Statistical Inference for High-Dimensional Models via Recursive Online-Score Estimation

Chengchun Shi[a], Rui Song[b], Wenbin Lu[b] and Runze Li[c]
[a]Department of Statistics, London School of Economics and Political Science, London; [b]Department of Statistics, North Carolina State University, Raleigh, NC; [c]Department of Statistics, The Pennsylvania State University, University Park, PA

Abstract

In this paper, we develop a new estimation and valid inference method for single or low-dimensional regression coefficients in high-dimensional generalized linear models. The number of the predictors is allowed to grow exponentially fast with respect to the sample size. The proposed estimator is computed by solving a score function. We recursively conduct model selection to reduce the dimensionality from high to a moderate scale and construct the score equation based on the selected variables. The proposed confidence interval (CI) achieves valid coverage without assuming consistency of the model selection procedure. When the selection consistency is achieved, we show the length of the proposed CI is asymptotically the same as the CI of the "oracle" method which works as well as if the support of the control variables were known. In addition, we prove the proposed CI is asymptotically narrower than the CIs constructed based on the de-sparsified Lasso estimator (van de Geer et al., 2014) and the decorrelated score statistic (Ning and Liu, 2017). Simulation studies and real data applications are presented to back up our theoretical findings.

Keywords: Confidence interval; Ultrahigh dimensions; Generalized linear models; Online estimation.

1 Introduction

Statistical inference for high-dimensional linear regression models has received more and more attention in the recent literature. Lee et al. (2016) proposed a valid post-selection-inference procedure for linear regression models. They targeted on the regression coefficients conditional on the model selected by the Lasso (Tibshirani, 1996), rather than the coefficients in the true model. The resulting confidence interval may change with the selected model and is hence difficult to interpret. Zhang and Zhang (2014) and Javanmard and Montanari (2014) proposed bias-corrected linear estimators based on the Lasso to form confidence intervals for individual regression coefficients. Liu and Yu (2013) and Liu et al. (2017) developed inference procedures by bootstrapping the Lasso+modified least squares estimator and the Lasso+partial ridge estimator, respectively. All these work, however, only considers linear regression models.

In this paper, we consider the class of generalized liner models (GLM, McCullagh and Nelder, 1989), which assumes the following conditional probability density function of the response Y_0 given the covariate vector $\boldsymbol{X}_0$,

$$\exp\left(\frac{Y_0\boldsymbol{X}_0^T\boldsymbol{\beta}_0 - b(\boldsymbol{\beta}_0^T\boldsymbol{X}_0)}{\phi_0}\right)c(Y_0), \tag{1}$$

for some $\boldsymbol{\beta}_0 = (\beta_{0,1}, \beta_{0,2}, \ldots, \beta_{0,p})^T \in \mathbb{R}^p$, some positive nuisance parameter ϕ_0 and some convex function $b(\cdot)$. We focus on constructing confidence intervals (CIs) for a univariate parameter of interest β_{0,j_0} for some $j_0 \in \{1, \ldots, p\}$. The main challenge in high-dimensional statistical inference lies in that the nonzero support set of the control variables (variables other than X_{0,j_0}) is unknown and needs to be estimated. Consider the following standard post-model-selection-inference procedure that first estimates the support of the controls based on some regularization methods, and then fits a generalized linear regression of the response on the variable of interest and the set of selected control variables. The validity of such a procedure typically relies on the perfect model selection at the first step, which is not guaranteed under the "small n, large p" settings.

Alternatively, one may apply sample-splitting estimation to allow for imperfect model selection. The idea of applying sample-splitting to high-dimensional statistical inference is implicitly contained in Wasserman and Roeder (2009). To construct CI for β_{0,j_0}, we can split the samples into two equal halves, use the first half to select the controls and evaluate β_{j_0} on the remaining second half of the data. Such methods are very similar to the single sample-splitting procedure described in Dezeure et al. (2015). However, the resulting CI will be approximately $\sqrt{2}$ times wider than the CI of our proposed procedure, since β_{0,j_0} is estimated based only on half of the samples. One can also average two such estimators by swapping the two sub-datasets that are split apart. However, the CI based on the aggregated estimator will fail when model selection consistencies are not guaranteed.

van de Geer et al. (2014) extended Zhang and Zhang (2014)'s methods to the GLM setup and proposed to construct CIs based on the de-sparsified Lasso estimator. Ning and Liu (2017) proposed to construct CIs for high-dimensional penalized M-estimators based on the decorrelated score statistic. These CIs are valid. However, the de-sparsified Lasso estimator and the decorrelated score statistic are computed by debiasing the Lasso estimator and the Rao's score test statistic, respectively. Their variances tend to increase after the de-biasing procedure, resulting in increased lengths of the corresponding CIs.

In this paper, we develop a new estimation and valid inference procedure for β_{0,j_0} under ultrahigh dimensional setting where the number of predictors p is allowed to grow exponentially fast with respect to the sample size. The idea originates from online learning algorithms for streaming datasets that recursively update estimators using new observations (see for example, Wang et al., 2016; Schifano et al., 2016). The proposed method differs from standard sample-splitting estimation. It divides the data into a series of non-overlapping "chunks". The target parameter β_{0,j_0} is estimated by solving a score equation. We first conduct model selection using a small chunk of data. Based on the selected control variables, we construct the score equation with the second chunk of data. Then we select the controls using the first two chunks of data and construct the estimating equation with the next chunk of data. We iterate this procedure until the last chunk of data is used. Note that we recursively conduct variable selection to construct the score equation. The accuracy of the proposed estimator gets improved with the dimensionality reduced from high to a moderate scale. As a result, we prove the Wald-type CI based on our estimator is asymptotically narrower than those based on the de-sparsified Lasso estimator and the decorrelated score statistic.

In addition, the proposed CI achieves valid coverage without assuming consistency of the model selection procedure. When the selection consistency is achieved, we show the length of the proposed CI is asymptotically the same as the CI of the "oracle" method which works as well as if the support of the control variables were known.

The rest of the paper is organized as follows. We consider a linear regression setup and introduce our methods in Section 2. In Section 3, we consider extensions to GLMs and investigate the asymptotic properties of the CI of our proposed procedure. Simulations studies are presented in Section 4.1 and Section 4.2. In Section 4.3, we apply the proposed method to a real dataset. Section 5 closes the paper with a summary and discusses some extensions of the proposed method. All the proofs are given in the supplementary material.

2 High-dimensional linear models

To better illustrate the idea, we begin by considering the following linear regression model:

$$Y_0 = \boldsymbol{X}_0^T \boldsymbol{\beta}_0 + \varepsilon_0,$$

where $\boldsymbol{\beta}_0 = (\beta_{0,1}, \beta_{0,2}, \ldots, \beta_{0,p})$ is a p-dimensional vector of regression coefficients, ε_0 is independent of the covariates $\boldsymbol{X}_0$ and satisfies $\mathrm{E}(\varepsilon_0) = 0$. Suppose $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ are a random sample from $(\boldsymbol{X}_0, Y_0)$ and the dimension p satisfies $\log p = o(n)$. We focus on constructing a confidence interval for a univariate parameter β_{0,j_0}. Extension to multi-dimensional parameters are given in Section 5.

Before presenting our approach, some notations are introduced. Let $\boldsymbol{\Sigma} = \mathrm{E}\boldsymbol{X}_0\boldsymbol{X}_0^T$ and $\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T/n$. For any $r \times q$ matrix $\boldsymbol{\Phi}$ and any sets $J_1 \subseteq [1, \ldots, r]$, $J_2 \subseteq [1, \ldots, q]$, we denoted by $\boldsymbol{\Phi}_{J_1,J_2}$ the submatrix of $\boldsymbol{\Phi}$ formed by rows in J_1 and columns in J_2. Similarly, for any q-dimensional vector $\boldsymbol{\psi}$, $\boldsymbol{\psi}_{J_1}$ stands for the subvector of $\boldsymbol{\psi}$ formed by elements in J_1. Let $|J_1|$ be the number of elements in J_1. Denote by $\mathcal{M}_{j_0} = \{j \neq j_0 : \beta_{0,j} \neq 0\}$. Let

$\mathbb{I} = \{1, \ldots, p\}$ and $\mathbb{I}_{j_0} = \mathbb{I} - \{j_0\}$. For any set $\mathcal{M} \subseteq \mathbb{I}_{j_0}$, define $\boldsymbol{\omega}_{\mathcal{M},j_0} = \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1} \boldsymbol{\Sigma}_{\mathcal{M},j_0}$ and

$$\sigma_{\mathcal{M},j_0}^2 = \boldsymbol{\Sigma}_{j_0,j_0} - \boldsymbol{\Sigma}_{\mathcal{M},j_0}^T \boldsymbol{\omega}_{\mathcal{M},j_0}.$$

Let $\|\mathbb{Z}\|_{\psi_p}$ be the Orlicz norm of any random variable $\mathbb{Z}$,

$$\|\mathbb{Z}\|_{\psi_p} \triangleq \inf_{c>0} \left\{ \mathrm{E} \exp\left(\frac{|\mathbb{Z}|^p}{c^p} \right) \leq 2 \right\}.$$

2.1 An online estimator

Before we present our algorithm, let us present the motivation of the online estimator. Suppose that we are interested in a constructing confidence interval for β_{0,j_0}, we construct an estimating equation for β_{0,j_0}. To this end, we propose to construct an estimating equation based on partial residual. Notice that

$$\mathrm{E}(Y_0 | \boldsymbol{X}_{0,\mathcal{M}_{j_0}}) = \beta_{0,j_0} \mathrm{E}(X_{0,j_0} | \boldsymbol{X}_{0,\mathcal{M}_{j_0}}) + \boldsymbol{\beta}_{0,\mathcal{M}_{j_0}}^T \boldsymbol{X}_{0,\mathcal{M}_{j_0}}.$$

Thus, it follows that

$$Y_0 - \mathrm{E}(Y_0 | \boldsymbol{X}_{0,\mathcal{M}_{j_0}}) = \beta_{0,j_0} \{X_{0,j_0} - \mathrm{E}(X_{0,j_0} | \boldsymbol{X}_{0,\mathcal{M}_{j_0}})\} + \varepsilon_0,$$

and we define the partial residual score equation

$$\sum_{t=1}^{n} \{X_{t,j_0} - \mathrm{E}(X_{t,j_0} | \boldsymbol{X}_{t,\mathcal{M}_{j_0}})\}(Y_t - \beta_{0,j_0} X_{t,j_0} - \boldsymbol{\beta}_{0,\mathcal{M}_{j_0}}^T \boldsymbol{X}_{t,\mathcal{M}_{j_0}}) = 0. \quad (2)$$

To use (2) for constructing statistical inference procedure for β_{0,j_0}, we need an estimate for $\boldsymbol{\beta}_{0,\mathcal{M}_{j_0}}$, $\mathcal{M}_{j_0}$ and $\mathrm{E}(X_{t,j_0} | \boldsymbol{X}_{t,\mathcal{M}_{j_0}})$. We propose using regularization methods, such as the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), MCP (Zhang, 2010) and Dantzig (Candes and Tao, 2007) etc., to obtain an initial estimate $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$. This corresponds to Step 2 in our proposed algorithm below. We may estimate $\mathcal{M}_{j_0}$ by (iterative) sure independence screening ((I)SIS, Fan and Lv, 2008) or some regularized regression procedure. Suppose

$\widehat{\mathcal{M}}$ is the selected model, then we can set

$$\widehat{\mathcal{M}}_{j_0} = \{j \in \widehat{\mathcal{M}} : j \neq j_0\},$$

as an estimate of $\mathcal{M}_{j_0}$. Due to ultrahigh dimensionality, some spuriously correlated predictors may be retained in the selected model (Fan et al., 2012), making it challenging to consistently estimate the variance of the solution computed by (2).

To address these concerns, we propose using data-splitting strategy for model selection and partial residual score evaluation. That is, we propose separately conducting model selection and evaluating the partial residual scores in (2) using different data subsets. Specifically, we use the sub-dataset $\mathcal{F}_t = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_t, Y_t)\}$ for model selection and update the contribution of the $(t+1)$-th sample $(\boldsymbol{X}_{t+1}, Y_{t+1})$ to the estimating equation by

$$\frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\{X_{t+1,j_0} - \widehat{\mathrm{E}}(X_{t+1,j_0}|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}})\}(Y - \beta_{0,j_0}X_{t+1,j_0} - \widetilde{\boldsymbol{\beta}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)}}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}),$$

where $\widehat{\mathcal{M}}_{j_0}^{(t)}$ denotes the model selected based on $\mathcal{F}_t$, $\hat{\sigma}^2_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ is the estimated variance of the residual $X_{t,j_0} - \mathrm{E}(X_{t,j_0}|\boldsymbol{X}_{t,\widehat{\mathcal{M}}_{j_0}^{(t)}})$ and $\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}$ is the subvector of $\widetilde{\boldsymbol{\beta}}$ formed by elements in $\widehat{\mathcal{M}}_{j_0}^{(t)}$. As a result, we propose using the following estimating equation

$$\sum_{t=s_n}^{n} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\{X_{t+1,j_0} - \widehat{\mathrm{E}}(X_{t+1,j_0}|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}})\}(Y - \beta_{0,j_0}X_{t+1,j_0} - \widetilde{\boldsymbol{\beta}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)}}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}) = 0, \quad (3)$$

where s_n is a pre-specified integer in order for us to do model selection reasonably well based on $\mathcal{F}_{s_n}$. This corresponds to Step 4 in our proposed algorithm below. The above estimating equation was motivated from the online estimator proposed in Luedtke and van der Laan (2016). The inclusion of the factor $1/\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ is necessary for theoretical development of asymptotic normality of the resulting estimate (See Step 4 of the proof of Theorem 2.1 for details). If one excludes the factor $1/\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ and set $\widehat{\mathcal{M}}_{j_0}^{(t)} = \mathbb{I}_{j_0}$, it leads to the decorrelated score function.

Finally, we may use a linear regression model to approximate $\mathrm{E}(X_{t,j_0}|\boldsymbol{X}_{t,\mathcal{M}})$ for any

$\mathcal{M} \subseteq \mathcal{M}_{j_0}$. This leads to its linear approximation $\boldsymbol{\omega}_{\mathcal{M},j_0}^T \boldsymbol{X}_{t,\mathcal{M}}$. The regression coefficients $\boldsymbol{\omega}_{\mathcal{M},j_0} = \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1} \boldsymbol{\Sigma}_{\mathcal{M},j_0}$ can be estimated by plugging the estimators $\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}$, $\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}$ for $\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}$ and $\boldsymbol{\Sigma}_{\mathcal{M},j_0}$. The estimating equation in (3) results in a root n consistent estimator regardless of whether the linear approximation is valid or not.

We can summarize our procedures in the following algorithm.

Step 1. Input $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ and an integer $1 < s_n < n$.

Step 2. Compute an initial estimator $\widetilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$, based on $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$.

Step 3. For $t = s_n, s_n + 1, \ldots, n-1$,

(i) Estimate $\mathcal{M}_{j_0}$ via some model selection procedure based on the sub-dataset $\mathcal{F}_t = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_t, Y_t)\}$. Denoted by $\widehat{\mathcal{M}}_{j_0}^{(t)}$ the corresponding estimator. We require $|\widehat{\mathcal{M}}_{j_0}^{(t)}| < n$, $j_0 \notin \widehat{\mathcal{M}}_{j_0}^{(t)}$.

(ii) Estimate $\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ by $\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} = \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},\widehat{\mathcal{M}}_{j_0}^{(t)}}^{-1} \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$.

(iii) Estimate $\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^2$ by $\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^2 = \widehat{\boldsymbol{\Sigma}}_{j_0,j_0} - \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$.

Step 4. Define $\bar{\beta}_{j_0}$ to be the solution to the following equation,

$$\sum_{t=s_n}^{n-1} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \widehat{Z}_{t+1,j_0} (Y_{t+1} - X_{t+1,j_0} \beta_{0,j_0} - \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}) = 0, \tag{4}$$

where $\widehat{Z}_{t+1,j_0} = X_{t+1,j_0} - \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}$.

Due to its nature, (4) is referred to as online-score equation in order to distinguish it from the decorrelated score equation in Ning and Liu (2017). Step 3 essentially is to recursively calculate $\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ and $\widehat{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ for Step 4. Thus, we refer this algorithm to as recursive online-score estimation (ROSE) algorithm.

Let

$$\Gamma_n = \frac{1}{n - s_n} \sum_{t=s_n}^{n-1} \frac{X_{t+1,j_0}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right).$$

Under certain conditions, we can show that

$$\sqrt{n-s_n}\Gamma_n(\bar{\beta}_{j_0}-\beta_{0,j_0}) = \frac{1}{\sqrt{n-s_n}}\sum_{t=s_n}^{n-1}\frac{\varepsilon_{t+1}}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\left(X_{t+1,j_0}-\boldsymbol{\omega}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)+o_p(1). \quad (5)$$

The first term on the right-hand-side (RHS) of (5) corresponds to a mean zero martingale with respect to the filtration $\{\sigma(\mathcal{F}_t) : t \geq s_n\}$ where $\sigma(\mathcal{F}_t)$ denotes the σ-algebra generated by $\mathcal{F}_t$. Note that

$$\mathrm{E}\left[\left\{\frac{1}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\left(X_{t+1,j_0}-\boldsymbol{\omega}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)\varepsilon_{t+1}\right\}^2 \middle| \mathcal{F}_t\right] = 1.$$

By the martingale central limit theorem, we have as $n - s_n \to \infty$,

$$\sqrt{n-s_n}\Gamma_n(\bar{\beta}_{j_0}-\beta_{0,j_0}) \xrightarrow{d} N(0,\sigma_0^2).$$

Therefore, a two-sided $1-\alpha$ CI for β_{0,j_0} is given by

$$\bar{\beta}_{j_0} \pm z_{\frac{\alpha}{2}}\frac{\Gamma_n^{-1}}{\sqrt{n-s_n}}\hat{\sigma}, \quad (6)$$

where $\hat{\sigma}$ is some consistent estimator for σ_0.

2.2 Refinements

The CI in (6) is asymptotically valid. However, it has one drawback. Its length is equal to

$$2z_{\frac{\alpha}{2}}\frac{\Gamma_n^{-1}}{\sqrt{n-s_n}}\hat{\sigma}. \quad (7)$$

In general, (7) increases as s_n increases. Nonetheless, s_n should be large enough to guarantee the sure screening property of $\widehat{\mathcal{M}}_{j_0}^{(s_n)}$. For small n, this will result in a large CI. To address these concerns, we propose the following refined estimator. The estimation procedure is described below.

Step 1. Input $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ and an integer $1 < s_n < n$.

Step 2. Compute an initial estimator $\widetilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$, based on $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$.

Step 3. Compute $\widehat{\mathcal{M}}_{j_0}^{(t)}$, $\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$, $\hat{\sigma}^2_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ for $t = s_n, \dots, n-1$ as described in Section 2.1.

Step 4. Estimate $\mathcal{M}_{j_0}$ based on the sub-dataset $\{(\boldsymbol{X}_{s_n+1}, Y_{s_n+1}), \dots, (\boldsymbol{X}_n, Y_n)\}$. The resulting estimator $\widehat{\mathcal{M}}_{j_0}^{(-s_n)}$ shall satisfy $|\widehat{\mathcal{M}}_{j_0}^{(-s_n)}| < n$, $j_0 \notin \widehat{\mathcal{M}}_{j_0}^{(-s_n)}$.

Step 5. Estimate $\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}$ by $\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0} = \widehat{\boldsymbol{\Sigma}}^{-1}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}$.

Step 6. Estimate $\sigma^2_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}$ by $\hat{\sigma}^2_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0} = \widehat{\boldsymbol{\Sigma}}_{j_0,j_0} - \widehat{\boldsymbol{\Sigma}}^T_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0} \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}$.

Step 7. Define $\hat{\beta}_{j_0}$ to be the solution to the following equation,

$$\sum_{t=0}^{s_n-1} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}} \widehat{Z}_{t+1,j_0}(Y_{t+1} - X_{t+1,j_0}\beta_{0,j_0} - \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}})$$
$$+ \sum_{t=s_n}^{n-1} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \widehat{Z}_{t+1,j_0}(Y_{t+1} - X_{t+1,j_0}\beta_{0,j_0} - \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}) = 0,$$

where $\widehat{Z}_{t+1,j_0} = X_{t+1,j_0} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}$ for $t = s_n, \dots, n-1$ and $\widehat{Z}_{t+1,j_0} = X_{t+1,j_0} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}}$ for $t = 0, \dots, s_n - 1$.

When $s_n = o(n)$, the first s_n terms in the estimating equation in Step 7 is negligible. As a result, $\hat{\beta}_{j_0}$ is asymptotically the same as $\bar{\beta}_{j_0}$. Define

$$\Gamma_n^* = \frac{1}{n} \left(\sum_{t=0}^{s_n-1} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} X_{t+1,j_0} \widehat{Z}_{t+1,j_0} + \sum_{t=s_n}^{n-1} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}} X_{t+1,j_0} \widehat{Z}_{t+1,j_0} \right). \tag{8}$$

Below, we prove

$$\hat{\beta}_{j_0} \pm z_{\frac{\alpha}{2}} \frac{\Gamma_n^{*-1}}{\sqrt{n}} \hat{\sigma}, \tag{9}$$

is a valid two-sided CI for β_{0,j_0}. We need the following conditions.

(A1) Assume $\widehat{\mathcal{M}}_{j_0}^{(n)}$ satisfies $\Pr(|\widehat{\mathcal{M}}_{j_0}^{(n)}| \le \kappa_n) = 1$ for some $1 \le \kappa_n = o(n)$. Besides,

$$\Pr\left(\mathcal{M}_{j_0} \subseteq \widehat{\mathcal{M}}_{j_0}^{(n)}\right) \ge 1 - O\left(\frac{1}{n^{\alpha_0}}\right),$$

for some constant $\alpha_0 > 1$.

(A2) Assume there exists some constant $\bar{c} > 0$ such that for any $\mathcal{M} \subseteq \mathbb{I}$ and $|\mathcal{M}| \le \kappa_n$, $\lambda_{\min}(\boldsymbol{\Sigma}_{j_0 \cup \mathcal{M}, j_0 \cup \mathcal{M}}) \ge \bar{c}$.

(A3) Assume there exists some constant $c_0 > 0$ such that $\|\boldsymbol{X}_0^T \boldsymbol{a}\|_{\psi_2} \le c_0 \|\boldsymbol{a}\|_2$ for any $\boldsymbol{a} \in \mathbb{R}^p$.

(A4) Assume (i) $\Pr(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \le \eta_n) \to 1$ for some $\eta_n > 0$; (ii) $\eta_n \sqrt{\kappa_n \log p} = o(1)$; (iii) $\Pr(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{M}_0^c} - \boldsymbol{\beta}_{0,\mathcal{M}_0^c}\|_1 \le k_0 \|\widetilde{\boldsymbol{\beta}}_{\mathcal{M}_0} - \boldsymbol{\beta}_{0,\mathcal{M}_0}\|_1) \to 1$ for some constant $k_0 > 0$, where $\mathcal{M}_0$ stands for the support of $\boldsymbol{\beta}_0$ and $\mathcal{M}_0^c$ denotes its complement.

(A5) Assume $\hat{\sigma} \overset{P}{\to} \sigma_0$.

Assumption (A1) essentially requires the sure screening property of the procedure for obtaining $\widehat{\mathcal{M}}_{j_0}^{(n)}$. Typically conditions to guarantee the sure screening property are weaker than those for the oracle property. Assume (A1) holds and $s_n \to \infty$. Then it follows from Bonferroni's inequality that

$$\Pr\left(\mathcal{M}_{j_0} \subseteq \bigcap_{t=s_n}^{n-1} \widehat{\mathcal{M}}_{j_0}^{(t+1)}\right) \ge 1 - O\left(\sum_{t=s_n}^{\infty} \frac{1}{t^{\alpha_0}}\right) \to 1.$$

Hence, all the selected models possess the sure screening property with probability tending to 1.

When $\mathcal{M}_{j_0}$ is estimated via SIS, we can show (A1) holds for any arbitrary $\alpha_0 > 1$ (see Theorem 1 in Fan and Lv, 2008). The validity of such sure screening property typically relies on certain minimum-signal-strength conditions on $\boldsymbol{\beta}_{0,\mathbb{I}_{j_0}}$. A counterexample is given in Section B.1.1 of the supplementary article where we show our CI is no longer valid when these conditions are violated. We note that van de Geer et al. (2014) and Ning and Liu (2017) do not require these conditions. However, these authors impose some additional assumptions on the design matrix. We discuss this further in Section B.1 of the supplementary article. Moreover, in Section 5.4, we present a variant of our method that is valid without the minimal-signal-strength conditions.

Condition (A2) is similar to the restricted eigenvalue condition (Bickel et al., 2009) imposed to derive the oracle inequalities for the Lasso estimator and the Dantzig selector. Condition (A3) requires $\boldsymbol{X}_0$ to be a sub-Gaussian vector. This condition is used in Ning and Liu (2017) and van de Geer et al. (2014) as well. See Section 4.1 of Ning and Liu (2017) and Condition (B1) in van de Geer et al. (2014) for details.

When $\widetilde{\boldsymbol{\beta}}$ is estimated via the Lasso or the Dantzig selector, then the first part of Condition (A4) holds with $\eta_n = c_n\sqrt{s^* \log p/n}$ where c_n is an arbitrary diverging sequence and s^* is the number of nonzero elements in $\boldsymbol{\beta}_0$. The second part holds as long as $\kappa_n s^* \log^2 p = o(n)$. Under (A1), we have $\kappa_n \geq s^* - 1$. This further implies $(s^*)^2 \log^2 p = o(n)$. Such a sample size requirement is consistent with those in van de Geer et al. (2014) and Ning and Liu (2017). See Condition (B2) of van de Geer et al. (2014), and Corollary 4.1 in Ning and Liu (2017) for details. The last condition in (A4) holds with $k_0 = 3$ for the Lasso estimator, $k_0 = 1$ for the Dantzig selector and $k_0 = 0$ for the non-convex penalized regression estimator (when the "oracle property" is achieved).

Condition (A5) holds when $\hat{\sigma}$ is computed by refitted cross-validation (Fan et al., 2012) or scaled lasso (Sun and Zhang, 2013). In Section B.3 of the supplementary article, we further introduce a simple plug-in estimator for σ_0^2 based on $\widetilde{\boldsymbol{\beta}}$ and show (A5) holds under (A3), (A4) and the conditions that $\log p = O(n^{2/3})$, $\mathrm{E}|\varepsilon_0|^3 = O(1)$. The last moment condition is also needed in the following theorem to guarantee the asymptotic normality of $\hat{\beta}_{j_0}$.

Theorem 2.1. *Under Conditions (A1)-(A5), assume $s_n \to \infty$, $s_n = o(n)$, $\kappa_n^2 \log p = O(n/\log^2 n)$ and $E|\varepsilon_0|^3 = O(1)$. Then, we have*

$$\frac{\sqrt{n}\Gamma_n^*(\hat{\beta}_{j_0} - \beta_{0,j_0})}{\hat{\sigma}} \xrightarrow{d} N(0,1),$$

where Γ_n^* is defined in (8).

3 High-dimensional generalized linear models

3.1 Estimation and inference

Suppose that $(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)$ is a random sample from $(\boldsymbol{X}_0, Y_0)$ in (1). The function $b(\cdot)$ is assumed to be thrice continuously differentiable. We further assume $b''(\cdot) > 0$ and $b'''(\cdot)$ is Lipschitz continuous. Denoted by $\mu(\cdot)$ the derivative of $b(\cdot)$. As in Section 2, our focus is to construct a CI for β_{0,j_0}. Let $\mathcal{M}_{j_0} = \{j \neq j_0 : \beta_{0,j} \neq 0\}$ and $\boldsymbol{\Sigma} = \mathrm{E}\boldsymbol{X}_0 b''(\boldsymbol{X}_0^T \boldsymbol{\beta}_0)\boldsymbol{X}_0^T$, we describe our estimating procedure below.

Step 1. Input $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ and an integer $1 < s_n < n$.

Step 2. Compute an initial estimator $\widetilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$. Compute

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i b''(\boldsymbol{X}_i^T \widetilde{\boldsymbol{\beta}})\boldsymbol{X}_i^T. \tag{10}$$

Step 3. For $t = s_n, s_n + 1, \dots, n-1$, estimate $\mathcal{M}_{j_0}$ based on the sub-dataset $\mathcal{F}_t = \{(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_t, Y_t)\}$. Denoted by $\widehat{\mathcal{M}}_{j_0}^{(t)}$ the corresponding estimator. We require $|\widehat{\mathcal{M}}_{j_0}^{(t)}| \le n$, $j_0 \notin \widehat{\mathcal{M}}_{j_0}^{(t)}$. Compute

$$\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0} = \widehat{\boldsymbol{\Sigma}}^{-1}_{\widehat{\mathcal{M}}_{j_0}^{(t)}, \widehat{\mathcal{M}}_{j_0}^{(t)}} \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0} \quad \text{and} \quad \hat{\sigma}^2_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0} = \widehat{\boldsymbol{\Sigma}}_{j_0, j_0} - \widehat{\boldsymbol{\Sigma}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0} \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0}.$$

Step 4. Estimate $\mathcal{M}_{j_0}$ based on the sub-dataset $\{(\boldsymbol{X}_{s_n+1}, Y_{s_n+1}), \dots, (\boldsymbol{X}_n, Y_n)\}$. Denoted by $\widehat{\mathcal{M}}_{j_0}^{(-s_n)}$ the resulting estimator. We require $|\widehat{\mathcal{M}}_{j_0}^{(-s_n)}| \le n$, $j_0 \notin \widehat{\mathcal{M}}_{j_0}^{(-s_n)}$. Compute

$$\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} = \widehat{\boldsymbol{\Sigma}}^{-1}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, \widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} \quad \text{and} \quad \hat{\sigma}^2_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} = \widehat{\boldsymbol{\Sigma}}_{j_0, j_0} - \widehat{\boldsymbol{\Sigma}}^T_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0}.$$

Step 5. Define $\hat{\beta}_{j_0}$ to be the solution to the following equation,

$$\sum_{t=0}^{s_n-1} \frac{\widehat{Z}_{t+1,j_0}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}} \left\{ Y_{t+1} - \mu \left(X_{t+1,j_0}\beta_{0,j_0} + \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \right) \right\}$$
$$+ \sum_{t=s_n}^{n-1} \frac{\widehat{Z}_{t+1,j_0}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left\{ Y_{t+1} - \mu \left(X_{t+1,j_0}\beta_{0,j_0} + \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right\} = 0,$$

where $\widehat{Z}_{t+1,j_0} = X_{t+1,j_0} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}}$ for $t = 0, \dots, s_n - 1$.

The estimating function in Step 5 can be solved via the Newton-Raphson method with the initial value $\hat{\beta}_{j_0}^{(0)} = \widetilde{\beta}_{j_0}$. More specifically, for $l = 1, 2, \dots$, we can iteratively update $\hat{\beta}_{j_0}$ by

$$\hat{\beta}_{j_0}^{(l)} = \hat{\beta}_{j_0}^{(l-1)} + \frac{\sum_{t=0}^{n-1} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \widehat{Z}_{t+1,j_0} \left\{ Y_{t+1} - \mu \left(X_{t+1,j_0}\hat{\beta}_{j_0}^{(l-1)} + \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \right) \right\}}{\sum_{t=0}^{n-1} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \widehat{Z}_{t+1,j_0} X_{t+1,j_0} b'' \left(X_{t+1,j_0}\hat{\beta}_{j_0}^{(l-1)} + \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} \right)}, \quad (11)$$

where we use a shorthand and write $\widehat{\mathcal{M}}_{j_0}^{(t)} = \widehat{\mathcal{M}}_{j_0}^{(-s_n)}$, for $t = 0, \dots, s_n - 1$. Define

$$\Gamma_n^{*,(l-1)} = \frac{1}{n} \sum_{t=0}^{n-1} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \widehat{Z}_{t+1,j_0} X_{t+1,j_0} b'' \left(X_{t+1,j_0}\hat{\beta}_{j_0}^{(l-1)} + \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} \right).$$

A two-sided $1 - \alpha$ CI for β_{0,j_0} is given by

$$\hat{\beta}_{j_0}^{(l)} \pm \frac{z_{\frac{\alpha}{2}} \hat{\phi}^{1/2}}{\sqrt{n}\Gamma_n^{*,(l-1)}}, \quad (12)$$

where $\hat{\phi}$ is some consistent estimator for ϕ_0. We state the following conditions.

(A1*) Assume $\widehat{\mathcal{M}}_{j_0}^{(n)}$ satisfies $\Pr(|\widehat{\mathcal{M}}_{j_0}^{(n)}| \le \kappa_n) = 1$ for some $1 \le \kappa_n = o(n)$. Besides, there exists some constant $\alpha_0 > 1$ such that

$$\Pr\left(\mathcal{M}_{j_0} \subseteq \widehat{\mathcal{M}}_{j_0}^{(n)} \right) \ge 1 - O\left(\frac{1}{n^{\alpha_0}} \right),$$

(A2*) Assume there exists some constant $\bar{c} > 0$ such that for any $\mathcal{M} \subseteq \mathbb{I}$ and $|\mathcal{M}| \le \kappa_n$, $\lambda_{\min}(\boldsymbol{\Sigma}_{j_0 \cup \mathcal{M}, j_0 \cup \mathcal{M}}) \ge \bar{c}$.

(A3*) Assume there exists some constant $c_0 > 0$ such that $\|\boldsymbol{X}_0^T \boldsymbol{a}\|_{\psi_2} \le c_0 \|\boldsymbol{a}\|_2$ for any $\boldsymbol{a} \in \mathbb{R}^p$.

(A4*) Assume $\max_{j \in [1,\dots,p]} |X_{0,j}| \le \omega_0$ for some constant $\omega_0 > 0$. Assume $|\boldsymbol{X}_0^T \boldsymbol{\beta}_0| \le \bar{\omega}$ for some constant $\bar{\omega} > 0$.

