 1, the martingale residuals provide a basis for specification tests in hazard models. For checking the Cox model, Lin, Wei and Ying (1993) proposed a class of goodness-of-fit tests based on the CUSUM process of martingale residuals. In this article, we propose new goodness-of-fit tests for the Cox model based on some components of the CUSUM martingale process, i.e., we are testing

$$H_0 : \lambda(t \mid X) = \lambda_0(t)exp(\beta^T X),\ a.s. \text{ for some } \beta \text{ and nonnegative } \lambda_0(t)$$

against its negation, where $\lambda_0(t)$ is an unspecified baseline hazard function. These components are obtained through a conditional PCD method, which fills the gap of PCA in the conditional model testing problem and is the main contribution of this paper. The components of the CUSUM martingale process play a similar role with the traditional PCs of the empirical process, hence behave as building blocks for goodness-of-fit tests. The difference is that these components are stochastic processes rather than random variables. Therefore we call them component processes to be more precise. Specifically, it consists of two steps to obtain the component processes, (i) derive PCD of the individual martingale process conditional on the covariables, (ii) sum up the obtained PCs in the first step w.r.t. the observations of the covariables. It turns out that the CUSUM martingale process can be decomposed into a decreasing weighted sum of the component processes. Since the PCD is in the time domain, the obtained components are sensitive when detecting certain deviations from the constant hazard ratio implied by the proportional hazard assumption, especially, higher-frequency deviations are more reflected in latter components. The omnibus test, which is based on the original CUSUM martingale process, down-weights the latter components heavily, thus, it has low power when detecting the high-frequency deviations. While we propose new goodness-of-fit tests, including tests based on each estimated component process and a Bonferroni test, which offset the power loss and outperform the omnibus test. Smooth tests that

based on reweighted sums of a few components are also constructed.

The conditional PCD method in this paper is applicable for any regression model that has a martingale interpretation, including conditional hazard models and transformation models. It also works for conditional distribution models, where the empirical process has a Brownian Bridge structure. However, we focus on the Cox model in the present paper. A brief introduction of the Cox model, together with some other important models in duration analysis and the omnibus test proposed by Lin, Wei and Ying (1993) is in section 2.2. Section 2.3 contains the main result: the conditional PCD, the asymptotic results of the component processes and the test statistics based on the component processes. Simulation studies illustrating the performance of our tests in the finite sample are presented in section 2.4.

2.2 Omnibus Test for the Cox Proportional Hazard Model

2.2.1 The Cox Model

This section is more or less the same with section 1.2.1 and 1.2.2, while here we repeat the basic setting of the Cox model in order to introduce some notations. In the framework of regression analysis with right-censored duration data, consider a sample $\{Z_i, \Delta_i, X_i\}, i = 1, \cdots, n$ of i.i.d. realizations of $\{Z, \Delta, X\}$. Here Z is the minimum of the non-negative failure and censoring time, which are denoted by T and C, i.e., $Z = min(T, C)$. The indicator $\Delta = \mathbb{1}_{\{T \leq C\}}$ contains the information indicating which of T and C is actually observed, and X is the covariable vector.

The hazard function, which is defined as the limit of conditional probability,

$$\lambda(t \mid X) = \lim_{h \to 0} h^{-1} P(t \leq T < t + h \mid T \geq t, X), \tag{2.1}$$

is assumed to have a multiplicative form in the Cox model, i.e.,

$$\lambda(t \mid X) = \lambda_0(t) exp(\beta^T X), \tag{2.2}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and the link function takes exponential form.

Another approach to the censored data regression model is based on the analysis of counting process. Define the following two processes

$$N(t) = \mathbb{1}_{\{Z \leq t, \Delta = 1\}},$$

$$Y(t) = \mathbb{1}_{\{Z \geq t\}}.$$

Here $N(t)$ is the counting process, and $Y(t)$ is the at-risk process. Applying the Doob-Meyer decomposition, there is a unique predictable process $A(t)$ such that $N(t) - A(t)$ is a martingale and $A(t)$ is called the compensator of $N(t)$. In the counting process approach, instead of modeling conditional hazard rate of T, the compensator process is modeled. Notice that the information contained in $\{Z, \Delta\}$ is equivalent to that contained in $\{N, Y\}$. Actually, these two approaches are equivalent under the conditional independence of T and C on X. To be more specific

$$M(t) = N(t) - \int_0^t Y(u) d\Lambda(u \mid X) \tag{2.3}$$

is a martingale process with the filtration $\mathscr{F}_t = \sigma\{X, N(u), Y(u+) : 0 \leq u \leq t\}$. Then modeling the compensator $\int_0^t Y(s) d\Lambda(s \mid X)$ is equivalent to modeling the conditional hazard.

The counting process counts the number of occurrence of the event, while the compensator captures its expected value. Thus, the martingale, as the difference of them, plays the same role with the error term in the mean regression model. The Doob-Meyer decomposition serves the same purpose with the projection decomposition, but instead of orthogonality between two random variables, we have martingale process.

Under the counting process framework, if the Cox specification is correct for a given sample, there exists a β and $\lambda_0(t)$, such that

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)exp(\beta^T X_i)\lambda_0(s)ds \quad i = 1, \cdots, n \tag{2.4}$$

are martingales. The corresponding martingale residuals are defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)exp(\hat{\beta}^T X_i)d\hat{\Lambda}_0(s), \tag{2.5}$$

where $\hat{\beta}$ is an estimator of β and $\hat{\Lambda}_0(t)$ is an estimator of the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. These martingale residuals provide a basis for goodness-of-fit test for the Cox model.

The estimation of the Cox model was suggested by Cox (1972, 1975) using partial likelihood inference. The partial likelihood score function for β is

$$U(\beta, \infty) = \sum_{i=1}^{n} \int_0^{\infty} \Big(X_i - \bar{X}(\beta, t)\Big) dN_i(t), \tag{2.6}$$

where

$$\bar{X}(\beta, t) = \frac{\sum_{i=1}^n Y_i(t)e^{\beta^T X_i} X_i}{\sum_{i=1}^n Y_i(t)e^{\beta^T X_i}}.$$

The partial likelihood estimator $\hat{\beta}$ is the solution to $U(\beta, \infty) = 0$. Under some mild regularity conditions, $n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution to a centered Gaussian variable with covariance matrix $\Sigma(\beta_0, \infty)^{-1}$. The matrix $\Sigma(\beta, t)$ is defined as

$$\Sigma(\beta, t) = \mathbb{E}\left(\int_0^t (X - \tilde{X}(\beta, s))^2 Y(s)e^{\beta^T X} d\Lambda_0(s)\right),$$

with

$$\tilde{X}(\beta, t) = \frac{\mathbb{E}\left(Y(t)e^{\beta^T X}X\right)}{\mathbb{E}\left(Y(t)e^{\beta^T X}\right)}$$

being the limit of $\bar{X}(\beta, t)$. The cumulative baseline hazard is estimated by the Breslow (1974) estimator

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u)e^{\hat{\beta}^T X_i}}. \tag{2.7}$$

2.2.2 Other Important Models in Duration Analysis

The Cox proportional hazard model assumes the conditional hazard rate of the duration time to be as the product of a baseline hazard and the covariable effect. In this sense, it is also called the multiplicative hazard model. Another important hazard model is the Aalen's additive hazard model, which is proposed by Aalen (1980) and the hazard rate is assumed to be a summation of the covariable effects. The multiplicative and additive hazard models are suitable for regression analysis of duration data, however, they are not the only important models in duration analysis. There are two general classes of models in duration analysis with regression, the transformation model and the accelerated failure time model. In fact, the Cox model is a special case of the transformation model.

A transformation model is

$$H(T) = -\beta^T X + \varepsilon, \tag{2.8}$$

with $H(\cdot)$ an unknown monotone transformation and ε an error term with a known distribution. The transformation model has a martingale interpretation, i.e., if we denote Λ_ε as the known cumulative hazard function of ε, then

$$M(t) = N(t) - \int_0^t Y(u) d\Lambda_\varepsilon(\beta^T X + H(u))$$

is a martingale. One special case is the Cox model, in which ε is taken to follow the extreme-value distribution with $\Lambda_\varepsilon(t) = e^t$ and the transformation is taken as $H(\cdot) = \ln(\Lambda_0(\cdot))$. Another special case is the proportional odds model, in which ε follows the standard logistic distribution.

The accelerated failure time model assumes

$$\log(T) = -\beta^T X + \varepsilon, \tag{2.9}$$

with an unspecified distribution of ε. It is just a transformed version of an ordinary linear model. The inference of the accelerated failure time model is not as easy as

that of the Cox model because of censorship. This is no direct martingale structure in (2.9). Although the parameter is easily interpreted as the effect on the mean of $\log(T)$ in the standard linear regression model, it is not so clear when T is under censorship. For the transformation model, Chen et al. (2002) have proposed an estimating equation approach based on the martingale structure. The estimation coincides with the partial-likelihood estimator in the special case of the Cox model. A brief review of the transformation model and accelerated failure time model can be found in Martinussen and Scheike (2007).

The method to construct a goodness-of-fit test in this paper is applicable to models that have a martingale structure, e.g., hazard models and transformation models. The tests we propose are therefore helpful with model selection for analysis of duration data. We demonstrate the method under the Cox model in section 2.3, and generate data from transformation models and accelerated failure time models as alternatives in the simulation in section 2.4, to study the power of our tests.

2.2.3 Omnibus Test

To test the specification of the Cox model, i.e., to test

$$H_0 : \lambda(t \mid X) = \lambda_0(t)exp(\beta^T X) \ a.s. \ \text{ for some } \beta \text{ and nonnegative } \lambda_0(t),$$

against all the possible alternatives, Lin, Wei and Ying (1993) proposed an omnibus test by considering the CUSUM process of martingale residuals

$$\hat{R}_n(t,x) := n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}} \hat{M}_i(t), \tag{2.10}$$

where $\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)exp(\hat{\beta}^T X_i)d\hat{\Lambda}_0(s), i = 1, \cdots, n$ are martingale residuals with partial likelihood estimator $\hat{\beta}$ and Breslow estimator $\hat{\Lambda}_0(t)$. They have shown that, under the null hypothesis, the process $\hat{R}_n(t,x)$ converges weakly to a centered Gaussian process $\hat{R}_\infty(t,x)$ in the space $D([0,\infty) \times [-1,1])$. Kolmogorov type statistic is constructed based on this process.

To simplify the notation, we write X as real-valued in the univariate case, however, it can be any vector and all the arguments in this chapter work for multivariate X. In the next section, we develop a decomposition of the process with the true value of the parameters

$$R_n(t,x) := n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} M_i(t), \tag{2.11}$$

into a countable sum of component processes and use these estimated component processes to construct new test statistics.

2.3 Tests based on Component Processes

2.3.1 Conditional Principal Component Decomposition

Notice that the process R_n in (2.11) is bivariate with non-independent components x and t. Hence, the direct Karhunen-Loève representation is not available in this case. Instead, we adopt a conditional PCD, namely, to do the PCD of individual martingales conditional on X, and then sum the decompositions up. We introduce the following assumptions.

(A1). T and C are independent conditional on X.

(A2). For each $\tau < \infty$, $P\{Y(\tau) = 1\} > 0$.

(A2). X is bounded, without loss of generality by 1.

(A4). C is independent of X.

(A5). $\Sigma(\beta_0, \infty) = \mathbb{E}\left[\int_0^\infty (X - \tilde{X}(\beta_0, s))^2 Y(s) e^{\beta_0^T X} \lambda_0(s) ds\right]$ is positive definite.

The first three assumptions are standard in the Cox model. The fourth one is needed to justify the consistency of the estimated conditional martingale variance. The last assumption is needed to get the asymptotic distribution of the partial likelihood estimator $\hat{\beta}$, see Anderson and Gill (1982), Theorem 4.2.

