Test for High Dimensionality of Random and Estimated Vectors

Abstract

Although the fast advancing computing techniques support rapid increase of the dimension of data or number of parameters, the corresponding statistical inference methods are few in the literature. The main reason for the failure of many classical inference methods is that the asymptotic properties of random vectors with infinite sample size may not hold when its dimension towards infinity and proportionally approaches the sample size. It has attracted many scholars to seek proper methods under this case we called "high dimensional settings", since 1958. Although these studies have provided some modified methods under different asymptotic relationships between the sample size and dimension, a more basic question about how to determine the dimension of a given data sample is high or low has not been considered. Therefore, this paper proposes a general test to distinguish high dimensional and classical settings for both random and estimated vectors. Results from a simulation study suggest that our test can work very well.

1 Introduction

Prior to 1950, most of practical problems consisted of a relatively large number of experimental units with a relatively small number of features which were measured (Rowell, 1976). Therefore, traditional theories and practice were limited to the "small dimension of variables and large sample size" scenario. This scenario naturally reflected the contemporary limitations of computers and graphical display. Over the last 25 years, however, Lindsay (2004) pointed out that the environment for practical problems has changed dramatically, with the huge evolution of data acquisition technologies and computing facilities. The main scenarios to be investigated steadily evolve into the "large dimension and small sample size," or in some cases "large dimension and large sample size." With the latest development of computing techniques, such as the neural network, this allows for data with much larger dimensions to be dealt with. Most of the latest large language models have contained more than 100 billions trainable parameters in Zhao (2023). Not only do the techniques develop rapidly, but the high-dimensional theories also advance dramatically. Theoretical studies have focused on two aspects: modify classical theories and investigate new theories.

Modification of the classical theories started early, since many classical methods will fail if the dimension is sufficiently large compared to the sample size. This type of failure was firstly noticed by Dempster (1958), who showed that the classical Hotelling's T-squared test became undefined as the number of variables became close to, or even exceeded the number of degrees of freedom within sample for estimation of the variance and co-variance matrix. A related simulation, which showed the failure of the Hotelling's T-squared test, was also proposed by Bai & Saranadasa (1996). On the other hand, some classical estimators and their properties have changed and need to be re-built under the high-dimensional scenario. The properties of M estimators in the linear regressions have been studied by Huber (1973), Portnoy (1984) and Portnoy (1991), while the properties of estimators in the non-linear regressions are focused by Portnoy (1988) and He & Shao (2000). Additionally, the classical F-test and likelihood-ratio test, which will fail if the dimension is large compared to the sample size, are corrected by Calhoun (2011) and Sur & Candès (2019), respectively.

On the other hand, an increasing number of novel and useful properties are found under the high-dimensional scenario. For instance, Jimenez & Landgrebe (1998) showed that, under the high-dimensional condition, the estimators in Lasso+mLS and Lasso+Ridge are asymptotically normal and nonzero parameters have the same asymptotic normal distribution when the zero parameters were known. Also, Chi et al. (2022) proved that the discontinuous regression is allowed in random forests algorithm if the dimension of data is high. The corresponding bias is bounded, which only correlates with the tree height and column subsampling parameter when the sample size is large enough. Additionally, Vershynin (2018) provided many practical theories for random vectors and matrices, such as estimating concentration of the norm, approximating isometries, etc. These theories only work when the number of coordinates of random vectors and the entries of random matrices are sufficiently large. Specifically, when the dimension of random vectors or matrices grows increasingly, some good properties start to appear. Overall, the "high dimensions" seems no longer just a "curse" and can even be utilized, although the "curse of dimensionality" was pointed out by Bellman (1957).

The idea behind high-dimensional theories is similar to that behind the central limit theorem (CLT) in which the normal distribution appears in the case of a sufficiently large sample size. In other words, under certain assumptions, statistical and probabilistic methodology that works for normal distributions can also be applicable to problems involving other types of distributions as the sample size approaches infinity. Such applicability sometimes collapses, given a finite sample in practice. In general, sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold. However, Lehmann (1999) showed the distribution of the arithmetic mean of independent random variables from a binomial distribution $B(N, p)$ is still not a satisfactory approximation of a normal distribution even when N is larger than 90, where the parameter p is equal to 0.05 and N is the sample size. It is also stated that the speed of convergence is dependent on the underlying distribution of the sample. Therefore, the general judgment criteria for an adequate sample size is not always reliable, which constitutes a major concern for the failure of CLT. Determining how large the sample size N is adequate to hold the CLT is essential in the observed finite sample. Similarly, one might ask how large the dimensions need to be for these high-dimensional theories to hold, which is also essential. This is what we aims to answer. The "high dimensions", without any clear definition in the literature, is defined, in this paper, as the scenario in which the high-dimensional theories hold or the high-dimensional properties appear.

Compared to the finite or low-dimensional scenario, the high-dimensional scenario becomes more common, which we are facing in reality but we usually do not realize it. Although many methods have been proposed for high dimensions; however, a more basic question about whether a given data sample is in high dimensions or not, has not been considered in the literature. In other words, there is no method which determines whether the high-dimensional theories can be applied for a given data sample. Like the threshold for the sample size in CLT, discussing a global threshold between the high and non-high dimensions is essential. Therefore, this paper provides a general testing method to distinguish high from non-high dimensions for random vectors. We classify the random vectors into two categories: vectors not being estimated (random vectors) and vectors being estimated (estimated vectors) for convenience. The null hypothesis in our paper is defined as

$$H_0 : \text{The dimension of the settings is high.}$$

The "settings" represent the random or estimated vectors for which one aims to test. Failure to reject the null hypothesis indicates that the "settings" are in high dimensions; and the high-dimensional theories can be applied. If the null hypothesis was rejected, the "settings" fall into the undefined non-high dimensions in which the availability of both classical and high-dimensional theories is unknown. Finally, we provide guidance to determine the threshold of high-dimensional settings based on a Monte Carlo study, which shows the performance of our proposed test.

In section 2, the test for high dimensionality of random vectors is presented. Section 3 & 4 show the test for estimated vectors especially from linear and non linear regression models, respectively. A Monte Carlo study is presented in section 5. Finally, section 6 concludes this paper.

Notation. As mentioned in section 1, N and D represent the sample size and the dimension of random and estimated vectors. D is allowed to approach N proportionally, but is less than N. In the test of random vectors, we work on the random vectors $\{\mathbf{V}_n\}_{n=1}^N$ whose realization is $\{\mathbf{v}_n\}_{n=1}^N$. In the case of estimated vectors, Y_n is the dependent variable while

$\mathbf{X}_n$ is the independent random vector for $n = 1, 2, ..., N$. y_n and $\mathbf{x}_n$ are the corresponding realizations respectively. Besides, the ℓ_2-norm is denoted by $\|.\|$. For a vector-valued function $f(\beta)$, define $\nabla_{f_n(\beta_0)} := \frac{\partial}{\partial\beta} f_n(\beta)\big|_{\beta=\beta_0}$ and $\mathbf{H}_{f_n(\beta_0)} := \frac{\partial^2}{\partial\beta\partial\beta^\top} f_n(\beta)\big|_{\beta=\beta_0}$.

2 Test of Random Vectors

In this section, we first present the test statistic for multivariate normal random vectors and its asymptotic property, and then extend the results to general random vectors. At the end of this section, we introduce how to apply the proposed test.

2.1 Multivariate Normal Vectors

2.1.1 Identity Covariance Matrix

Suppose there is a sequence of i.i.d. (independent and identically distributed) D-dimension random vectors $\{\mathbf{V}_n\}_{n=1}^N$ such that $\mathbf{V}_n \sim \mathcal{N}(0, I_{D\times D})$ for each n. Our test statistic is defined as the following:

$$T_1 := \sqrt{\frac{D}{2}} \left(\frac{1}{D} \left\| \sqrt{N} \left(\frac{1}{N} \sum_{n=1}^N \mathbf{V}_n \right) \right\|^2 - 1 \right).$$

It is easy to show that $T_1 \xrightarrow{d} \mathcal{N}(0,1)$. Rewrite $\left\| \sqrt{N}(\frac{1}{N}\sum_{n=1}^N \mathbf{V}_n) \right\|^2$ as $\sum_{d=1}^D \left(\frac{1}{\sqrt{N}} \sum_{n=1}^N \mathbf{V}_{n,d} \right)^2$ where $\mathbf{V}_{n,d}$ is the dth element of the vector $\mathbf{V}_n$. Since we know that $\frac{1}{\sqrt{N}}\sum_{n=1}^N \mathbf{V}_{n,d} \sim \mathcal{N}(0,1)$ is independent of N, and $\frac{1}{\sqrt{N}}\sum_{n=1}^N \mathbf{V}_{n,d}$ is independent w.r.t. index d, it implies that the central limit theorem gives us

$$T_1 \xrightarrow{d} \mathcal{N}(0,1).$$

2.1.2 Non-Identity Matrix

Now, the covariance matrix of $\mathbf{V}_n$ is Ω_D, i.e., $\mathbf{V}_n \sim \mathcal{N}(0, \Omega_D)$ for each n. Standardize $\mathbf{V}_n$ by

$$\tilde{\mathbf{V}}_n := \Omega_D^{-\frac{1}{2}} V_n \sim \mathcal{N}(0, I_{D\times D}).$$

This is identical to the case in the section 2.1.1. As a result, there is

$$T_2 := \sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N}\Omega_D^{-\frac{1}{2}}\mathbf{V}_n\right)\right\|^2 - 1\right) \xrightarrow{d} \mathcal{N}(0,1).$$

2.2 General Vectors

In this section, a more general case without normality assumption is considered. Let $\{\mathbf{V}_n\}_{n=1}^N$ be a sequence of independent random vectors in R^D, such that for $n = 1, 2, ..., N$,

1. $\mathbb{E}\left[\mathbf{V}_n\right] = \alpha,$
2. $\mathbb{E}\left[\left(\mathbf{V}_n - \alpha\right)\left(\mathbf{V}_n - \alpha\right)^\top\right] = \Omega_D.$

2.2.1 α is known

If α is known, vectors can be easily centralized. Define the normalized sum with centralization as

$$S_N^V := \frac{1}{\sqrt{N}}\sum_{n=1}^{N}(\mathbf{V}_n - \alpha)$$

The high dimensional central limit theorem from Chernozhukov (2017) will be applied, since the dimension D is no longer constant. As a result of it, the following assumptions are required.

