Asymptotics of the adaptive elastic net estimation for conditional heteroscedastic time series models

Authors: YUANYUAN LIAO
– Department of Mathematics, Nanjing University, Nanjing, PR China

LIHONG WANG
– Department of Mathematics, Nanjing University, Nanjing, PR China (lhwang@nju.edu.cn)

Abstract:

- In this paper we propose an iteratively reweighted adaptive elastic net estimation method for conditional heteroscedastic time series models. The sign consistency and the asymptotic normality of the estimator are investigated. Compared with the Lasso method, the elastic net is more efficient for autoregressive time series models, because it benefits not only from the selection of the Lasso but also from the grouping effect inherited from the ridge penalty. The Monte Carlo simulation studies based on an AR-ARCH model are reported to assess the finite-sample performance of the proposed elastic net method.

Key-Words:

- *Adaptive elastic net; AR-ARCH models; asymptotic normality; iteratively reweighted algorithm; sign consistency.*

AMS Subject Classification:

- 62J07, 62M10.

1. INTRODUCTION

The Lasso introduced in [10] is a shrinkage and selection method for linear regression models. As variable selection is of increasing importance in big data analysis, the lasso is much more appealing owing to its sparse representation. However, the literature about the penalization techniques mainly deals with homoscedastic linear regression models, see, e.g., [2], [5], [9], [15], [16], and [18], among others. The investigation of the Lasso type estimator for heteroscedastic models started relatively late. Recently, [11] and [12] analysed the weighted lasso type estimators in a linear heteroscedastic regression model setting. [17] derived an iteratively reweighted adaptive lasso algorithm for time series models under conditional heteroscedasticity, and proved that the resulting estimator has sign consistency and asymptotic normality. The proposed method can be applied to various AR-ARCH type processes.

In this paper, we generalize the results of [17] to the adaptive elastic net method. That is, we consider the model similar to the one used by [17], but suggest the use of an iteratively reweighted adaptive elastic net algorithm. The elastic net introduced by [19] is a convex combination of the Lasso and ridge penalty. The ridge part of the penalty shrinks the estimated coefficients of all the variables and induces coefficients of correlated variables to be close to one another. The Lasso part of the penalty shrinks and selects the coefficients of the variables. As discussed in [4], the elastic net benefits from the selection of the Lasso, as well as from the finite-sample grouping effect inherited from the ridge penalty. This makes the elastic net particularly useful for estimating the autoregressive time series models, since this estimation procedure leaves out irrelevant variables but does not exclude correlated variables that may be relevant as part of a group.

In the next section, we introduce the iteratively reweighted adaptive elastic net algorithm for high-dimensional sparse linear regression models under conditional heteroscedasticity. The sign consistency and the asymptotic normality of the weighted adaptive elastic net estimators of the parameters are also addressed. Section 3 gives the Monte Carlo simulations based on a specific AR-ARCH model, evaluating and comparing the performance of the proposed adaptive elastic net algorithm and the adaptive Lasso method. The proof of the theorem is given in Appendix.

Throughout the paper, all limits are taken as $n \to \infty$, unless specified otherwise. The symbol C denotes an absolute positive constant whose value may vary at each occurrence. $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution, $\xrightarrow{\mathcal{P}}$ denotes convergence in probability, Z stands for a standard normal random variable. For any two real sequences $\{a_n\}$ and $\{b_n\}$, $a_n \sim b_n$ means that there are constants $c > 0$ and $C < \infty$ such that $c \leq a_n/b_n \leq C$ for all sufficiently large n.

2. THE ITERATIVELY REWEIGHTED ADAPTIVE ELASTIC NET ALGORITHM

We now introduce the model and the basic ideas of the algorithm. The model discussed here is similar to the one used by [14] and [17]. We consider a stationary random process $Y_t \in \mathbb{R}$ and a possibly infinite vector of covariates of stationary processes, $X_{t,\infty} = (X_{t,1}, X_{t,2}, \ldots)'$, $t \in \mathbb{Z}$, $\mathbb{Z} := \{0, \pm 1, \pm 2, \ldots\}$, obeying the model

$$Y_t = X'_{t,\infty}\beta^0_\infty + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{2.1}$$

where $\beta^0_\infty = (\beta^0_1, \beta^0_2, \ldots)'$ satisfying $\sum_{i=1}^\infty |\beta^0_i|^2 < +\infty$, ε_t is zero mean and independent of the covariates $X_{t,\infty}$, and

$$\varepsilon_t = \sigma_t Z_t, \;\; \sigma_t = g(\alpha^0_\infty; L_{\infty,t}), \;\; t \in \mathbb{Z},$$

where $Z_t, t \in \mathbb{Z}$, are i.i.d. standardized r.v.'s, g is a positive function, $L_{\infty,t} = (L_{1,t}, L_{2,t}, \ldots)$ is a possibly infinite vector of covariates of stationary processes $L_{i,t}$, $t \in \mathbb{Z}$, and $\alpha^0_\infty = (\alpha^0_1, \alpha^0_2, \ldots)'$ is a parameter vector. Here the covariates $X_{t,\infty}$ and $L_{\infty,t}$ can contain lagged versions of Y_t and $(\varepsilon_t, \sigma_t)$, respectively, which allows flexible modelling of autoregressive processes and a class of conditional variance models such as GARCH type models.

