Integrated conditional moment test and beyond: 
when the number of covariates is divergent *

Falong Tan¹ and Lixing Zhu²

¹ College of Finance and Statistics, Hunan University, Changsha, China
² Department of Mathematics, Hong Kong Baptist University, Hong Kong

Abstract

The classic integrated conditional moment test is a promising method for testing regression model misspecification. However, it severely suffers from the curse of dimensionality. To extend it to handle the testing problem for parametric multi-index models with diverging number of covariates, we investigate three issues in inference in this paper. First, we study the consistency and asymptotically linear representation of the least squares estimator of the parameter matrix at faster rates of divergence than those in the literature for nonlinear models. Second, we propose, via sufficient dimension reduction techniques, an adaptive-to-model version of the integrated conditional moment test. We study the asymptotic properties of the new test under both the null and alternative hypothesis to examine its ability of significance level maintenance and its sensitivity to the global and local alternatives that are distinct from the null at the fastest possible rate in hypothesis testing. Third, we derive the consistency of the bootstrap approximation for the new test in the diverging dimension setting. The numerical studies show that the new test can very much enhance the performance of the original ICM test in high-dimensional scenarios. We also apply the test to a real data set for illustrations.

Key words: Adaptive-to-model test, dimension reduction, integrated conditional moment test, least squares estimation, model checking, projection-based technique.

*Lixing Zhu is a Chair professor of Department of Mathematics at Hong Kong Baptist University, Hong Kong, China. He was supported by a grant from the University Grants Council of Hong Kong, Hong Kong, China.
1 Introduction

To test for model misspecification in regression analysis, integrated conditional moment (ICM) test (Bierens 1982; Bierens and Ploberger 1997) is a classic approach that attracts, when the dimension \( p \) of the covariate vector is fixed, many follow-up studies and develops many variants in the literature. The research described in this paper is motivated by extending it to handle the goodness-of-fit testing for parametric multiple-index regression models when the dimension \( p \) of the covariate vector is large regarding it as a diverging number as the sample size \( n \) goes to infinity. As the classic ICM test suffers severely from the curse of dimensionality that will be described below, we then try to solve three problems in inference. First, we will study at which rate of divergence of \( p \), the asymptotically linear presentation of the least squares estimation of the parameters in the regression model can hold. Second, we develop an adaptive-to-model version of the ICM test such that it can greatly avoid the curse of dimensionality. Third, as the limiting null distribution is in general intractable, we investigate the consistency of bootstrap approximation in this diverging setting. As everything in this paper relies on the sample size \( n \), we stress \( n \) in the notations.

We begin with describing the model checking problem in detail. Consider a parametric multiple-index model

\[ Y = g(\beta_0^\top X, \vartheta_0) + \varepsilon \quad (1.1) \]

where \( Y \) represents the real-valued response variable with the \( p \)-dimensional explanatory variable \( X \), \( g \) is a given smooth function, \( \vartheta_0 \in \mathbb{R}^l \) and \( \beta_0 = (\beta_{10}, \ldots, \beta_{d0}) \in \mathbb{R}^{p \times d} \) is the unknown matrix of regression parameters, \( \varepsilon = Y - E(Y|X) \) is the unpredictable part of \( Y \) given \( X \), and the notation \( \top \) denotes the transpose. Without loss of generality, we assume that \( \beta_{10}, \ldots, \beta_{d0} \) are orthogonal.

When the dimension \( p \) of covariates is fixed, there are a number of proposals in the literature for testing the parametric regression models consistently which can also be applied to test the null hypothesis of model (1.1). We just list a few such as Härde and Mammen (1993), Zheng (1996), Dette (1999), Fan and Huang (2001), Horowitz and Spokoiny (2001), Koul and Ni (2004) and van Keilegom et al. (2008) by using nonparametric estimation methods. They are called local smoothing tests. Some of tests in this class have tractable limiting null distributions, but the use of nonparametric regression estimations usually cause them to suffer severely from the curse of dimensionality in high-dimensional cases. Guo et al. (2016) gave some detailed comments on this issue. The empirical process-based tests are constructed in terms of converting conditional expectations \( E(\varepsilon|X) = 0 \) to infinite and parametric unconditional orthogonality restrictions,
that is
\[
E(\varepsilon|X) = 0 \iff E[\varepsilon w(X,t)] = 0 \quad \forall \, t \in \Gamma,
\]
where \(\Gamma\) is some proper space. There exist several parametric families \(w(X,t)\) such that the equivalence (1.2) holds; see Bierens and Ploberger (1997) and Escanciano (2006b) for more details on the primitive conditions of \(w(\cdot,t)\) to satisfy this equivalence. The indicator function \(w(X,t) = I(X \leq t)\) is commonly used as a weight function in the literature; see, e.g., Stute (1997), Stute, González Manteiga and Presedo Quindimil (1998), Stute, Thies and Zhu (1998), Zhu (2003), Khmadladze and Koul (2004), Stute, Xu and Zhu (2008), among many others.

Bierens (1982) used the characteristic function \(w(X,t) = \exp(it^\top X)\) as the weight function and constructed an integrated conditional moment (ICM) test, where \(i = \sqrt{-1}\) denotes the imaginary unit. The test statistic is given by
\[
ICM_n = \int_{\Gamma} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \hat{\eta}_j \exp(it^\top \Phi(X_j)) \right|^2 d\mu(t),
\]
where \(\hat{\eta}_j\) is the residual, \(\Gamma\) is a compact subset in \(\mathbb{R}^p\) with non-empty interior, \(\Phi(\cdot)\) is a bounded smoothing function, and \(\mu\) is a probability measure on \(\Gamma\). Note that \(ICM_n\) integrates over a compact subset \(\Gamma \subset \mathbb{R}^p\). It is well known that high-dimensional numerical integrations are extremely difficult to handle in practice. When \(p\) is large, the computation of the integral in (1.3) or its approximation becomes difficult, time consuming and expensive. Escanciano (2006a) suggested to choose the standard normal measure in (1.3) and the high-dimensional numerical integrations can be avoided. The test statistic \(ICM_n\) then becomes
\[
ICM_n = \int_{\mathbb{R}^p} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \hat{\eta}_j \exp(it^\top X_j) \right|^2 \phi(t) dt = \frac{1}{n} \sum_{j,k=1}^{n} \hat{\eta}_j \hat{\eta}_k \exp(-\frac{1}{2} \|X_j - X_k\|^2),
\]
where \(\phi(t)\) is the standard normal density on \(\mathbb{R}^p\). Escanciano (2006a, 2006b) commented that Bierens’ test (1982) is less impacted by the dimension \(p\) as it is based on one-dimensional projections. Although it does perform well in many cases, we found that Bierens’ test (1982) still suffers from the dimensionality problem in some other cases. More specifically, when \(p\) is large it has difficulty to maintain the significance level and is less powerful. We give a simple simulation study to show this phenomenon by the following toy example:
\[
H_1 : Y = \beta_0^\top X + a \exp(-\beta_0^\top X^2) + \varepsilon;
\]
where \(\beta_0 = (1, \cdots, 1)^\top/\sqrt{p}\) and \(X\) is \(N(0, I_p)\) independent with the standard normal distributed error \(\varepsilon\). The significance level is set to be \(\alpha = 0.05\). The constant \(a = 0\) corresponds to the null and \(a \neq 0\) to the alternatives. The empirical sizes and powers of Bierens’ test (1982) \(ICM_n\) are presented in Figure 1. We can find that not only the empirical powers drop quickly as the
dimension $p$ increases, but also the empirical sizes are almost zero for large $p$. This means that the $ICM_n$ test does not work even when the dimension is moderate.

To better understand this phenomenon in large dimension scenarios, we take a special case for illustration. Suppose that the components $\{X_{ik} : k = 1, \cdots, p\}$ of $X_i$ are independent and identically distributed. Note that $\|X_i - X_j\|^2 = \sum_{k=1}^{p} (X_{ik} - X_{jk})^2$. If the dimension $p$ tends to infinity, it follows that $(1/p)\|X_i - X_j\|^2$ tends to $2\text{Var}(X_{i1})$. Thus $\exp\{-\|X_i - X_j\|^2/2\}$ tends to 0 with probability 1. This means that the Bierens’ test statistic tends to be small under both the null and alternatives for large $p$.

The purpose of this paper is to extend the Bierens’ (1982) ICM test to the diverging dimension setting and simultaneously avoid the curse of dimensionality. As we need the residuals $Y_j - g(\hat{\beta}^\top X_j, \hat{\vartheta})$ to construct the test statistic where $\hat{\beta}$ and $\hat{\vartheta}$ are estimators of $\beta_0$ and $\vartheta_0$ respectively, we first study the asymptotic properties of the estimator $(\hat{\beta}, \hat{\vartheta})$ under both the null and alternatives in the divergent dimension cases. This problem has been well investigated for linear regression models, see Huber (1973), Portnoy (1984, 1985) and Zou and Zhang (2009), etc. Nevertheless, for nonlinear models, it has rarely been studied in the literature. Under some wild conditions, Tan and Zhu (2019a) obtained the estimation consistency and the asymptotically linear representation of $(\hat{\beta}, \hat{\vartheta})$ at the rates $p = o(n^{1/4})$ and $p = o(n^{1/5})$ of divergence respectively. In this paper, with the help of some high dimensional empirical processes techniques developed in Tan and Zhu (2019b), we will greatly improve these rates of divergence to $p = o(n^{1/2})$ and $p = o(n^{1/3}/\log n)$ and obtain the uniformly asymptotically linear representation of $(\hat{\beta}, \hat{\vartheta})$. These improvements enhance our ability to handle higher dimension in model checking problems. Under the conditions we design, the rates are the best in the relevant research in the literature to the best of our knowledge. Shi et al. (2019) obtained similar results for generalized linear models with fixed design in the diverging setting. But, as we actually deal with general
nonlinear models with random design here, their approach is significantly different from ours.

To attack the curse of dimensionality, we notice that the main reason that the Bierens’ (1982) test statistic $ICM_n$ does not work is that the weight function $\exp(-\|X_j - X_k\|^2/2)$ in (1.4) degenerates to zero at an exponential rate when the dimension $p$ tends to infinity. To avoid this problem, we adopt projection pursuit and sufficient dimension reduction (SDR) techniques to reduce the original dimension $p$ to a much low dimension and then develop an adaptive-to-model version of the classic $ICM_n$ test under the above convergence rates. The asymptotic properties of the proposed test are investigated under both the null and alternative hypothesis when $p = o(n^{1/3}/\log n)$. We also show that the test is consistent against all global alternatives and can detect the local alternatives converging to the null at the parametric rate $1/\sqrt{n}$ in this diverging dimensional scenario. This means that the test is sensitive to the local alternatives at a parametric rate even when the dimension $p$ is divergent. As the limiting null distribution of the proposed test admits a complicated structure, its critical values generally cannot be tabulated in practice. Thus we propose a bootstrap method to approximate the asymptotic null distribution of the test. We also show that the consistency of the bootstrap approximation remains valid when $p = o(n^{1/3}/\log n)$. This extends the results of Stute, González Manteiga and Presedo Quindimil (1998) to the diverging dimension cases.

To implement the adaptive-to-model test construction, we note that model (1.1) under the null admits a dimension reduction structure. Thus to make full use of the model structure under both the null and the alternative hypothesis, consider the following general alternative model

$$Y = G(B^T X) + \varepsilon, \quad (1.5)$$

where $E(\varepsilon|X) = 0$, $G(\cdot)$ is an unknown smooth function, and $B$ is a $p \times q$ orthonormal matrix such that $S_{E(Y|X)} = \text{span}(B)$. Here $S_{E(Y|X)}$ is the central mean space of $Y$ with respect to $X$ which will be specified in Section 3. Note that this model is even more general than the nonparametric model $Y = G(X) + \varepsilon$ that is a special case when $B$ is an $p \times p$ identity matrix with $q = p$.

The rest of this article is organised as follows. In Section 2 we present the asymptotic results of the least squares estimation of the parameters under the null and alternative models. Section 3 describes the basic test construction. Since the sufficient dimension reduction techniques are crucial for the model adaptation property, we give a short review in this section. To alleviate the computational complexity of the test statistic, we also discuss the choice of the weight functions in this section. Section 4 is devoted to present the asymptotic properties of the test under both the null and alternatives. In Section 5, the simulation results are reported and a real data analysis is used as an illustration of application. Appendix contains the technical proofs for the
theoretical results.

2 Parametric estimation

In this section, we consider parameter estimation for the model (1.1), and also study its asymptotic properties in the diverging dimension setting. Recall the null model (1.1)

\[ Y = g(\beta_0^T X, \vartheta_0) + \varepsilon. \]

To estimate the unknown parameters \((\beta_0, \vartheta_0)\), we here restrict ourselves to the ordinary least squares method. For notational simplicity, define \(\theta = [\text{vec}(\beta)^\top, \vartheta_0^\top]^\top\) where \(\text{vec}(\beta) = [\beta_1^\top, \cdots, \beta_d^\top]^\top\).

Then we rewrite the null hypothesis \(H_0\) in the following parametric regression format:

\[ Y = g(\theta_0, X) + \varepsilon \text{ for some } \theta_0 \in \mathbb{R}^m, \]

where \(m = pd + l\), \(E(\varepsilon|X) = 0\) and the number \(p\) diverges as the sample size \(n\) tends to infinity.

Next we consider the ordinary least squares estimator of \(\theta_0\) and its asymptotic properties. Let

\[ \hat{\theta} = \arg \min_{\theta \in \mathbb{R}^m} \sum_{i=1}^n [Y_i - g(\theta, X_i)]^2. \] (2.1)

To analyse the asymptotic properties of \(\hat{\theta}\), set

\[ \tilde{\theta}_0 = \arg \min_{\theta \in \mathbb{R}^m} E[Y - g(\theta, X)]^2. \] (2.2)

To obtain the asymptotic properties of \(\hat{\theta}\) when the dimension of \(\theta\) is divergent, we need some regularity conditions. In the following, \(C\) always stands for a constant which may be different from place to place.

(A1). Suppose the function \(g(\theta, x)\) admits second derivatives in \(\theta\). Let

\[ \psi(\theta, x, y) = [y - g(\theta, x)]g'(\theta, x) = [y - g(\theta, x)] \frac{\partial g(\theta, x)}{\partial \theta}, \]

\[ \psi_j(\theta, x, y) = [y - g(\theta, x)]g_j'(\theta, x) = [y - g(\theta, x)] \frac{\partial g(\theta, x)}{\partial \theta_j}. \]

Assume that (i) \(E|\psi_j(\tilde{\theta}_0, X, Y)|^2 < C\) for all \(j = 1, \cdots, m\); (ii) \(\left| \frac{\partial \psi_j}{\partial \theta_i}(\theta, x, y) \right| < F(x, y)\) with \(E|F(X, Y)|^2 < \infty\) for all \(i, j = 1, \cdots, m\) and \(\theta \in U(\tilde{\theta}_0)\), where \(U(\tilde{\theta}_0)\) is some neighborhood of \(\tilde{\theta}_0\).

(A2). Let \(P_{\psi_\theta} = E[\psi(\theta, X, Y)]\) and \(P_{\psi_j\theta} = E[\psi_j(\theta, X, Y)]\). Suppose that \(P_{\psi_\theta}\) is twice differentiable in the neighborhood \(U(\tilde{\theta}_0)\) of \(\tilde{\theta}_0\). Set

\[ \hat{P}_{\psi_\theta} = \frac{\partial P_{\psi_\theta}}{\partial \theta}, \hat{P}_{\psi_j} = \frac{\partial P_{\psi_\theta}}{\partial \theta_j}, \text{ and } \hat{P}_{\psi_j\theta} = \frac{\partial P_{\psi_j\theta}}{\partial \theta_j}. \]
Define $\Sigma = -\hat{P}\psi_{\theta_0}$. Assume that the matrix $\Sigma$ is nonsingular and satisfies the following condition

$$0 < \underline{\lambda} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < \bar{\lambda} < \infty,$$

where $\underline{\lambda}$ and $\bar{\lambda}$ are two constants that do not depend on $n$ and $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote the smallest and largest eigenvalue of $\Sigma$ respectively. Furthermore, let $\lambda_1(\hat{P}\psi_{j\theta}), \cdots, \lambda_m(\hat{P}\psi_{j\theta})$ be the eigenvalues of the matrix $\hat{P}\psi_{j\theta}$ and assume

$$\max_{1 \leq i,j \leq m} |\lambda_i(\hat{P}\psi_{j\theta})| \leq C, \quad \forall \theta \in U(\tilde{\theta}_0)$$

where $C$ is a constant that is free of $n$.

(A3). The vector $\tilde{\theta}_0$ is the unique minimizer of (2.2). Thus it is easy to see that $\tilde{\theta}_0 = \theta_0$ under the model (1.1).

The regularity conditions (A1) and (A3) are standard for the nonlinear least squares estimation, see, e.g., Jennrich (1969) and White (1981). (A2) is similar to the regularity condition on the information matrix $I_n$ of the likelihood function proposed in Fan and Peng (2004) which facilitated the theoretical derivations. Although we only need the asymptotic property of the parameters for multiple index models in this paper, it also holds for general parametric regression models.

**Theorem 2.1.** Suppose the regularity conditions (A1)-(A3) hold. If $p^2/n \to 0$, then it follows that $\hat{\theta}$ is a norm consistent estimator of $\tilde{\theta}_0$ in the sense that $\|\hat{\theta} - \tilde{\theta}_0\| = O_p(\sqrt{p/n})$ where $\| \cdot \|$ denotes the Frobenius norm.

For the asymptotically linear representation of $\hat{\theta} - \tilde{\theta}_0$, we need some extra regularity conditions. We say a random variable $X \in \mathbb{R}$ is sub-Gaussian with variance proxy $\sigma^2$ if

$$E\{\exp[u(X - EX)]\} \leq \exp(\frac{\sigma^2 u^2}{2}), \quad \forall u \in \mathbb{R}.$$

(A4). The function $\psi_j(\theta, x)$ admits second derivatives in $\theta$ and $|\frac{\partial^2 \psi_j}{\partial \theta_i \partial \theta_l}(\theta, x, y)| < F(x, y)$ with $E|F(X, Y)|^2 < \infty$ for all $\theta \in U(\tilde{\theta}_0)$ and $i, j = 1, \cdots, m$. Moreover, $v^\top \frac{\partial \psi_j}{\partial \theta}(\theta_0, X, Y)v$ is sub-Gaussian with variance proxy $\sigma^2$ for all $v \in S^{m-1}$, where $S^{m-1}$ is the unit sphere in $\mathbb{R}^m$.

The “sub-Gaussian” condition is also used in Cai et al. (2010) to obtain the optimal rates of convergence for large covariance matrix estimation. It seems too restrictive in high dimensional settings. However, in sufficient dimension reduction, it can be rather reasonable in the following sense. When $p \to \infty$ as $n \to \infty$, Hall and Li (1993) proved that the linear combinations of the covariates are approximately normally distributed. Consequently, in the high dimension setting,
the random variable $v^\top \partial \psi / \partial \theta (\hat{\theta}_0, X, Y)v$ approximately has a Gaussian tail that is equivalent to the fact that $v^\top \partial \psi / \partial \theta (\hat{\theta}_0, X, Y)v$ is approximately sub-Gaussian. See Section 2 of Wainwright (2019) for more details.

**Theorem 2.2.** Suppose the regularity conditions (A1)-(A4) hold. If $(p \log n)^3/n \to 0$, it follows that
\[
\sqrt{n} \alpha^\top (\hat{\theta} - \hat{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [Y_i - g(\hat{\theta}_0, X_i)] \alpha^\top \Sigma^{-1} g'(\hat{\theta}_0, X_i) + o_p(1),
\]
(2.3)
where the remaining term $o_p(1)$ is uniformly in $\alpha \in S^{m-1}$.

The convergence rate $O_p(\sqrt{p/n})$ in Theorem 2.1 is in line with the results of the $M$-estimator studied by Huber (1973) and Portnoy (1984) when the number of parameters is divergent. But they only considered linear regression models $g(\theta, X) = \theta^\top X$. In Theorem 2.2 we obtain the uniformly asymptotically linear representation of $\hat{\theta} - \hat{\theta}_0$. This result is much stronger than that in Portnoy (1985) who also only considered linear regression models. Furthermore, in the linear regression setting, we can derive, when $p = o(n)$ and $p = o(n^{1/2})$, the norm consistency and the uniformly asymptotically linear representation of $\hat{\theta}$ to $\theta_0$.

**Theorem 2.3.** Suppose the regularity conditions (A1)-(A3) hold for linear regression models and $v^\top X$ is sub-Gaussian with the variance proxy $\sigma^2$ for all $v \in S^{p-1}$. If $p/n \to 0$, then we have $\|\hat{\theta} - \theta_0\| = O_p(\sqrt{p/n})$ where $\|\cdot\|$ denotes the Frobenious norm. Moreover, if $p^2/n \to 0$, it follows that
\[
\sqrt{n} \alpha^\top (\hat{\theta} - \hat{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - \hat{\theta}_0^\top X_i) \alpha^\top \Sigma^{-1} X_i + o_p(1),
\]
where $\Sigma = E(XX^\top)$ and the term $o_p(1)$ is uniformly in $\alpha \in S^{p-1}$.