There are constant b, a sequence of constants $B_N \geq 1$ and a covariance estimator $\hat{\Omega}_{N,D}$, which satisfy the following conditions:

(a) $\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left(\mathbf{V}_{n,d} - \alpha_d\right)^2\right] \geq b$ for all $d = 1, 2, ..., D,$

(b) $\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\left|\mathbf{V}_{n,d} - \alpha_d\right|^{2+k}\right] \leq B_N^k$ for all $d = 1, 2, ..., D$ and $k = 1, 2,$

(c) $\mathbb{E}\left[exp\left(\left|\mathbf{V}_{n,d} - \alpha_d\right|/B_N\right)\right] \leq 2$ for all $d = 1, 2, ..., D$ and $n = 1, 2, ..., N,$

(d) $\left(\frac{B_N^2 \log^7(DN)}{N}\right) = o_p(1),$

(e) $\left\|\hat{\Omega}_{N,D}^{-\frac{1}{2}} - \Omega_D^{-\frac{1}{2}}\right\|^2 = o_p\left(\frac{1}{\sqrt{D}}\right).$

Based on the central limit theorem from Chernozhukov (2017),

$$\sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\hat{\Omega}_{N,D}^{-\frac{1}{2}}S_N^V\right\|^2-1\right)\xrightarrow{d}\mathcal{N}\left(0,1\right).$$

Proof: Suppose $\mathbf{W}_1,\mathbf{W}_2,...,\mathbf{W}_N$ are independent centered normal random vectors in R^D, such that each $\mathbf{W}_n$ has the same covariance matrix as $\mathbf{V}_n$, that is,

$$\mathbf{W}_n\sim\mathcal{N}\left(0,\Omega_D\right)$$

with the normalized sum $S_N^W:=\frac{1}{\sqrt{N}}\sum_{n=1}^N\mathbf{W}_n$. Under the conditions (a)-(d), it can be shown that S_N^V converges to S_N^W in distribution. As a result of $\Omega_D^{-\frac{1}{2}}\mathbf{W}_n\sim\mathcal{N}\left(0,I_{D\times D}\right)$, section 2.1 gives us

$$\sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\frac{1}{\sqrt{N}}\sum_{n=1}^N\Omega_D^{-\frac{1}{2}}\mathbf{W}_n\right\|^2-1\right)\xrightarrow{d}\mathcal{N}\left(0,1\right).$$

Based on the continuous mapping theorem and condition (e), there is

$$T_3:=\sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\hat{\Omega}_{N,D}^{-\frac{1}{2}}S_N^V\right\|_2^2-1\right)\xrightarrow{d}\mathcal{N}\left(0,1\right).$$

□

2.2.2 α is unknown - multiplier bootstrap

Once α is unknown, we need additional multipliers for normalization. Suppose $\{e_n\}_{n=1}^N$ is a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables which are independent of $\{\mathbf{V}_n\}_{n=1}^N$. Define the different normalized sum from the one in section 2.2.1 as

$$S_N^{eV}:=\frac{1}{\sqrt{N}}\sum_{n=1}^N e_n(\mathbf{V}_n-\bar{\mathbf{V}}_N)$$

where $\bar{\mathbf{V}}_N=\frac{1}{N}\sum_{n=1}^N\mathbf{V}_n$. Under conditions (a)-(d) in section 2.2.1, *theorem* 4.1 from Chernozhukov (2017) implies that S_N^{eV} converges in distribution to S_N^W. Based on the similar arguments in section 2.2.1,

$$T_4:=\sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\hat{\Omega}_{N,D}^{-\frac{1}{2}}S_N^{eV}\right\|^2-1\right)\xrightarrow{d}\mathcal{N}\left(0,1\right).$$

Remark: There are four different test statistics introduced in this section for different cases of random vectors respectively, but they are foundationally similar. The only difference is the standardization. We introduce how to use the test for the most general case in section 2.2.2. Suppose one observes a realization $\{\mathbf{v}_n\}_{n=1}^N$ of $\{\mathbf{V}_n\}_{n=1}^N$ and aims to check whether this data is under high dimensional settings or not.

(i) Generate a sample $e_1, e_2, ..., e_N$ from $\mathcal{N}(0,1)$.

(ii) Calculate the estimates of mean $\bar{\mathbf{v}}_N$ and covariance matrix $\hat{\Omega}$ from $\{\mathbf{v}_n\}_{n=1}^N$.

(iii) Compute the test statistic $t_4 = \sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\hat{\Omega}^{-\frac{1}{2}}\frac{1}{\sqrt{N}}\sum_{n=1}^N e_n(\mathbf{v}_n - \bar{\mathbf{v}}_N)\right\|^2 - 1\right)$.

If $|t_4| > 1.96$ (at 5% level of significance), we reject the null hypothesis, which means this data sample is under classical settings. Therefore, the classical inference methods, such as classical Hotelling's T-squared test, can be applied. However, once $|t_4| \leq 1.96$, this data sample is under high dimensional settings. One should apply the high dimensional versions of the classical methods.

3 Test for Estimated Vectors from Linear Regressions

In this section, the test for high dimensionality of estimated vectors from the linear regression model is introduced. Firstly, we define the linear regression model

$$\mathrm{Y}_n = \mathbf{X}_n^\top \beta_0 + \epsilon_n, n = 1, 2, ..., N$$

where $\{\mathbf{X}_n, \mathrm{Y}_n\}_{n=1}^N$ is a sequence of independent vectors with length $D+1$, $\{\epsilon_n\}_{n=1}^N$ is a sequence of i.i.d. random variables with zero means and a constant variance, and β_0 is an unknown $D \times 1$ vector. The traditional linear model has the following assumptions:

1. $\{\mathbf{X}_n, \mathrm{Y}_n\}_{n=1}^N$ are i.i.d.,
2. $\mathbf{\Sigma} = \mathbb{E}[\mathbf{X}_n\mathbf{X}_n^\top]$ exists and $\|\mathbf{\Sigma}\|$ is bounded,
3. $\mathbb{E}[\epsilon_n|\mathbf{X}_n] = 0$,
4. $\mathbb{E}[\epsilon_n^2|\mathbf{X}_n] = \sigma^2$.

Define $\mathbb{X}$ as an $N \times D$ matrix with $\mathbf{X}_n^\top$ as each row. Additionally, we require the following assumptions.

Assumption 3.1 There is a positive constant B, such that for all n,

$$\frac{\mathbf{X}_n^\top \mathbf{X}_n}{D} \leq B.$$

Assumption 3.2 There are positive constants b and c, such that the minimum and maximum eigenvalues of $\frac{\mathbb{X}^\top \mathbb{X}}{N}$ satisfy

$$\lambda_{min}\left(\frac{\mathbb{X}^\top \mathbb{X}}{N}\right) \geq b,\ \lambda_{max}\left(\frac{\mathbb{X}^\top \mathbb{X}}{N}\right) \leq c.$$

Assumption 3.3 (Conditions for high dimensional CLT) Suppose there are constant e and a sequence of constants $B_N \geq 1$ such that

$$\begin{aligned}
&(a)\ \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left[(\mathbf{X}_n \epsilon_n)_d^2\right] \geq e && \text{for all } d = 1, 2, ..., D,\\
&(b)\ \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left[(\mathbf{X}_n \epsilon_n)_d^{2+k}\right] \leq B_N^k && \text{for all } d = 1, 2, ..., D \text{ and } k = 1, 2,\\
&(c)\ \mathbb{E}\left[\exp\left(\left|\left(\mathbf{X}_n \epsilon_n\right)_d\right| / B_N\right)\right] \leq 2 && \text{for all } d = 1, 2, .., D \text{ and } n = 1, 2, ..., N,\\
&(d)\ \left(\frac{B_N^2 \log^7(DN)}{N}\right) = o_p(1).
\end{aligned}$$

Assumption 3.4

$$\left\| \left(\frac{\mathbb{X}^\top \mathbb{X}}{N}\right)^{-1/2} - \mathbf{\Sigma}^{-1/2} \right\| = o_p\left(\frac{1}{\sqrt{D}}\right).$$

Before the test statistic, we need to ensure that the OLS (ordinary least square) estimator $\hat{\beta}$ we applied in this section is valid. Define

$$\hat{\beta} = \arg\min_{\beta} \sum_{n=1}^{N} \left(\mathrm{Y}_n - \mathbf{X}_n^\top \beta\right)^2.$$

Lemma 3.1. (Consistency) Under *Assumptions* 3.1 & 3.2, $\|\hat{\beta}$ - $\beta_0\|^2 = O_p(\frac{D}{N})$.

Remark: OLS estimator is just one method we used for convenience. There have been other novel estimation methods which could be valid under high dimensional settings and may provide a faster rate of convergence. This paper focuses on the idea of testing the high dimesionality, so the attempt of other estimators can be an extension of our test.

Lemma 3.2. Under *Assumptions* 3.1 - 3.4, we have

$$\sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\sigma^{-1}\left(N^{-1}\mathbb{X}^{\top}\mathbb{X}\right)^{1/2}\sqrt{N}\left(\hat{\beta}-\beta_0\right)\right\|_2^2-1\right)\xrightarrow{d}\mathcal{N}\left(0,1\right).$$

3.1 Test Statistic

Lemma 3.2 could be an idea of calculating the test statistic for high dimensions; however, in practice, the true value of β is unknown, so we use the average of the bootstrapped estimates replacing the true value to obtain a proper statistic. The bootstrapping method is implemented as follows:

(i) Estimate $\hat{\beta}$ (OLS) and $\hat{\epsilon}_n$ where $\hat{\epsilon}_n := \mathrm{Y}_n - \mathbf{X}_n^{\top}\hat{\beta}$ for $n = 1, 2, ..., N$.