Let us begin with the PCD of the martingale $M(t)$ conditional on X. Suppose

the counting process has continuous compensator, which is equivalent to having continuous $\Lambda_0(t)$, the conditional covariance of $M(t)$ conditional on X is

$$\begin{aligned}
\mathbb{E}(M(s)M(t) \mid X) &= \mathbb{E}\Big[\int_0^{s\wedge t} Y(u)e^{\beta^T X}\lambda_0(u)du \mid X\Big] \\
&= \int_0^{s\wedge t} \mathbb{E}[Y(u) \mid X]e^{\beta^T X}\lambda_0(u)du \\
&= \int_0^{s\wedge t} P(T \geq u, C \geq u \mid X)e^{\beta^T X}\lambda_0(u)du \\
&= \int_0^{s\wedge t} P(T \geq u \mid X)P(C \geq u \mid X)e^{\beta^T X}\lambda_0(u)du \\
&= \int_0^{s\wedge t} exp(-\Lambda_0(u)e^{\beta^T X})P(C \geq u)e^{\beta^T X}\lambda_0(u)du. \qquad (2.12)
\end{aligned}$$

The first equation follows from martingale properties, see Fleming and Harrington (1991), Theorem 2.5.1. The last two equations follow respectively from assumption (A1) and (A3). Let us denote the conditional covariance function by

$$H(s \wedge t, x) = \mathbb{E}(M(s)M(t) \mid X = x), \qquad (2.13)$$

with

$$H(t, x) := \mathbb{E}(M^2(t) \mid X = x) = \int_0^t exp(-\Lambda_0(u)e^{\beta^T X})P(C \geq u)e^{\beta^T X}\lambda_0(u)du.$$

Notice that function H is non-decreasing in t, and $H(0, x) = 0$, $H(\infty, x) \leq 1$.

Remark. Suppose we do not have censorship, then

$$H(t, x) = 1 - exp(-\Lambda_0(t)e^{\beta^T x}) = F_T(t \mid X = x),$$

where $F_T(t \mid X)$ is the conditional distribution function of T conditional on X. The conditional covariance of the martingale equals to the conditional distribution of T. In this case, $H(\infty, x) = 1$.

Let

$$\mu_j = \frac{4}{\pi^2(2j-1)^2}, \quad \varphi_j(t) = \sqrt{2}sin\frac{(2j-1)\pi t}{2}, \quad j = 1, 2, \cdots$$

be the eigenvalues and eigenfunctions of the standard Brownian Motion with covariance structure $K(s,t) = s \wedge t$. For each x, let f_j be the transformation

$$f_j(t,x) = \varphi_j(H(t,x)/H(\infty,x)).$$

Then, for each x, $\{f_j(\cdot,x)\}_{j=1}^{\infty}$ form an orthonormal basis of a subspace of the Hilbert space $L^2(\mathbb{R}^+, H(\cdot,x)/H(\infty,x))$ of all square integrable functions on $\mathbb{R}^+$ with the inner product

$$\langle \rho, g \rangle_x = \int_{\mathbb{R}^+} \rho(t)g(t)\frac{H(dt,x)}{H(\infty,x)},$$

since

$$\begin{aligned}\langle f_j, f_h \rangle_x &= \int_{\mathbb{R}^+} \varphi_j\Big(\frac{H(t,x)}{H(\infty,x)}\Big)\varphi_h\Big(\frac{H(t,x)}{H(\infty,x)}\Big)\frac{H(dt,x)}{H(\infty,x)} \\ &= \int_0^1 \varphi_j(u)\varphi_h(u)du = \begin{cases} 1 & j = h \\ 0 & j \neq h \end{cases}.\end{aligned}$$

Moreover, $\{f_j(\cdot,x)\}_{j=1}^{\infty}$ are the eigenfunctions of the covariance kernel $H(s \wedge t, x)/H(\infty,x)$ with associated eigenvalues $\{\mu_j\}_{j=1}^{\infty}$, i.e.,

$$\int_{\mathbb{R}^+} \frac{H(s \wedge t, x)}{H(\infty,x)} f_j(s,x)\frac{H(ds,x)}{H(\infty,x)} = \mu_j f_j(t,x).$$

By Mercer's theorem, the covariance function can be decomposed as

$$\frac{H(s \wedge t, x)}{H(\infty,x)} = \sum_{j=1}^{\infty} \mu_j f_j(s,x) f_j(t,x).$$

Since $H(s \wedge t, x)/H(\infty,x)$ is the conditional covariance function of $(H(\infty,x))^{-1/2}M_i(t)$ given $X_i = x$, we have the PCD of individual maringales as

$$(H(\infty,X_i))^{-1/2}M_i(t) = \sum_{j=1}^{\infty} {\mu_j}^{1/2} z_{ij} f_j(t,X_i) \;\; a.s., \; i = 1, \cdots, n \tag{2.14}$$

where

$$\begin{aligned} z_{ij} &:= {\mu_j}^{-1/2}\langle (H(\infty,X_i))^{-1/2}M_i, f_j(\cdot,X_i)\rangle_{X_i} \\ &= {\mu_j}^{-1/2}\int_{\mathbb{R}^+} (H(\infty,X_i))^{-3/2}M_i(t)f_j(t,X_i)H(dt,X_i). \end{aligned}$$

The z_{ij} is the j'th principal component of $(H(\infty, X_i))^{-1/2} M_i(t)$ conditional on X_i. For each j and $j \neq h$, it has the following properties

$$\begin{aligned} \mathbb{E}(z_{ij} \mid X_i) &= 0, \\ \mathbb{E}(z_{ij}^2 \mid X_i) &= 1, \\ \mathbb{E}(z_{ij} z_{ih} \mid X_i) &= 0. \end{aligned} \tag{2.15}$$

Hence, plugging the PCD (2.14) into (2.11), the process R_n has the decomposition as

$$\begin{aligned} R_n(t,x) &= n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \left[(H(\infty, X_i))^{1/2} \sum_{j=1}^{\infty} {\mu_j}^{1/2} z_{ij} f_j(t, X_i) \right] \\ &= \sum_{j=1}^{\infty} {\mu_j}^{1/2} \left[n^{-1/2} \sum_{i=1}^{n} z_{ij} \mathbb{1}_{\{X_i \leq x\}} (H(\infty, X_i))^{1/2} f_j(t, X_i) \right]. \end{aligned}$$

We call the term in the bracket the j'th component process of R_n and denote it as

$$c_{n,j}(t,x) := n^{-1/2} \sum_{i=1}^{n} z_{ij} \mathbb{1}_{\{X_i \leq x\}} (H(\infty, X_i))^{1/2} f_j(t, X_i). \tag{2.16}$$

Thus, we have the following proposition.

Proposition 1 *Under the null hypothesis and* $(A1)$*-*$(A5)$*, the CUSUM of martingale processes* (2.11) *can be decomposed into a weighted sum of component processes, i.e.,*

$$R_n(t,x) = \sum_{j=1}^{\infty} {\mu_j}^{1/2} c_{n,j}(t,x). \tag{2.17}$$

The weights are the square root of the standard Brownian Motion eigenvalues.

2.3.2 Asymptotic Theory of Component Processes

In this section, we develop asymptotic results of the component processes and their estimated ones. From (2.16), each component process is a sum of i.i.d. centered random functions with variance

$$H_j(t,x) := \mathbb{E}\left[\mathbb{1}_{\{X \leq x\}} H(\infty, X) f_j^2(t, X) \right] = \int_{-\infty}^{x} H(\infty, s) f_j^2(t,s) F_X(ds),$$

where $F_X(\cdot)$ denotes the distribution function of X. Together with a tightness result, we have the following theorem.

Theorem 1 *Under the null hypothesis and (A1)-(A5), for each j, the process $c_{n,j}(t,x)$ converges weakly to a centered Gaussian process in the space $D([0,\infty) \times [-1,1])$,*

$$c_{n,j} \xrightarrow{d} c_{\infty,j}.$$

The limit Gaussian process $c_{\infty,j}$ has covariance structure

$$K(t_1,t_2,x_1,x_2) = \int_{-\infty}^{x_1 \wedge x_2} H(\infty,s) f_j(t_1,s) f_j(t_2,s) F(ds).$$

Moreover, $c_{\infty,j}$ and $c_{\infty,h}$ are independent for $j \neq h$.

The expression of the component process (2.16) can be rewritten as a sum of i.i.d. martingale integrals. Let us define another function g_j corresponding to f_j as

$$g_j(t,x) := \phi_j(H(t,x)/T(\infty,x)) = \sqrt{2} cos \frac{(2j-1)H(t,x)/H(\infty,x)}{2}.$$

By integration by parts, the component processes equal to

$$c_{n,j}(t,x) = n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} \mathbb{1}_{\{X_i \leq x\}} f_j(t,X_i) g_j(s,X_i) dM_i(s). \tag{2.18}$$

It is convenient to write statistics as martingale integrals in duration analysis. Now we consider the component process after estimation, i.e., the process

$$\hat{c}_{n,j}(t,x) := n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} \mathbb{1}_{\{X_i \leq x\}} \hat{f}_j(t,X_i) \hat{g}_j(s,X_i) d\hat{M}_i(s).$$

Here

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) exp(\hat{\beta}^T X_i) d\hat{\Lambda}_0(s),$$

$$\hat{f}_j(t,x) = \varphi_j(\hat{H}(t,x)/\hat{H}(\infty,x)),$$

and

$$\hat{g}_j(t,x) = \phi_j(\hat{H}(t,x)/\hat{H}(\infty,x)).$$

As remarked earlier, it involves the estimators of β and Λ_0 and the estimator of the conditional covariance function $H(t,x)$. For $\hat{H}(t,x)$, recall that

$$H(t,x) = \int_0^t exp(-\Lambda_0(u)e^{\beta^T X})P(C \geq u)e^{\beta^T X}\lambda_0(u)du.$$

A natural consistent estimator is

$$\hat{H}(t,x) = \int_0^t exp(-\hat{\Lambda}_0(u)e^{\hat{\beta}x})(1-\hat{G}(u-))e^{\hat{\beta}x}d\hat{\Lambda}_0(u), \tag{2.19}$$

where $\hat{G}$ is the Kaplan-Meier estimator of the distribution function of C.

In Appendix, it is shown that $\hat{c}_{n,j}(t,x)$ has the same asymptotic distribution as

$$\begin{aligned}\tilde{c}_{n,j}(t,x) \;:=\; & n^{-1/2}\sum_{i=1}^n \int_0^\infty \Big[\mathbb{1}_{\{X_i\leq x\}}f_j(t,X_i)g_j(s,X_i) - \tilde{l}_j(\beta_0,t,x,s)\Big]dM_i(s) \\ & -A_j(t,x)\Sigma(\beta_0,\infty)^{-1}n^{-1/2}\sum_{i=1}^n\int_0^\infty (X_i - \tilde{X}(\beta_0,s))dM_i(s),\end{aligned}$$

with $\tilde{X}(\beta,t)$ and $\Sigma(\beta,t)$ defined in section 2.2.1, and

$$\tilde{l}_j(\beta,t,x,s) = \frac{\mathbb{E}[Y(s)e^{\beta^T X}\mathbb{1}_{\{X\leq x\}}f_j(t,X)g_j(s,X)]}{\mathbb{E}[Y(s)e^{\beta^T X}]},$$

$$A_j(t,x) = \mathbb{E}\left[\int_0^\infty Y(s)e^{\beta_0^T X}(X-\tilde{X}(\beta_0,s))\lambda_0(s)\mathbb{1}_{\{X\leq x\}}f_j(t,X)g_j(s,X)ds\right].$$

The process $\tilde{c}_{n,j}(t,x)$ can also be rewritten as an i.i.d. sum of martingale integrals

$$\tilde{c}_{n,j}(t,x) = n^{-1/2}\sum_{i=1}^n\int_0^\infty h_{ij}(\beta_0,t,x,s)dM_i(s), \tag{2.20}$$

with

$$h_{ij}(\beta,t,x,s) = \mathbb{1}_{\{X_i\leq x\}}f_j(t,X_i)g_j(s,X_i) - \tilde{l}_j(\beta,t,x,s) - A_j(t,x)\Sigma(\beta,\infty)^{-1}(X_i - \tilde{X}(\beta,s)). \tag{2.21}$$

The asymptotic distribution of $\hat{c}_{n,j}(t,x)$ is shown by the following theorem.