The observed data consists of $(\mathbf{X}_n, \mathbf{Y}_n)$, where

$$\mathbf{Y}_n = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \;\; \mathbf{X}_n = \begin{pmatrix} X_{1,1} & \cdots & X_{1,p_n} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,p_n} \end{pmatrix}, \;\; \boldsymbol{\beta}^0_n = \begin{pmatrix} \beta^0_1 \\ \vdots \\ \beta^0_{p_n} \end{pmatrix}, \;\; \boldsymbol{\varepsilon}^0_n = \mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta}^0_n,$$

where p_n is the number of possible parameters which increases with sample size n, $\boldsymbol{\beta}^0_n$ is the restriction of β^0_∞ to its first p_n coordinates, $\boldsymbol{\varepsilon}^0_n = (\varepsilon^0_1, ..., \varepsilon^0_n)'$.

The fact $\sum_{i=1}^\infty |\beta^0_i|^2 < +\infty$ implies that there is a positive sequence a_n decreasing to zero such that $\lim_{n\to\infty} P(\max_{1\le t\le n} |\varepsilon^0_t - \varepsilon_t| < a_n) \to 1$ holds. Thus, for a sufficiently large n we can approximately write

$$\varepsilon^0_t = \sigma_t Z_t, \;\; \sigma_t = g_n(\boldsymbol{\alpha}^0_n; \mathbf{L}^0_{n,t}), \;\; 1 \le t \le n, \tag{2.2}$$

here $\boldsymbol{\alpha}^0_n$ and $\mathbf{L}^0_{n,t}$ are the restrictions of α^0_∞ and $L_{\infty,t}$ to their first p_n coordinates, respectively, and g_n is the restriction of g that corresponds to $\boldsymbol{\alpha}^0_n$ and $\mathbf{L}^0_{n,t}$. Without loss of generality we assume that only q_n of the p_n parameters are non-zero. That is, $\boldsymbol{\beta}^0_n = (\beta^0_1, \ldots, \beta^0_{q_n}, 0, \ldots, 0)' = (\boldsymbol{\beta}^0_n(1)', \mathbf{0}')'$. In a similar manner, $\mathbf{X}_n = (\mathbf{X}_n(1), \mathbf{X}_n(2))$ and $\mathbf{X}_{t,n} = (\mathbf{X}_{t,n}(1)', \mathbf{X}_{t,n}(2)')'$, where $\mathbf{X}_{t,n}$ is the t-th row of $\mathbf{X}_n$.

We now introduce the adaptive elastic net algorithm based on an iteratively reweighted technique which is similar to the approaches in [7], [8], and [17]. Rewrite Model (2.1) as

$$\widetilde{Y}_t = \widetilde{\mathbf{X}}'_{t,n}\boldsymbol{\beta}^0_n + Z_t, \;\; 1 \le t \le n, \tag{2.3}$$

where $\widetilde{Y}_t = \frac{1}{\sigma_t} Y_t$, $\widetilde{\mathbf{X}}_{t,n} = \frac{1}{\sigma_t}\mathbf{X}_{t,n}$. It is obvious that the error Z_t is homoscedastic.

Since we have no a priori information about the conditional standard deviation σ_t, at first step we assume homoscedasticity. Then we use a weighted adaptive elastic net algorithm to estimate $\boldsymbol{\beta}_n^0$ in each iteration step. That is,

$$\begin{aligned} &\boldsymbol{\beta}_{n,elastic}(\lambda_n, \gamma_n, w_n) \\ (2.4) \qquad &= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta})' W_n^2 (\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta}) + \lambda_n \|\Sigma_1 \boldsymbol{\beta}\|_1 + \gamma_n \|\Sigma_2^{1/2}\boldsymbol{\beta}\|_2^2, \end{aligned}$$

where $\lambda_n \geq 0$, $\gamma_n \geq 0$, $\Sigma_1 = diag(v_n)$, $v_n = (v_{n,1}, \ldots, v_{n,p_n}) = |\boldsymbol{\beta}_{n,init1}|^{-\tau_1}$, $\Sigma_2 = diag(u_n)$, $u_n = (u_{n,1}, \ldots, u_{n,p_n}) = |\boldsymbol{\beta}_{n,init2}|^{-\tau_2}$, $\boldsymbol{\beta}_{n,init1}$ and $\boldsymbol{\beta}_{n,init2}$ are two initial estimators of $\boldsymbol{\beta}_n^0$ for some $\tau_1 \geq 0$ and $\tau_2 \geq 0$, and $W_n = diag(w_n)$, $w_n = (w_{n,1}, ..., w_{n,n}) = (\hat{\sigma}_{n,1}^{-1}, ..., \hat{\sigma}_{n,n}^{-1})$, $\hat{\sigma}_{n,t}$ is a suitable estimator of σ_t. Moreover, let $\hat{\boldsymbol{\alpha}}_n$ $(\boldsymbol{\beta}_{n,elastic}; \mathbf{X}_n, \mathbf{Y}_n)$ and $\hat{\mathbf{L}}_{n,t}(\boldsymbol{\beta}_{n,elastic}; \mathbf{X}_n, \mathbf{Y}_n)$ be the suitable known plug-in estimators for $\boldsymbol{\alpha}_n^0$ and $\mathbf{L}_{n,t}^0$, respectively. For relevant literature on estimation methods for the conditional variance part, see e.g. [7], [8], [17], and the references therein. For example, if the error process is an ARCH(p) model as in the simulation studies of Section 3, the usual maximum likelihood methods can be applied to estimate the unknown parameters of the conditional variance part based on the residuals from step 2 of the following algorithm.