3 adaptive-to-model version of the ICM test

Now we restate the null hypothesis as
\[
H_0 : E(Y|X) = g(\beta_0^\top X, \vartheta_0) \quad \text{for some } \beta_0 \in \mathbb{R}^{p \times d} \text{ and } \vartheta_0 \in \mathbb{R}^l.
\]
While the alternative hypothesis is that for any $\beta \in \mathbb{R}^{p \times d}$ and $\vartheta \in \mathbb{R}^l$,
\[
H_1 : E(Y|X) = G(B^\top X) \neq g(\beta^\top X, \vartheta),
\]
where $G$ is an unknown smoothing function, $B$ is a $p \times q$ orthonormal matrix satisfying $S_{E(Y|X)} = \text{span}(B)$. Here $S_{E(Y|X)}$ is the central mean space of $Y$ with respect to $X$ which is defined as the
intersection of all subspaces \( \text{span}(A) \) such that \( Y \perp E(Y|X)|A^\top X \) where \( \perp \) means statistical independence and \( \text{span}(A) \) means the subspace spanned by the columns of \( A \). Under mild conditions, such a subspace \( S_{E(Y|X)} \) always exists (see Cook and Li (2002)). The dimension of \( S_{E(Y|X)} \) is called the structural dimension which is \( q \) under the alternatives. Similarly, under the null we have \( S_{E(Y|X)} = \text{span}(B) = \text{span}(\beta_0) \) with a structural dimension \( d \). Here we use the same notation \( B \) under both the null and alternatives which should be different in these two cases.

Recall that \( \varepsilon = Y - E(Y|X) \) and
\[
(\tilde{\beta}_0, \tilde{\vartheta}_0) = \arg\min_{\beta \in \mathbb{R}^{p \times d}, \vartheta \in \mathbb{R}^l} E[Y - g(\beta^\top X, \vartheta)]^2.
\]
Let \( \eta = Y - g(\tilde{\beta}_0^\top X, \tilde{\vartheta}_0) \) and \( \widehat{\mathcal{B}} = (\tilde{\beta}_0, B) \). Under the null hypothesis, we have \( \tilde{\beta}_0 = \beta_0 \) and \( \text{span}(B) = \text{span}(\beta_0) \). Consequently,
\[
E(\eta|\tilde{\mathcal{B}}^\top X) = E(Y|\tilde{\beta}_0^\top X, B^\top X) - g(\tilde{\beta}_0^\top X, \tilde{\vartheta}_0) = E(Y|\beta_0^\top X) - g(\beta_0^\top X, \vartheta_0) = 0.
\]
Under the alternative, we have \( \eta = G(B^\top X) - g(\tilde{\beta}_0^\top X, \tilde{\vartheta}_0) + \varepsilon \). Then we obtain that
\[
E(\eta|\tilde{\mathcal{B}}^\top X) = G(B^\top X) - g(\tilde{\beta}_0^\top X, \tilde{\vartheta}_0) \neq 0.
\]
By Theorem 1 of Bierens (1982), under the null, we have \( E[\eta \exp(it^\top \tilde{\mathcal{B}}^\top X)] = 0 \) for all \( t \in \mathbb{R}^{2d} \). Then it follows that
\[
\int_{\mathbb{R}^{2d}} |E[\eta \exp(it^\top \tilde{\mathcal{B}}^\top X)]|^2 \varphi(t) dt = 0. \tag{3.1}
\]
While under the alternative, there exist \( t_0 \in \mathbb{R}^{d+q} \) such that \( E[\eta \exp(it_0^\top \tilde{\mathcal{B}}^\top X)] \neq 0 \). Therefore,
\[
\int_{\mathbb{R}^{d+q}} |E[\eta \exp(it^\top \tilde{\mathcal{B}}^\top X)]|^2 \varphi(t) dt > 0, \tag{3.2}
\]
where \( \varphi(t) \) denotes a positive weight function which will be specified later. Thereby we reject the null hypothesis for “large value” of the empirical version of the left-hand side of (2.1). Let \( \{(X_1, Y_1), \cdots, (X_n, Y_n)\} \) be a random sample from the distribution of \( (X, Y) \). Then we propose an adaptive-to-model integrated conditional moment test statistic as
\[
\hat{V}_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n |Y_j - g(\tilde{\beta}^\top X_j, \tilde{\vartheta}_0)| \exp(it^\top \tilde{\mathcal{B}}^\top X_j), \tag{3.3}
\]
\[
AICM_n = \int_{\mathbb{R}^{d+q}} |\hat{V}_n(t)|^2 \varphi(t) dt, \tag{3.4}
\]
where \( \tilde{\mathcal{B}} = (\tilde{\beta}, \tilde{B}) \), \( \tilde{B} \) is a sufficient dimension reduction estimator of \( B \) with an estimated structural dimension \( \tilde{q} \) of \( q \), and \( \tilde{\beta} \) is a norm consistent estimators of \( \tilde{\beta}_0 \). In this paper, we restrict ourselves to the ordinary least square estimator of \( \tilde{\beta}_0 \).
It is worth mentioning that Guo et al. (2016) first used sufficient dimension reduction techniques to construct a goodness of fit test for parametric single-index models when the dimension $p$ of the covariates is fixed. But they only used the matrix $B$ rather than $(\tilde{\beta}_0, B)$ as we consider here. To make sure $E(\eta | B^\top X) \neq 0$ under the alternative hypotheses, they need an extra condition that $\tilde{\beta}_0 \in \text{span}(B)$. Note that this extra condition does not always hold. Thus we use the matrix $\tilde{B}$ in the conditional distribution to avoid this extra condition.

Another test related to ours is that of Stute and Zhu (2002), who developed a dimension reduction type test for generalized linear models in the fixed dimension cases. Their test statistic is based on the fact that $E(\eta | \beta^\top X) = 0$ under the null hypothesis, where $\beta_0$ is the fixed direction involved in the generalized linear model. Stute and Zhu’s (2002) test has been turned out to be very powerful for large $p$ in some cases. However, it is not omnibus as it only used one projection direction in the test statistic. Escanciano (2006a) gave a detail comment on this issue. While our test combines these two methods of Guo et al. (2016) and Stute and Zhu (2002) to avoid their shortcomings at the cost of dealing with one more dimension than that in Guo et al. (2016)’s test.

3.1 The choice of $\varphi(t)$

The choice of weight functions $\varphi(t)$ is flexible. Recall that our test statistic $AICM_n = \int_{\mathbb{R}^{d+q}} |\tilde{V}_n(t)|^2 \varphi(t) dt$. Suppose that $\varphi(t) = \varphi(-t)$ and then some elementary calculations lead to

$$AICM_n = \frac{1}{n} \int_{\mathbb{R}^{d+q}} |\sum_{j=1}^n \tilde{\eta}_j \exp(it^\top \tilde{B}^\top X_j)|^2 \varphi(t) dt$$

$$= \frac{1}{n} \sum_{j,k=1}^n \tilde{\eta}_j \tilde{\eta}_k \int_{\mathbb{R}^{d+q}} \exp\{it^\top (\tilde{B}^\top X_j - \tilde{B}^\top X_k)\} \varphi(t) dt$$

$$= \frac{1}{n} \sum_{j,k=1}^n \tilde{\eta}_j \tilde{\eta}_k \int_{\mathbb{R}^{d+q}} \cos\{t^\top (\tilde{B}^\top X_j - \tilde{B}^\top X_k)\} \varphi(t) dt$$

where $\tilde{\eta}_j = Y_j - g(\hat{\beta}^\top X_j, \hat{\vartheta}_j)$. Let $K_\varphi(x) = \int \cos(t^\top x) \varphi(t) dt$, then it follows that

$$AICM_n = \frac{1}{n} \sum_{j,k=1}^n \tilde{\eta}_j \tilde{\eta}_k K_\varphi(\tilde{B}^\top X_j - \tilde{B}^\top X_k).$$

Yet the calculation of $K_\varphi(x)$ is still complex in high dimension settings, even when we use sufficient dimension reduction techniques here. To facilitate the calculation of the test statistic, a close form of the function $K_\varphi(x)$ is preferred. There is a large class of weight functions $\varphi(t)$ available for this purpose. For instance, if we choose the density function $\phi(t)$ of a standard
Gaussian distribution as a weight function, then

$$K_{\phi}(x) = \int_{\mathbb{R}^{d+\hat{q}}} \cos(t^\top x)\phi(t)dt = \exp\left(-\frac{1}{2}\|x\|^2\right),$$

where $\|\cdot\|$ denotes the Frobenius norm. Consequently, the test statistic can be stated as

$$AICM_n = \frac{1}{n} \sum_{j,k=1}^{n} \hat{\eta}_j \hat{\eta}_k \exp\{-\frac{1}{2}(\|\hat{B}^\top X_j - \hat{B}^\top X_k\|^2 + \|\hat{\beta}^\top X_j - \hat{\beta}^\top X_k\|^2)\}. \quad (3.5)$$

Note that the characteristic function of a spherically symmetric distribution has a form of $\psi(\|x\|)$ (see, Chapter 2 of Fang, Kotz and Ng (1990)). Thus the density function of any spherically symmetric distribution is also suitable as a weight function. In the simulation studies, we will use the density function of a multivariate standard Gaussian distribution.

Unlike Bierens’s (1982) original test statistic $ICM_n$ in (1.3), we here utilize the sufficient dimension reduction and projection pursuit techniques to reduce the original dimension $p$ to $d + \hat{q}$. Then the exponential weight in (3.5) would not deteriorate to zero as $p$ tends to infinity. In practice, when $d + \hat{q}$ is much smaller than $p$, the dimensionality difficulty will be largely alleviated. We will show the advantages in the simulations.

### 3.2 Model adaptation property

To achieve the model adaptation, we need the sufficient dimension reduction (SDR) techniques to identify the structural dimension $q$ and the central mean subspace $S_{E(Y|X)}$. When $p$ is fixed, there are several methods in the literature to identify the central mean subspace $S_{E(Y|X)}$, such as principal Hessian directions (pHd, Li (1992)). However, when $p$ is divergent, we have no corresponding asymptotic results about the estimated structural dimension $\hat{q}$ and orthonormal matrix $\hat{B}$ for these methods. To overcome this difficulty, we consider the central subspace $S_{Y|X}$ instead of the central mean subspace $S_{E(Y|X)}$. The central subspace $S_{Y|X}$ defined as the intersection of all subspaces span($A$) such that $Y \perp \perp X|A^\top X$. It is easy to see that $S_{E(Y|X)} \subset S_{Y|X}$. Thus we further assume that $S_{E(Y|X)} = S_{Y|X}$. This can be achieved when the error terms $\varepsilon$ under the null and alternatives have dimension reduction structures: $\varepsilon = \sigma_1(\beta_0^\top X)\tilde{\varepsilon}$ and $\varepsilon = \sigma_2(B^\top X)\tilde{\varepsilon}$ with $\tilde{\varepsilon} \perp \perp X$ respectively. More details about this issue can be found in Tan and Zhu (2019a).

When $p$ is fixed, there exist many estimation approaches available in the literature to identify the central subspace $S_{Y|X}$, such as sliced inverse regression (SIR, Li (1991)), sliced average variance estimation (SAVE, Cook and Weisberg (1991)), minimum average variance estimation (MAVE, Xia et.al. (2002)), directional regression (DR, Li and Wang, (2007)), discretization-expectation estimation (DEE, Zhu, et al. (2010a)) etc. Zhu, Miao, and Peng (2006) first
discussed the asymptotic properties of SIR in the divergent dimension setting. In this paper we adapt cumulative slicing estimation (CSE, Zhu, Zhu, and Feng (2010b)) to identify the central subspace $\mathcal{S}_{Y|X}$, as it allows the dimension $p$ to diverge to infinity at a rate $p^2/n \to 0$ and is very easy to implement in practice. The basic procedure of CSE is given below.

For simplicity, we assume $E(X) = 0$ and $\text{Var}(X) = I_p$. Under the linearity condition (see Li (1991)), it can be shown that $E[Xh(Y)] \in \mathcal{S}_{Y|X}$ for any function $h(\cdot)$. Therefore, we obtain infinity amount of vectors in $\mathcal{S}_{Y|X}$. Zhu et al. (2010b) suggested the determining class of indicator functions $\mathcal{H} = \{h_t(Y) = I(Y \leq t) : t \in \mathbb{R}\}$ to substitute $h(\cdot)$. Then define a target matrix as follows

$$
M = \int E[Xh_t(Y)]E[X^T h_t(Y)]dF_Y(t),
$$

(3.6)

where $F_Y(t)$ is the cumulative distribution function of $Y$. If the rank of $M$ is $q$, Zhu et al. (2010b) showed that $\text{span}(M) = \mathcal{S}_{Y|X}$. Let $Z_i$ be the standardized $X_i$ and $\hat{\alpha}_t = \frac{1}{n} \sum_{i=1}^{n} Z_i I(Y_i \leq t)$. Then the sample version of $M$ is given by

$$
\hat{M} = \frac{1}{n} \sum_{j=1}^{n} \hat{\alpha}_{Y_j} \hat{\alpha}^T_{Y_j}.
$$

(3.7)

If the structural dimension $q$ is known, the estimator $\hat{B}(q)$ of $B$ consists of the eigenvectors corresponding to the largest $q$ eigenvalues of $\hat{M}$.

Yet we need to estimate the structural dimension $q$. Inspired by the idea of Xia, Xu and Zhu (2015), we suggest a minimum ridge-type eigenvalue ratio estimator (MRER) to identify $q$. Let $\{\hat{\lambda}_j, 1 \leq j \leq p\}$ and $\{\lambda_j, 1 \leq j \leq p\}$ be the eigenvalues of the matrix $\hat{M}$ and $M$ respectively. Suppose $\hat{\lambda}_{j+1} \leq \hat{\lambda}_i$ and $\lambda_{j+1} \leq \lambda_j$. Since $\text{rank}(M) = q$, we have

$$
\lambda_p = \cdots = \lambda_{q+1} = 0 < \lambda_q \leq \cdots \leq \lambda_1.
$$

Hence we estimate the structural dimension $q$ by

$$
\hat{q} = \arg \min_{1 \leq j \leq p-1} \left\{ j : \frac{\hat{\lambda}_j^2 + c_n}{\hat{\lambda}_j^2 + c_n} \leq \frac{\lambda_j^2 + c_n}{\lambda_j^2 + c_n} \right\}.
$$

(3.8)

Here $c_n$ is a positive constant relying on the sample size $n$. The following result is a slight extension of Tan and Zhu (2019a) that shows the consistency of MRER and model adaptation to the underlying models, if $c_n$ is selected appropriately.

**Theorem 3.1.** Suppose the regularity conditions of Theorem 3 in Zhu et al. (2010b) hold and let $\hat{B}(q)$ be a matrix whose columns are the eigenvectors associating with the largest $q$ eigenvalues of $\hat{M}$. If $0 < c < \lambda_q \leq \lambda_1 \leq C < \infty$ and $c_n = \log n/n$, then it follows that

(1) under $H_0$, we have $\mathbb{P}(\hat{q} = d) \to 1$ and $\|\hat{B}(d) - B\| = O_p(\sqrt{p/n})$,
(2) under $H_1$, we have $\mathbb{P}(\hat{q} = q) \to 1$ and $\|\hat{B}(q) - B\| = O_p\left(\sqrt{pq/n}\right)$, where $c$ and $C$ are two positive constants free of $n$ and $B$ under the null $H_0$ satisfied that $\text{span}(B) = \text{span}(\beta_0)$.

4 Asymptotic properties of the test statistic

4.1 Limiting null distribution

Recall that

$$AICM_n = \int_{\mathbb{R}^{d+q}} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \hat{\eta}_j \exp(it^T \hat{B}^T X_j) \right|^2 \varphi(t) dt,$$

where $\hat{\eta}_j = Y_j - g(\hat{\beta}^T X_j, \hat{\vartheta}_j)$. To facilitate the derivation of the asymptotical properties, we define the following empirical process

$$\hat{V}_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \hat{\eta}_j [\cos(t^T \hat{B}^T X_j) + \sin(t^T \hat{B}^T X_j)].$$

If $\varphi(t) = \varphi(-t)$, then it follows that

$$AICM_n = \int_{\mathbb{R}^{d+q}} |\hat{V}_n^1(t)|^2 \varphi(t) dt.$$

To obtain the large-sample properties of $AICM_n$, we need some regularity conditions. Also recall that $g(\beta^T X, \vartheta) = g(\beta^T_1 X, \cdots, \beta^T_d X, \vartheta)$. Put

$$g'_i(t_1, \cdots, t_d, \vartheta) = \frac{\partial g}{\partial t_i}(t_1, \cdots, t_d, \vartheta) \quad \text{for } i = 1, \cdots, d.$$

(B1) Assume that $|g'_i(\beta^T_1 x, \cdots, \beta^T_d x, \vartheta)| \leq F(x)$ for all $\beta_1, \cdots, \beta_d, \vartheta$ and $i = 1, \cdots, d$. Further, assume that

$$\lambda_{\text{max}}\{E[F(X)XX^T]\} \leq C \quad \text{and} \quad \lambda_{\text{max}}\{E[\varepsilon^2 F(X)XX^T]\} \leq C,$$

where $\lambda_{\text{max}}(\cdot)$ is the largest eigenvalue of the matrix and $C$ is a positive constant.

(B2) The weight function $\varphi(t)$ is positive and satisfies that $\varphi(t) = \varphi(-t)$, $\int_{\mathbb{R}^{d+q}} \varphi(t) dt < \infty$, $\int_{\mathbb{R}^{d+q}} \|t\|^6 \varphi(t) dt < \infty$, and $\int_{\mathbb{R}^{2d}} \|M(t)\|^2 \varphi(t) dt = O(1)$. Here $M(t)$ is a shift term which is given by $M(t) = E\{g'_i(\theta_0, X)[\cos(t^T \mathfrak{B}_0^T X) + \sin(t^T \mathfrak{B}_0^T X)]\}$ and $\mathfrak{B}_0 = (\beta_0, B)$.

(B3) Assume that $E(\varepsilon^8) < \infty$ and $E(X^8_{jk}) \leq C$ where $X_{jk}$ is the $k$-component of $X_j$. 

13
The regularity condition (B1) is similar as that in condition (A2) which is usually used in the diverging dimensional statistical inference, see Fan and Peng (2004) and Zhang and Huang (2008) for instance. Condition (B2) is necessary for the convergence of the remainder in the decomposition of the test statistic in the diverging dimension scenarios. Condition (B3) is standard in model checking literature, see, e.g., Stute (1997) and Escanciano (2006a).

Now we can obtain the asymptotic distribution of $AICM_n$ under the null hypothesis. By Theorem 3.1, $\mathbb{P}(\hat{q} = d) \to 1$ under the null hypothesis. Thus we only work on the event $\{\hat{q} = d\}$. Consequently, $\hat{B} = [\hat{\beta}_0, \hat{B}(d)]$ and $AICM_n$ becomes

$$AICM_n = \int_{\mathbb{R}^d} |\hat{V}_n^1(t)|^2 \varphi(t) dt.$$ 

Under the regularity conditions (A1)-(A4) and (B1)-(B3) and on the event $\{\hat{q} = d\}$, we can show that under the null hypothesis, $\hat{V}_n^1(t) = V_n^1(t) + R_n(t)$, (4.1)

where

$$V_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j [\cos(t^\top B_0^\top X_j) + \sin(t^\top B_0^\top X_j) - M(t^\top \Sigma^{-1} g'(\beta_0, X_j)],$$

and $R_n(t)$ is a remainder satisfying

$$\int_{\mathbb{R}^d} |R_n(t)|^2 \varphi(t) dt = o_p(1).$$

The proof of (4.1) will be given in the Appendix. Then we can obtain the following result.

**Theorem 4.1.** Suppose that the regularity conditions (A1)-(A4) and (B1)-(B3) hold. If $(p \log(n))^3/n \to 0$, under the null $H_0$, we have in distribution

$$AICM_n \to \int_{\mathbb{R}^2} |V_\infty^1(t)|^2 \varphi(t) dt,$$

(4.2)

where $V_\infty^1(t)$ is a zero-mean Gaussian process with a covariance function $K(s, t)$ which is the point-wise limit of $K_n(s, t)$. Here $K_n(s, t)$ is the covariance function of $V_n^1(t)$, that is,

$$K_n(s, t) = \text{Cov}(V_n^1(s), V_n^2(t))$$

$$= E\{\varepsilon^2 [\cos((s - t)^\top B_0^\top X) + \sin((s + t)^\top B_0^\top X)] \}$$

$$- M(t)^\top \Sigma^{-1} E\{\varepsilon^2 [\cos(s^\top B_0^\top X) + \sin(s^\top B_0^\top X)] g'(\beta_0^\top X) \}$$

$$- M(s)^\top \Sigma^{-1} E\{\varepsilon^2 [\cos(t^\top B_0^\top X) + \sin(t^\top B_0^\top X)] g'(\beta_0^\top X) \}$$

$$+ M(s)^\top \Sigma^{-1} E\{\varepsilon^2 g'(\theta_0, X) XX^\top \Sigma^{-1} M(t) \}. $$
4.2 Limiting distribution under the alternative hypotheses

Now we discuss the asymptotic property of $AICM_n$ under the alternative hypotheses. Consider the following sequence of alternative hypotheses

$$H_{1n}: Y_n = g(\beta_0^* X, \theta_0) + r_n G(B^T X) + \varepsilon,$$  \hspace{1cm} (4.3)

where $E(\varepsilon|X) = 0$, $G(B^T X)$ is a random variable satisfying $E[G(B^T X)] = 0$ and $\mathbb{P}(G(B^T X) = 0) < 1$. The convergence rate $r_n$ satisfies $r_n = 1/\sqrt{n}$ or $\sqrt{n}r_n \to \infty$. To obtain the asymptotical distribution of $AICM_n$ under the alternatives (4.3), we first derive the asymptotic properties of the estimators $\hat{q}$ and $\hat{\beta}$, when the dimension $p$ diverges to infinity.

**Theorem 4.2.** Suppose that the regularity conditions of Theorem 3 in Zhu et al. (2010b) hold. Let $\hat{B}(d)$ be the eigenvectors associated with the largest $d$ eigenvalues of $\hat{M}_n$. If $p r_n \to 0$ and $c_n = r_n^2 \log r_n^{-2}$, then under $H_{1n}$, we have $\mathbb{P}(\hat{q} = d) \to 1$ and $\|\hat{B}(d) - B_L\| = O_p(\sqrt{p/r_n})$. Here $B_L$ is a $p \times d$ orthonormal matrix satisfying $\text{span}(B_L) = \text{span}(\beta_0)$.