(ii) Independent of the observed data $D_N := \{(\mathrm{Y}_n, \mathbf{X}_n) : 1 \leq n \leq N\}$, use a random number generator to generate $R_1, R_2, ..., R_N \overset{iid}{\sim}$ Rademacher. Define $\mathrm{Y}_n^* := \mathbf{X}_n^{\top}\hat{\beta} + R_n\hat{\epsilon}_n$ for $n = 1, 2, ..., N$.

(iii) Re-estimate β using the data $\{(\mathrm{Y}_n^*, \mathbf{X}_n) : 1 \leq n \leq N\}$. Denote it by β^*.

(iv) Repeat steps (ii)-(iii) B times. We calculate the following and denote it as t_k^*

$$t_k^* := \sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\left[\left(\mathbb{X}^{\top}\mathbb{X}\right)^{-1}\left(\sum_{n=1}^{N}\hat{\epsilon}_n^2\mathbf{X}_n\mathbf{X}_n^{\top}\right)\left(\mathbb{X}^{\top}\mathbb{X}\right)^{-1}\right]^{-1/2}\sqrt{B}\left(\frac{1}{B}\sum_{b=1}^{B}\beta_b^*-\hat{\beta}\right)\right\|_2^2-1\right).$$

(v) Repeating steps (ii)-(iv) K times produces a sequence $\{t_k^*\}_{k=1}^K$. The test statistic is

$$T_{N,D,K,B}^* := K^{\frac{1}{2}}\sup_{z\in R}\left|K^{-1}\sum_{k=1}^{K}\mathbf{1}_{(-\infty,z]}(t_k^*)-\Phi(z)\right| \tag{1}$$

where $\Phi(.)$ is the CDF of the standard normal distribution.

Theorem 3.1 Under *Assumptions* 3.1 - 3.3, after setting B, K such that $\frac{K}{D} = o_p(1)$ and $\frac{\sqrt{K}D^{\frac{7}{4}}}{\sqrt{B}} = o_p(1)$, there is

$$T_{N,D,K,B}^* \xrightarrow{d} \sup_{z\in R}\left|B\left(\Phi\left(z\right)\right)\right|$$

where $B(.)$ is the Brownian bridge.

Remark: For one realization $\{\mathbf{x}_n, \mathrm{y}_n\}_{n=1}^N$ of $\{\mathbf{X}_n, \mathrm{Y}_n\}_{n=1}^N$, in order to check whether this data is under high dimensional settings or not, one can follow the above bootstrapping method and obtain the test statistic. If the statistic does pass the Kolmogorov–Smirnov test of standard normality, then the null hypothesis is not rejected. Therefore, any classical inference methods for this realization $\{\mathbf{x}_n, \mathrm{y}_n\}_{n=1}^N$ could be undefined. The high dimensional versions of classical methods should be considered.

4 Test for Estimated Vectors from Non Linear Regressions

This section extends the result of linear regression to non linear case. The non linear model is defined as

$$\mathrm{Y}_n = f(\mathbf{X}_n, \beta) + \epsilon_n, \ n = 1, 2, ..., N$$

where $\{\mathrm{Y}_n, \mathbf{X}_n\}_{n=1}^N$ is a sequence of independent vectors with length $J+1$, $\{\epsilon_n\}_{n=1}^N$ is a sequence of i.i.d. random variables with zero means and a constant variance, β is a $D \times 1$ vector of unknown parameters, and f is a known non linear function. In general, we assume

1. $\{\mathrm{Y}_n, \mathbf{X}_n\}_{n=1}^N$ are independent,
2. $\mathbf{\Sigma} = \mathbb{E}[\nabla_{f_n(\beta_0)} \nabla_{f_n(\beta_0)}^\top]$ has a bounded norm,
3. $\mathbb{E}[\epsilon_n | \mathbf{X}_n] = 0$,
4. $\mathbb{E}[\epsilon_n^2 | \mathbf{X}_n] = \sigma^2$.

Additionally, the following assumptions are required.

Assumption 4.1 There are constants B and δ such that

$$\sum_{d=1}^{D} \sum_{n=1}^{N} \left(\nabla_{f_{n,d}(\beta)}\right)^2 \leq BDN \text{ for all } \beta \text{ with } \|\beta - \beta_0\| \leq \delta.$$

Assumption 4.2 There is a constant δ_2 such that all eigenvalues of

$$\frac{1}{N} \sum_{n=1}^{N} \nabla_{f_{n,d}(\tilde{\beta})} \nabla_{f_{n,d}(\beta)}^\top$$

are non-negative for all β with $\|\beta - \beta_0\| \leq \delta_2$ and $\tilde{\beta}$, between β_0 and β.

Assumption 4.3 Suppose that there are constants b and a and a sequence of constants $B_N \geq 1$ such that

$$
\begin{aligned}
&(a)\ \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[t_{n,d}^{2}\right] \geq b && \text{for all } d = 1, 2, ..., D,\\
&(b)\ \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[t_{n,d}^{2+k}\right] \leq B_N^k && \text{for all } d = 1, 2, ..., D \text{ and } k = 1, 2,\\
&(c)\ \mathbb{E}\left[\exp\left(\frac{|t_{n,d}|}{B_N}\right)\right] \leq 2 && \text{for all } d = 1, 2, .., D \text{ and } n = 1, 2, ..., N,\\
&(d)\ \left(\frac{B_N^2 \log^7(DN)}{N}\right) \to 0,
\end{aligned}
$$

where $t_{n,d} = (\sigma^{-1}\Omega^{-\frac{1}{2}}\epsilon_n \nabla_{f_n(\beta_0)})_{(d)}$.

Assumption 4.4 $\frac{1}{N}\sum_{n=1}^{N}\nabla_{f_n(\beta)}\nabla_{f_n(\beta)}^{\top}$ converges uniformly in β in an open neighborhood of β_0 with respect to the L^2 norm, which is

$$
\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla_{f_n(\beta)}\nabla_{f_n(\beta)}^{\top} - \frac{1}{N}\sum_{n=1}^{N}\nabla_{f_n(\beta_0)}\nabla_{f_n(\beta_0)}^{\top}\right\| = o_p\left(\frac{1}{D^{1/2}}\right)
$$

for any β with $\|\beta - \beta_0\|^2 = O_p\left(\frac{D}{N}\right)$.

Assumption 4.5 For any β with $\|\beta - \beta_0\| = O_p\left(\frac{D}{N}\right)$, we have

$$
\begin{aligned}
\left\|\frac{1}{N}\sum_{n=1}^{N}\left[f_n(\beta_0) - f_n(\beta)\right]\mathbf{H}_{f_n(\beta)}\right\| &= o_p\left(\frac{1}{D^{1/2}}\right),\\
\left\|\frac{1}{N}\sum_{n=1}^{N}\epsilon_n \mathbf{H}_{f_n(\beta)}\right\| &= o_p\left(\frac{1}{D^{1/2}}\right).
\end{aligned}
$$

Assumption 4.6

$$
\begin{aligned}
\left\|\frac{1}{N}\sum_{n=1}^{N}\nabla_{f_n(\beta_0)}\nabla_{f_n(\beta_0)}^{\top} - \mathbf{\Sigma}\right\| &= o_p\left(\frac{1}{D^{1/2}}\right),\\
\left\|\left(\frac{1}{N}\sum_{n=1}^{N}\nabla_{f_n(\beta_0)}\nabla_{f_n(\beta_0)}^{\top}\right)^{1/2} - \mathbf{\Sigma}^{1/2}\right\| &= o_p\left(\frac{1}{D^{1/2}}\right)
\end{aligned}
$$

for any β such that $\|\beta - \beta_0\|^2 = O_p(\frac{D}{N})$.

Similar to *Lemma* 3.1, we need the consistency of the least square estimator,

$$
\hat{\beta} = \arg\min_{\beta}\sum_{n=1}^{N}\left(\mathrm{Y}_n - f(\mathbf{X}_n, \beta)\right)^2.
$$

Lemma 4.1 (Consistency) Under *Assumptions* 4.1 & 4.2, $\|\hat{\beta}-\beta\|^2 = O_p(\frac{D}{N})$.

Lemma 4.2 Under *Assumptions* 4.1 - 4.6, we have

$$\sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\sigma^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\nabla_{f_n(\hat{\beta})}\nabla_{f_n(\hat{\beta})}^{\top}\right)^{\frac{1}{2}}\sqrt{N}\left(\hat{\beta}-\beta_0\right)\right\|^2-1\right)\xrightarrow{d}\mathcal{N}(0,1).$$

4.1 Test Statistic

The test statistic via bootstrapping method is obtained as follows:

(i) Estimate $\hat{\beta}$ (LS) and $\hat{\epsilon}_n$ where $\hat{\epsilon}_n := \mathrm{Y}_n - f(\mathbf{X}_n^{\top}\hat{\beta})$, $n = 1, 2, ..., N$.

(ii) Independent of the observed data $D_n := \{(\mathrm{Y}_n, \mathbf{X}_n) : 1 \leq n \leq N\}$, use a random number generator to generate $R_1, R_2, ..., R_N \overset{iid}{\sim}$ Rademacher. Define $\mathrm{Y}_n^* := f(\mathbf{X}_n^{\top}\hat{\beta}) + R_n\hat{\epsilon}_n, n = 1, 2, ..., N$.

(iii) Re-estimate β using the data $\{(\mathrm{Y}_n^*, \mathbf{X}_n) : 1 \leq n \leq N\}$. Denote it by β^*.