The iteratively reweighted adaptive elastic net algorithm:

1. Let $k = 1$, $w_n^{[0]} = \mathbf{1}$. Determine the initial values of v_n,u_n, λ_n and γ_n.

2. Calculate the estimator $\boldsymbol{\beta}_n^{[k]} = \boldsymbol{\beta}_{n,elastic}(\lambda_n, \gamma_n, w_n^{[k-1]})$ of $\boldsymbol{\beta}_n^0$ for Model (2.3) using the weighted adaptive elastic net algorithm (2.4), compute the residuals $\boldsymbol{\varepsilon}_n^{[k]} = \mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta}_n^{[k]}$.

3. Estimate the conditional variances $\sigma_{n,t}^{[k]} = g_n(\boldsymbol{\alpha}_n^{[k]}; \mathbf{L}_{n,t}^{[k]})$, $1 \leq t \leq n$, where $\boldsymbol{\alpha}_n^{[k]} = \hat{\boldsymbol{\alpha}}_n(\boldsymbol{\beta}_n^{[k]}; \mathbf{X}_n, \mathbf{Y}_n)$, $\mathbf{L}_{n,t}^{[k]} = \hat{\mathbf{L}}_{n,t}(\boldsymbol{\beta}_n^{[k]}; \mathbf{X}_n, \mathbf{Y}_n)$ based on Model (2.2) and the residuals from step 2.

4. Calculate new weights $w_{n,t}^{[k]} = g_n(\boldsymbol{\alpha}_n^{[k]}; \mathbf{L}_{n,t}^{[k]})^{-1}$. Let $w_n^{[k]} = (w_{n,1}^{[k]}, ..., w_{n,n}^{[k]})$.

5. Let $k = k + 1$ and back to step 2 until a specified stopping criterion is satisfied. Return estimate $\boldsymbol{\beta}_n^{[k]}$.

As stated in [17], a plausible stopping criterion should measure the convergence of $\sigma_n^{[k]}$, where $\sigma_n^{[k]} = (\sigma_{n,1}^{[k]}, ..., \sigma_{n,n}^{[k]})'$. One can stop the algorithm if $\|\sigma_n^{[k]} - \sigma_n^{[k-1]}\|_2 < \zeta$ for some small $\zeta > 0$. It is suggested that, under certain conditions, $k = 2$ is sufficient to get an optimal estimator if n is large.

For the two initial estimators $\boldsymbol{\beta}_{n,init1}$ and $\boldsymbol{\beta}_{n,init2}$, as stated in [17], there are several options available. When $p_n < n$, one can simply choose the OLS estimator. Alternatively, one can select the lasso estimator as $\boldsymbol{\beta}_{n,init1}$, the ridge regression estimator as $\boldsymbol{\beta}_{n,init2}$, or set both $\boldsymbol{\beta}_{n,init1}$ and $\boldsymbol{\beta}_{n,init2}$ equal the elastic net estimator.

Next we show the sign consistency and asymptotic normality of the non-vanishing components of $\boldsymbol{\beta}_n^{[k]}$. Let $b_n = \min\{|\boldsymbol{\beta}_n^0(1)|\}$, $W_n^{[k]} = diag(w_n^{[k]})$, $\widetilde{\mathbf{X}}_n^{[k]} = W_n^{[k-1]}\mathbf{X}_n$, $\widetilde{\mathbf{Y}}_n^{[k]} = W_n^{[k-1]}\mathbf{Y}_n$, $\widetilde{\Gamma}_n^{[k]} = \frac{1}{n}(\widetilde{\mathbf{X}}_n^{[k]})'\widetilde{\mathbf{X}}_n^{[k]}$, $\Gamma_n = \widetilde{\Gamma}_n^{[1]} = \frac{1}{n}\mathbf{X}_n'\mathbf{X}_n$. Let W_n^0 and $\widetilde{\Gamma}_n^0$ be the true matrices, and the submatrices to $\boldsymbol{\beta}_n^0(1)$ are denoted as $\widetilde{\Gamma}_n^{[k]}(1)$, $\widetilde{\Gamma}_n^0(1)$, $\Gamma_n(1)$, $\Sigma_1(1)$ and $\Sigma_2(1)$. Similarly to [12] and [17], we require the following assumptions.

Assumption (A):

(A1) $\{Y_t, X_{t,1}, \ldots, X_{t,m}, \sigma_t\}_{t\in\mathbb{Z}}$ is weakly stationary for all $m \geq 1$, $\{Z_t\}_{t\in\mathbb{Z}}$ is an i.i.d. standardized random sequence and $E(Z_t^4) < \infty$, Z_t is independent of $X_{t,\infty}$ for any $t \in \mathbb{Z}$, and $E(\sigma_t^4) < \infty$.

(A2) $E(X_{t,i}^2) = 1$ for any $i \geq 1$ and $t \in \mathbb{Z}$.

(A3) There is a positive sequence $\{\upsilon_n\}$ such that $\max_{1\leq t\leq n} \|\mathbf{X}_{t,n}(1)\|_2 = \mathcal{O}_p(\upsilon_n\sqrt{q_n})$.