It is worth to mention that under the local alternatives $H_{1n}$ with $r_n = o_p(1/p)$, the estimated structural dimension $\hat{q}$ is not equal to the true structure dimension, but to $d$ asymptotically. This means $\hat{q}$ does not an consistent estimator of the true structural dimension in this case. A special case is that if $r_n = 1/\sqrt{n}$, it follows that $\mathbb{P}(\hat{q} = d) \to 1$ and $\|\hat{B}(d) - B_L\| = O_p(\sqrt{p/n})$. Yet we need to derive the asymptotic properties of $(\hat{\beta}, \hat{\theta})$ with respective to $(\beta_0, \theta_0)$. Recall that $\hat{\theta} = [\text{vec}(\hat{\beta})^\top, \hat{\theta}^\top]^\top$ and $\theta_0 = [\text{vec}(\beta_0)^\top, \theta_0^\top]^\top$ where $\text{vec}(\beta) = [\beta_1^\top, \cdots, \beta_p^\top]^\top$.

**Theorem 4.3.** Suppose the regularity conditions (A1)-(A3) and (4.3) hold. If $p r_n \to 0$, then $\hat{\theta}$ is a norm consistent estimator for $\theta_0$ with $\|\hat{\theta} - \theta_0\| = O_p(\sqrt{p/r_n})$. Moreover, if the regularity conditions (A1)-(A4) hold and $n(p \log n)^3 r_n^4 \to 0$, then under the alternatives (4.3), we have

$$\sqrt{n}\alpha^\top(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \alpha^\top \Sigma^{-1} g'(\theta_0, X_j) + \sqrt{n}r_n \alpha^\top \Sigma^{-1} E[g'(\theta_0, X_j)G(B^T X_j)] + o_p(1),$$  \hspace{1cm} (4.4)

where the term $o_p(1)$ is uniformly in $\alpha \in S^{p-1}$.

Under the alternatives $H_{1n}$ with $r_n = 1/\sqrt{n}$, if $p^2/n \to 0$, then we have $\hat{\theta}$ is a norm consistent estimator for $\theta_0$ with $\|\hat{\theta} - \theta_0\| = O_p(\sqrt{p/n})$. Moreover, if $(p \log n)^3/n \to 0$, it follows that

$$\sqrt{n}\alpha^\top(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \alpha^\top \Sigma^{-1} g'(\theta_0, X_j) + \alpha^\top \Sigma^{-1} E[g'(\theta_0, X)G(B^T X)] + o_p(1).$$

**Theorem 4.4.** Suppose the regularity conditions (A1)-(A4) and (B1)-(B3) hold.

(1) If $(p \log n)^3/n \to 0$, under the global alternative $H_1$, we have in probability

$$\frac{1}{n} \int_{\mathbb{R}^{d+q}} |\tilde{V}_n(t)|^2 \varphi(t) dt \to C_1,$$
where $C_1$ is a positive constant. This means $AICM_n$ diverges to infinity at the rate of $n$;

(2) If $n(p \log n)^3 r_n^4 \to 0$, under the local alternatives $H_{1n}$ with $\sqrt{n} r_n \to \infty$, we have in probability

$$\frac{1}{nr_n^2} \int_{R^{d+q}} |\hat{V}_n(t)|^2 \varphi(t) dt \to C_2,$$

where $C_2$ is a positive constant. This means $AICM_n$ diverges to infinity at the rate of $nr_n^2$;

(3) If $(p \log n)^3/n \to 0$, under the local alternatives $H_{1n}$ with $r_n = 1/\sqrt{n}$, we have in distribution

$$\int_{R^{d+q}} |\hat{V}_n(t)|^2 \varphi(t) dt \to \int_{R^{2d}} |V_\infty^1(t) + L_1(t) - L_2(t)|^2 \varphi(t) dt,$$

where $V_\infty^1$ is a zero-mean Gaussian process given by (4.2) and $L_1(t)$ and $L_2(t)$ are the uniformly limits of $L_{n1}(t)$ and $L_{n2}(t)$, respectively. Two functions $L_{n1}(t)$ and $L_{n2}(t)$ are given by

$$L_{n1}(t) = M^\top(t) \Sigma^{-1} E[g'(\theta_0, X) G(B^\top X)]$$

$$L_{n2}(t) = E\{G(B^\top X)|\cos(t^\top B_{0}^\top X) + \sin(t^\top B_{0}^\top X)\}.$$  

This means $AICM_n$ is still sensitive to the local alternatives distinct from the null at the rate of $1/\sqrt{n}$.

5 Bootstrap approximation

Note that limiting null distribution of our test statistic $AICM_n$ depends on the parameters $\beta_0$ and the matrix $B$, and thus is not tractable for the critical value determination. A typical method used in the literature is the wild bootstrap proposed by Wu (1986). Stute, González Manteiga and Presedo Quindimil (1998) proved that the wild bootstrap yields a valid approximation of residual marked empirical processes based on the indictor functions, when the dimension of covariates is fixed. We now show that the wild bootstrap still works in the diverging dimension setting. More specially, set

$$X_j^* = X_j \quad Y_j^* = g(\hat{\theta}, X_j) + \hat{\eta}_j V_j^*, \quad j = 1, \ldots, n,$$

where $\hat{\eta}_j$ is the residual and $\{V_j^*\}_{j=1}^n$ are i.i.d. bounded random variables with zero mean and unit variance, independent of the original sample $\{(X_i, Y_i)\}_{i=1}^n$. An often used example of $\{V_j^*\}_{j=1}^n$ is the i.i.d. Bernoulli variates with

$$P(V_j^* = 1 - \frac{\sqrt{5}}{2}) = \frac{1 + 2\sqrt{5}}{2\sqrt{5}}, \quad P(V_j^* = 1 + \frac{\sqrt{5}}{2}) = 1 - \frac{1 + 2\sqrt{5}}{2\sqrt{5}}.$$  

For other examples of $\{V_j^*\}_{j=1}^n$, one can refer to Mammen (1993).
Let $\hat{\theta}^*$ be the bootstrap estimator obtained by the ordinary least square based on the bootstrap sample $\{(X_j, Y_j^*)\}_{j=1}^n$. Then we approximate the limiting null distribution of $AICM_n$ by that of

$$AICM^*_n = \int_{\mathbb{R}^{d+q}} |\tilde{V}^*_n(t)|^2 \varphi(t) dt,$$

where

$$\tilde{V}^*_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [Y_j^* - g(\hat{\theta}^*, X_j)] \exp(it^\top \mathbf{B}^\top X_j).$$

To determine the critical value in practice, repeat the above process a large number times, say $B$ times. For a given a nominal level $\tau$, the critical value is determined by the upper $\tau$ quantile of the bootstrap distribution $\{\tilde{V}^*_nk : k = 1, \ldots, B\}$.

In the next theorem we establish the validity of the wild bootstrap in the diverging dimension setting.

**Theorem 5.1.** Suppose that the bootstrap sample is generated from the wild bootstrap and the regularity conditions (A1)-(A4) holds.

(i) If $(p \log n)^3/n \to 0$, under the null hypothesis $H_0$ or under the alternative hypothesis with $r_n = 1/\sqrt{n}$, we have with probability 1,

$$AICM^*_n \longrightarrow \int_{\mathbb{R}^{2d}} |V^1_\infty(t)|^2 \varphi(t) dt \quad \text{in distribution},$$

where $V^1_\infty$ have the same distribution as the Gaussian process $V^1_\infty$ given in Theorem 4.1.

(ii) If $n(p \log n)^2r_n^4 \to 0$, under the local alternatives $H_{1n}$ with $\sqrt{n}r_n \to \infty$, the result in (i) continues to hold.

(iii) If $(p \log n)^3/n \to 0$, under the global alternative $H_1$, the distribution of $AICM^*_n$ converge to a finite limiting distribution which may different from the limiting null distribution.

### 6 Numerical studies

#### 6.1 Simulations

In this subsection we conduct some numerical studies to show the performance of the proposed test in finite samples when the dimension of covariates is relatively large. From the theoretical view in this paper, we set $p = \lfloor 3n^{1/3} \rfloor - 5$ with $n = 100, 200, 400$ and 600. We also compare our test with some existing competitors proposed in the literature, although most of them dealt with fixed dimension.
Based on the standard normal density function, the Bierens’ (1982) test statistic becomes

\[ ICM_n = \int_{\mathbb{R}^p} \left| \frac{1}{n} \sum_{j=1}^{n} \hat{\eta}_j \exp(it^\top X_j) \right|^2 \phi(t) dt = \frac{1}{n} \sum_{j,k=1}^{n} \hat{\eta}_j \hat{\eta}_k \exp\left(-\frac{1}{2}||X_j - X_k||^2\right), \]

where \( \hat{\eta}_j \) is the residual and \( \phi(t) \) is the standard normal density on \( \mathbb{R}^p \). Escanciano (2006a) also use this test statistic for comparison.

Zheng (1996) proposed a local smoothing test for parametric regression models as

\[ T_{ZH}^n = \frac{1}{n} \sum_{i \neq j} K\left(\frac{(X_i - X_j)/h}{\hat{\eta}_i \hat{\eta}_j} \right), \]

where \( K(u) = (15/16)(1 - u^2)^2 I(|u| \leq 1) \) as the kernel function and the bandwidth \( h = 1.5n^{-1/(4+p)} \).

Escanciano (2006a) developed a global smoothing test for parametric regression models based on a projected residual marked empirical process. The test statistic is defined as follows,

\[ PCvM_n = \frac{1}{n} \sum_{i,j,r} \hat{\eta}_i \hat{\eta}_j \int_{S^{p-1}} I(\alpha^\top X_i \leq \alpha^\top X_r)I(\alpha^\top X_j \leq \alpha^\top X_r) d\alpha, \]

where \( S^{p-1} = \{ \alpha \in \mathbb{R}^p : ||\alpha|| = 1 \} \) and \( d\alpha \) denotes the uniform density on the unit sphere \( S^{p-1} \). The critical value of Escanciano’s (2006a) test is determined by the wild bootstrap.

Stute and Zhu (2002) proposed a dimension reduction-based test for generalized linear models based on residual marked empirical processes. A martingale transformation leads the test to be asymptotically distribution-free. Their test statistic is given by

\[ T_{SZ}^n = \frac{1}{\psi_n(x_0)} \int_{-\infty}^{x_0} |\hat{T}_n R_n^1|^2 \sigma_n^2 dF_n, \]

where

\[ R_n^1(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\eta}_i I(\hat{\beta}^\top X_i \leq u); \]

\[ \hat{T}_n R_n^1(u) = R_n^1(u) - \int_{-\infty}^{u} \hat{a}_n(z)^\top \hat{A}_n^{-1}(z) \left( \int_{z}^{\infty} \hat{a}_n(v) R_n^1(dv) \right) \sigma_n^2(z) F_n(dz). \]

More details of the definitions of \( \hat{T}_n R_n^1 \) can be found in their paper. Under the right specification of the generalized linear model and some regularity conditions,

\[ T_{SZ}^n \to \int_0^1 B^2(u) du \quad \text{in distribution}, \]

where \( B(\cdot) \) is the standard Brownian motion.
Guo, Wang and Zhu (2016) introduced a model-adaptive local smoothing test for parametric single index models that largely alleviate the dimensionality problem, although they also considered in the fixed dimension cases. Their test statistic is given by
\[
T_{GWZ}^n = \frac{h^{1/2} \sum_{i \neq j} \hat{\eta}_i \hat{\eta}_j K(\hat{B}^T(X_i - X_j)/h)}{(2 \sum_{i \neq j} \hat{\eta}_i^2 \hat{\eta}_j^2 K^2(\hat{B}^T(X_i - X_j)/h))^{1/2}},
\]
where the kernel function \( K(u) = (15/16)(1 - u^2)^2 I(|u| \leq 1) \) and the bandwidth \( h = 1.5n^{1/(4+\hat{q})} \) as suggested in Guo, Wang and Zhu (2016) and \( \hat{B} \) is a sufficient dimension estimator of \( B \) with an estimated structural dimension \( \hat{q} \) of \( q \).

Recently, Tan and Zhu (2019a) proposed a projected adaptive-to-model test for parametric single index models where they also allow the dimension \( p \) to diverge with the sample size \( n \). Their test statistic is given by
\[
ACM^2_n = \frac{1}{\hat{\psi}_n(u_0)^2} \int_{-\infty}^{u_0} \sup_{\hat{\alpha} \in S_+^2} |\hat{T}_n V_n(\hat{\alpha}, u)|^2 \hat{\sigma}_n^2(u) F_n(du),
\]
where
\[
\hat{T}_n V_n(\hat{\alpha}, u) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \hat{\eta}_i I(\hat{\alpha}^T \hat{B}^T X_i \leq u) - \frac{1}{n^{3/2}} \sum_{i,j=1}^{n} I(\hat{\alpha}^T \hat{B}^T X_i \leq u) \hat{a}_n(\hat{\alpha}^T \hat{B}^T X_i)^T \hat{A}_n^{-1}(\hat{\alpha}^T \hat{B}^T X_i) \hat{a}_n(\hat{\alpha}^T \hat{B}^T X_j) I(\hat{\alpha}^T \hat{B}^T X_j \geq \hat{\alpha}^T \hat{B}^T X_i) \hat{\eta}_j \hat{\sigma}_n^2(\hat{\alpha}^T \hat{B}^T X_i).
\]
For the quantities \( \hat{\psi}(u_0), \hat{\sigma}_n^2, \hat{a}_n \) and \( \hat{A}_n \), one can refer their paper for details. Under the null hypothesis and some regularity conditions, we have
\[
ACM^2_n \rightarrow \int_0^1 B^2(u) du \quad \text{in distribution},
\]
Thus this test is asymptotically distribution-free and its critical values can be tabulated.

In the simulations that follows, \( a = 0 \) corresponds to the null while \( a \neq 0 \) to the alternatives. The significance level is \( \alpha = 0.05 \). The simulation results are based on the average of 1000 replications and the bootstrap approximation of \( B = 500 \) replications.

**Study 1.** Consider the following regression models
\[
H_{11}: Y = \beta_0^T X + a \exp(-\beta_0^T X)^2 + \varepsilon;
H_{12}: Y = \beta_0^T X + a \cos(0.6\pi \beta_0^T X) + \varepsilon;
H_{13}: Y = \beta_1^T X + a (\beta_2^T X)^2 + \varepsilon;
H_{14}: Y = \beta_1^T X + a \exp(\beta_2^T X) + \varepsilon;
\]
where $\beta_0 = (1, \cdots, 1)^T / \sqrt{p}, \beta_1 = (1, \ldots, 1, 0, \ldots, 0)^T / \sqrt{p_1}$ and $\beta_2 = (0, \ldots, 0, 1, \ldots, 1)^T / \sqrt{p_1}$ with $p_1 = \lceil p/2 \rceil$. The covariate $X$ is $N(0, I_p)$ independent of the standard Gaussian error term $\varepsilon$.

Note that $H_{12}$ is a high-frequency model and the others are low-frequency models. The structure dimension $q = 1$ under both the null and alternative hypotheses in the models $H_{11}$ and $H_{12}$, while the structure dimension $q = 2$ under the alternative hypothesis in models $H_{13}$ and $H_{14}$.

The empirical sizes and powers are presented in Tables 1-4. First we can see that the Bierens’ (1982) original test $ICM_n$ performs the worst among these tests. It almost has no empirical powers at all and its empirical sizes are far away from the nominal sizes in the large dimension cases. In our experience, when the dimension $p$ of $X$ is smaller than 5, $ICM_n$ can maintain the significance level very well and the empirical power grows very quickly. But for higher dimension, the Bierens’s (1982) test does not work well. For the other tests, we observe that $AICM_n$, $ACM_n$, $T_n^{SZ}$, and $T_n^{GWZ}$ can control the empirical sizes very well. The empirical sizes of $PCvM_n$ are also close to the significance level, but slightly unstable in some cases. While $T_n^{SZ}$ can not maintain the significance level in most cases and is generally conservative with smaller empirical sizes. For the empirical power, we can see that $AICM_n$, $ACM_n$, $T_n^{SZ}$, and $PCvM_n$ all perform very well for low frequency models $H_{11}$, $H_{13}$ and $H_{14}$, whereas $T_n^{GWZ}$ behaves slightly worse for the three low frequency models. For the high frequency model $H_{12}$, the test $T_n^{GWZ}$ beats all other competitors except for the new test $AICM_n$. This is somewhat surprising as local smoothing tests such as $T_n^{GWZ}$ usually performs better for high frequency models and global smoothing tests work better for low frequency models. While the new test which can be viewed as a global smoothing test seems also to work well for high frequency models. In contrast, Zheng’s (1996) test $T_n^{ZH}$ which is a typical local smoothing test has no empirical powers in all cases. This validates the well-known results that the traditional local smoothing tests suffer severely from the curse of dimensionality.

The hypothetical models in study 1 are all single-index models. Next we consider multiple-index models in the second simulation study. As the tests $ACM_n$, $T_n^{SZ}$ and $T_n^{GWZ}$ only dealt with parametric single index models, we only compare our new test with $PCvM_n$, $ICM_n$ and $T_n^{ZH}$.

**Study 2.** Generate data from the following models:

$$H_{21} : Y = \beta_1^T X + \exp(\beta_2^T X) + a(\beta_2^T X)^2 + \varepsilon;$$

$$H_{22} : Y = \beta_1^T X + \exp(\beta_2^T X) + a \cos(0.6 \pi \beta_2^T X) + \varepsilon;$$

$$H_{23} : Y = \beta_1^T X + \exp(\beta_2^T X) + a(\beta_1^T X)(\beta_2^T X) + \varepsilon;$$

Tables 1 – 4 about here
where $\beta_1, \beta_2, \varepsilon$ and $X$ are the same as in Study 1.

The simulation results are presented in Tables 5-7. We can observe that our test $AICM_n$ and $PCvM_n$ perform much better than the other two. While Bierens’ (1982) test $ICM_n$ again does not work at all and Zheng’s (1996) test $T_{n}^{ZH}$ can not maintain the nominal level and has no empirical powers in both cases. For the empirical size, both our test $AICM_n$ and $PCvM_n$ are slightly conservative with larger empirical sizes. This may be due to the inaccurate estimation of the related parameters when $p$ is large. The empirical powers of $AICM_n$ and $PCvM_n$ both grow fast under both the low frequency model $H_{21}$ and the high frequency model $H_{22}$. In model $H_{23}$, our test has much better power performance than $PCvM_n$.

In summary, the simulation results show that the proposed test performs well and can detect different alternative hypotheses for large $p$ paradigms. In low frequency alternatives, the new test has the best power performance in high-dimensional cases among the global smoothing tests proposed by Bierens’ (1982), Stute and Zhu (2002), Escanciano (2006a), and Tan and Zhu (2019a). While surprisingly, the new test, which can be viewed as a global smoothing test, also performs best or is comparable to the best one among the local smoothing tests proposed by Zheng (1996) and Guo, Wang and Zhu (2016).

6.2 A real data example

In this subsection we apply the test to the Boston house-price data set that is first analysed by Harrison and Rubinfeld (1978). This data set can be obtained through the website [http://lib.stat.cmu.edu/datasets/boston](http://lib.stat.cmu.edu/datasets/boston). It contains 506 cases on one response variable: Median value of owner-occupied homes in 1000’s (MEDV) $Y$ and 13 predictors: per capita crime rate by town (CRIM) $X_1$, proportion of residential land zoned for lots over 25,000 sq.ft. (ZN) $X_2$, proportion of non-retail business acres per town (INDUS) $X_3$, Charles River dummy variable (CHAS, 1 if tract bounds river; 0 otherwise) $X_4$, nitric oxides concentration (NOX parts per 10 million) $X_5$, average number of rooms per dwelling (RM) $X_6$, proportion of owner-occupied units built prior to 1940 (AGE) $X_7$, weighted distances to five Boston employment centres (DIS) $X_8$, index of accessibility to radial highways (RAD) $X_9$, full-value property-tax rate per $ 10,000 (TAX) X_{10}$, pupil-teacher ratio by town (PTRATIO) $X_{11}$, $1000(B_k - 0.63)^2$ where $B_k$ is the proportion of blacks by town (B) $X_{12}$, % lower status of the population (LSTAT) $X_{13}$. For easy explanation, we standardize all variables separately. Since the dimension of the predictor $13 \approx 506^{0.4119}$, it seems reasonable to apply our method. To establish the relationship between
the response \( Y \) and the covariates \( X = (X_1, \cdots, X_{13})^\top \), we first apply the dimension reduction techniques to the data set. When the cumulative slicing estimation (CSE) is used, we find that the estimated structural dimension of this data set is \( \hat{q} = 1 \), which indicates that \( Y \) may be conditional independent of \( X \) given the projected covariates \( \hat{\beta}_1^\top X \) where

\[
\hat{\beta}_1 = (-0.0804, 0.0611, 0.0211, 0.0676, -0.3120, 0.1513, -0.1820, -0.3701, 0.3999, -0.3404, -0.2876, 0.1233, -0.5663)^\top.
\]

The scatter plot of \( Y \) against \( \hat{\beta}_1^\top X \) is presented in Figure 2.

Figure 2 about here

From this Figure, a linear regression model seem to be plausible to fit the data set. We then apply our test to see whether these exists a model misspecification. The value of the test statistic \( AICM_n \) is about 0.9639 and the \( p \)-value is about 0.01. Thus a linear regression model is not adequate to predict the response. Figure 3 presents the scatter plot of residuals from the linear regression model for \((Y, X)\) against \( \hat{\beta}_1^\top X \). It also suggests that a linear relationship between \( Y \) and \( X \) may be not reasonable.

Figure 3 about here

To find the relationship between \( Y \) and \( X \), we give a more thorough search of the projected covariates. Consider the second projected covariates \( \hat{\beta}_2^\top X \) and the scatter plot of \( Y \) against \((\hat{\beta}_1^\top X, \hat{\beta}_2^\top X)\) is presented in Figure 4.