(iv) Repeat steps (ii)-(iii) B times. Calculate the following and denote it as t_k^*

$$t_k^* := \sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\hat{\mathbf{\Sigma}}^{-1/2}\sqrt{B}\left(\frac{1}{B}\sum_{b=1}^{B}\beta_b^* - \hat{\beta}\right)\right\|^2 - 1\right)$$

where $\hat{\mathbf{\Sigma}} = \left(\sum_{n=1}^{N}\nabla_{f_n(\hat{\beta})}\nabla_{f_n(\hat{\beta})}^{\top}\right)^{-1}\left(\sum_{n=1}^{N}\hat{\epsilon}_n^2\nabla_{f_n(\hat{\beta})}\nabla_{f_n(\hat{\beta})}^{\top}\right)\left(\sum_{n=1}^{N}\nabla_{f_n(\hat{\beta})}\nabla_{f_n(\hat{\beta})}^{\top}\right)^{-1}$.

(v) Repeating steps (ii)-(iv) K times provides a sequence of $\{t_k^*\}_{k=1}^{K}$. The test statistic is

$$R_{N,D,K,B}^* := K^{\frac{1}{2}}\sup_{z\in R}|K^{-1}\sum_{k=1}^{K}\mathbf{1}_{(-\infty,z]}(t_k^*) - \Phi(z)| \tag{2}$$

where $\Phi(.)$ is the CDF of the standard normal distribution.

Theorem 4.1. Under *Assumptions* 4.1 - 4.5, let B and K satisfy $\sqrt{\frac{K}{D}} = o_p(1)$ and $\frac{\sqrt{K}D^{\frac{7}{4}}}{\sqrt{B}N^{\frac{3}{2}}} = o_p(1)$, which gives

$$R_{N,D,K,B}^* \xrightarrow{d} \sup_{z\in R}|B(\Phi(z))|$$

where $B(.)$ is the Brownian bridge.

Remark: The idea in non-linear case is similar to that in linear case, but more assumptions (4.4 & 4.5) are required especially for the covariance matrix estimator, since we use the simplest one. Other covariance matrix estimators for high dimensional settings could be tried to release some assumptions as an extension of our test.

4.2 Example: Logistic Regressions

The required assumptions seem complicated for the general non-linear case; however, for the logistic regression, only one assumption is needed. This assumption has been studied by Sur & Candès (2019).

Assumption 4.7. $\mathbf{X}_n \sim \mathcal{N}(\mathbf{0}, N^{-1}\mathbf{I}_D)$, $\mathbb{E}[\epsilon_n^2|\mathbf{X}_n] = \frac{1}{N}$

Theorem 4.2. Under *Assumption* 4.7, for the logistic regression model, after letting B and K satisfy $\sqrt{\frac{K}{D}} = o_p(1)$ and $\frac{\sqrt{K}D^{\frac{7}{4}}}{\sqrt{B}N^{\frac{3}{2}}} = o_p(1)$, we have

$$R^*_{N,D,K,B} \overset{d}{\to} \sup_{z\in R} |B(\Phi(z))|$$

where $B(.)$ is the Brownian bridge.

5 Monte Carlo Study

In this section, we present a Monte Carlo study to illustrate the size and power of our test. We consider different designs for three cases: random vectors, estimated vectors from the linear regression model, and estimated vectors from the non-linear regression model. We set different sample sizes N, from 100 to 2000. Within each case, the dimension D is chosen from 2 to 100. To keep the running time manageable, these results are based on 300 simulations.

For random vectors, suppose there is a sequence of D-dimension random vectors $\{\mathbf{V}_n\}_{n=1}^N$.

We considered four different designs.

$$
\begin{aligned}
\text{Design i. } & \mathbf{V}_n \sim \mathcal{N}(0, I_{D\times D}) \\
\text{Design ii. } & \mathbf{V}_n \sim \mathcal{N}(\alpha, I_{D\times D}) \\
& \alpha = (1, 2, ..., D)' \\
\text{Design iii. } & \mathbf{V}_n \sim \mathcal{N}(\alpha, \Omega_D) \\
& \alpha = (1, 2, ..., D)' \\
& \Omega_{D,ii} = 1, \text{where } i = 1, 2, ...D. \\
& \Omega_{D,ij} = 0.8, \text{where } i, j = 1, 2, ...D, i \neq j. \\
\text{Design iv. } & \mathbf{V}_n \sim \mathcal{N}(\alpha, \Omega_D) \\
& \alpha = (1, 2, ..., D)' \\
& \Omega_{D,ii} = 1, \text{where } i = 1, 2, ...D. \\
& \Omega_{D,ij} = 0.8, \text{where } i = D, j = D-1 \text{ or } i = D-1, j = D.
\end{aligned}
$$

For the first three designs, we know the true values of the mean and variance-covariance matrix for the distribution of $\{\mathbf{V}_n\}_{n=1}^N$.

For the fourth design, we estimate the mean of the distribution as

$$\alpha = \frac{1}{N}\sum_{n=1}^{N} \mathbf{V}_n.$$

Ω_D is a bandable covariance matrix. With this class of high dimensional covariance matrices, Bickel and Levina (2008b) introduced the following estimators:

$$
\begin{aligned}
\mathcal{U}_\alpha(M_0, M) = \{\mathbf{\Sigma} : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq \boldsymbol{M} k^{-\alpha} \quad \text{for all} \quad k > 0 \\
0 < & M_0^{-1} \leq \lambda_{\min}(\mathbf{\Omega}) \leq \lambda_{\max}(\mathbf{\Omega}) \leq \boldsymbol{M}_0\} \\
\hat{\mathbf{\Omega}}_k = & B_k(\mathbf{S}) {:=} (s_{ij}\mathbf{1}(|i{-}j|{\leq}k))
\end{aligned}
$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ are the minimum and maximum eigenvalues of a matrix, α is a constant, and $\mathbf{1}(\cdot)$ is an indicator function. They prove that if $k = k_n \asymp (N^{-1}\log p)^{-1/(2(\alpha+1))}$, then uniformly over the class $\mathcal{U}_\alpha$, for Gaussian or light-tailed data,

$$\|\hat{\mathbf{\Sigma}}_{k_n} - \mathbf{\Sigma}\| = O_P\left(\left(\frac{\log(D)}{N}\right)^{\alpha/(2(\alpha+1))}\right) = \|\hat{\mathbf{\Sigma}}_{k_n}^{-1} - \mathbf{\Sigma}^{-1}\|.$$

For estimated vectors from the linear regressions, the data generating process (DGP) is

$$\mathrm{Y}_n = \mathbf{X}_n^\top \beta_0 + \epsilon_n$$

where $\beta_0 = (0, 0, ..., 0)^\top$ and $\epsilon_n \sim \mathcal{N}(0, 1)$. For the distribution of $\mathbf{X}_n$, three designs are considered.

Design v.	$\mathbf{X}_n \sim \mathcal{N}(0, I_{D\times D})$	for all $n = 1, 2, ..., N.$
Design vi.	$\mathbf{X}_n \sim \chi_1^2$	for all $n = 1, 2, ..., N.$
Design vii.	$\mathbf{X}_n \sim \mathcal{N}(0, I_{D\times D})$	for all $n < \frac{N}{2},$
	$\mathbf{X}_n \sim \chi_1^2$	for all $n \geq \frac{N}{2}.$

For the non-linear case, since the restrictions on the non-linear function and the regressors are strict, we only consider one design (Design viii).

$$\mathrm{Y}_n = \frac{1}{1 + e^{-\mathbf{X}_n^\top \beta_0}} + \epsilon_n$$

where $\mathbf{X}_n \sim \mathcal{N}\left(0, \frac{1}{N}\right)$, $\epsilon_n \sim \mathcal{N}\left(0, \frac{1}{N}\right)$ and $\beta_0 = (0, 0, ..., 0)^\top$.

In Tables 1 & 2, since we know the true values of the mean and covariance matrix for the random vectors, the empirical rejection rates in Designs i,ii, and iii share the same results: the empirical rejection rates approach 1 as D decreases to 2; the empirical rejection rates do not decrease to 0.05 when N is less than or equal to 200 and even D increases to 100; the empirical rejection rates are less than 0.1 when D is 30 and N is 2000. When the sample sizes are 300, 500, 1000, 2000, we obtain the following results. In Design i, the "dividing line" of 0.1 for the empirical rejection rates are 50, 40, 30, 30, respectively; In Design ii, the "dividing line" of 0.1 for the empirical rejection rates are 70, 70, 30, 30, respectively; In Design iii, the "dividing line" of 0.1 for the empirical rejection rates are 80, 70, 60, 30, respectively. In Design iv, since we estimate the values of the mean and covariance matrix for the random vectors, we obtain a different conclusion: the empirical rejection rates do not decrease to 0.05 when N is less than or equal to 300 and even D increases to 100; the "dividing line" of 0.1 for the empirical rejection rates, when sample size is 2000, is 60, which is quite larger than that in the first three designs.

The results in Tables 3 & 4 also suggest that our test for estimated vectors works well. For Design v, when the independent random vector $\mathbf{X}_n$ is normal, the empirical rejection rates approach 1 as D decreases to 2. Additionally, we fail to reject the null hypothesis when the dimension D increases up to 100, as the empirical rejection rates are close to the nominal value 0.05 (at 5% level of significance). This is true for all sample sizes. The dimension $D = 100$ seems to be the cutoff, the "dividing line" of 0.05, for judging the high dimensional settings under our specific designs. Even when we change the distribution of $\mathbf{X}_n$ to the Chi-squared distribution in Design vi or to the combination of the normal and Chi-squared distribution in Design vii, the results are still consistent. For the non-linear case in Design viii, our test performs well under $N = 500, 1000, 2000$. However, when $N = 100$, the rejection rate decreases first but increase rapidly up to 1 as D rises beyond 30. This issue is caused by the least square estimator which does not provide good estimates for such small sample sizes and such large dimensions. As we mentioned, one can try other estimators as extensions of our test. Overall, these results suggest that our test appears to have a good size and power for moderate sample sizes.