(A4) There are constants $a_1 > 0$ and $a_2 > 0$ such that

$$\lim_{n\to\infty} P(a_1 \min\{|\boldsymbol{\beta}_{n,init1}(1)|^{\tau_1}\} < b_n) = 0, \quad \lim_{n\to\infty} P(a_2 \min\{|\boldsymbol{\beta}_{n,init2}(1)|^{\tau_2}\} < b_n) = 0.$$

(A5) There exists a positive sequence $\{r_n\}$ with $r_n \to \infty$ such that

$$\lim_{n\to\infty} P(\max\{|\beta_{n,init1}(2)|^{\tau_1}\} \geq {r_n}^{-1}) = 0.$$

(A6) There are positive constants $\lambda_{0,min} < \lambda_{0,max}$ and $\lambda_{1,min}$ such that the eigenvalues satisfy

$$\lim_{n\to\infty} P(\lambda_{0,min} < \lambda_{min}(\Gamma_n(1)) \leq \lambda_{max}(\Gamma_n(1)) < \lambda_{0,max}) = 1,$$

and

$$\lim_{n\to\infty} P(\lambda_{1,min} < \lambda_{min}(\widetilde{\Gamma}_n^0(1)) \leq \lambda_{max}(\widetilde{\Gamma}_n^0(1))) = 1.$$

(A7) There are constants $0 < \lambda_{2,min}$ and $\lambda_{3,min} > 0$ such that the eigenvalues satisfy

$$\lim_{n\to\infty} P(\lambda_{2,min} < \lambda_{min}(D_n) \leq \lambda_{max}(D_n)) = 1,$$

and

$$\lim_{n\to\infty} P(\lambda_{3,min} < \lambda_{min}(E_n) \leq \lambda_{max}(E_n)) = 1,$$

where

$$D_n = \left(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\right)^{-1}\widetilde{\Gamma}_n^0(1)\left(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\right)^{-1},$$

$$E_n = \Big(\Gamma_n(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}\Gamma_n(1)\Big(\Gamma_n(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}.$$

(A8) There is a positive constant σ_{min} such that

$$0 < \sigma_{min} < g_n(\hat{\boldsymbol{\alpha}}_n(\boldsymbol{\beta}_n; \mathbf{X}_n, \mathbf{Y}_n), \hat{\mathbf{L}}_{n,t}(\boldsymbol{\beta}_n; \mathbf{X}_n, \mathbf{Y}_n)), 1 \le t \le n$$

for all large enough n and $\boldsymbol{\beta}_n$ in an open neighbourhood of $\boldsymbol{\beta}_n^0$.

(A9) For all n and any $1 \le t \le n$, the estimators $\hat{\boldsymbol{\alpha}}_n$ and $\hat{\mathbf{L}}_{n,t}$ are consistent for $\boldsymbol{\alpha}_n^0$ and $\mathbf{L}_{n,t}^0$, and there is a sequence $\{h_n\}$ with $h_n n^{-1/2} \to 0$ such that

$$\max_{1\le t\le n} |g(\boldsymbol{\alpha}_\infty^0; \mathbf{L}_{\infty,t})^{-2} - g_n(\hat{\boldsymbol{\alpha}}_n(\boldsymbol{\beta}_n^0; \mathbf{X}_n, \mathbf{Y}_n), \hat{\mathbf{L}}_{n,t}(\boldsymbol{\beta}_n^0; \mathbf{X}_n, \mathbf{Y}_n))^{-2}| = \mathcal{O}_p(h_n/\sqrt{n}).$$

(A10) There are positive constants C_1, C_2 and d with $1 \le d \le 2$ such that, for any $t \in Z$,

$$P(|\varepsilon_t| > x) \le C_1 \exp(-C_2 x^d).$$

(A11)

① $\frac{(\log n)^{I\{d=1\}}(\log(1+q_n))^{1/d}}{\sqrt{n}b_n} \to 0,$ ② $\frac{h_n}{\sqrt{n}b_n} \to 0,$

③ $\frac{\lambda_n\sqrt{q_n}}{\sqrt{n}b_n} \to 0,$ ④ $\frac{\sqrt{n}(\log n)^{I\{d=1\}}(\log(1+p_n-q_n))^{1/d}}{\lambda_n r_n} \to 0,$

⑤ $\frac{h_n\sqrt{n}}{\lambda_n r_n} \to 0,$ ⑥ $\frac{\sqrt{q_n}}{b_n r_n} \to 0,$

⑦ $\frac{\upsilon_n\sqrt{q_n}}{\sqrt{n}} \to 0,$ ⑧ $\frac{h_n\sqrt{q_n}}{\sqrt{n}} \to 0,$

⑨ $\frac{\gamma_n\left((\log n)^{I\{d=1\}}+h_n\right)}{\sqrt{n}b_n} \to 0.$

Similar assumptions as in (A1)-(A11) are also imposed in [17] to study the asymptotic behaviour of the iteratively reweighted adaptive lasso algorithm. Assumption (A1) is standard for variable selection in a time series setting. Assumption (A2) is the usual scale standardization required in a lasso setting without loss of generality (see e.g. [6]), because $\{X_{t,i}\}$ is stationary and hence its mean and variance are constants. Assumption (A3) characterises the structure of regressors. For instance, if $\{X_{t,n}(1)\}$ is stationary and $\boldsymbol{\beta}_n^0$ contains a finite number of non-zero components, then we can choose $\upsilon_n = O_P(1)$ for Assumption (A3) to hold. Assumptions (A4) and (A5) actually assume that the weights υ_n and u_n are not too large for $\beta_j^0 \neq 0$ and not too small for $\beta_j^0 = 0$. They also mean that the initial estimators can distinguish between zero and non-zero components of the parameter vector well. For the Lasso initial estimators, Assumptions