Theorem 2 *Under the null hypothesis and (A1)-(A5), for each $j = 1,2,\cdots$, the process $\hat{c}_{n,j}(t,x)$ converges weakly to a centered Gaussian process in the space $D([0,\infty)\times[-1,1])$,*

$$\hat{c}_{n,j} \xrightarrow{d} \tilde{c}_{\infty,j}.$$

The limit Gaussian process $\tilde{c}_{\infty,j}(t,x)$ *has covariance structure*

$$K(t_1,t_2,x_1,x_2)=\mathbb{E}\left[\int_0^\infty h_j(\beta_0,t_1,x_1,s)h_j(\beta_0,t_2,x_2,s)Y(s)e^{\beta_0^T X}\lambda_0(s)ds\right].$$

In addition to the single component process, finite weighted sum of some component processes can also be used for model checking. Consider the first m component processes with weight $w=\{w_j\}_{j=1}^m$, i.e., the process $\sum_{j=1}^m w_j\hat{c}_{n,j}(t,x)$. It has the same asymptotic distribution with the following process

$$\sum_{j=1}^m w_j\tilde{c}_{n,j}(t,x)=n^{-1/2}\sum_{i=1}^n\int_0^\infty\left(\sum_{j=1}^m w_j h_{ij}(\beta_0,t,x,s)\right)dM_i(s). \tag{2.22}$$

Its asymptotic distribution is given in the following theorem.

Theorem 3 *Under the null hypothesis and* $(A1)$*-*$(A5)$*, for any given weight* $w=\{w_j\}_{j=1}^m$*, the process* $\sum_{j=1}^m w_j\hat{c}_{n,j}(t,x)$ *converges weakly to a centered Gaussian process in the space* $D([0,\infty)\times[-1,1])$*,*

$$\sum_{j=1}^m w_j\hat{c}_{n,j}\xrightarrow{d}\tilde{c}_\infty^w.$$

The limit Gaussian process $\tilde{c}_\infty^w(t,x)$ *has covariance structure*

$$K(t_1,t_2,x_1,x_2)=\mathbb{E}\left[\int_0^\infty\left(\sum_{j=1}^m w_j h_j(\beta_0,t_1,x_1,s)\right)\left(\sum_{j=1}^m w_j h_j(\beta_0,t_2,x_2,s)\right)Y(s)e^{\beta_0^T X}\lambda_0(s)ds\right].$$

2.3.3 Test Statistics

The omnibus tests in Lin, Wei and Ying (1993) are based on the original CUSUM martingale residual process. By the continuous mapping theorem, we have the following asymptotic distribution of the Kolmogorov-Smirnov and Cramér-von Mises type statistics

$$KS_o=\sup_{t,x}\left|\hat{R}_n(t,x)\right|\xrightarrow{d}\sup_{t,x}\left|R_\infty(t,x)\right|,$$

$$CvM_o = \int \left[\hat{R}_n(t,x)\right]^2 \hat{F}_X(dx)dt \xrightarrow{d} \int \left[R_\infty(t,x)\right]^2 F_X(dx)dt.$$

Here $\hat{F}_X(\cdot)$ is the empirical distribution of X.

The component processes we derived provide a basis of new specification tests for the Cox model. We propose to construct Kolmogorov-Smirnov and Cramér-von Mises type statistics based on each component process, i.e., for each $j = 1, 2, \cdots$, we have the following, what we call, component tests,

$$KS_{nj} = \sup_{t,x} \left|\hat{c}_{n,j}(t,x)\right| \xrightarrow{d} \sup_{t,x} \left|\tilde{c}_{\infty,j}(t,x)\right|,$$

$$CvM_{nj} = \int \left[\hat{c}_{n,j}(t,x)\right]^2 \hat{F}_X(dx)dt \xrightarrow{d} \int \left[\tilde{c}_{\infty,j}(t,x)\right]^2 F_X(dx)dt.$$

Note that in (3.6), the weight for the j'th component process is $\mu_j^{1/2}$ that decreases very rapidly in j. In consequence, the latter components are down-weighted in the original process. In fact, each component reflects certain aspect of a deviation from the null hypothesis. For example, high-frequency deviations are more reflected in the latter components. Therefore, the omnibus test, which gives low weights to latter components, has low power, while the tests based on latter components are specially designed for such high-frequency alternatives. In practice, the data should not be very frequent, hence we can focus on the first few components, say no more than ten in general.

In addition, smooth test statistics based on the reweighted sum of component processes can be constructed. If we give the components with equal weights and consider the sum of the first m components, the Kolmogorov and Cramér-von Mises type statistics, for some fixed m, can be constructed as

$$KS_{nm} = \sup_{t,x} \left|\sum_{j=1}^{m} w_j \hat{c}_{n,j}(t,x)\right| \xrightarrow{d} \sup_{t,x} \left|\tilde{c}^w_\infty(t,x)\right|,$$

$$CvM_{nm} = \int \Big[\sum_{j=1}^{m} w_j \hat{c}_{n,j}(t,x)\Big]^2 \hat{F}_X(dx)dt \xrightarrow{d} \int \left[\tilde{c}^w_\infty(t,x)\right]^2 F_X(dx)dt.$$

The smooth tests provide a compromise between the omnibus tests and the tests based on one component. The smooth test is the one that takes $w = (1, \cdots, 1)$. The

test based on the j'th component process is the one that takes w as the j'th unit vector, i.e., $w = (0, \cdots, 1, \cdots, 0)$. However, the problem is that one has to choose a suitable w before model checking.

Actually, we can take into account the information together from some of the component processes by considering a Bonferroni test that behaves as an intersection of the component tests. Specifically, we run the first m component tests and record the decision for each one. Then we accept H_0 if all the m tests accept, and reject H_0 if any of them gives us a rejection. Let $T^1, T^2, \cdots, T^m$ be the first m component tests with common size x. The Bonferroni test T^b is

$$T^b = \begin{cases} 0 & \text{if } T^1 = \cdots = T^m = 0, \\ 1 & \text{o.w.} \end{cases}$$

The probability of Bonferroni test to accept under H_0 is $P_0(T^1 = 0, \cdots, T^m = 0)$, and it admits the following inequality

$$\begin{aligned} P_0(T^1 = 0, \cdots, T^m = 0) &\geq P_0(T^1 = 0) + \cdots + P_0(T^m = 0) - (m-1) \\ &= (1-x) + \cdots + (1-x) - (m-1) \\ &= 1 - mx. \end{aligned}$$

For a significant level α, we could choose $x = \alpha/m$, then the size of the Bonferroni test will be

$$1 - P_0(T^b = 0) = 1 - P_0(T^1 = 0, \cdots, T^m = 0) \leq mx = \alpha,$$

i.e., the Bonferroni test has a bounded size of α.

At last, to approximate the limit distribution $\tilde{c}_{\infty,j}(t, x)$, we follow the suggestion of Lin, Wei and Ying (1993) through Monte Carlo simulations. Recall from the expression (2.20), $\tilde{c}_{n,j}(t, x)$ is a martingale integral. To approximate its asymptotic distribution, the integrand $h_i(\beta_0, t, x, s)$ can be replaced by its consistent estimator, but we do not know the distribution form of the martingale $M_i(t)$. Lin, Wei and Ying (1993) suggested to replace $M_i(t)$ by a similar process which has a known

distribution. The candidate is $N_i(t)G_i$, where N_i is the observed counting process and $\{G_i; i = 1, \cdots, n\}$ is a random sample of standard normal variables. Noticing the martingale property $\mathbb{E}[M^2(t)] = \mathbb{E}[N(t)]$, the process $M_i(t)$ and $N_i(t)G_i$ have the same variance function. Finally replace all the unknown quantities in $h_i(\beta_0, t, x, s)$ by their consistent estimators, i.e., replace $\beta, \Lambda_0(t), f_j(t,x), g_j(t,x)$ by $\hat{\beta}, \hat{\Lambda}_0(t), \varphi_j(\hat{H}(t,x)/\hat{H}(\infty,x)), \phi_j(\hat{H}(t,x)/\hat{H}(\infty,x))$ and replace $\tilde{X}(\beta,t), \tilde{l}(\beta,t,x,s)$ by their sample analogies. Given the observed data, the distribution of the process after replacement is the same with $\tilde{c}_{n,j}(t,x)$ in the limit.

2.3.4 Other Weight Functions

In this section, we discuss briefly the application of the conditional PCD method in different testing problems for the Cox model. Recall the process R_n is a weighted sum of all $M_i(t)$'s with weight function the indicator $\mathbb{1}_{\{X_i \le x\}}$. The conditional PCD for R_n carries out decomposition of $M_i(t)$ conditional on X_i and the weight function $\mathbb{1}_{\{X_i \le x\}}$ only shows up at the second step after the decomposition. Hence the conditional PCD approach also works for any other process with a different weight function as a function of X_i.

For instance, if we replace the indicator by the identity function of X_i, we end up with the score process

$$U(\beta, t) = \sum_{i-1}^{n} X_i M_i(t).$$

Applying the conditional PCD, we have

$$\begin{aligned} n^{-1/2}U(\beta,t) &= n^{-1/2}\sum_{i=1}^{n} X_i \left((H(\infty, X_i))^{1/2} \sum_{j=1}^{\infty} {\mu_j}^{1/2} z_{ij} f_j(t, X_i) \right) \\ &= \sum_{j=1}^{\infty} {\mu_j}^{1/2} \left[n^{-1/2} \sum_{i=1}^{n} z_{ij} X_i (H(\infty, X_i))^{1/2} f_j(t, X_i) \right] \\ &= \sum_{j=1}^{\infty} {\mu_j}^{1/2} \left[n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} X_i f_j(t, X_i) g_j(s, X_i) dM_i(s) \right]. \end{aligned} \tag{2.23}$$

The j's component process of the score process is the one in the square bracket of

the last equation. Compared to the previous $c_{n,j}$ the only difference is by making a change of the weight function from $\mathbb{1}_{\{X_i \leq x\}}$ to X_i.

Omnibus test based on $U(\hat{\beta}, t)$, for instance, the KS supreme test, is consistent against nonproportional hazards alternative (Wei, 1984), while constructing component tests, smooth tests and Bonferroni tests in the same way as in section 2.3.3 based on its component processes could be more informative and will improve efficiency when checking against certain alternatives. Similar to $c_{n,j}$, each component process of $n^{-1/2}U(\beta, t)$ reveals certain high-frequency deviation from the constant hazard ratio implied by the proportional hazard assumption.

The weight function $\mathbb{1}_{\{X_i \leq x\}}$ ensures consistency of the omnibus test based on R_n against all the possible deviations from the Cox specification since it covers every $x \in \mathbb{R}$. However, sometimes it is too much to consider every real value, one might only be interested in the correctness of specification on a finite partition, especially for discrete X, where a natural partition exists. Suppose we want to test the specification on a partition $\{\mathcal{A}_l\}_{l=1}^{L}$ of the real line, the bivariate process then reduces to a finite collection of univariate processes

$$\gamma_{n,l}(t) = n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in \mathcal{A}_l\}} M_i(t), \ l = 1, \cdots, L.$$

Similarly, for each l, we have

$$\gamma_{n,l}(t) = \sum_{j=1}^{\infty} {\mu_j}^{1/2} \left[n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} \mathbb{1}_{\{X_i \in \mathcal{A}_l\}} f_j(t, X_i) g_j(s, X_i) dM_i(s) \right]. \tag{2.24}$$

It is again possible to improve the efficiency of tests using the component processes in the square bracket.

The last weight function we shall introduce is $\mathbb{1}_{\{\beta^T X_i \leq z\}}$, which is equivalent to $\mathbb{1}_{\{X_i \leq x\}}$ since the Cox model assumes linear functional form of X. This weight function is especially useful in the multivariate X case. Now we have the process

$$\eta_n(\beta, z, t) = n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{\beta^T X_i \leq z\}} M_i(t),$$

and its decomposition

$$\eta_n(\beta, z, t) = \sum_{j=1}^{\infty} \mu_j{}^{1/2} \left[n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} \mathbb{1}_{\{\beta^T X_i \leq z\}} f_j(t, X_i) g_j(s, X_i) dM_i(s) \right]. \quad (2.25)$$

The further analysis on $\eta_n(\beta, z, t)$ depends on the choice of β and will not be discussed here.