Table 1: Empirical rejection rates in designs for **random vectors**: Part I

	Design i						Design ii					
	Sample size N						Sample size N					
D	100	200	300	500	1000	2000	100	200	300	500	1000	2000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.628	0.588	0.606	0.592	0.568	0.558	0.618	0.589	0.556	0.564	0.602	0.505
4	0.476	0.452	0.470	0.470	0.406	0.502	0.517	0.473	0.469	0.477	0.385	0.498
5	0.468	0.428	0.394	0.388	0.366	0.376	0.511	0.432	0.396	0.367	0.365	0.385
6	0.400	0.360	0.340	0.368	0.376	0.320	0.414	0.352	0.331	0.404	0.341	0.345
7	0.364	0.354	0.288	0.298	0.300	0.278	0.394	0.375	0.265	0.284	0.290	0.288
8	0.302	0.276	0.324	0.232	0.262	0.296	0.280	0.288	0.299	0.239	0.262	0.313
9	0.336	0.264	0.264	0.248	0.272	0.232	0.350	0.285	0.285	0.240	0.283	0.218
10	0.276	0.252	0.224	0.212	0.184	0.218	0.250	0.264	0.224	0.211	0.199	0.223
20	0.216	0.220	0.152	0.118	0.130	0.114	0.221	0.236	0.161	0.120	0.122	0.108
30	0.184	0.162	0.146	0.126	0.096	0.092	0.198	0.147	0.146	0.118	0.095	0.096
40	0.164	0.120	0.136	0.090	0.092	0.082	0.164	0.119	0.124	0.093	0.084	0.075
50	0.174	0.118	0.100	0.070	0.095	0.080	0.175	0.108	0.106	0.074	0.092	0.074
60	0.188	0.118	0.088	0.105	0.070	0.098	0.194	0.118	0.112	0.108	0.074	0.097
70	0.150	0.088	0.098	0.078	0.072	0.072	0.148	0.094	0.090	0.072	0.067	0.070
80	0.154	0.106	0.089	0.070	0.110	0.060	0.164	0.112	0.081	0.069	0.098	0.059
90	0.142	0.128	0.082	0.094	0.074	0.064	0.146	0.131	0.079	0.093	0.067	0.070
100	0.154	0.108	0.098	0.080	0.058	0.052	0.144	0.099	0.095	0.075	0.058	0.050

NOTE: The columns report the fraction of simulations for which the p-value is less than 0.05. We set $B = 500$.

Table 2: Empirical rejection rates in designs for **random vectors**: Part II

	Design iii						Design iv					
	Sample size N						Sample size N					
D	100	200	300	500	1000	2000	100	200	300	500	1000	2000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.685	0.637	0.552	0.620	0.542	0.549	0.684	0.601	0.723	0.684	0.661	0.655
4	0.481	0.492	0.462	0.515	0.390	0.522	0.484	0.488	0.494	0.511	0.480	0.520
5	0.484	0.431	0.410	0.401	0.342	0.348	0.493	0.440	0.405	0.433	0.406	0.387
6	0.440	0.336	0.308	0.373	0.405	0.331	0.468	0.405	0.364	0.398	0.382	0.335
7	0.341	0.345	0.286	0.327	0.279	0.298	0.373	0.367	0.302	0.301	0.354	0.331
8	0.311	0.269	0.304	0.229	0.261	0.274	0.332	0.303	0.346	0.249	0.268	0.342
9	0.342	0.250	0.258	0.252	0.258	0.222	0.362	0.277	0.285	0.255	0.323	0.276
10	0.282	0.240	0.239	0.232	0.192	0.211	0.308	0.255	0.235	0.247	0.185	0.220
20	0.220	0.203	0.164	0.127	0.138	0.109	0.223	0.249	0.174	0.129	0.144	0.121
30	0.187	0.147	0.144	0.121	0.090	0.086	0.211	0.168	0.166	0.135	0.108	0.106
40	0.161	0.110	0.139	0.089	0.096	0.085	0.167	0.142	0.157	0.098	0.100	0.115
50	0.179	0.107	0.091	0.067	0.121	0.082	0.192	0.130	0.116	0.079	0.129	0.103
60	0.185	0.126	0.092	0.115	0.070	0.095	0.208	0.126	0.105	0.117	0.079	0.091
70	0.138	0.090	0.103	0.077	0.066	0.069	0.156	0.093	0.107	0.091	0.103	0.075
80	0.143	0.101	0.088	0.070	0.109	0.065	0.159	0.111	0.128	0.083	0.083	0.062
90	0.143	0.139	0.084	0.103	0.070	0.066	0.168	0.153	0.089	0.099	0.080	0.072
100	0.148	0.112	0.102	0.073	0.055	0.049	0.162	0.121	0.112	0.082	0.061	0.055

NOTE: The columns report the fraction of simulations for which the p-value is less than 0.05. We set $B = 500$.

Table 3: Empirical rejection rates in designs for estimated vectors from **linear** regressions

	Design v				Design vi				Design vii			
	Sample size N				Sample size N				Sample size N			
D	100	500	1000	2000	100	500	1000	2000	100	500	1000	2000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0.607	0.600	0.623	0.580	0.583	0.547	0.550	0.597	0.570	0.607	0.603	0.550
4	0.467	0.470	0.460	0.420	0.427	0.440	0.430	0.420	0.470	0.467	0.463	0.423
5	0.387	0.363	0.407	0.380	0.340	0.373	0.363	0.380	0.380	0.400	0.380	0.377
6	0.297	0.313	0.287	0.300	0.333	0.313	0.313	0.283	0.300	0.283	0.290	0.337
7	0.293	0.260	0.243	0.283	0.230	0.297	0.240	0.270	0.283	0.290	0.287	0.223
8	0.213	0.227	0.193	0.230	0.207	0.260	0.243	0.233	0.233	0.217	0.207	0.233
9	0.250	0.250	0.250	0.230	0.227	0.177	0.213	0.240	0.200	0.230	0.223	0.223
10	0.223	0.190	0.203	0.217	0.190	0.200	0.203	0.213	0.260	0.197	0.200	0.270
20	0.160	0.110	0.117	0.140	0.110	0.127	0.153	0.130	0.117	0.177	0.107	0.143
30	0.133	0.110	0.130	0.097	0.113	0.083	0.100	0.077	0.093	0.093	0.083	0.097
40	0.080	0.090	0.080	0.093	0.077	0.067	0.097	0.097	0.080	0.130	0.063	0.080
50	0.080	0.050	0.113	0.080	0.080	0.070	0.087	0.080	0.063	0.073	0.070	0.093
60	0.077	0.100	0.087	0.060	0.087	0.060	0.067	0.057	0.077	0.083	0.063	0.053
70	0.070	0.070	0.073	0.057	0.080	0.073	0.050	0.047	0.067	0.043	0.060	0.040
80	0.073	0.067	0.050	0.063	0.073	0.077	0.063	0.057	0.050	0.060	0.057	0.073
90	0.050	0.050	0.047	0.060	0.067	0.060	0.057	0.060	0.030	0.053	0.053	0.060
100	-	0.047	0.057	0.057	-	0.050	0.053	0.057	-	0.057	0.047	0.053

NOTE: The columns report the fraction of simulations for which the p-value is less than 0.05. We set $B = 500$ and $K = 100$ for Designs v-vii.

Table 4: Empirical rejection rates for estimated vectors from **non-linear** regressions

	Design viii			
	Sample size N			
D	100	500	1000	2000
2	1.000	1.000	1.000	1.000
3	0.583	0.573	0.553	0.568
4	0.403	0.436	0.413	0.447
5	0.307	0.360	0.390	0.354
6	0.213	0.317	0.330	0.341
7	0.190	0.233	0.260	0.268
8	0.123	0.230	0.257	0.245
9	0.077	0.200	0.247	0.172
10	0.097	0.223	0.160	0.128
20	0.083	0.130	0.140	0.156
30	1.000	0.113	0.137	0.079
40	1.000	0.067	0.083	0.075
50	1.000	0.083	0.077	0.068
60	1.000	0.090	0.087	0.061
70	1.000	0.047	0.063	0.065
80	1.000	0.057	0.063	0.057
90	1.000	0.060	0.053	0.044
100	-	0.053	0.037	0.041

NOTE: We set $B = 500$ and $K = 100$ for Design viii.

6 Conclusion

We have shown how to test for high dimensionality of random and estimated vectors especially from the linear and non-linear regression models. One can use this method to check whether their vectors are under high dimensional or classical settings and then decide to use some high dimensional versions of inference methods from existing literature. Results from the simulation study suggest that our test can work well for both random and estimated vectors. Some concerns about applying the least square estimator in our test could be raised, as there have been some advanced estimation methods in the literature. We suggest this as a potential extension of our test by replacing the least square estimator by other estimators.