(A4) and (A5) can be derived from sharp thresholds and sign consistency of the Lasso estimate under some additional mild assumptions (see, e.g., [13] and [16]). Assumption (A6) is needed to address heteroscedasticity in high-dimensional regression models (see, for example, [3]). Since we deal with the weighted adaptive elastic net algorithm, additional similar assumptions such as (A7) are also needed here. It is worth mentioning that, under certain conditions, $D_n - \widetilde{\Gamma}_n^0(1) \to 0$ and $E_n - \Gamma_n(1) \to 0$ as $n \to \infty$. Assumptions (A8) and (A9) are standard in heteroscedastic regression and Assumption (A10) excludes heavy-tailed errors.

Assumption (A11) postulate properties required for deriving the asymptotics of the proposed estimator. As a simple example, to better understand Assumption (A11) assume b_n to be fixed and $d = 1$, ① and ② permit $h_n \sim 1$ and $q_n \sim n^{1/2+\delta}$ for any $0 < \delta < 1/4$. With these choices we can choose $\lambda_n \sim n^{1/4-\delta}$, $r_n \sim n^{1/2+\delta}$, $\upsilon_n \sim 1$ and $\gamma_n \sim n^{1/4-\delta}$ by ③, ④, ⑦ and ⑨, and p_n can grow with every polynomial order. Obviously these selections satisfy Assumption (A11), and also Assumptions (A3)-(A5) and (A9). Moreover, by ④ and ⑨, we obtain $\frac{\gamma_n}{b_n \lambda_n r_n} \to 0$ as $n \to \infty$.

The following theorem shows the sign consistency and asymptotic normality of the estimator. The proof will be given in the Appendix. The sign consistency introduced by [16] is stronger than the usual selection consistency which only requires the zeros to be matched, but not the signs. The reason for using sign consistency is to avoid dealing with situations where a model is estimated with matching zeros but reversed signs.

Theorem 2.1. *Under Assumption (A), it holds for all $k \geq 1$ that*

(1) (sign consistency)

$$\lim_{n\to\infty} P\big(sign(\boldsymbol{\beta}_n^{[k]}) = sign(\boldsymbol{\beta}_n^0)\big) = 1,$$

where $sign(\cdot)$ maps positive entry to 1, negative entry to -1 and zero to zero, that is, $\boldsymbol{\beta}_n^{[k]}$ asymptotically matches the zeros and signs of $\boldsymbol{\beta}_n^0$ with probability one.

(2) (asymptotic normality)

$$\sqrt{n}(s_n(k))^{-1}\boldsymbol{\xi}_n'(\boldsymbol{\beta}_n^{[k]}(1) - \boldsymbol{\beta}_n^0(1)) \xrightarrow{\mathcal{D}} Z,$$

where $\boldsymbol{\xi}_n \in \mathbb{R}^{q_n}$ with $\|\boldsymbol{\xi}_n\|_2 = 1$, $s_n^2(1) = \boldsymbol{\xi}_n' E_n \boldsymbol{\xi}_n$ and $s_n^2(k) = \boldsymbol{\xi}_n' D_n \boldsymbol{\xi}_n$ for $k \geq 2$.

3. SIMULATION STUDIES

In this section, we provide simulation studies to check the finite sample performance of the iteratively reweighted adaptive elastic net algorithm (IRAEN) for an AR-ARCH model. The comparison with the iteratively reweighted adaptive Lasso algorithm (IRAL) introduced in [17] is also considered.

We consider the following AR-ARCH model

$$Y_t = \sum_{i \in I} \phi_i Y_{t-i} + \varepsilon_t,$$

and

$$\varepsilon_t = \sigma_t Z_t, \ \ \sigma_t = \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2},$$

where the true values of the parameters are $\alpha_0 = 0.02$ and $\alpha_1 = \alpha_2 = 0.49$, $Z_t \sim N(0,1)$, $\phi_i = 0.95(\phi^{-1} - 1)\phi^{\sqrt{i}}$, $\phi = 0.85$, $I = \{1, 4, 9, 16, \ldots\}$. It is easy to see that $\sum_{i \in I} |\phi_i| = 0.95 < 1$ and $\sum_{i \in I} \phi_i^2 < \infty$ which imply the stationarity of Y_t. Note that, by the properties of the AR-ARCH model, $EY_t^2 = E\sigma_t^2 = \alpha_0/(1 - \alpha_1 - \alpha_2) = 1$. This implies that Assumption (A2) is satisfied.

Let $p_n = [2\sqrt{n}]$ and $q_n = [\sqrt{p_n}]$, where n is the sample size. For example, when $n = 500$, $p_n = 44$, $q_n = 6$ and $I = \{1, 4, 9, 16, 25, 36\}$. If $n = 1000$, then $p_n = 63$, $q_n = 7$ and $I = \{1, 4, 9, 16, 25, 36, 49\}$.

After generating data from the above AR-ARCH model with sample size $n = 500$ and $n = 1000$, respectively, we use two methods, IRAEN and IRAL, to estimate the parameters ϕ_i and to check the sign consistency of the estimators. In the simulations, we use the C_p criterion to choose the appropriate λ_n and γ_n. The two initial estimators $\boldsymbol{\beta}_{n,init1}$ and $\boldsymbol{\beta}_{n,init2}$ are chosen to be the OLS estimator.