(A5*) Assume (i) $\Pr(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \le \eta_n) \to 1$ for some $\eta_n > 0$; (ii) $\eta_n \sqrt{\kappa_n \log p} = o(1)$ and $\sqrt{n}\eta_n^2 = o(1)$; (iii) $\Pr(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{M}_0^c} - \boldsymbol{\beta}_{0,\mathcal{M}_0^c}\|_1 \le k_0 \|\widetilde{\boldsymbol{\beta}}_{\mathcal{M}_0} - \boldsymbol{\beta}_{0,\mathcal{M}_0}\|_1) \to 1$ for some constant $k_0 > 0$.

(A6*) Assume $\|Y_0 - \mu(\boldsymbol{X}_0^T \boldsymbol{\beta}_0)\|_{\psi_1|\boldsymbol{X}_0}$ is uniformly bounded for all $\boldsymbol{X}_0$, where $\|\cdot\|_{\psi_1|\boldsymbol{X}_0}$ denotes the Orlicz norm conditional on $\boldsymbol{X}_0$.

(A7*) Assume $\hat{\phi} \xrightarrow{P} \phi_0$.

Conditions (A1*)-(A3*) are very similar to (A1)-(A3). In (A4*), for technical convenience, we assume $X_{0,j}$'s and $\boldsymbol{X}_0^T \boldsymbol{\beta}_0$ are bounded. In (A5*), we further assume $\sqrt{n}\eta_n^2 = o(1)$. Note that such assumption doesn't appear in (A4). This is because we focus on a more general class of models here. Assume (A4*) holds. Then (A6*) is automatically satisfied for logistic and Poisson regression models. In logistic or Poisson regression models, we have $\phi_0 = 1$. Condition (A7*) thus automatically holds by setting $\hat{\phi} = 1$.

Theorem 3.1. *Assume (A1*)-(A7*) hold. Assume $s_n \to \infty$, $s_n = o(n)$, $\kappa_n^{5/2} \log p = O(n/\log^2 n)$ and $\kappa_n^3 = O(n)$. Then, for any fixed $l \ge 1$, we have*

$$\frac{\sqrt{n}\Gamma_n^{*,(l-1)}}{\hat{\phi}^{1/2}} (\hat{\beta}_{j_0}^{(l)} - \beta_{0,j_0}) \xrightarrow{d} N(0,1).$$

Theorem 3.1 proves the validity of the two-sided CI in (12), for any $l \ge 1$. When $l = 1$, $\hat{\beta}_{j_0}^{(l)}$ corresponds to the solution of the first-order approximation of the score equation. We note that Bickel (1975) and Ning and Liu (2017) used a similar one-step approximation to ensure the consistency of the resulting estimator. In practice, we can update $\hat{\beta}_{j_0}^{(l)}$ for a few Newton steps. In our numerical experiments, we find that $\hat{\beta}_{j_0}^{(l)}$ converges very fast and it suffices to set $l = 3, 4$ or 5.

3.2 Asymptotic efficiency

Theorem 3.1 proves the validity of the CI in (12). The length of the CI is given by

$$\mathrm{L}(\hat{\beta}_{j_0}^{(l)}, \alpha) = 2z_{\alpha/2}\frac{\hat{\phi}^{1/2}}{\Gamma_n^{*,(l-1)}\sqrt{n}}. \tag{13}$$

Under the given conditions in Theorem 3.1, it follows from the law of large numbers for martingales (Csörgö, 1968) that

$$\Gamma_n^{*,(l-1)} = \frac{s_n}{n}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0} + \frac{1}{n}\left(\sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\right) + o_p(1), \tag{14}$$

where

$$\sigma^2_{\mathcal{M},j_0} = \boldsymbol{\Sigma}_{j_0,j_0} - \boldsymbol{\Sigma}^T_{\mathcal{M},j_0}\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},j_0},$$

for any $\mathcal{M} \subseteq \mathbb{I}_{j_0}$.

By Assumption (A1*) and (A2*), we have almost surely,

$$\sigma^2_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0} \geq \bar{c} \ \text{ and } \ \sigma^2_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \geq \bar{c}, \quad \forall t = s_n, \ldots, n-1. \tag{15}$$

Under (A7*), $\hat{\phi}$ is consistent. This together with (13)-(15) yields

$$\sqrt{n}\mathrm{L}(\hat{\beta}_{j_0}, \alpha) = \frac{2z_{\alpha/2}\phi_0^{1/2}}{s_n\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}/n + \sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}/n} + o_p(1). \tag{16}$$

Based on (16), we compare the length of the CI of the proposed method with the desparsified Lasso method, the decorrelated score method and the "oracle" method below.

3.2.1 Comparison with the de-sparsified Lasso and the decorrelated score

Consider the Lasso estimator

$$\widehat{\boldsymbol{\beta}}^L = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{n} \sum_{i=1}^n \{ b(\boldsymbol{X}_i^T \boldsymbol{\beta}) - Y_i \boldsymbol{X}_i^T \boldsymbol{\beta} \} + \lambda_n \|\boldsymbol{\beta}\|_1 \right).$$

The de-sparsified Lasso estimator is defined by

$$\widehat{\boldsymbol{\beta}}^{DL} = \widehat{\boldsymbol{\beta}}^L + \widehat{\boldsymbol{\Theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\boldsymbol{X}_i^T \{ Y_i - \mu(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}^L) \} \right) \right\},$$

where the matrix $\widehat{\boldsymbol{\Theta}}$ is computed by the nodewise Lasso (see Section 3.1.1 in van de Geer et al., 2014). Theorem 3.1 in van de Geer et al. (2014) proved that

$$\sqrt{n}(\hat{\beta}_{j_0}^{DL} - \beta_{0,j_0}) / \sqrt{\boldsymbol{e}_{j_0,p}^T \widehat{\boldsymbol{\Omega}} \boldsymbol{e}_{j_0,p}} \sim N(0,1), \tag{17}$$

where

$$\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Theta}} \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i \{ Y_i - \mu(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}^L) \}^2 \boldsymbol{X}_i^T \right) \widehat{\boldsymbol{\Theta}}^T,$$

and

$$\boldsymbol{e}_{j_1,j_2} = (\underbrace{0,\dots,0}_{j_1-1}, 1, \underbrace{0,\dots,0}_{j_2-j_1}).$$

for any integer $1 \le j_1 < j_2$.

Based on the de-sparsified Lasso estimator, the corresponding CI for β_{0,j_0} is given by

$$\left[\hat{\beta}_{j_0}^{DL} - z_{\frac{\alpha}{2}} \frac{\sqrt{\boldsymbol{e}_{j_0,p}^T \widehat{\boldsymbol{\Omega}} \boldsymbol{e}_{j_0,p}}}{\sqrt{n}}, \hat{\beta}_{j_0}^{DL} + z_{\frac{\alpha}{2}} \frac{\sqrt{\boldsymbol{e}_{j_0,p}^T \widehat{\boldsymbol{\Omega}} \boldsymbol{e}_{j_0,p}}}{\sqrt{n}} \right], \tag{18}$$

Moreover, it follows from Theorem 3.2 in van de Geer et al. (2014) that

$$\boldsymbol{e}_{j_0,p}^T \widehat{\boldsymbol{\Omega}} \boldsymbol{e}_{j_0,p} = \boldsymbol{e}_{j_0,p}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_{j_0,p} \phi_0 + o_p(1).$$

Therefore, the length of (18) satisfies

$$\sqrt{n} L(\hat{\beta}_{j_0}^{DL}, \alpha) = 2 z_{\alpha/2} \phi_0^{1/2} \sqrt{\boldsymbol{e}_{j_0,p}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_{j_0,p}} + o_p(1). \tag{19}$$

Ning and Liu (2017) proposed to construct the CI for high-dimensional parameters in GLM based on the one-step estimator that solves a first-order approximation of the decorrelated score equation. Specifically, the one-step estimator is given by

$$\hat{\beta}_{j_0}^{DS} = \tilde{\beta}_{j_0} + \frac{\sum_{t=0}^{n-1} (X_{t+1,j_0} - \widehat{\boldsymbol{w}}^T \boldsymbol{X}_{t+1,\mathbb{I}_{j_0}}) \left\{ Y_{t+1} - \mu\left(\boldsymbol{X}_{t+1}^T \widetilde{\boldsymbol{\beta}} \right) \right\}}{\sum_{t=0}^{n-1} (X_{t+1,j_0} - \widehat{\boldsymbol{w}}^T \boldsymbol{X}_{t+1,\mathbb{I}_{j_0}}) X_{t+1,j_0} b''\left(\boldsymbol{X}_{t+1}^T \widetilde{\boldsymbol{\beta}} \right)}, \tag{20}$$

where $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{w}}$ are some consistent estimators for $\boldsymbol{\beta}_0$ and $\boldsymbol{\Sigma}_{\mathbb{I}_{j_0},\mathbb{I}_{j_0}}^{-1} \boldsymbol{\Sigma}_{\mathbb{I}_{j_0},j_0}$, respectively. The corresponding CI is given by

$$\left[\hat{\beta}_{j_0}^{DS} - z_{\frac{\alpha}{2}} \hat{\phi}^{1/2} (\hat{\sigma}_s)^{-1/2}, \hat{\beta}_{j_0}^{DS} + z_{\frac{\alpha}{2}} \hat{\phi}^{1/2} (\hat{\sigma}_s)^{-1/2} \right],$$

where $\hat{\sigma}_s$ is the denominator of the second term on the RHS of (20) and $\hat{\phi}$ is some consistent estimator for ϕ_0. Under certain conditions, we can show

$$\sqrt{n} L(\hat{\beta}_{j_0}^{DS}, \alpha) = \frac{2 z_{\frac{\alpha}{2}} \phi_0^{1/2}}{\sigma_{\mathbb{I}_{j_0},j_0}} + o_p(1) = 2 z_{\frac{\alpha}{2}} \phi_0^{1/2} \sqrt{\boldsymbol{e}_{j_0,p}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_{j_0,p}} + o_p(1), \tag{21}$$

where the last equality follows from the matrix inversion formula (see Lemma A.5 in the supplementary material). This together with (19) implies that the lengths of CIs based on $\hat{\beta}_{j_0}^{DL}$ and $\hat{\beta}_{j_0}^{DS}$ are asymptotically the equivalent.

For any $\mathcal{M} \subseteq \mathbb{I}_{j_0}$, let

$$
\begin{aligned}
\boldsymbol{\xi}_{\mathcal{M},j_0} &= \mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}^T \boldsymbol{X}_{0,\mathcal{M}}) b''(\boldsymbol{X}_i^T \boldsymbol{\beta}_0)(\boldsymbol{X}_{0,(\mathcal{M}\cup\{j_0\})^c} - \boldsymbol{\Sigma}_{(\mathcal{M}\cup\{j_0\})^c,\mathcal{M}} \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1} \boldsymbol{X}_{0,\mathcal{M}}) \\
&= \boldsymbol{\Sigma}_{(\mathcal{M}\cup\{j_0\})^c,j_0} - \boldsymbol{\Sigma}_{(\mathcal{M}\cup\{j_0\})^c,\mathcal{M}} \boldsymbol{\omega}_{\mathcal{M},j_0}.
\end{aligned}
$$

We have the following results.

Theorem 3.2. *Assume* (16), (19), (21), *(A3*) and (A4*) hold. Let* $\bar{k} = \sup_{|z|\le\bar{\omega}} b''(z)$. *Then for any* $0 < \alpha < 1$, $l \geq 1$,

$$
\begin{aligned}
& \sqrt{n} L(\hat{\beta}_{j_0}^{DL}, \alpha) = \sqrt{n} L(\hat{\beta}_{j_0}^{DS}, \alpha) + o_p(1) \\
\geq\; & \sqrt{n} L(\hat{\beta}_{j_0}^{(l)}, \alpha) + \frac{\phi_0^{1/2} z_{\alpha/2}}{\bar{k}^{3/2} c_0^5} \left(\frac{s_n}{n} \|\boldsymbol{\xi}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}\|_2^2 + \frac{1}{n} \sum_{t=s_n}^{n-1} \|\boldsymbol{\xi}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\|_2^2 \right) + o_p(1).
\end{aligned}
$$

Theorem 3.2 implies that the proposed CI is asymptotically shorter than those based on the de-sparsified Lasso and the decorrelated score statistic. The difference depends on the L_2 norm of $\boldsymbol{\xi}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$, which measures the partial dependence between X_{0,j_0} and $\boldsymbol{X}_{0,(\widehat{\mathcal{M}}_{j_0}^{(t)}\cup j_0)^c}$, after adjusted by $\boldsymbol{X}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}$. For linear regression models, we have $\boldsymbol{\xi}_{\mathcal{M},j_0} = 0$ when X_{0,j_0} is independent of other predictors. However, $\|\boldsymbol{\xi}_{\mathcal{M},j_0}\|_2$ can be positive when X_{0,j_0} is partially correlated with $\boldsymbol{X}_{0,(\mathcal{M}\cup j_0)^c}$ given $\boldsymbol{X}_{0,\mathcal{M}}$.

Although our method yields narrower CI on average, its validity relies on certain minimal-signal-strength conditions on $\boldsymbol{\beta}_{0,\mathbb{I}_{j_0}}$, as discussed in Section 2.2. This is a potential disadvantage of our method. Moreover, our procedure can be more time consuming than the existing methods, as it requires to recursively estimate the support set based on different data subsets. A variant of our method is proposed in Section 3.3 to reduce the computational cost.

3.2.2 Comparison with the oracle method

We compare the proposed CI with the CI of the oracle method. The oracle knew the set $\mathcal{M}_{j_0}$ ahead of time. It estimates β_{0,j_0} by $\hat{\beta}^{oracle}_{j_0}$ defined as

$$(\hat{\beta}^{oracle}_{j_0}, \hat{\boldsymbol{\beta}}^{oracle}_{\mathcal{M}_{j_0}}) = \arg\min_{(\beta_{j_0}, \boldsymbol{\beta}_{\mathcal{M}_{j_0}})} \frac{1}{n}\sum_{i=1}^{n} \left(b(X_{i,j_0}\beta_{j_0} + \boldsymbol{X}^T_{i,\mathcal{M}_{j_0}}\boldsymbol{\beta}_{\mathcal{M}_{j_0}}) - Y_i(X_{i,j_0}\beta_{j_0} + \boldsymbol{X}^T_{i,\mathcal{M}_{j_0}}\boldsymbol{\beta}_{\mathcal{M}_{j_0}})\right).$$

Let

$$\widehat{\boldsymbol{\Sigma}}^{oracle} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i b''(X_{i,j_0}\hat{\beta}^{oracle}_{j_0} + \boldsymbol{X}^T_{i,\mathcal{M}_{j_0}}\hat{\boldsymbol{\beta}}^{oracle}_{\mathcal{M}_{j_0}})\boldsymbol{X}_i.$$

The asymptotic variance of $\sqrt{n}\hat{\beta}^{oracle}_{j_0}$ can be consistently estimated by

$$\hat{\phi}\boldsymbol{e}^T_{1,|\mathcal{M}_{j_0}|+1} \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^{oracle}_{j_0,j_0} & \widehat{\boldsymbol{\Sigma}}^{oracle}_{j_0,\mathcal{M}_{j_0}} \\ \widehat{\boldsymbol{\Sigma}}^{oracle}_{\mathcal{M}_{j_0},j_0} & \widehat{\boldsymbol{\Sigma}}^{oracle}_{\mathcal{M}_{j_0},\mathcal{M}_{j_0}} \end{pmatrix}^{-1} \boldsymbol{e}_{1,|\mathcal{M}_{j_0}|+1} = \hat{\phi}\left\{\widehat{\boldsymbol{\Sigma}}^{oracle}_{j_0,j_0} - \widehat{\boldsymbol{\Sigma}}^{oracle}_{j_0,\mathcal{M}_{j_0}}\left(\widehat{\boldsymbol{\Sigma}}^{oracle}_{\mathcal{M}_{j_0},\mathcal{M}_{j_0}}\right)^{-1}\widehat{\boldsymbol{\Sigma}}^{oracle}_{\mathcal{M}_{j_0},j_0}\right\}^{-1},$$

where the equality follows by the matrix inversion formula (see Lemma A.5). Let

$$\hat{\sigma}^{oracle}_{\mathcal{M}_{j_0},j_0} = \sqrt{\widehat{\boldsymbol{\Sigma}}^{oracle}_{j_0,j_0} - \widehat{\boldsymbol{\Sigma}}^{oracle}_{j_0,\mathcal{M}_{j_0}}\left(\widehat{\boldsymbol{\Sigma}}^{oracle}_{\mathcal{M}_{j_0},\mathcal{M}_{j_0}}\right)^{-1}\widehat{\boldsymbol{\Sigma}}^{oracle}_{\mathcal{M}_{j_0},j_0}}.$$

The corresponding confidence interval is given by

$$\left[\hat{\beta}^{oracle}_{j_0} - z_{\frac{\alpha}{2}}\hat{\phi}^{1/2}/(\sqrt{n}\hat{\sigma}^{oracle}_{\mathcal{M}_{j_0},j_0}), \hat{\beta}^{oracle}_{j_0} + z_{\frac{\alpha}{2}}\hat{\phi}^{1/2}/(\sqrt{n}\hat{\sigma}^{oracle}_{\mathcal{M}_{j_0},j_0})\right], \tag{22}$$

where $\hat{\phi}$ is some constant estimator for ϕ_0.

Under certain conditions, the length of (22) satisfies

$$\sqrt{n}\mathrm{L}(\hat{\beta}^{oracle}_{j_0}, \alpha) = \frac{2z_{\alpha/2}\phi_0^{1/2}}{\sigma_{\mathcal{M}_{j_0},j_0}} + o_p(1). \tag{23}$$

Theorem 3.3. *Assume* (16) *and* (23) *hold. Assume* $s_n \to \infty$, $n - s_n \to \infty$, *and there* *exists some* $\alpha_0 > 1$ *such that*

$$Pr\left(\mathcal{M}_{j_0} = \widehat{\mathcal{M}}_{j_0}^{(n)}\right) \geq 1 - O\left(\frac{1}{n^{\alpha_0}}\right). \tag{24}$$

Then for any $0 < \alpha < 1$, $l \geq 1$, *we have*

$$\sqrt{n}L(\hat{\beta}_{j_0}^{(l)}, \alpha) = \sqrt{n}L(\hat{\beta}_{j_0}^{oracle}, \alpha) + o_p(1).$$

Condition (24) in Theorem 3.3 requires the variable selection procedure to be consistent. Under this condition, we prove the "oracle" property of our method, which means that the length of the proposed CI is asymptotically equivalent to the CI of the oracle method.

3.3 Computationally efficient procedure

The proposed estimation procedure in Section 3.1 requires to estimate $\mathcal{M}_{j_0}$ $(n - s_n + 1)$ times. This will be time consuming for large n. To address this concern, for a given integer $S > 1$, we can compute $\widehat{\mathcal{M}}_{j_0}^{(t)}$ approximately $(n - s_n)/S$ times based on the sub-dataset $\mathcal{F}_t$ for $t = s_n, s_n + S, s_n + 2S, \ldots, s_n + \lfloor (n - 1 - s_n)/S \rfloor S$ where $\lfloor z \rfloor$ denotes the largest integer smaller than or equal to z. For any $s_n < t < n$, define

$$\widehat{\mathcal{M}}_{j_0}^{(t)} = \widehat{\mathcal{M}}_{j_0}^{(s_n + l_0 S)},$$

for some nonnegative integer l_0 such that $s_n + l_0 S \leq t < s_n + (l_0 + 1)S$. The resulting estimator $\hat{\beta}_{j_0}^{(l)}$ is computed by (11). The corresponding CI can be similarly derived as in (12).

4 Numerical examples

4.1 Linear regression

In this section, we conduct some simulation studies to examine the performance of the proposed CI in high dimensional linear regression models. Suppose that $\{\boldsymbol{X}_i, Y_i\}$, $i =$

$1, \ldots, n$ is a sample from the following model:

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}_0 + \varepsilon_i, \tag{25}$$

where $\varepsilon_i \sim N(0, 1)$, $\boldsymbol{X}_i \sim N(0, \boldsymbol{\Sigma})$.

Consider the following four settings: (A) $n = 100$, $\beta_{0,1} = \beta_{0,2} = 1.0$ and $\beta_{0,j} = 0$ for $j > 2$; (B) $n = 100$, $\beta_{0,1} = \beta_{0,2} = 2.0$ and $\beta_{0,j} = 0$ for $j > 2$; (C) $n = 200$, $\beta_{0,1} = 2.0$, $\beta_{0,2} = -2.0$ and $\beta_{0,j} = 0$ for $j > 2$; (D) $n = 200$, $\beta_{0,1} = \beta_{0,2} = \beta_{0,3} = \beta_{0,4} = \beta_{0,5} = 1.0$ and $\beta_{0,j} > 0$ for $j > 5$. For each setting, we set $p = 1000$, and consider two different covariance matrices $\boldsymbol{\Sigma}$, corresponding to $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $\boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$. This yields a total of 8 scenarios. For the first three settings, the objective is to construct 95% two-sided CIs for $\beta_{0,2}$ and $\beta_{0,3}$. For the last setting, we aim to construct 95% two-sided CIs for $\beta_{0,3}$, $\beta_{0,4}$, $\beta_{0,5}$ and $\beta_{0,6}$. Comparison is made among the following CIs:

(i) The proposed CI in (9), labeled by ROSE in Tables 1 and 2;

(ii) The CI constructed by the de-sparsified Lasso (DLASSO) method;

(iii) The CI constructed by the Bootstrap Lasso+Partial Ridge (BLPR) method (Liu et al., 2017);

(iv) The CI constructed by the simple sample-splitting (S3) method.

To calculate the CI in (9), we set $s_n = \lfloor 2n/\log(n) \rfloor$. Such a choice of s_n satisfies the conditions in Theorem 2.1. The set $\mathcal{M}_{j_0}$ is estimated by ISIS. The estimation procedure is implemented by the R package `SIS` (Saldana and Feng, 2016). To compute the initial estimator $\widetilde{\boldsymbol{\beta}}$, we first apply ISIS based on all observations and then fit a penalized linear regression model using the R package `ncvreg` (Breheny and Huang, 2011) with SCAD penalty function for the variables picked by ISIS. The variance estimator $\hat{\sigma}$ is computed by refitted cross-validation. We implement the CI in (ii) by the R package `hdi` (Dezeure et al., 2015). BLPR estimates β_{0,j_0} by the Lasso+Partial Ridge (LPR) estimator. More specifically, it first uses the Lasso to select important predictors and then refit the model using partial ridge regression based on the selected variables. The corresponding CI for β_{0,j_0} is constructed by bootstrapping the LPR estimator. We implement the BLPR method by the R package `HDCI`. To compute the CI in (iv), we randomly split the samples into two equal halves, use ISIS to estimate the support of control variables and construct the CI based on the remaining second half of the data. In Table 1 and 2, we report the empirical coverage probability (ECP) and average length (AL) of these CIs. Results are averaged over 500 simulations.

It can be seen from Table 1 that ECPs of our procedure and the S3 method are close to the nominal level in all cases. However, CIs constructed by the S3 method are approximately $\sqrt{2}$ times wider than our proposed method, according to Table 2. As commented in the introduction, this is because S3 only uses half of the samples to evaluate β_{0,j_0}.

Under the settings where $\boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$, ECPs of the DLASSO method are significantly smaller than the nominal level. For example, in Setting (A) and (B), ECPs of the DLASSO method are smaller than 90% when $\boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$. Under the settings where $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, CIs constructed by the DLASSO method have approximately nominal coverage probabilities. However, we note these CIs are wider than the proposed CIs in all cases. Take Setting (D) as an example. When $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, ALs of the DLASSO method are approximately 10% larger than the proposed method.

We note that BLPR yields very narrow CIs for zero parameters. For nonzero parameters however, the CIs based on the BLPR method are much wider than the proposed CIs in all cases. Moreover, under the settings where $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, ECPs of the BLPR method are significantly smaller than the nominal level for nearly all nonzero parameters.

4.2 Logistic regression

We generate $\{\boldsymbol{X}_i, Y_i\}_{i=1,\dots,n}$ from the following logistic regression model

$$\texttt{logit}\{\Pr(Y_i = 1|\boldsymbol{X}_i)\} = \boldsymbol{X}_i^T \boldsymbol{\beta}_0,$$

where $\texttt{logit}(z) = \log\{z/(1-z)\}$ for $0 < z < 1$.

We consider two settings: (A) $n = 500$, $\beta_{0,1} = 2.0$, $\beta_{0,2} = -2.0$ and $\beta_{0,j} = 0$ for $j > 2$; (B) $n = 600$, $\beta_{0,1} = \beta_{0,2} = \beta_{0,3} = \beta_{0,4} = \beta_{0,5} = 1.0$, $\beta_{0,j} = 0$ for $j > 6$. As in Section 4.1, we set $p = 1000$ and consider two different covariance matrices, $\boldsymbol{\Sigma} = \boldsymbol{I}$ and

Table 1: ECP (%) of the CIs with standard errors in parenthesis

Setting (A)		ROSE	DLASSO	BLPR	S3
$\boldsymbol{\Sigma} = \boldsymbol{I}_p$	β_2	93.0 (1.1)	94.0 (1.1)	83.0 (1.7)	94.0 (1.1)
	β_3	96.4 (0.8)	96.0 (0.9)	97.4 (0.7)	95.2 (1.0)
$\boldsymbol{\Sigma} = \{0.5^{\lvert i-j \rvert}\}_{i,j}$	β_2	93.6 (1.1)	89.0 (1.4)	92.0 (1.2)	94.6 (1.0)
	β_3	94.6 (1.0)	86.0 (1.6)	95.4 (0.9)	93.4 (1.1)
Setting (B)		ROSE	DLASSO	BLPR	S3
$\boldsymbol{\Sigma} = \boldsymbol{I}_p$	β_2	94.0 (1.1)	94.0 (1.1)	87.0 (1.5)	94.4 (1.0)
	β_3	96.8 (0.8)	96.0 (0.9)	97.4 (0.7)	95.2 (1.0)
$\boldsymbol{\Sigma} = \{0.5^{\lvert i-j \rvert}\}_{i,j}$	β_2	93.8 (1.1)	89.0 (1.4)	93.4 (1.1)	94.4 (1.0)
	β_3	95.6 (0.9)	85.6 (1.6)	96.8 (0.8)	95.2 (1.0)
Setting (C)		ROSE	DLASSO	BLPR	S3
$\boldsymbol{\Sigma} = \boldsymbol{I}_p$	β_2	95.6 (0.9)	95.8 (0.9)	94.6 (1.0)	93.8 (1.1)
	β_3	94.8 (1.0)	95.2 (1.0)	96.0 (0.9)	96.6 (0.8)
$\boldsymbol{\Sigma} = \{0.5^{\lvert i-j \rvert}\}_{i,j}$	β_2	94.8 (1.0)	76.4 (1.9)	90.8 (4.0)	96.4 (0.8)
	β_3	94.0 (1.1)	92.0 (1.2)	95.6 (0.9)	93.6 (1.1)
Setting (D)		ROSE	DLASSO	BLPR	S3
$\boldsymbol{\Sigma} = \boldsymbol{I}_p$	β_3	94.8 (1.0)	94.0 (1.1)	92.6 (1.2)	95.2 (1.0)
	β_4	93.8 (1.1)	93.4 (1.1)	91.0 (1.3)	93.8 (1.1)
	β_5	96.2 (0.9)	95.4 (0.9)	91.0 (1.3)	95.2 (1.0)
	β_6	94.4 (1.0)	95.2 (1.0)	95.0 (1.0)	95.6 (0.9)
$\boldsymbol{\Sigma} = \{0.5^{\lvert i-j \rvert}\}_{i,j}$	β_3	96.0 (0.9)	81.2 (1.7)	93.2 (1.1)	94.6 (1.0)
	β_4	93.4 (1.1)	82.6 (1.7)	94.8 (1.0)	93.8 (1.1)
	β_5	94.6 (1.0)	91.0 (1.2)	93.6 (1.1)	96.4 (0.8)
	β_6	93.8 (1.1)	91.6 (1.3)	95.0 (1.0)	95.4 (0.9)

Table 2: AL of the CIs with standard errors in parenthesis (numbers reported in the table are multiplied by 100)

Setting (A)		ROSE	DLASSO	BLPR	S3
$\boldsymbol{\Sigma} = \boldsymbol{I}_p$	β_2	42.2 (0.3)	45.4 (0.3)	88.5 (1.4)	63.1 (0.5)
	β_3	42.4 (0.3)	45.1 (0.2)	2.6 (0.1)	64.1 (0.5)
$\boldsymbol{\Sigma} = \{0.5^{\lvert i-j \rvert}\}_{i,j}$	β_2	48.0 (0.3)	47.5 (0.2)	106.2 (1.9)	72.4 (0.5)
	β_3	49.1 (0.3)	47.8 (0.2)	5.4 (0.3)	74.8 (0.6)
Setting (B)		ROSE	DLASSO	BLPR	S3
$\boldsymbol{\Sigma} = \boldsymbol{I}_p$	β_2	40.6 (0.2)	45.4 (0.2)	154.8 (3.4)	60.2 (0.5)
	β_3	40.7 (0.2)	45.1 (0.2)	3.4 (0.1)	60.5 (0.4)
$\boldsymbol{\Sigma} = \{0.5^{\lvert i-j \rvert}\}_{i,j}$	β_2	46.7 (0.2)	47.6 (0.2)	193.5 (4.4)	69.2 (0.5)
	β_3	47.6 (0.3)	47.8 (0.2)	6.8 (0.3)	70.9 (0.6)
Setting (C)		ROSE	DLASSO	BLPR	S3
$\boldsymbol{\Sigma} = \boldsymbol{I}_p$	β_2	27.9 (0.1)	30.1 (0.1)	142.8 (3.3)	40.0 (0.2)
	β_3	28.0 (0.1)	30.2 (0.1)	4.7 (0.1)	40.5 (0.2)
$\boldsymbol{\Sigma} = \{0.5^{\lvert i-j \rvert}\}_{i,j}$	β_2	32.2 (0.1)	34.5 (0.1)	161.6 (2.8)	46.3 (0.2)
	β_3	32.4 (0.1)	34.7 (0.1)	9.8 (0.3)	46.6 (0.2)
Setting (D)		ROSE	DLASSO	BLPR	S3
$\boldsymbol{\Sigma} = \boldsymbol{I}_p$	β_3	28.4 (0.1)	31.4 (0.1)	65.6 (1.6)	42.7 (0.2)
	β_4	28.5 (0.1)	31.4 (0.1)	65.6 (1.6)	42.6 (0.2)
	β_5	28.3 (0.1)	31.3 (0.1)	63.9 (1.6)	42.3 (0.2)
	β_6	28.5 (0.1)	31.4 (0.1)	3.8 (0.1)	42.7 (0.2)
$\boldsymbol{\Sigma} = \{0.5^{\lvert i-j \rvert}\}_{i,j}$	β_3	37.0 (0.1)	34.0 (0.1)	55.1 (1.3)	55.6 (0.3)
	β_4	36.9 (0.1)	33.9 (0.1)	68.9 (1.4)	55.7 (0.3)
	β_5	33.2 (0.1)	33.9 (0.1)	84.5 (1.6)	50.3 (0.2)
	β_6	33.1 (0.1)	33.8 (0.1)	6.5 (0.2)	50.1 (0.3)

Table 3: ECP and AL of the CIs, with standard errors in parenthesis

Setting (A)		ROSE		DLASSO		S3	
$\boldsymbol{\Sigma}$		ECP(%)	AL*100	ECP(%)	AL*100	ECP(%)	AL*100
$\boldsymbol{I}_p$	β_2	95.8 (0.9)	80.5 (0.3)	13.0 (1.5)	53.4 (0.2)	93.6 (1.1)	118.8 (0.8)
	β_3	94.8 (1.0)	51.2 (0.1)	97.2 (0.7)	45.2 (0.1)	93.8 (1.1)	75.5 (0.3)
$\{0.5^{\lvert i-j\rvert}\}_{i,j}$	β_2	95.8 (0.9)	77.1 (0.3)	26.6 (2.0)	52.9 (0.1)	95.2 (1.0)	113.0 (0.6)
	β_3	94.8 (1.0)	52.9 (0.1)	95.4 (0.9)	47.5 (0.1)	95.8 (0.9)	76.9 (0.2)
Setting (B)		ROSE		DLASSO		S3	
$\boldsymbol{I}_p$	β_3	94.0 (1.1)	51.1 (0.1)	30.8 (1.1)	39.9 (0.1)	95.2 (1.0)	76.2 (0.3)
	β_4	93.2 (1.1)	50.9 (0.1)	26.0 (1.1)	39.8 (0.1)	95.6 (0.9)	75.7 (0.3)
	β_5	96.2 (0.9)	50.9 (0.1)	30.6 (0.9)	39.8 (0.1)	94.8 (1.0)	76.2 (0.3)
	β_6	93.4 (1.0)	43.9 (0.1)	96.4 (1.0)	38.1 (0.1)	92.4 (1.2)	64.9 (0.2)
$\{0.5^{\lvert i-j\rvert}\}_{i,j}$	β_3	95.6 (0.9)	71.5 (0.2)	88.2 (1.4)	55.7 (0.2)	95.0 (1.0)	107.5 (0.5)
	β_4	93.2 (1.1)	71.5 (0.2)	84.6 (1.6)	54.9 (0.2)	93.8 (1.1)	108.0 (0.5)
	β_5	93.8 (1.1)	65.5 (0.2)	67.6 (2.1)	53.1 (0.1)	93.6 (1.1)	99.1 (0.5)
	β_6	94.0 (1.1)	58.5 (0.2)	95.0 (1.0)	50.8 (0.1)	94.0 (1.1)	88.2 (0.4)

$\boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j=1,\ldots,p}$. The objective is to construct two-sided CIs for $\beta_{0,2}, \beta_{0,3}$ in Setting (A) and $\beta_{0,3}, \beta_{0,4}, \beta_{0,5}, \beta_{0,6}$ in Setting (B).