2.4 Simulation Study

As discussed earlier, the accelerated failure time model and transformation model provide general frameworks for studying the covariable effects of duration data. In our simulation study, we take several alternatives from these models to study the power of our tests. We consider the following DGPs with explanations afterwards.

Cox:

$$\lambda(t \mid X) = \lambda_0(t) exp(\beta^T X).$$

DGP1: Weibull hazard rate

$$\lambda(t \mid X) = (0.2X) t^{0.2X-1}.$$

DGP2: Log-normal Model

$$\ln(T) = -\beta^T X + \epsilon.$$

Here we take ϵ as a standard normal variable. This model is a special case of accelerated failure time models.

DGP3: Transformation Model

$$\Lambda_0(T) e^{\beta^T X} = Pareto,$$

where $Pareto$ is a standard Pareto variable, which has hazard rate x^{-1} for $x > 1$.

DGP4: Transformation Model

$$\Lambda_0(T)e^{\beta^T X} = A_1,$$

where A_1 is a positive random variable with hazard rate $\lambda(t) = 1 + sin(3\pi t/2)$.

DGP5: Transformation Model

$$\Lambda_0(T)e^{\beta^T X} = A_2,$$

where A_2 is a positive random variable with hazard rate $\lambda(t) = 1 + cos(3\pi t/2)$.

DGP6: Transformation Model

$$\Lambda_0(T)e^{\beta^T X} = A_3,$$

where A_3 is a positive random variable with hazard rate $\lambda(t) = 1 + sin(5\pi t/2)$.

DGP7: Transformation Model

$$\Lambda_0(T)e^{\beta^T X} = A_4,$$

where A_4 is a positive random variable with hazard rate $\lambda(t) = 1 + cos(5\pi t/2)$.

DGP1 is the Weibull hazard model, in which the hazard for different values of the covariable is non-proportional. DGP2 is a commonly used model in economics, and it belongs to the accelerated failure time models. DGP3-7 are transformation models with unspecified transformation $\ln(\Lambda_0(\cdot))$. The Cox model, as a special case of a transformation model, can be expressed as

$$\Lambda_0(T)e^{\beta^T X} = E,$$

where E is the standard exponential variable with constant hazard rate. In DGP3, we replace the exponential by a Pareto variable which has decreasing hazard rate.

For DGP4-7, we call them high-frequency alternatives, in the sense that the variable A_1, A_2, A_3, A_4 have periodic hazard rates rather than constants.

We take $\beta = 0.2$, $\lambda_0(t) = 1$, $\Lambda_0(t) = t$, and $X = 0, 1, \cdots, 9$ with equal proportions. The censoring variable in each case is drawn from a uniform distribution such that the percentage of censorship is around 30%. We run for sample size $n = 50, 100, 150$, and use 1000 realizations of the Gaussian process to estimate the distribution of each statistic. We run 1000 replications for each DGP.

The results of the omnibus test, smooth test and Bonferroni test are shown in Table 2.1 and 2.2. The omnibus test is based on $\hat{R}_n(t, x)$. The smooth test is based on the reweighted sum of the first five component processes with equal weights. The Bonferroni test is based on the first five component tests. In all cases, the Bonferroni test behaves better than the omnibus test, especially for DGP 4-7, the high-frequency alternatives, for which the omnibus test has no power at all. In general, the smooth test also behaves better than the omnibus test.

Table 2.3 and 2.4 show the results of the first five component tests based on a single component process. We use bold type to indicate the test that has the largest power. From these results, it is clear to see how the obtained components reflect certain deviations. When the alternative gets more frequent in the time domain, the test based on the latter component behaves better.

Table 2.1: Estimated size and power of KS tests at 5%

	Cox			DGP1			DGP2			DGP3		
	$n = 50$	100	150	$n = 50$	100	150	$n = 50$	100	150	$n = 50$	100	150
omnibus	0.025	0.038	0.035	0.654	0.994	1.000	0.043	0.078	0.125	0.241	0.795	0.979
smooth	0.025	0.041	0.037	0.539	0.984	1.000	0.058	0.171	0.240	0.184	0.767	0.988
Bonferroni	0.034	0.046	0.037	0.737	0.998	1.000	0.072	0.244	0.377	0.727	0.994	1.000

	DGP4			DGP5			DGP6			DGP7		
	$n = 50$	100	150	$n = 50$	100	150	$n = 50$	100	150	$n = 50$	100	150
omnibus	0.022	0.071	0.095	0.085	0.213	0.348	0.030	0.063	0.076	0.039	0.080	0.139
smooth	0.106	0.231	0.354	0.163	0.547	0.801	0.073	0.256	0.476	0.102	0.332	0.509
Bonferroni	0.159	0.451	0.696	0.304	0.770	0.941	0.146	0.440	0.744	0.194	0.520	0.736

Table 2.2: Estimated size and power of CvM tests at 5%

	Cox			DGP1			DGP2			DGP3		
	$n = 50$	100	150	$n = 50$	100	150	$n = 50$	100	150	$n = 50$	100	150
omnibus	0.041	0.050	0.036	0.276	0.626	0.931	0.046	0.073	0.103	0.086	0.264	0.492
smooth	0.020	0.038	0.035	0.589	0.968	0.999	0.049	0.105	0.133	0.130	0.348	0.616
Bonferroni	0.030	0.052	0.035	0.813	1.000	1.000	0.086	0.268	0.419	0.782	0.993	1.000

	DGP4			DGP5			DGP6			DGP7		
	$n = 50$	100	150	$n = 50$	100	150	$n = 50$	100	150	$n = 50$	100	150
omnibus	0.037	0.040	0.038	0.025	0.038	0.031	0.035	0.036	0.035	0.032	0.036	0.024
smooth	0.116	0.203	0.231	0.159	0.476	0.640	0.050	0.162	0.245	0.076	0.216	0.383
Bonferroni	0.163	0.449	0.696	0.305	0.777	0.945	0.146	0.433	0.728	0.209	0.526	0.736

Table 2.3: Estimated size and power of KS component tests at 5%

	Cox			DGP1			DGP2			DGP3		
	$n=50$	100	150	$n=50$	100	150	$n=50$	100	150	$n=50$	100	150
$1st$	0.037	0.041	0.039	0.413	0.858	0.968	0.036	0.049	0.080	0.114	0.419	0.651
$2nd$	0.031	0.039	0.046	**0.911**	**0.999**	**1.000**	**0.155**	**0.340**	**0.468**	**0.881**	**0.999**	**1.000**
$3rd$	0.032	0.048	0.053	0.097	0.125	0.129	0.104	0.177	0.268	0.381	0.725	0.913
$4th$	0.042	0.051	0.055	0.059	0.272	0.702	0.028	0.049	0.043	0.049	0.068	0.128
$5th$	0.045	0.047	0.045	0.139	0.166	0.203	0.038	0.072	0.091	0.048	0.119	0.218

	DGP4			DGP5			DGP6			DGP7		
	$n=50$	100	150	$n=50$	100	150	$n=50$	100	150	$n=50$	100	150
$1st$	0.043	0.047	0.031	0.042	0.045	0.050	0.035	0.043	0.043	0.040	0.041	0.039
$2nd$	0.085	0.168	0.230	0.087	0.205	0.325	0.048	0.072	0.110	0.062	0.075	0.116
$3rd$	**0.251**	**0.548**	**0.782**	0.200	0.380	0.568	0.109	0.185	0.232	0.186	0.374	0.521
$4th$	0.096	0.160	0.204	**0.384**	**0.744**	**0.884**	**0.230**	**0.553**	**0.798**	0.198	0.359	0.495
$5th$	0.084	0.222	0.374	0.167	0.346	0.438	0.119	0.222	0.268	**0.213**	**0.462**	**0.554**

Table 2.4: Estimated size and power of CvM component tests at 5%

	Cox			DGP1			DGP2			DGP3		
	$n=50$	100	150	$n=50$	100	150	$n=50$	100	150	$n=50$	100	150
$1st$	0.044	0.050	0.046	0.617	0.960	0.998	0.038	0.081	0.121	0.249	0.671	0.884
$2nd$	0.038	0.044	0.044	**0.927**	**1.000**	**1.000**	**0.150**	**0.336**	**0.491**	**0.871**	**0.993**	**0.999**
$3rd$	0.036	0.056	0.057	0.124	0.137	0.164	0.105	0.197	0.299	0.346	0.663	0.882
$4th$	0.035	0.039	0.053	0.061	0.252	0.656	0.031	0.050	0.057	0.042	0.044	0.047
$5th$	0.041	0.047	0.041	0.148	0.172	0.201	0.046	0.069	0.088	0.038	0.059	0.065

	DGP4			DGP5			DGP6			DGP7		
	$n=50$	100	150	$n=50$	100	150	$n=50$	100	150	$n=50$	100	150
$1st$	0.038	0.041	0.044	0.032	0.047	0.045	0.041	0.047	0.037	0.042	0.052	0.042
$2nd$	0.061	0.109	0.140	0.118	0.291	0.436	0.042	0.060	0.096	0.065	0.071	0.099
$3rd$	**0.246**	**0.537**	**0.765**	0.161	0.287	0.419	0.107	0.154	0.193	0.164	0.360	0.522
$4th$	0.091	0.158	0.188	**0.358**	**0.763**	**0.889**	**0.206**	**0.545**	**0.797**	0.183	0.345	0.484
$5th$	0.102	0.210	0.359	0.160	0.340	0.428	0.115	0.231	0.265	**0.224**	**0.470**	**0.579**

2.5 Conclusion

We have used conditional PCD method to decompose the CUSUM martingale process in the hazard model with regression. The component processes provide a basis of more powerful specification tests. The decomposition is in the time domain, and each component process reflects certain deviations from the proportional hazard assumption.

However, these components do not help very much when the deviations come from misspecifications of the covariate effect, for example, missing variables or wrong link functions. To have more powerful tests against these deviations, the decomposition of the process $R_n(t,x)$ in the covariable domain is required. This leads us to the decomposition that has been developed in chapter 1. Hence, together with the result in chapter 1, we are able to obtain the components of $R_n(t,x)$ in both directions and these components help to improve the efficiency of the specification test in every respect.

2.6 Appendix: Proofs

Proof of Theorem 1:

Note that each f_j and g_j are bounded and differentiable. The tightness of $c_{n,j}$ follows from Lemma 1 in Lin, Wei and Ying (1993). It then follows from the multivariate CLT that the process converges weakly to a centered Gaussian process. The independence between $c_{\infty,j}$ and $c_{\infty,h}$ comes from the Gaussian property and conditional uncorrelation between z_{ij} and z_{ih}.