3.1. The iteratively reweighted adaptive elastic net algorithm

To apply the proposed iteratively reweighted elastic net algorithm, we consider two cases: the homoscedastic case ($k = 1$) and the heteroscedastic case with one additional replication ($k = 2$).

For the $k = 1$ case, Table 1 reports the estimation results for two sample sizes $n = 500$ and $n = 1000$ based on 1000 replications. We hope that the covariates with nonzero coefficients (relevant parameters) can be selected from the estimation procedure, but the covariates with zero coefficients (irrelevant parameters) shouldn't be included. Table 1 shows the proportions of both the relevant and irrelevant included parameters of all estimated parameters for the homoscedastic case. Proportion 1 (the accuracy rate) denotes the proportion of the relevant included parameters and Proportion 2 (the error rate) is the proportion of the irrelevant included parameters. The number of times each parameter has been selected during 1000 simulations are also reported.

It is seen from Table 1 that the accuracy rate increases with larger sample size n, while the error rate decreases in n. This is consistent with the theoretical results in Theorem 1.

In a similar way, we apply the proposed iteratively reweighted elastic net algorithm with $k = 2$. Proportions of both the relevant and irrelevant included parameters

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
n=500	866	691	622	834	568	506
n=1000	932	721	668	902	575	542
	ϕ_7	ϕ_8	ϕ_9	...	Proportion 1	Proportion 2
n=500	454	482	864	...	81.25%	33.61%
n=1000	502	502	927	...	89.41%	29.66%

Table 1: Proportions of relevant included parameters and irrelevant included parameters for the case of $k = 1$ using IRAEN.

of all estimated parameters and the number of times each parameter has been selected during 1000 simulations for the heteroscedastic case are given in Table 2. Inspection of Table 2 reveals that, as in the $k = 1$ case, the accuracy rate increases with larger sample size n, while the error rate decreases in n. Comparing two tables, we conclude that the heteroscedastic case with $k = 2$ has better selection properties than the homoscedastic case $k = 1$ for the conditional heteroscedastic models.

Moreover, the plots in Figure 1 show the selection results for both the $k = 1$ and $k = 2$ cases from one simulation with $n = 500$. For each plot, the vertical axis represents the values of the estimated coefficients, the horizontal axis (bottom) represents the values of $\ln \lambda_n$, and the top shows the numbers of the non-zero coefficients selected for different values of $\ln \lambda_n$. The 44 curves illustrate the change of the values of 44 estimated coefficients with $\ln \lambda_n$ changing. Note that there are only six non-zero positive coefficients in the true model. It can be seen that, when $k = 2$, these six coefficients tend to zero from positive side, while when $k = 1$, there exist some coefficients tending to zero from negative side, which means that no matter what value $\ln \lambda_n$ takes, the sign consistency may not be satisfied. This is consistent with the conclusions drawn from the comparison of Tables 1 and 2. Figure 1 again visually displays that the heteroscedastic algorithm with $k = 2$ outperforms its homoscedastic counterpart.

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
n =500	938	415	395	969	383	363
n =1000	992	407	371	999	377	287
	ϕ_7	ϕ_8	ϕ_9	...	Proportion 1	Proportion 2
n=500	322	345	979	...	94.92%	29.54%
n=1000	316	294	999	...	98.77%	21.98%

Table 2: Proportions of relevant included parameters and irrelevant included parameters for the case of $k = 2$ using IRAEN

3.2. The iteratively reweighted adaptive Lasso algorithm

Next we report the estimation results using the iteratively reweighted adaptive Lasso algorithm. Proportions of both the relevant and irrelevant included parameters and the number of times each parameter has been selected during 1000 simulations for

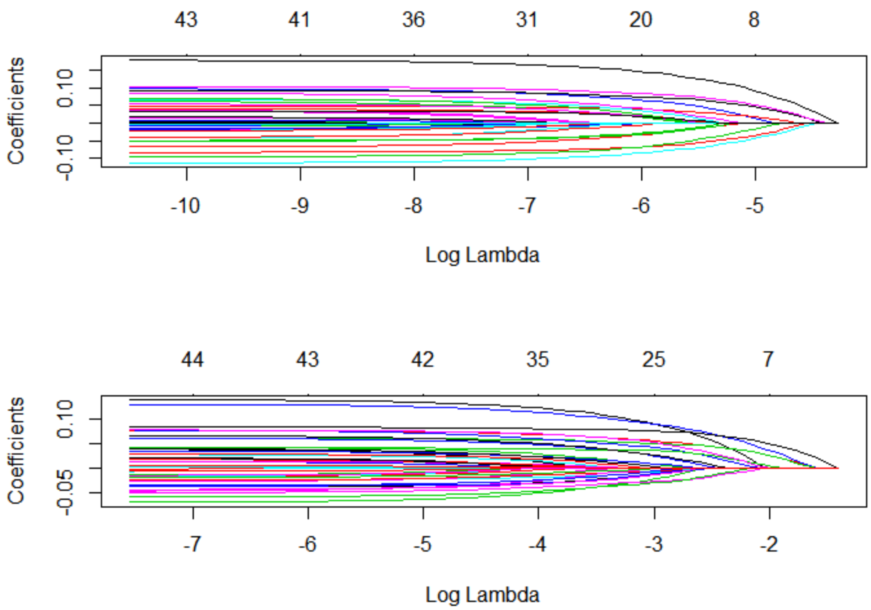

Figure 1: Estimated coefficients for different λ_n values with $n = 500$ and $k = 1$ (upper) or $k = 2$ (lower) using IRAEN

the homoscedastic case and the heteroscedastic case with $n = 500$ and $n = 1000$ are given in Tables 3 and 4, respectively. Figure 2 shows the selection results for both the $k = 1$ and $k = 2$ cases from one simulation with $n = 500$. Similarly to the IRAEN algorithm, Tables 3-4 and Figure 2 indicate that the heteroscedastic algorithm with $k = 2$ outperforms its homoscedastic counterpart.