To implement the proposed CI in (12), we set $s_n = \lfloor 2n/\log(n) \rfloor$ and $l = 5$. We use the R package `SIS` and estimate $\mathcal{M}_{j_0}$ by ISIS. The initial estimator $\widetilde{\boldsymbol{\beta}}$ is computed by fitting a penalized logistic regression model with SCAD penalty function for the variables picked by ISIS. We implement the penalized logistic regression by the R package `ncvreg`. For Setting (B), we update $\widehat{\mathcal{M}}_{j_0}^{(t)}$ using the method discussed in Section 3.3 with $S = 2$.

We further compare the proposed CI with the CI constructed by the DLASSO method and the S3 method. In Table 3, we report the ECP and AL of the proposed CI and the CIs constructed by DLASSO and S3. It can been seen that DLASSO performs poorly for nonzero parameters. On the contrary, ECPs of the proposed CIs are close to the nominal level in almost all cases. In addition, our CIs are much narrower than those based on the S3 method in all cases.

4.3 Real data analysis

We apply the proposed methods to a real dataset riboflavin (vitamin B2) production in Bacillus subtilis. This dataset is provided by DSM (Kaiseraugst, Switzerland) and is publicly available in the R package `hdi`. It consists of a response variable which is the logarithm of the riboflavin production rate and 4088 predictors measuring the logarithm of the expression level of 4088 genes. There are a total of 71 observations. We model this data with a linear regression model, center the response and standardize all the covariates before analysis. To identify genes that are significantly associated with the response, we construct CIs for each individual coefficient and apply Bonferroni's method for multiple adjustment. We compare the proposed method with the de-sparsified Lasso method and implement both methods as discussed in Section 4.1. At the 5% significance level, the proposed method finds three important genes (the 1588th, 3154th and 4004th) while the de-sparsified Lasso procedure claims no variables are significant.

5 Discussion

5.1 Statistical inference via online estimation

In this paper, we develop an online estimation procedure for high-dimensional statistical inference, to account for model selection uncertainty in subsequent inferences. Such an online inference method can be applied to some other non-regular problems as well. Variations of this approach has been used by Luedtke and van der Laan (2016) to provide a CI for the mean outcome under a non-unique optimal treatment regime, and by Luedtke and van der Laan (2017) to construct a CI for the maximal absolute correlation between responses and covariates.

5.2 Multi-dimensional extensions

We focus on constructing CIs for a single regression coefficient in GLMs. The proposed procedure can be naturally extended to form confidence regions for multi-dimensional parameters as well. Let $\mathbb{J}_0$ be an arbitrary subset of $\mathbb{I}$ with $|\mathbb{J}_0| > 1$. The confidence region for $\boldsymbol{\beta}_{0,\mathbb{J}_0}$ can be constructed as follows.

Let $\mathcal{M}_{\mathbb{J}_0} = \{j \notin \mathbb{J}_0 : \beta_{0,j_0} \neq 0\}$. We first estimate $\mathcal{M}_{\mathbb{J}_0}$ by some model selection procedure based on the sub-dataset $\mathcal{F}_t = \{(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_t, Y_t)\}$ for $t = s_n, \dots, n-1$

and $\{(\boldsymbol{X}_{s_n+1}, Y_{s_n+1}), \ldots, (\boldsymbol{X}_n, Y_n)\}$. Denoted by $\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(s_n)}$, $\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(s_n+1)}, \ldots, \widehat{\mathcal{M}}_{\mathbb{J}_0}^{(n-1)}$ and $\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)}$ the corresponding estimators. We calculate $\widehat{\boldsymbol{\Sigma}}$ as in (10) based on some consistent initial estimator $\widetilde{\boldsymbol{\beta}}$ and compute

$$\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)},\mathbb{J}_0} = \widehat{\boldsymbol{\Sigma}}^{-1}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)},\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)}} \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)},\mathbb{J}_0}, \qquad \widehat{\boldsymbol{\sigma}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)},\mathbb{J}_0} = \left(\widehat{\boldsymbol{\Sigma}}_{\mathbb{J}_0,\mathbb{J}_0} - \widehat{\boldsymbol{\Sigma}}^T_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)},\mathbb{J}_0} \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)},\mathbb{J}_0}\right)^{1/2},$$

for $t = s_n, \ldots, n-1$ and

$$\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)},\mathbb{J}_0} = \widehat{\boldsymbol{\Sigma}}^{-1}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)},\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)}} \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)},\mathbb{J}_0}, \qquad \widehat{\boldsymbol{\sigma}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)},\mathbb{J}_0} = \left(\widehat{\boldsymbol{\Sigma}}_{\mathbb{J}_0,\mathbb{J}_0} - \widehat{\boldsymbol{\Sigma}}^T_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)},\mathbb{J}_0} \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)},\mathbb{J}_0}\right)^{1/2}.$$

Consider the following score equation:

$$\begin{aligned} &\sum_{t=0}^{s_n-1} \widehat{\boldsymbol{\sigma}}^{-1}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)},\mathbb{J}_0} \widehat{\boldsymbol{Z}}^T_{t+1,\mathbb{J}_0} \left\{Y_{t+1} - \mu\left(\boldsymbol{X}_{t+1,\mathbb{J}_0}\boldsymbol{\beta}_{0,\mathbb{J}_0} + \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)}}\right)\right\} \\ + &\sum_{t=s_n}^{n-1} \widehat{\boldsymbol{\sigma}}^{-1}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)},\mathbb{J}_0} \widehat{\boldsymbol{Z}}^T_{t+1,\mathbb{J}_0} \left\{Y_{t+1} - \mu\left(\boldsymbol{X}_{t+1,\mathbb{J}_0}\boldsymbol{\beta}_{0,\mathbb{J}_0} + \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)}}\right)\right\} = 0, \end{aligned}$$

where $\widehat{\boldsymbol{Z}}_{t+1,\mathbb{J}_0} = \boldsymbol{X}_{t+1,\mathbb{J}_0} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)},\mathbb{J}_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)}}$ for $t = s_n, \ldots, n-1$ and $\widehat{\boldsymbol{Z}}_{t+1,\mathbb{J}_0} = \boldsymbol{X}_{t+1,\mathbb{J}_0} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)},\mathbb{J}_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)}}$ for $t = 0, \ldots, s_n - 1$. The estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{J}_0}$ can be computed by solving the score equation via Newton's method with initial value $\widetilde{\boldsymbol{\beta}}_{\mathcal{J}_0}$. The corresponding $1-\alpha$ 100% confidence region is given by

$$\left\{\boldsymbol{\beta}_{\mathbb{J}_0} \in \mathbb{R}^{|\mathbb{J}_0|} : n(\boldsymbol{\beta}_{\mathbb{J}_0} - \widehat{\boldsymbol{\beta}}_{\mathbb{J}_0})^T (\boldsymbol{\Gamma}^*_n)^T \boldsymbol{\Gamma}^*_n (\boldsymbol{\beta}_{\mathbb{J}_0} - \widehat{\boldsymbol{\beta}}_{\mathbb{J}_0}) \hat{\phi} \le \chi^2_\alpha(|\mathbb{J}_0|)\right\}, \tag{26}$$

where $\hat{\phi}$ denotes some constant estimator for ϕ_0, $\chi^2_\alpha(|\mathbb{J}_0|)$ is the upper α-quantile of a central χ^2 distribution with $|\mathbb{J}_0|$ degrees of freedom, and

$$\begin{aligned} \boldsymbol{\Gamma}^*_n &= \frac{1}{n}\sum_{t=0}^{s_n-1} \widehat{\boldsymbol{\sigma}}^{-1}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)},\mathbb{J}_0} \widehat{\boldsymbol{Z}}^T_{t+1,\mathbb{J}_0} b''\left(\boldsymbol{X}_{t+1,\mathbb{J}_0}\widehat{\boldsymbol{\beta}}_{0,\mathbb{J}_0} + \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(-s_n)}}\right) \\ &+ \frac{1}{n}\sum_{t=s_n}^{n-1} \widehat{\boldsymbol{\sigma}}^{-1}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)},\mathbb{J}_0} \widehat{\boldsymbol{Z}}^T_{t+1,\mathbb{J}_0} b''\left(\boldsymbol{X}_{t+1,\mathbb{J}_0}\widehat{\boldsymbol{\beta}}_{0,\mathbb{J}_0} + \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)}} \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{\mathbb{J}_0}^{(t)}}\right). \end{aligned}$$

To guarantee the validity of (26), the number of elements in $\mathbb{J}_0$ needs to be much smaller than n. It would be interesting to construct confidence regions for the entire regression coefficient vector $\boldsymbol{\beta}_0$ based on some multiple comparison procedures. However, this is beyond the scope of the current paper.

5.3 Extension to generic penalized M-estimators

The proposed method can also be extended beyond the class of GLMs to a general framework with a convex loss function. Specifically, given a high-dimensional random vector $\boldsymbol{U}_0$, define

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathrm{E}\ell(\boldsymbol{U}_0, \boldsymbol{\beta}),$$

for some convex loss function ℓ. An initial estimator for $\boldsymbol{\beta}_0$ can be computed by minimizing

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left(\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{U}_i, \boldsymbol{\beta}) + \sum_{j=1}^{p} \rho_\lambda(|\beta_j|) \right), \tag{27}$$

where $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ are i.i.d random vectors generated according as $\boldsymbol{U}_0$, and $\rho_\lambda(\cdot)$ denotes some penalty function. In addition to estimating the regression coefficients in GLMs, such a generic framework includes some other important applications such as estimation of the precision matrix in Gaussian graphical models (as illustrated in Section 2.1.4 of Ning and Liu, 2017).

Here, we focus on constructing the CI for a univariate parameter β_{0,j_0}. Let $\widehat{\mathcal{M}}_{j_0}^{(t)}$ denote the estimated support of the control variables based on $\{\boldsymbol{U}_i\}_{i=1}^t$ for $t = s_n, s_n+1, \ldots, n-1$ and $\widehat{\mathcal{M}}_{j_0}^{(-s_n)}$ the estimated support based on $\{\boldsymbol{U}_i\}_{i=s_n+1}^n$. Define

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\ell(\boldsymbol{U}_i, \widetilde{\boldsymbol{\beta}}),$$

where $\widetilde{\boldsymbol{\beta}}$ corresponds to the initial estimator in (27). Given $\widehat{\boldsymbol{\Sigma}}$, we compute $\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0}$, and $\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0}$ as in Section 3.1. For any model $\mathcal{M} \subseteq \mathbb{I}_{j_0}$, suppose we have some consistent estimator $\hat{\sigma}^2_{\mathcal{M},j_0}$ for

$$\sigma^2_{\mathcal{M},j_0} = \mathrm{E}\left(\frac{\partial \ell(\boldsymbol{U}_0, \boldsymbol{\beta}_0)}{\partial \beta_{j_0}} - \boldsymbol{\omega}^T_{\mathcal{M},j_0}\frac{\partial \ell(\boldsymbol{U}_0, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_{\mathcal{M}}}\right)^2.$$

For any $c \in \mathbb{R}$, $\mathcal{M} \subseteq \mathbb{I}_{j_0}$ and $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{M}|}$, we define a p-dimensional vector $\boldsymbol{\theta} = \mathrm{h}(c, \mathcal{M}, \boldsymbol{\alpha})$ such that $\theta_{j_0} = c$, $\boldsymbol{\theta}_{\mathcal{M}} = \boldsymbol{\alpha}$ and $\boldsymbol{\theta}_{\mathcal{M}^c - \{j_0\}} = \mathbf{0}$. Let $\widehat{\mathcal{M}}^{(t)}_{j_0} = \widehat{\mathcal{M}}^{(-s_n)}_{j_0}$ for $t = 0, 1, \ldots, s_n - 1$ and $\hat{\beta}^{(0)}_{j_0} = \widetilde{\beta}_{j_0}$, we update $\hat{\beta}_{j_0}$ as

$$\hat{\beta}^{(l)}_{j_0} = \hat{\beta}^{(l-1)}_{j_0} - \underbrace{\frac{\sum_{t=0}^{n-1} \frac{1}{n\hat{\sigma}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} \left(\frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\hat{\beta}^{(l-1)}_{j_0}, \widehat{\mathcal{M}}^{(t)}_{j_0}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}))}{\partial \beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}^{(t_0)}_{j_0},j_0} \frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\hat{\beta}^{(l-1)}_{j_0}, \widehat{\mathcal{M}}^{(t)}_{j_0}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}))}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}} \right)}{\sum_{t=0}^{n-1} \frac{1}{n\hat{\sigma}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} \left(\frac{\partial^2 \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\hat{\beta}^{(l-1)}_{j_0}, \widehat{\mathcal{M}}^{(t)}_{j_0}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}))}{\partial \beta^2_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}^{(t_0)}_{j_0},j_0} \frac{\partial^2 \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\hat{\beta}^{(l-1)}_{j_0}, \widehat{\mathcal{M}}^{(t)}_{j_0}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}))}{\partial \beta_{j_0} \partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}} \right)}}_{\Gamma^{*,(l-1)}_n},$$

for $l = 1, 2, \ldots$ The corresponding CI for β_{0,j_0} is given by

$$\hat{\beta}^{(l)}_{j_0} \pm \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}\Gamma^{*,(l-1)}_n}.$$

In Section C of the supplementary article, we sketch a few lines to show that the above CI achieves nominal coverage under certain conditions.

5.4 Doubly-robust procedure

The proposed ROSE algorithm constructs the score equation for β_{0,j_0} by recursively estimating the support of control variables. As commented in (2.2), such a procedure requires certain minimal-signal-strength conditions on $\boldsymbol{\beta}_{0,\mathbb{I}_{j_0}}$.

We now introduce a variant of our method that is valid even when the minimal-signal-strength conditions fail. At the t-th iteration, instead of estimating $\mathcal{M}_{j_0}$ only, we might apply another variable selection procedure to estimate the support of $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0}$ based on $\mathcal{F}_t$ and set $\widehat{\mathcal{M}}^{(t)}_{j_0}$ to be a union of the two sets of important variables selected. The resulting CI is doubly-robust in the sense that it achieves nominal coverage as long as either $\boldsymbol{\beta}_{0,\mathbb{I}_{j_0}}$ satisfies certain minimal-signal-strength conditions, or the ℓ_2 norm of weak signals in $\boldsymbol{\beta}_{0,\mathbb{I}_{j_0}}$ and $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0}$ is $o(n^{-1/4})$. The latter condition allows the existence of weak signals in $\boldsymbol{\beta}_{0,\mathbb{I}_{j_0}}$. It automatically holds when variables with signals larger than or proportional to $(n/\log\log n)^{-1/4}(s^*)^{-1/2}$ can be consistently identified by the model selection procedure. In addition, it is considerably weaker than the zonal assumption (Bühlmann and Mandozzi, 2014) that requires the strength of weak signals to be $o(n^{-1/2})$. More detailed discussions are given in Section B.1.3 of the supplementary article.

Acknowledgment

The authors thank the AE, and the reviewers for their constructive comments, which have led to a significant improvement of the earlier version of this article. This work was supported by NSF grants DMS 1555244 and 1820702, and NIH grant NIH P01CA142538. Li also received partial travel support from NNSFC 11690015 during his visit at Nankai University. The content is solely the responsibility of the authors and does not necessarily represent the official views of NSF or NIH.

References

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc. 70*, 428–434.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist. 37*(4), 1705–1732.

Breheny, P. and J. Huang (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat. 5*(1), 232–253.

Bühlmann, P. and J. Mandozzi (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Comput. Statist. 29*(3-4), 407–430.

Candes, E. and T. Tao (2007). The dantzig selector: statistical estimation when p is much larger than n (with discussions). *Ann. Stat. 35*, 2313–2404.

Csörgö, M. (1968). On the strong law of large numbers and the central limit theorem for martingales. *Trans. Amer. Math. Soc. 131*, 259–275.

Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen (2015). High-dimensional inference: confidence intervals, p-values and `r`-software `hdi`. *Statist. Sci. 30*(4), 533–558.

Fan, J., S. Guo, and N. Hao (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 74*(1), 37–65.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 70*(5), 849–911.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res. 15*, 2869–2909.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist. 44*(3), 907–927.

Liu, H., X. Xu, and J. J. Li (2017). A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models. *arXiv preprint arXiv:1706.02150*.

Liu, H. and B. Yu (2013). Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electron. J. Stat. 7*, 3124–3169.

Luedtke, A. R. and M. J. van der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist. 44*(2), 713–742.

Luedtke, A. R. and M. J. van der Laan (2017). Parametric-rate inference for one-sided differentiable parameters. *J. Amer. Statist. Assoc.* (just-accepted).

McCullagh and Nelder (1989). *Generalized Linear Models*. Chapman and Hall.

Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist. 45*(1), 158–195.

Saldana, D. F. and Y. Feng (2016). Sis: An r package for sure independence screening in ultrahigh dimensional statistical models. *Journal of Statistical Software*, to appear.

Schifano, E. D., J. Wu, C. Wang, J. Yan, and M.-H. Chen (2016). Online updating of statistical inference in the big data setting. *Technometrics 58*(3), 393–403.

Sun, T. and C.-H. Zhang (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res. 14*, 3385–3418.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 58*, 267–288.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist. 42*(3), 1166–1202.

Wang, C., M.-H. Chen, E. Schifano, J. Wu, and J. Yan (2016). Statistical methods and computing for big data. *Stat. Interface 9*(4), 399–414.

Wasserman, L. and K. Roeder (2009). High-dimensional variable selection. *Ann. Statist. 37*(5A), 2178–2201.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*(2), 894–942.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 76*(1), 217–242.

Supplement to "Statistical Inference for High-Dimensional Models via Recursive Online-Score Estimation"

Chengchun Shi[a], Rui Song[b], Wenbin Lu[b] and Runze Li[c]
[a]Department of Statistics, London School of Economics and Political Science, London; [b]Department of Statistics, North Carolina State University, Raleigh, NC; [c]Department of Statistics, The Pennsylvania State University, University Park, PA

This supplementary article is organized as follows. In Section A, we present proofs of Theorem 2.1, Theorem 3.1, Theorem 3.2, Theorem 3.3, Lemma A.1, Lemma A.2, Lemma A.3 and Lemma A.4. In Section B, we provide detailed discussions on our technical conditions and compare them with those imposed in the existing literature. In Section C, we give more details on the extensions to the generic penalized M-estimators.

A Proofs

A.1 Proof of Theorem 2.1

Before proving Theorem 2.1, we present the following lemmas whose proofs are given in the supplementary material.

Lemma A.1. *Under conditions in Theorem 2.1, we have*

$$\min_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sigma_{\mathcal{M},j_0} \ge \sqrt{\bar{c}}, \qquad \max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 \le (\bar{c})^{-1/2} c_0, \tag{A.1}$$

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} E\left\{ \frac{1}{\sigma_{\mathcal{M},j_0}^4} \left(X_{0,j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}^T \boldsymbol{X}_{0,\mathcal{M}} \right)^4 \right\} \le \frac{c_0^4}{\bar{c}^2} \left(1 + \frac{c_0^2}{\bar{c}} \right)^2, \tag{A.2}$$

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ |\mathcal{M}|\le\kappa_n}} E\|X_{0,\mathcal{M}}\|_2^2 \le \kappa_n c_0^2, \tag{A.3}$$

where $\bar{c}$ and c_0 are defined in Condition (A2) and (A3), respectively. Moreover, we have with probability tending to 1 that

$$\max_{j\in[1,\dots,p]} |X_{0,j}| \leq \sqrt{3c_0^2 \max(\log p, \log n)}. \tag{A.4}$$

Lemma A.2. *Under conditions in Theorem 2.1, the following events hold with probability tending to 1,*

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 \leq \bar{c}_0 \left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}}\right), \tag{A.5}$$

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} |\hat{\sigma}_{\mathcal{M},j_0} - \sigma_{\mathcal{M},j_0}| \leq \bar{c}_0 \left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}}\right), \tag{A.6}$$

for some constant $\bar{c}_0 > 0$, where $\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}$ and $\hat{\sigma}^2_{\mathcal{M},j_0} = \widehat{\boldsymbol{\Sigma}}_{j_0,j_0} - \widehat{\boldsymbol{\Sigma}}^T_{\mathcal{M},j_0}\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0}$,

$$\overline{\boldsymbol{\omega}}_{\mathcal{M},j_0} = \boldsymbol{\omega}_{\mathcal{M},j_0} + \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}\{\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0} - \boldsymbol{\Sigma}_{\mathcal{M},j_0} - (\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}} - \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}\}.$$

In addition, we have

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \overline{\boldsymbol{\omega}}_{\mathcal{M},j_0}\|_2 = O_p\left(\frac{\kappa_n \log p}{n}\right). \tag{A.7}$$

Lemma A.3. *Under conditions in Theorem 2.1, the following events hold with probability tending to 1,*

$$\max_{j\in[1,\dots,p]} \left|\sum_{t=s_n}^{n-1} \frac{X_{t+1,j}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right) I(j\in\widehat{\mathcal{M}}_{j_0}^{(t)})\right| \leq \bar{c}_*\sqrt{\log p}, \tag{A.8}$$

$$\sum_{t=s_n}^{n-1} \left\|\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)\right\|_2^2 \leq \bar{c}_* n\eta_n^2, \tag{A.9}$$

for some constant $\bar{c}_* > 0$.

Proof of Theorem 2.1: Under (A1), it follows from the Bonferroni's inequality that

$$\Pr\left(\mathcal{M}_{j_0} \subseteq \bigcap_{t=s_n}^{n-1} \widehat{\mathcal{M}}_{j_0}^{(t+1)}\right) \geq 1 - O\left(\sum_{t=s_n}^{\infty}\frac{1}{t^{\alpha_0}}\right) \to 1. \tag{A.10}$$

Besides,

$$\Pr\left(\mathcal{M}_{j_0} \subseteq \widehat{\mathcal{M}}_{j_0}^{(-s_n)}\right) \to 1. \tag{A.11}$$

Under the events defined in the left-hand-side (LHS) of (A.10) and (A.11), we have

$$\sqrt{n}\Gamma_n^*(\hat{\beta}_{j_0} - \beta_{0,j_0}) = \underbrace{\sum_{t=s_n}^{n-1} \frac{\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}}_{I_1} - \underbrace{\sum_{t=s_n}^{n-1} \frac{\widehat{Z}_{t+1,j_0}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)}_{I_2}$$
$$- \underbrace{\sum_{t=0}^{s_n-1} \frac{\widehat{Z}_{t+1,j_0}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}}\right)}_{I_3} + \underbrace{\sum_{t=0}^{s_n-1} \frac{\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}}}_{I_4}.$$

In the following, we break the proof into four steps. In the first three steps, we show $I_j = o_p(1)$, for $j = 2, 3, 4$, respectively. In the last step, we prove

$$I_1 \xrightarrow{d} N(0, \sigma_0^2).$$

By Assumption (A6) and Slutsky's theorem, we have

$$\frac{\sqrt{n}\Gamma_n^*(\hat{\beta}_{j_0} - \beta_{0,j_0})}{\hat{\sigma}} \xrightarrow{d} N(0, 1).$$

The assertion therefore follows.

Step 1: Let

$$I_2^* = \sum_{t=s_n}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right) \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right).$$

$|I_2 - I_2^*|$ is upper bounded by

$$\sum_{t=s_n}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left| \left(\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\right)^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right| \left| \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right) \right|.$$

Under the given conditions, we have $\kappa_n \log p = o(n)$. Under the events defined in (A.1) and

(A.6), we have for sufficiently large n,

$$\min_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \hat{\sigma}_{\mathcal{M},j_0} \geq \min_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \sigma_{\mathcal{M},j_0} - o(1) \geq \frac{\sqrt{\bar{c}}}{2},$$

and hence

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \frac{1}{\hat{\sigma}_{\mathcal{M},j_0}} \leq \frac{2}{\sqrt{\bar{c}}}. \tag{A.12}$$

Under the events defined in Condition (A1) and (A.12), $|I_2 - I_2^*|$ can be upper bounded by

$$\begin{aligned} &\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \frac{1}{\hat{\sigma}_{\mathcal{M},j_0}} \sum_{t=s_n}^{n-1} \frac{1}{\sqrt{n}} \left| \left(\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \right)^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right| \left| \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right| \\ &\leq \frac{2}{\sqrt{\bar{c}n}} \sum_{t=s_n}^{n-1} \left| \left(\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \right)^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right| \left| \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right|. \end{aligned}$$

By Cauchy-Schwarz inequality, we have with probability tending to 1 that

$$|I_2 - I_2^*| \leq 2(\bar{c})^{-1/2}\sqrt{nI_2^{(1)}I_2^{(2)}}, \tag{A.13}$$

where

$$\begin{aligned} I_2^{(1)} &= \frac{1}{n}\sum_{t=s_n}^{n-1} \left| \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right|^2, \\ I_2^{(2)} &= \frac{1}{n}\sum_{t=s_n}^{n-1} \left| \left(\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \right)^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right|^2. \end{aligned}$$

It follows from (A.9) that

$$I_2^{(1)} = O(\eta_n^2), \tag{A.14}$$

with probability tending to 1.

For any $a, b \in \mathbb{R}$, we have by Cauchy-Schwarz inequality that $(a+b)^2 \leq 2a^2 + 2b^2$. It follows that $I_2^{(2)} \leq 2I_2^{(3)} + 2I_2^{(4)}$ where

$$\begin{aligned} I_2^{(3)} &= \frac{1}{n}\sum_{t=s_n}^{n-1}\left|\left(\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \overline{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\right)^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right|^2 \\ I_2^{(4)} &= \frac{1}{n}\sum_{t=s_n}^{n-1}\left|\left(\overline{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\right)^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right|^2. \end{aligned}$$

By Condition (A1), (A.7) and Cauchy-Schwarz inequality, $I_2^{(3)}$ can be bounded by

$$\frac{1}{n}\sum_{t=s_n}^{n-1}\|\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \overline{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\|_2^2\|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\|_2^2 = O_p(n^{-2}\kappa_n^2\log^2 p)\frac{1}{n}\sum_{t=s_n}^{n-1}\|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\|_2^2.$$

In the following, we show

$$\frac{1}{n}\sum_{t=s_n}^{n-1}\|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\|_2^2 = O_p(\kappa_n). \tag{A.15}$$

This further implies

$$I_2^{(3)} = O_p\left(\frac{\kappa_n^3\log^2 p}{n^2}\right) = O_p\left(\frac{\kappa_n\log p}{n}\right), \tag{A.16}$$

under the condition that $\kappa_n^2\log p = O(n/\log^2 n)$. To prove (A.15), it suffices to show

$$\frac{1}{n}\sum_{t=s_n}^{n-1}\mathrm{E}\|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\|_2^2 = O(\kappa_n). \tag{A.17}$$

Since X_{t+1} and $\widehat{\mathcal{M}}_{j_0}^{(t)}$ is independent, we have by Condition (A1) that

$$\frac{1}{n}\sum_{t=s_n}^{n-1}\mathrm{E}\|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\|_2^2 \leq \frac{1}{n}\sum_{t=s_n}^{n-1}\sup_{\substack{\mathcal{M}\leq[1,\ldots,p]\\|\mathcal{M}|\leq\kappa_n}}\mathrm{E}\|\boldsymbol{X}_{t+1,\mathcal{M}}\|_2^2.$$

The RHS is $O(\kappa_n)$ by (A.3).

Consider $I_2^{(4)}$. For $i = 1,\ldots,n$ and any $\mathcal{M} \subseteq [1,\ldots,p]$, define

$$\overline{\boldsymbol{\omega}}_{\mathcal{M},j_0}^{(-i)} = \boldsymbol{\omega}_{\mathcal{M},j_0} + \frac{n-1}{n}\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}\{\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}^{(-i)} - \boldsymbol{\Sigma}_{\mathcal{M},j_0} - (\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{(-i)} - \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}\},$$

where

$$\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}^{(-i)} = \frac{1}{n-1}\sum_{l\neq i}\boldsymbol{X}_{l,\mathcal{M}}X_{l,j_0} \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}^{(-i)} = \frac{1}{n-1}\sum_{l\neq i}\boldsymbol{X}_{l,\mathcal{M}}\boldsymbol{X}_{l,\mathcal{M}}^T.$$

It follows that

$$\|\overline{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \overline{\boldsymbol{\omega}}_{\mathcal{M},j_0}^{(-i)}\|_2 \leq \frac{1}{n}\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}\{\boldsymbol{X}_{l,\mathcal{M}}X_{l,j_0} - \boldsymbol{\Sigma}_{\mathcal{M},j_0} - (\boldsymbol{X}_{l,\mathcal{M}}\boldsymbol{X}_{l,\mathcal{M}}^T - \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}\}.$$

Under the event defined in (A.4), it follows from Condition (A2) and (A.1) that

$$\max_{\substack{\mathcal{M}\in[1,\dots,p],i\in[1,\dots,n]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \|\overline{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \overline{\boldsymbol{\omega}}_{\mathcal{M},j_0}^{(-i)}\|_2 = O\left(\frac{\kappa_n\log p + \kappa_n\log n}{n}\right). \tag{A.18}$$

The condition $\kappa_n^2\log p = O(n/\log^2 n)$ implies that $O_p(n^{-1}\kappa_n\log p) = o_p(n^{-1/2}\sqrt{\kappa_n\log p})$. By (A.7), we have with probability tending to 1 that

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \overline{\boldsymbol{\omega}}_{\mathcal{M},j_0}\|_2 \leq \left(\frac{\kappa_n\log p}{n}\right)^{1/2}.$$

Combining this together with (A.5), (A.18) and the condition $\kappa_n^2\log p = O(n/\log^2 n)$ yields

$$\begin{aligned}&\max_{\substack{\mathcal{M}\in[1,\dots,p],i\in[1,\dots,n]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \|\boldsymbol{\omega}_{\mathcal{M},j_0} - \overline{\boldsymbol{\omega}}_{\mathcal{M},j_0}^{(-i)}\|_2 \qquad\qquad (\text{A.19})\\ =\;& O\left(\frac{\kappa_n\log p + \kappa_n\log n}{n} + \frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right) = O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right),\end{aligned}$$

with probability tending to 1. Define

$$\begin{aligned} I_2^{(5)} &= \frac{1}{n}\sum_{t=s_n}^{n-1}\left|\left(\overline{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^{(-t-1)} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\right)^T\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right|^2,\\ I_2^{(6)} &= \frac{1}{n}\sum_{t=s_n}^{n-1}\left|\left(\overline{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \overline{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^{(-t-1)}\right)^T\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right|^2. \end{aligned}$$

By Cauchy-Schwarz inequality, we can similarly show $I_2^{(4)} \leq 2I_2^{(5)} + 2I_2^{(6)}$. Using similar arguments in bounding $I_2^{(3)}$, we can similarly show that

$$I_2^{(6)} = O_p\left(\frac{\kappa_n^3(\log^2 p + \log^2 n)}{n^2}\right), \tag{A.20}$$

by (A.15) and (A.18). Under the events defined in Condition (A1) and (A.19), we have

$$I_2^{(5)} \leq \frac{1}{n}\sum_{t=s_n}^{n-1}\left|\left(\overline{\boldsymbol{\omega}}^{(-t-1)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}\right)^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}\right|^2 \mathcal{I}\left\{\|\overline{\boldsymbol{\omega}}^{(-t-1)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}\|_2 = O\left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}}\right)\right\},$$

where $\mathcal{I}\{\cdot\}$ denote the indicator function. Since $\boldsymbol{X}_{t+1}$ is independent of $\overline{\boldsymbol{\omega}}^{(-t-1)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}$ and $\widehat{\mathcal{M}}^{(t)}_{j_0}$, we have

$$\mathrm{E}I_2^{(5)} = \frac{1}{n}\sum_{t=s_n}^{n-1}\mathrm{E}\left\|\boldsymbol{\Sigma}^{1/2}_{\widehat{\mathcal{M}}^{(t)}_{j_0},\widehat{\mathcal{M}}^{(t)}_{j_0}}\left(\overline{\boldsymbol{\omega}}^{(-t-1)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}\right)\right\|_2^2 \mathcal{I}\left\{\|\overline{\boldsymbol{\omega}}^{(-t-1)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}\|_2 = O\left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}}\right)\right\}.$$

For any random variable Z, it follows from the definition of the Orlicz norm that

$$1 + \mathrm{E}\frac{Z^2}{\|Z\|_{\psi_2}^2} \leq \mathrm{E}\exp\left(\frac{Z^2}{\|Z\|_{\psi_2}^2}\right) \leq 2,$$

and hence

$$\mathrm{E}Z^2 \leq \|Z\|_{\psi_2}^2. \tag{A.21}$$

Note that $\boldsymbol{\Sigma}$ is positive definite, we have by Condition (A3) that

$$\|\boldsymbol{\Sigma}\|_2 = \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^p\\ \|\boldsymbol{a}\|_2=1}} \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a} = \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^p\\ \|\boldsymbol{a}\|_2=1}} \mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_0|^2 \leq \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^p\\ \|\boldsymbol{a}\|_2=1}} \|\boldsymbol{a}^T\boldsymbol{X}_0\|_{\psi_2}^2 \leq c_0^2, \tag{A.22}$$

It follows from (A.22) that

$$\begin{aligned}\mathrm{E}I_2^{(5)} \leq \frac{1}{n}\sum_{t=s_n}^{n-1}\mathrm{E}\lambda_{\max}(\boldsymbol{\Sigma}_{\widehat{\mathcal{M}}^{(t)}_{j_0},\widehat{\mathcal{M}}^{(t)}_{j_0}})\|\overline{\boldsymbol{\omega}}^{(-t-1)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}\|_2^2\mathcal{I}\left\{\|\overline{\boldsymbol{\omega}}^{-t-1}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}\|_2 = O\left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}}\right)\right\}\\ = O(n^{-1}\kappa_n \log p),\end{aligned}$$

with probability tending to 1. This further implies $I_2^{(5)} = O_p(n^{-1}\kappa_n \log p)$. Combining this together with (A.20) and the condition $\kappa_n^2 \log p = O(n/\log^2 n)$ yields that

$$I_2^{(4)} = O_p\left(\frac{\kappa_n \log p}{n} + \frac{\kappa_n^3(\log^2 p + \log^2 n)}{n^2}\right) = O_p\left(\frac{\kappa_n \log p}{n}\right).$$

This together with (A.13), (A.14) and (A.16) yields that

$$|I_2 - I_2^*| = O_p(\eta_n\sqrt{\kappa_n \log p}).$$

It follows from Condition (A4) that $|I_2 - I_2^*| = o_p(1)$.