Proof of the asymptotic equivalence of $\hat{c}_{n,j}(t,x)$ and $\tilde{c}_{n,j}(t,x)$:

The asymptotic properties of $\hat{\beta}$ and $\hat{\Lambda}_0$ is given by Tsiatis (1981) and Andersen and Gill (1982). By taking the Taylor's expansion of $\hat{c}_{n,j}(t,x)$ and the score function

$U(\beta)$ at β_0, we have

$$\begin{aligned}
\hat{c}_{n,j}(t,x) &= n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \mathbb{1}_{\{X_i\le x\}}\hat{f}_j(t,X_i)\hat{g}_j(s,X_i)dM_i(s) \\
&\quad -n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \frac{\sum_{i=1}^n Y_i(s)e^{\beta_0^T X_i}\mathbb{1}_{\{X_i\le x\}}\hat{f}_j(t,X_i)\hat{g}_j(s,X_i)}{\sum_{i=1}^n Y_i(s)e^{\beta_0^T X_i}}dM_i(s) \\
&\quad -n^{-1}\sum_{i=1}^{n}\int_0^\infty Y_i(s)e^{\beta_0^T X_i}(X_i-\bar{X}(\beta_0,s))\lambda_0(s)\mathbb{1}_{\{X_i\le x\}}\hat{f}_j(t,X_i)\hat{g}_j(s,X_i)ds \\
&\quad \times \Sigma(\beta_0)^{-1}n^{-1/2}\sum_{i=1}^{n}\int_0^\infty (X_i-\bar{X}(\beta_0,s))dM_i(s) \\
&\quad +o_p(1).
\end{aligned}$$

By the strong consistency of $\hat{\beta}$, $\hat{\Lambda}_0$ and the Kaplan-Meier estimator, together with the continuous mapping theorem, $\hat{f}_j$ and $\hat{g}_j$ are strongly consistent. Hence, for the first term on the right-hand side of the above equation, by the martingale property and the strong consistency and boundness of $\hat{f}_j$ and $\hat{g}_j$, we have

$$\begin{aligned}
& E\Big[n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \Big(\mathbb{1}_{\{X_i\le x\}}\hat{f}_j(t,X_i)\hat{g}_j(s,X_i)-\mathbb{1}_{\{X_i\le x\}}f_j(t,X_i)g_j(s,X_i)\Big)dM_i(s)\Big]^2 \\
= \;& E\Big[n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \Big(\mathbb{1}_{\{X_i\le x\}}\hat{f}_j(t,X_i)\hat{g}_j(s,X_i)-\mathbb{1}_{\{X_i\le x\}}f_j(t,X_i)g_j(s,X_i)\Big)^2 Y_i(s)e^{\beta^T X_i}\lambda_0(s)ds\Big] \\
= \;& \int_0^\infty E\Big[\Big(\mathbb{1}_{\{X_i\le x\}}\hat{f}_j(t,X_i)\hat{g}_j(s,X_i)-\mathbb{1}_{\{X_i\le x\}}f_j(t,X_i)g_j(s,X_i)\Big)^2 Y_i(s)e^{\beta^T X_i}\Big]\lambda_0(s)ds \\
\to \;& 0,
\end{aligned}$$

thus

$$n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \Big(\mathbb{1}_{\{X_i\le x\}}\hat{f}_j(t,X_i)\hat{g}_j(s,X_i)-\mathbb{1}_{\{X_i\le x\}}f_j(t,X_i)g_j(s,X_i)\Big)dM_i(s)=o_p(1).$$

The same argument for the second term, since from the strong consistency of $\hat{f}_j$ and $\hat{g}_j$ and the uniform SLLN, we have

$$n^{-1}\sum_{i=1}^{n}Y_i(s)e^{\beta_0^T X_i}\mathbb{1}_{\{X_i\le x\}}(\hat{f}_j(t,X_i)\hat{g}_j(s,X_i)-f_j(t,X_i)g_j(s,X_i))=o_p(1).$$

For the third term, we have

$$n^{-1}\sum_{i=1}^{n}\int_0^\infty Y_i(s)e^{\beta_0^T X_i}(X_i-\bar{X}(\beta_0,s))\lambda_0(s)\mathbb{1}_{\{X_i\le x\}}(\hat{f}_j(t,X_i)\hat{g}_j(s,X_i)-f_j(t,X_i)g_j(s,X_i))ds=o_p(1),$$

and

$$n^{-1/2}\sum_{i=1}^{n}\int_0^{\infty}(X_i-\bar{X}(\beta_0,s))dM_i(s)\xrightarrow{d}N(0,\Sigma(\beta_0)).$$

Thus, $\hat{c}_{n,j}(t,x)$ and $\tilde{c}_{n,j}(t,x)$ have the same asymptotic distribution.

Proof of Theorem 2:

To show the tightness of $\hat{c}_{n,j}(t,x)$, it suffices to show the tightness of $\tilde{c}_{n,j}(t,x)$. Recall

$$\begin{aligned}\tilde{c}_{n,j}(t,x) &= n^{-1/2}\sum_{i=1}^{n}\int_0^{\infty}\Big[\mathbb{1}_{\{X_i\le x\}}f_j(t,X_i)g_j(s,X_i)-\tilde{l}(\beta_0,t,x,s)\Big]dM_i(s)\\ &\quad -A(t,x)\Sigma(\beta_0)^{-1}n^{-1/2}\sum_{i=1}^{n}\int_0^{\infty}(X_i-\tilde{X}(\beta_0,s))dM_i(s).\end{aligned}$$

From Lemma 1 in Lin, Wei and Ying (1993), the first term is tight. The second term is tight since

$$n^{-1/2}\sum_{i=1}^{n}\int_0^{\infty}(X_i-\tilde{X}(\beta_0,s))dM_i(s)$$

converges in distribution. It then follows from the multivariate CLT that $\hat{c}_{n,j}(t,x)$ converges weakly to a centered Gaussian process.

Chapter 3

Goodness-of-Fit Tests for Conditional Distributions

3.1 Introduction

In this article, we propose new specification tests for parametric conditional distribution models, i.e., we want to test

$$H_0 : F(\cdot \mid X) = F(\cdot \mid X, \theta) \ a.s. \ \text{ for some } \theta \in \Theta.$$

The covariable X can be a multidimensional vector and its distribution is unspecified. The alternative can be omnibus or directional. In the nonparametric testing literature, for example, Andrews (1997) proposed an omnibus test, which is consistent against all possible deviations, based on a CUSUM process of single event processes. His test is not distribution-free and is implemented by a parametric bootstrap. For the same question, Delgado and Stute (2008) provided a class of asymptotically distribution-free tests based on PCD of the multivariate empirical process. They used the Rosenblatt transformation to obtain independence between components of the empirical process and then applied Khmaladze martingale method to remove the effect caused by estimation. The distribution-free property together with the independence structure makes PCA for the conditional model possible.

In this article, we conduct PCA for testing conditional distributions in a different way. We propose new goodness-of-fit tests based on some components of the CUSUM process in Andrews (1997). These components are obtained through a conditional PCD method, which works for a general class of conditional models and fills the gap of PCA in testing conditional models. The components of the CUSUM process play a similar role with the classical PCs of the empirical process, hence behave as building blocks for goodness-of-fit tests. The difference is that these components are stochastic processes rather than random variables. Therefore we call them component processes to be more precise. Specifically, it consists of two steps to obtain the component processes, (i) derive PCD of the centered single event process conditional on the covariables, (ii) sum up the obtained PCs in the first step w.r.t. the observations of the covariables. It turns out that the CUSUM process can

be decomposed into a decreasing weighted sum of its component processes. Since the PCD is in the response variable domain, the obtained components are sensitive when detecting certain deviations from the specified distribution, especially, higher-frequency deviations are more reflected in latter components. The omnibus test, which is based on the original CUSUM process, down-weights the latter components heavily, thus, it has low power when detecting the high-frequency deviations. While we propose new goodness-of-fit tests, including tests based on each estimated component process and a Bonferroni test, which offset the power loss and outperform the omnibus test. Smooth tests based on reweighted sums of a few components are also constructed.

The conditional PCD method in this paper is applicable to a general class of conditional models. It will be our further work to clarify its application range. The structure of this chapter will be as follows. A brief introduction of Andrew's test is in section 3.2. Section 3.3 contains the main result: the conditional PCD, the asymptotic results of the component processes and the test statistics based on the component processes. Simulation studies illustrating the performance of our tests in the finite sample are presented in section 3.4.

3.2 Omnibus Test for Conditional Distributions

In the framework of regression analysis, consider a sample $\{Y_i, X_i\}, i = 1, \cdots, n$ of i.i.d. realizations of $\{Y, X\}$. Y is the real-valued response variable and X is the covariable vector. We are interested in the conditional distribution of Y conditional on X. To test the H_0, Andrews (1997) considered a CUSUM type process

$$R_n(y, x) := n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} M_i(y), \tag{3.1}$$

where

$$M_i(y) = \mathbb{1}_{\{Y_i \leq y\}} - F\left(y \mid X_i, \theta\right), \tag{3.2}$$

is the single event process after centering. Note that although we continue to take the notation M as in chapter2, $M_i(y)$ is not a martingale any more, but actually has the same covariance kernel with the Brownian Bridge. R_n is a CUSUM process of M_i's w.r.t. X_i's and it can be viewed as a CUSUM version of the classical empirical process for unconditional distribution. Andrews proposed an omnibus Kolmogorov-Smirnov type statistic based on the estimated process

$$\hat{R}_n(y,x) := n^{-1/2}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}\hat{M}_i(y),$$

where

$$\hat{M}_i(y) = \mathbb{1}_{\{Y_i \le y\}} - F\left(y \mid X_i, \hat{\theta}\right)$$

and $\hat{\theta}$ is an estimator of θ. Andrews's test takes the form of $\sup_{y,x}\left|\hat{R}_n(y,x)\right|$ and he has provided the weak convergence result of the test statistic under the pseudo-metric defined as (3.6) in his paper. The following assumptions of the conditional distribution model and the estimation are also required.

(A1). $F(y \mid X_i, \theta)$ is differentiable in θ on a neighborhood of θ_0, $\forall i \ge 1$.

(A2). $\sup_{(y,x)\in\mathbb{R}^2}\sup_{\theta:\|\theta-\theta_0\|\le r_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}F(y \mid X_i,\theta)\mathbb{1}_{\{X_i\le x\}} - \Delta_0(y,x)\right\| \to 0,\ a.s.$

for all sequences of positive constants $\{r_n : n \ge 1\}$ such that $r_n \to 0$, where $\Delta_0(y,z) = \int(\partial/\partial\theta)F(y \mid s,\theta_0)\mathbb{1}_{\{s\le x\}}dF_X(s)$ and F_X denotes the marginal distribution of X.

(A3). $\sup_{(y,x)\in\mathbb{R}^2}\|\Delta_0(y,x)\| < \infty$

and $\Delta_0(\cdot)$ is uniformly continuous on $\mathbb{R}^2$.

(A4). The parametric estimator has an expression as

$$n^{1/2}\left(\hat{\theta}-\theta_0\right) = n^{-1/2}\sum_{i=1}^{n} l(X_i,Y_i,\theta_0) + o_p(1),\ \text{conditional on } X\ a.s.$$

for some function l such that it is a measurable function and satisfies $\int l(y,x,\theta_0)F(dy \mid x,\theta_0) = 0$ for all x in the support of X and $\int l_0(x)F_X(dx) < \infty$, where $l_0(x) = \int\|l(y,x,\theta_0)\|^{2+\varepsilon}F(dy \mid x,\theta_0)$ for some $\varepsilon > 0$.

In regard to the assumption on the estimation (A4), for example, the maximum likelihood estimator satisfies. Under the null hypothesis and the above assumptions, $\hat{R}_n$ converges weakly to a Gaussian process and the KS test statistic has a limit distribution. In order to estimate the limit distribution of the test, he proposed a valid parametric bootstrap procedure.

In this paper, to simplify the notation, we only consider the univariate case, i.e., real-valued X, however, it can be any vector and all the arguments in this chapter work for multivariate X. In next section, we develop a decomposition of the process with the true value of the parameters $R_n(y, x)$ into a countable sum of component processes, and use these estimated component processes to construct new test statistics.

3.3 Tests based on Component Processes

3.3.1 Conditional Principal Component Analysis

Notice that the process R_n in (3.1) is bivariate with dependent components y and t. Hence, the explicit Karhunen-Loève representation of R_n is not available. We apply the conditional PCD method that has been developed in chapter 2 to do the decomposition in two steps, namely, first to get PCD of the single event process conditional on X, and then sum the obtained PCs up w.r.t. the observations of X.

Let us begin with the PCD of $M(y)$ conditional on X. Under the null model, the conditional covariance kernel of $M(y)$ conditional on X equals to the covariance kernel of a transformed Brownian Bridge, i.e.,

$$\mathbb{E}(M(y_1)M(y_2) \mid X) = F(y_1 \wedge y_2 \mid X, \theta_0) - F(y_1 \mid X, \theta_0)F(y_2 \mid X, \theta_0) \tag{3.3}$$

For each x the transformation function can be defined as

$$T(y, x) := F(y \mid X = x, \theta_0), \tag{3.4}$$

with the true value of the parameters. Then

$$\mathbb{E}(M(y_1)M(y_2) \mid X = x) = K(T(y_1, x), T(y_2, x)), \tag{3.5}$$

where $K(s,t) = s \wedge t - st$ is the covariance kernel of standard Brownian Bridge.