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
n=500	871	682	619	863	515	501
n=1000	927	710	612	893	555	519
	ϕ_7	ϕ_8	ϕ_9	...	Proportion 1	Proportion 2
n=500	481	452	827	...	80.20%	31.95%
n=1000	506	471	931	...	88.37%	27.80%

Table 3: Proportions of relevant included parameters and irrelevant included parameters for the case of $k = 1$ using IRAL

Comparing Tables 1 and 2 with Tables 3 and 4, it is clear that the IRAEN algorithm proposed in this paper uniformly improves the accuracy rate as compared to the IRAL method, while the error rate is increased as a price to pay for using IRAEN algorithm. This implies that the IRAL method excludes irrelevant variables more thoroughly. It is also consistent with the conclusions of [19]. That is, if the covariates have grouping effect (a group of variables among which the pairwise correlations are very high), then the IRAL algorithm tends to arbitrarily select only one variable from the group, while the IRAEN algorithm has the capacity of selecting groups of correlated variables. Generally

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
n=500	930	399	358	971	338	296
n=1000	995	349	325	997	305	277
	ϕ_7	ϕ_8	ϕ_9	...	Proportion 1	Proportion 2
n=500	327	313	974	...	93.40%	26.35%
n=1000	253	252	1000	...	98.66%	19.98%

Table 4: Proportions of relevant included parameters and irrelevant included parameters for the case of $k = 2$ using IRAL

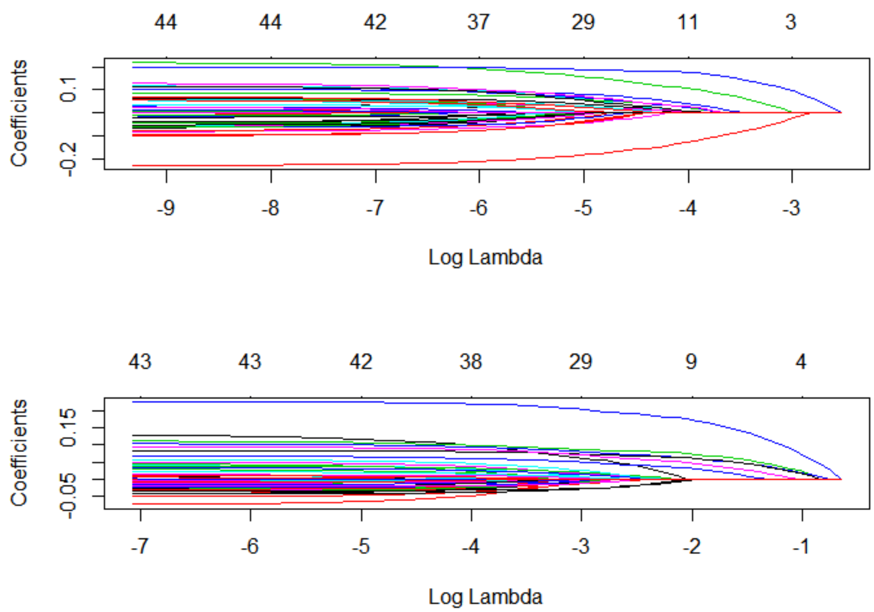

Figure 2: Estimated coefficients for different λ_n values with $n = 500$ and $k = 1$ (upper) or $k = 2$ (lower) using IRAL

speaking, the IRAEN algorithm produces a sparse model with good estimation accuracy, while encouraging a grouping effect. This makes the IRAEN algorithm particularly useful for estimating the models containing several correlated variables such as the AR-ARCH type processes.

ACKNOWLEDGMENTS

This work has been supported by National Natural Science Foundation of China (NSFC) Grants 11671194 and 11501287.

REFERENCES

[1] BROWN, B.M. (1971). Martingale central limit theorems, *The Annals of Mathematical Statistics,* **42**, 59–66.

[2] CHEN, K. and CHAN, K. (2011). Subset ARMA selection via the adaptive Lasso, *Statistics and Its Interface*, **4**, 197–205.

[3] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annls of statististics*, **32**, 928–961.

[4] FURMAN, Y. (2014). VAR estimation with the adaptive elastic net, Available at SSRN: https://ssrn.com/abstract=2456510 or http://dx.doi.org/10.2139/ssrn.2456510

[5] HSU, N., HUNG, H. and CHANG, Y. (2008). Subset selection for vector autoregressive processes using Lasso, *Computational Statistics and Data Analysis*, **52**, 3645–3657.

[6] HUANG, J., MA, S. and ZHANG, C.H. (2008). Adaptive lasso for sparse high-dimensional regression, *Statistica Sinica,* **18**, 1603–1618.

[7] LING, S. (2007). Self-weighted and local quasi-maximum likelihood estimators for arma-garch/igarch models, *Journal of Econometrics*, **140**, 849–873.