Figure 4 about here

From this figure, we can see that the second projected direction \( \hat{\beta}_2^\top X \) is not necessary as the plot along this direction is nearly identical. This means the projection of the data onto the space \( \hat{\beta}_1^\top X \) already contain almost all the regression information of the model structure. Further more, from Figure 3, it seems that there exists an exponential relationship between \( Y \) and \( \hat{\beta}_1^\top X \). Thus we consider \( \log(Y) \) as the response variable that is also considered in Harrison and Rubinfeld (1978). Thus we use the following model to fit this data set:

\[
\log(Y) = \beta^\top X + \varepsilon. \tag{6.1}
\]

When applying our test for the above model \((6.1)\), we obtain that the value of the test statistic \( AICM_n \) is about 0.1536 and the \( p \)-value is about 0.368. A scatter plot of residuals from model \((6.1)\) against the fitted values also does not reveal any trend given in Figure 5. Thus this model is plausible.

Figure 5 about here
7 Discussions

In this article we discuss three issues in statistical inference. First we extend Bierens’ (1982) classic ICM test, via a model adaptation strategy, to the diverging dimension scenarios for parametric multiple index models. In fixed dimension scenarios, it is generally believed that the ICM test is less sensitive to the dimension $p$ of covariates due to its projection-based nature. However, we note that even when $p$ is moderate, Bierens’ original ICM test does not work at all. Thus, to attack the dimensionality problem, we construct an adaptive-to-model version of ICM test in terms of sufficient dimension reduction and projection techniques. The numerical studies suggest that the proposed test largely alleviates the adverse impact of the dimensionality in the sense that it can well maintain the significance level and enhance power performance. We also show that a large class of weight functions $\varphi(t)$ give the corresponding test statistics with the nice property of computational simplicity. Thus the test can be easily implemented in practice. Second, we obtain the uniformly asymptotically linear presentation of the least squares estimation of the parameters in the nonlinear regression model at a divergent rate $p = o(n^{1/3}/\log n)$. This result can be useful for further studies in inference. Third, the study on the consistency of the wild bootstrap approximation in this diverging setting can also be used in other hypothesis testing problems. Note that the test we proposed here is for the hypothetical models with dimension reduction structures. It is of great interest to investigate goodness of fit testing for hypothetical models without such a structure in the divergent dimension setting. The relevant studies are ongoing.

8 Appendix

Proof of Theorem 2.1. Let $\theta = \hat{\theta}_0 + \zeta$ and $F(\zeta) = \sum_{i=1}^{n} [Y_i - g(\hat{\theta}_0 + \zeta, X_i)]g'(\hat{\theta}_0 + \zeta, X_i)$. Then it suffices to show that there is a root $\hat{\zeta}$ of $F(\zeta)$ such that $\|\hat{\zeta}\|^2 = O_p(m/n)$. According to the results in (6.3.4) of Ortega and Rheinboldt (1970), it in turn needs to prove that $\zeta^T F(\zeta) < 0$ for $\|\zeta\|^2 = C m/n$ where $C$ is some large enough constant.

Recall $\psi_\theta(x, y) = [y - g(\theta, x)]g'(\theta, x)$. Then $F(\zeta) = \sum_{i=1}^{n} \psi_{\hat{\theta}_0 + \zeta}(X_i, Y_i)$. Set

$$P_n \psi_\theta = \frac{1}{n} \sum_{i=1}^{n} \psi_\theta(X_i, Y_i) \quad \text{and} \quad P \psi_\theta = E[\psi_\theta(X, Y)].$$

Since $P \psi_{\hat{\theta}_0} = 0$, it follows that

$$\zeta^T F(\zeta) = n \zeta^T P_n \psi_{\hat{\theta}_0 + \zeta}$$
Then it follows that
\[
\frac{\zeta}{\zeta}^\top [P_n(\psi_{\hat{\theta}_0} + \zeta) - P(\psi_{\hat{\theta}_0} + \zeta) - P(\psi_{\tilde{\theta}_0} + \zeta)] + \frac{\zeta}{\zeta}^\top P_n\psi_{\tilde{\theta}_0} + n\zeta^\top P\psi_{\hat{\theta}_0 + \zeta}
\]

For the first term in \(\zeta^\top F(\zeta)\), we have
\[
E\{\zeta^\top [P_n(\psi_{\hat{\theta}_0} + \zeta) - P(\psi_{\hat{\theta}_0} + \zeta) - P(\psi_{\tilde{\theta}_0} + \zeta)]\}^2
\]
\[
= \frac{1}{n} Var\{\zeta^\top [\psi(\hat{\theta}_0 + \zeta, X, Y) - \psi(\hat{\theta}_0, X, Y)]\}
\]
\[
\leq \frac{1}{n} E|\zeta^\top [\psi(\hat{\theta}_0 + \zeta, X, Y) - \psi(\hat{\theta}_0, X, Y)]|^2
\]
\[
= \frac{1}{n} E|\zeta^\top \partial \psi(\tilde{\theta}_1, X, Y)|\zeta|^2 \leq C\frac{m^2}{n}\|\zeta\|^4,
\]
where \(\tilde{\theta}_1\) lies between \(\tilde{\theta}_0 + \zeta\) and \(\tilde{\theta}_0\). Therefore,
\[
n\zeta^\top [P_n(\psi_{\hat{\theta}_0} + \zeta) - P(\psi_{\hat{\theta}_0} + \zeta) - P(\psi_{\tilde{\theta}_0} + \zeta)] = \sqrt{nm^2}\|\zeta\|^2 O_p(1).
\]

Note that \(P\psi_{\tilde{\theta}_0} = 0\). Then it follows that
\[
E\|P_n\psi_{\tilde{\theta}_0}\|^2 = E\|P_n\psi_{\tilde{\theta}_0}\|^2 = \frac{1}{n} \sum_{j=1}^m E\|P_{\psi_{\tilde{\theta}_0}}\|^2 \leq \frac{m}{n} C
\]

Thus we obtain that \(n\zeta^\top P_n\psi_{\tilde{\theta}_0} = \sqrt{nm}\|\zeta\| O_p(1)\).

Next we consider the term \(n\zeta^\top P\psi_{\hat{\theta}_0 + \zeta}\). Applying Taylor expansion,
\[
n\zeta^\top P\psi_{\hat{\theta}_0 + \zeta} = n\zeta^\top \tilde{\psi}_{\hat{\theta}_0} \zeta + n\zeta^\top \zeta + n\zeta^\top \chi(\tilde{\psi}_\theta |_{\theta = \tilde{\theta}_2})\zeta
\]
\[
= -n\zeta^\top \Sigma \zeta + n\zeta^\top \chi(\tilde{\psi}_\theta |_{\theta = \tilde{\theta}_2})\zeta,
\]
where \(\tilde{\theta}_2\) lies between \(\tilde{\theta}_0 + \zeta\) and \(\tilde{\theta}_0\). By Assumption 2, we obtain
\[
\|\zeta^\top (\tilde{\psi}_\theta |_{\theta = \tilde{\theta}_2})\zeta\|^2 \leq \sum_{j=1}^m \max_{1 \leq i \leq m} |\lambda_i(\tilde{\psi}_\theta |_{\theta = \tilde{\theta}_2})|^2 \|\zeta\|^4 \leq Cm\|\zeta\|^4.
\]

Then it follows that
\[
n\zeta^\top P\psi_{\hat{\theta}_0 + \zeta} = -n\zeta^\top \Sigma \zeta + n\sqrt{m}\|\zeta\|^3 O_p(1).
\]

Let \(\zeta = \sqrt{m/n} U\). Altogether we obtain
\[
\zeta^\top F(\zeta) = \frac{\sqrt{mn^2}\|\zeta\|^2 O_p(1) + \sqrt{mn}\|\zeta\| O_p(1) - n\zeta^\top \Sigma \zeta + \sqrt{mn}\|\zeta\|^3 O_p(1)}{\sqrt{n}}
\]
\[
= \frac{m^2}{\sqrt{n}}\|U\|^2 O_p(1) + m\|U\| O_p(1) - mU^\top \Sigma U + \frac{m^2}{\sqrt{n}}\|U\|^3 O_p(1)
\]
\[
\leq m\|U\| O_p(1) - m\lambda_{\min}(\Sigma)\|U\|^2 + \frac{m^2}{\sqrt{n}}\|U\|^2 O_p(1) + \frac{m^2}{\sqrt{n}}\|U\|^3 O_p(1)
\]
\[ = m\|U\|\{O_p(1) - \lambda_{\min}(\Sigma)\|U\| + \frac{m}{\sqrt{n}}\|U\|O_p(1) + \frac{m}{\sqrt{n}}\|U\|^2O_p(1)\} \]

If \(m^2/n \to 0\) and \(\|U\| = C\) is large enough, for any \(\epsilon > 0\), we have

\[ \mathbb{P}(\zeta^T F(\zeta) < 0) \geq \mathbb{P}\{O_p(1) - \lambda_{\min}(\Sigma)\|U\| + \|U\|o_p(1) < 0\} \geq 1 - \epsilon. \]

Thus our result follows from (6.3.4) of Ortega and Rheinboldt (1970). Hence we complete the proof. \(\Box\)

**Proof of Theorem 2.2.** We use the same notations as those in the proof for Theorem 2.1. Let \(G_n\psi_\theta = \sqrt{n}(P_n\psi_\theta - P\psi_\theta)\) and \(H = \{\theta: \theta \in \mathbb{R}^m, \|\theta - \tilde{\theta}_0\| \leq C \sqrt{(m \log n)/n}\}\). First we show that \(G_n\alpha^T(\psi_\theta - \psi_{\tilde{\theta}_0}) = o_p(1)\) uniformly in \(\alpha \in S^{m-1}\). Fix \(\epsilon > 0\). Then we have

\[ \mathbb{P}(\sup_{\alpha \in S^{m-1}} |G_n\alpha^T(\psi_\theta - \psi_{\tilde{\theta}_0})| > 8\epsilon) \leq \mathbb{P}(\sup_{\alpha \in S^{m-1}} |G_n\alpha^T(\psi_\theta - \psi_{\tilde{\theta}_0})| > 8\epsilon, \tilde{\theta} \in H) + \mathbb{P}(\tilde{\theta} \notin H) \]

Since \(\tilde{\theta} - \tilde{\theta}_0 = O_p(\sqrt{m/n})\), it suffices to show that

\[ \mathbb{P}(\sup_{\alpha \in S^{m-1}, \tilde{\theta} \in H} |G_n\alpha^T(\psi_\theta - \psi_{\tilde{\theta}_0})| > 8\epsilon) \to 0. \]

Set \(\mathcal{F} = \{\alpha^T[\psi_\theta(x, y) - \psi_{\tilde{\theta}_0}(x, y)] : \alpha \in S^{m-1}, \tilde{\theta} \in H\}\). By Assumption (A4), we obtain that

\[
\begin{align*}
|\alpha^T[\psi_\theta(x, y) - \psi_{\tilde{\theta}_0}(x, y)]| & = |\alpha^T[\psi(\theta, x, y) - \psi(\tilde{\theta}_0, x, y)]| \\leq & |\alpha^T\frac{\partial \psi}{\partial \theta}(\tilde{\theta}_0, x, y)(\theta - \tilde{\theta}_0) + \frac{1}{2}(\theta - \tilde{\theta}_0)^T\{\alpha^T\frac{\partial \psi}{\partial \theta} + \frac{\partial \psi}{\partial \theta}\}(\tilde{\theta}_1, x, y)(\theta - \tilde{\theta}_0)\}| \\
& \leq |\theta - \tilde{\theta}_0||\|\frac{\partial \psi}{\partial \theta}(\tilde{\theta}_0, x, y)\||_2 + m^{3/2}\|\theta - \tilde{\theta}_0\|^2F_2(x, y) \\
& \leq |\theta - \tilde{\theta}_0||\|\frac{\partial \psi}{\partial \theta}(\tilde{\theta}_0, x, y)\||_2 + m^2\sqrt{(\log n)/n}F_2(x, y) 
\end{align*}
\]

where \(\tilde{\theta}_1\) lies between \(\theta\) and \(\tilde{\theta}_0\) and \(\|\cdot\|_2\) denotes the operator norm of a matrix in the sense that

\[ \|A\|_2 = \sup_{\|v\|=1} \|Av\|. \]

Let \(F(x, y) = \|\frac{\partial \psi}{\partial \theta}(\tilde{\theta}_0, x, y)\||_2 + m^2\sqrt{(\log n)/n}F_2(x, y)\). Next we show that \(E|F(X, Y)|^2 < Cm\).

Since \((m \log n)^3/n \to 0\), it follows that \(E[m^2\sqrt{(\log n)/n}F_2(x, y)]^2 \leq Cm\). It remains to show that \(E[\|\frac{\partial \psi}{\partial \theta}(\tilde{\theta}_0, x, y)\||_2]^2 \leq Cm\). According to the arguments of Lemma 3 of Cai et al. (2010), there exists vectors \(\{v_j \in S^{m-1}, 1 \leq j \leq 7^m\}\) such that

\[
\|A\|_2 \leq 4 \sup_{1 \leq j \leq 7^m} |v_j^TAv_j| < 7^m.
\]

25
for any symmetric matrix $A$. Set $Z_j = v_j^\top \frac{\partial \psi}{\partial \theta}(\tilde{\theta}_0, X, Y)v_j$. Then it follows that

$$E \| \frac{\partial \psi}{\partial \theta}(\tilde{\theta}_0, X, Y)\|_2^2 \leq 4E[ \sup_{1 \leq j \leq 7^m} |Z_j|^2] \leq 8E[ \sup_{1 \leq j \leq 7^m} |Z_j - EZ_j|^2] + 8E[ \sup_{1 \leq j \leq 7^m} |EZ_j|^2]$$

Note that $EZ_j = -v_j^\top \Sigma v_j$. Thus

$$8E[ \sup_{1 \leq j \leq 7^m} |EZ_j|^2] \leq C.$$

For the term $E[\sup_{1 \leq j \leq 7^m} |Z_j - EZ_j|^2]$, we have

$$E[\sup_{1 \leq j \leq 7^m} |Z_j - EZ_j|^2] = \frac{1}{w} E \log \{ \exp( \sup_{1 \leq j \leq 7^m} w |Z_j - EZ_j|^2) \}$$

$$\leq \frac{1}{w} \log \{ \sum_{j=1}^{7^m} E[\exp(w |Z_j - EZ_j|^2)] \}.$$

where $w$ is any positive number. Since $Z_j - EZ_j$ is sub-Gaussian with proxy variance $\sigma^2$, by Lemma 1.3 of Rigollet and Hütter (2017), we obtain that

$$E[\exp(w |Z_j - EZ_j|^2)] = \int_0^\infty P\{\exp(w |Z_j - EZ_j|^2) > t\} dt$$

$$= 1 + \int_1^\infty P\{|Z_j - EZ_j| > \sqrt{\log \frac{t}{w}}\} dt$$

$$\leq 1 + 2 \int_1^\infty \exp\left(-\frac{\log t}{2w\sigma^2}\right) dt$$

$$= 1 + 2 \int_1^\infty t^{-\frac{1}{2w\sigma^2}} dt.$$

For any $w$ with $2w\sigma^2 < 1$, we have $\int_1^\infty t^{-\frac{1}{2w\sigma^2}} dt < \infty$. Then it follows that

$$E[\sup_{1 \leq j \leq 7^m} |Z_j - EZ_j|^2] \leq Cm.$$

Thus we obtain $E[\|\frac{\partial \psi}{\partial \theta}(\tilde{\theta}_0, x, y)\|_2^2] \leq Cm$ and then $E[F(X, Y)]^2 < Cm$.

Now we can show that

$$P(\sup_{\alpha \in S^{m-1}, \theta \in H} |G_n[\alpha^\top (\psi_{\theta} - \psi_{\tilde{\theta}_0})]| > 8\epsilon) \to 0.$$

Let $N(\epsilon, F, L_1(Q))$ be the covering number of $F$ with respect to the seminorm $L_1(Q)$ and let $N(\epsilon, H, \| \cdot \|)$ and $N(\epsilon, S^{m-1}, \| \cdot \|)$ be the covering number of $H$ and $S^{m-1}$ respectively, where $\| \cdot \|$ denotes the Frobenious norm. Since

$$|\alpha^\top [\psi_{\theta}(x, y) - \psi_{\tilde{\theta}_0}(x, y)]| \leq \|\theta - \tilde{\theta}_0\| F(x, y),$$

26
Since (similar to the arguments of Lemma 1.18 of Rigollet and Hüttner (2017)), we have

\[ N(\epsilon, H, \| \cdot \|) \leq N(\epsilon, S^{m-1}, \| \cdot \|) \]

Similiar to the arguments of Lemma 1.18 of Rigollet and Hüttner (2017), we have

\[ N(\epsilon, H, \| \cdot \|) \leq \left( \frac{3}{\epsilon} \sqrt{\frac{m \log n}{n}} \right)^m \text{ and } N(\epsilon, S^{m-1}, \| \cdot \|) \leq \left( \frac{3}{\epsilon} \right)^m, \]

whence

\[ N(2\epsilon QF, F, L_1(Q)) \leq \left( \frac{9}{\epsilon^2} \sqrt{\frac{m \log n}{n}} \right)^m. \]

Let \( \epsilon_n = \epsilon / \sqrt{n} \) and \( \delta_n = \sqrt{(m \log n)/n} \). Since \( \| \theta - \hat{\theta}_0 \| \leq \delta_n \), it follows that

\[ \text{Var}\{P_n[\alpha^\top (\psi_y - \psi_{\hat{\theta}_0})]\} = \frac{E[\| \theta - \hat{\theta}_0 \|^2 F(x,y)^2]}{16\epsilon^2} \leq \frac{C(m^2 \log n)/n}{16\epsilon^2} \to 0. \]

Similar to the arguments of Lemma 2 of Tan and Zhu (2019b), we have

\[ \mathbb{P}(\sup_{\alpha \in S^{m-1}, \theta \in H} |G_n[\alpha^\top (\psi_y - \psi_{\hat{\theta}_0})]| > 8\epsilon) = \mathbb{P}(\sup_{\alpha \in S^{m-1}, \theta \in H} |P_n[\alpha^\top (\psi_y - \psi_{\hat{\theta}_0})] - P[\alpha^\top (\psi_y - \psi_{\hat{\theta}_0})]| > 8\epsilon_n) \leq 4E\left\{ 2N(\epsilon_n, F, L_1(P_n)) \exp\left\{ -\frac{1}{2} n\epsilon_n^2 / (\delta_n^2 P_n F^2) \right\} \wedge 1 \right\} \leq 8E\left\{ \left( \frac{36\delta_n(P_n F)^2}{\epsilon_n^2} \right)^m \exp\left\{ -\frac{1}{2} n\epsilon_n^2 / (\delta_n^2 P_n F^2) \right\} \wedge 1 \right\} \leq 8 \left( \frac{36\delta_n \tau_n^2}{\epsilon_n^2} \right)^m \exp\left\{ -\frac{1}{2} n\epsilon_n^2 / (\delta_n^2 \tau_n^2) \right\} + 4\mathbb{P}(P_n F^2 > \tau_n^2). \]

Let \( \tau_n = \sqrt{m \log n} \). Then it follows that

\[ \mathbb{P}(P_n F^2 > \tau_n^2) \leq \frac{E(P_n F^2)}{\tau_n^2} = \frac{E[F(X,Y)^2]}{\tau_n^2} \to 0. \]

Since \((m \log n)^2/n \to 0\), it follows that

\[ \left( \frac{36\delta_n \tau_n^2}{\epsilon_n^2} \right)^m \exp\left\{ -\frac{1}{2} n\epsilon_n^2 / (\delta_n^2 \tau_n^2) \right\} \to 0. \]

Thus we obtain that \( \mathbb{P}(\sup_{\alpha \in S^{m-1}, \theta \in H} |G_n[\alpha^\top (\psi_y - \psi_{\hat{\theta}_0})]| > 8\epsilon) = o(1) \) and then \( G_n[\alpha^\top (\psi_{\hat{\theta}} - \psi_{\hat{\theta}_0})] = o_p(1) \) uniformly in \( \alpha \in S^{m-1} \).

Note \( P_n \psi_{\hat{\theta}} = 0 \) and \( P \psi_{\hat{\theta}_0} = 0 \). Then it follows that

\[
G_n(\alpha^\top \psi_{\hat{\theta}}) = \sqrt{n}(P_n \alpha^\top \psi_{\hat{\theta}} - P \alpha^\top \psi_{\hat{\theta}}) = -\sqrt{n}(P \alpha^\top \psi_{\hat{\theta}} - P \alpha^\top \psi_{\hat{\theta}_0}) = G_n(\alpha^\top \psi_{\hat{\theta}_0}) + G_n[\alpha^\top (\psi_{\hat{\theta}} - \psi_{\hat{\theta}_0})] = G_n(\alpha^\top \psi_{\hat{\theta}_0}) + o_p(1)
\]
Thus we obtain \( G_n(\alpha^\top \psi_{\tilde{\theta}_0}) = -\sqrt{n}(P\alpha^\top \psi_{\tilde{\theta}} - P\alpha^\top \psi_{\theta_0}) + o_p(1) \). Replacing \( \alpha \) with \( \Sigma^{-1}\alpha \), it follows that
\[
\alpha^\top \Sigma^{-1} G_n(\psi_{\tilde{\theta}_0}) = -\sqrt{n}\alpha^\top \Sigma^{-1}(P\psi_{\tilde{\theta}} - P\psi_{\theta_0}) + o_p(1).
\]
Applying Taylor expansion, we have
\[
P\psi_{\tilde{\theta}} - P\psi_{\theta_0} = \hat{P}\psi_{\theta|\theta=\tilde{\theta}_0}(\hat{\theta} - \tilde{\theta}_0) + (\hat{\theta} - \tilde{\theta}_0)^\top (\hat{P}\psi_{\theta|\theta=\tilde{\theta}_0})(\hat{\theta} - \tilde{\theta}_0)
\]
where \( \tilde{\theta}_2 \) lies between \( \hat{\theta} \) and \( \tilde{\theta}_0 \). According to Assumption 2, we have
\[
(\hat{\theta} - \tilde{\theta}_0)^\top (\hat{P}\psi_{\theta|\theta=\tilde{\theta}_2})(\hat{\theta} - \tilde{\theta}_0) = \frac{m^3/2}{n} O_p(1),
\]
whence,
\[
\sqrt{n}\alpha^\top (\hat{\theta} - \tilde{\theta}_0) = \alpha^\top \Sigma^{-1} G_n(\psi_{\tilde{\theta}_0}) + \sqrt{\frac{m^3}{n}} O_p(1) + o_p(1).
\]
Since \( (m^3 \log n)/n \to 0 \), it follows that
\[
\alpha^\top (\hat{\theta} - \tilde{\theta}_0) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(\theta, X_i)] \alpha^\top \Sigma^{-1} g'(\theta, X_i) + o_p(\frac{1}{\sqrt{n}}).
\]
Hence we complete the proof.