Let

$$I_2^{**} = \sum_{t=s_n}^{n-1} \frac{1}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right).$$

By definition, we have

$$|I_2^* - I_2^{**}| = \sum_{t=s_n}^{n-1} \frac{|\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}|}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left| \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right|.$$

Using similar arguments in bounding $I_2 - I_2^*$, we can show

$$|I_2^* - I_2^{**}| \le \frac{2\bar{c}_0}{\bar{c}} \left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}} \right) \sum_{t=s_n}^{n-1} \left| X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right| \left| \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right|.$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\sum_{t=s_n}^{n-1} \left| X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right| \left| \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right| \\ \le \; & \left(\sum_{t=s_n}^{n-1} \left| X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right|^2 \right)^{1/2} \left(\sum_{t=s_n}^{n-1} \left| \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right|^2 \right)^{1/2}. \end{aligned}$$

By (A.1), (A.21) and Condition (A3), we can show

$$\sum_{t=s_n}^{n-1} \mathrm{E} \left| X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right|^2 = O(n),$$

and hence

$$\sum_{t=s_n}^{n-1} \left| X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right|^2 = O_p(n). \tag{A.23}$$

This together with (A.9) yields

$$\sum_{t=s_n}^{n-1} \left| X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right| \left| \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right| = O_p(\sqrt{n}\eta_n).$$

It follows that

$$|I_2^* - I_2^{**}| = O_p(\eta_n \sqrt{\kappa_n \log p}),$$

which is $o_p(1)$ under (A4).

Thus, to prove $I_2 = o_p(1)$, it suffices to show $I_2^{**} = o_p(1)$. Define

$$I_{2,j} = \sum_{t=s_n}^{n-1} \frac{1}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) X_{t+1,j} \mathcal{I}(j \in \widehat{\mathcal{M}}_{j_0}^{(t)}).$$

We have $I_2^{**} = \sum_{j=1}^{p} I_{2,j}(\widetilde{\beta}_j - \beta_{0,j})$. Therefore,

$$|I_2^{**}| \le \max_{j \in [1,\dots,p]} |I_{2,j}| \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1. \tag{A.24}$$

Let $\sigma(\mathcal{F}_t)$ be the σ-algebra generated by $\{(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_t, Y_t)\}$. Then, each $I_{2,j}$ forms a mean zero martingale with respect to $\sigma(\mathcal{F}_t)$. To see this, note that $\widehat{\mathcal{M}}_{j_0}^{(t)}$ is fixed given $\mathcal{F}_t$. If $j \notin \mathcal{M}_{j_0}^{(t)}$, we have

$$\begin{aligned} & \mathrm{E}\left\{ \frac{1}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) X_{t+1,j} \mathcal{I}(j \in \widehat{\mathcal{M}}_{j_0}^{(t)}) | \mathcal{F}_t \right\} \\ = & \mathrm{E}\left\{ \frac{1}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) X_{t+1,j} | \mathcal{F}_t \right\} \mathcal{I}(j \in \widehat{\mathcal{M}}_{j_0}^{(t)}) = 0. \end{aligned}$$

If $j \in \mathcal{M}_{j_0}^{(t)}$, then we have

$$\mathrm{E}\left\{ \frac{X_{t+1,j}}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) | \mathcal{F}_t \right\} = \frac{\boldsymbol{\Sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \boldsymbol{\Sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} = 0.$$

By some exponential inequalities for martingales, we show in (A.8) that $\Pr(\max_j |I_{2,j}| \ge \bar{c}_* \sqrt{\log p}) \to 0$. It follows from Condition (A4) that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \le (k_0+1) \|\widetilde{\boldsymbol{\beta}}_{\mathcal{M}_0} - \boldsymbol{\beta}_{0,\mathcal{M}_0}\|_1 \le \sqrt{|\mathcal{M}_0|}(k_0+1) \|\widetilde{\boldsymbol{\beta}}_{\mathcal{M}_0} - \boldsymbol{\beta}_{0,\mathcal{M}_0}\|_2 \le \sqrt{|\mathcal{M}_0|}(k_0+1)\eta_n$, with probability tending to 1. Under (A1), we have $|\mathcal{M}_0| \le \kappa_n - 1$. It follows that

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = O(\sqrt{\kappa_n} \eta_n), \tag{A.25}$$

with probability tending to 1. Since $\eta_n\sqrt{\kappa_n \log p} = o(1)$, we have $\max_j |I_{2,j}| \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = o_p(1)$. This together with (A.24) gives $I_2^{**} = o_p(1)$.

Step 2: Using similar arguments in Step 1, we can show that I_3 is asymptotically equivalent to I_3^{**}, defined as

$$I_3^{**} = \sum_{t=0}^{s_n-1} \frac{1}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}} Z_{t+1,j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \right),$$

where $Z_{t+1,j_0} = X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}}$. Hence, it suffices to show $I_3^{**} = o_p(1)$. Note that $|I_3^{**}|$ is upper bounded by

$$|I_3^{**}| \le \max_{j\in[1,\dots,p]} |I_{3,j}| \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1, \tag{A.26}$$

where

$$I_{3,j} = \sum_{t=0}^{s_n-1} \frac{1}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \right) X_{t+1,j} \mathcal{I}(j \in \widehat{\mathcal{M}}_{j_0}^{(-s_n)}).$$

Given $\{(\boldsymbol{X}_{s_n+1}, Y_{s_n+1}), \dots, (\boldsymbol{X}_n, Y_n)\}$, the set $\widehat{\mathcal{M}}_{j_0}^{(-s_n)}$ is fixed. For any $j \in [1,\dots,p]$, $I_{3,j}$ corresponds to a sum of mean zero i.i.d random variables. Similar to the proof of Lemma A.3, we can show

$$\Pr(\max_j |I_{3,j}| \le c_* \sqrt{\log p}) \to 1,$$

for some constant $c_* > 0$ that is independent of $\widehat{\mathcal{M}}_{j_0}^{(-s_n)}$. By (A.26) and Condition (A4), we have $|I_3^{**}| \to 0$ with probability tending to 1. This proves $I_3 = o_p(1)$.

Step 3: Let

$$I_4^* = \sum_{t=0}^{s_n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}^T \boldsymbol{X}_{t+1,\mathcal{M}_{j_0}^{(-s_n)}} \right) \varepsilon_{t+1},$$

we have

$$I_4 - I_4^* = \sum_{t=0}^{s_n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}} \left(\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} - \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \right) \varepsilon_{t+1}.$$

We first show $I_4 - I_4^* = o_p(1)$. Since $\varepsilon_1, \ldots, \varepsilon_{s_n}$ are independent of $\{\boldsymbol{X}_i\}_{i=1}^n$, it follows from the Chebyshev's inequality that

$$
\begin{aligned}
& \Pr\left(|I_4 - I_4^*| > t^* | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \varepsilon_{s_n+1}, \ldots, \varepsilon_n\right) \\
\leq\ & \frac{1}{(t^*)^2} \mathrm{E}\{(I_4 - I_4^*)^2 | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \varepsilon_{s_n+1}, \ldots, \varepsilon_n\} \\
\leq\ & \sum_{t=0}^{s_n-1} \frac{\sigma_0^2}{(t^*)^2 n \hat{\sigma}^2_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0}} \left\{ \left(\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} - \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} \right)^T \boldsymbol{X}_{t+1, \widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \right\}^2 . \qquad \text{(A.27)}
\end{aligned}
$$

By (A.5), (A.12) and (A.15), we can show that

$$
\sum_{t=0}^{s_n-1} \frac{\sigma_0^2}{n \hat{\sigma}^2_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0}} \left\{ \left(\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} - \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} \right)^T \boldsymbol{X}_{t+1, \widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \right\}^2 = O_p\left(\frac{\kappa_n^2 \log p}{n} \right).
$$

In view of (A.27), this further implies that $|I_4 - I_4^*| = O_p(n^{-1/2}\kappa_n \sqrt{\log p})$. Under the given conditions, we obtain that $I_4 - I_4^* = o_p(1)$. Similarly, we can show $I_4^* - I_4^{**} = o_p(1)$, where

$$
I_4^{**} = \sum_{t=0}^{s_n-1} \frac{1}{\sqrt{n} \sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0}} \left(X_{t+1, j_0} - \boldsymbol{\omega}^T_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} \boldsymbol{X}_{t+1, \widehat{\mathcal{M}}_{j_0}^{(-s_n)}} \right) \varepsilon_{t+1}.
$$

Thus, it suffices to show $I_4^{**} = o_p(1)$. Note that we have

$$
\mathrm{E}[(I_4^{**})^2 | \{(\boldsymbol{X}_{s_n+1}, \varepsilon_{s_n+1}), \ldots, (\boldsymbol{X}_n, \varepsilon_n)\}] = \frac{s_n \sigma_0^2}{n},
$$

and hence $\mathrm{E}(I_4^{**})^2 \leq s_n \sigma_0^2 / n$. Since $s_n = o(n)$, it follows from Chebyshev's inequality that $I_4^{**} = o_p(1)$.

Step 4: For $t = s_n, \ldots, n-1$, define

$$
\mathcal{A}_t = \left\{ \hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0} \leq \sqrt{\bar{c}}/2 \right\} \bigcap \left\{ \left\| \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0} \right\|_2 \leq \bar{c}_0 \frac{\sqrt{\kappa_n \log p}}{\sqrt{n}} \right\}.
$$

By (A.5), (A.12) and Condition (A1), we have $\Pr(\cap_{t=s_n}^{n-1} \mathcal{A}_t) \to 1$. Hence, we have $\Pr(I_1 =$

$I_1^*) \to 1$ and $\Pr(I_1^{**} = I_1^{***}) \to 1$ where

$$\begin{aligned}
I_1^* &= \sum_{t=s_n}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \varepsilon_{t+1} \mathcal{I}(\mathcal{A}_t), \\
I_1^{**} &= \sum_{t=s_n}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \varepsilon_{t+1} \mathcal{I}(\mathcal{A}_t), \\
I_1^{***} &= \sum_{t=s_n}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \varepsilon_{t+1}.
\end{aligned}$$

In the following, we prove $I_1^* = I_1^{**} + o_p(1)$. This further implies $I_1 = I_1^{***} + o_p(1)$. For any $t_0 > 0$,

$$\Pr\left(|I_1^* - I_1^{**}| > t_0\right) \le \frac{1}{nt_0^2} \mathrm{E} \left\{ \sum_{t=s_n}^{n-1} \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \right)^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \varepsilon_{t+1} \mathcal{I}(\mathcal{A}_t) \right\}^2$$

$$\begin{aligned}
&= \frac{1}{nt_0^2} \mathrm{E} \sum_{t=s_n}^{n-1} \left\{ \frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \right)^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \varepsilon_{t+1} \mathcal{I}(\mathcal{A}_t) \right\}^2 \\
&\le \frac{\sigma_0^2}{nt_0^2} \mathrm{E} \sum_{t=s_n}^{n-1} \left(\frac{\|\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\|_2^2}{\hat{\sigma}^2_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \mathcal{I}(\mathcal{A}_t) \|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\|_2^2 \right) \\
&\le \frac{4\bar{c}_0^2\sigma_0^2}{\bar{c}nt_0^2} \left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}} \right)^2 \mathrm{E} \sum_{t=s_n}^{n-1} \|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\|_2^2 = O(n^{-1}\kappa_n^2 \log p) = o(1),
\end{aligned}$$

where the second equality is due to (A.17) and the last equality is due to the condition that $\kappa_n^2 \log p = O(n/\log^2 n)$. This implies $I_1^* = I_1^{**} + o_p(1)$ and hence $I_1 = I_1^{***} + o_p(1)$. Similarly, we can show I_1 is asymptotically equivalent to

$$I_1^{****} = \frac{1}{\sqrt{n}} \sum_{t=s_n}^{n-1} \frac{1}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \varepsilon_{t+1}.$$

Observe that I_1^{****} is a mean zero martingale with respect to the filtration $\{\sigma(\mathcal{F}_t)\}_t$. Since $s_n = o(n)$, we have

$$\sum_{t=s_n}^{n-1} \mathrm{E} \left[\left\{ \frac{\varepsilon_{t+1}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}^T_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right\}^2 |\mathcal{F}_t \right] = \frac{n - s_n}{n} \sigma_0^2 \to \sigma_0^2.$$

Let $Z_{t+1,j_0} = X_{t+1,j_0} - \boldsymbol{\omega}^T_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}$ for $t \geq s_n$. It follows from Condition (A1) and (A.2) that

$$\mathrm{E}\left\{ \frac{1}{\sigma^4_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}^T_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}} \right)^4 |\mathcal{F}_t \right\} \leq \frac{c_0^4}{\bar{c}^2} \left(1 + \frac{c_0^2}{\bar{c}} \right)^2.$$

By Hölder's inequality, we have

$$\mathrm{E}\left(\frac{1}{\sigma^3_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} |Z_{t+1,j_0}|^3 |\mathcal{F}_t \right) \leq \left\{ \mathrm{E}\left(\frac{1}{\sigma^4_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} Z^4_{t+1,j_0} |\mathcal{F}_t \right) \right\}^{3/4} \leq \frac{c_0^3}{\bar{c}^{3/2}} \left(1 + \frac{c_0^2}{\bar{c}} \right)^{3/2}.$$

By condition, $\mathrm{E}|\varepsilon_{t+1}|^3 = O(1)$. Since ε_0 and $\boldsymbol{X}_0$ are independent, we have

$$\mathrm{E}\left(\frac{1}{\sigma^3_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} |Z_{t+1}|^3 |\varepsilon_{t+1}|^3 |\mathcal{F}_t \right) \leq \mathrm{E}|\varepsilon_{t+1}|^3 \mathrm{E}\left(\frac{1}{\sigma^3_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} |Z_{t+1}|^3 |\mathcal{F}_t \right) \leq \bar{c}_{**},$$

for some constant $\bar{c}_{**} > 0$. Therefore, for any $\delta_0 > 0$, it follows from Markov's inequality that

$$\begin{aligned} &\sum_{t=s_n}^{n-1} \mathrm{E}\left\{ \left| \frac{\varepsilon_{t+1} Z_{t+1,j_0}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} \right|^2 \mathcal{I}\left(\left| \frac{\varepsilon_{t+1} Z_{t+1,j_0}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} \right| \geq \delta_0 \right) |\mathcal{F}_t \right\} \\ &\leq \sum_{t=s_n}^{n-1} \frac{1}{n^{3/2}\delta_0} \mathrm{E}\left\{ \left| \frac{Z_{t+1,j_0}\varepsilon_{t+1}}{\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} \right|^3 |\mathcal{F}_t \right\} \leq \frac{\bar{c}_{**}}{\sqrt{n}\delta_0} \to 0. \end{aligned}$$

This verifies the Lindeberg's condition for I_1^{****}. It follows from the martingale central limit theorem that

$$I_1^{****} \xrightarrow{d} N(0, \sigma_0^2).$$

As a result, we have $I_1 \xrightarrow{d} N(0, \sigma_0^2)$. This completes the proof.

A.2 Proof of Theorem 3.1

We use a shorthand and write $\widehat{\mathcal{M}}_{j_0}^{(t)} = \widehat{\mathcal{M}}_{j_0}^{(-s_n)}$ for $t = 0, \dots, s_n - 1$. Let

$$\widehat{\boldsymbol{\Sigma}}^* = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)\boldsymbol{X}_i^T \quad \text{and} \quad \widehat{\boldsymbol{\Psi}}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)\boldsymbol{X}_i^T X_{i,j},$$

for any $j \in \{1, 2, \dots, p\}$. For any $\mathcal{M} \subseteq \mathbb{I}$, define

$$\begin{aligned}
\boldsymbol{\omega}_{\mathcal{M},j_0} &= \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}\boldsymbol{\Sigma}_{\mathcal{M},j_0}, \qquad \sigma^2_{\mathcal{M},j_0} = \boldsymbol{\Sigma}_{j_0,j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{\Sigma}_{\mathcal{M},j_0},\\
\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} &= \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}, \qquad \hat{\sigma}^2_{\mathcal{M},j_0} = \widehat{\boldsymbol{\Sigma}}_{j_0,j_0} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}^T\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0},\\
\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0} &= \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{*-1}\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0}, \qquad \hat{\sigma}^{*2}_{\mathcal{M},j_0} = \widehat{\boldsymbol{\Sigma}}^*_{j_0,j_0} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}^{*T}\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0},\\
\widetilde{\boldsymbol{\omega}}_{\mathcal{M},j_0} &= \widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0} + \sum_{j=1}^{p} \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{*-1}\left(\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},j_0} + \widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},\mathcal{M}}\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0}\right)(\widetilde{\beta}_j - \beta_{0,j}),\\
\widehat{Z}^*_{t+1,j_0} &= X_{t+1,j_0} - \widehat{\boldsymbol{\omega}}^{*T}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}, \quad \widetilde{Z}_{t+1,j_0} = X_{t+1,j_0} - \widetilde{\boldsymbol{\omega}}^{T}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}},\\
\hat{\xi}^{(j)}_{\mathcal{M},j_0} &= \widehat{\boldsymbol{\Psi}}^{(j)}_{j_0,j_0} - \widehat{\boldsymbol{\omega}}^{*T}_{\mathcal{M},j_0}\left(2\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},j_0} + \widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},\mathcal{M}}\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0}\right),\\
\tilde{\sigma}^2_{\mathcal{M},j_0} &= \hat{\sigma}^{*2}_{\mathcal{M},j_0} + \sum_{j=1}^{p} \hat{\xi}^{(j)}_{\mathcal{M},j_0}(\widetilde{\beta}_j - \beta_{0,j}).
\end{aligned}$$

Here, $\widetilde{\boldsymbol{\omega}}_{\mathcal{M},j_0}$ and $\tilde{\sigma}_{\mathcal{M},j_0}$ correspond to first-order approximations of $\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0}$ and $\hat{\sigma}_{\mathcal{M},j_0}$ around $\boldsymbol{\beta}_0$. We introduce the following lemmas before proving Theorem 3.1. The proof of Lemma A.4 is given in Section A.8.

Lemma A.4. *Under conditions in Theorem 3.1, we have*

$$\min_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M}, |\mathcal{M}|\le\kappa_n}} \sigma_{\mathcal{M},j_0} \ge \sqrt{\bar{c}}, \qquad \max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M}, |\mathcal{M}|\le\kappa_n}} \|\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 \le (\bar{c})^{-1/2}c_0, \tag{A.28}$$

where $\bar{c}$ and c_0 are defined in Condition (A2) and (A3*). Besides, the following events* *hold with probability tending to* 1,

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0}-\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 \le \bar{c}_0\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}+\eta_n\right), \tag{A.29}$$

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} |\hat{\sigma}_{\mathcal{M},j_0}-\sigma_{\mathcal{M},j_0}| \le \bar{c}_0\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}+\eta_n\right), \tag{A.30}$$

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0}-\widetilde{\boldsymbol{\omega}}_{\mathcal{M},j_0}\|_2 \le \bar{c}_0\eta_n^2, \qquad \max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \left|\hat{\sigma}^2_{\mathcal{M},j_0}-\tilde{\sigma}^2_{\mathcal{M},j_0}\right| \le \bar{c}_0\eta_n^2, \tag{A.31}$$

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \left|\hat{\sigma}^2_{\mathcal{M},j_0}-\hat{\sigma}^{*2}_{\mathcal{M},j_0}\right| \le \bar{c}_0\eta_n, \tag{A.32}$$

for some constant $\bar{c}_0>0$*. Moreover, we have*

$$\sum_{t=0}^{n-1}\frac{\widetilde{Z}_{t+1,j_0}\varepsilon_{t+1}}{\sqrt{n}}\left(\frac{1}{\hat{\sigma}^*_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}-\frac{\sum_j \hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}(\tilde{\beta}_j-\beta_{0,j})}{\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(-sn)}_{j_0},j_0}}\right)=\sum_{t=0}^{n-1}\frac{\widehat{Z}^*_{t+1,j_0}\varepsilon_{t+1}}{\sqrt{n}\hat{\sigma}^*_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}+o_p(1).$$

Similar to (A.25), we have

$$\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1=O(\sqrt{\kappa_n}\eta_n), \tag{A.33}$$

with probability tending to 1, under Condition (A5*).

For simplicity, we only consider the case where $l=1$. When $l>1$, assume we've shown the asymptotic normality of $\hat{\beta}^{(l-1)}_{j_0}$. Under the given conditions, we can show $\Gamma_n^{*,(l-2)}$ is lower bounded by $\sqrt{\bar{c}}/2$, with probability tending to 1. This implies $\hat{\beta}^{(l-1)}_{j_0}$ converges to β_{0,j_0} at a rate of $O_p(n^{-1/2})$. As a result, the estimator $\widehat{\boldsymbol{\beta}}^{(l-1)}=\widetilde{\boldsymbol{\beta}}+\boldsymbol{e}_{j_0,p}(\hat{\beta}^{(l-1)}_{j_0}-\widetilde{\beta}_{j_0})$ also satisfies the conditions in (A5*). The asymptotic normality of $\hat{\beta}^{(l)}_{j_0}$ can be similarly derived.

In the following, we omit the superscript and write $\hat{\beta}^{(1)}_{j_0}$ and $\Gamma_n^{*,(0)}$ as $\hat{\beta}_{j_0}$ and Γ_n^*. Let $\varepsilon_i=Y_i-\mu(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)$ for $i=0,1,\dots,n$. By definition, we have

$$\sqrt{n}\Gamma_n^*(\hat{\beta}_{j_0}-\widetilde{\beta}_{j_0})=\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}\widehat{Z}_{t+1,j_0}\left\{Y_{t+1}-\mu\left(X_{t+1}\tilde{\beta}_{0,j_0}+\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}\right)\right\} \tag{A.34}$$

By Condition (A1), we can show the following events occur with probability tending to 1,

$$\mathcal{M}_{j_0}\subseteq\widehat{\mathcal{M}}^{(t)}_{j_0},\quad |\widehat{\mathcal{M}}^{(t)}_{j_0}|\le\kappa_n,\quad t=0,\dots,n-1. \tag{A.35}$$

Besides, similar to (A.6) and (A.12), we can show

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} |\hat{\sigma}^{*2}_{\mathcal{M},j_0}-\sigma^2_{\mathcal{M},j_0}| \le \bar{c}_0\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right), \tag{A.36}$$

for some constant $\bar{c}_0>0$, and

$$\min_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \hat{\sigma}_{\mathcal{M},j_0} \ge \sqrt{\bar{c}}/2 \quad\text{and}\quad \min_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \hat{\sigma}^{*}_{\mathcal{M},j_0} \ge \sqrt{\bar{c}}/2, \tag{A.37}$$

with probability tending to 1.

Under the events defined in (A.35), we have for $t=0,1,\dots,n-1$,

$$\boldsymbol{X}_{t+1}^T\boldsymbol{\beta}_0 = X_{t+1,j_0}\beta_{0,j_0}+\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}\boldsymbol{\beta}_{0,\widehat{\mathcal{M}}^{(t)}_{j_0}}.$$

Hence, using a second order Taylor expansion, we have

$$\begin{aligned} &\mu(\boldsymbol{X}_{t+1}^T\boldsymbol{\beta}_0)=\mu\left(X_{t+1,j_0}\beta_{0,j_0}+\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}\boldsymbol{\beta}_{0,\widehat{\mathcal{M}}^{(t)}_{j_0}}\right)\\ =\;& \mu\left(X_{t+1,j_0}\widetilde{\beta}_{j_0}+\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}\right)+b''\left(X_{t+1,j_0}\widetilde{\beta}_{j_0}+\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}\right)\\ \times\;& \left(X_{t+1,j_0}(\beta_{0,j_0}-\widetilde{\beta}_{j_0})+\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}(\boldsymbol{\beta}_{0,\widehat{\mathcal{M}}^{(t)}_{j_0}}-\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}_{j_0}})\right)\\ +\;& \frac{1}{2}b'''\left(\boldsymbol{X}^T_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}\widetilde{\boldsymbol{\beta}}^*_t\right)\left(\boldsymbol{X}^T_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}(\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}-\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}})\right)^2, \end{aligned}$$

for some $\widetilde{\boldsymbol{\beta}}^*_t\in\mathbb{R}^{1+|\widehat{\mathcal{M}}^{(t)}_{j_0}|}$ lying on the line segment joining $\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}$ and $\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}$. Let R^*_t be the second order Remainder term. Under the events defined in (A.35), we have

$$\begin{aligned} &\left|\boldsymbol{X}^T_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}\widetilde{\boldsymbol{\beta}}^*_t\right| \le \left|\boldsymbol{X}^T_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}\right|+\left|\boldsymbol{X}^T_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}(\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}-\widetilde{\boldsymbol{\beta}}^*_t)\right|\\ =\;& |\boldsymbol{X}^T_{t+1}\boldsymbol{\beta}_0|+\left|\boldsymbol{X}^T_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}(\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}-\widetilde{\boldsymbol{\beta}}^*_t)\right| \le \bar{\omega}+\omega_0\left\|\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}-\widetilde{\boldsymbol{\beta}}^*_t\right\|_1\\ \le\;& \bar{\omega}+\omega_0\left\|\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}\right\|_1, \end{aligned}$$

where the second inequality is due to Condition (A4*). By Condition (A5*) and (A.33), we have with probability tending to 1,

$$\omega_0\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \le \omega_0\eta_n^{(1)} \le \bar{\omega}.$$

Since $b'''(\cdot)$ is continuous, $\sup_{|z|\le 2\bar{\omega}}|b'''(z)|$ is upper bounded by some constant $c_* > 0$. Therefore, we have with probability tending to 1 that

$$\max_{t=0,\dots,n-1}\left|b'''\left(\boldsymbol{X}^T_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}\widetilde{\boldsymbol{\beta}}^*_t\right)\right| \le c_*. \tag{A.38}$$

Under the event defined in (A.38), we have

$$|R^*_t| \le \frac{c_*}{2}\left|\boldsymbol{X}^T_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}(\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}}-\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}^{(t)}_{j_0}})\right|^2, \tag{A.39}$$

for any t. Note that

$$\left|\widehat{Z}_{t+1,j_0}\right| \le |X_{t+1,j_0}| + \left\|\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}-\boldsymbol{\omega}_{\widehat{\mathcal{M}}^{(t)}_{j_0}}\right\|_2\|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}\|_2 + |\boldsymbol{\omega}^T_{\widehat{\mathcal{M}}^{(t)}_{j_0}}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}|.$$

By Condition (A1*) and (A4*), we have almost surely,

$$|\widehat{Z}_{t+1,j_0}| \le \omega_0 + \sqrt{\kappa_n}\omega_0\max_{\substack{\mathcal{M}\subseteq\mathbb{I}\\|\mathcal{M}|\le\kappa_n}}\|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0}-\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 + |\boldsymbol{\omega}^T_{\widehat{\mathcal{M}}^{(t)}_{j_0}}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}|. \tag{A.40}$$

The second term on the RHS of (A.40) is $o(1)$ with probability tending to 1, by (A.29) and Condition (A5*). Thus, we have with probability tending to 1 that

$$|\widehat{Z}_{t+1,j_0}| \le 2\omega_0 + |\boldsymbol{\omega}^T_{\widehat{\mathcal{M}}^{(t)}_{j_0}}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}|, \quad \forall t. \tag{A.41}$$

Under the events defined in (A.35), (A.37), (A.39) and (A.41), we have

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\widehat{Z}_{t+1,j_0}\left\{Y_{t+1}-\mu\left(X_{t+1}\tilde{\beta}_{0,j_0}+\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)\right\}\right.\\
- \quad & \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}-\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{\widehat{Z}_{t+1,j_0}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}b''\left(\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)\\
\times \quad & \left.\left(X_{t+1,j_0}(\beta_{0,j_0}-\tilde{\beta}_{j_0})+\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T(\boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}-\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}})\right)\right| \le \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}|R_t^*||\widehat{Z}_{t+1,j_0}|\\
\le \quad & \frac{c_*}{\sqrt{n\bar{c}}}\sum_{t=0}^{n-1}(2\omega_0+|\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}|)\left|\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T(\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}-\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}})\right|^2. \quad \text{(A.42)}
\end{aligned}$$

Similar to (A.9), we can show

$$\sum_{t=0}^{n-1}\left|\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T(\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}-\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}})\right|^2 = O(n\eta_n^2),$$

$$\sum_{t=0}^{n-1}|\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}|\left|\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T(\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}-\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}})\right|^2 = O(n\eta_n^2),$$

with probability tending to 1. It follows that

$$\frac{c_*}{\sqrt{n\bar{c}}}\sum_{t=0}^{n-1}(2\omega_0+|\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}|)\left|\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T(\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}-\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}})\right|^2 = o_p(1), \quad \text{(A.43)}$$

under the condition $\sqrt{n}\eta_n^2 = o(1)$ in (A5*). Hence, we've shown

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{1}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\widehat{Z}_{t+1,j_0}\left\{Y_{t+1}-\mu\left(X_{t+1}\tilde{\beta}_{0,j_0}+\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)\right\}\\
= \quad & \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{\widehat{Z}_{t+1,j_0}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}b''\left(\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}-\boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)\\
+ \quad & \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}+\sqrt{n}\Gamma_n^*(\beta_{0,j_0}-\tilde{\beta}_{j_0})+o_p(1).
\end{aligned}$$

In view of (A.34), we have

$$\sqrt{n}\Gamma_n^*(\hat{\beta}_{j_0}-\beta_{0,j_0}) = o_p(1) + \underbrace{\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}}_{I_1}$$

$$+ \quad \underbrace{\frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{\widehat{Z}_{t+1,j_0}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} b''\left(\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}-\boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)}_{I_2}.$$

In the following, we break the proof into two steps. In the first step, we prove $I_2 = o_p(1)$. In the second step, we show $I_1 \xrightarrow{d} N(0,\phi_0)$. This implies $\sqrt{n}\Gamma_n^*(\hat{\beta}_{j_0}-\beta_{0,j_0}) \xrightarrow{d} N(0,\phi_0)$. By Condition (A7*), $\hat{\phi}$ is consistent to ϕ_0. It follows from Slutsky's theorem that

$$\frac{\sqrt{n}\Gamma_n^*(\hat{\beta}_{j_0}-\beta_{0,j_0})}{\hat{\phi}^{1/2}} \xrightarrow{d} N(0,1).$$

The proof is hence completed.

Step 1: Under the events defined in (A.35), using a first order Taylor expansion, we have

$$b''\left(\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}\right) = b''(\boldsymbol{X}_{t+1}^T\boldsymbol{\beta}_0)$$

$$+ \quad \underbrace{\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\left(\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}-\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}\right) b'''\left(\boldsymbol{X}_{t+1,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\widetilde{\boldsymbol{\beta}}_t^{**}\right)}_{R_t^{**}},$$

for some $\widetilde{\boldsymbol{\beta}}_t^{**} \in \mathbb{R}^{1+|\widehat{\mathcal{M}}_{j_0}^{(t)}|}$ lying on the line segment joining $\boldsymbol{\beta}_{0,\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}$ and $\widetilde{\boldsymbol{\beta}}_{\{j_0\}\cup\widehat{\mathcal{M}}_{j_0}^{(t)}}$. Let

$$I_2^* = \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{\widehat{Z}_{t+1,j_0}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} b''\left(\boldsymbol{X}_{t+1}^T\boldsymbol{\beta}_0\right)\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}-\boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right),$$

we have

$$|I_2 - I_2^*| \le \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\frac{|\widehat{Z}_{t+1,j_0}|}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}|R_t^{**}|\left|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}-\boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}\right)\right|_2.$$

Similar to (A.39), (A.42) and (A.43), we can show

$$|I_2 - I_2^*| \leq \frac{c_*}{\sqrt{n}} \sum_{t=0}^{n-1} \frac{|\widehat{Z}_{t+1,j_0}|}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left| \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right|_2^2 = o(1),$$

with probability tending to 1. Thus, to prove $I_2 = o_p(1)$, it suffices to show $I_2^* = o_p(1)$.

Similar to the proof of Theorem 2.1, we can show under the given conditions that

$$\left| \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \frac{\widehat{Z}_{t+1,j_0} - Z_{t+1,j_0}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} b'' \left(\boldsymbol{X}_{t+1}^T \boldsymbol{\beta}_0 \right) \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) \right| = o_p(1),$$

where $Z_{t+1,j_0} = X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}$, and

$$\frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \left(\frac{Z_{t+1,j_0}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} - \frac{Z_{t+1,j_0}}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \right) b'' \left(\boldsymbol{X}_{t+1}^T \boldsymbol{\beta}_0 \right) \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) = o_p(1).$$

This implies $I_2^* = I_2^{**} + o_p(1)$, where

$$I_2^{**} = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \frac{Z_{t+1,j_0}}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} b'' \left(\boldsymbol{X}_{t+1}^T \boldsymbol{\beta}_0 \right) \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \left(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right).$$

Note that I_2^{**} can be further bounded from above by $\max_j |I_{2,j}| \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1$ where

$$I_{2,j} = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \frac{Z_{t+1,j_0}}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} b'' \left(\boldsymbol{X}_{t+1}^T \boldsymbol{\beta}_0 \right) \boldsymbol{X}_{t+1,j} \mathcal{I} \left(j \in \widehat{\mathcal{M}}_{j_0}^{(t)} \right).$$

Similar to Lemma A.3, we can show $\max_j |I_{2,j}| = O_p(\sqrt{\log p})$. This together with (A.33) and Condition (A5*) implies $\max_j |I_{2,j}| \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = o_p(1)$ and hence $I_2^{**} = o_p(1)$. This proves $I_2 = o_p(1)$.