Appendices

Proof of Lemma 3.1 Let's first define $F(\beta) = \sum_{n=1}^{N} \mathbf{X}_n \left(\epsilon_n - \mathbf{X}_n^\top \beta\right)$. WLOG assume $\beta = 0$. Then there is a root $\hat{\beta}$ of the equation $F(\beta) = 0$ satisfying $\|\hat{\beta}\|^2 = O_p(D/N)$ if the result 6.3.4 of Ortega and Rheinboldt (1970) holds here. It equivalently to show that $\beta^\top F(\beta) < 0$ for $\|\beta\|^2 = BD/N$.

$$\beta^\top F(\beta) = \sum_{n=1}^{N} \mathbf{X}_n^\top \beta \left(\epsilon_n - \mathbf{X}_n^\top \beta\right) = \sum_{n=1}^{N} \mathbf{X}_n^\top \beta \epsilon_n - \beta^\top \mathbf{X}_n \mathbf{X}_n^\top \beta =: M_1 - M_2.$$

We know that $M_1 \leq \|\beta\| \|\sum_{n=1}^{N} \mathbf{X}_n \epsilon_n\|$. Under *assumption* 3.3,

$$\mathbb{E}\left[\|\sum_{n=1}^{N} \mathbf{X}_n \epsilon_n\|^2\right] = \sum_{d=1}^{D} \sum_{n=1}^{N} \mathbf{X}_{n,d}^2 \mathbb{E}\left[\epsilon_n^2\right] \leq BND.$$

Therefore, Chebychev's inequality implies that for any $\epsilon > 0$ there is a constant C_1, such that for all n

$$\mathbb{P}\left\{M_1 \leq C_1 \sqrt{ND} \|\beta\| \text{ for all } \beta\right\} \geq 1 - \epsilon.$$

For M_2, *assumption* 3.2 implies that there is a constant δ such that

$$\sum_{n=1}^{N} \beta^\top \mathbf{X}_n \mathbf{X}_n^\top \beta \geq bN \|\beta\|^2$$

for all β with $\|\beta\| \leq \delta$. As a result, there is N_0 such that for $n \geq N_0$

$$\mathbb{P}\{M_1 - M_2 \leq C_1 \sqrt{ND} \|\beta\| - bN \|\beta\|^2 \text{ for all } \beta \text{ with } \|\beta\| \leq \delta\} \geq 1 - \epsilon.$$

Let $\sqrt{C} = 2C_1/b$ and choose $N_0^\top > N$ such that $C\left(D/N^\top\right) \leq \delta^2$, for $n \geq N_0^\top$ then

$$\begin{aligned} &\mathbb{P}\left\{M_1 - M_2 \leq 0 \text{ for all } \beta \text{ with } \|\beta\|^2 = CD/N\right\} \\ \geq &\mathbb{P}\left\{M_1 - M_2 \leq -\frac{1}{2} BbD \text{ for all } \beta \text{ with } \|\beta\|^2 = CD/N\right\} \geq 1 - \epsilon. \end{aligned}$$

□

Proof of Lemma 3.2 Let's consider $\sigma^{-1}\left(N^{-1}\mathbb{X}^\top\mathbb{X}\right)^{\frac{1}{2}} \sqrt{N}\left(\hat{\beta} - \beta_0\right)$ first.

$$\begin{aligned} &\sigma^{-1}\left(N^{-1}\mathbb{X}^\top\mathbb{X}\right)^{\frac{1}{2}} \sqrt{N}\left(\hat{\beta} - \beta_0\right) = \sigma^{-1}\left(N^{-1}\mathbb{X}^\top\mathbb{X}\right)^{-1/2} \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbf{X}_n \epsilon_n \\ =&\sigma^{-1}\mathbf{\Sigma}^{-1/2} \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbf{X}_n \epsilon_n + \sigma^{-1}\left(\left(N^{-1}\mathbb{X}^\top\mathbb{X}\right)^{-1/2} - \mathbf{\Sigma}^{-1/2}\right) \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \mathbf{X}_n \epsilon_n \\ =&: (a) + (b). \end{aligned}$$

For (a), *assumption* 3.3 implies that $\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbf{X}_n\epsilon_n$ converges to W which follows a multivariate normal distribution with the covariance matrix $\sigma^2\mathbf{\Sigma}$, so $\sigma^{-1}\mathbf{\Sigma}^{-1/2}\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbf{X}_n\epsilon_n$ will converge to standard multivariate normal distribution. Then

$$\sqrt{\frac{D}{2}}\left(\frac{1}{D}\|\sigma^{-1}\left(N^{-1}\mathbb{X}^\top\mathbb{X}\right)^{1/2}\sqrt{N}\left(\hat{\beta}-\beta_0\right)\|_2^2-1\right)=\frac{\|\left(a\right)\|^2-D}{\sqrt{2D}}+\frac{e}{\sqrt{D}}$$

where e consists of $\|\left(b\right)\|^2$ and $\left(a\right)^\top\left(b\right)$. *Assumption* 3.4 provides that $\|\left(b\right)\|^2=o_p\left(\sqrt{D}\right)$; and

$$\left(a\right)^\top\left(b\right)\leq\|\left(a\right)\|\|\left(b\right)\|=o_p\left(\sqrt{D}\right).$$

As a result, by continuous mapping theorem, we obtain the desired result. □

Proof of Theorem 3.1 Rewrite the $T^*_{N,D,K,B}$,

$$\begin{aligned}T^*_{N,D,K,B}&=\sqrt{K}\sup_{z\in R}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{1}_{(-\infty,z]}\left(t_k^*\right)-\Phi\left(z\right)\right)\\&=\sqrt{K}\sup_{z\in R}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{1}_{(-\infty,z]}\left(t_k^*\right)-\mathbb{E}\left(\mathbf{1}_{(-\infty,z]}\left(t_k^*\right)\right)+\mathbb{E}\left(\mathbf{1}_{(-\infty,z]}\left(t_k^*\right)\right)-\Phi\left(z\right)\right)\\&=\sqrt{K}\sup_{z\in R}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{1}_{(-\infty,z]}\left(t_k^*\right)-\mathbb{E}\left(\mathbf{1}_{(-\infty,z]}\left(t_k^*\right)\right)\right)+K^{\frac{1}{2}}\sup_{z\in R}\left(\mathbb{E}\left(\mathbf{1}_{(-\infty,z]}\left(t_k^*\right)\right)-\Phi\left(z\right)\right)\\&=:(3.a)+(3.b)\end{aligned}$$

Firstly, let's show that $t_k^*\xrightarrow{d}\mathcal{N}(0,1)$.

For each β_b^*, we know that

$$\beta_b^*=\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathrm{Y}_b^*=\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\left(\mathbb{X}\hat{\beta}+R_N^b\hat{\epsilon}\right)=\hat{\beta}+\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\sum_{n=1}^{N}\mathbf{X}_nR_n^b\hat{\epsilon}_n.$$

From above, conditional on $D_N:=\{\{\mathrm{Y}_n,\mathbf{X}_n\}:n=1,2,...,N\}$, we have $\mathbb{E}\left[\beta_b^*\right]=\hat{\beta}$ and

$$\begin{aligned}\mathrm{var}\left(\beta_b^*\right)&=\mathbb{E}\left[\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top R_N^b\hat{\epsilon}\hat{\epsilon}^\top {R_N^b}^\top\mathbb{X}\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\right]\\&=\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{E}\left[\sum_{n=1}^{N}\left(R_n^b\right)^2\hat{\epsilon}_n^2\mathbf{X}_n\mathbf{X}_n^\top\right]\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\\&=\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\left(\sum_{n=1}^{N}\hat{\epsilon}_n^2\mathbf{X}_n\mathbf{X}_n^\top\right)\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}.\end{aligned}$$

As a result, conditional on D_N, high dimensional CLT (under *assumption* 3.3) implies that

$$\left[\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\left(\sum_{n=1}^{N}\hat{\epsilon}_n^2\mathbf{X}_n\mathbf{X}_n^\top\right)\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\right]^{-1/2}\sqrt{B}\left(\frac{1}{B}\sum_{b=1}^{B}\beta_b^*-\hat{\beta}\right)\overset{d}{=}W$$

where $W\sim\mathcal{N}\left(0,I_{D\times D}\right)$. Therefore, by continuous mapping theorem, there is $t_k^*\xrightarrow{d}N(0,1)$.

For $(3.a)$, since the class $\mathbf{C}=\{(-\infty,z]:z\in R\}$ is a Donsker class. It is known that $(3.a)$ converges weakly in $l^\infty(R)$ to a Brownian bridge $B\left(\mathbb{P}\{t_k^*<z\}\right)$, and so converges to $B(\Phi(z))$ as $D,B\to\infty$ because $\mathbb{P}\{t_k^*<z\}\to\Phi(z)$.

For $(3.b)$,

$$(3.b)=K^{\frac{1}{2}}\left(\mathbb{P}\{t_k^*<z\}-\Phi\left(z\right)\right).$$

We aim to show that under some conditions, there is $(3.b)=o_p\left(1\right)$. Since we know that $\mathbb{P}\{t_k^*<z\}-\Phi\left(z\right)\to 0$ as $D,B\to\infty$, once the convergence rate is obtained, K can be set comparatively small to achieve our goal. We can rewrite $(3.b)$ as

$$\begin{aligned}(3.b)=&\sqrt{K}\left(\mathbb{P}\{t_k^*<z\}-\mathbb{P}\{t^0<z\}+\mathbb{P}\{t^0<z\}-\Phi\left(z\right)\right)\\=&\sqrt{K}\left(\mathbb{P}\{t_k^*<z\}-\mathbb{P}\{t^0<z\}\right)+K^{\frac{1}{2}}\left(\mathbb{P}\{t^0<z\}-\Phi\left(z\right)\right)\\=:&\,(3.b.1)+(3.b.2)\end{aligned}$$

where $t_0=\sqrt{\frac{D}{2}}\left(\frac{1}{D}\sum_{d=1}^{D}Z_d^2-1\right)$ with $Z_d\sim\mathcal{N}\left(0,1\right)$ for $d=1,2,...,D$. Based on Berry–Esseen theorem, we have that

$$(3.b.2)\leq\frac{C_1\mathbb{E}\left[|Z_d|^3\right]}{\sqrt{D}}\leq C_1\sqrt{\frac{K}{D}}$$

for a constant C_1. Similarly, for $(3.b.1)$, there is a constant C_2 such that

$$\begin{aligned}(3.b.1)\leq&C_2\frac{K^{\frac{1}{2}}D^{\frac{1}{4}}}{B^{\frac{3}{2}}}\sum_{b=1}^{B}\mathbb{E}\left[\left\|\left[\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\left(\mathbb{X}^\top\mathcal{D}\mathbb{X}\right)\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\right]^{-1/2}\left(\beta_b^*-\hat{\beta}\right)\right\|^3\right]\\=&C_2\frac{\sqrt{K}D^{\frac{1}{4}}}{\sqrt{B}}\mathbb{E}\left[\left\|\left[\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\left(\mathbb{X}^\top\mathcal{D}\mathbb{X}\right)\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\right]^{-1/2}\left(\beta_b^*-\hat{\beta}\right)\right\|^3\right]\end{aligned}$$

where $\mathcal{D}:=\text{Diag}\left(\hat{\epsilon}_1^2,\hat{\epsilon}_2^2,...,\hat{\epsilon}_N^2\right)$. By *assumption* 3.2,

$$\begin{aligned}&\mathbb{E}\left[\left\|\left[\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\left(\mathbb{X}^\top\mathcal{D}\mathbb{X}\right)\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\right]^{-1/2}\left(\beta_b^*-\hat{\beta}\right)\right\|^3\right]\\&\leq\left\|N^{-1}\mathbb{X}^\top\mathbb{X}\right\|^3\left\|\left(N^{-1}\mathbb{X}^\top\mathcal{D}\mathbb{X}\right)^{-1/2}\right\|^3\mathbb{E}\left[\left\|\sqrt{N}\left(\beta_b^*-\hat{\beta}\right)\right\|^3\right]\leq C_3D^{\frac{3}{2}}\end{aligned}$$

for a constant C_3. Therefore, we have $(3.b) = o_p(1)$ for B and K with $\frac{\sqrt{K}D^{\frac{7}{4}}}{\sqrt{B}} \to 0$ and $\frac{K}{D} \to 0$. □