**Proof of Theorem 2.3.** We use the same arguments for Theorem 2.1. For the linear model, we have \( \psi_{\theta}(X, Y) = X(Y - \theta^\top X) \). Thus
\[
\psi_{\bar{\theta}_0 + \zeta}(X, Y) - \psi_{\bar{\theta}_0}(X, Y) = -XX^\top \zeta,
\]
whence
\[
P_n(\psi_{\bar{\theta}_0 + \zeta} - \psi_{\bar{\theta}_0}) - P(\psi_{\bar{\theta}_0 + \zeta} - \psi_{\bar{\theta}_0}) = -(\frac{1}{n} \sum_{i=1}^{n} X_iX_i^\top - EXX^\top)\zeta.
\]
Next we show that
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} X_iX_i^\top - EXX^\top \right\|_2 = O_p(\sqrt{\frac{p}{n}}),
\]
where \( \| \cdot \|_2 \) is the operator norm of a matrix \( A \) that is define in the proof of Theorem 2.2. Set \( EX = \mu \). Then it follows that
\[
\frac{1}{n} \sum_{i=1}^{n} X_iX_i^\top - EXX^\top = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^\top - E(X - \mu)(X - \mu)^\top + \frac{1}{n} \sum_{i=1}^{n} \mu X_i^\top - \mu \mu^\top + \frac{1}{n} \sum_{i=1}^{n} X_i \mu^\top - \mu \mu^\top =: A_1 + A_2 + A_3.
\]
According to the proof of Lemma 3 of Cai et al. (2010), there exist vectors \( \{v_j \in \mathbb{S}^{p-1}, 1 \leq j \leq 7^p\} \) such that
\[
\|A\|_2 \leq 4 \sup_{1 \leq j \leq 7^p} |v_j^T Av_j|
\]
for any symmetric matrix \( A \), where \( \mathbb{S}^{p-1} \) is the unit sphere with respect to the Frobenious norm. Then it follows that
\[
\mathbb{P}(\|\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^\top - E(X - \mu)(X - \mu)^\top\|_2 > t) \leq \mathbb{P}(\sup_{1 \leq j \leq 7^p} |\frac{1}{n} \sum_{i=1}^{n} v_j^T (X_i - \mu)(X_i - \mu)^\top v_j - Ev_j^T (X - \mu)(X - \mu)^\top v_j| > t) = \mathbb{P}(\sup_{1 \leq j \leq 7^p} |\frac{1}{n} \sum_{i=1}^{n} |v_j^T (X_i - \mu)|^2 - Ev_j^T (X - \mu)|^2| > t).
\]
Note that \( v^T (X - \mu) \) is sub-Gaussian with proxy variance \( \sigma^2 \). By Lemma 1.12 of Rigollet and Hütter (2017), we obtain that \( |v^T (X - \mu)|^2 - E|v^T (X - \mu)|^2 \) is sub-exponential with parameter \( \lambda = 16\sigma^2 \). By Berstein’s inequality, see e.g., Theorem 1.13 of Rigollet and Hütter (2017),
\[
\mathbb{P}(\frac{1}{n} \sum_{i=1}^{n} |v^T (X_i - \mu)|^2 - E|v^T (X - \mu)|^2| > t) < 2 \exp\{-\frac{1}{2} n(t^2/\lambda^2)\},
\]
where \( a \wedge b \equiv \min(a, b) \). Let \( t = \sqrt{p/nM} \) and then \( t/\lambda < 1 \) for \( n \) large enough. Thus \( t^2/\lambda^2 < t/\lambda \).
Therefore,
\[
\mathbb{P}(\|\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^\top - E(X - \mu)(X - \mu)^\top\|_2 > \sqrt{\frac{p}{nM}}) \leq \frac{7^p}{nM} \mathbb{P}(\|\frac{1}{n} \sum_{i=1}^{n} |v_j^T (X_i - \mu)|^2 - Ev_j^T (X - \mu)|^2| > \sqrt{\frac{p}{nM}}) \leq \frac{7^p}{nM} \cdot 2 \exp\{-\frac{pM^2}{2\lambda^2}\} = 2 \exp\{p(\log 7 - \frac{M^2}{2\lambda^2})\}.
\]
Then \( \|A_1\|_2 = O_p(\sqrt{p/n}) \).

For the second term \( A_2 \), we have
\[
\mathbb{P}(\|\frac{1}{n} \sum_{i=1}^{n} \mu(X_i - \mu)^\top\|_2 > \sqrt{\frac{p}{nM}}) \leq \frac{7^p}{nM} \mathbb{P}(\|\frac{1}{n} \sum_{i=1}^{n} v_j^T \mu(X_i - \mu)^\top v_j| > \sqrt{\frac{p}{nM}}) = \frac{7^p}{nM} \mathbb{P}(\|\frac{M}{n |v_j^\top \mu|}\|_2 > \sqrt{\frac{p}{nM}}).
\]

29
Since \( v_j^\top (X_i - \mu) \) is sub-Gaussian with proxy variance \( \sigma^2 \), by Corollary 1.7 of Rigollet and Hütter (2017), we obtain
\[
\mathbb{P}(\frac{1}{n} \sum_{i=1}^{n} v_j^\top (X_i - \mu) > t) \leq 2 \exp(-\frac{n t^2}{2\sigma^2}),
\]
whence,
\[
\mathbb{P}(\| \frac{1}{n} \sum_{i=1}^{n} \mu(X_i - \mu)^\top \|_2 > \sqrt{\frac{p}{n}} M) \leq 2 \exp\{p(\log 7 - \frac{M^2}{2\sigma^2|v_j^\top \mu|^2})\}.
\]
Since \( v_j^\top E(X - \mu)(X - \mu)^\top v_j \geq 0 \), it follows that \( v_j^\top \mu^\top v_j \leq v_j^\top E(XX^\top) v_j \). Note that \( \Sigma = E(XX^\top) \). Thus we obtain that
\[
|v_j^\top \mu|^2 \leq v_j^\top \Sigma v_j \leq \lambda_{\max}(\Sigma) \leq C.
\]
Therefore
\[
\mathbb{P}(\| \frac{1}{n} \sum_{i=1}^{n} \mu(X_i - \mu)^\top \|_2 > \sqrt{\frac{p}{n}} M) \leq 2 \exp\{p(\log 7 - \frac{M^2}{2C\sigma^2})\}.
\]
Hence \( \|A_2\|_2 = O_p(\sqrt{p/n}) \). Similarly, we have \( \|A_3\|_2 = O_p(\sqrt{p/n}) \). Consequently,
\[
n\zeta^\top [P_n(\psi_{\theta_0 + \zeta} - \psi_{\theta_0}) - P(\psi_{\theta_0 + \zeta} - \psi_{\theta_0})] = \sqrt{pM}\|\zeta\|^2 O_p(1).
\]
Similar to the arguments for Theorem 2.1, we derive
\[
n\zeta^\top P_n\psi_{\theta_0} = \sqrt{pM}\|\zeta\| O_p(1) \quad \text{and} \quad n\zeta^\top P\psi_{\theta_0 + \zeta} = -n\zeta^\top \Sigma \zeta.
\]
Let \( \zeta = \sqrt{p/n} U \). If \( p/n \to 0 \), then we have
\[
\zeta^\top F(\zeta) = \sqrt{pM}\|\zeta\| O_p(1) - n\zeta^\top \Sigma \zeta + \sqrt{pM}\|\zeta\|^2 O_p(1)
\leq p\|U\| (O_p(1) - n\zeta^\top \Sigma \zeta + \sqrt{pM}\|\zeta\|^2 O_p(1))
= p\|U\| \{O_p(1) - \zeta^\top \Sigma \zeta + \sqrt{pM}\|\zeta\|^2 O_p(1) \}.
\]
The rest of the proof follows the same arguments for Theorem 2.1. Hence we obtain the norm consistency of \( \hat{\theta} \) to \( \theta_0 \).

Now we derive the uniformly asymptotically linear representation of \( \hat{\theta} - \theta_0 \). Recall that \( \psi_\theta(x, y) = x(y - \theta^\top x) \) and \( \Sigma = E(XX^\top) \). Since
\[
\psi_\theta(x, y) - \psi_{\theta_0}(x, y) = -xx^\top (\theta - \theta_0),
\]
it follows that
\[
G_n[\alpha^\top \Sigma^{-1}(\psi_{\theta} - \psi_{\theta_0})] = \sqrt{n} \alpha^\top \Sigma^{-1}[P_n(\psi_{\theta} - \psi_{\theta_0}) - P(\psi_{\theta} - \psi_{\theta_0})]
\]
\[ = -\sqrt{n}\alpha^\top\Sigma^{-1}\left(\frac{1}{n} \sum_{i=1}^{n} X_iX_i^\top - \Sigma\right)(\hat{\theta} - \tilde{\theta}_0). \]

Consequently,
\[
|G_n[\alpha^\top\Sigma^{-1}(\psi_{\hat{\theta}} - \psi_{\tilde{\theta}_0})]| \leq \sqrt{n}\|\alpha^\top\Sigma^{-1}\| \cdot \|\frac{1}{n} \sum_{i=1}^{n} X_iX_i^\top - \Sigma\|_2 \cdot \|\hat{\theta} - \tilde{\theta}_0\| = O_p\left(\frac{p^2}{n}\right) = o_p(1).
\]

Thus \(G_n[\alpha^\top\Sigma^{-1}(\psi_{\hat{\theta}} - \psi_{\tilde{\theta}_0})] = o_p(1)\) uniformly in \(\alpha \in S^{q-1}\). Note \(P_n\psi_{\hat{\theta}} = 0\) and \(P\psi_{\tilde{\theta}_0} = 0\). Then we have
\[
\alpha^\top\Sigma^{-1}G_n\psi_{\hat{\theta}} = \alpha^\top\Sigma^{-1}G_n\psi_{\tilde{\theta}_0} + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - \tilde{\theta}_0^\top X_i)\alpha^\top\Sigma^{-1}X_i + o_p(1)
\]
whence,
\[
\sqrt{n}\alpha^\top(\hat{\theta} - \tilde{\theta}_0) = \alpha^\top\Sigma^{-1}\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - \tilde{\theta}_0^\top X_i)X_i + o_p(1).
\]
Hence we complete the proof. \(\Box\)

**Proof of Theorem 3.1.** This proof is similar to that for proving Proposition 3 in Tan and Zhu (2019a). Thus we omit the details here.

**Proof of Theorem 4.1.** Under the null hypothesis, Theorem 3.1 shows that \(P(\hat{q} = d) \to 1\). Thus we only need to work on the event \(\{\hat{q} = d\}\). Recall that
\[
\hat{V}_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [Y_j - g(\hat{\beta}^\top X_j, \hat{\theta})][\cos(t^\top \hat{\mathbf{B}}^\top X_j) + \sin(t^\top \hat{\mathbf{B}}^\top X_j)].
\]
Here \(t \in \mathbb{R}^{2d}\) and \(\hat{\mathbf{B}} = (\hat{\beta}, \hat{B}) \in \mathbb{R}^{p \times 2d}\) is a random matrix. Let \(\hat{\theta} = [\text{vec}(\hat{\beta})^\top, \hat{\theta}^\top]^\top\) as we have defined before. Then we rewrite \(\hat{V}_n^1(t)\) as
\[
\hat{V}_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [Y_j - g(\hat{\theta}, X_j)][\cos(t^\top \hat{\mathbf{B}}^\top X_j) + \sin(t^\top \hat{\mathbf{B}}^\top X_j)].
\]
In the following we decompose the proof into three steps.

**Step 1.** To show that
\[
\hat{V}_n^1(t) = V_n^1(t) + R_n(t) \quad \text{and} \quad \int_{\mathbb{R}^2} |R_n(t)|^2 \varphi(t)dt = o_p(1).
\]
Here $R_n(t)$ is a remainder and

$$V_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j [\cos(t^\top \mathbb{B}_0^\top X_j) + \sin(t^\top \mathbb{B}_0^\top X_j)] - M(t)^\top \Sigma^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j g'(\theta_0, X_j) X_j$$

with $\mathbb{B}_0 = (\beta_0, \mathbb{B})$ and $M(t) = E\{g'(\theta_0, X)[\cos(t^\top \mathbb{B}_0^\top X) + \sin(t^\top \mathbb{B}_0^\top X)]\}$.

Decompose $\hat{V}_n^1(t)$ as

$$\hat{V}_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j [\cos(t^\top \mathbb{B}_0^\top X_j) + \sin(t^\top \mathbb{B}_0^\top X_j)]$$

$$- \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [g(\hat\theta, X_j) - g(\theta_0, X_j)][\cos(t^\top \mathbb{B}_0^\top X_j) + \sin(t^\top \mathbb{B}_0^\top X_j)]$$

$$= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j [\cos(t^\top \mathbb{B}_0^\top X_j) + \sin(t^\top \mathbb{B}_0^\top X_j)]$$

$$+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j [\cos(t^\top \mathbb{B}_0^\top X_j) - \cos(t^\top \mathbb{B}_0^\top X_j) + \sin(t^\top \mathbb{B}_0^\top X_j) - \sin(t^\top \mathbb{B}_0^\top X_j)]$$

$$- \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [g(\hat\theta, X_j) - g(\theta_0, X_j)][\cos(t^\top \mathbb{B}_0^\top X_j) + \sin(t^\top \mathbb{B}_0^\top X_j)]$$

$$- \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [g(\hat\theta, X_j) - g(\theta_0, X_j)][\cos(t^\top \mathbb{B}_0^\top X_j) - \cos(t^\top \mathbb{B}_0^\top X_j) + \sin(t^\top \mathbb{B}_0^\top X_j) - \sin(t^\top \mathbb{B}_0^\top X_j)]$$

$$=: V_{n1}(t) + V_{n2}(t) - V_{n3}(t) - V_{n4}(t),$$

First we show that

$$\int_{\mathbb{R}^2} V_{n2}(t)^2 \varphi(t) dt = o_p(1) \quad \text{and} \quad \int_{\mathbb{R}^2} V_{n4}(t)^2 \varphi(t) dt = o_p(1).$$

Decompose further $V_{n2}(t)$ as

$$V_{n2}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j [\cos(t^\top \mathbb{B}_0^\top X_j) - \cos(t^\top \mathbb{B}_0^\top X_j)] + \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j [\sin(t^\top \mathbb{B}_0^\top X_j) - \sin(t^\top \mathbb{B}_0^\top X_j)]$$

$$=: V_{n21}(t) + V_{n22}(t)$$

For the term $V_{n21}(t)$, by Taylor expansion, we have

$$V_{n21}(t) = - \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j (t^\top \mathbb{B}_0^\top X_j - t^\top \mathbb{B}_0^\top X_j) \sin(t^\top \mathbb{B}_0^\top X_j)$$

32
where \( \mathfrak{B}_1 \) lies between \( \mathfrak{B} \) and \( \mathfrak{B}_0 \). Then it follows that

\[
\int_{\mathbb{R}^d} |V_{n211}(t)|^2 \varphi(t) dt \leq \| \mathfrak{B} - \mathfrak{B}_0 \|^2 \int_{\mathbb{R}^d} \| t \|^2 \cdot \| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \epsilon_j \sin(t^\top \mathfrak{B}_0 X_j) X_j \| \varphi(t) dt.
\]

Note that

\[
E \left\{ \int_{\mathbb{R}^d} \| t \|^2 \cdot \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \epsilon_j \sin(t^\top \mathfrak{B}_0 X_j) X_j \| \varphi(t) dt \right\} = \int_{\mathbb{R}^d} \| t \|^2 E \left\{ \| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \epsilon_j \sin(t^\top \mathfrak{B}_0 X_j) X_j \| \varphi(t) dt \right\} = \int_{\mathbb{R}^d} \| t \|^2 E \{ \| X \|^2 \epsilon^2 \sin(t^\top \mathfrak{B}_0 X)^2 \} \varphi(t) dt \leq C_p.
\]

Consequently, we obtain that

\[
\int_{\mathbb{R}^d} |V_{n211}|^2 \varphi(t) dt = \frac{p^2}{n} O_p(1) = o_p(1).
\]

Similarly, we can obtain

\[
\int_{\mathbb{R}^d} |V_{n212}(t)|^2 \varphi(t) dt = \frac{p^4}{n^2} O_p(1) = o_p(1).
\]

For the third term \( V_{n213}(t) \) in \( V_{n21}(t) \), we have

\[
\int_{\mathbb{R}^d} |V_{n213}(t)|^2 \varphi(t) dt = \int_{\mathbb{R}^d} \left| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \epsilon_j (t^\top \mathfrak{B} X_j - t^\top \mathfrak{B}_0 X_j)^3 \sin(t^\top \mathfrak{B}_1 X_j) \right|^2 \varphi(t) dt \leq \int_{\mathbb{R}^d} \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \| \epsilon_j \| \cdot \| t^\top \mathfrak{B} X_j - t^\top \mathfrak{B}_0 X_j \|^3 \right\}^2 \varphi(t) dt \leq \| \mathfrak{B} - \mathfrak{B}_0 \|^6 \cdot \int_{\mathbb{R}^d} \| t \|^6 \varphi(t) dt \cdot \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \| \epsilon_j \| \| X_j \|^3 \right\}^2.
\]

Since \( E|\epsilon_j|^8 < C \) and \( E|X_{ji}|^8 < C \) for \( 1 \leq i \leq p \), it follows that

\[
E \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \| \epsilon_j \| \| X_j \|^3 \right\}^2 < Cnp^3.
\]
Combining this with $\|\hat{\mathcal{B}} - \mathcal{B}_0\| = O_p(\sqrt{p/n})$, we obtain
\[
\int_{\mathbb{R}^{2d}} |V_{n213}(t)|^2 \varphi(t)dt = \frac{p^6}{n^2} O_p(1) = o_p(1).
\]
Consequently, we have $\int_{\mathbb{R}^{2d}} |V_{n21}(t)|^2 \varphi(t)dt = o_p(1)$. By the same arguments for the term $V_{n22}$, we have $\int_{\mathbb{R}^{2d}} |V_{n22}(t)|^2 \varphi(t)dt = o_p(1)$. Then it follows that
\[
\int_{\mathbb{R}^{2d}} |V_{n2}(t)|^2 \varphi(t)dt = o_p(1).
\]

Now we consider the term $\int_{\mathbb{R}^{2d}} V_{n4}(t)^2 \varphi(t)dt$. Decompose $V_{n4}(t)$ as follows,
\[
V_{n4}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [(\hat{\theta} - \theta_0)^\top g'(\theta_0, X_j) + \frac{1}{2} (\hat{\theta} - \theta_0)^\top g''(\theta_1, X_j)(\hat{\theta} - \theta_0)] \times \left[-t^\top (\hat{\mathcal{B}} - \mathcal{B}_0)^\top X_j \sin(t^\top \mathcal{B}_0^\top X_j) - \frac{1}{2} (t^\top \mathcal{B}^\top X_j - t^\top \mathcal{B}_0^\top X_j)^2 \sin(t^\top \mathcal{B}_1^\top X_j) \right]
\]
\[
= -\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta} - \theta_0)^\top g'(\theta_0, X_j) X_j^\top \sin(t^\top \mathcal{B}_0^\top X_j)(\hat{\mathcal{B}} - \mathcal{B}_0) t
-\frac{1}{2\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta} - \theta_0)^\top g''(\theta_1, X_j)(\hat{\theta} - \theta_0) t^\top (\hat{\mathcal{B}} - \mathcal{B}_0)^\top X_j \sin(t^\top \mathcal{B}_1^\top X_j)
-\frac{1}{2\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta} - \theta_0)^\top g'(\theta_0, X_j)(t^\top \mathcal{B}^\top X_j - t^\top \mathcal{B}_0^\top X_j)^2 \sin(t^\top \mathcal{B}_1^\top X_j)
-\frac{1}{4\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta} - \theta_0)^\top g''(\theta_1, X_j)(\hat{\theta} - \theta_0)(t^\top \mathcal{B}^\top X_j - t^\top \mathcal{B}_0^\top X_j)^2 \sin(t^\top \mathcal{B}_1^\top X_j)
=: -V_{n41}(t) - V_{n42}(t) - V_{n43}(t) - V_{n44}(t),
\]
where $\theta_1$ lies between $\hat{\theta}$ and $\theta_0$ and $\mathcal{B}_1$ lies between $\mathcal{B}$ and $\mathcal{B}_0$. By the same arguments as that for the term $V_{n213}$, we have
\[
\int_{\mathbb{R}^{2d}} |V_{n41j}|^2 \varphi(t)dt = o_p(1), \quad \text{for } j = 2, 3, 4.
\]
For the term $V_{n41}$, set
\[
M_1(t) = E[g'(\theta_0, X)X^\top \sin(t^\top \mathcal{B}_0^\top X)].
\]
Then we have

\[
\int_{\mathbb{R}^{2d}} |V_{n411}(t)|^2 \varphi(t) dt \\
\leq 2n \int_{\mathbb{R}^{2d}} |(\hat{\theta} - \theta_0)^\top \{ \frac{1}{n} \sum_{j=1}^n g'(\theta_0, X_j)X_j^\top \sin(t^\top \mathbf{B}_0^\top X_j) - M_1(t) \} \times \\
(\mathbf{B} - \mathbf{B}_0) t^2 \varphi(t) dt + 2n \int_{\mathbb{R}^{2d}} |(\hat{\theta} - \theta_0)^\top M_1(t)(\mathbf{B} - \mathbf{B}_0) t^2 \varphi(t) dt \\
\leq 2n ||\hat{\theta} - \theta_0||^2 ||\mathbf{B} - \mathbf{B}_0||^2 \int_{\mathbb{R}^{2d}} \frac{1}{n} \sum_{j=1}^n g'(\theta_0, X_j)X_j^\top \sin(t^\top \mathbf{B}_0^\top X_j) - M_1(t) \|^2 \times \\
||t||^2 \varphi(t) dt + Cn ||\hat{\theta} - \theta_0||^2 ||\mathbf{B} - \mathbf{B}_0||^2.
\]