Step 2: By Taylor's theorem, we have for any $\mathcal{M} \subseteq \mathbb{I}_{j_0}$,

$$\frac{1}{\hat{\sigma}_{\mathcal{M},j_0}} - \frac{1}{\hat{\sigma}_{\mathcal{M},j_0}^*} + \frac{\hat{\sigma}_{\mathcal{M},j_0}^2 - \hat{\sigma}_{\mathcal{M},j_0}^{*2}}{\hat{\sigma}_{\mathcal{M},j_0}^{*3}} = \frac{(\hat{\sigma}_{\mathcal{M},j_0}^2 - \hat{\sigma}_{\mathcal{M},j_0}^{*2})^2}{2\{\rho_{\mathcal{M}} \hat{\sigma}_{\mathcal{M},j_0} + (1 - \rho_{\mathcal{M}}) \hat{\sigma}_{\mathcal{M},j_0}^*\}^5},$$

for some $0 < \rho_{\mathcal{M}} < 1$. By (A.32) and (A.37), the second-order remainder term satisfies

$$\max_{\mathcal{M} \subseteq \mathbb{I}_{j_0}, |\mathcal{M}| \leq \kappa_n} \left| \frac{(\hat{\sigma}_{\mathcal{M},j_0}^2 - \hat{\sigma}_{\mathcal{M},j_0}^{*2})^2}{2\{\rho_{\mathcal{M}} \hat{\sigma}_{\mathcal{M},j_0} + (1 - \rho_{\mathcal{M}}) \hat{\sigma}_{\mathcal{M},j_0}^*\}^5} \right| \leq \frac{16 \bar{c}_0^2 \eta_n^2}{\bar{c}^{5/2}}, \tag{A.44}$$

with probability tending to 1.

Besides, it follows from (A.31) and (A.37) that

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\left|\frac{\hat{\sigma}^2_{\mathcal{M},j_0}-\hat{\sigma}^{*2}_{\mathcal{M},j_0}-\sum_j\hat{\xi}^{(j)}_{\mathcal{M},j_0}(\tilde{\beta}_j-\beta_{0,j})}{\hat{\sigma}^{*3}_{\mathcal{M},j_0}}\right| = \max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\left|\frac{\hat{\sigma}^2_{\mathcal{M},j_0}-\tilde{\sigma}^2_{\mathcal{M},j_0}}{\hat{\sigma}^{*3}_{\mathcal{M},j_0}}\right| \le \frac{8\bar{c}_0\eta_n^2}{\bar{c}^{3/2}},$$

with probability tending to 1. Combining this together with (A.44) yields

$$\Pr\left(\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\left|\frac{1}{\hat{\sigma}_{\mathcal{M},j_0}}-\frac{1}{\hat{\sigma}^*_{\mathcal{M},j_0}}+\frac{\sum_j\hat{\xi}^{(j)}_{\mathcal{M},j_0}(\tilde{\beta}_j-\beta_{0,j})}{\hat{\sigma}^{*3}_{\mathcal{M},j_0}}\right|\le\bar{c}_1\eta_n^2\right)\to 1,$$

for some constant $\bar{c}_1>0$. By Condition (A1*), we have

$$\frac{1}{\sqrt{n}}\left|\sum_{t=0}^{n-1}\frac{\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}}{\hat{\sigma}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}-\sum_{t=0}^{n-1}\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}\left(\frac{1}{\hat{\sigma}^*_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}-\frac{\sum_j\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}(\tilde{\beta}_j-\beta_{0,j})}{\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}\right)\right|$$
$$\le\sqrt{n}\bar{c}_1\eta_n^2\max_t\frac{1}{n}\sum_{t=0}^{n-1}|\varepsilon_{t+1}||\widehat{Z}_{t+1,j_0}|,$$

with probability tending to 1. Similar to (A.23), we can show $\sum_{t=0}^{n-1}|\boldsymbol{\omega}^T_{\widehat{\mathcal{M}}^{(t)}_{j_0}}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}|^2=O_p(n)$. This together with (A.41) and Cauchy-Schwarz inequality yields

$$\sum_{t=0}^{n-1}|\widehat{Z}_{t+1,j_0}|^2\le 8n\omega_0^2+2\sum_{t=0}^{n-1}|\boldsymbol{\omega}^T_{\widehat{\mathcal{M}}^{(t)}_{j_0}}\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}}|^2=O_p(n). \tag{A.45}$$

In addition, we have $\sum_{t=0}^{n-1}\varepsilon^2_{t+1}=O_p(n)$, under (A6*). It follows from Cauchy-Schwarz inequality that

$$\frac{1}{n}\sum_{t=0}^{n-1}|\varepsilon_{t+1}||\widehat{Z}_{t+1,j_0}|\le\left(\frac{1}{n}\sum_{t=0}^{n-1}|\widehat{Z}_{t+1,j_0}|^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=0}^{n-1}\varepsilon^2_{t+1}\right)^{1/2}=O_p(1),$$

and hence

$$\frac{1}{\sqrt{n}}\left|\sum_{t=0}^{n-1}\frac{\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}}{\hat{\sigma}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}-\sum_{t=0}^{n-1}\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}\left(\frac{1}{\hat{\sigma}^*_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}-\frac{\sum_j\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}(\tilde{\beta}_j-\beta_{0,j})}{\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}\right)\right|=O_p(\sqrt{n}\eta_n^2)$$
$$=o_p(1),$$

under Condition (A5*).

Using similar arguments in bounding $I_2^{(2)}$ in the proof of Theorem 2.1, we can show

$$\frac{1}{\sqrt{n}}\left|\sum_{t=0}^{n-1}(\widehat{Z}_{t+1,j_0}-\widetilde{Z}_{t+1,j_0})\varepsilon_{t+1}\left(\frac{1}{\hat{\sigma}^*_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}-\frac{\sum_j \hat{\xi}^{(j)}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}(\tilde{\beta}_j-\beta_{0,j})}{\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\right)\right| = o_p(1).$$

It follows that

$$\left|\sum_{t=0}^{n-1}\left\{\frac{\widehat{Z}_{t+1,j_0}\varepsilon_{t+1}}{\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}-\widetilde{Z}_{t+1,j_0}\varepsilon_{t+1}\left(\frac{1}{\hat{\sigma}^*_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}-\frac{\sum_j \hat{\xi}^{(j)}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}(\tilde{\beta}_j-\beta_{0,j})}{\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\right)\right\}\right| = o_p(\sqrt{n}).$$

Therefore, we've shown $I_1 = I_1^* + o_p(1)$ where

$$I_1^* = \frac{1}{\sqrt{n}}\sum_{t=0}^{n-1}\widetilde{Z}_{t+1,j_0}\varepsilon_{t+1}\left(\frac{1}{\hat{\sigma}^*_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}-\frac{\sum_j \hat{\xi}^{(j)}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}(\tilde{\beta}_j-\beta_{0,j})}{\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}}\right).$$

In Lemma A.4, we further show I_1^* is equivalent to

$$I_1^{**} \equiv \sqrt{n}\sum_{t=0}^{n-1}\frac{\widehat{Z}^*_{t+1,j_0}\varepsilon_{t+1}}{\hat{\sigma}^*_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}.$$

Hence, we have $I_1 = I_1^{**} + o_p(1)$. Unlike $\widetilde{Z}_{t+1,j_0}$ and $\tilde{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$, $\widehat{Z}^*_{t+1,j_0}$ and $\hat{\sigma}^*_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ didn't depend on the initial estimator $\widetilde{\boldsymbol{\beta}}$. As a result, $\widehat{Z}^*_{t+1,j_0}$ and $\hat{\sigma}^*_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ are fixed given $\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_n\}$ and $\widehat{\mathcal{M}}_{j_0}^{(t)}$. Following the arguments in the proof of Theorem 2.1, we can show

$$I_1^{**} = \frac{1}{\sqrt{n}}\sum_{t=s_n}^{n-1}\frac{Z_{t+1,j_0}\varepsilon_{t+1}}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} + o_p(1) \text{ and } \frac{1}{\sqrt{n}}\sum_{t=s_n}^{n-1}\frac{Z_{t+1,j_0}\varepsilon_{t+1}}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \stackrel{d}{\to} N(0,\phi_0).$$

By Slutsky's theorem, we have $I_1 \stackrel{d}{\to} N(0,\phi_0)$. The proof is hence completed.

A.3 Proof of Theorem 3.2

Recall that $\mathbb{I} = [1,\dots,p]$ and $\mathbb{I}_{j_0} = \mathbb{I} - \{j_0\}$. By (19) and Lemma A.5, we have

$$\sqrt{n}\mathrm{L}(\hat{\beta}^{DL}_{j_0},\alpha) = 2z_{\frac{\alpha}{2}}\sqrt{\phi_0 \boldsymbol{e}^T_{j_0,p}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_{j_0,p}} + o_p(1) = \frac{2z_{\frac{\alpha}{2}}\sqrt{\phi_0}}{\sigma_{\mathbb{I}_{j_0},j_0}} + o_p(1). \qquad \text{(A.46)}$$

It follows from (16) that

$$\sqrt{n}\mathrm{L}(\hat{\beta}_{j_0}, \alpha) = \frac{2z_{\alpha/2}\sqrt{\phi_0}}{s_n\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0}/n + \sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0}/n} + o_p(1). \tag{A.47}$$

With some calculations, we have

$$\begin{aligned} &\frac{2z_{\alpha/2}\sqrt{\phi_0}}{\sigma_{\mathbb{I}_{j_0}, j_0}} - \frac{2z_{\alpha/2}\sqrt{\phi_0}}{s_n\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0}/n + \sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0}/n} \\ = \;& 2z_{\alpha/2}\sqrt{\phi_0}\frac{s_n\{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} - \sigma_{\mathbb{I}_{j_0}, j_0}\}/n + \sum_{t=s_n}^{n-1}\{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0} - \sigma_{\mathbb{I}_{j_0}, j_0}\}/n}{\sigma_{\mathbb{I}_{j_0}, j_0}\{s_n\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0}/n + \sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0}/n\}}. \end{aligned} \tag{A.48}$$

For any $\mathcal{M} \subseteq \mathbb{I}_{j_0}$, we have

$$\begin{aligned} \sigma_{\mathcal{M}, j_0}^2 &= \mathrm{E}|X_{0,j_0} - \boldsymbol{\omega}_{\mathcal{M}, j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|^2 b''(\boldsymbol{X}_0^T\boldsymbol{\beta}_0) = \arg\min_{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}}\mathrm{E}|X_{0,j_0} - \boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^2 b''(\boldsymbol{X}_0^T\boldsymbol{\beta}_0) \\ &\geq \arg\min_{\boldsymbol{a}\in\mathbb{R}^{p-1}}\mathrm{E}|X_{0,j_0} - \boldsymbol{a}^T\boldsymbol{X}_{0,\mathbb{I}_{j_0}}|^2 b''(\boldsymbol{X}_0^T\boldsymbol{\beta}_0) = \sigma_{\mathbb{I}_{j_0}, j_0}^2. \end{aligned}$$

This shows $\sigma_{\mathcal{M}, j_0} \geq \sigma_{\mathbb{I}_{j_0}, j_0}$ for any $\mathcal{M} \subseteq \mathbb{I}_{j_0}$. Hence, the numerator of the RHS of (A.48) is nonnegative.

On the other hand, by Condition (A4*), we have $|\boldsymbol{X}_0^T\boldsymbol{\beta}_0| \leq \bar{\omega}$ and hence $b''(\boldsymbol{X}_0^T\boldsymbol{\beta}_0) \leq \bar{k}$. Therefore,

$$\sigma_{\mathcal{M}, j_0}^2 = \arg\min_{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}}\mathrm{E}|X_{0,j_0} - \boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^2 b''(\boldsymbol{X}_0^T\boldsymbol{\beta}_0) \leq \bar{k}\mathrm{E}|X_{0,j_0}^2| \leq \bar{k}\|X_{0,j_0}\|_{\psi_2}^2 = \bar{k}c_0^2, \tag{A.49}$$

where the last inequality is due to Condition (A3*). This implies

$$\begin{aligned} &\frac{2z_{\alpha/2}\sqrt{\phi_0}}{\sigma_{\mathbb{I}_{j_0}, j_0}} - \frac{2z_{\alpha/2}\sqrt{\phi_0}}{s_n\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0}/n + \sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0}/n} \\ \geq \;& \frac{2z_{\alpha/2}\sqrt{\phi_0}}{\bar{k}c_0^2}\left(s_n\{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, j_0} - \sigma_{\mathbb{I}_{j_0}, j_0}\}/n + \sum_{t=s_n}^{n-1}\{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)}, j_0} - \sigma_{\mathbb{I}_{j_0}, j_0}\}/n\right). \end{aligned} \tag{A.50}$$

Besides, it follows from (A.49) that

$$\sigma_{\mathcal{M}, j_0} - \sigma_{\mathbb{I}_{j_0}, j_0} = \frac{\sigma_{\mathcal{M}, j_0}^2 - \sigma_{\mathbb{I}_{j_0}, j_0}^2}{\sigma_{\mathcal{M}, j_0} + \sigma_{\mathbb{I}_{j_0}, j_0}} \geq \frac{\sigma_{\mathcal{M}, j_0}^2 - \sigma_{\mathbb{I}_{j_0}, j_0}^2}{2\sqrt{\bar{k}}c_0},$$

for any $\mathcal{M} \subseteq \mathbb{I}_{j_0}$. This together with (A.50) gives

$$\frac{2z_{\alpha/2}\sqrt{\phi_0}}{\sigma_{\mathbb{I}_{j_0},j_0}} - \frac{2z_{\alpha/2}\sqrt{\phi_0}}{s_n\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}/n + \sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}/n} \quad \text{(A.51)}$$
$$\geq \frac{z_{\alpha/2}\sqrt{\phi_0}}{\bar{k}^{3/2}c_0^3}\left(s_n\{\sigma^2_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0} - \sigma^2_{\mathbb{I}_{j_0},j_0}\}/n + \sum_{t=s_n}^{n-1}\{\sigma^2_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0} - \sigma^2_{\mathbb{I}_{j_0},j_0}\}/n\right).$$

For any $\mathcal{M} \subseteq \mathbb{I}_{j_0}$, define

$$\boldsymbol{\Omega}_{\mathcal{M},j_0} = \left(\boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathbb{I}_{j_0}\cap\mathcal{M}^c} - \boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathcal{M}}\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},\mathbb{I}_{j_0}\cap\mathcal{M}^c}\right)^{-1}.$$

It follows from Lemma A.5 that

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}} & \boldsymbol{\Sigma}_{\mathcal{M},\mathbb{I}_{j_0}\cap\mathcal{M}^c} \\ \boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathcal{M}} & \boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathbb{I}_{j_0}\cap\mathcal{M}^c} \end{pmatrix}^{-1} - \begin{pmatrix} \boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{pmatrix}$$
$$= \begin{pmatrix} \boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},\mathbb{I}_{j_0}\cap\mathcal{M}^c}\boldsymbol{\Omega}_{\mathcal{M},j_0}\boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathcal{M}}\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}} & -\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},\mathbb{I}_{j_0}\cap\mathcal{M}^c}\boldsymbol{\Omega}_{\mathcal{M},j_0} \\ -\boldsymbol{\Omega}_{\mathcal{M},j_0}\boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathcal{M}}\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}} & \boldsymbol{\Omega}_{\mathcal{M},j_0} \end{pmatrix}.$$

Therefore,

$$\begin{aligned}
&\boldsymbol{\Sigma}^T_{\mathbb{I}_{j_0},j_0}\boldsymbol{\Sigma}^{-1}_{\mathbb{I}_{j_0},\mathbb{I}_{j_0}}\boldsymbol{\Sigma}_{\mathbb{I}_{j_0},j_0} - \boldsymbol{\Sigma}^T_{\mathcal{M},j_0}\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},j_0} = (\boldsymbol{\Sigma}_{j_0,\mathcal{M}}, \boldsymbol{\Sigma}_{j_0,\mathbb{I}_{j_0}\cap\mathcal{M}^c}) \\
\times &\begin{pmatrix} \boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},\mathbb{I}_{j_0}\cap\mathcal{M}^c}\boldsymbol{\Omega}_{\mathcal{M},j_0}\boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathcal{M}}\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}} & -\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},\mathbb{I}_{j_0}\cap\mathcal{M}^c}\boldsymbol{\Omega}_{\mathcal{M},j_0} \\ -\boldsymbol{\Omega}_{\mathcal{M},j_0}\boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathcal{M}}\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}} & \boldsymbol{\Omega}_{\mathcal{M},j_0} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{M},j_0} \\ \boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,j_0} \end{pmatrix} \\
= &(\boldsymbol{\Sigma}_{j_0,\mathbb{I}_{j_0}\cap\mathcal{M}^c} - \boldsymbol{\omega}^T_{\mathcal{M},j_0}\boldsymbol{\Sigma}_{\mathcal{M},\mathbb{I}_{j_0}\cap\mathcal{M}^c})\boldsymbol{\Omega}_{\mathcal{M},j_0}(\boldsymbol{\Sigma}_{j_0,\mathbb{I}_{j_0}\cap\mathcal{M}^c} - \boldsymbol{\omega}^T_{\mathcal{M},j_0}\boldsymbol{\Sigma}_{\mathcal{M},\mathbb{I}_{j_0}\cap\mathcal{M}^c})^T \\
\geq &\lambda_{\min}(\boldsymbol{\Omega}_{\mathcal{M},j_0})\|\boldsymbol{\Sigma}_{j_0,\mathbb{I}_{j_0}\cap\mathcal{M}^c} - \boldsymbol{\omega}^T_{\mathcal{M},j_0}\boldsymbol{\Sigma}_{\mathcal{M},\mathbb{I}_{j_0}\cap\mathcal{M}^c}\|_2^2 = \lambda_{\min}(\boldsymbol{\Omega}_{\mathcal{M},j_0})\|\boldsymbol{\xi}_{\mathcal{M},j_0}\|_2^2.
\end{aligned}$$

By definition, we have

$$\lambda_{\min}(\boldsymbol{\Omega}_{\mathcal{M},j_0}) \geq \lambda_{\min}\left\{\left(\boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathbb{I}_{j_0}\cap\mathcal{M}^c}\right)^{-1}\right\} = \left\{\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathbb{I}_{j_0}\cap\mathcal{M}^c,\mathbb{I}_{j_0}\cap\mathcal{M}^c}\right)\right\}^{-1} \geq \{\lambda_{\max}(\boldsymbol{\Sigma})\}^{-1} \geq \frac{1}{c_0^2},$$

where the last inequality follows from (A.22). It follows that

$$\boldsymbol{\Sigma}^T_{\mathbb{I}_{j_0},j_0}\boldsymbol{\Sigma}^{-1}_{\mathbb{I}_{j_0},\mathbb{I}_{j_0}}\boldsymbol{\Sigma}_{\mathbb{I}_{j_0},j_0} - \boldsymbol{\Sigma}^T_{\mathcal{M},j_0}\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},j_0} \geq \frac{1}{c_0^2}\|\boldsymbol{\xi}_{\mathcal{M},j_0}\|_2^2.$$

Note that we have

$$\sigma^2_{\mathcal{M},j_0} - \sigma^2_{\mathbb{I}_{j_0},j_0} = \boldsymbol{\Sigma}_{j_0,j_0} - \boldsymbol{\Sigma}^c_{\mathcal{M},j_0}\boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},j_0} - (\boldsymbol{\Sigma}_{j_0,j_0} - \boldsymbol{\Sigma}^c_{\mathbb{I}_{j_0},j_0}\boldsymbol{\Sigma}^{-1}_{\mathbb{I}_{j_0},\mathbb{I}_{j_0}}\boldsymbol{\Sigma}_{\mathbb{I}_{j_0},j_0}).$$

This further implies

$$\sigma^2_{\mathcal{M},j_0} - \sigma^2_{\mathbb{I}_{j_0},j_0} \geq \frac{1}{c_0^2}\|\boldsymbol{\xi}_{\mathcal{M},j_0}\|_2^2,$$

for any $\mathcal{M} \subseteq \mathbb{I}_{j_0}$. By (A.51), we have

$$\frac{2z_{\alpha/2}\sqrt{\phi_0}}{\sigma_{\mathbb{I}_{j_0},j_0}} - \frac{2z_{\alpha/2}\sqrt{\phi_0}}{s_n\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}/n + \sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}/n} \geq \frac{\sqrt{\phi_0}z_{\alpha/2}}{\bar{k}^{3/2}c_0^5}\left(\frac{s_n}{n}\|\boldsymbol{\xi}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}\|_2^2 + \frac{1}{n}\sum_{t=s_n}^{n-1}\|\boldsymbol{\xi}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\|_2^2\right).$$

In view of (A.46) and (A.47), we've shown

$$\sqrt{n}\mathrm{L}(\hat{\beta}_{j_0}^{DL},\alpha) \geq \sqrt{n}\mathrm{L}(\hat{\beta}_{j_0},\alpha) + \frac{\sqrt{\phi_0}z_{\alpha/2}}{\bar{k}^{3/2}c_0^5}\left(\frac{s_n}{n}\|\boldsymbol{\xi}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}\|_2^2 + \frac{1}{n}\sum_{t=s_n}^{n-1}\|\boldsymbol{\xi}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\|_2^2\right) + o_p(1).$$

The proof is completed by noting that $\sqrt{n}\mathrm{L}(\hat{\beta}_{j_0}^{DL},\alpha) = \sqrt{n}\mathrm{L}(\hat{\beta}_{j_0}^{DS},\alpha) + o_p(1)$.

A.4 Proof of Theorem 3.3

Under the given conditions, using similar arguments in (A.10), we can show the following event occurs with probability tending to 1,

$$\widehat{\mathcal{M}}_{j_0}^{(-s_n)} = \widehat{\mathcal{M}}_{j_0}^{(s_n)} = \cdots = \widehat{\mathcal{M}}_{j_0}^{(n)} = \mathcal{M}_{j_0}.$$

Under these events, we have

$$\frac{2z_{\alpha/2}\sqrt{\phi_0}}{s_n\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}/n + \sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}/n} = \frac{2z_{\alpha/2}\sqrt{\phi_0}}{\sigma_{\mathcal{M}_{j_0},j_0}}. \tag{A.52}$$

By (16) and (23), for any sufficiently small $\varepsilon_0 > 0$, the following events occur with probability tending to 1,

$$\limsup_n \left|\sqrt{n}\mathrm{L}(\hat{\beta}_{j_0}^{(l)},\alpha) - \frac{2z_{\alpha/2}\sqrt{\phi_0}}{s_n\sigma_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}/n + \sum_{t=s_n}^{n-1}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}/n}\right| \leq \frac{\varepsilon_0}{2}, \tag{A.53}$$

$$\limsup_n \left|\sqrt{n}\mathrm{L}(\hat{\beta}_{j_0}^{oracle},\alpha) - \frac{2z_{\alpha/2}\sqrt{\phi_0}}{\sigma_{\mathcal{M}_{j_0},j_0}}\right| \leq \frac{\varepsilon_0}{2}. \tag{A.54}$$

Conditional on the events defined in (A.52)-(A.54), we have

$$\limsup_n \left| \sqrt{n}\mathrm{L}(\hat{\beta}_{j_0}^{(l)}, \alpha) - \sqrt{n}\mathrm{L}(\hat{\beta}_{j_0}^{oracle}, \alpha) \right| \leq \varepsilon_0.$$

The proof is hence completed.

A.5 Proof of Lemma A.1

We first prove (A.1). Condition (A2) states that

$$\min_{\substack{\mathcal{M}\subseteq[1,\dots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\leq\kappa_n}} \inf_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|+1} \\ \|\boldsymbol{a}\|_2\geq 1}} \boldsymbol{a}^T \boldsymbol{\Sigma}_{j_0\cup\mathcal{M}, j_0\cup\mathcal{M}} \boldsymbol{a} \geq \bar{c}. \tag{A.55}$$

Note that

$$\begin{aligned} \sigma_{\mathcal{M},j_0}^2 &= \boldsymbol{\Sigma}_{j_0,j_0} - \boldsymbol{\Sigma}_{j_0,\mathcal{M}} \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1} \boldsymbol{\Sigma}_{\mathcal{M},j_0} = (1, -\boldsymbol{\omega}_{\mathcal{M},j_0}) \begin{pmatrix} \boldsymbol{\Sigma}_{j_0,j_0} & \boldsymbol{\Sigma}_{j_0,\mathcal{M}} \\ \boldsymbol{\Sigma}_{\mathcal{M},j_0} & \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}} \end{pmatrix} \begin{pmatrix} 1 \\ -\boldsymbol{\omega}_{\mathcal{M},j_0}^T \end{pmatrix} \\ &\geq \inf_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|+1} \\ \|\boldsymbol{a}\|_2\geq 1}} \boldsymbol{a}^T \boldsymbol{\Sigma}_{j_0\cup\mathcal{M}, j_0\cup\mathcal{M}} \boldsymbol{a}. \end{aligned}$$

By (A.55), this implies

$$\min_{\substack{\mathcal{M}\subseteq[1,\dots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\leq\kappa_n}} \sigma_{\mathcal{M},j_0}^2 \geq \bar{c}, \tag{A.56}$$

and hence

$$\min_{\substack{\mathcal{M}\subseteq[1,\dots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\leq\kappa_n}} \sigma_{\mathcal{M},j_0} \geq \sqrt{\bar{c}}.$$

It follows from (A.21) and Assumption (A3) implies that

$$\boldsymbol{\Sigma}_{j_0,j_0} = \mathrm{E}X_{0,j_0}^2 \leq \|X_{0,j_0}\|_{\psi_2}^2 \leq c_0^2.$$

In view of (A.56), this further implies that

$$\boldsymbol{\Sigma}_{j_0,\mathcal{M}} \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1} \boldsymbol{\Sigma}_{\mathcal{M},j_0} = \boldsymbol{\Sigma}_{j_0,j_0} - \sigma_{\mathcal{M},j_0}^2 \leq c_0^2.$$

Note that $\mathbf{\Sigma}_{j_0,\mathcal{M}}\mathbf{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}\mathbf{\Sigma}_{\mathcal{M},j_0} = \boldsymbol{\omega}_{\mathcal{M},j_0}^T\mathbf{\Sigma}_{\mathcal{M},\mathcal{M}}\boldsymbol{\omega}_{\mathcal{M},j_0}$. Hence, we have

$$\|\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2^2 \leq \frac{c_0^2}{\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{M},\mathcal{M}})} \leq \frac{c_0^2}{\bar{c}}, \tag{A.57}$$

where the last inequality is due to Condition (A2). Therefore, (A.1) is proven.

Similar to (A.21), we can show for any random variable Z,

$$\mathrm{E}Z^4 \leq 2\|Z\|_{\psi_2}^4. \tag{A.58}$$

It follows from Condition (A3) that

$$\begin{aligned}\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \mathrm{E}|X_{0,j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|^4 &\leq 2 \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \|X_{0,j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}\|_{\psi_2}^4 \\ &\leq 2c_0^4 \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \|(1,\boldsymbol{\omega}_{\mathcal{M},j_0}^T)^T\|_2^4 \leq 2c_0^4(1+\bar{c}^{-1}c_0^2)^2.\end{aligned} \tag{A.59}$$

Moreover, by (A.1),

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \frac{1}{\sigma_{\mathcal{M},j_0}} \leq \frac{1}{\sqrt{\bar{c}}}.$$

Thus, we have

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\leq\kappa_n}} \frac{1}{\sigma_{\mathcal{M},j_0}^4}\mathrm{E}|X_{0,j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|^4 \leq \frac{2c_0^4}{\bar{c}^2}\left(1+\frac{c_0^2}{\bar{c}}\right)^2.$$

For any random variable Z with $\|Z\|_{\psi_2} \leq \omega$, it follows from the definition of the Orlicz norm that $\|Z\|_{\psi_1} \leq \omega^2$. Under Condition (A3), this implies

$$\max_j \|X_{0,j}^2\|_{\psi_1} \leq \max_j(\|X_{0,j}\|_{\psi_2})^2 \leq c_0^2. \tag{A.60}$$

For any random variable Z, we have $\mathrm{E}|Z| \leq \|Z\|_{\psi_1}$. This together with (A.60) yields that

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ |\mathcal{M}|\leq\kappa_n}} \mathrm{E}\|X_{0,\mathcal{M}}\|_2^2 = \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ |\mathcal{M}|\leq\kappa_n}} \sum_{j\in\mathcal{M}} \mathrm{E}X_{0,j}^2 \leq \kappa_n \max_j \mathrm{E}X_{0,j}^2 \leq \kappa_n \max_j \|X_{0,j}^2\|_{\psi_1} \leq \kappa_n c_0^2. \tag{A.61}$$

Finally, notice that

$$\Pr\left(|X_{i,j}| > \sqrt{3c_0^2\max(\log p,\log n)}\right) \le \exp\{-3\max(\log p,\log n)\}\mathrm{E}\exp(|X_{i,j}|^2/c_0^2)$$
$$\le 2\exp\{-3\max(\log p,\log n)\} \le 2\min(p^{-3},n^{-3}),$$

where the first inequality follows from Markov's inequality and the second inequality follows from the definition of the Orlicz norm. Now it follows from Bonferroni's inequality that

$$\Pr\left(\max_{i,j}|X_{i,j}| > \sqrt{3c_0^2\max(\log p,\log n)}\right) \le 2pn\min(p^{-3},n^{-3}) = 2\min(np^{-2},pn^{-2}) \to 0.$$

The proof is hence completed.

A.6 Proof of Lemma A.2

We first prove (A.5). Note that

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 = \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0} - \boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\boldsymbol{\Sigma}_{\mathcal{M},j_0}\|_2$$
$$\le \underbrace{\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0} - \boldsymbol{\Sigma}_{\mathcal{M},j_0})\|_2}_{\eta_1} + \underbrace{\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|(\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}} - \boldsymbol{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}})\boldsymbol{\Sigma}_{\mathcal{M},j_0}\|_2}_{\eta_2}.$$

Hence, it suffices to show that with probability tending to 1,

$$\eta_1 = O\left(\frac{\sqrt{\kappa_n}\log p}{\sqrt{n}}\right) \quad \text{and} \quad \eta_2 = O\left(\frac{\sqrt{\kappa_n}\log p}{\sqrt{n}}\right).$$

Upper bound for η_1: Since

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0} - \boldsymbol{\Sigma}_{\mathcal{M},j_0})\|_2$$
$$\le \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}\|_2 \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0} - \boldsymbol{\Sigma}_{\mathcal{M},j_0})\|_2,$$

it suffices to show with probability tending to 1 that,

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}\|_2 = O(1), \tag{A.62}$$

and

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\boldsymbol{\Sigma}_{\mathcal{M},j_0})\|_2 = O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right). \tag{A.63}$$

Note that $\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}$ is symmetric. To prove (A.62), it is equivalent to show that the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}$ are uniformly bounded with probability tending to 1. Hence, it suffices to prove

$$\min_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \lambda_{\min}\left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}\right) > \frac{\bar{c}}{2}, \tag{A.64}$$

with probability tending to 1.