Proof of Lemma 4.1 WLOG let's assume $\beta_0 = 0$. Similar to *Lemma 3.1*, it suffices to show that $\beta^\top F(\beta) < 0$ for $\|\beta\|^2 = BD/N$ where $F(\beta) = \sum_{n=1}^N \nabla_{f_n(\beta)} (\mathrm{Y}_n - f(\mathbf{X}_n, \beta))$.

$$\begin{aligned}
\beta^\top F(\beta) &= \sum_{n=1}^N \nabla_{f_n(\beta)}^\top \beta \left(\epsilon_n + f(\mathbf{X}_n, 0) - f(\mathbf{X}_n, \beta)\right) \\
&= \sum_{n=1}^N \nabla_{f_n(\beta)}^\top \beta \epsilon_n - \sum_{n=1}^N \nabla_{f_n(\beta)}^\top \beta \left(f(\mathbf{X}_n, \beta) - f(\mathbf{X}_n, 0)\right) \\
&=: M_1 - M_2.
\end{aligned}$$

For M_1, $M_1 \leq \|\beta\| \| \sum_{n=1}^N \nabla_{f_n(\beta)} \epsilon_n\|$ and under *assumption* 4.1, there is

$$\mathbb{E}\| \sum_{n=1}^N \nabla_{f_n(\beta)} \epsilon_n\|^2 = \sum_{d=1}^D \sum_{n=1}^N \nabla_{f_{n,d}(\beta)}^2 \mathbb{E}\epsilon_n^2 \leq BND.$$

Therefore, Chebychev's inequality implies that for any $\epsilon > 0$ there is a constant C_1 such that

$$\mathbb{P}\{M_1 \leq C_1 \sqrt{ND} \|\beta\| \text{ for all } \beta\} \geq 1 - \epsilon.$$

For M_2,

$$\begin{aligned}
M_2 = \sum_{n=1}^N \nabla_{f_n(\beta)}^\top \beta \left(f(\mathbf{X}_n, \beta) - f(\mathbf{X}_n, 0)\right) &= \sum_{n=1}^N \nabla_{f_n(\beta)}^\top \beta \nabla_{f_n\left(\tilde{\beta}\right)}^\top \beta \\
= \sum_{n=1}^N \beta^\top \nabla_{f_n(\beta)} \nabla_{f_n\left(\tilde{\beta}\right)}^\top \beta &\geq C_2 N \|\beta\|^2 \ (\textit{assumption } 4.2)
\end{aligned}$$

for all β with $\|\beta\| \leq \delta_2$.

Thus, there is N such that for $n \geq N$

$$\mathbb{P}\{M_1 - M_2 \leq C_1 \sqrt{ND} \|\beta\| - C_2 N \|\beta\|^2 \text{ for all } \beta \text{ with } \|\beta\| \leq \delta\} \geq 1 - 2\epsilon.$$

Let $\sqrt{C} = 2C_1/C_2$ and choose $N^\top > N$, so that $C\left(D/N^\top\right) \leq \delta^2$ for $n \geq N^\top$. Then we have

$$\begin{aligned}
&\mathbb{P}\{M_1 - M_2 \leq 0 \text{ for all } \beta \text{ with } \|\beta\|^2 = CD/N\} \\
\geq &\mathbb{P}\{M_1 - M_2 \leq -1/2BC_2D \text{ for all } \beta \text{ with } \|\beta\|^2 = CD/N\} \geq 1 - 2\epsilon.
\end{aligned}$$

□

Proof of Lemma 4.2 Recall

$$S_N(\beta) = \sum_{n=1}^{N} \left(\mathrm{Y}_n - f(\mathbf{X}_n, \beta)\right)^2.$$

By taking Taylor expansion for $\frac{\partial S_N}{\partial \beta}\Big|_{\hat{\beta}}$ around the true value β_0, there is

$$\frac{\partial S_N}{\partial \beta}\bigg|_{\hat{\beta}} = \frac{\partial S_N}{\partial \beta}\bigg|_{\beta_0} + \frac{\partial^2 S_N}{\partial \beta \partial \beta^\top}\bigg|_{\beta^*} \left(\hat{\beta} - \beta_0\right)$$

where β_0 is the true value. Since the $\hat{\beta}$ is the minimizer of $S_N(\beta)$, the left hand side of the above equation is zero. Therefore, we obtain

$$-\frac{1}{N} \frac{\partial^2 S_N}{\partial \beta \partial \beta^\top}\bigg|_{\beta^*} \sqrt{N}\left(\hat{\beta} - \beta_0\right) = \frac{1}{\sqrt{N}} \frac{\partial S_N}{\partial \beta}\bigg|_{\beta_0}. \tag{3}$$

Let's first consider $\frac{1}{N} \frac{\partial^2 S_N}{\partial \beta \partial \beta^\top}\Big|_{\beta^*}$. Differentiating it with respect to β yields

$$\frac{1}{N} \frac{\partial^2 S_N}{\partial \beta \partial \beta^\top} = \frac{2}{N} \sum_{n=1}^{N} \nabla_{f_n(\beta)} \nabla_{f_n(\beta)}^\top - \frac{2}{N} \sum_{n=1}^{N} \left[f_n(\beta_0) - f_n(\beta)\right] \mathbf{H}_{f_n(\beta)} - \frac{2}{N} \sum_{n=1}^{N} \epsilon_n \mathbf{H}_{f_n(\beta)}.$$

Then, we can rewrite $\frac{1}{N} \frac{\partial^2 S_N}{\partial \beta \partial \beta^\top}\Big|_{\beta^*}$ to $2\mathbf{\Sigma} + (i) + (ii) + (iii) + (iv)$ in which

$$\begin{aligned}
(i) &= 2\frac{1}{N} \sum_{n=1}^{N} \nabla_{f_n(\beta_0)} \nabla_{f_n(\beta_0)}^\top - 2\mathbf{\Sigma}, \\
(ii) &= 2\frac{1}{N} \sum_{n=1}^{N} \nabla_{f_n(\beta^*)} \nabla_{f_n(\beta^*)}^\top - 2\frac{1}{N} \sum_{n=1}^{N} \nabla_{f_n(\beta_0)} \nabla_{f_n(\beta_0)}^\top, \\
(iii) &= -\frac{2}{N} \sum_{n=1}^{N} \left[f_n(\beta_0) - f_n(\beta^*)\right] \mathbf{H}_{f_n(\beta^*)}, \\
(iv) &= -\frac{2}{N} \sum_{n=1}^{N} \epsilon_n \mathbf{H}_{f_n(\beta^*)}.
\end{aligned}$$

As a result, (3) becomes

$$\sqrt{N}\left(\hat{\beta} - \beta_0\right) = (2\mathbf{\Sigma})^{-1} \frac{1}{\sqrt{N}} \frac{\partial S_N}{\partial \beta}\bigg|_{\beta_0} + e$$

where $e = (2\Omega)^{-1}((i)+(ii)+(iii)+(iv))\sqrt{N}\left(\hat{\beta}-\beta_0\right)$.

Now,

$$\begin{aligned}\sigma^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\nabla_{f_n(\hat{\beta})}\nabla_{f_n(\hat{\beta})}^{\top}\right)^{1/2}\sqrt{N}\left(\hat{\beta}-\beta_0\right) =&\sigma^{-1}\mathbf{\Sigma}^{1/2}(2\mathbf{\Sigma})^{-1}\frac{1}{\sqrt{N}}\frac{\partial S_N}{\partial\beta}\bigg|_{\beta_0}\\ &+\sigma^{-1}\left(\hat{\mathbf{\Sigma}}^{1/2}-\mathbf{\Sigma}^{1/2}\right)(2\mathbf{\Sigma})^{-1}\frac{1}{\sqrt{N}}\frac{\partial S_N}{\partial\beta}\bigg|_{\beta_0}\\ &+\sigma^{-1}\mathbf{\Sigma}^{1/2}e\\ =&:(4.a)+(4.b)+(4.c)\end{aligned}$$

where $\hat{\mathbf{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}\nabla_{f_n(\hat{\beta})}\nabla_{f_n(\hat{\beta})}^{\top}$.

Therefore, we can write $\sqrt{\frac{D}{2}}\left(\frac{1}{D}\left\|\sigma^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\nabla_{f_n(\hat{\beta})}\nabla_{f_n(\hat{\beta})}^{\top}\right)^{\frac{1}{2}}\sqrt{N}\left(\hat{\beta}-\beta_0\right)\right\|^2-1\right)$ as

$$\sqrt{\frac{D}{2}}\left(\frac{1}{D}\|\,(4.a)\,\|^2-1\right)+\frac{e^*}{\sqrt{D}}$$

where e^* consists of the sum of squared norms of $(4.b)$,$(4.c)$ plus the sum of the inner products $|\,(4.a)^{\top}(4.b)\,|$,$|\,(4.a)^{\top}(4.c)\,|$,$|\,(4.b)^{\top}(4.c)\,|$.