Note that

\[
E \left( \int_{\mathbb{R}^{2d}} \frac{1}{n} \sum_{j=1}^n g'(\theta_0, X_j)X_j^\top \sin(t^\top \mathbf{B}_0^\top X_j) - M_1(t) \|^2 ||t||^2 \varphi(t) dt \right) \\
= \sum_{k=1}^m \sum_{l=1}^p \int_{\mathbb{R}^{2d}} E \left[ \frac{1}{n} \sum_{j=1}^n g_k(\theta_0, X_j)X_{jl} \sin(t^\top \mathbf{B}_0^\top X_j) - M_{1kl}(t) \|^2 ||t||^2 \varphi(t) dt \right] \\
\leq \frac{1}{n} \sum_{k=1}^m \sum_{l=1}^p E |g'_k(\theta_0, X_j)X_{jl}|^2 \leq C\frac{p^2}{n}.
\]

Thus we obtain \( \int_{\mathbb{R}^{2d}} |V_{n411}(t)|^2 \varphi(t) dt = (p^2/n)O_p(1) = o_p(1) \) and then \( \int_{\mathbb{R}^{2d}} |V_{n41}(t)|^2 \varphi(t) dt = o_p(1) \). The same arguments show that \( \int_{\mathbb{R}^{2d}} |V_{n42}(t)|^2 \varphi(t) dt = o_p(1) \). Hence we have

\[
\int_{\mathbb{R}^{2d}} |V_{n4}(t)|^2 \varphi(t) dt = o_p(1).
\]

For the term \( V_{n3}(t) \), recall that

\[
V_{n3}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n [g(\hat{\theta}, X_j) - g(\theta_0, X_j)] \{ \cos(t^\top \mathbf{B}_0^\top X_j) + \sin(t^\top \mathbf{B}_0^\top X_j) \}.
\]

Decompose \( V_{n3}(t) \) as follows,

\[
V_{n3}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n (\hat{\theta} - \theta_0)^\top g'(\theta_0, X_j) \{ \cos(t^\top \mathbf{B}_0^\top X_j) + \sin(t^\top \mathbf{B}_0^\top X_j) \} \\
+ \frac{1}{2\sqrt{n}} \sum_{j=1}^n (\hat{\theta} - \theta_0)^\top g''(\theta_0, X_j)(\hat{\theta} - \theta_0) \{ \cos(t^\top \mathbf{B}_0^\top X_j) + \sin(t^\top \mathbf{B}_0^\top X_j) \} \\
+ \frac{1}{6\sqrt{n}} \sum_{j=1}^n (\hat{\theta} - \theta_0)^\top ([\hat{\theta} - \theta_0]^\top g'''(\theta_2, X_j)(\hat{\theta} - \theta_0)] \{ \cos(t^\top \mathbf{B}_0^\top X_j) + \sin(t^\top \mathbf{B}_0^\top X_j) \}
\]

35
By Theorem 2.2, we have
\[ V_{n32}(t) + \frac{1}{2} V_{n31}(t) + \frac{1}{6} V_{n33}(t), \]
where \( \theta_2 \) lies between \( \hat{\theta} \) and \( \theta_0 \). Similar to the arguments as that for \( V_{n213} \) and \( V_{n411} \), we have
\[
\begin{align*}
\int_{\mathbb{R}^2d} |V_{n33}(t)|^2 \varphi(t)dt &= \frac{n^0}{n^2} O_p(1) = o_p(1); \\
\int_{\mathbb{R}^2d} |V_{n32}(t)|^2 \varphi(t)dt &= \frac{n^2}{n} O_p(1) = o_p(1);
\end{align*}
\]
For the first term \( V_{n31}(t) \) in \( V_{n3}(t) \), recall that
\[ M(t) = E\{g'(\theta_0, X)\{\cos(t^\top \mathcal{B}_0^\top X) + \sin(t^\top \mathcal{B}_0^\top X)\}\}. \]
Then it follows that
\[ V_{n31}(t) = \sqrt{n}(\hat{\theta} - \theta_0)^\top \{\frac{1}{n} \sum_{j=1}^{n} g'(\theta_0, X_j)\{\cos(t^\top \mathcal{B}_0^\top X_j) + \sin(t^\top \mathcal{B}_0^\top X_j)\} - M(t)\} \]
\[ = \sqrt{n}(\hat{\theta} - \theta_0)^\top M(t) \]
Note that
\[
\begin{align*}
\int_{\mathbb{R}^2d} |V_{n311}(t)|^2 \varphi(t)dt &\leq n||\hat{\theta} - \theta_0||^2 \int_{\mathbb{R}^2d} \left|\frac{1}{n} \sum_{j=1}^{n} g'(\theta_0, X_j)\{\cos(t^\top \mathcal{B}_0^\top X_j) + \sin(t^\top \mathcal{B}_0^\top X_j)\} - M(t)\right|^2 \varphi(t)dt \\
&\leq \frac{1}{n} \sum_{k=1}^{m} E|g'_k(\theta_0, X)|^2 \leq C_{n}^2.
\end{align*}
\]
Thus we obtain that
\[ \int_{\mathbb{R}^2d} |V_{n311}(t)|^2 \varphi(t)dt = \frac{n^2}{n} O_p(1) = o_p(1). \]
By Theorem 2.2, we have
\[ \sqrt{n}M(t)^\top (\hat{\theta} - \theta_0) = M(t)^\top \Sigma^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j g'(\theta_0, X_j) + o_p(1), \]
where $o_p(1)$ is uniformly in $t$. Altogether we obtain

\[ \dot{V}_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j [\cos(t^\top B_0^\top X_j) + \sin(t^\top B_0^\top X_j)] + M(t)^\top \Sigma^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j g'(\theta_0, X_j) + R_n(t) \]

\[ =: V_n^1(t) + R_n(t), \]

where the remainder $R_n(t)$ satisfies $\int_{\mathbb{R}^{2d}} |R_n(t)|^2 \varphi(t)dt = o_p(1)$.

**Step 2.** In this step we will show that

\[ V_n^1(t) \rightarrow V_\infty^1(t) \quad \text{in distribution in the space } C(\Gamma), \]

where $\Gamma$ is any compact subset in $\mathbb{R}^{2d}$, $C(\Gamma)$ is the space of real-valued continuous functions on $\Gamma$, and $V_\infty^1(t)$ is a zero-mean Gaussian process with a covariance function $\psi(s,t)$. By the Continuous Mapping Theorem, we obtain

\[ \int_\Gamma |V_n^1(t)|^2 \varphi(t)dt \rightarrow \int_\Gamma |V_\infty^1(t)|^2 \varphi(t)dt \quad \text{in distribution}. \]

First we show that $V_n^1(t)$ is asymptotically tight. Recall that

\[ V_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j [\cos(s^\top B_0^\top X_j) + \sin(s^\top B_0^\top X_j)] - M(t)^\top \Sigma^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j g'(\theta_0, X_j). \]

For any $s,t \in \Gamma$, we have

\[ E|V_n^1(s) - V_n^1(t)|^2 \]

\[ \leq 3E \left[ \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j [\cos(s^\top B_0^\top X_j) - \cos(t^\top B_0^\top X_j)]^2 + \right. \]

\[ \left. 3E \left[ \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j [\sin(s^\top B_0^\top X_j) - \sin(t^\top B_0^\top X_j)]^2 + \right. \right. \]

\[ \left. 3E[(M(s) - M(t))^\top \Sigma^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^n \varepsilon_j g'(\theta_0, X_j)]^2 \right. \]

\[ = 3E[\varepsilon [\cos(s^\top B_0^\top X) - \cos(t^\top B_0^\top X)]^2 + \]

\[ 3E[\varepsilon [\sin(s^\top B_0^\top X) - \sin(t^\top B_0^\top X)]^2 + \]

\[ 3[M(s) - M(t)]^\top \Sigma^{-1} E[\varepsilon^2 g'(\theta_0, X)g'(\theta_0, X)^\top] \Sigma^{-1} [M(s) - M(t)] \]

\[ \leq 6E[\varepsilon^2(s^\top B_0^\top X - t^\top B_0^\top X)^2] + C\|M(s) - M(t)\|^2 \]

\[ \leq 6\|s - t\|^2 E[\varepsilon^2 \|B_0^\top X\|^2] + C\left\| \frac{\partial M}{\partial t} (\xi)(s - t) \right\|^2, \]

37
where $\xi$ lies between $s$ and $t$. Note that

$$
\frac{\partial M}{\partial t}(\xi) = E\{g'(\theta_0, X)[\cos(\xi^T \mathbf{B}_0^T X) - \sin(\xi^T \mathbf{B}_0^T X)]X^T \mathbf{B}_0\}.
$$

Thus we have

$$
\|\frac{\partial M}{\partial t}(\xi)(s - t)\|^2 \leq C\|(s - t)\|^2.
$$

By condition (B1), we have $E[\varepsilon^2\|\mathbf{B}_0^T X\|^2] \leq C$. Then it follows that

$$
E|V_n^1(s) - V_n^1(t)|^2 \leq C\|(s - t)\|^2.
$$

By Theorem 12.3 of Billingsley (1968), we obtain that $V_n^1(t)$ is asymptotically tight.

Next we consider the convergence of finite-dimensional distributions of $V_n^1(t)$. Putting

$$
Z_{nj}(t) = \frac{1}{\sqrt{n}}\varepsilon_j[\cos(t^T \mathbf{B}_0^T X_j) + \sin(t^T \mathbf{B}_0^T X_j) - M(t)^T \Sigma^{-1} g'(\theta_0, X_j)].
$$

For any $t_1, \cdots, t_m \in \Gamma$, let $Z_{nj} = [Z_{nj}(t_1), \cdots, Z_{nj}(t_m)]^T$. For any $\delta > 0$, following the same arguments in Theorem 3.1 of Tan and Zhu (2019a), we obtain that

$$
\sum_{j=1}^{n} E[\|Z_{nj}\|^2 I(\|Z_{nj}\| > \delta)] = O(\sqrt{p^2/n}) = o(1).
$$

For the covariance matrix $\sum_{j=1}^{n} \text{Cov}(Z_{nj})$, we only need to calculate $\sum_{j=1}^{n} \text{Cov}\{Z_{nj}(s), Z_{nj}(t)\}$. It is easy to see that

$$
\sum_{j=1}^{n} \text{Cov}\{Z_{nj}(s), Z_{nj}(t)\}
$$

$$
= E\{\varepsilon^2[\cos((s - t)^T \mathbf{B}_0^T X) + \sin((s + t)^T \mathbf{B}_0^T X)]\} - M(t)^T \Sigma^{-1} E\{\varepsilon^2[\cos(s^T \mathbf{B}_0^T X) + \sin(s^T \mathbf{B}_0^T X)]g'(\theta_0, X)\} - M(s)^T \Sigma^{-1} E\{\varepsilon^2[\cos(t^T \mathbf{B}_0^T X) + \sin(t^T \mathbf{B}_0^T X)]g'(\beta_0, X)\} + M(s)^T \Sigma^{-1} E[\varepsilon^2 g'(\theta_0, X)g'(\theta_0, X)]\Sigma^{-1} M(t).
$$

Note that $\sum_{j=1}^{n} \text{Cov}[Z_{nj}(s), Z_{nj}(t)] = K_n(s, t)$ and $K_n(s, t) \to K(s, t)$ point-wisely in $(s, t)$. Then it follows that $Z_{nj}$ satisfies conditions of Lindeberg-Feller Central limit theorem and then convergence of finite-dimensional distribution of $V_n^1(t)$ holds. Hence we obtain that

$$
V_n^1(t) \to V_{\infty}^1(t) \quad \text{in distribution in the space } C(\Gamma),
$$

whence

$$
\int_{\Gamma} |V_n^1(t)|^2 \varphi(t) dt \to \int_{\Gamma} |V_{\infty}^1(t)|^2 \varphi(t) dt \quad \text{in distribution}.
$$
Step 3. In this step we will show that

$$\int_{\mathbb{R}^d} |V_n^1(t)|^2 \varphi(t) dt \longrightarrow \int_{\mathbb{R}^d} |V_\infty^1(t)|^2 \varphi(t) dt$$
in distribution.

Fix \( \varepsilon > 0 \). As \( \int_{\mathbb{R}^d} K(t,t) \varphi(t) dt < \infty \), there exists a compact subset \( \Gamma \subset \mathbb{R}^{2d} \) such that

$$\int_{\Gamma^c} K(t,t) \varphi(t) dt < \frac{\varepsilon^2}{2},$$

where \( \Gamma^c \) is the complementary set of \( \Gamma \) in \( \mathbb{R}^{2d} \). Note that \( K(t,t) = E|V_\infty^1(t)|^2 \). Then it follows that

$$E\left( \int_{\Gamma^c} V_\infty^1(t)^2 \varphi(t) dt \right) = \int_{\Gamma^c} E|V_\infty^1(t)|^2 \varphi(t) dt < \frac{\varepsilon^2}{2}.$$

Recall that

$$E|V_n^1(t)|^2 = K_n(t,t) \quad \text{and} \quad K_n(t,t) \longrightarrow K(t,t).$$

Then it follows that

$$E\left( \int_{\Gamma^c} V_n^1(t)^2 \varphi(t) dt \right) = \int_{\Gamma^c} K_n(t,t) \varphi(t) dt \longrightarrow \int_{\Gamma^c} K(t,t) \varphi(t) dt < \frac{\varepsilon^2}{2},$$

whence

$$E\left( \int_{\Gamma^c} V_n^1(t)^2 \varphi(t) dt \right) < \varepsilon^2, \quad \text{for } n \text{ large enough.}$$

Putting

$$V_{n1} = \int_{\Gamma^c} V_n^1(t)^2 \varphi(t) dt, \quad V_{n2} = \int_{\Gamma} V_n^1(t)^2 \varphi(t) dt,$$

$$V_1 = \int_{\Gamma^c} V_\infty^1(t)^2 \varphi(t) dt, \quad V_2 = \int_{\Gamma} V_\infty^1(t)^2 \varphi(t) dt.$$ 

Thus we have \( E(V_1) < \varepsilon^2/2 \), \( E(V_{n1}) < \varepsilon^2 \) for \( n \) large enough, and \( V_{n2} \rightarrow V_2 \) in distribution.

Then it follows that

$$\mathbb{P}(V_1 + V_2 \leq t - \varepsilon) - \varepsilon$$
$$\leq \mathbb{P}(V_2 \leq t - \varepsilon) - \varepsilon$$
$$= \liminf_{n \rightarrow \infty} \mathbb{P}(V_{n2}^1 \leq t - \varepsilon) - \varepsilon$$
$$= \liminf_{n \rightarrow \infty} \{ \mathbb{P}(V_{n1}^1 + V_{n2}^1 \leq t) + \mathbb{P}(V_{n1}^1 \geq \varepsilon) \} - \varepsilon$$
$$\leq \liminf_{n \rightarrow \infty} \mathbb{P}(V_{n1}^1 + V_{n2}^1 \leq t) + \varepsilon - \varepsilon, \quad \text{for } n \text{ large enough}$$
$$\leq \limsup_{n \rightarrow \infty} \mathbb{P}(V_{n1}^1 + V_{n2}^1 \leq t)$$
$$\leq \limsup_{n \rightarrow \infty} \mathbb{P}(V_{n2}^1 \leq t)$$
$$\leq \mathbb{P}(V_2 \leq t)$$
$$\leq \mathbb{P}(V_1 + V_2 \leq t + \varepsilon) + \mathbb{P}(V_1 \geq \varepsilon)$$

39
\[ \leq \mathbb{P}(V_1 + V_2 \leq t + \varepsilon) + \frac{\varepsilon}{2} \]

Let \( \varepsilon \) tend to zero, we obtain that

\[ \lim_{n \to \infty} \mathbb{P}(V_{n1}^1 + V_{n2}^1 \leq t) = \mathbb{P}(V_1 + V_2 \leq t). \]

Consequently,

\[ \int_{\mathbb{R}^d} |V_{n1}^1(t)|^2 \varphi(t) dt \longrightarrow \int_{\mathbb{R}^d} |V_{\infty}^1(t)|^2 \varphi(t) dt \quad \text{in distribution.} \]

In Step 1 we have shown

\[ \dot{V}_{n1}^1(t) = V_{n1}^1(t) + R_n(t) \quad \text{and} \quad \int_{\mathbb{R}^d} |R_n(t)|^2 \varphi(t) dt = o_p(1). \]

Thus we have

\[ \int_{\mathbb{R}^d} |\dot{V}_{n1}^1(t)|^2 \varphi(t) dt \longrightarrow \int_{\mathbb{R}^d} |V_{\infty}^1(t)|^2 \varphi(t) dt \quad \text{in distribution.} \]

Hence we complete the proof. \( \Box \)

**Proof of Theorem 4.2**: The proof follows the same line as that in the Proposition 4 in Tan and Zhu (2019a). Thus we omit it here. \( \Box \)

**Proof of Theorem 4.3**: We use the same notations as that in Theorem 2.1. Let \( F(\zeta) = \sum_{i=1}^n [Y_i - g(\theta_0 + \zeta, X_i)]g'(\theta_0 + \zeta, X_i) \) and \( \psi_0(x, y) = [y - g(\theta, x)]g'(\theta, x) \). Then we have

\[ \zeta^T F(\zeta) = n\zeta^T [P_n(\psi_{\theta_0+\zeta} - \psi_{\theta_0}) - P(\psi_{\theta_0+\zeta} - \psi_{\theta_0})] + n\zeta^T P_n \psi_{\theta_0} + n\zeta^T P(\psi_{\theta_0+\zeta} - \psi_{\theta_0}) \]

Similar to the arguments in Theorem 2.1, we have

\[ n\zeta^T (P_n \psi_{\theta_0} - P \psi_{\theta_0}) = \sqrt{np}\|\zeta\|O_p(1). \]
\[ n\zeta^T P(\psi_{\theta_0+\zeta} - \psi_{\theta_0}) = n\zeta^T P \psi_{\theta_0} \zeta + \sqrt{np}\|\zeta\|^3O_p(1). \]
\[ n\zeta^T [P_n(\psi_{\theta_0+\zeta} - \psi_{\theta_0}) - P(\psi_{\theta_0+\zeta} - \psi_{\theta_0})] = \sqrt{np}\|\zeta\|^2O_p(1). \]

Note that \( Y_n = g(\theta_0, X) + r_n G(B^T X) + \varepsilon \) and

\[ \dot{P} \psi_{\theta_0} = -E[g'(\theta_0, X)g'(\theta_0, X)^T] + E\{[Y_n - g(\theta_0, X)]g''(\theta_0, X)\}. \]

Then it follows that \( \dot{P} \psi_{\theta_0} = -\Sigma + r_n E[G(B^T X)g''(\theta_0, X)]. \) Therefore,

\[ n\zeta^T P(\psi_{\theta_0+\zeta} - \psi_{\theta_0}) = -n\zeta^T \Sigma \zeta + nr_n \zeta^T E[G(B^T X)g''(\theta_0, X)] \zeta + \sqrt{np}\|\zeta\|^3O_p(1) \]
\[ = -n\zeta^T \Sigma \zeta + nr_n p\|\zeta\|^2O_p(1) + \sqrt{np}\|\zeta\|^3O_p(1). \]
For $n\zeta^TP\psi_{\theta_0}$, we have

\[
n\zeta^TP\psi_{\theta_0} = n\zeta^T E\{[Y_n - g(\theta_0, X)]g'(\theta_0, X)\} \\
= nr_n\zeta^T E[G(B^TX)g'(\theta_0, X)] = nr_n\sqrt{p}\|\zeta\|O_p(1)
\]

Let $\zeta = \sqrt{pr_n}U$. Then we obtain

\[
\zeta^TF(\zeta) = \sqrt{np^2}\|\zeta\|^2O_p(1) + \sqrt{np}\|\zeta\|O_p(1) + nr_n\sqrt{p}\|\zeta\|O_p(1) \\
- n\zeta^T \Sigma \zeta + nr_n p\|\zeta\|^2 O_p(1) + \sqrt{np}\|\zeta\|^2 O_p(1) \\
\leq \sqrt{np^2r_n^2}\|U\|^2O_p(1) + \sqrt{npnr_n}\|U\|O_p(1) + npn^2r_n^2\|U\|^2O_p(1) \\
- np^2r_n\lambda_{\text{min}}(\Sigma)\|U\|^2 + np^2r_n^3\|U\|^2O_p(1) + np^2r_n^2\|U\|^3O_p(1) \\
= npn^2r_n\|U\|\left\{\frac{p}{\sqrt{np}}\|U\|O_p(1) + \frac{1}{\sqrt{nr_n}}O_p(1) + O_p(1) - \lambda_{\text{min}}(\Sigma)\|U\| \right. \\
+pr_n\|U\|O_p(1) + nr_n\|U\|^2O_p(1)\}
\]

Note that $p^2/n \to 0$ and $pr_n \to 0$. Following the same arguments in Theorem 2.1, we have

$\|\hat{\theta} - \theta_0\| = O_p(\sqrt{pr_n})$.