Observe that

$$\begin{aligned}
\min_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \lambda_{\min}\left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}\right) &= \inf_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \inf_{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}} \boldsymbol{a}^T\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}\boldsymbol{a} \\
\ge \min_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \min_{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}} \boldsymbol{a}^T\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}\boldsymbol{a} &- \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \max_{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}} \left|\boldsymbol{a}^T\left(\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}\right)\boldsymbol{a}\right|.
\end{aligned}$$

By Condition (A2), the first term on the second line is greater than or equal to $\bar{c}$. Since $\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}$ is symmetric, the second term can be bounded by

$$\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^p\\ \|\boldsymbol{a}\|_2=1,\|\boldsymbol{a}\|_0\le\kappa_n}} |\boldsymbol{a}^T(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}})\boldsymbol{a}|. \tag{A.65}$$

Define the stochastic process

$$\mathbb{X}(\boldsymbol{a}) = \boldsymbol{a}^T(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{a}.$$

For any $\boldsymbol{a}_1,\boldsymbol{a}_2\in\mathbb{R}^p$ with $\|\boldsymbol{a}_1\|_2,\|\boldsymbol{a}_2\|_2\le 1$, we have

$$|\mathbb{X}(\boldsymbol{a}_1)-\mathbb{X}(\boldsymbol{a}_2)| \le |(\boldsymbol{a}_1-\boldsymbol{a}_2)^T(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})(\boldsymbol{a}_1+\boldsymbol{a}_2)|,$$

since $\boldsymbol{a}_2^T(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{a}_1 = \boldsymbol{a}_1^T(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{a}_2$, by the symmetricity of the matrix $\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}$. Recall that $\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma} = \sum_{i=1}^n(\boldsymbol{X}_i\boldsymbol{X}_i^T - \mathrm{E}\boldsymbol{X}_0\boldsymbol{X}_0^T)/n$. It follows from Condition (A3) and Cauchy-Schwarz inequality that

$$
\begin{aligned}
\|(\boldsymbol{a}_1-\boldsymbol{a}_2)^T\boldsymbol{X}_0\boldsymbol{X}_0^T(\boldsymbol{a}_1+\boldsymbol{a}_2)\|_{\psi_1} &\le \sqrt{2}\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2 \sup_{\substack{\boldsymbol{a}_3,\boldsymbol{a}_4\in\mathbb{R}^p\\ \|\boldsymbol{a}_3\|_2,\|\boldsymbol{a}_4\|_2\le 1}} \|\boldsymbol{a}_3^T\boldsymbol{X}_0\boldsymbol{X}_0^T\boldsymbol{a}_4\|_{\psi_1}\\
&\le \sqrt{2}\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2 \sup_{\substack{\boldsymbol{a}_3,\boldsymbol{a}_4\in\mathbb{R}^p\\ \|\boldsymbol{a}_3\|_2,\|\boldsymbol{a}_4\|_2\le 1}} \frac{\|(\boldsymbol{a}_3^T\boldsymbol{X}_0)^2+(\boldsymbol{a}_4^T\boldsymbol{X}_0)^2\|_{\psi_1}}{2}\\
&\le \sqrt{2}\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2 \sup_{\substack{\boldsymbol{a}_3,\boldsymbol{a}_4\in\mathbb{R}^p\\ \|\boldsymbol{a}_3\|_2,\|\boldsymbol{a}_4\|_2\le 1}} \frac{\|(\boldsymbol{a}_3^T\boldsymbol{X}_0)^2\|_{\psi_1}+\|(\boldsymbol{a}_4^T\boldsymbol{X}_0)^2\|_{\psi_1}}{2}\\
&\le \sqrt{2}\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2 \sup_{\substack{\boldsymbol{a}_3,\boldsymbol{a}_4\in\mathbb{R}^p\\ \|\boldsymbol{a}_3\|_2,\|\boldsymbol{a}_4\|_2\le 1}} \frac{\|\boldsymbol{a}_3^T\boldsymbol{X}_0\|_{\psi_2}^2+\|\boldsymbol{a}_4^T\boldsymbol{X}_0\|_{\psi_2}^2}{2} \le \sqrt{2}c_0^2\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2. \quad \text{(A.66)}
\end{aligned}
$$

By Jensen's inequality, we have

$$
\begin{aligned}
\|(\boldsymbol{a}_1-\boldsymbol{a}_2)^T(\boldsymbol{X}_0\boldsymbol{X}_0^T-\mathrm{E}\boldsymbol{X}_0\boldsymbol{X}_0^T)(\boldsymbol{a}_1+\boldsymbol{a}_2)\|_{\psi_1} &\le 2\|(\boldsymbol{a}_1-\boldsymbol{a}_2)^T\boldsymbol{X}_0\boldsymbol{X}_0^T(\boldsymbol{a}_1+\boldsymbol{a}_2)\|_{\psi_1}\\
&\le 2\sqrt{2}c_0^2\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2.
\end{aligned}
$$

It follows from Bernstein's inequality (Theorem 3.1, Klartag and Mendelson, 2005) that

$$
\Pr(|\mathbb{X}(\boldsymbol{a}_1)-\mathbb{X}(\boldsymbol{a}_2)|>t)\le 2\exp\left\{-O(1)\min\left(\frac{nt^2}{\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2^2},\frac{nt}{\|\boldsymbol{a}_1-\boldsymbol{a}_2\|_2}\right)\right\},
$$

for some positive constant $O(1)$ that is independent of $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$. Let $\mathbb{S}=\{\boldsymbol{a}\in\mathbb{R}^p : \|\boldsymbol{a}\|_2\le 1, \|\boldsymbol{a}\|_0\le\kappa_n\}$. It follows from Theorem 1.2.7 of Talagrand (2005) that

$$
\mathrm{E}\sup_{\boldsymbol{a}_1,\boldsymbol{a}_2\in\mathbb{S}}|\mathbb{X}(\boldsymbol{a}_1)-\mathbb{X}(\boldsymbol{a}_2)| = O\{\gamma_2(\mathbb{S}, n^{-1/2}\|\cdot\|_2)+\gamma_1(\mathbb{S}, n^{-1}\|\cdot\|_2)\},
$$

where the definitions of the γ_p-functionals are given in Definition 1.2.5 of Talagrand (2005). Since $\mathbb{X}(\mathbf{0}_p)=0$, we have

$$
\mathrm{E}\sup_{\boldsymbol{a}\in\mathbb{S}}|\mathbb{X}(\boldsymbol{a})| = O\{\gamma_2(\mathbb{S}, n^{-1/2}\|\cdot\|_2)+\gamma_1(\mathbb{S}, n^{-1}\|\cdot\|_2)\}. \quad \text{(A.67)}
$$

By Lemma 2.3 of Mendelson et al. (2008), for any $0\le\varepsilon\le 1/2$, there exists an ε-cover of $\mathbb{S}$ with cardinality at most $(5/2\varepsilon)^{\kappa_n}\binom{p}{\kappa_n}$. Using similar arguments in proving Lemma G.8

of Shi et al. (2018), we can show that

$$\gamma_2(\mathbb{S}, n^{-1/2}\|\cdot\|_2) \le n^{-1/2}\gamma_2(\mathbb{S}, \|\cdot\|_2) = O(n^{-1/2}\sqrt{\kappa_n \log p}),$$
$$\gamma_1(\mathbb{S}, n^{-1}\|\cdot\|_2) \le n^{-1}\gamma_1(\mathbb{S}, \|\cdot\|_2) = O(n^{-1}\kappa_n \log p).$$

Under the given conditions, it follows from (A.67) that $\mathrm{E}\sup_{\boldsymbol{a}\in\mathbb{S}} |\mathbb{X}(\boldsymbol{a})| = O(n^{-1/2}\sqrt{\kappa_n \log p})$. By Markov's inequality, we obtain $\sup_{\boldsymbol{a}\in\mathbb{S}} |\mathbb{X}(\boldsymbol{a})| = O_p(n^{-1/2}\sqrt{\kappa_n \log p}) = o_p(1)$ and hence (A.65) is $o_p(1)$. Under Condition (A2), we have

$$\min_{\substack{\mathcal{M}\subseteq[1,\dots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\le\kappa_n}} \min_{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}} \boldsymbol{a}^T \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}} \boldsymbol{a} \ge \bar{c}.$$

Assertion (A.62) thus follows.

Recall that $\widehat{\boldsymbol{\Sigma}}_{j_1,j_2} - \boldsymbol{\Sigma}_{j_1,j_2} = \sum_i (X_{i,j_1}X_{i,j_2} - \mathrm{E}X_{0,j_1}X_{0,j_2})/n$. Combining (A.60) with Cauchy-Schwarz inequality, we have

$$\|X_{0,j_1}X_{0,j_2}\|_{\psi_1} \le \frac{\|X_{0,j_1}^2 + X_{0,j_2}^2\|_{\psi_1}}{2} \le \frac{\|X_{0,j_1}^2\|_{\psi_1}}{2} + \frac{\|X_{0,j_2}^2\|_{\psi_1}}{2} \le c_0^2, \tag{A.68}$$

for all $j_1, j_2 \in [1, \dots, p]$. By Jensen's inequality, we have

$$\mathrm{E}\exp\left(\frac{\mathrm{E}|X_{0,j_1}X_{0,j_2|}}{\omega_0^2}\right) \le \mathrm{E}\exp\left(\frac{|X_{0,j_1}X_{0,j_2|}}{\omega_0^2}\right) \le 2.$$

This implies $\|\mathrm{E}X_{0,j_0}X_{0,j_2}\|_{\psi_1} \le c_0^2$, $\forall j_1, j_2$. Combining this together with (A.68) gives

$$\|X_{0,j_1}X_{0,j_2} - \mathrm{E}X_{0,j_0}X_{0,j_2}\|_{\psi_1} \le \|X_{0,j_1}X_{0,j_2}\|_{\psi_1} + \|\mathrm{E}X_{0,j_1}X_{0,j_2}\|_{\psi_1} \le 2c_0^2.$$

Therefore, it follows from Bernstein's inequality that

$$\max_{1\le j_1,j_2\le p} \Pr\left(\left|\sum_i (X_{i,j_1}X_{i,j_2} - \boldsymbol{\Sigma}_{j_1,j_2})\right| \ge t\right) \le 2\exp\left(-O(1)\min\left(\frac{t^2}{4nc_0^2}, \frac{t}{2c_0}\right)\right), \tag{A.69}$$

for any $t > 0$, where $O(1)$ denotes some positive constant.

Take $t_0 = 3\sqrt{n\log p}c_0/\sqrt{c_1}$. Since $\log p = o(n)$, we have for sufficiently large n,

$$\frac{t_0^2}{4nc_0^2} = \frac{9\log p}{4c_1} \ll \frac{3\sqrt{n\log p}}{2\sqrt{c_1}} = \frac{t_0}{2c_0}.$$

It follows from (A.69) that

$$\max_{j_1,j_2} \Pr\left(\left|\sum_i (X_{i,j_1}X_{i,j_2} - \mathbf{\Sigma}_{j_1,j_2})\right| \geq t_0\right) \leq 2\exp\left(-\frac{c_1 t_0^2}{4nc_0^2}\right) \leq 2\exp\left(-\frac{9\log p}{4}\right).$$

By Bonferroni's inequality, we have

$$\begin{aligned} & \Pr\left(\max_{j_1,j_2\in[1,\ldots,p]} \left|\sum_i (X_{i,j_1}X_{i,j_2} - \mathbf{\Sigma}_{j_1,j_2})\right| \geq t_0\right) \qquad \text{(A.70)} \\ \leq & \sum_{j_1,j_2\in[1,\ldots,p]} \Pr\left(\left|\sum_i (X_{i,j_1}X_{i,j_2} - \mathbf{\Sigma}_{j_1,j_2})\right| \geq t_0\right) \\ \leq & \; p^2 2\exp\left(-\frac{9\log p}{4}\right) = 2\exp\left(-\frac{9\log p}{4} + 2\log p\right) = 2\exp\left(-\frac{\log p}{4}\right) \to 0. \end{aligned}$$

Under the event defined in (A.70), we have

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\leq\kappa_n}} \|(\widehat{\mathbf{\Sigma}}_{\mathcal{M},j_0} - \mathbf{\Sigma}_{\mathcal{M},j_0})\|_2 \leq \sqrt{\kappa_n} \max_{j_1,j_2\in[1,\ldots,p]} \left|\widehat{\mathbf{\Sigma}}_{j_1,j_2} - \mathbf{\Sigma}_{j_1,j_2}\right| \leq \frac{\sqrt{\kappa_n} t_0}{n}.$$

This proves (A.63).

Upper bound for η_2: Observe that

$$\begin{aligned} & \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\leq\kappa_n}} \|(\widehat{\mathbf{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}} - \mathbf{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}})\mathbf{\Sigma}_{\mathcal{M},j_0}\|_2 \\ = & \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\leq\kappa_n}} \|\widehat{\mathbf{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}(\widehat{\mathbf{\Sigma}}_{\mathcal{M},\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M},\mathcal{M}})\mathbf{\Sigma}^{-1}_{\mathcal{M},\mathcal{M}}\mathbf{\Sigma}_{\mathcal{M},j_0}\|_2 \\ \leq & \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\leq\kappa_n}} \|\widehat{\mathbf{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}\|_2 \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\leq\kappa_n}} \|(\widehat{\mathbf{\Sigma}}_{\mathcal{M},\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2. \end{aligned}$$

By (A.62), it suffices to show

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p] \\ j_0\notin\mathcal{M}, |\mathcal{M}|\leq\kappa_n}} \|(\widehat{\mathbf{\Sigma}}_{\mathcal{M},\mathcal{M}} - \mathbf{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 = O\left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}}\right), \qquad \text{(A.71)}$$

with probability tending to 1.

LHS of (A.71) can be upper bounded by

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} |\boldsymbol{a}^T(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}|.$$

For any subset $\mathcal{M}$ such that $j_0\notin\mathcal{M}$, $|\mathcal{M}|\le\kappa_n$, define the stochastic process

$$T_{\mathcal{M}}(\boldsymbol{a})=\frac{1}{n}\sum_{i=1}^n g_{\mathcal{M}}(\boldsymbol{X}_i,\boldsymbol{a})=\frac{1}{n}\sum_{i=1}^n \boldsymbol{a}^T\left(\boldsymbol{X}_{i,\mathcal{M}}\boldsymbol{X}_{i,\mathcal{M}}^T-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}\right)\boldsymbol{\omega}_{\mathcal{M},j_0}.$$

Using similar arguments in bounding (A.65), we can show

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \mathrm{E}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} |T_{\mathcal{M}}(\boldsymbol{a})|=O(n^{-1/2}\sqrt{\kappa_n}). \tag{A.72}$$

The envelope function of |g| is bounded by

$$G_{\mathcal{M}}(\boldsymbol{X}_i)\stackrel{\Delta}{=}\|\boldsymbol{X}_{i,\mathcal{M}}\|_2|\boldsymbol{X}_{i,\mathcal{M}}^T\boldsymbol{\omega}_{\mathcal{M},j_0}|_2+\|\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}\|_2\|\boldsymbol{\omega}_{\mathcal{M},j_0}\|.$$

Combing (A.22) together with (A.1), we have

$$G_{\mathcal{M}}(\boldsymbol{X}_i)\le\|\boldsymbol{X}_{i,\mathcal{M}}\|_2|\boldsymbol{X}_{i,\mathcal{M}}^T\boldsymbol{\omega}_{\mathcal{M},j_0}|_2+c_0^3/\sqrt{\bar{c}}.$$

The $\|\cdot\|_{\psi_1}$ Orlicz norm of G can be upper bounded by

$$\begin{aligned}\|G_{\mathcal{M}}(\boldsymbol{X}_i)\|_{\psi_1}&\le\|c_0^3/\sqrt{\bar{c}}\|_{\psi_1}+\|\|\boldsymbol{X}_{i,\mathcal{M}}\|_2|\boldsymbol{X}_{i,\mathcal{M}}^T\boldsymbol{\omega}_{\mathcal{M},j_0}|\|_{\psi_1}\\ &\le c_0^3/\sqrt{\bar{c}}+\|\|\boldsymbol{X}_{i,\mathcal{M}}\|_2|\boldsymbol{X}_{i,\mathcal{M}}^T\boldsymbol{\omega}_{\mathcal{M},j_0}|\|_{\psi_1}.\end{aligned} \tag{A.73}$$

Notice that

$$\begin{aligned}\|\|\boldsymbol{X}_{i,\mathcal{M}}\|_2|\boldsymbol{X}_{i,\mathcal{M}}^T\boldsymbol{\omega}_{\mathcal{M},j_0}|_2\|_{\psi_1}&\le\left\|\frac{\|\boldsymbol{X}_{i,\mathcal{M}}\|_2^2}{2\sqrt{\kappa_n}}+\frac{\sqrt{\kappa_n}|\boldsymbol{X}_{i,\mathcal{M}}^T\boldsymbol{\omega}_{\mathcal{M},j_0}|^2}{2}\right\|_{\psi_1}\\ &\le\frac{\|\|\boldsymbol{X}_{i,\mathcal{M}}\|_2^2\|_{\psi_1}}{2\sqrt{\kappa_n}}+\frac{\sqrt{\kappa_n}\|(\boldsymbol{X}_{i,\mathcal{M}}^T\boldsymbol{\omega}_{\mathcal{M},j_0})^2\|_{\psi_1}}{2}\\ &\le\frac{\sum_{j\in\mathcal{M}}\|X_{i,j}^2\|_{\psi_1}}{2\sqrt{\kappa_n}}+\frac{\sqrt{\kappa_n}\|\boldsymbol{X}_{i,\mathcal{M}}^T\boldsymbol{\omega}_{\mathcal{M},j_0}\|_{\psi_2}^2}{2}=O(\sqrt{\kappa_n}),\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality, the equality follows from

(A.60), (A.1) and Condition (A3). This together with (A.73) yields that

$$\max_{i\in[1,\ldots,n]} \|G_{\mathcal{M}}(\boldsymbol{X}_i)\|_{\psi_1} = O(\sqrt{\kappa_n}).$$

Hence, it follows from Lemma 2.2.2 in van der Vaart and Wellner (1996) that

$$\left\|\max_{i\in[1,\ldots,n]} |G_{\mathcal{M}}(\boldsymbol{X}_i)|\right\|_{\psi_1} \le K_1 \log(1+n) \max_{i\in[1,\ldots,n]} \|G_{\mathcal{M}}(\boldsymbol{X}_i)\|_{\psi_1} = O(\sqrt{\kappa_n}\log n), \tag{A.74}$$

for some constant K_1 that is independent of $\mathcal{M}$.

Moreover, it follows from Cauchy-Schwarz inequality that

$$\begin{aligned}
&\sigma_*^2 \equiv \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} \mathrm{E} g_{\mathcal{M}}(\boldsymbol{X}_0,\boldsymbol{a})^2 \\
\le\ & \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} \mathrm{E}|\boldsymbol{a}^T(\boldsymbol{X}_{0,\mathcal{M}}\boldsymbol{X}_{0,\mathcal{M}}^T - \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}|^2 \\
\le\ & \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} \mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}\boldsymbol{X}_{0,\mathcal{M}}^T\boldsymbol{\omega}_{\mathcal{M},j_0}|^2 \\
\le\ & \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} \sqrt{\mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^4} \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sqrt{\mathrm{E}|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|^4}.
\end{aligned}$$

Using similar arguments in proving (A.59), we can show

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} \sqrt{\mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^4} \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sqrt{\mathrm{E}|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|^4} = O(1),$$

and hence $\sigma_*^2 = O(1)$.

Therefore, it follows from Theorem 4 in Adamczak (2008) that there exists some constant $K_2, K_3 > 0$ such that

$$\begin{aligned}
& \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \Pr\left(\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} |T_{\mathcal{M}}(\boldsymbol{a})| - \frac{3}{2}\mathrm{E}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} |T_{\mathcal{M}}(\boldsymbol{a})| \ge \frac{t}{n}\right) \\
\le\ & \exp\left(-\frac{t^2}{3n\sigma_*^2}\right) + 3\exp\left(-\frac{t}{K_2\sqrt{\kappa_n}\log n}\right) \\
\le\ & \exp\left(-\frac{t^2}{3K_3 n}\right) + 3\exp\left(-\frac{t}{K_2\sqrt{\kappa_n}\log n}\right), \qquad \forall t > 0.
\end{aligned}$$

Define

$$t_0 = \max\left(2\sqrt{K_3 n\kappa_n \log p}, \frac{4}{3}K_2\kappa_n^{3/2}\log p \log n\right),$$

we have

$$\begin{aligned}
&\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \Pr\left(\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}}|T_{\mathcal{M}}(\boldsymbol{a})| - \frac{3}{2}\mathrm{E}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}}|T_{\mathcal{M}}(\boldsymbol{a})| \ge \frac{t_0}{n}\right)\\
\le\ & \exp\left(-\frac{4K_3 n\kappa_n \log p}{3nK_3}\right) + 3\exp\left(-\frac{4K_2\kappa_n^2\log p\log(n+1)}{3K_2\kappa_n\log(1+n)}\right) \le 4\exp\left(-\frac{4}{3}\kappa_n\log p\right).
\end{aligned}$$

The number of subset $\mathcal{M}$ with less than or equal to κ_n elements is upper bounded by $C_p^{\kappa_n} \le p^{\kappa_n}$. Hence, it follows from Bonferroni's inequality that

$$\begin{aligned}
&\Pr\left(\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}}|T_{\mathcal{M}}(\boldsymbol{a})| - \frac{3}{2}\mathrm{E}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}}|T_{\mathcal{M}}(\boldsymbol{a})| \ge \frac{t_0}{n}\right)\\
\le\ & p^{\kappa_n}\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \Pr\left(\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}}|T_{\mathcal{M}}(\boldsymbol{a})| - \frac{3}{2}\mathrm{E}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}}|T_{\mathcal{M}}(\boldsymbol{a})| \ge \frac{t_0}{n}\right)\\
\le\ & 4p^{\kappa_n}\exp\left(-\frac{4}{3}\kappa_n\log p\right) = 4\exp\left(-\frac{4}{3}\kappa_n\log p + \kappa_n\log p\right) = 4\exp\left(-\frac{1}{3}\kappa_n\log p\right) \to 0.
\end{aligned}$$

This together with (A.72) implies that

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}}|T_{\mathcal{M}}(\boldsymbol{a})| \le O(1)n^{-1/2}\sqrt{\kappa_n} + \frac{t_0}{n}, \tag{A.75}$$

with probability tending to 1, where $O(1)$ denotes some positive constant.

Under the given conditions, we have

$$\frac{4}{3}K_2\kappa_n^{3/2}\log p\log n = O(\sqrt{n\kappa_n\log p}),$$

and hence $t_0 = O(\sqrt{n\kappa_n\log p})$. Under the event defined in (A.75), we have for sufficiently large n,

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\ \|\boldsymbol{a}\|_2=1}} |T_{\mathcal{M}}(\boldsymbol{a})| = O\left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}}\right).$$

This proves (A.71). The upper bound for η_2 is thus given.

Consider (A.6). Assume for now, we've shown

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} |\hat{\sigma}^2_{\mathcal{M},j_0} - \sigma^2_{\mathcal{M},j_0}| = O\left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}}\right), \tag{A.76}$$

with probability tending to 1. Then, under the event defined in (A.76), we have

$$\begin{aligned} \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} |\hat{\sigma}_{\mathcal{M},j_0} - \sigma_{\mathcal{M},j_0}| &= \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \frac{|\hat{\sigma}^2_{\mathcal{M},j_0} - \sigma^2_{\mathcal{M},j_0}|}{|\hat{\sigma}_{\mathcal{M},j_0} + \sigma_{\mathcal{M},j_0}|} \\ \le \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \frac{|\hat{\sigma}^2_{\mathcal{M},j_0} - \sigma^2_{\mathcal{M},j_0}|}{|\sigma_{\mathcal{M},j_0}|} &\le \frac{1}{\sqrt{\bar{c}}} \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} |\hat{\sigma}^2_{\mathcal{M},j_0} - \sigma^2_{\mathcal{M},j_0}| = O\left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}}\right), \end{aligned}$$

where the last inequality follows from (A.1) and the last equality is due to (A.76). Hence, it suffices to show (A.76).

By definition, we have

$$\begin{aligned} & |\hat{\sigma}^2_{\mathcal{M},j_0} - \sigma^2_{\mathcal{M},j_0}| \le |\widehat{\boldsymbol{\Sigma}}_{j_0,j_0} - \boldsymbol{\Sigma}_{j_0,j_0}| + |\widehat{\boldsymbol{\Sigma}}^T_{\mathcal{M},j_0}\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \boldsymbol{\Sigma}^T_{\mathcal{M},j_0}\boldsymbol{\omega}_{\mathcal{M},j_0}| \qquad \text{(A.77)} \\ \le\ & |\widehat{\boldsymbol{\Sigma}}_{j_0,j_0} - \boldsymbol{\Sigma}_{j_0,j_0}| + |\boldsymbol{\Sigma}^T_{\mathcal{M},j_0}(\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0})| + |(\boldsymbol{\Sigma}_{\mathcal{M},j_0} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0})^T\boldsymbol{\omega}_{\mathcal{M},j_0}| \\ +\ & |(\boldsymbol{\Sigma}_{\mathcal{M},j_0} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0})^T(\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0})| \le |\widehat{\boldsymbol{\Sigma}}_{j_0,j_0} - \boldsymbol{\Sigma}_{j_0,j_0}| \\ +\ & \|\boldsymbol{\Sigma}_{\mathcal{M},j_0}\|_2\|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 + \|\boldsymbol{\Sigma}_{\mathcal{M},j_0} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}\|_2\|\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 \\ +\ & \|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}\|_2\|\boldsymbol{\Sigma}_{\mathcal{M},j_0} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}\|_2. \end{aligned}$$

It follows from (A.1), (A.5), (A.63) and (A.70) that with probability tending to 1,

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\boldsymbol{\Sigma}_{\mathcal{M},j_0}-\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}\|_2 = O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right), \tag{A.78}$$

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\boldsymbol{\omega}_{\mathcal{M},j_0}-\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0}\|_2 = O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right), \tag{A.79}$$

$$|\widehat{\boldsymbol{\Sigma}}_{j_0,j_0}-\boldsymbol{\Sigma}_{j_0,j_0}| = O\left(\frac{\sqrt{\log p}}{\sqrt{n}}\right), \quad \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 = O(1). \tag{A.80}$$

By Condition (A2), $\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}$ is invertible for any subset $\mathcal{M}$ such that $|\mathcal{M}|\le\kappa_n$. Hence, it follows from (A.22) that

$$\min_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}) \ge c_0^{-1/2}.$$

Using similar arguments in proving (A.1), we can show that

$$\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \|\boldsymbol{\Sigma}_{\mathcal{M},j_0}\|_2^2 \le \max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \frac{c_0^2}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1})} \le c_0^{5/2}.$$

Under the given conditions, we have $\kappa_n\log p = o(n)$. Under the events defined in (A.77)-(A.79), we obtain that

$$\begin{aligned}
&\max_{\substack{\mathcal{M}\subseteq[1,\ldots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} |\hat{\sigma}^2_{\mathcal{M},j_0}-\sigma^2_{\mathcal{M},j_0}| \le O\left(\frac{\sqrt{\log p}}{\sqrt{n}}\right) + c_0^{3/2}O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right)\\
+\ & O(1)O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right) + O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right)O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right) = O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right).
\end{aligned}$$

This proves (A.76).

We now focus on proving (A.7). By definition, we have

$$\begin{aligned}
&\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0}-\boldsymbol{\omega}_{\mathcal{M},j_0} = \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\boldsymbol{\Sigma}_{\mathcal{M},j_0}) + (\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1})\boldsymbol{\Sigma}_{\mathcal{M},j_0}\\
=\ & \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}\{\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\boldsymbol{\Sigma}_{\mathcal{M},j_0}-(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}\},
\end{aligned}$$

for any $\mathcal{M}$ and hence the LHS of (A.7) can be upper bounded by

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \left\|(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1})\{\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\boldsymbol{\Sigma}_{\mathcal{M},j_0}-(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}\}\right\|_2. \quad (A.81)$$

It suffices to provide an upper bound for (A.81). Using similar arguments in bounding η_1 and η_2, we can show the following event occurs with probability tending to 1,

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\boldsymbol{\Sigma}_{\mathcal{M},j_0}-(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}\right\|_2 = O\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right). \quad (A.82)$$

Notice that

$$\begin{aligned}
&\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}\right\|_2\\
\le\ &\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}\right\|_2 \max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}\right\|_2 \max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \left\|\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}\right\|_2.
\end{aligned}$$

To bound η_2, we have shown that

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}\right\|_2 = O_p\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right).$$

By (A.64) and Condition (A2), we obtain

$$\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1}\right\|_2 = O_p\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right). \quad (A.83)$$

Combining this together with (A.82) and Cauchy-Schwarz inequality yields that

$$\begin{aligned}
\max_{\substack{\mathcal{M}\subseteq[1,\dots,p]\\ j_0\notin\mathcal{M},|\mathcal{M}|\le\kappa_n}} &\left\|(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}^{-1}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}^{-1})\{\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\boldsymbol{\Sigma}_{\mathcal{M},j_0}-(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}\}\right\|_2\\
&= O_p\left(\frac{\kappa_n\log p}{n}\right).
\end{aligned}$$

This completes the proof.

A.7 Proof of Lemma A.3

Let

$$I_{2,j} = \frac{1}{\sqrt{n}} \sum_{t=s_n}^{n-1} \frac{1}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) X_{t+1,j} \mathcal{I}(j \in \widehat{\mathcal{M}}_{j_0}^{(t)}).$$

Similar to (A.66), we can show

$$\left\| \frac{1}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) X_{t+1,j} \mathcal{I}(j \in \widehat{\mathcal{M}}_{j_0}^{(t)}) \right\|_{\psi_1|\mathcal{F}_t} \le \frac{c_0^2}{\sqrt{\bar{c}}} \left(1 + \frac{c_0^2}{\bar{c}} \right),$$

almost surely, under Condition (A1). For any random variable Z, it follows from the definition of the Orlicz norm that $1+\mathrm{E}|Z|^k/\|Z\|_{\psi_1}^k \le \mathrm{E}\exp(|Z|/\|Z\|_{\psi_1}) = 2$ for any integer $k > 0$ and hence $\mathrm{E}|Z|^k \le k!\|Z\|_{\psi_1}^k$. As a result, we have

$$\mathrm{E}\left\{ \left| \frac{1}{\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right) X_{t+1,j} \mathcal{I}(j \in \widehat{\mathcal{M}}_{j_0}^{(t)}) \right|^k \middle| \mathcal{F}_t \right\} \le k! \frac{c_0^{2k}}{\bar{c}^{k/2}} \left(1 + \frac{c_0^2}{\bar{c}} \right)^k, \tag{A.84}$$

almost surely, for any $k \ge 1$.

Let $c_0^* = \bar{c}^{-1/2} c_0^2 (1 + \bar{c} c_0^2)$. It follows from Theorem 9.12 in de la Peña et al. (2009) that

$$\Pr\left(|I_{2,j}^{**}| > z, \sum_{t=s_n}^{n-1} \mathrm{E} \frac{1}{\sigma^2_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left\{ \left(X_{t+1,j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^T \boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}} \right)^2 X_{t+1,j}^2 | \mathcal{F}_t \right\} \le 2n(c_0^*)^2 \right) \le 2\exp\left(-\frac{z^2}{2(2(c_0^*)^2 + c_0^* z/\sqrt{n})} \right), \quad \forall z > 0.$$

In view of (A.84), we have

$$\Pr\left(|I_{2,j}^{**}| > z \right) \le 2\exp\left(-\frac{z^2}{2(2(c_0^*)^2 + c_0^* z/\sqrt{n})} \right).$$

Let $z_0 = 3c_0^* \sqrt{\log p}$, we have by the condition $\log p = o(n)$ that

$$\Pr\left(|I_{2,j}^{**}| > z_0 \right) \le 2\exp\left(-\frac{9\log p}{4 + 6n^{-1/2}\sqrt{\log p}} \right) \le 2\exp\left(-\frac{3}{2}\log p \right),$$

for sufficiently large n. It follows from Bonferroni's inequality that

$$\Pr\left(\max_j |I_{2,j}^{**}| > z_0\right) \le \sum_j \Pr\left(|I_{2,j}^{**}| > z_0\right) \le 2p\exp\left(-\frac{3}{2}\log p\right) = 2\exp\left(-\frac{1}{2}\log p\right) \to 0.$$

This completes the first part of the proof.

For $t \in [s_n, \dots, n-1]$ and $j \in [1, \dots, p]$, let $W_{t+1,j} = X_{t+1,j}\mathcal{I}(j \in \widehat{\mathcal{M}}^{(t)})$ and $\boldsymbol{W}_{t+1} = (W_{t+1,1}, \cdots, W_{t+1,p})^T$. It follows that

$$\begin{aligned}\frac{1}{n}\sum_{t=s_n}^{n-1} \|\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}^{(t)}}(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}^{(t)}})\|_2^2 &= \frac{1}{n}\sum_{t=s_n}^{n-1}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \boldsymbol{W}_{t+1}\boldsymbol{W}_{t+1}^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &\le \underbrace{\frac{1}{n}\sum_{t=s_n}^{n-1}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathrm{E}(\boldsymbol{W}_{t+1}\boldsymbol{W}_{t+1}^T|\mathcal{F}_t)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}_{\eta_3} \\ &+ \underbrace{\left|\frac{1}{n}\sum_{t=s_n}^{n-1}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\{\boldsymbol{W}_{t+1}\boldsymbol{W}_{t+1}^T - \mathrm{E}(\boldsymbol{W}_{t+1}\boldsymbol{W}_{t+1}^T|\mathcal{F}_t)\}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right|}_{\eta_4}.\end{aligned}$$

Notice that

$$\begin{aligned}\eta_3 &= \frac{1}{n}\sum_{t=s_n}^{n-1}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathrm{E}(\boldsymbol{W}_{t+1}\boldsymbol{W}_{t+1}^T|\mathcal{F}_t)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &= \frac{1}{n}\sum_{t=s_n}^{n-1}(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}^{(t)}})^T \boldsymbol{\Sigma}_{\widehat{\mathcal{M}}^{(t)},\widehat{\mathcal{M}}^{(t)}}(\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}^{(t)}}) \le \lambda_{\max}(\boldsymbol{\Sigma}_{\widehat{\mathcal{M}}^{(t)},\widehat{\mathcal{M}}^{(t)}})\|\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}^{(t)}}\|_2^2.\end{aligned}$$

It follows from Condition (A1), (A4) and (A.22) that $\eta_3 = O(\eta_n^2)$, with probability tending to 1. As for η_4, we have

$$\eta_4 \le \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1^2 \max_{j_1,j_2}\left|\frac{1}{n}\sum_{t=s_n}^{n-1}\{W_{t+1,j_1}W_{t+1,j_2} - \mathrm{E}(W_{t+1,j_1}W_{t+1,j_2}|\mathcal{F}_t)\}\right|. \tag{A.85}$$

For any j_1, j_2, $\sum_{t=s_n}^{n-1}\{W_{t+1,j_1}W_{t+1,j_2} - \mathrm{E}(W_{t+1,j_1}W_{t+1,j_2}|\mathcal{F}_t)$ forms a mean zero martingale with respect to the filtration $\{\sigma(\mathcal{F}_t)\}$. Using similar arguments in bounding $\max_j |\mathcal{I}_{2,j}^{**}|$, we can show the following holds with probability tending to 1,

$$\max_{j_1,j_2}\left|\frac{1}{n}\sum_{t=s_n}^{n-1}\{W_{t+1,j_1}W_{t+1,j_2} - \mathrm{E}(W_{t+1,j_1}W_{t+1,j_2}|\mathcal{F}_t)\}\right| = O(n^{-1/2}\sqrt{\log p}). \tag{A.86}$$

Combining (A.25) with (A.85), (A.86) and the condition $\kappa_n^2 \log p = O(n/\log^2 n)$ yields that

$$\eta_4 = O\left(\eta_n^2 n^{-1/2} \kappa_n \sqrt{\log p}\right) = O(\eta_n^2),$$

with probability tending to 1. (A.9) is hence proven.