Under *assumption* 4.3, high dimensional CLT and continuous mapping theorem imply that

$$\sqrt{\frac{D}{2}}\left(\frac{1}{D}\|\,(4.a)\,\|^2-1\right)\xrightarrow{d}\mathcal{N}(0,1).$$

We are left to show that $e^* = o_p(\sqrt{D})$. *Assumptions* 4.4-4.7 are sufficient to yield $e^* = o_p(\sqrt{D})$. □

Proof of Theorem 4.1 The proof of *Theorem* 4.1 is a simple combination of the proofs for *Theorem* 3.1 and *Lemma* 4.2. □

Proof of Theorem 4.2

(Consistency) WLOG let's assume $\beta_0 = 0$. Similar to *Lemma 3.1*, it suffices to show that $\beta^{\top}F(\beta) < 0$ for $\|\beta\|^2 = BD/N$ where $F(\beta) = \sum_{n=1}^{N}\nabla_{f_n(\beta)}(\mathrm{Y}_n - f(\mathbf{X}_n,\beta))$.

$$\begin{aligned}\beta^{\top}F(\beta) &= \sum_{n=1}^{N}\nabla_{f_n(\beta)}^{\top}\beta\left(\epsilon_n + f(\mathbf{X}_n,0) - f(\mathbf{X}_n,\beta)\right)\\ &= \sum_{n=1}^{N}\nabla_{f_n(\beta)}^{\top}\beta\epsilon_n - \sum_{n=1}^{N}\nabla_{f_n(\beta)}^{\top}\beta\left(f(\mathbf{X}_n,\beta)-f(\mathbf{X}_n,0)\right)\\ &=: M_1 - M_2.\end{aligned}$$

For M_1, $M_1 \leq \|\beta\| \| \sum_{n=1}^{N} \nabla_{f_n(\beta)} \epsilon_n \|$, and there is a constant C_1 such that

$$\mathbb{E}\left[\| \sum_{n=1}^{N} \nabla_{f_n(\beta)} \epsilon_n \|^2 \right] = \sum_{d=1}^{D} \sum_{n=1}^{N} \nabla^2_{f_{n,d}(\beta)} \mathbb{E}\left[\epsilon_n^2\right] \leq C_1 \sum_{d=1}^{D} \sum_{n=1}^{N} \frac{1}{N} \frac{1}{N} \leq C_1 \frac{D}{N}.$$

Therefore, Chebychev's inequality implies that for any $\epsilon > 0$ there is a constant C_1 such that

$$\mathbb{P}\left\{ M_1 \leq C_1 \sqrt{\frac{D}{N}} \|\beta\| \text{ for all } \beta \right\} \geq 1 - \epsilon$$

For M_2,

$$\begin{aligned} M_2 =& \sum_{n=1}^{N} \nabla^{\top}_{f_n(\beta)} \beta \left(f\left(\mathbf{X}_n, \beta\right) - f\left(\mathbf{X}_n, 0\right) \right) = \sum_{n=1}^{N} \nabla^{\top}_{f_n(\beta)} \beta \nabla^{\top}_{f_n\left(\tilde{\beta}\right)} \beta = \sum_{n=1}^{N} \beta^{\top} \nabla_{f_n(\beta)} \nabla^{\top}_{f_n\left(\tilde{\beta}\right)} \beta \\ =&: \beta^{\top} \mathbb{X}^{\top} \mathcal{D}_1 \mathbb{X} \beta = \|\mathcal{D}_1^{1/2} \mathbb{X} \beta\|^2 = \|\mathcal{D}_1^{1/2}\|^2 \|\mathbb{X}\|^2 \|\beta\|^2 \end{aligned}$$

where $\mathcal{D}_1 = \text{Diag}\left(\frac{e^{x_1^{\top}\beta}}{\left(1+e^{x_1^{\top}\beta}\right)^2} \frac{e^{x_1^{\top}\tilde{\beta}}}{\left(1+e^{x_1^{\top}\tilde{\beta}}\right)^2}, ..., \frac{e^{x_N^{\top}\beta}}{\left(1+e^{x_N^{\top}\beta}\right)^2} \frac{e^{x_N^{\top}\tilde{\beta}}}{\left(1+e^{x_N^{\top}\tilde{\beta}}\right)^2} \right)$. *Exercise* 4.4.7 in (Vershynin (2018)) implies that there is a constant C_2 such that $\|\mathbb{X}\| \geq C_2$ with high probability. Together with upper bound for M_1, the lower bound for M_2 implies that $\beta^{\top} F(\beta) < 0$ with high probability.

Now, we left to show that *assumption* 4.3-4.5 hold. *Assumption* 4.3 is trivial. For *assumption* 4.4,

$$\left\| \frac{1}{N} \sum_{n=1}^{N} \nabla_{f_n(\beta)} \nabla^{\top}_{f_n(\beta)} - \frac{1}{N} \sum_{n=1}^{N} \nabla_{f_n(\beta_0)} \nabla^{\top}_{f_n(\beta_0)} \right\| = \frac{1}{N} \|\mathbb{X}^{\top} \mathcal{D}_2 \mathbb{X}\|$$

where $\mathcal{D}_2 = \text{Diag}\left(\frac{\left(e^{x_1^{\top}\beta}\right)^2}{\left(1+e^{x_1^{\top}\beta}\right)^4} - \frac{\left(e^{x_1^{\top}\beta_0}\right)^2}{\left(1+e^{x_1^{\top}\beta_0}\right)^4}, ..., \frac{\left(e^{x_N^{\top}\beta}\right)^2}{\left(1+e^{x_N^{\top}\beta}\right)^4} - \frac{\left(e^{x_N^{\top}\beta_0}\right)^2}{\left(1+e^{x_N^{\top}\beta_0}\right)^4} \right)$. *Theorem* 4.4.5 in (Vershynin (2018)) implies that $\|\mathbf{X}\|$ is bounded with high probability, so we have $\left\| \frac{1}{N} \sum_{n=1}^{N} \nabla_{f_n(\beta)} \nabla^{\top}_{f_n(\beta)} - \frac{1}{N} \sum_{n=1}^{N} \nabla_{f_n(\beta_0)} \nabla^{\top}_{f_n(\beta_0)} \right\| \leq O_p(\frac{1}{N}) = o_p(\frac{1}{\sqrt{D}})$. For *assumption* 4.5, one can use the similar strategy to obtain the desired upper bound.

References

Bai, Z., & Saranadasa, H. (1996). EFFECT OF HIGH DIMENSION: BY AN EXAMPLE OF A TWO SAMPLE PROBLEM. *Statistica Sinica*, 6(2), 311–329. http://www.jstor.org/stable/24306018

Bellman, R. (1957). Dynamic programming. –. *Princeton University Press.*

Calhoun, G. (2011). Hypothesis testing in linear regression when k / n is large. *Journal of Econometrics*, 165(2), 163–174. https://doi.org/10.1016/j.jeconom.2011.07.003

Chernozhukov, V., Chetverikov, D., & Kato, K. (2017). CENTRAL LIMIT THEOREMS AND BOOTSTRAP IN HIGH DIMENSIONS. *The Annals of Probability*, 45(4), 2309–2352. https://doi.org/10.1214/16-AOP1113

Chi, C. M., Vossler, P., Fan, Y., & Lv, J. (2022). ASYMPTOTIC PROPERTIES OF HIGH-DIMENSIONAL RANDOM FORESTS. *The Annals of Statistics*, 50(6), 3415–3438. https://doi.org/10.1214/22-AOS2234

Dempster, A. P. (1958). A High Dimensional Two Sample Significance Test. *The Annals of Mathematical Statistics*, 29(4), 995–1010. https://doi.org/10.1214/aoms/1177706437

He, X., & Shao, Q.-M. (2000). On Parameters of Increasing Dimensions. *Journal of Multivariate Analysis*, 73(1), 120–135. https://doi.org/10.1006/jmva.1999.1873

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5), 799–821. https://doi.org/10.1214/aos/1176342503

Jimenez, L. O., & Landgrebe, D. A. (1998). Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. IEEE Transactions on Systems, Man and Cybernetics. Part C, *Applications and Reviews*, 28(1), 39–54. https://doi.org/10.1109/5326.661089

Lehmann, E. L. (Erich L. (1999). Elements of large-sample theory. *Springer.*

Lindsay B G, Kettenring J, Siegmund D O. (2004). A report on the future of statistics

Portnoy, S. (1984). Asymptotic Behavior of M-Estimators of p Regression Parameters when p^2/n is Large. I. Consistency. *The Annals of Statistics*, 12(4), 1298–1309. https://doi.org/10.1214/aos/1176346793

Portnoy, S. (1988). Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity. *The Annals of Statistics*, 16(1), 356–366. https://doi.org/10.1214/aos/1176350710

Portnoy, S. (1991). Correction: Asymptotic Behavior of M Estimators of p Regression Parameters when p^2/n is Large: II. Normal Approximation. *The Annals of Statistics*, 19(4), 2282–2282. https://doi.org/10.1214/aos/1176348403

Rowell J G, Walters D E. (1976). Analysing data with repeated observations on each experimental unit[J]. *The Journal of Agricultural Science*, 1976, 87(2): 423-432.

Sur, P., & Candès, E. J. (2019), The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled Chi-square. *Probability Theory and Related Fields*, 175(1–2), 487–558. https://doi.org/10.1007/s00440-018-00896-9

Sur, P., & Candès, E. J. (2019), A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences - PNAS*, 116(29), 14516–14525. https://doi.org/10.1073/pnas.1810420116

Vershynin, R. (2018). High-dimensional probability: an introduction with applications in data science. *Cambridge University Press.*

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). A Survey of Large Language Models.. https://doi.org/10.48550/arxiv.2303.1822