Next we show the asymptotically linear representation of $\hat{\theta} - \theta_0$ under the alternatives $H_{1n}$. Following the same line of the proof in Theorem 2.2, if $n(p \log n)^3 r_n^2 \to 0$, we have

\[
G_n[\alpha^T(\psi_{\hat{\theta}} - \psi_{\theta_0})] = o_p(1),
\]

uniformly in $\alpha \in \mathcal{S}^{q-1}$. Then it follows that

\[
G_n(\alpha^T \psi_{\hat{\theta}}) = \sqrt{n}(P_n \alpha^T \psi_{\hat{\theta}} - Pa^T \psi_{\theta_0}) = -\sqrt{n}(Pa^T \psi_{\hat{\theta}} - Pa^T \psi_{\theta_0}) - \sqrt{n}P\alpha^T \psi_{\theta_0} \\
= G_n(\alpha^T \psi_{\theta_0}) + G_n[\alpha^T(\psi_{\hat{\theta}} - \psi_{\theta_0})] = G_n(\alpha^T \psi_{\theta_0}) + o_p(1)
\]

Replacing $\alpha$ with $\Sigma^{-1}\alpha$, we have

\[
\sqrt{n}P_n(\alpha^T \Sigma^{-1} \psi_{\theta_0}) = -\sqrt{n}\alpha^T \Sigma^{-1}(P\psi_{\hat{\theta}} - P\psi_{\theta_0}) + o_p(1).
\]

Applying the Taylor’s expansion, we have

\[
P\psi_{\hat{\theta}} - P\psi_{\theta_0} = \hat{P}\psi_{\theta_0}|_{\theta = \theta_0}(\hat{\theta} - \theta_0) + (\hat{\theta} - \theta_0)^T(\hat{P}\psi_{\theta_0}|_{\theta = \hat{\theta}_3})(\hat{\theta} - \theta_0) \\
= \{-\Sigma + r_n E[G(B^TX)g''(\theta_0, X)]\}(\hat{\theta} - \theta_0) + (\hat{\theta} - \theta_0)^T(\hat{P}\psi_{\theta_0}|_{\theta = \hat{\theta}_3})(\hat{\theta} - \theta_0) \\
= -\Sigma(\hat{\theta} - \theta_0) + r_n^2 p^{3/2}O(1).
\]

where $\Sigma = E[g'(\theta_0, X)g'(\theta_0, X)^T]$ and $\hat{\theta}_3$ lies between $\hat{\theta}$ and $\theta_0$. Therefore,

\[
\sqrt{n}\alpha^T \Sigma^{-1}(P\psi_{\hat{\theta}} - P\psi_{\theta_0}) = -\sqrt{n}\alpha^T(\hat{\theta} - \theta_0) + \sqrt{np^3r_n^2}O(1)
\]
Since

\[
P_n(\psi_{q_0}) = \frac{1}{n} \sum_{j=1}^{n} [Y_{nj} - g(\theta_0, X_j)] g'(\theta_0, X_j)
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j g'(\theta_0, X_j) + \frac{1}{n} \sum_{j=1}^{n} r_n G(B^\top X_j) g'(\theta_0, X_j)
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j g'(\theta_0, X_j) + r_n E [G(B^\top X) g'(\theta_0, X)] + \sqrt{n} r_n O_p(1),
\]

it follows that

\[
\sqrt{n} \alpha^\top (\hat{\theta} - \theta_0) = \alpha^\top \Sigma^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j g'(\theta_0, X_j) + \sqrt{n} r_n \alpha^\top \Sigma^{-1} E [G(B^\top X) g'(\theta_0, X)] + o_p(1).
\]

Then we complete the proof. \[\square\]

**Proof of Theorem 4.4** (1) Under the global alternative hypothesis \(H_1\), we have \(P(\hat{q} = q) \to 1\). Thus we only work on the event \(\{\hat{q} = q\}\). Recall that

\[
\hat{V}_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [Y_j - g(\hat{\theta}, X_j)] \left[ \cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j) \right].
\]

Since \(Y_j = G(B^\top X_j) + \varepsilon_j\) under \(H_1\), it follows that

\[
\frac{1}{\sqrt{n}} \hat{V}_n^1(t) = \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)]
\]

\[
- \frac{1}{n} \sum_{j=1}^{n} [g(\hat{\theta}, X_j) - g(\hat{\theta}_0, X_j)] [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)]
\]

\[
+ \frac{1}{n} \sum_{j=1}^{n} [G(B^\top X_j) - g(\hat{\theta}_0, X_j)] [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)]
\]

\[
=: T_{n1}(t) - T_{n2}(t) + T_{n3}(t),
\]

Following the same line of the arguments in the Theorem 3.1, we have

\[
\int_{\mathbb{R}^{d+q}} |T_{n1}(t)|^2 \varphi(t) dt = o_p(1) \quad \text{and} \quad \int_{\mathbb{R}^{d+q}} |T_{n2}(t)|^2 \varphi(t) dt = o_p(1).
\]

For the third term \(T_{n3}(t)\) in \((1/\sqrt{n})\hat{V}_n^1(t)\), we have

\[
T_{n3}(t) = \frac{1}{n} \sum_{j=1}^{n} [G(B^\top X_j) - g(\hat{\theta}_0, X_j)] [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)]
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} [G(B^\top X_j) - g(\hat{\theta}_0, X_j)] [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)]
\]

42
Then we have
\[
-\frac{1}{n} \sum_{j=1}^{n} \left[ G(B^\top X_j) - g(\hat{\theta}_0, X_j) \right] [\cos(t^\top \hat{B}^\top X_j) - \cos(t^\top \bar{B}^\top X_j)]
\]
\[
-\frac{1}{n} \sum_{j=1}^{n} \left[ G(B^\top X_j) - g(\hat{\theta}_0, X_j) \right] [\sin(t^\top \hat{B}^\top X_j) - \sin(t^\top \bar{B}^\top X_j)]
\]
\[
=: \ T_{n31}(t) - T_{n32}(t) - T_{n33}(t)
\]
where \( \hat{B} = (\hat{\beta}_0, B) \). Similar to the argument for the term \( V_{n21}(t) \) in the proof of Theorem 4.1, we have
\[
\int_{\mathbb{R}^{d+q}} |T_{n3k}(t)|^2 \varphi(t) dt = O_p \left( \frac{p^2}{n} \right) = o_p(1), \quad k = 2, 3.
\]
Let \( L_n(t) = E[T_{n31}(t)] \). Then it follows that
\[
E \left( \int_{\mathbb{R}^{d+q}} |T_{n31}(t) - L_n(t)|^2 \varphi(t) dt \right)
\]
\[
= \frac{1}{n} \int_{\mathbb{R}^{d+q}} Var \left\{ \left[ G(B^\top X_j) - g(\hat{\theta}_0, X_j) \right] [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)] \right\} \varphi(t) dt
\]
\[
\leq \frac{1}{n} E \left[ G(B^\top X_j) - g(\hat{\theta}_0, X_j) \right]^2 \leq C \frac{1}{n}.
\]
Then we have \( \int_{\mathbb{R}^{d+q}} |T_{n31}(t) - L_n(t)|^2 \varphi(t) dt = o_p(1) \). Altogether we obtain that
\[
\frac{1}{\sqrt{n}} \hat{V}_n(t) = L_n(t) + R_n(t),
\]
where \( R_n(t) \) is a remainder satisfying \( \int_{\mathbb{R}^{d+q}} |R_n(t)|^2 \varphi(t) dt = o_p(1) \). Therefore, we have in probability
\[
\int_{\mathbb{R}^{d+q}} \left| \frac{1}{\sqrt{n}} \hat{V}_n(t) \right|^2 \varphi(t) dt = \int_{\mathbb{R}^{d+q}} |L_n(t)|^2 \varphi(t) dt + o_p(1) \to C_1 > 0.
\]

(2) We only work on the event \( \{ \hat{q} = d \} \) as \( \mathbb{P}(\hat{q} = d) \to 1 \) under the alternatives \( H_{n1} \) with \( pr_n \to 0 \). Since \( Y_{nj} = g(\hat{\theta}, X_j) + r_n G(B^\top X_j) + \varepsilon_j \), we decompose \( \hat{V}_n(t) \) as follows,
\[
\hat{V}_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left[ Y_j - g(\hat{\theta}, X_j) \right] [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)]
\]
\[
= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)]
\]
\[
- \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left[ g(\hat{\theta}, X_j) - g(\hat{\theta}_0, X_j) \right] [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)]
\]
\[
+ \sqrt{n} r_n \frac{1}{n} \sum_{j=1}^{n} G(B^\top X_j) [\cos(t^\top \hat{B}^\top X_j) + \sin(t^\top \hat{B}^\top X_j)]
\]
\[
=: W_{n1}(t) - W_{n2}(t) + W_{n3}(t)
\]
For the first term $W_{n1}(t)$ in $\hat{V}_n^1(t)$, we have

$$W_{n1}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \cos(t^\top \mathfrak{B}_0^\top X_j) + \sin(t^\top \mathfrak{B}_0^\top X_j)$$

$$+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \cos(t^\top \mathfrak{B}^\top X_j) - \cos(t^\top \mathfrak{B}_0^\top X_j) + \sin(t^\top \mathfrak{B}^\top X_j) - \sin(t^\top \mathfrak{B}_0^\top X_j)$$

$$=: W_{n11}(t) + W_{n12}(t)$$

Then it is easy to see that

$$E \left( \int_{\mathbb{R}^2d} |W_{n11}(t)|^2 \varphi(t) dt \right) \leq 2E(\varepsilon^2) \int_{\mathbb{R}^2d} \varphi(t) dt.$$

Thus we obtain that $\int_{\mathbb{R}^2d} |W_{n11}(t)|^2 \varphi(t) dt = O_p(1)$. By the same arguments for the term $V_{n21}(t)$ in Theorem 4.1, we have

$$\int_{\mathbb{R}^2d} |W_{n12}(t)|^2 \varphi(t) dt = O_p(pr_n^2) + O_p(n(pr_n)^4) + O_p(n(pr_n)^6).$$

Note that $n(p \log n)^3 r_n^4 \to 0$. Then it follows that

$$\int_{\mathbb{R}^2d} \frac{1}{\sqrt{nr_n}} W_{n1}(t)|^2 \varphi(t) dt = o_p(1).$$

For the second term $W_{n2}(t)$ in $\hat{V}_n^1(t)$, we have

$$W_{n2}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [g(\hat{\theta}, X_j) - g(\theta_0, X_j)] \cos(t^\top \mathfrak{B}_0^\top X_j) + \sin(t^\top \mathfrak{B}_0^\top X_j)$$

$$+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [g(\hat{\theta}, X_j) - g(\theta_0, X_j)] \cos(t^\top \mathfrak{B}^\top X_j) - \cos(t^\top \mathfrak{B}_0^\top X_j)$$

$$+ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [g(\hat{\theta}, X_j) - g(\theta_0, X_j)] \sin(t^\top \mathfrak{B}^\top X_j) - \sin(t^\top \mathfrak{B}_0^\top X_j)$$

$$=: W_{n21}(t) + W_{n22}(t) + W_{n23}(t)$$

For the term $W_{n21}(t)$, we have

$$W_{n21}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta} - \theta_0)^\top g'(\theta_0, X_j) \cos(t^\top \mathfrak{B}_0^\top X_j) + \sin(t^\top \mathfrak{B}_0^\top X_j)$$

$$+ \frac{1}{2\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta} - \theta_0)^\top g''(\theta_0, X_j)(\hat{\theta} - \theta_0) \cos(t^\top \mathfrak{B}_0^\top X_j) + \sin(t^\top \mathfrak{B}_0^\top X_j)$$

$$+ \frac{1}{6\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta} - \theta_0)^\top [(\hat{\theta} - \theta_0)^\top g'''(\theta_0, X_j)(\hat{\theta} - \theta_0)] \cos(t^\top \mathfrak{B}_0^\top X_j) + \sin(t^\top \mathfrak{B}_0^\top X_j)$$
where $\theta_3$ lies between $\hat{\theta}$ and $\theta_0$. Similar to the arguments for the term $V_{n3}(t)$ in the proof of Theorem 4.1, we obtain
\[
\int_{\mathbb{R}^{2d}} |W_{n212}(t)|^2 \varphi(t) dt = O_p(np^2 r_n^4) + O_p(pr_n)^4
\]
\[
\int_{\mathbb{R}^{2d}} |W_{n213}(t)|^2 \varphi(t) dt = O_p(np^6 r_n^6)
\]
For the term $W_{n211}(t)$, recall that $M(t) = \{ g'(\theta_0, X)[\cos(t^T \mathbf{W}_0^T X) + \sin(t^T \mathbf{W}_0^T X)] \}$. By the same arguments in $V_{n31}(t)$ in the proof of Theorem 4.1, we have
\[
W_{n211}(t) = \sqrt{n}(\hat{\theta} - \theta_0)^\top M(t) + R_{n211}(t) \quad \text{and} \quad \int_{\mathbb{R}^{2d}} |R_{n211}(t)|^2 \varphi(t) dt = O_p(pr_n)^2.
\]
By Theorem 4.3, it follows that
\[
\sqrt{n}M(t)^\top (\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j M(t)^\top \Sigma^{-1} g'(\theta_0, X_j) + \sqrt{n}r_n M(t)^\top \Sigma^{-1} E[g'(\theta_0, X)G(B^\top X)] + o_p(1),
\]
where $o_p(1)$ is uniformly in $t$. Note that $\int_{\mathbb{R}^{2d}} E[\varepsilon M(t)^\top \Sigma^{-1} g'(\theta_0, X)]^2 \varphi(t) dt = O(1)$. Then it follows that
\[
\int_{\mathbb{R}^{2d}} \frac{1}{\sqrt{n}r_n} W_{n211}(t)^2 \varphi(t) dt = \int_{\mathbb{R}^{2d}} |M(t)^\top \Sigma^{-1} E[g'(\theta_0, X)G(B^\top X)]|^2 \varphi(t) dt + o_p(1).
\]
Therefore we obtain
\[
\frac{1}{\sqrt{n}r_n} W_{n2}(t) = L_{n1}(t) + R_{n2}(t) \quad \text{and} \quad \int_{\mathbb{R}^{2d}} |R_{n2}(t)|^2 \varphi(t) dt = o_p(1).
\]
where $L_{n1}(t) = M(t)^\top \Sigma^{-1} E[g'(\theta_0, X)G(B^\top X)]$.

For the third term $W_{n3}(t)$ in $\tilde{V}_n^\top(t)$, we decompose it as follows,
\[
\frac{1}{\sqrt{n}r_n} W_{n3}(t) = \frac{1}{n} \sum_{j=1}^{n} G(B^\top X_j)[\cos(t^T \mathbf{W}_0^T X_j) + \sin(t^T \mathbf{W}_0^T X_j)]
\]
\[
+ \frac{1}{n} \sum_{j=1}^{n} G(B^\top X_j)[\cos(t^T \hat{\mathbf{W}}^\top X_j) - \cos(t^T \mathbf{W}_0^T X_j)]
\]
\[
+ \frac{1}{n} \sum_{j=1}^{n} G(B^\top X_j)[\sin(t^T \hat{\mathbf{W}}^\top X_j) - \sin(t^T \mathbf{W}_0^T X_j)]
\]
\[
=: \frac{1}{\sqrt{n}r_n} W_{n31}(t) + \frac{1}{\sqrt{n}r_n} W_{n32}(t) + \frac{1}{\sqrt{n}r_n} W_{n33}(t).
\]
Let $L_{n2}(t) = E\{G(B^\top X)[\cos(t^\top \mathbf{2}_0^\top X) + \sin(t^\top \mathbf{2}_0^\top X)]\}$. It is easy to see that

$$\int_{\mathbb{R}^d} \frac{1}{\sqrt{n}r_n} W_{n31}(t) - L_{n2}(t) \varphi(t) dt = O_p\left(\frac{1}{n}\right).$$

For the terms $W_{n32}(t)$ and $W_{n33}(t)$, by the Taylor expansion, we obtain

$$\int_{\mathbb{R}^d} \frac{1}{\sqrt{n}r_n} W_{n3k}(t)^2 \varphi(t) dt = O_p(pr_n)^2 + O_p(pr_n)^4, \text{ for } k = 2, 3.$$

Then it follows that

$$\int_{\mathbb{R}^d} \frac{1}{\sqrt{n}r_n} W_{n3}(t) = L_{n2}(t) + R_{n3}(t) \quad \text{and} \quad \int_{\mathbb{R}^d} |R_{n3}(t)|^2 \varphi(t) dt = o_p(1).$$

Altogether we obtain that

$$\int_{\mathbb{R}^d} \frac{1}{\sqrt{n}r_n} \hat{V}_n^1(t)^2 \varphi(t) dt = \int_{\mathbb{R}^d} |L_{n1}(t) + L_{n2}(t)|^2 \varphi(t) dt + o_p(1) \longrightarrow C_2 > 0.$$

(3) Following the same line as that in (2) with $r_n = 1/\sqrt{n}$, we have

$$\hat{V}_n^1(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j [\cos(t^\top \mathbf{2}_0^\top X_j) + \sin(t^\top \mathbf{2}_0^\top X_j) - M(t)^\top \Sigma^{-1} g(\theta_0, X_j)]$$

$$- L_{n1}(t) + L_{n2}(t) + R_n^1(t),$$

where $R_n^1(t)$ is a remainder with $\int_{\mathbb{R}^d} |R_n^1(t)|^2 \varphi(t) dt = o_p(1)$ and

$$L_{n1}(t) = M(t)^\top \Sigma^{-1} E[g'(\theta_0, X)G(B^\top X)]$$

$$L_{n2}(t) = E\{G(B^\top X)[\cos(t^\top \mathbf{2}_0^\top X) + \sin(t^\top \mathbf{2}_0^\top X)]\}.$$

The rest of proof is the same as that in the proof of step 2 and 3 of Theorem 4.1. Therefore we obtain

$$\int_{\mathbb{R}^d} |\hat{V}_n^1(t)|^2 \varphi(t) dt \rightarrow \int_{\mathbb{R}^d} |V_{\infty}^1(t) + L_1(t) - L_2(t)|^2 \varphi(t) dt,$$

where $V_{\infty}^1(t)$ is a zero-mean Gaussian process given in Theorem 4.1 and $L_1(t)$ and $L_2(t)$ are the uniformly limits of $L_{n1}(t)$ and $L_{n2}(t)$ respectively.

**Proof of Theorem 5.1.** We only prove the theorem under the null hypothesis. The arguments under the alternatives are similar and thus we omit it here. Let $P^*$ be the probability measure induced by the wild bootstrap resampling conditional on the original sample \{(X_i, Y_i) : i = 1, \ldots, n\} and let $E^*$ be the expectation under $P^*$. Recall that

$$\hat{V}_n^{1*}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [Y_j^* - g(\hat{\theta}^*, X_j)] [\cos(t^\top \mathbf{2}^\top X_j) + \sin(t^\top \mathbf{2}^\top X_j)].$$
Decompose $\hat{V}^{1*}_{n}(t)$ as follows,

$$
\hat{V}^{1*}_{n}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [\varepsilon_j^* + g(\hat{\theta}, X_j) - g(\hat{\theta}^*, X_j)] [\cos(t^T \mathbf{\hat{\theta}}^T X_j) + \sin(t^T \mathbf{\hat{\theta}}^T X_j)]
$$

$$
= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j^* [\cos(t^T \mathbf{\hat{\theta}}^T X_j) + \sin(t^T \mathbf{\hat{\theta}}^T X_j)] 
$$

$$
- \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [g(\hat{\theta}^*, X_j) - g(\hat{\theta}, X_j)] [\cos(t^T \mathbf{\hat{\theta}}^T X_j) + \sin(t^T \mathbf{\hat{\theta}}^T X_j)] 
$$

$$
=: \hat{V}^{1*}_{n1}(t) - \hat{V}^{1*}_{n2}(t)
$$

For the term $\hat{V}^{1*}_{n1}(t)$, we can decompose it as follows

$$
\hat{V}^{1*}_{n1}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j^* [\cos(t^T \mathbf{\hat{\theta}}^T_0 X_j) + \sin(t^T \mathbf{\hat{\theta}}^T_0 X_j)] + R^{1*}_{n1}(t).
$$

Note that $E^*(\varepsilon_j^*) = 0$ and $E^*(\varepsilon_j^*)^2 = \hat{\varepsilon}_j^2$. Then it follows that

$$
E^* \left( \int_{\mathbb{R}^d} |R^{1*}_{n1}(t)|^2 \varphi(t)dt \right)
$$

$$
= \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j^2 [\cos(t^T \mathbf{\hat{\theta}}^T_0 X_j) - \cos(t^T \mathbf{\hat{\theta}}^T_0 X_j) + \sin(t^T \mathbf{\hat{\theta}}^T_0 X_j) - \sin(t^T \mathbf{\hat{\theta}}^T_0 X_j)]^2
$$

$$
= o_p(1).
$$

Thus we obtain that $\int_{\mathbb{R}^d} |R^{1*}_{n1}(t)|^2 \varphi(t)dt = o_p(1)$.