A.8 Proof of Lemma A.4

Assertion (A.28) can be proven in a similar manner as (A.1). We omit its proof for brevity. To prove (A.29) and (A.30), we first show the following events occur with probability tending to 1,

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}} - \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}\right\|_2 \le \frac{\bar{c}_*\sqrt{\kappa_n \log p}}{\sqrt{n}} + \bar{c}_*\eta_n, \tag{A.87}$$

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0} - \boldsymbol{\Sigma}_{\mathcal{M},j_0}\right\|_2 \le \frac{\bar{c}_*\sqrt{\kappa_n \log p}}{\sqrt{n}} + \bar{c}_*\eta_n, \tag{A.88}$$

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \left\|\left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}} - \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}\right)\boldsymbol{\omega}_{\mathcal{M},j_0}\right\|_2 \le \frac{\bar{c}_*\sqrt{\kappa_n \log p}}{\sqrt{n}} + \bar{c}_*\eta_n. \tag{A.89}$$

Using similar arguments in the proof of Lemma A.2, we can show that there exists some constant $\bar{c}_{**} > 0$ such that the following events occur with probability tending to 1,

$$\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n} \left\|\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}} - \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}\right\|_2 \le \frac{\bar{c}_{**}\sqrt{\kappa_n \log p}}{\sqrt{n}},$$
$$\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n} \left\|\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0} - \boldsymbol{\Sigma}_{\mathcal{M},j_0}\right\|_2 \le \frac{\bar{c}_{**}\sqrt{\kappa_n \log p}}{\sqrt{n}},$$
$$\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n} \left\|\left(\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}} - \boldsymbol{\Sigma}_{\mathcal{M},\mathcal{M}}\right)\boldsymbol{\omega}_{\mathcal{M},j_0}\right\|_2 \le \frac{\bar{c}_{**}\sqrt{\kappa_n \log p}}{\sqrt{n}}.$$

Therefore, it suffices to show the following events occur with probability tending to 1,

$$\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n} \left\|\left(\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}\right)\boldsymbol{\omega}_{\mathcal{M},j_0}\right\|_2 \le \bar{c}_{***}\eta_n, \tag{A.90}$$

$$\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n} \left\|\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}\right\|_2 \le \bar{c}_{***}\eta_n, \tag{A.91}$$

$$\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n} \left\|\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}\right\|_2 \le \bar{c}_{***}\eta_n, \tag{A.92}$$

for some constant $\bar{c}_{***} > 0$.

Similar to (A.39), we can show that with probability tending to 1 that

$$|b''(\boldsymbol{X}_i^T\widetilde{\boldsymbol{\beta}}) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)| \le c_*|\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)|,$$

where the constant c_* is defined in (A.38). With some calculations, we have

$$\begin{aligned}
&\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \left\|\left(\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}\right)\boldsymbol{\omega}_{\mathcal{M},j_0}\right\|_2 \qquad \text{(A.93)}\\
\le\ & \max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2\le1}} \left|\boldsymbol{a}^T\left(\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}} - \widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}\right)\boldsymbol{\omega}_{\mathcal{M},j_0}\right|\\
\le\ & \max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2\le1}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}||\boldsymbol{\omega}^T_{\mathcal{M},j_0}\boldsymbol{X}_{i,\mathcal{M}}||b''(\boldsymbol{X}_i^T\widetilde{\boldsymbol{\beta}}) - b''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)|\\
\le\ & c_*\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2\le1}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}||\boldsymbol{\omega}^T_{\mathcal{M},j_0}\boldsymbol{X}_{i,\mathcal{M}}||\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)| \le c_*\sqrt{\eta_5\eta_6},
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality and

$$\eta_5 = \max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2\le1}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}|^2|\boldsymbol{\omega}^T_{\mathcal{M},j_0}\boldsymbol{X}_{i,\mathcal{M}}|,$$

$$\eta_6 = \max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \frac{1}{n}\sum_{i=1}^n |\boldsymbol{\omega}^T_{\mathcal{M},j_0}\boldsymbol{X}_{i,\mathcal{M}}||\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)|^2.$$

Consider η_5. By (A.28), (A3*) and (A4*), we have for any $\mathcal{M}$ and $\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}$ that

$$\||\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^2|\boldsymbol{X}^T_{0,\mathcal{M}}\boldsymbol{\omega}_{\mathcal{M},j_0}|\|_{\psi_1} \le \bar{c}^{-1/2}c_0\sqrt{\kappa_n}\omega_0\||\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^2\|_{\psi_1} \le \bar{c}^{-1/2}c_0^3\sqrt{\kappa_n}\omega_0.$$

Using similar arguments in bounding (A.65), we can show for any $\mathcal{M}$ with $|\mathcal{M}|\le\kappa_n$, we have that

$$\mathrm{E}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2\le1}}\left(\frac{1}{n}\sum_{i=1}^n|\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}|^2|\boldsymbol{\omega}^T_{\mathcal{M},j_0}\boldsymbol{X}_{i,\mathcal{M}}| - \mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^2|\boldsymbol{\omega}^T_{\mathcal{M},j_0}\boldsymbol{X}_{0,\mathcal{M}}|\right) = o(1),$$

under the given conditions on κ_n. Hence, using similar arguments in proving (A.75), we can show

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2\le 1}}\left(\frac{1}{n}\sum_{i=1}^n|\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}|^2|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{i,\mathcal{M}}|-\mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^2|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|\right) \tag{A.94}$$

$$= O_p\left(\sqrt{\frac{\kappa_n\log p}{n}}+\frac{\kappa_n^2\log p\log n}{n}\right),$$

which is $o_p(1)$ under the condition that $\kappa_n^{5/2}\log p = O(n/\log^2 n)$. In addition, similar to (A.59), we can show

$$\sqrt{\mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^4}(\mathrm{E}|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|^4)^{1/4}=O(1),\quad\forall\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\text{ with }\|\boldsymbol{a}\|_2=1,$$

by (A.4) and Condition (A3*). It follows from Cauchy-Schwarz inequality that

$$\mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^2|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|\le\sqrt{\mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^4}\sqrt{\mathrm{E}|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|^2}$$
$$\le\quad\sqrt{\mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,\mathcal{M}}|^4}(\mathrm{E}|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{0,\mathcal{M}}|^4)^{1/4}=O(1),\quad\forall\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\text{ with }\|\boldsymbol{a}\|_2=1.$$

This together with (A.94) yields that

$$\eta_5 = O(1), \tag{A.95}$$

with probability tending to 1.

Recall that s^* is the number of nonzero elements in $\boldsymbol{\beta}_0$. Under Condition (A5*), it follows from Lemma G.9 of Shi et al. (2018) that

$$\eta_6\le(k_0+2)^2\max_{|\mathcal{M}|\le s^*}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s^*}\\\|\boldsymbol{a}\|_2\le 1}}\frac{1}{n}\sum_{i=1}^n|\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}|^2|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{i,\mathcal{M}}|\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2,$$

with probability tending to 1. Condition (A1*) implies that $s^*\le\kappa_n$. It follows that

$$\eta_6\le(k_0+2)^2\max_{|\mathcal{M}|\le \kappa_n}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{\kappa_n}\\\|\boldsymbol{a}\|_2\le 1}}\frac{1}{n}\sum_{i=1}^n|\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}|^2|\boldsymbol{\omega}_{\mathcal{M},j_0}^T\boldsymbol{X}_{i,\mathcal{M}}|\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2,$$

with probability tending to 1. Similar to (A.95), we can show

$$\eta_6 = O(\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2),$$

with probability tending to 1. Under (A5*), we obtain

$$\eta_6 = O(\eta_n^2), \tag{A.96}$$

with probability tending to 1. This together with (A.93) and (A.95) proves (A.90). Similarly, we can show (A.91) and (A.92) hold. We omit the technical details to save space.

This further implies (A.87)-(A.89) hold. Based on these results, following the arguments in the proof of Lemma A.2, we can show (A.29) and (A.30) hold. Besides, based on (A.90)-(A.92), we can similarly show (A.32) holds.

Now, we focus on proving (A.31). Similar to (A.83), we can show

$$\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n}\left\|\widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}\right\|_2 = O\left(\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n}\left\|\widehat{\boldsymbol{\Sigma}}^{*}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}\right\|_2\right),$$

with probability tending to 1. In view of (A.91), we obtain

$$\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n}\left\|\widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}\right\|_2 \le \bar{c}_0^*\eta_n, \tag{A.97}$$

for some constant $\bar{c}_0^* > 0$, with probability tending to 1.

For any $\mathcal{M}\subseteq\mathbb{I}_{j_0}$, we have

$$\begin{aligned}
\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0}-\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0} &= \underbrace{\widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0})}_{I_1^*}+(\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}})\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0}\\
+&\underbrace{(\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}})(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0})}_{I_2^*} = I_1^*+\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}})\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0}\\
+&I_2^* = I_1^*+I_2^*+\underbrace{(\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}})(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}})\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0}}_{I_3^*}+\underbrace{\widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}})\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0}}_{I_4^*}
\end{aligned}$$

By (A.92) and (A.97), it is immediate to see that $|I_2^*|$ is upper bounded by $\bar{c}_0^*\bar{c}_{***}\eta_n^2$, with probability tending to 1.

Similar to (A.5), we can show

$$\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n}\|\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0}-\boldsymbol{\omega}_{\mathcal{M},j_0}\|_2 = O_p\left(\frac{\sqrt{\kappa_n\log p}}{\sqrt{n}}\right) = o_p(1). \tag{A.98}$$

By (A.28), this further implies that

$$\Pr\left(\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n}\|\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0}\|_2 \le 2(\bar{c})^{-1/2}c_0\right)\to 1. \tag{A.99}$$

This together with (A.91) and (A.97) yields that

$$\Pr\left(|I_3^*| \le \frac{4\bar{c}^2_{***}\eta_n^2}{\bar{c}^2}2(\bar{c})^{-1/2}c_0\right)\to 1.$$

Recall that

$$\widetilde{\boldsymbol{\omega}}_{\mathcal{M},j_0} = \widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0} + \sum_{j=1}^p \widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}}\left(\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},j_0} + \widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},\mathcal{M}}\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0}\right)(\widetilde{\beta}_j - \beta_{0,j}).$$

Hence, in order to prove

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\|\widehat{\boldsymbol{\omega}}_{\mathcal{M},j_0} - \widetilde{\boldsymbol{\omega}}_{\mathcal{M},j_0}\|_2 \le \bar{c}_0\eta_n^2, \tag{A.100}$$

it suffices to show the following events occur with probability tending to 1,

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\left\|I_1^* - \sum_{j=1}^p \widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}}\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},j_0}(\tilde{\beta}_j - \beta_{0,j})\right\|_2 = O(\eta_n^2), \tag{A.101}$$

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\left\|I_4^* - \sum_{j=1}^p \widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}}\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},\mathcal{M}}\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0}(\tilde{\beta}_j - \beta_{0,j})\right\|_2 = O(\eta_n^2). \tag{A.102}$$

We first prove (A.101). Similar to (A.62), we can show

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\|\widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}}\|_2 = O(1),$$

with probability tending to 1. By the definition of $\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},j_0}$, it suffices to show

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0} - \widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0} - \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_{i,\mathcal{M}}b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)X_{i,j_0}\{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}\right\|_2 \le c_1^*\eta_n^2, \tag{A.103}$$

for some constant $c_1^* > 0$, with probability tending to 1. This is equivalent to show

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2=1}}\left|\boldsymbol{a}^T\left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0}-\frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_{i,\mathcal{M}}b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)X_{i,j_0}\{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}\right)\right|\le c_1^*\eta_n^2,$$

with probability tending to 1. For any $\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}$, it follows from Taylor's theorem that

$$\boldsymbol{a}^T\left(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0}\right)=\frac{1}{n}\sum_{i=1}^n\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}^*_{\boldsymbol{a}})X_{i,j_0}\{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\},$$

for some $\boldsymbol{\beta}_a^*$ lying on the line segment joining $\boldsymbol{\beta}_0$ and $\widetilde{\boldsymbol{\beta}}$. The function b''' is Lipschitz continuous. Similar to (A.39), we can show

$$\sup_{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}}|b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}_a^*)-b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)|\le L_0|\boldsymbol{X}_i^T(\boldsymbol{\beta}_0-\boldsymbol{\beta}_a^*)|\le L_0|\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0)|,$$

for some constant $L_0 > 0$, with probability tending to 1. This together with Condition (A4*) yields that

$$\begin{aligned}\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2=1}}&\left\|\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},j_0}-\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},j_0}-\frac{1}{n}\sum_{i=1}^n\boldsymbol{X}_{i,\mathcal{M}}b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)X_{i,j_0}\{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}\right\|_2\\&\le L_0\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2=1}}\left|\frac{1}{n}\sum_{i=1}^n|\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}||X_{i,j_0}|\{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}^2\right|\\&\le L_0\omega_0\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2=1}}\left|\frac{1}{n}\sum_{i=1}^n|\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}}|\{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\}^2\right|,\quad\text{(A.104)}\end{aligned}$$

with probability tending to 1. Now (A.103) can be proven in a similar manner as (A.96). This further implies (A.101) holds.

The proof of (A.102) is more involved. Define

$$I_4^{**}=\widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}}(\widehat{\boldsymbol{\Sigma}}_{\mathcal{M},\mathcal{M}}-\widehat{\boldsymbol{\Sigma}}^*_{\mathcal{M},\mathcal{M}})\boldsymbol{\omega}_{\mathcal{M},j_0}$$

Using similar arguments in proving (A.101), we can show

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \left\| I_4^{**} - \sum_{j=1}^{p} \widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}} \widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},\mathcal{M}} \boldsymbol{\omega}_{\mathcal{M},j_0}(\tilde{\beta}_j - \beta_{0,j}) \right\| \tag{A.105}$$

$$\le O(1) \max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2=1}} \left| \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{a}^T \boldsymbol{X}_{i,\mathcal{M}} \boldsymbol{X}_{i,\mathcal{M}}^T \boldsymbol{\omega}_{\mathcal{M},j_0}| \{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}^2 \right|,$$

with probability tending to 1, where $O(1)$ denotes some positive constant. Using similar arguments in proving (A.75) and (A.96), we can show the last term is upper bounded by $O(\eta_n^2)$ with probability tending to 1, under the condition that $\kappa_n^3 = O(n), \kappa_n^{5/2}\log p = O(n/\log^2 n)$. Hence, to prove (A.102), it suffice to show

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \left\| I_4^{*} - I_4^{**} - \sum_{j=1}^{p} \widehat{\boldsymbol{\Sigma}}^{*-1}_{\mathcal{M},\mathcal{M}} \widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},\mathcal{M}} (\widehat{\boldsymbol{\omega}}^*_{\mathcal{M},j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0})(\tilde{\beta}_j - \beta_{0,j}) \right\| = o_p(\eta_n^2). \tag{A.106}$$

Using similar arguments in proving (A.105), we have by (A.98) that the LHS of (A.106) can be upper bounded by

$$|R_0| \max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2=1}} \left| \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{a}^T \boldsymbol{X}_{i,\mathcal{M}}| \{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}^2 \right|.$$

for some random variable $R_0 = O_p(n^{-1/2}\kappa_n\sqrt{\log p})$. Under the condition that $\kappa_n^{5/2}\log p = O(n/\log^2 n)$, we can show similarly that the above expression is $o_p(\eta_n^2)$. This proves (A.106). As a result, (A.102) and (A.100) are proven. Similarly, we can show

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \left\| \hat{\sigma}^2_{\mathcal{M},j_0} - \tilde{\sigma}^2_{\mathcal{M},j_0} \right\|_2 \le \bar{c}_0 \eta_n^2.$$

This together with (A.100) proves (A.31). We omit the details to save space.

Finally, we show

$$\sum_{t=0}^{n-1} \frac{\widetilde{Z}_{t+1,j_0}\varepsilon_{t+1}}{\sqrt{n}} \left(\frac{1}{\hat{\sigma}^*_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} - \frac{\sum_j \hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}(\tilde{\beta}_j - \beta_{0,j})}{\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(-sn)}_{j_0},j_0}} \right) = \sum_{t=0}^{n-1} \frac{\widehat{Z}^*_{t+1,j_0}\varepsilon_{t+1}}{\sqrt{n}\hat{\sigma}^*_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} + o_p(1). \tag{A.107}$$

With some calculations, we have

$$
\begin{aligned}
& \sum_{t=0}^{n-1} \frac{\widetilde{Z}_{t+1,j_0}\varepsilon_{t+1}}{\sqrt{n}} \left(\frac{1}{\hat{\sigma}^*_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} - \frac{\sum_j \hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} (\tilde{\beta}_j - \beta_{0,j})}{\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} \right) - \sum_{t=0}^{n-1} \frac{\widehat{Z}^*_{t+1,j_0}\varepsilon_{t+1}}{\sqrt{n}\hat{\sigma}^*_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} \\
= & \underbrace{\sum_{j=1}^{p} \left(\sum_{t=0}^{n-1} \frac{\widehat{Z}^*_{t+1,j_0}\varepsilon_{t+1}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} (\tilde{\beta}_j - \beta_{0,j}) \right)}_{\eta_1^*} + \underbrace{\sum_{t=0}^{n-1} \frac{(\widetilde{Z}_{t+1,j_0} - \widehat{Z}^*_{t+1,j_0})\varepsilon_{t+1}}{\sqrt{n}\hat{\sigma}^*_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}}_{\eta_2^*} \\
+ & \underbrace{\sum_{j=1}^{p} \left(\sum_{t=0}^{n-1} \frac{(\widetilde{Z}_{t+1,j_0} - \widehat{Z}^*_{t+1,j_0})\varepsilon_{t+1}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} (\tilde{\beta}_j - \beta_{0,j}) \right)}_{\eta_3^*}.
\end{aligned}
$$

In the following, we first prove $\eta_1^* = o_p(1)$. Note that $|\eta_1^*| \le \max_j |\eta_{1,j}^*| \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1$ where

$$
\eta_{1,j}^* = \sum_{t=0}^{n-1} \frac{\widehat{Z}^*_{t+1,j_0}\varepsilon_{t+1}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}.
$$

By Condition (A5*), it suffices to show $\max_j |\eta_{1,j}^*| = O_p(\sqrt{\log p})$.

It follows from the Lipschitz continuity of $b'''(\cdot)$ that

$$
|b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0) - b'''(0)| \le L_0 |\boldsymbol{X}_i^T\boldsymbol{\beta}_0|.
$$

Hence, under Condition (A4*), $\max_{1\le i\le n} |b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)|$ is bounded by some universal constant. Since $X_{0,j}$'s are uniformly bounded, we obtain

$$
\max_{1\le j\le p} |\widehat{\boldsymbol{\Psi}}^{(j)}_{j_0,j_0}| = \max_{1\le j\le p} \left| \frac{1}{n} \sum_i X_{i,j}^3 b'''(\boldsymbol{X}_i^T\boldsymbol{\beta}_0) \right| = O(1). \tag{A.108}
$$

Similarly, we can show

$$
\begin{aligned}
& \max_{1\le j\le p} \max_{\substack{\mathcal{M}\subseteq \mathbb{I}_{j_0} \\ |\mathcal{M}|\le \kappa_n}} \|\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},\mathcal{M}}\|_2 = \max_{1\le j\le p} \max_{\substack{\mathcal{M}\subseteq \mathbb{I}_{j_0} \\ |\mathcal{M}|\le \kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|} \\ \|\boldsymbol{a}\|_2=1}} |\boldsymbol{a}^T\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},\mathcal{M}}\boldsymbol{a}| \\
\le \;& O(1) \max_{\substack{\mathcal{M}\subseteq \mathbb{I}_{j_0} \\ |\mathcal{M}|\le \kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|} \\ \|\boldsymbol{a}\|_2=1}} \left| \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}})^2 \right|,
\end{aligned}
$$

where $O(1)$ denotes some positive constant. Using similar arguments in bounding η_5, we can show

$$\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}|}\\\|\boldsymbol{a}\|_2=1}} \left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}^T\boldsymbol{X}_{i,\mathcal{M}})^2\right| = O(1),$$

with probability tending to 1 and hence,

$$\max_{1\le j\le p}\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},\mathcal{M}}\|_2 = O(1), \tag{A.109}$$

with probability tending to 1. Similarly, we can show

$$\max_{1\le j\le p}\max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} \|\widehat{\boldsymbol{\Psi}}^{(j)}_{\mathcal{M},j_0}\|_2 = O(1), \tag{A.110}$$

with probability tending to 1. This together with (A.99), (A.108) and (A.109) yields

$$\max_{1\le j\le p}\max_{\mathcal{M}\subseteq\mathbb{I}_{j_0},|\mathcal{M}|\le\kappa_n} |\hat{\xi}^{(j)}_{\mathcal{M},j_0}| = O(1), \tag{A.111}$$

with probability tending to 1.

Note that

$$\eta^*_{1,j} = \underbrace{\sum_{t=0}^{s_n-1}\frac{\widehat{Z}^*_{t+1,j_0}\varepsilon_{t+1}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}}_{\eta^{**}_{1,j}} + \underbrace{\sum_{t=s_n}^{n-1}\frac{\widehat{Z}^*_{t+1,j_0}\varepsilon_{t+1}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}}_{\eta^{***}_{1,j}}.$$

We first prove $\max_j|\eta^{***}_{1,j}| = O_p(\sqrt{\kappa_n\log p})$.

Define $\sigma(\mathcal{F}^*_t) = \sigma(\boldsymbol{X}_1,\boldsymbol{X}_2,\dots,\boldsymbol{X}_n,Y_1,Y_2,\dots,Y_t)$, $\eta^{***}_{1,j}$ corresponds to a mean-zero martingale with respect to the filtration $\{\sigma(\mathcal{F}^*_t): t\ge s_n\}$. By Condition (A1*) and (A4*), we have for any $t=0,\dots,n-1$,

$$|\widehat{Z}^*_{t+1,j_0}| \le \omega_0 \max_{\substack{\mathcal{M}\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}|\le\kappa_n}} (1+\sqrt{\kappa_n}\|\widehat{\omega}^*_{\mathcal{M},j_0}\|_2). \tag{A.112}$$

Let

$$\bar{c}_n^{(j)} \equiv \omega_0 n^{-1/2} \max_{1\le j\le p} \max_{\substack{\mathcal{M}\subseteq \mathbb{I}_{j_0} \\ |\mathcal{M}|\le \kappa_n}} \frac{|(1+\sqrt{\kappa}_n \|\widehat{\omega}^*_{\mathcal{M},j_0}\|_2)|}{|\hat{\sigma}^*_{\mathcal{M},j_0}|} |\hat{\xi}^{(j)}_{\mathcal{M},j_0}|.$$

Under Condition (A6*), $\|\varepsilon_{t+1}\|_{\psi_1|\mathcal{F}_t^*} \le L^*$ for some constant $L^*>0$. Similar to (A.84), we can show

$$\mathrm{E}\{|\varepsilon_{t+1}|^k|\mathcal{F}_t^*\} \le k!(L^*)^k,$$

for any k and t. By Condition (A1*) and (A.112), we have

$$\mathrm{E}\left\{\left|\frac{\widehat{Z}^*_{t+1,j_0}\varepsilon_{t+1}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}\right|^k \middle| \mathcal{F}_t^*\right\} \le \left(\frac{\widehat{Z}^*_{t+1,j_0}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}\right)^2 k!(L^*)^k(\bar{c}_n^{(j)})^{k-2}, \tag{A.113}$$

for any j,t and $k\ge 2$.

Let

$$V_n^{(j)} = 2\sum_{t=s_n}^{n-1}\left(\frac{\widehat{Z}^*_{t+1,j_0}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}\right)^2.$$

Similar to (A.41) and (A.12), we can show $\sum_{t=s_n}^{n-1}(\widehat{Z}^*_{t+1,j_0})^2 = O_p(n)$ and $\min_t \hat{\sigma}^{*3}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} \le 2/\sqrt{\bar{c}}$, respectively. It follows from (A.111) that

$$\max_j V_n^{(j)} = O_p(1). \tag{A.114}$$

It follows from Theorem 9.12 in de la Peña et al. (2009) that for any $1\le j\le p$,

$$\Pr(|\eta^{***}_{1,j}|>z, V_n^{(j)}\le \bar{z}) \le 2\exp\left(-\frac{z^2}{2(L^*)^2\bar{z}+2L^*\bar{c}_n^{(j)}z}\right) \le 2\exp\left(-\frac{z^2}{\max\{4(L^*)^2\bar{z}, 4L^*\bar{c}_n^{(j)}z\}}\right).$$

Take $z_0^{(j)} = \max(3L^*\sqrt{\bar{z}\log p}, 8L^*\bar{c}_n^{(j)}\log p)$, we have

$$\Pr(|\eta^{***}_{1,j}|>z_0^{(j)}, V_n^{(j)}\le n\bar{z}) \le 2\exp(-2\log p) = \frac{2}{p^2}.$$

It follows from Bonferroni's inequality that

$$\Pr\left(\bigcap_{j=1}^{p}\left\{|\eta_{1,j}^{***}|>z_0^{(j)}\right\}, \max_j V_n^{(j)}\le n\bar{z}\right)\le\sum_{j=1}^{p}\Pr(|\eta_{1,j}^{***}|>z_0^{(j)})=\frac{2}{p}\to 0.$$

By (A.114), for any $\epsilon>0$, there exists some $\bar{z}>0$ such that $\Pr(\max_j V_n^{(j)}\le n\bar{z})\ge 1-\epsilon$.

This implies that

$$\max_{j=1}^{p}|\eta_{1,j}^{***}|\le\max_{j=1}^{p} z_0^{(j)}, \tag{A.115}$$

with probability tending to $1-\epsilon$. By (A.37) and (A.99), we have $\max_j \bar{c}_n^{(j)}=O_p(\sqrt{\kappa_n})$ and hence

$$\max_j|\eta_{1,j}^{***}|=O_p(\sqrt{\log p}), \tag{A.116}$$

by (A.115) and the condition that $\kappa_n^{5/2}\log p=O(n/\log^2 n)$.

Recall that

$$\eta_{1,j}^{**}=\sum_{t=0}^{s_n-1}\frac{\widehat{Z}^{*}_{t+1,j_0}\varepsilon_{t+1}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}}.$$

Given $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ and $Y_{s_n+1},\dots,Y_n$, each term in $\eta_{1,j}^{**}$ is independent of others. Using similar arguments, we can show $\max_j|\eta_{1,j}^{**}|=O_p(\sqrt{\log p})$. This together with (A.116) gives $\max_j|\eta_{1,j}^{*}|=O_p(\sqrt{\log p})$. By Condition (A5*), we obtain $|\eta_1^*|\le\max_j\max_j|\eta_{1,j}^*|\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1=o_p(1)$. Similarly, we can show $\eta_2^*=o_p(1)$. It remains to show $\eta_3^*=o_p(1)$.

Note that $|\eta_3^*|$ can be upper bounded by $\max_j|\eta_{3,j}^*|\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1$ where

$$\eta_{3,j}^{*}=\sum_{t=0}^{n-1}\frac{(\widetilde{Z}_{t+1,j_0}-\widehat{Z}^{*}_{t+1,j_0})\varepsilon_{t+1}\hat{\xi}^{(j)}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}{\sqrt{n}\hat{\sigma}^{*3}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}.$$

Since

$$\widehat{Z}^{*}_{t+1,j_0}-\widetilde{Z}_{t+1,j_0}=\sum_{j=1}^{p}\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}\widehat{\boldsymbol{\Sigma}}^{*-1}_{\widehat{\mathcal{M}}_{j_0}^{(t)},\widehat{\mathcal{M}}_{j_0}^{(t)}}\left(\widehat{\boldsymbol{\Psi}}^{(j)}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}+\widehat{\boldsymbol{\Psi}}^{(j)}_{\widehat{\mathcal{M}}_{j_0}^{(t)},\widehat{\mathcal{M}}_{j_0}^{(t)}}\widehat{\boldsymbol{\omega}}^{*}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}\right)(\widetilde{\beta}_j-\beta_{0,j}).$$

Using similar arguments in proving $\max_j \eta_{1,j}^{**} = O_p(\sqrt{\log p})$, we can show

$$\max_{j_1,j_2}\left|\sum_{t=0}^{n-1}\frac{\boldsymbol{X}_{t+1,\widehat{\mathcal{M}}_{j_0}^{(t)}}^T\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},\widehat{\mathcal{M}}_{j_0}^{(t)}}^{*-1}\left(\widehat{\boldsymbol{\Psi}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^{(j_1)}+\widehat{\boldsymbol{\Psi}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},\widehat{\mathcal{M}}_{j_0}^{(t)}}^{(j_1)}\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^{*}\right)\varepsilon_{t+1}\hat{\xi}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^{(j_2)}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}^{*3}}\right| = O_p(\sqrt{\log p}).$$

Hence, we have $|\eta_3^*| = O_p(\sqrt{\log p}(\sqrt{\kappa_n}\eta_n)^2) = o_p(1)$, by (A5*). The proof is hence completed.

A.9 Technical lemmas

Lemma A.5. *For any positive definite matrix*

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} \end{pmatrix},$$

denote its inverse matrix as $\boldsymbol{\Omega}$ *and partition it into* $\boldsymbol{\Omega}_{11}, \ldots, \boldsymbol{\Omega}_{22}$ *accordingly. Then,*

$$\boldsymbol{\Omega}_{11} = (\boldsymbol{\Psi}_{11} - \boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_{22}^{-1}\boldsymbol{\Psi}_{21})^{-1}.$$

Besides, let $\boldsymbol{\Psi}_* = \boldsymbol{\Psi}_{22} - \boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12}$*, we have*

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Psi}_{11}^{-1} + \boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_*^{-1}\boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1} & -\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_*^{-1} \\ -\boldsymbol{\Psi}_*^{-1}\boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1} & \boldsymbol{\Psi}_* \end{pmatrix}.$$

B More on the technical conditions

B.1 More on (A1) and (A1*)

The validity of the sure screening property assumed in (A1) or (A1*) relies typically on the following minimum-signal-strength condition:

$$\min_{j \in \mathcal{M}_{j_0}} |\beta_{0,j}| \geq \sigma_n^*, \tag{B.1}$$

for some monotonically nonincreasing sequence $\{\sigma_n^*\}_n$ that satisfies $\sigma_n^* \gg n^{-1/2}$ and $\sigma_n^* \to 0$ as $n \to \infty$. Although such conditions are not assumed in van de Geer et al. (2014) or Ning and Liu (2017), these authors imposed some additional assumptions on the design matrix. For instance, consider the decorrelated score statistic proposed by Ning and Liu (2017). For linear regression models, its validity depends on the sparsity of a high-dimensional vector $\boldsymbol{w}^*$. When the covariates follow a Gaussian graphical model, the sparsity assumption on $\boldsymbol{w}^*$ requires the degree of a particular node in the graph to be relatively small. See Remark 6 of Ning and Liu (2017) for details.

B.1.1 A counterexample

Assume $\boldsymbol{X}_0 \sim N(0, \{\rho^{|i-j|}\}_{i,j=1,\dots,p})$ for some $0<\rho<1$, $Y_0 = \boldsymbol{X}_0^T \boldsymbol{\beta}_0 + \varepsilon_0$ where $\varepsilon_0 \sim N(0,1)$ that is independent of $\boldsymbol{X}_0$ and $\beta_{0,1}=0$, $\beta_{0,2}=n^{-1/2}$, $\beta_{0,j}=0$ for all $j>2$. The minimum-signal-strength condition (B.1) is thus violated. Our goal is to construct a CI for $\beta_{0,1}$.

Suppose we use SIS to determine the set of important variables based on their marginal correlations with the response. Specifically, set

$$\begin{aligned} \widehat{\mathcal{M}}_1^{(t)} &= \left\{ j \ge 2 : t^{-1} \sum_{i=1}^{t} |Y_i X_{i,j}| \ge \sigma_t \right\}, \qquad \forall s_n \le t < n, \\ \widehat{\mathcal{M}}_1^{(-s_n)} &= \left\{ j \ge 2 : (n-s_n)^{-1} \sum_{i=s_n+1}^{n} |Y_i X_{i,j}| \ge \sigma_{n-s_n} \right\}, \end{aligned}$$

for some sequence $\{\sigma\}_n$ that satisfies $\sigma_n \gg n^{-1/2}\log^{1/2} n$.

Notice that for any $j \ge 2$, we have $\mathrm{E} Y_0 X_{0,j} = n^{-1/2} \mathrm{E} X_{0,2} X_{0,j} = n^{-1/2} \rho^{j-2}$. Using Bernstein's inequality, we can show that the following events occur with probability tending to 1 that

$$\begin{aligned} t^{-1} \sum_{i=1}^{t} |Y_i X_{i,j}| &\le O(1) t^{-1/2} \sqrt{\log t + \log p}, \qquad \forall s_n \le t < n, 2 \le j \le p, \\ (n-s_n)^{-1} \sum_{i=s_n+1}^{n} |Y_i X_{i,j}| &\le O(1)(n-s_n)^{-1/2} \sqrt{\log(n-s_n) + \log p}, \quad \forall 2 \le j \le p, \end{aligned}$$

where $O(1)$ denotes some positive constant. Suppose $p = O(n)$ and the sequence s_n is set to be $\lfloor \epsilon n \rfloor$ for some $0 < \epsilon < 1$. It follows that

$$\begin{aligned} \max_{s_n \le t \le n, 2 \le j \le p} t^{-1/2} \sqrt{\log t + \log p} &= O(n^{-1/2} \log^{1/2} n), \\ (n-s_n)^{-1/2} \sqrt{\log(n-s_n) + \log p} &= O(n^{-1/2} \log^{1/2} n). \end{aligned}$$

Hence, for sufficiently large n, we have $\widehat{\mathcal{M}}_1^{(-s_n)} = \widehat{\mathcal{M}}_1^{(s_n)} = \widehat{\mathcal{M}}_1^{(s_n+1)} = \widehat{\mathcal{M}}_1^{(n-1)} = \emptyset$, with probability tending to 1.

As a result, our score equation for $\beta_{0,1}$ is given by

$$\sum_{i=1}^{n} X_{i,1}(Y_i - X_{i,1}^T \beta_{0,1}) = 0,$$

with probability tending to 1. Therefore, the proposed CI for $\beta_{0,1}$ equals

$$\left[\hat{\beta}_1 - z_{\alpha/2} n^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} X_{i,1}^2 \right)^{-1/2}, \hat{\beta}_1 + z_{\alpha/2} n^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} X_{i,1}^2 \right)^{-1/2} \right],$$

where

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} X_{i,1}^2 \right)^{-1} \left(\sum_{i=1}^{n} X_{i,1} Y_i \right).$$

It follows that

$$\left(\sum_{i=1}^{n} X_{i,1}^2 \right)^{1/2} (\hat{\beta}_1 - \beta_{0,1}) = \left(\sum_{i=1}^{n} X_{i,1}^2 \right)^{-1/2} \left(\sum_{i=1}^{n} X_{i,1}(Y_i - X_{i,1}\beta_{0,1}) \right)$$
$$= \left(\sum_{i=1}^{n} X_{i,1}^2 \right)^{-1/2} \left(\sum_{i=1}^{n} X_{i,1}\varepsilon_i \right) + n^{-1/2} \left(\sum_{i=1}^{n} X_{i,1}^2 \right)^{-1/2} \left(\sum_{i=1}^{n} X_{i,1} X_{i,2} \right). \tag{B.2}$$

By the central limit theorem, the first term on the RHS of (B.2) converges to $N(0,1)$ in distribution. The second term converges to ρ, according to the law of large numbers. Hence, our CI is not valid as long as $\rho > 0$. This implies that the minimal-signal-strength condition is necessary to guarantee the validity of our procedure.