Notice that function T is non-decreasing in y, and $T(-\infty, x) = 0$, $T(\infty, x) = 1$. The eigenfunctions of $M(t)$ can be obtained through transformation. Let

$$\mu_j = \frac{1}{(\pi j)^2}, \quad \varphi_j(y) = \sqrt{2} sin(j\pi y), \quad j = 1, 2, \cdots$$

be the eigenvalues and eigenfunctions of the standard Brownian Bridge with covariance kernel $K(s,t)$. For each x, let f_j be the transformation

$$f_j(y, x) := \varphi_j(T(y, x)).$$

Therefore, for each fixed x, $\{f_j(\cdot, x)\}_{j=1}^{\infty}$ form an orthonormal basis of a subspace of $L^2(\mathbb{R}, T(\cdot, x))$, the Hilbert space of all square integrable functions on $\mathbb{R}$ with inner product

$$\langle \rho, g \rangle_x = \int_{\mathbb{R}} \rho(y) g(y) T(dy, x),$$

since

$$\begin{aligned} \langle f_j, f_h \rangle_x &= \int_{\mathbb{R}} \varphi_j \left(T(y, x)\right) \varphi_h \left(T(y, x)\right) T(dy, x) \\ &= \int_0^1 \varphi_j(u) \varphi_h(u) du = \begin{cases} 1 & j = h \\ 0 & j \neq h \end{cases}. \end{aligned}$$

Moreover, $\{f_j(\cdot, x)\}_{j=1}^{\infty}$ are the eigenfunctions of the covariance kernel $K(T(y_1, x), T(y_2, x))$ with associated eigenvalues $\{\mu_j\}_{j=1}^{\infty}$, i.e.,

$$\int_{\mathbb{R}} K(T(y_1, x), T(y_2, x)) f_j(y_1, x) T(dy_1, x) = \mu_j f_j(y_2, x).$$

By Mercer's theorem, the covariance kernel can be decomposed as

$$K(T(y_1, x), T(y_2, x)) = \sum_{j=1}^{\infty} \mu_j f_j(y_1, x) f_j(y_2, x). \tag{3.6}$$

The Karhunen-Loève representation of the centered single event process is

$$M_i(y) = \sum_{j=1}^{\infty} {\mu_j}^{1/2} z_{ij} f_j(y, X_i) \quad a.s., \ i = 1, \cdots, n \tag{3.7}$$

where

$$\begin{aligned} z_{ij} :=& {\mu_j}^{-1/2} \langle M_i, f_j(\cdot, X_i) \rangle_{X_i} \\ =& {\mu_j}^{-1/2} \int_{\mathbb{R}} M_i(y) f_j(y, X_i) T(dy, X_i). \end{aligned} \tag{3.8}$$

The z_{ij} is the j'th conditional principal component of $M_i(y)$ conditional on X_i. For each j and $j \neq h$, it has the following properties

$$\begin{aligned} \mathbb{E}(z_{ij} \mid X_i) &= 0, \\ \mathbb{E}(z_{ij}^2 \mid X_i) &= 1, \\ \mathbb{E}(z_{ij} z_{ih} \mid X_i) &= 0. \end{aligned} \tag{3.9}$$

That is, for each i, the PCs have conditional zero mean and unit variance and are uncorrelated with each other conditional on X. The next step is to sum up the obtained conditional PCs in the same way as how $R_n(y, x)$ has summed up $M_i(y)$. Simply by plugging the conditional PCD (3.7) into (3.1), we have the decomposition of R_n,

$$\begin{aligned} R_n(y, x) &= n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \left(\sum_{j=1}^{\infty} {\mu_j}^{1/2} z_{ij} f_j(y, X_i) \right) \\ &= \sum_{j=1}^{\infty} {\mu_j}^{1/2} \left[n^{-1/2} \sum_{i=1}^{n} z_{ij} \mathbb{1}_{\{X_i \leq x\}} f_j(y, X_i) \right]. \end{aligned}$$

We call the term in the square bracket the j'th component process of R_n and denote it as

$$c_{n,j}(y, x) := n^{-1/2} \sum_{i=1}^{n} z_{ij} \mathbb{1}_{\{X_i \leq x\}} f_j(y, X_i). \tag{3.10}$$

This result is summarized in the following proposition.

Proposition 1 *Under the null hypothesis, the processes* (3.1) *can be decomposed into a weighted sum of component processes, i.e.,*

$$R_n(y, x) = \sum_{j=1}^{\infty} {\mu_j}^{1/2} c_{n,j}(y, x). \tag{3.11}$$

The weights are the square root of the standard Brownian Bridge eigenvalues.

Actually, each PC z_{ij} in this distribution case takes an explicit form of

$$z_{ij} = \sqrt{2}\cos\left(j\pi T(Y_i, X_i)\right). \tag{3.12}$$

To see it, let us first define the cosin function g_j corresponding to f_j as

$$g_j(y, x) := \phi_j(T(y, x)) = \sqrt{2}cos\left(j\pi T(y, x)\right).$$

By applying integration by parts to the integral (3.8), we have

$$z_{ij} = \int_{-\infty}^{\infty} g_j(s, X_i)dM_i(s) = g_j(Y_i, X_i).$$

Repeating (3.9) we have

$$\mathbb{E}(g_j(Y_i, X_i) \mid X_i) = 0,$$

$$\mathbb{E}(g_j^2(Y_i, X_i) \mid X_i) = 1,$$

$$\mathbb{E}(g_j(Y_i, X_i)g_h(Y_i, X_i) \mid X_i) = 0.$$

Furthermore, for each $j \geq 1$, $g_j(Y_i, X_i)$ have the same conditional distribution on X_i, the Fourier transform of which equal to the Bessel-function of order zero. In another word, g_j is a conditional distribution-free transformation. We can rewrite the component processes as

$$c_{n,j}(y, x) = n^{-1/2}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} f_j(y, X_i)g_j(Y_i, X_i). \tag{3.13}$$

3.3.2 Asymptotic Theory of Component Processes

In this section, we develop asymptotic results of the component processes and the estimated ones. From (3.13), each component process is a sum of i.i.d. centered random functions with variance

$$H_j(y, x) := \mathbb{E}\left[\mathbb{1}_{\{X \leq x\}} f_j^2(y, X)\right] = \int_{-\infty}^{x} f_j^2(y, s)F_X(ds),$$

where $F_X(\cdot)$ denotes the distribution function of X. Following Andrews (1997), we have the below theorem.

Theorem 1 *Under the null hypothesis and (A1)-(A3), for each j, the process $c_{n,j}(y,x)$ converges weakly to a centered Gaussian process*

$$c_{n,j} \xrightarrow{d} c_{\infty,j}.$$

The limit Gaussian process $c_{\infty,j}$ has covariance structure

$$K(y_1,y_2,x_1,x_2) = \int_{-\infty}^{x_1\wedge x_2} f_j(y_1,s)f_j(y_2,s)F_X(ds).$$

Moreover, $c_{\infty,j}$ and $c_{\infty,h}$ are independent for $j \neq h$.

Now we consider the component process after estimation, i.e., the process

$$\hat{c}_{n,j}(y,x) := n^{-1/2}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\leq x\}}\hat{f}_j(y,X_i)\hat{g}_j(Y_i,X_i).$$

Here

$$\hat{f}_j(y,x) = \varphi_j\left(\hat{T}(y,x)\right),$$

and

$$\hat{g}_j(y,x) = \phi_j\left(\hat{T}(y,x)\right),$$

with $\hat{T}(y,x)$ being an estimator of the function $T(y,x)$. A natural consistent one is

$$\hat{T}(y,x) = F\left(y \mid X = x, \hat{\theta}\right). \tag{3.14}$$

In Appendix, it is shown that $\hat{c}_{n,j}(y,x)$ has the same asymptotic distribution as

$$\tilde{c}_{n,j}(y,x) := c_{n,j}(y,x) - A_j(y,x)n^{-1/2}\sum_{i=1}^{n} l(X_i,Y_i,\theta_0),$$

where

$$A_j(y,x) = {\mu_j}^{-1/2}\mathbb{E}\left[\mathbb{1}_{\{X\leq x\}}f_j(y,X)\int_{\mathbb{R}} F_\theta(y \mid X,\theta_0)f_j(y,X)T(dy,X)\right]$$

with F_θ denoting the partial derivative w.r.t. θ.

Theorem 2 *Under the null hypothesis and* $(A1)$-$(A4)$*, for each* $j = 1, 2, \cdots$*, the process* $\hat{c}_{n,j}(y, x)$ *converges weakly to a centered Gaussian process*

$$\hat{c}_{n,j} \xrightarrow{d} \tilde{c}_{\infty,j}.$$

The limit Gaussian process $\tilde{c}_{\infty,j}(y, x)$ *has covariance structure*

$$K(y_1, y_2, x_1, x_2) = \mathbb{E}\Big[\mathbb{1}_{\{X \le x_1\}} f_j(y_1, X) g_j(Y, X) - A_j(y_1, x_1) l(X, Y, \theta_0), \\ \mathbb{1}_{\{X \le x_2\}} f_j(y_2, X) g_j(Y, X) - A_j(y_2, x_2) l(X, Y, \theta_0)\Big].$$

In addition to the single component process, finite weighted sum of some component processes can also be used for model checking. Consider the first m component processes with weight $w = \{w_j\}_{j=1}^m$, i.e., the process $\sum_{j=1}^m w_j \hat{c}_{n,j}(y, x)$. It has the same asymptotic distribution with the process $\sum_{j=1}^m w_j \tilde{c}_{n,j}(y, x)$. Its asymptotic distribution is given in the following theorem.

Theorem 3 *Under the null hypothesis and* $(A1)$-$(A4)$*, for any given weight* $w = \{w_j\}_{j=1}^m$*, the process* $\sum_{j=1}^m w_j \hat{c}_{n,j}(y, x)$ *converges weakly to a centered Gaussian process*

$$\sum_{j=1}^m w_j \hat{c}_{n,j} \xrightarrow{d} \tilde{c}_\infty^w.$$

The limit Gaussian process $\tilde{c}_\infty^w(y, x)$ *has covariance structure*

$$K(y_1, y_2, x_1, x_2) = \mathbb{E}\Bigg[\sum_{j=1}^m w_j \left(\mathbb{1}_{\{X \le x_1\}} f_j(y_1, X) g_j(Y, X) - A_j(y_1, x_1) l(X, Y, \theta_0)\right), \\ \sum_{j=1}^m w_j \left(\mathbb{1}_{\{X \le x_2\}} f_j(y_2, X) g_j(Y, X) - A_j(y_2, x_2) l(X, Y, \theta_0)\right)\Bigg].$$

3.3.3 Test Statistics

The omnibus test in Andrews (1997) is based on the original CUSUM process. By the continuous mapping theorem, we have the following asymptotic distribution of

the Kolmogorov-Smirnov type statistics

$$KS_o = \sup_{y,x} \left| \hat{R}_n(y,x) \right| \xrightarrow{d} \sup_{y,x} \left| R_\infty(y,x) \right|.$$

The component processes we derived provide a basis of new specification tests for conditional distributions. We propose to construct Kolmogorov-Smirnov type statistics based on each component process, i.e., for each $j = 1, 2, \cdots$, we have the following, what we call, component tests,

$$KS_{nj} = \sup_{y,x} \left| \hat{c}_{n,j}(y,x) \right| \xrightarrow{d} \sup_{y,x} \left| \tilde{c}_{\infty,j}(y,x) \right|.$$

Note that in (3.11), the weight for the j'th component process is $\mu_j^{1/2}$ that decreases very rapidly in j. In consequence, the latter components are down-weighted in the original process. In fact, each component reflects certain aspect of a deviation from the null hypothesis. For example, high-frequency deviations are more reflected in the latter components. Therefore, the omnibus test, which gives low weights to latter components, has low power, while the tests based on latter components are specially designed for such high-frequency alternatives. In practice, the data should not be very frequent, hence we can focus on the first few components, say no more than ten in general.