For the term $\hat{V}^{1*}_{n2}(t)$, applying the Taylor expansion, we have

$$
\hat{V}^{1*}_{n2}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta}^* - \hat{\theta})^T g'(\hat{\theta}, X_j) [\cos(t^T \mathbf{\hat{\theta}}^T X_j) + \sin(t^T \mathbf{\hat{\theta}}^T X_j)]
$$

$$
+ \frac{1}{2\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta}^* - \hat{\theta})^T g''(\hat{\theta}, X_j)(\hat{\theta}^* - \hat{\theta}) [\cos(t^T \mathbf{\hat{\theta}}^T X_j) + \sin(t^T \mathbf{\hat{\theta}}^T X_j)]
$$

$$
+ \frac{1}{6\sqrt{n}} \sum_{j=1}^{n} (\hat{\theta}^* - \hat{\theta})^T [g''(\hat{\theta}, X_j)(\hat{\theta}^* - \hat{\theta})^T g'''(\hat{\theta}, X_j)(\hat{\theta}^* - \hat{\theta})] [\cos(t^T \mathbf{\hat{\theta}}^T X_j) + \sin(t^T \mathbf{\hat{\theta}}^T X_j)]
$$

$$
=: \hat{V}^{1*}_{n21}(t) + \hat{V}^{1*}_{n22}(t) + \hat{V}^{1*}_{n23}(t)
$$

where $\hat{\theta}_2$ lies between $\hat{\theta}^*$ and $\hat{\theta}$. Similar to the arguments for $V_{n32}(t)$ and $V_{n33}(t)$, we have in probability

$$
\int_{\mathbb{R}^d} |\hat{V}^{1*}_{n2k}(t)|^2 \varphi(t)dt = O_p \left( \frac{p^3}{n} \right) \quad \text{for } k = 2, 3.
$$

47
For the term $\hat{V}_{n21}^{1*}(t)$, we obtain
\[
\hat{V}_{n21}^{1*}(t) = \sqrt{n}(\hat{\theta}^* - \hat{\theta})^\top \frac{1}{n} \sum_{j=1}^{n} g'(\theta_0, X_j)[\cos(t^\top \mathcal{G}_0^\top X_j) + \sin(t^\top \mathcal{G}_0^\top X_j)] + R_{n211}^{1*}(t)
\]
\[
= \sqrt{n}(\hat{\theta}^* - \hat{\theta})^\top M(t) + R_{n211}^{1*}(t) + R_{n212}^{1*}(t),
\]
where $M(t) = E\{g'(\theta_0, X_j)[\cos(t^\top \mathcal{G}_0^\top X_j) + \sin(t^\top \mathcal{G}_0^\top X_j)]\}$ and the remainders $R_{n211}^{1*}(t)$ and $R_{n212}^{1*}(t)$ satisfy
\[
\int_{\mathbb{R}^d} |R_{n211}^{1*}(t)|^2 \varphi(t) dt = O_p^*(p^3/n) \quad \text{and} \quad \int_{\mathbb{R}^d} |R_{n212}^{1*}(t)|^2 \varphi(t) dt = O_p^*(p^2/n).
\]

Now we need the asymptotically linear expansion of $(\hat{\theta}^* - \hat{\theta})$. Following the arguments in Theorem 2.1 and 2.2, we obtain that
\[
\sqrt{n} \alpha^\top (\hat{\theta}^* - \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j^* \alpha^\top \Sigma_n(\hat{\theta})^{-1} g'(\hat{\theta}, X_j) + o_p^*(1),
\]
where $\Sigma_n(\hat{\theta}) = (1/n) \sum_{j=1}^{n} g'(\hat{\theta}, X_j)g'(\hat{\theta}, X_j)^\top$. Then it follows that
\[
\sqrt{n}M(t)^\top (\hat{\theta}^* - \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j^* M(t)^\top \Sigma_n(\hat{\theta})^{-1} g'(\hat{\theta}, X_j) + o_p^*(1).
\]

Consequently we obtain
\[
\hat{V}_{n2}^{1*}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j^* M(t)^\top \Sigma^{-1} g'(\theta_0, X_j) + R_{n2}^{1*}(t),
\]
where $\Sigma = E[g'(\theta_0, X)g'(\theta_0, X)^\top]$ and the remainder $R_{n2}^{1*}(t)$ satisfies
\[
\int_{\mathbb{R}^d} |R_{n2}^{1*}(t)|^2 \varphi(t) dt = o_p^*(1).
\]

Altogether we obtain that
\[
\hat{V}_n^{1*}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j^* [\cos(t^\top \mathcal{G}_0^\top X_j) + \sin(t^\top \mathcal{G}_0^\top X_j) - M(t)^\top \Sigma^{-1} g'(\theta_0, X_j)] + R_{n}^{1*}(t)
\]
\[
= V_n^{1*}(t) + R_{n}^{1*}(t).
\]
Here $R_{n}^{1*}(t) = R_{n1}^{1*}(t) + R_{n2}^{1*}(t)$ and then $\int_{\mathbb{R}^d} |R_{n}^{1*}(t)|^2 \varphi(t) dt = o_p^*(1)$.

It remains to derive the limiting distribution of $V_n^{1*}(t)$ under the probability measure $P^*$. Thus we need to deal with the covariance structure, convergence of the finite dimensional distributions of $V_n^{1*}(t)$ and the tightness under the induced bootstrap probability measure $P^*$. Similar
to the arguments in Lemma A.1-A.3 in Stute, González Manteiga and Presedo Quindimil (1998), we have in probability,

\[ \hat{V}_n^1(t) \longrightarrow V_1^\infty(t) \quad \text{in distribution in } C(\Gamma), \]

where \( \Gamma \) is a compact subset of \( \mathbb{R}^{2d} \) and \( V_1^\infty \) have the same distribution as the Gaussian process \( V_1 \) given in Theorem 4.1. The rest of proof follows the same line as that in the arguments of Theorem 4.1 with the expectation \( E \) replaced by the bootstrap expectation \( E^* \). Hence we complete the proof.

□

References

[1] Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics*, 20, 105-134.

[2] Bierens, H. J. and W. Ploberger (1997). Asymptotic theory of integrated conditional moment test. *Econometrica*, 65, 1129-1151.

[3] Billingsley, P. (1968) *Convergence of probability measures*. New York: Wiley.

[4] Cai, T. T., Zhang, C. H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38, 2118-2144.

[5] Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30, 455-474.

[6] Cook, R. D. and Weisberg, S. (1991). Discussion of Sliced inverse regression for dimension reduction, by K. C. Li. *Journal of the American Statistical Association*, 86, 316-342.

[7] Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimates. *The Annals of Statistics*. 27, 1012-1050.

[8] Escanciano, J. C. (2006a). A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22, 1030-1051.

[9] Escanciano, J. C. (2006b). Goodness-of-Fit tests for linear and nonlinear time series models. *Journal of the American Statistical Association*, 101, 531-541.

[10] Fan, J. Q. and Huang, L. S. (2001). Goodness-of-fit tests for parametric regression models, *Journal of the American Statistical Association*, 96, 640-652.
[11] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics, 32*, 928-961.

[12] Fang, K. T., Kotz, S., Ng, K. W. (1990) *Symmetric Multivariate and Related Distributions.* Springer-Science+Business Media, B.V.

[13] Guo, X., Wang, T. and Zhu, L. X. (2016). Model checking for generalized linear models: a dimension-reduction model-adaptive approach. *Journal of the Royal Statistical Society: Series B, 78*, 1013-1035.

[14] Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The annals of Statistics, 867*-889.

[15] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics, 21*, 1926-1947.

[16] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management, 5*, 81-102.

[17] Horowitz, J. L. and V. G. Spokoiny. (2001). An adaptive, rate-optimal test of a parametric mean- regression model against a nonparametric alternative. *Econometrica, 69*, 599-631.

[18] Huber, P. J. (1973) Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics, 799*-821.

[19] Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics, 40*, 633-643.

[20] Khmaladze, E V. and Koul, H. L. (2004). Martingale transforms goodness-of-fit tests in regression models. *The Annals of Statistics, 32*, 995-1034.

[21] Koul, H. L. and Ni, P. P. (2004). Minimum distance regression model checking. *Journal of Statistical Planning and Inference, 119*, 109-141.

[22] Li, K. C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association, 86*, 316-327.

[23] Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Steins lemma. *Journal of the American Statistical Association, 87*, 1025-1039.

[24] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association, 102*, 997-1008.
[25] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics, 21*, 255-285.

[26] Ortega, J. M., and Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*. Society for Industrial and Applied Mathematics.

[27] Portnoy S. (1984). Asymptotic behavior of M-estimators of p regression parameters when \( p^2/n \) is large. I. Consistency. *The Annals of Statistics, 1298-1309.*

[28] Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when \( p^2/n \) is large; II. Normal approximation. *The Annals of Statistics, 1403-1417.*

[29] Rigollet, P. and Hütter, J. C. (2017). *High dimensional statistics*. Unpublished Manuscript.

[30] Shi, C. C., Song, R., Chen, Z. and Li, R. Z. (2019). Linear hypothesis testing for high dimensional generalized linear models. *The Annals of Statistics, To appear*

[31] Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics. 25*, 613-641.

[32] Stute, W., Gonzales-Manteiga, W. and Presedo-Quindimil, M. (1998). Bootstrap approximation in model checks for regression. *Journal of the American Statistical Association. 93*, 141-149.

[33] Stute, W., Thies, S. and Zhu, L. X. (1998). Model checks for regression: An innovation process approach. *The Annals of Statistics. 26*, 1916-1934.

[34] Stute,W., Xu, W. L. and Zhu, L. X. (2008). Model diagnosis for parametric regression in high dimensional spaces. Biometrika. *95*, 1-17.

[35] Stute, W. and Zhu, L. X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics. 29*, 535-545.

[36] Tan, F. L. and Zhu, L. X. (2019a). An adaptive-to-model test for regressions with a diverging number of parameters. *The Annals of Statistics. To appear*

[37] Tan, F. L. and Zhu, L. X. (2019b). Supplement to Adaptive-to-model checking for regressions with diverging number of predictors. DOI:10.1214/18-AOS173SUPP.

[38] van Keilegom, I., González-Manteiga, W. and Sánchez Sellero, C. (2008). Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *TEST, 17*, 401-415.
[39] White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, **76**, 419-433.

[40] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.

[41] Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, **14**, 1261-1295.

[42] Xia, Y. C., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society: Series B*, **64**, 363-410.

[43] Xia, Q., Xu, W. L. and Zhu, L. X. (2015). Consistently determining the number of factors in multivariate volatility modelling. *Statistica Sinica*, **25**, 1025-1044.

[44] Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**, 1567-1594.

[45] Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, **75**, 263-289.

[46] Zhu, L. X. (2003). Model checking of dimension-reduction type for regression. *Statistica Sinica*, **13**, 283-296.

[47] Zhu, L. X., Miao, B. Q. and Peng, H. (2006). On Sliced Inverse Regression with High Dimensional Covariates. *Journal of the American Statistical Association*, **101**, 630-643.

[48] Zhu, L. P., Zhu, L. X. , Ferré, L. and Wang, T. (2010a). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, **97**, 295-304.

[49] Zhu, L. P., Zhu, L. X. and Feng, Z. H. (2010b). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, **105**, 1455-1466.

[50] Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, **37**, 17331751.
Table 1: Empirical sizes and powers of the tests for $H_0$ vs. $H_{11}$.

|         | a  | n=100 | n=200 | n=400 | n=600 |
|-------|----|-------|-------|-------|-------|
|       | p=8| p=12  | p=17  | p=20  |       |
| $AICM_n$ | 0.00 | 0.0580 | 0.0540 | 0.0410 | 0.0560 |
|        | 0.25 | 0.3670 | 0.6260 | 0.8710 | 0.9770 |
| $ACM_n$  | 0.00 | 0.0440 | 0.0495 | 0.0570 | 0.0510 |
|         | 0.25 | 0.2720 | 0.4960 | 0.8125 | 0.9335 |
| $PCVM_n$ | 0.00 | 0.0660 | 0.0530 | 0.0540 | 0.0650 |
|         | 0.25 | 0.3080 | 0.5140 | 0.8330 | 0.9350 |
| $T_{nS}^Z$ | 0.00 | 0.0475 | 0.0450 | 0.0450 | 0.0450 |
|         | 0.25 | 0.2665 | 0.4945 | 0.7970 | 0.9315 |
| $ICM_n$  | 0.00 | 0.0020 | 0.0000 | 0.0000 | 0.0000 |
|         | 0.25 | 0.0200 | 0.0000 | 0.0000 | 0.0000 |
| $T_{nGW}^Z$ | 0.00 | 0.0345 | 0.0580 | 0.0490 | 0.0605 |
|         | 0.25 | 0.2550 | 0.4435 | 0.7570 | 0.9100 |
| $T_{nZH}^Z$ | 0.00 | 0.0305 | 0.0210 | 0.0250 | 0.0285 |
|         | 0.25 | 0.0490 | 0.0370 | 0.0225 | 0.0255 |

Table 2: Empirical sizes and powers of the tests for $H_0$ vs. $H_{12}$.

|         | a  | n=100 | n=200 | n=400 | n=600 |
|-------|----|-------|-------|-------|-------|
|       | p=8| p=12  | p=17  | p=20  |       |
| $AICM_n$ | 0.00 | 0.0530 | 0.0630 | 0.0550 | 0.0510 |
|        | 0.50 | 0.5830 | 0.8970 | 0.9990 | 1.0000 |
| $ACM_n$  | 0.00 | 0.0435 | 0.0595 | 0.0495 | 0.0480 |
|         | 0.50 | 0.1465 | 0.3045 | 0.7255 | 0.9580 |
| $PCVM_n$ | 0.00 | 0.0790 | 0.0470 | 0.0500 | 0.0790 |
|         | 0.50 | 0.1650 | 0.2330 | 0.3860 | 0.5060 |
| $T_{nS}^Z$ | 0.00 | 0.0575 | 0.0475 | 0.0470 | 0.0545 |
|         | 0.50 | 0.1455 | 0.3390 | 0.7415 | 0.9585 |
| $ICM_n$  | 0.00 | 0.0010 | 0.0000 | 0.0000 | 0.0000 |
|         | 0.50 | 0.0250 | 0.0000 | 0.0000 | 0.0000 |
| $T_{nGW}^Z$ | 0.00 | 0.0410 | 0.0440 | 0.0455 | 0.0400 |
|         | 0.50 | 0.5645 | 0.9090 | 0.9965 | 1.0000 |
| $T_{nZH}^Z$ | 0.00 | 0.0385 | 0.0325 | 0.0200 | 0.0195 |
|         | 0.50 | 0.0715 | 0.0570 | 0.0420 | 0.0330 |
Table 3: Empirical sizes and powers of the tests for $H_0$ vs. $H_{13}$.

|           | $a$ | $n=100$ | $n=200$ | $n=400$ | $n=600$ |
|-----------|-----|---------|---------|---------|---------|
|           |     | $p=8$   | $p=12$  | $p=17$  | $p=20$  |
| $AICM_n$  | 0.00| 0.0510  | 0.0500  | 0.0500  | 0.0650  |
|           | 0.25| 0.5620  | 0.8530  | 0.9920  | 1.0000  |
| $ACM_n$   | 0.00| 0.0480  | 0.0555  | 0.0440  | 0.0495  |
|           | 0.25| 0.5830  | 0.8965  | 0.9955  | 1.0000  |
| $PCvM_n$  | 0.00| 0.0800  | 0.0570  | 0.0650  | 0.0690  |
|           | 0.25| 0.6090  | 0.9020  | 0.9960  | 1.0000  |
| $T_n^{SZ}$| 0.00| 0.0490  | 0.0525  | 0.0590  | 0.0495  |
|           | 0.25| 0.5820  | 0.8885  | 0.9960  | 1.0000  |
| $ICM_n$   | 0.00| 0.0020  | 0.0000  | 0.0000  | 0.0000  |
|           | 0.25| 0.0120  | 0.0000  | 0.0000  | 0.0000  |
| $T_n^{GWZ}$| 0.00| 0.0545  | 0.0550  | 0.0510  | 0.0550  |
|           | 0.25| 0.3735  | 0.6900  | 0.9500  | 0.9960  |
| $T_n^{ZH}$| 0.00| 0.0245  | 0.0290  | 0.0235  | 0.0180  |
|           | 0.25| 0.0525  | 0.0405  | 0.0270  | 0.0255  |

Table 4: Empirical sizes and powers of the tests for $H_0$ vs. $H_{14}$.

|           | $a$ | $n=100$ | $n=200$ | $n=400$ | $n=600$ |
|-----------|-----|---------|---------|---------|---------|
|           |     | $p=8$   | $p=12$  | $p=17$  | $p=20$  |
| $AICM_n$  | 0.0 | 0.0640  | 0.0520  | 0.0540  | 0.0500  |
|           | 0.1 | 0.3360  | 0.5920  | 0.8630  | 0.9670  |
| $ACM_n$   | 0.0 | 0.0505  | 0.0470  | 0.0475  | 0.0575  |
|           | 0.1 | 0.3490  | 0.5900  | 0.8975  | 0.9685  |
| $PCvM_n$  | 0.0 | 0.0470  | 0.0590  | 0.0560  | 0.0530  |
|           | 0.1 | 0.3600  | 0.6510  | 0.9010  | 0.9780  |
| $T_n^{SZ}$| 0.0 | 0.0515  | 0.0530  | 0.0510  | 0.0530  |
|           | 0.1 | 0.3325  | 0.5830  | 0.8860  | 0.9765  |
| $ICM_n$   | 0.0 | 0.0000  | 0.0000  | 0.0000  | 0.0000  |
|           | 0.1 | 0.0170  | 0.0000  | 0.0000  | 0.0000  |
| $T_n^{GWZ}$| 0.0 | 0.0465  | 0.0565  | 0.0560  | 0.0515  |
|           | 0.1 | 0.1955  | 0.3720  | 0.6610  | 0.8395  |
| $T_n^{ZH}$| 0.0 | 0.0280  | 0.0225  | 0.0170  | 0.0200  |
|           | 0.1 | 0.0335  | 0.0285  | 0.0290  | 0.0175  |
Table 5: Empirical sizes and powers of the tests $AICM_n$, $PCvM_n$, $ICM_n$, and $T_{ZH}^n$ for $H_{21}$.

| a  | n=100 | n=200 | n=400 | n=600 |
|----|-------|-------|-------|-------|
|    | p=8   | p=12  | p=17  | p=20  |
| $AICM_n$ | 0.0  | 0.0650 | 0.0640 | 0.0580 | 0.0650 |
|      | 0.5  | 0.2850 | 0.6390 | 0.9780 | 1.0000 |
| $PCvM_n$ | 0.0  | 0.0680 | 0.0750 | 0.0680 | 0.0760 |
|      | 0.5  | 0.2390 | 0.5030 | 0.8020 | 0.9820 |
| $ICM_n$  | 0.0  | 0.0170 | 0.0000 | 0.0000 | 0.0000 |
|        | 0.5  | 0.0460 | 0.0000 | 0.0000 | 0.0000 |
| $T_{ZH}^n$ | 0.0  | 0.0215 | 0.0220 | 0.0270 | 0.0170 |
|        | 0.5  | 0.0405 | 0.0335 | 0.0325 | 0.0345 |

Table 6: Empirical sizes and powers of the tests $AICM_n$, $PCvM_n$, $ICM_n$, and $T_{ZH}^n$ for $H_{22}$.

| a  | n=100 | n=200 | n=400 | n=600 |
|----|-------|-------|-------|-------|
|    | p=8   | p=12  | p=17  | p=20  |
| $AICM_n$ | 0.0  | 0.0700 | 0.0790 | 0.0560 | 0.0720 |
|      | 0.5  | 0.4030 | 0.7780 | 0.9880 | 1.0000 |
| $PCvM_n$ | 0.0  | 0.0880 | 0.0770 | 0.0610 | 0.0500 |
|      | 0.5  | 0.3080 | 0.5420 | 0.8100 | 0.9110 |
| $ICM_n$  | 0.0  | 0.0170 | 0.0000 | 0.0000 | 0.0000 |
|        | 0.5  | 0.0880 | 0.0000 | 0.0000 | 0.0000 |
| $T_{ZH}^n$ | 0.0  | 0.0230 | 0.0230 | 0.0240 | 0.0220 |
|        | 0.5  | 0.0565 | 0.0500 | 0.0525 | 0.0335 |

Table 7: Empirical sizes and powers of the tests $AICM_n$, $PCvM_n$, $ICM_n$, and $T_{ZH}^n$ for $H_{23}$.

| a  | n=100 | n=200 | n=400 | n=600 |
|----|-------|-------|-------|-------|
|    | p=8   | p=12  | p=17  | p=20  |
| $AICM_n$ | 0.00 | 0.0790 | 0.0740 | 0.0700 | 0.0610 |
|      | 0.75 | 0.3060 | 0.5250 | 0.8390 | 0.9680 |
| $PCvM_n$ | 0.00 | 0.0690 | 0.0730 | 0.0660 | 0.0630 |
|      | 0.75 | 0.2170 | 0.2820 | 0.3320 | 0.3730 |
| $ICM_n$  | 0.00 | 0.0090 | 0.0000 | 0.0000 | 0.0000 |
|        | 0.75 | 0.0230 | 0.0000 | 0.0000 | 0.0000 |
| $T_{ZH}^n$ | 0.00 | 0.0270 | 0.0270 | 0.0225 | 0.0225 |
|        | 0.75 | 0.0380 | 0.0365 | 0.0375 | 0.0280 |
Figure 2: Scatter plot of $Y$ versus the projected covariate $\hat{\beta}_1^T X$ in which the direction $\hat{\beta}_1$ is obtained by CSE.

Figure 3: Scatter plot of Residuals from the linear model for $(Y, X)$ versus the projected covariate $\hat{\beta}_1$ in which the direction $\hat{\beta}_1^T X$ is obtained by CSE.
Figure 4: Scatter plot of $Y$ versus the projected covariates $(\hat{\beta}_1^T X, \hat{\beta}_2^T X)$ in which the directions $(\hat{\beta}_1, \hat{\beta}_2)$ are obtained by CSE.

Figure 5: Scatter plot of Residuals from model (6.1) versus the fitted values $\hat{\beta}_n^T X$. 

57