B.1.2 Extension to many small but weak signals

Moreover, one could relax the minimum-signal-strength condition in (B.1) by assuming there are many small but weak signals in $\boldsymbol{\beta}_0$. Specifically, assume $\mathcal{M}_{j_0}$ is a union of two disjoint subsets $\mathcal{M}_{j_0}^*$ and $\mathcal{M}_{j_0}^{**}$ such that

$$\mathcal{M}_{j_0}^* = \{ j \in \mathbb{I}_{j_0} : |\beta_{0,j}| \geq \sigma_n^* \}, \tag{B.3}$$

and

$$\mathcal{M}_{j_0}^{**} = \mathcal{M}_{j_0} \cap (\mathcal{M}_{j_0}^{*})^c \quad \text{with} \quad \|\boldsymbol{\beta}_{0,\mathcal{M}_{j_0}^{**}}\|_2 = O(n^{-\kappa^*}), \tag{B.4}$$

for some sequence $n^{-1/2} \ll \sigma_n^* \ll 1$ and some constant $\kappa^* > 1/2$. We require $|\mathcal{M}_{j_0}^*|$ is much smaller than n while $|\mathcal{M}_{j_0}^{**}|$ can be much larger than the sample size. Such conditions are very similar to the zonal assumption imposed by Bühlmann and Mandozzi (2014). When (B.3) and (B.4) hold, Condition (A1) or (A1*) can then be replaced by the following:

(A1**) Assume $\widehat{\mathcal{M}}_{j_0}^{(n)}$ satisfies $\Pr(|\widehat{\mathcal{M}}_{j_0}^{(n)}| \le \kappa_n) = 1$ for some $1 \le \kappa_n = o(n)$. Besides,

$$\Pr\left(\mathcal{M}_{j_0}^* \subseteq \widehat{\mathcal{M}}_{j_0}^{(n)}\right) \ge 1 - O\left(\frac{1}{n^{\alpha_0}}\right),$$

for some constant $\alpha_0 > 1$.

That is, we require the selected model will contain all those strong signals with probability tending to 1. This assumption can be satisfied under the condition in (B.3). In the following, we sketch a few lines to show the proposed method works. For simplicity, we focus on linear regression models.

By (A1**) and Bonferroni's inequality, the following event occurs with probability tending to 1,

$$\mathcal{M}_{j_0}^* \subseteq \bigcap_{t=s_n}^{n-1} \widehat{\mathcal{M}}_{j_0}^{(t+1)}. \tag{B.5}$$

Under the event defined in (B.5), we have

$$\begin{aligned} &\sqrt{n}\Gamma_n^*(\hat{\beta}_{j_0} - \beta_{0,j_0}) = I_1 + I_2 + I_3 + I_4 \\ +\ & \sum_{t=0}^{s_n-1} \frac{\widehat{Z}_{t+1,j_0} \boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}_{j_0}^{(-s_n)})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}_{j_0}^{(-s_n)})^c}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(-s_n)},j_0}} + \sum_{t=s_n}^{n-1} \frac{\widehat{Z}_{t+1,j_0} \boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}_{j_0}^{(t)})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}_{j_0}^{(t)})^c}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}, \end{aligned}$$

and

$$\left|\sum_{t=0}^{s_n-1} \frac{\widehat{Z}_{t+1,j_0}\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(-s_n)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(-s_n)}_{j_0})^c}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}^{(-s_n)}_{j_0},j_0}}\right| \le \sum_{t=0}^{s_n-1} \frac{|\widehat{Z}_{t+1,j_0}||\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(-s_n)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(-s_n)}_{j_0})^c}|}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}^{(-s_n)}_{j_0},j_0}},$$

$$\left|\sum_{t=s_n}^{n-1} \frac{\widehat{Z}_{t+1,j_0}\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(-s_n)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(-s_n)}_{j_0})^c}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}^{(-s_n)}_{j_0},j_0}}\right| \le \sum_{t=s_n}^{n-1} \frac{|\widehat{Z}_{t+1,j_0}||\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(-s_n)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(-s_n)}_{j_0})^c}|}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}},$$

where I_1, I_2, I_3 and I_4 are defined in Section A.1, $\widehat{Z}_{t+1,j_0}$, $\widehat{\mathcal{M}}^{(-s_n)}_{j_0}$, $\widehat{\mathcal{M}}^{(t)}_{j_0}$, $\hat{\sigma}_{\widehat{\mathcal{M}}^{(-s_n)}_{j_0},j_0}$ and $\hat{\sigma}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}$ are defined in Section 2.

Note that we have shown in Section A.1 that $I_1 + I_2 + I_3 + I_4$ is asymptotically normal. It suffices to show

$$\sum_{t=0}^{n} \frac{|\widehat{Z}_{t+1,j_0}||\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}|}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}^{(-s_n)}_{j_0},j_0}} = o_p(1), \tag{B.6}$$

where $\mathcal{M}^{(t)}_{j_0} = \mathcal{M}^{(-s_n)}_{j_0}$, for $t = 0, \dots, s_n - 1$. Under the event defined in (A.12) and (A1**), the LHS of (B.6) is upper bounded by

$$I_5 \equiv \sum_{t=0}^{n-1} \frac{2|\widehat{Z}_{t+1,j_0}||\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}|}{\sqrt{\bar{c}n}}.$$

By Cauchy-Schwarz inequality, we have

$$I_5 \le \frac{2}{\sqrt{\bar{c}n}} \left(\sum_{t=0}^{n-1} |\widehat{Z}_{t+1,j_0}|^2\right)^{1/2} \left(\sum_{t=0}^{n-1} |\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}|^2\right)^{1/2}. \tag{B.7}$$

Similar to (A.45), we can show $\sum_{t=0}^{n-1} |\widehat{Z}_{t+1,j_0}|^2 = O_p(n)$ under the given conditions in Theorem 2.1. Under (A1**), we have $\|\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\|_2 \le \|\boldsymbol{\beta}_{0,\mathcal{M}^{**}_{j_0}}\|_2$, almost surely for any $t = 0, 1, \dots, n-1$. This together with (A.22) and (B.4) yields that

$$\mathrm{E}\sum_{t=0}^{n-1} |\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}|^2 \le n\lambda_{\max}(\boldsymbol{\Sigma})\mathrm{E}\|\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\|_2^2 = O(n^{1-2\kappa^*}) = o(1). \tag{B.8}$$

By Markov's inequality, we obtain $\sum_{t=0}^{n-1} |\boldsymbol{X}^T_{t+1,\mathcal{M}^{**}_{j_0}}\boldsymbol{\beta}_{0,\mathcal{M}^{**}_{j_0}}|^2 = o_p(1)$. In view of (B.7), we have shown $I_5 = o_p(1)$. The proof is hence completed.

B.1.3 Additional details regarding the doubly-robust procedure

To better understand the proposed algorithm in Section 5.4, we decompose $\mathcal{M}_{j_0}$ into $\mathcal{M}_{j_0}^*$ and $\mathcal{M}_{j_0}^{**}$ as in Section B.1.2, where $\mathcal{M}_{j_0}^*$ denotes the set of strong signals that satisfies (B.3) and $\mathcal{M}_{j_0}^{**} = \mathcal{M}_{j_0} \cap (\mathcal{M}_{j_0}^*)^c$ is the set of weak signals.

In case the set $\mathcal{M}_{j_0}^{**}$ is nonempty, we can apply another model selection procedure to estimate the support of $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0}$ (denoted by $\mathcal{M}_{\boldsymbol{\omega}}$), in order to gain some robustness. For linear regression models, $\mathcal{M}_{\boldsymbol{\omega}}$ can be estimated by (I)SIS or regularized regression, with X_{i,j_0}'s being the responses and $\boldsymbol{X}_{i,\mathbb{I}_{j_0}}$'s being the covariates. Similarly, we decompose $\mathcal{M}_{\boldsymbol{\omega}}$ into $\mathcal{M}_{\boldsymbol{\omega}}^*$ and $\mathcal{M}_{\boldsymbol{\omega}}^{**}$, corresponding to the set of strong and weak signals in $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0}$, respectively.

Let $\widehat{\mathcal{M}}_{j_0}^{1,(t)}, s_n \le t < n$ and $\widehat{\mathcal{M}}_{j_0}^{1,(-s_n)}$ denote the estimated supports of $\boldsymbol{\beta}_{0,\mathbb{I}_{j_0}}$, and $\widehat{\mathcal{M}}_{j_0}^{2,(t)}, s_n \le t < n$ and $\widehat{\mathcal{M}}_{j_0}^{2,(-s_n)}$ the estimated supports of $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0}$. We will assume the following occurs with probability tending to 1,

$$\mathcal{M}_{j_0}^* \subseteq \left\{ \bigcap_{t=s_n}^{n} \widehat{\mathcal{M}}_{j_0}^{1,(t)} \right\} \bigcap \widehat{\mathcal{M}}_{j_0}^{1,(-s_n)} \quad \text{and} \quad \mathcal{M}_{\boldsymbol{\omega}}^* \subseteq \left\{ \bigcap_{t=s_n}^{n} \widehat{\mathcal{M}}_{j_0}^{2,(t)} \right\} \bigcap \widehat{\mathcal{M}}_{j_0}^{2,(-s_n)}.$$

Set $\widehat{\mathcal{M}}_{j_0}^{(t)} = \widehat{\mathcal{M}}_{j_0}^{1,(t)} \cup \widehat{\mathcal{M}}_{j_0}^{2,(t)}$, for $s_n \le t < n$ and $\widehat{\mathcal{M}}_{j_0}^{(-s_n)} = \widehat{\mathcal{M}}_{j_0}^{1,(-s_n)} \cup \widehat{\mathcal{M}}_{j_0}^{2,(-s_n)}$. We propose to use the union of these two sets in our algorithm to construct the CI for β_{0,j_0}. The number of elements in $\widehat{\mathcal{M}}_{j_0}^{(-s_n)}, \widehat{\mathcal{M}}_{j_0}^{(s_n)}, \cdots, \widehat{\mathcal{M}}_{j_0}^{(n-1)}$ shall be bounded by κ_n, almost surely. We require $\eta_n \sqrt{\kappa_n \log p} = o(1)$, $\kappa_n^2 \log p = O(n/\log^2 n)$ and $\kappa_n^2 \log^2 p = O(n)$. In the following, we focus on linear regression models and show the resulting CI for β_{0,j_0} is valid as long as either one of the following two conditions holds:

(i) $\mathcal{M}_{j_0}^{**} = \emptyset$.

(ii) $\|\boldsymbol{\beta}_{0,\mathcal{M}_{j_0}^{**}}\|_2 = o(n^{-1/4})$ and $\|\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}_{j_0}^{**}}\|_2 = o(n^{-1/4})$, where $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}_{j_0}^{**}}$ is the sub-vector of $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0}$ formed by elements in $\mathcal{M}_{j_0}^{**}$.

When (i) holds, the assertion can be proven in similar manner as Theorem 2.1. Consider the case where (ii) holds. Using similar arguments in Section B.1.2, it suffices to show

$$\sum_{t=0}^{n-1} \frac{\widehat{Z}_{t+1,j_0} \boldsymbol{X}_{t+1,(\widehat{\mathcal{M}}_{j_0}^{(t)})^c}^T \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}_{j_0}^{(t)})^c}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} = o_p(1). \tag{B.9}$$

We decompose the LHS of (B.9) into $I_6 + I_7 + I_8$ where

$$I_6 = \sum_{t=0}^{n-1} \left(\frac{\widehat{Z}_{t+1,j_0} \boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} - \frac{\widehat{Z}_{t+1,j_0} \boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} \right),$$

$$I_7 = \sum_{t=0}^{n-1} \frac{\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}} (\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}) \boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}},$$

$$I_8 = \sum_{t=0}^{n-1} \frac{Z_{t+1,j_0} \boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}}.$$

Under the event defined in (A.6) and (A.12), we have almost surely that

$$|I_6| \le \frac{2\bar{c}_0}{\bar{c}\sqrt{n}} \left(\frac{\sqrt{\kappa_n \log p}}{\sqrt{n}} \right) \sum_{t=0}^{n-1} |\widehat{Z}_{t+1,j_0} \boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}|.$$

Using similar arguments in bounding I_5 in Section B.1.2, we can show

$$\sum_{t=0}^{n-1} |\widehat{Z}_{t+1,j_0} \boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}| = o_p(n^{3/4}).$$

Under the condition $\kappa_n^2 \log^2 p = O(n)$, it follows that $I_6 = o_p(1)$.

Using similar arguments in bounding $I_2^{(2)}$ in the proof of Theorem 2.1, we have

$$\sum_{t=0}^{n-1} |\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}^{(t)}_{j_0}} (\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0} - \boldsymbol{\omega}_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0})|^2 = O_p(\kappa_n \log p).$$

In addition, similar to (B.8), we can show

$$\sum_{t=0}^{n-1} |\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}|^2 = o_p(n^{-1/2}). \tag{B.10}$$

By (A.1) and Cauchy-Schwarz inequality, we obtain $I_7 = o_p(n^{-1/4}\sqrt{\kappa_n \log p}) = o_p(1)$, under the condition that $\kappa_n^2 \log^2 p = O(n)$.

It remains to show $I_8 = o_p(1)$. Since $\mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}^T_{\mathbb{I}_{j_0},j_0} \boldsymbol{X}_{0,\mathbb{I}_{j_0}}) \boldsymbol{X}_{0,\mathbb{I}_{j_0}} = 0$, we have for any $\mathcal{M} \subseteq \mathbb{I}_{j_0}$ that $\mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}^T_{\mathbb{I}_{j_0},j_0} \boldsymbol{X}_{0,\mathbb{I}_{j_0}}) \boldsymbol{X}_{0,\mathcal{M}} = 0$. Thus, for any $\mathcal{M}$ that contains $\mathcal{M}_{\boldsymbol{\omega}}$, we have

$$\mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}}^T \boldsymbol{X}_{0,\mathcal{M}})\boldsymbol{X}_{0,\mathcal{M}} = 0,$$

where $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}}$ is the sub-vector of $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0}$ formed by elements in $\mathcal{M}$. This further implies $\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}} = \boldsymbol{\omega}_{\mathcal{M},j_0}$, for any $\mathcal{M}$ that contains $\mathcal{M}_{\boldsymbol{\omega}}$, and hence

$$\mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}_{\mathcal{M},j_0}^T \boldsymbol{X}_{0,\mathcal{M}})\boldsymbol{X}_{0,\mathbb{I}_{j_0}} = 0.$$

For an arbitrary set $\mathcal{M}^*$ that contains $\mathcal{M}_{\boldsymbol{\omega}}^*$, define $\mathcal{M}^{**} = \mathcal{M}^* \cup \mathcal{M}_{\boldsymbol{\omega}}^{**}$. It follows that

$$\mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^{**}}^T \boldsymbol{X}_{0,\mathcal{M}^{**}})\boldsymbol{X}_{0,\mathbb{I}_{j_0}} = 0, \tag{B.11}$$

and hence $\mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^*}^T \boldsymbol{X}_{0,\mathcal{M}^*} - \boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^{**}-\mathcal{M}^*}^T \boldsymbol{X}_{0,\mathcal{M}^{**}-\mathcal{M}^*})\boldsymbol{X}_{0,\mathcal{M}^*} = 0$. By (A.22), (ii) and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\sup_{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}^*|},\|\boldsymbol{a}\|_2=1} \mathrm{E}|\boldsymbol{a}^T \boldsymbol{X}_{0,\mathcal{M}^*}||\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^{**}-\mathcal{M}^*}^T \boldsymbol{X}_{0,\mathcal{M}^{**}-\mathcal{M}^*}| \\
\le &\left(\sup_{\boldsymbol{a}\in\mathbb{R}^{|\mathcal{M}^*|},\|\boldsymbol{a}\|_2=1} \mathrm{E}|\boldsymbol{a}^T \boldsymbol{X}_{0,\mathcal{M}^*}|^2\right)^{1/2} \left(\mathrm{E}|\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^{**}-\mathcal{M}^*}^T \boldsymbol{X}_{0,\mathcal{M}^{**}-\mathcal{M}^*}|^2\right)^{1/2} \\
&\le \lambda_{\max}(\boldsymbol{\Sigma})\|\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^{**}-\mathcal{M}^*}\|_2 \le \lambda_{\max}(\boldsymbol{\Sigma})\|\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}_{\boldsymbol{\omega}}^{**}}\|_2 = o(n^{-1/4}).
\end{aligned}$$

This yields $\|\mathrm{E}\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^{**}-\mathcal{M}^*}^T \boldsymbol{X}_{0,\mathcal{M}^{**}-\mathcal{M}^*} \boldsymbol{X}_{0,\mathcal{M}^*}\|_2 = o(n^{-1/4})$ and hence

$$\|\mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^*}^T \boldsymbol{X}_{0,\mathcal{M}^*})\boldsymbol{X}_{0,\mathcal{M}^*}\|_2 = o(n^{-1/4}).$$

Notice that $\mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}_{\mathcal{M}^*,j_0}^T \boldsymbol{X}_{0,\mathcal{M}^*})\boldsymbol{X}_{0,\mathcal{M}^*} = 0$. For any $\mathcal{M}^*$ that satisfies $|\mathcal{M}^*| \le \kappa_n$, it follows from Condition (A2) that

$$\begin{aligned}
\|\boldsymbol{\omega}_{\mathcal{M}^*,j_0} - \boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^*}\|_2 &\le \frac{\|\mathrm{E}(\boldsymbol{\omega}_{\mathcal{M}^*,j_0} - \boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^*})^T \boldsymbol{X}_{0,\mathcal{M}^*(\boldsymbol{X}_{0,\mathcal{M}^*})^T}\|_2}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{M}^*,\mathcal{M}^*})} \\
&= \frac{\|\mathrm{E}(X_{0,j_0} - \boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^*}^T \boldsymbol{X}_{0,\mathcal{M}^*})\boldsymbol{X}_{0,\mathcal{M}^*}\|_2}{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{M}^*,\mathcal{M}^*})} = o(n^{-1/4}).
\end{aligned}$$

To summarize, we have shown that

$$\max_{\substack{\mathcal{M}^*\subseteq\mathbb{I}_{j_0}\\|\mathcal{M}^*|\le\kappa_n,\mathcal{M}^*_{\boldsymbol{\omega}}\subseteq\mathcal{M}^*}} \|\boldsymbol{\omega}_{\mathcal{M}^*,j_0}-\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\mathcal{M}^*}\|_2 = o(n^{-1/4}).$$

Under the given conditions, we obtain

$$\max_{t\in\{0,1,\dots,n-1\}} \|\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{(t)},j_0}-\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}}\|_2 = o(n^{-1/4}),$$

almost surely. By (A.22), this yields

$$\sum_{t=0}^{n-1}\mathrm{E}|\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{(t)}}(\boldsymbol{\omega}_{\widehat{\mathcal{M}}_{(t)},j_0}-\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}})|^2 = o(\sqrt{n}).$$

Similarly, we can show

$$\sum_{t=0}^{n-1}\mathrm{E}|\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{(t)}}\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}}-\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}}\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}}|^2 = o(\sqrt{n}).$$

This together with (A.1), (B.10) and Cauchy-Schwarz inequality yields that

$$\sum_{t=0}^{n-1}\frac{(\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{(t)}}\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}}-\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}}\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}})\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} = o_p(1).$$

Thus, to show $I_8 = o_p(1)$, it suffices to show

$$\sum_{t=0}^{n-1}\frac{(X_{t+1,j_0}-\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}}\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}})\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} = o_p(1).$$

We first show

$$\sum_{t=s_n}^{n-1}\frac{(X_{t+1,j_0}-\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}}\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}})\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} = o_p(1). \tag{B.12}$$

By (B.11), the LHS of (B.12) forms a mean zero martingale with respect to the filtration $\{\sigma(\mathcal{F}_t) : t \ge s_n\}$. Moreover, it follows from (ii) that $\|\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\|_2 = o(1)$ and hence

$$\frac{1}{n}\sum_{t=s_n}^{n-1}\frac{\mathrm{E}\{(X_{t+1,j_0}-\boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}}\boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}})^2(\boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}\boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c})^2|\mathcal{F}_t\}}{\sigma^2_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} = o(1).$$

This proves (B.12). Similarly, we can show

$$\sum_{t=0}^{s_n-1} \frac{(X_{t+1,j_0} - \boldsymbol{X}^T_{t+1,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}} \boldsymbol{\omega}_{\mathbb{I}_{j_0},j_0,\widehat{\mathcal{M}}_{(t)}\cup\mathcal{M}^{**}_{\boldsymbol{\omega}}}) \boldsymbol{X}^T_{t+1,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c} \boldsymbol{\beta}_{0,(\widehat{\mathcal{M}}^{(t)}_{j_0})^c}}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}^{(t)}_{j_0},j_0}} = o_p(1).$$

The proof is hence completed.

B.2 More on (A2) and (A2*)

Condition (A2) requires $\lambda_{\min}(\boldsymbol{\Sigma}_{j_0\cup\mathcal{M},j_0\cup\mathcal{M}}) \geq \bar{c}$ for some constant $\bar{c} > 0$ and any $\mathcal{M} \subseteq \mathbb{I}$ and $|\mathcal{M}| \leq \kappa_n$, where $\boldsymbol{\Sigma} = \mathrm{E}\boldsymbol{X}_0\boldsymbol{X}_0^T$. This condition is similar to the restricted eigenvalue condition (Bickel et al., 2009) used to derive the oracle inequalities of the Lasso estimator and the Dantzig selector. Notice that this condition is weaker compared to the one used in van de Geer et al. (2014) or Ning and Liu (2017), which requires the minimum eigenvalue of $\boldsymbol{\Sigma}$ to be strictly positive. See Section 4.1 of Ning and Liu (2017), Condition (A2) and (B3) in van de Geer et al. (2014) for details.

B.3 More on (A5)

In this section, we provide a consistent estimator for σ_0^2. Specifically, define

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\widetilde{\boldsymbol{\beta}})^2.$$

In the following, we show $|\hat{\sigma}^2 - \sigma_0^2| = o_p(1)$. Notice that

$$\begin{aligned} |\hat{\sigma}^2 - \sigma_0^2| &= \left|\frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}_0 - \boldsymbol{X}_i^T\widetilde{\boldsymbol{\beta}})^2 - \sigma_0^2\right| \\ = & \left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}) + \frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}^2 - \sigma_0^2\right| \\ \leq & \left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \sigma_0^2\right| + \left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{X}_i^T(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}})\right| + \frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{X}_i^T(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}^2. \qquad \text{(B.13)} \end{aligned}$$

Under the condition $\mathrm{E}|\varepsilon_0|^3 = O(1)$, the first term on the RHS of (B.13) is $o_p(1)$ by the law of large numbers.

Suppose we can show

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i \right\|_\infty = O_p(\sqrt{\log p}). \tag{B.14}$$

It follows from Condition (A4) and (A.25) that the second term on the RHS of (B.13) is $o_p(1)$, since

$$\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i^T (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}) \right| \le \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i \right\|_\infty \|\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}\|_1 = O_p(\eta_n \sqrt{\kappa_n \log p}) = o_p(1).$$

The third term is $O_p(\eta_n^2)$ by (A.9). Under the given conditions, it is $o_p(1)$.

Therefore, to complete the proof, it suffices to show (B.14), or equivalently,

$$\mathrm{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i \right\|_\infty = O(\sqrt{\log p}), \tag{B.15}$$

by Markov's inequality. It follows from Lemma A.3 in Chernozhukov et al. (2013) that

$$\mathrm{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i \right\|_\infty \le O(1)(\sigma \sqrt{\log p} + \mathbb{M} \log p),$$

where $O(1)$ denotes some positive constant, $\sigma^2 = \max_{j \in \{1,\dots,p\}} \sum_{i=1}^{n} \mathrm{E}\varepsilon_i^2 X_{i,j}^2 / n^2$ and $\mathbb{M}^2 = \mathrm{E} \max_{1 \le j \le p} \max_{1 \le i \le n} X_{i,j}^2 \varepsilon_i^2 / n^2$. Notice that

$$\sigma^2 = \max_{1 \le j \le p} n^{-1} \mathrm{E}\varepsilon_0^2 X_{0,j}^2 = n^{-1} \sigma_0^2 \max_{1 \le j \le p} \mathrm{E}X_{0,j}^2 \le n^{-1} \sigma_0^2 \max_{1 \le j \le p} \|X_{0,j}\|_{\psi_2}^2 \le n^{-1} \sigma_0^2 c_0^2,$$

where the last equality is due to the independence between ε_0 and $\boldsymbol{X}_0$, the first inequality is due to the fact that $\mathrm{E}|Z|^2 \le \|Z^2\|_{\psi_1} = \|Z\|_{\psi_2}^2$ for any random variable Z and the last inequality is due to (A3).

Similarly, we can show

$$\begin{aligned} n^2 \mathbb{M}^2 &= \mathrm{E} \max_{1 \le i \le n} \varepsilon_i^2 \mathrm{E} \max_{\substack{1 \le i \le n \\ 1 \le j \le p}} X_{i,j}^2 \le \left(\mathrm{E} \max_{1 \le i \le n} |\varepsilon_i|^3 \right)^{2/3} \mathrm{E} \max_{\substack{1 \le i \le n \\ 1 \le j \le p}} X_{i,j}^2 \\ &\le \left(\mathrm{E} \sum_{1 \le i \le n} |\varepsilon_i|^3 \right)^{2/3} \mathrm{E} \max_{\substack{1 \le i \le n \\ 1 \le j \le p}} X_{i,j}^2 \le (\mathrm{E}|\varepsilon_0|^3)^{2/3} n^{2/3} \mathrm{E} \max_{\substack{1 \le i \le n \\ 1 \le j \le p}} X_{i,j}^2, \end{aligned}$$

where the first equality is due to the independence between ε_0 and $\boldsymbol{X}_0$ and the first inequality follows from Hölder's inequality. Using similar arguments in (A.60), (A.61) and (A.74), we can show

$$\begin{aligned}\mathrm{E}\max_{\substack{1\le i\le n\\1\le j\le p}} X_{i,j}^2 &\le \|\max_{\substack{1\le i\le n\\1\le j\le p}} X_{i,j}^2\|_{\psi_1} = K_1\{\log(1+pn)\}\max_{\substack{1\le i\le n\\1\le j\le p}}\|X_{i,j}^2\|_{\psi_1}\\ &= K_1\{\log(1+pn)\}\max_{\substack{1\le i\le n\\1\le j\le p}}\|X_{i,j}\|_{\psi_2}^2 = O(\log p+\log n),\end{aligned}$$

by (A3). Under the condition $\mathrm{E}|\varepsilon_0|^3 = O(1)$, it follows that $\mathbb{M}^2 = O(n^{-4/3}\log p + n^{-4/3}\log n)$. Therefore, we obtain

$$\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\boldsymbol{X}_i\right\|_\infty = O(n^{-1/2}\log^{1/2}p) + O(n^{-2/3}\log^{3/2}p) + O(n^{-2/3}\log p\sqrt{\log n}).$$

Under the condition that $\log p = O(n^{2/3})$, we have $\mathrm{E}\|\sum_{i=1}^n\varepsilon_i\boldsymbol{X}_i/n\|_\infty = O(\sqrt{\log p})$. This proves (B.15). The proof is hence completed.

C Additional details regarding extensions to generic M-estimators

It this section, we sketch a few lines to show that the CI proposed in Section 5.3 is valid. It suffices to show that

$$\sqrt{n}\Gamma_n^{*,(l-1)}(\hat{\beta}_{j_0}^{(l)} - \beta_{0,j_0}) \xrightarrow{d} N(0,1).$$

It follows from Taylor's theorem that

$$\begin{aligned}&\sum_{t=0}^{n-1}\frac{1}{n\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\left(\frac{\partial\ell(\boldsymbol{U}_{t+1},\mathrm{h}(\beta_{0,j_0},\widehat{\mathcal{M}}_{j_0}^{(t)},\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial\beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0}\frac{\partial\ell(\boldsymbol{U}_{t+1},\mathrm{h}(\beta_{0,j_0},\widehat{\mathcal{M}}_{j_0}^{(t)},\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial\boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}}\right)\\ =\;&\sum_{t=0}^{n-1}\frac{1}{n\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}}\left(\frac{\partial\ell(\boldsymbol{U}_{t+1},\mathrm{h}(\hat{\beta}_{j_0}^{(l-1)},\widehat{\mathcal{M}}_{j_0}^{(t)},\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial\beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0}\frac{\partial\ell(\boldsymbol{U}_{t+1},\mathrm{h}(\hat{\beta}_{j_0}^{(l-1)},\widehat{\mathcal{M}}_{j_0}^{(t)},\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial\boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}}\right)\\ +\;&\Gamma_n^{*,(l-1)}(\beta_{0,j_0} - \hat{\beta}_{j_0}^{l-1}) + \mathrm{Rem},\end{aligned}$$

where the second-order remainder term satisfies Rem $= o_p(n^{-1/2})$ under certain local smoothness assumption on the loss function ℓ.

By the definition of $\hat{\beta}_{j_0}^{(l)}$, we obtain that

$$
\begin{aligned}
&\sqrt{n}\Gamma_n^{*,(l-1)}(\hat{\beta}_{j_0}^{(l)} - \beta_{0,j_0}) = o_p(1) \\
+ &\sum_{t=0}^{n-1} \frac{1}{n\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right).
\end{aligned}
$$

It suffices to show

$$
\sum_{t=0}^{n-1} \frac{1}{n\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right) \xrightarrow{d} N(0,1).
$$

Under certain local smoothness assumptions on ℓ, it follows from Taylor's theorem that

$$
\begin{aligned}
&\sum_{t=0}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right) \\
= &\sum_{t=0}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right) \\
+ &\sum_{t=0}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial^2 \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \beta_{j_0} \partial \boldsymbol{\beta}^T_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial^2 \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} \partial \boldsymbol{\beta}^T_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right) \\
\times &\ (\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}) + o_p(n^{-1/2}).
\end{aligned}
$$

Suppose the model selection procedure satisfies the sure screening property. Then we have

$\mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}) = \boldsymbol{\beta}_0$, $\forall t$ and hence

$$\sum_{t=0}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right)$$

$$+ \underbrace{\sum_{t=0}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial \ell(\boldsymbol{U}_{t+1}, \boldsymbol{\beta}_0)}{\partial \beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial \ell(\boldsymbol{U}_{t+1}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right)}_{\zeta_1}$$

$$+ \underbrace{\sum_{t=0}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial^2 \ell(\boldsymbol{U}_{t+1}, \boldsymbol{\beta}_0)}{\partial \beta_{j_0} \partial \boldsymbol{\beta}^T_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial^2 \ell(\boldsymbol{U}_{t+1}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} \partial \boldsymbol{\beta}^T_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right) (\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}} - \boldsymbol{\beta}_{0,\widehat{\mathcal{M}}_{j_0}^{(t)}})}_{\zeta_2} + o_p(n^{-1/2}).$$

Using similar arguments in the proof of Theorem 3.1, we can show that $\zeta_2 = o_p(1)$ when $\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ and $\widehat{\boldsymbol{\omega}}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}$ satisfy certain uniform convergence rates, and

$$\zeta_1 = \sum_{t=s_n}^{n-1} \frac{1}{\sqrt{n}\sigma_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial \ell(\boldsymbol{U}_{t+1}, \boldsymbol{\beta}_0)}{\partial \beta_{j_0}} - \boldsymbol{\omega}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial \ell(\boldsymbol{U}_{t+1}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right) + o_p(1).$$

The first term on the RHS of the above expression is asymptotically normal under certain regularity conditions, according to the martingale central limit theorem. Thus, we obtain

$$\sum_{t=0}^{n-1} \frac{1}{\sqrt{n}\hat{\sigma}_{\widehat{\mathcal{M}}_{j_0}^{(t)},j_0}} \left(\frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \beta_{j_0}} - \widehat{\boldsymbol{\omega}}^T_{\widehat{\mathcal{M}}_{j_0}^{(t_0)},j_0} \frac{\partial \ell(\boldsymbol{U}_{t+1}, \mathrm{h}(\beta_{0,j_0}, \widehat{\mathcal{M}}_{j_0}^{(t)}, \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}))}{\partial \boldsymbol{\beta}_{\widehat{\mathcal{M}}_{j_0}^{(t)}}} \right) \xrightarrow{d} N(0,1),$$

by Slutsky's theorem.

References

Adamczak, R. a. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab. 13*, no. 34, 1000–1034.

Bühlmann, P. and J. Mandozzi (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Comput. Statist. 29*(3-4), 407–430.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Testing many moment inequalities. *arXiv preprint arXiv:1312.7614*.

de la Peña, V. H., T. L. Lai, and Q.-M. Shao (2009). *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin. Limit theory and statistical applications.

Klartag, B. and S. Mendelson (2005). Empirical processes and random projections. *J. Funct. Anal. 225*(1), 229–245.

Mendelson, S., A. Pajor, and N. Tomczak-Jaegermann (2008). Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx. 28*(3), 277–289.

Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist. 45*(1), 158–195.

Shi, C., A. Fan, R. Song, and W. Lu (2018). High-dimensional A-learning for optimal dynamic treatment regimes. *Ann. Statist. 46*(3), 925–957.

Talagrand, M. (2005). *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin. Upper and lower bounds of stochastic processes.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist. 42*(3), 1166–1202.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. New York: Springer.