In addition, smooth test statistics based on the reweighted sum of component processes can be constructed. If we give the components with equal weights and consider the sum of the first m components, the KS type statistics, for some fixed m, can be constructed as

$$KS_{nm} = \sup_{y,x} \left| \sum_{j=1}^{m} w_j \hat{c}_{n,j}(y,x) \right| \xrightarrow{d} \sup_{y,x} \left| \tilde{c}^w_\infty(y,x) \right|,$$

The smooth tests provide a compromise between the omnibus tests and the tests based on one component. The smooth test is the one that takes $w = (1, \cdots, 1)$. The test based on the j'th component process is the one that takes w as the j'th unit vector, i.e., $w = (0, \cdots, 1, \cdots, 0)$. However, the problem is that one has to choose a suitable w before model checking.

Actually, we can take into account the information together from some component processes by considering a Bonferroni test that behaves as an intersection of the component tests. Specifically, we run the first m component tests and record the decision for each one. Then we accept H_0 if all the m tests accept and reject H_0 if any of them gives us a rejection. Let $T_1, T_2, \cdots, T_m$ be the first m component tests with common size x. The Bonferroni test T^b is

$$T^b = \begin{cases} 0 & \text{if } T_1 = \cdots = T_m = 0, \\ 1 & \text{o.w.} \end{cases}$$

The probability of Bonferroni test to accept under H_0 is $P_0(T_1 = 0, \cdots, T_m = 0)$, and it admits the following inequality

$$\begin{aligned} P_0(T_1 = 0, \cdots, T_m = 0) &\geq P_0(T_1 = 0) + \cdots + P_0(T_m = 0) - (m-1) \\ &= (1-x) + \cdots + (1-x) - (m-1) \\ &= 1 - mx. \end{aligned}$$

For a significant level α, we could choose $x = \alpha/m$, then the size of the Bonferroni test will be

$$1 - P_0(T^b = 0) = 1 - P_0(T_1 = 0, \cdots, T_m = 0) \leq mx = \alpha,$$

i.e., the Bonferroni test has a bounded size of α.

Finally, to approximate the distribution of $\tilde{c}_{\infty,j}(y, x)$ and $\tilde{c}^w_\infty(y, x)$, we follow the suggestion of Andrews (1997) to run a parametric bootstrap procedure.

3.3.4 Other Weight Functions

As we have discussed in chapter 2, the weight function $\mathbb{1}_{\{X_i \leq x\}}$ can be replaced by any other function of X_i. For instance, if it is replaced by $\mathbb{1}_{\{X_i \in \mathcal{A}_l\}}$, where $\{\mathcal{A}_l\}_{l=1}^L$ is a partition of the real line, the omnibus test based on the processes

$$\gamma_{n,l}(y) = n^{-1/2} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in \mathcal{A}_l\}} M_i(y), \ l = 1, \cdots, L,$$

is consistent against all the possible deviations happened on the particular partition. By applying the conditional PCD, for each l, we have the decomposition of each process

$$\begin{aligned}\gamma_{n,l}(y) &= n^{-1/2}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\in\mathcal{A}_l\}}\left(\sum_{j=1}^{\infty}{\mu_j}^{1/2}z_{ij}f_j(y,X_i)\right)\\ &= \sum_{j=1}^{\infty}{\mu_j}^{1/2}\left[n^{-1/2}\sum_{i=1}^{n}z_{ij}\mathbb{1}_{\{X_i\in\mathcal{A}_l\}}f_j(y,X_i)\right]\\ &= \sum_{j=1}^{\infty}{\mu_j}^{1/2}\left[n^{-1/2}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\in\mathcal{A}_l\}}f_j(y,X_i)g_j(Y_i,X_i)\right]. \qquad (3.15)\end{aligned}$$

The component processes of $\gamma_{n,l}(y)$ are the ones in the square bracket of (3.15). The only difference from the previous $c_{n,j}$ is by making a change of the weight function from $\mathbb{1}_{\{X_i\le x\}}$ to $\mathbb{1}_{\{X_i\in\mathcal{A}_l\}}$. Although the omnibus test based on $\gamma_{n,l}(y)$'s, e.g., the KS supreme test, is consistent against misspecifications on the partition, constructing component tests, smooth tests and Bonferroni tests in the same way as in section 3.3.3 based on its component processes could be more informative and will improve efficiency when checking against certain departures. Similar to $c_{n,j}$, each component process of $\gamma_{n,l}(y)$ reveals certain high-frequency deviation from the null hypothesis.

Another weight function worth to mention is $\mathbb{1}_{\{\beta^T X_i\le z\}}$, which is especially useful in the multivariate X case. Now we have the process

$$\eta_n(\beta,z,y) = n^{-1/2}\sum_{i=1}^{n}\mathbb{1}_{\{\beta^T X_i\le z\}}M_i(y),$$

and its decomposition

$$\eta_n(\beta,z,y) = \sum_{j=1}^{\infty}{\mu_j}^{1/2}\left[n^{-1/2}\sum_{i=1}^{n}\mathbb{1}_{\{\beta^T X_i\le z\}}f_j(y,X_i)g_j(Y_i,X_i)\right]. \qquad (3.16)$$

The further analysis on $\eta_n(\beta,z,y)$ depends on the choice of β and will be discussed elsewhere.

3.4 Simulation Study

In this section, we study the behavior of our tests in the finite sample by considering three special cases. In all the cases, we run the simulation for sample size $n = 25, 50, 75$, and generate 1000 bootstrap samples to estimate the distribution of each statistic. We run 1000 replications for each DGP.

We consider conditional normal models with conditional mean and either fitted variance or fixed one. In testing the unconditional normal distribution, Durbin and Knott (1972) have suggested to examining each PC based on the observation that the first PC is sensitive to the mean shift, while the second PC is sensitive to the variance shift, and same patterns for the third and fourth PCs to skewness and kurtosis shifts. In our simulation, we generate DGPs that have mean shifts, variance shifts, skewness shifts and kurtosis shifts, respectively, from the conditional mean normal distribution, to see how the component processes serve in testing goodness-of-fit.

We observe similar patterns in the conditional normal case to that in Durbin and Knott (1972) for unconditional normal, i.e., the first four component tests are specialists for testing mean shift, variance shift, skewness shift and kurtosis shift, respectively. In each of the following three cases, we show the performance of Andrews's omnibus test, smooth test, Bonferroni test and the first four component tests. The smooth tests are based on the reweighted sum of the first four component processes with equal weights. The Bonferroni tests are also based on the first four component processes. We use bold type to indicate the component test that has the largest power.

(a). Conditional mean model,

$$H_0 : Y \mid X \sim \mathcal{N}(\beta_0 + \beta_1 X, 1).$$

The parameters are β_1 and β_2.

DGP1:

$$Y \mid X \sim \mathcal{N}(1 + X, 1).$$

DGP2: Variance shift,

$$Y \mid X \sim \mathcal{N}(1 + X, 1.5).$$

DGP3: Variance shift,

$$Y \mid X \sim \mathcal{N}(1 + X, 2).$$

Table 3.1: Estimated size and power of KS tests at 5%

	DGP1			DGP2			DGP3		
	$n = 25$	50	75	$n = 25$	50	75	$n = 25$	50	75
omnibus	0.051	0.044	0.050	0.072	0.089	0.157	0.130	0.263	0.432
smooth	0.025	0.032	0.031	0.078	0.148	0.261	0.204	0.465	0.696
Bonferroni	0.040	0.023	0.037	0.063	0.119	0.190	0.122	0.352	0.563

Table 3.2: Estimated size and power of KS component tests at 5%

	DGP1			DGP2			DGP3		
	$n = 25$	50	75	$n = 25$	50	75	$n = 25$	50	75
$1st$	0.043	0.046	0.052	0.067	0.073	0.087	0.096	0.083	0.123
$2nd$	0.037	0.030	0.033	0.067	**0.156**	**0.270**	**0.191**	**0.430**	**0.666**
$3rd$	0.037	0.033	0.043	0.030	0.049	0.027	0.022	0.023	0.027
$4th$	0.037	0.037	0.037	0.050	0.074	0.111	0.081	0.226	0.335

From the result in Table 3.1 and 3.2, for the variance shift in a conditional normal distribution with conditional mean and fixed variance, the test based on the second component process behaves the best. Whereas the first and third component tests have little power, this is due to the symmetry of normal distributions.

(b). Conditional mean with fitted variance,

$$H_0 : Y \mid X \sim \mathcal{N}(\beta_0 + \beta_1 X, \theta).$$

The parameters are β_1, β_2 and θ.

DGP4: Mean shift,

$$Y \mid X \sim \mathcal{N}(1 + X + \sin(2\pi X), 1).$$

DGP5: Mean shift,

$$Y \mid X \sim \mathcal{N}(1 + X + 1.5\sin(2\pi X), 1).$$

DGP6: Heteroscedasticity,

$$Y \mid X \sim \mathcal{N}(1 + X, 6(X - 0.5)^2 + 0.5).$$

DGP7: Heteroscedasticity,

$$Y \mid X \sim \mathcal{N}(1 + X, 12(X - 0.5)^2).$$

Table 3.3: Estimated power of KS tests at 5%

	DGP4			DGP5			DGP6			DGP7		
	$n = 25$	50	75	$n = 25$	50	75	$n = 25$	50	75	$n = 25$	50	75
omnibus	0.054	0.073	0.067	0.070	0.080	0.111	0.052	0.063	0.053	0.197	0.500	0.736
smooth	0.021	0.026	0.020	0.010	0.007	0.012	0.035	0.033	0.047	0.162	0.408	0.679
Bonferroni	0.035	0.073	0.116	0.055	0.147	0.275	0.063	0.062	0.085	0.310	0.688	0.907

Table 3.4: Estimated power of KS component tests at 5%

	DGP4			DGP5			DGP6			DGP7		
	$n = 25$	50	75	$n = 25$	50	75	$n = 25$	50	75	$n = 25$	50	75
$1st$	**0.071**	**0.131**	**0.230**	**0.116**	**0.307**	**0.550**	0.062	0.069	0.076	0.127	0.159	0.181
$2nd$	0.036	0.036	0.040	0.019	0.027	0.029	0.053	**0.091**	**0.125**	**0.359**	**0.749**	**0.947**
$3rd$	0.026	0.035	0.033	0.032	0.035	0.028	0.032	0.049	0.034	0.087	0.134	0.137
$4th$	0.033	0.033	0.039	0.017	0.028	0.029	0.045	0.027	0.039	0.129	0.225	0.390

Under the null hypothesis of a conditional normal model with conditional mean and fitted variance, the first two departures we take are from a nonlinear mean shift and the next two departures are from the conditional variance, which is called heteroscedasticity. As shown in Table 3.3 and 3.4, the first component test is a specialist for checking mean shift, and it behaves much better than the omnibus test,

while the second component test is a specialist for checking variance shift. Besides, the Bonferroni tests also behave better than the omnibus tests.

(c). Conditional mean with fitted variance,

$$H_0 : Y \mid X \sim \mathcal{N}(\beta_0 + \beta_1 X, \theta).$$

The parameters are β_1, β_2 and θ.

We take the alternatives that have the below conditional distribution function

$$F(y \mid x, \beta_0, \beta_1, \theta, \gamma_3, \gamma_4) = \Phi(y-1-x)+\gamma_3 \sin(3\pi\Phi(y-1-x))+\gamma_4 \sin(4\pi\Phi(y-1-x)),$$

where $\Phi(\cdot)$ is the distribution function of standard normal variable. This distribution is a conditional version of (8.4) in DKT and γ_3 and γ_4 indicate deviations in skewness and kurtosis from the normal distribution.

DGP8: Skewness shift, $\gamma_3 = 0.1, \gamma_4 = 0$.

DGP9: Skewness shift, $\gamma_3 = 0.2, \gamma_4 = 0$.

DGP10: Kurtosis shift, $\gamma_3 = 0, \gamma_4 = 0.1$.

DGP11: Kurtosis shift, $\gamma_3 = 0, \gamma_4 = 0.2$.

Again under the null hypothesis of a conditional normal model with conditional mean and fitted variance, we consider two departures from skewness and two departures from kurtosis. The results in Table 3.5 and 3.6 shows that the third and fourth component tests are specialists for checking skewness and kurtosis shifts, respectively, and they have much larger power than the omnibus tests. The Bonferroni tests also outperform the omnibus tests very much. These results are consistent with the ones in Durbin and Knott (1972), except that we have extended the testing for distributions to conditional cases.

Table 3.5: Estimated power of KS tests at 5%

	DGP8			DGP9			DGP10			DGP11		
	$n=25$	50	75	$n=25$	50	75	$n=25$	50	75	$n=25$	50	75
omnibus	0.067	0.213	0.469	0.084	0.297	0.525	0.102	0.173	0.261	0.114	0.193	0.274
smooth	0.243	0.677	0.923	0.500	0.912	0.991	0.194	0.630	0.894	0.265	0.690	0.919
Bonferroni	0.430	0.860	0.985	0.753	0.994	1.000	0.256	0.701	0.900	0.277	0.681	0.888

Table 3.6: Estimated power of KS component tests at 5%

	DGP8			DGP9			DGP10			DGP11		
	$n=25$	50	75	$n=25$	50	75	$n=25$	50	75	$n=25$	50	75
1*st*	0.176	0.449	0.675	0.272	0.592	0.804	0.051	0.049	0.042	0.041	0.046	0.035
2*nd*	0.179	0.305	0.419	0.346	0.571	0.683	0.074	0.131	0.174	0.055	0.070	0.085
3*rd*	**0.471**	**0.846**	**0.954**	**0.677**	**0.911**	**0.965**	0.107	0.167	0.187	0.011	0.144	0.157
4*th*	0.136	0.314	0.434	0.343	0.638	0.825	**0.332**	**0.701**	**0.914**	**0.335**	**0.723**	**0.902**

3.5 Conclusion

We have used the conditional PCD method to decompose the CUSUM process of centered single event process in conditional distribution models. The component processes provide a basis of more powerful specification tests. The decomposition is in the support of the response variable and each component process reflects certain deviations from the null hypothesis. The special roles of the component processes have been shown through simulations in the conditional normal models. We observe that the test based on the latter component process is capable to detect a shift in higher order moment.

3.6 Appendix: Proofs

Proof of the asymptotic equivalence of $\hat{c}_{n,j}(y,x)$ and $\tilde{c}_{n,j}(y,x)$:

First note that the sine and cosin functions are bounded and differentiable. Since $\hat{\theta}$ is root n-consistent, $\hat{f}_j$ and $\hat{g}_j$ are uniformly consistent.

We can write

$$\begin{aligned}
&\hat{c}_{n,j}(y,x) - n^{-1/2}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}} f_j(t, X_i) g_j(Y_i, X_i) \\
&= \left(\hat{c}_{n,j}(y,x) - n^{-1/2}\sum_{i=1}^{n} \int_0^{\infty} \mathbb{1}_{\{X_i \le x\}} \hat{f}_j(t, X_i)\hat{g}_j(s, X_i) dM_i(s) \right) \\
&+ \left(n^{-1/2}\sum_{i=1}^{n} \int_0^{\infty} \mathbb{1}_{\{X_i \le x\}} \hat{f}_j(t, X_i)\hat{g}_j(s, X_i) dM_i(s) - n^{-1/2}\sum_{i=1}^{n} \int_0^{\infty} \mathbb{1}_{\{X_i \le x\}} f_j(t, X_i) g_j(s, X_i) dM_i(s) \right).
\end{aligned}$$

For the first differnce, take the Taylor's expansion of $\hat{c}_{n,j}(y,x)$ and then we have

$$\begin{aligned}
\hat{c}_{n,j}(y,x) &= n^{-1/2}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{1}_{\{X_i \le x\}} \hat{f}_j(t, X_i)\hat{g}_j(s, X_i) dM_i(s) \\
&\quad -{\mu_j}^{-1/2} n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{X \le x\}} \hat{f}_j(y, X) \int_{-\infty}^{\infty} F_\theta(y \mid X, \theta_0) \hat{f}_j(y, X) \hat{T}(dy, X) \\
&\quad \times n^{-1/2} \sum_{i=1}^{n} l(X_i, Y_i, \theta_0) + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \mathbb{1}_{\{X_i \le x\}} \hat{f}_j(t, X_i)\hat{g}_j(s, X_i) dM_i(s) \\
&\quad -A_j(y,x) n^{-1/2} \sum_{i=1}^{n} l(X_i, Y_i, \theta_0) + o_p(1)
\end{aligned}$$

For the second difference, let us denote

$$K(s) = s - s^2,$$

then

$$\begin{aligned}
& E\Big[n^{-1/2}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \Big(\mathbb{1}_{\{X_i \le x\}} \hat{f}_j(y, X_i)\hat{g}_j(s, X_i) - \mathbb{1}_{\{X_i \le x\}} f_j(y, X_i) g_j(s, X_i) \Big) dM_i(s) \Big]^2 \\
= \; & E\Big[n^{-1/2}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \Big(\mathbb{1}_{\{X_i \le x\}} \hat{f}_j(y, X_i)\hat{g}_j(s, X_i) - \mathbb{1}_{\{X_i \le x\}} f_j(y, X_i) g_j(s, X_i) \Big)^2 K(T(ds, X_i)) \Big] \\
= \; & \int_{-\infty}^{\infty} E\Big[\Big(\mathbb{1}_{\{X_i \le x\}} \hat{f}_j(y, X_i)\hat{g}_j(s, X_i) - \mathbb{1}_{\{X_i \le x\}} f_j(y, X_i) g_j(s, X_i) \Big)^2 K(T(ds, X_i)) \Big] \\
\to \; & 0.
\end{aligned}$$

Hence the second difference is negligible, i.e.,

$$n^{-1/2}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \Big(\mathbb{1}_{\{X_i \le x\}} \hat{f}_j(y, X_i)\hat{g}_j(s, X_i) - \mathbb{1}_{\{X_i \le x\}} f_j(y, X_i) g_j(s, X_i) \Big) dM_i(s) = o_p(1).$$

Thus,

$$\hat{c}_{n,j}(y,x) = n^{-1/2}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} f_j(t, X_i) g_j(Y_i, X_i) - A_j(y,x) n^{-1/2} \sum_{i=1}^{n} l(X_i, Y_i, \theta_0) + o_p(1).$$

The weak convergence of $c_{n,j}(y,x)$ and $\hat{c}_{n,j}(y,x)$ is then easy to obtain.

Bibliography

[1] Aalen, O. O. (1980). A model for non-parametric regression analysis of life times. *Mathematical Statistics and Probability Theory (eds W. Klonecki, A. Kozek and J. Rosinski)*, 1–25. Lecture Notes in Statistics, vol. 2, Springer-Verlag, New York.

[2] Aalen, O. O., Borgan, O. and Gjessing, H. (2008). *Survival and event history analysis: a process point of view.* Springer Science & Business Media.

[3] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100–1120.

[4] Andrews, D. W. (1997). A conditional Kolmogorov test. *Econometrica*, 1097-1128.

[5] Anh, T. L. and Stute, W. (2012). Principal Component Analysis of Martingale Residuals. *Indian Statist. Assoc.*

[6] Bai, J. (2003). Testing parametric conditional distributions of dynamic models. *Review of Economics and Statistics*. 85(3), 531-549.

[7] Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika*, 65–74.

[8] Bickel, P. J., Klaassen, C. A., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models.*Baltimore: Johns Hopkins University Press.

[9] Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *The Annals of Mathematical Statistics*, 1656–1670.

[10] Bierens, H. J. and Ploberger, W. (1971). Asymptotic theory of integrated conditional moment tests. *Econometrica*, 1129-1151.

[11] Billingsley, P. (2013). *Convergence of probability measures*, John Wiley & Sons.

[12] Breslow, N. (1974). Covariance analysis of censored duration data. *Biometrics*, 89–99.

[13] Chen, K. and Jin, Z. and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 659–668.

[14] Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 835–845.

[15] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187–220.

[16] Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

[17] Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian journal of statistics*, 1–23.

[18] Delgado, M. A. and Stute, W. (2008). Distribution-free specification tests of conditional models. *Journal of Econometrics*, 37–55.

[19] Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics*, 1(2), 279-290.

[20] Durbin, J. and Knott, M. (1972). Components of Cramér-von Mises statistics I. *Journal of the Royal Statistical Society. Series B (Methodological)*, 290–307.

[21] Durbin, J., Knott, M. and Taylor, C. C. (1975). Components of Cramér-von Mises statistics. II. *Journal of the Royal Statistical Society. Series B (Methodological)*, 216–237.

[22] Escanciano, J. C. (2009). On the lack of power of omnibus specification tests. *Econometric Theory*, 25(1), 162-194.

[23] Eubank, R. L., and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *The Annals of Statistics*, 1412-1425.

[24] Eubank, R. L., and LaRiccia, V. N. (1992). Asymptotic comparison of Cramer-von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *The Annals of Statistics*, 20(4), 2071-2086. .

[25] Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and duration analysis*. Wiley, New York.

[26] Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3), 515-526.

[27] Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv fr matematik*, 1(3), 195-277.

[28] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21(4), 1926-1947.

[29] Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69(3), 553-566.

[30] Huber, P. J. (1985). Projection pursuit. *The annals of Statistics*, 435-475.

[31] Inglot, T. and Ledwina, T. (1996). Asymptotic optimality of data-driven Neyman's tests for uniformity. *The Annals of Statistics*, 24(5), 1982-2019.

[32] Janssen, A. (2000). Global power functions of goodness of fit tests. *The Annals of Statistics.*, 28(1), 239-253.

[33] Jennrich, R. I. (1969). Asymptotic properties of non-linear least-squares estimators. *Ann. Math. Statist.*, 40(2) 633-643.

[34] Kac, M. and Siegert, A. J. F. (1947). An explicit representation of a stationary Gaussian process. *Ann. Math. Statist.*, 18(3), 438-442.

[35] Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, (Vol. 360). John Wiley & Sons.

[36] Kallenberg, W. C. and Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association*, 92(439), 1094-1104.

[37] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457-481.

[38] Khmaladze, E. V. (1981). Martingale approach to the goodness-of-fit tests. *Theory of Probability and Applications*, 26, 246-265.

[39] Khmaladze, E. V. (1993). Goodness of fit problem and scanning innovation martingales. *The Annals of Statistics*, 21(2), 798-829.

[40] Koul, H. L. and Stute, W. (1999). Nonparametric model checks for time series. *The Annals of Statistics*, 27(1), 204-236.

[41] Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, 89(427), 1000-1005.

[42] Lin, D. Y. (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *Journal of the american statistical association*, 86(415), 725-728.

[43] Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American statistical Association*, 84(408), 1074-1078.

[44] Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572.

[45] Martinussen, T. and Scheike, T. H. (2007). *Dynamic regression models for duration data*. Springer Science & Business Media.

[46] Marzec, L. and Marzec, P. (1997). Generalized martingale-residual processes for goodness-of-fit inference in Cox's type regression models. *The Annals of Statistics*, 25(2), 683-714.

[47] Neyman, J. (1937). Smooth test for goodness of fit. *Scandinavian Actuarial Journal*, 1937(3-4), 149-199.

[48] Pollard, D. (2012). *Convergence of stochastic processes*. Springer Science & Business Media.

[49] Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, 65(1), 167-179.

[50] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3), 470-472.

[51] Schoenfeld, D. A. (1977). Asymptotic properties of tests based on linear combinations of the orthogonal components of the Cramér-von Mises statistic. *The Annals of Statistics*, 1017-1026.

[52] Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, 67(1):145–153.

[53] Schoenfeld, D. (1980). Tests based on linear combinations of the orthogonal components of the Cramér-von Mises statistic when parameters are estimated. *The Annals of Statistics*, 8(5), 1017-1022.

[54] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 239–241.

[55] Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, 45(1):89–103.

[56] Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, 613–641.

[57] Stute, W., Thies, S. and Zhu, L. X. (1998). Model checks for regression: an innovation process approach. *The Annals of Statistics*, 26(5), 1916-1934.

[58] Stute, W., Xu, W. L. and Zhu, L. X. (2008). Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika*, 95(2), 451-467.

[59] Stute, W. and Zhu, L. X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics*, 29(3), 535-545.

[60] Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 147–160.

[61] Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *The Annals of Statistics*, 93–108.

[62] Wei, L. J. (1984). Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association*, 649-652.

[63] Zheng, J. X. (2000). A consistent test of conditional parametric distributions. *Econometric Theory*, 16(5), 667-691.