[8] MAK, T., WONG, H. and LI, W. (1997). Estimation of nonlinear time series with conditional heteroskedastic variance by iteratively weighted least squares, *Computational Statistics and Data Analysis*, **24**, 169–178.

[9] SCHOMAKER, M. (2012). Shrinkage averaging estimation, *Statistical Papers*, **53**, 1015–1034.

[10] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, **58**, 267–288.

[11] WAGENER, J. and DETTE, H. (2012). Bridge estimators and the adaptive lasso under heteroscedasticity, *Mathematical Methods of Statistics*, **21**, 109–126.

[12] WAGENER, J. and DETTE, H. (2013). The adaptive lasso in high-dimensional sparse heteroscedastic models, *Mathematical Methods of Statistics*, **22**, 137–154.

[13] WAINWRIGHT, M.J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (Lasso), *IEEE Transactions on Information Theory*, **55**, 2183–2202.

[14] YOON, Y.J., PARK, C. and LEE, T. (2013). Penalized regression models with autoregressive error terms, *Journal of Statistical Computation and Simulation*, **83**, 1756–1772.

[15] YÜZBAŞI, B., AHMED, S.E. and GÜNGÖR, M. (2017). Improved penalty strategies in linear regression models, *REVSTAT - Statistical Journal*, **15**, 2, 251–276.

[16] ZHAO, P. and YU, B. (2006). On model selection consistency of lasso, *The Journal of Machine Learning Research*, **7**, 2541–2563.

[17] ZIEL, F. (2016). Iteratively reweighted adaptive lasso for conditional heteroscedastic time series with applications to AR-ARCH type processes. *Computational Statistics and Data Analysis*, **100**, 773–793.

[18] ZOU, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

[19] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **67**, 301–320.

APPENDIX

Proof of Theorem 2.1: The basic ideas of the proof are mainly from [12] and [17]. Since we are dealing with the elastic net algorithm, we need some extra steps to achieve our goal.

Let $\|X\|_{\psi_d} = \inf\left\{C > 0 \middle| E[\psi_d(|X|/C)] \leq 1\right\}$ be the Orlicz norm of a random variable X, where $\psi_d(x) = \exp(x^d) - 1$, $1 \leq d \leq 2$. Denote $e_{n,j}$ the jth unit vector in $\mathbb{R}^{q_n}$. For any vector a and b, $a =_s b$ means that $sign(a) = sign(b)$. Let $k \geq 2$, the case $k = 1$ can be proved in a similar way.

(I) The sign consistency

The Karush-Kuhn-Tucker (KKT) conditions yield that $(\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta})'(W_n^{[k-1]})^2 (\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta}) + \lambda_n\|\Sigma_1\boldsymbol{\beta}\|_1 + \gamma_n\|\Sigma_2^{1/2}\boldsymbol{\beta}\|_2^2$ is minimised by $\boldsymbol{\beta} = (\boldsymbol{\beta}(1)', \mathbf{0}')'$ if and only if

$$\mathbf{X}_j^{0\prime}(W_n^{[k-1]})^2(\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta}) - \gamma_n u_{n,j}\beta_j = \frac{\lambda_n}{2}v_{n,j}sign(\beta_j), \text{ if } \beta_j \neq 0, \tag{3.1}$$

$$|\mathbf{X}_j^{0\prime}(W_n^{[k-1]})^2(\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta})| < \frac{\lambda_n}{2}v_{n,j}, \text{ if } \beta_j = 0, \tag{3.2}$$

where $\mathbf{X}_j^0$ is the j-th column of $\mathbf{X}_n$. Let

$$\begin{aligned}\delta_n^{[k]}(1) \;=\; &\boldsymbol{\beta}_n^0(1) + \frac{1}{n}\Big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}\mathbf{X}_n(1)'(W_n^{[k-1]})^2\boldsymbol{\varepsilon}_n^0 \\ &-\frac{\lambda_n}{2n}\Big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}s_n^0(1),\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\beta}_n^{[k]}(1) \;=\; &\Big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}\widetilde{\Gamma}_n^{[k]}(1)\boldsymbol{\beta}_n^0(1) \\ &+\frac{1}{n}\Big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}\mathbf{X}_n(1)'(W_n^{[k-1]})^2\boldsymbol{\varepsilon}_n^0 \\ &-\frac{\lambda_n}{2n}\Big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}s_n^0(1),\end{aligned} \tag{3.3}$$

where $s_n^0(1) = \Sigma_1(1)sign(\boldsymbol{\beta}_n^0(1))$. In addition, let $\delta_n^{[k]} = (\delta_n^{[k]}(1)', \mathbf{0}')'$ and $\boldsymbol{\beta}_n^{[k]} = (\boldsymbol{\beta}_n^{[k]}(1)', \mathbf{0}')'$.

First we show

$$\lim_{n\to\infty} P(\boldsymbol{\beta}_n^0 \neq_s \delta_n^{[k]}) = 0. \tag{3.4}$$

Let $\eta_{1,j} = e_{n,j}'\big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\mathbf{X}_n(1)'(W_n^{[k-1]})^2\boldsymbol{\varepsilon}_n^0$, $\eta_{2,j} = e_{n,j}'\big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}s_n^0(1)$, and let $A_1 = \{\frac{1}{n}|\eta_{1,j}| \geq \frac{1}{2}|\beta_j^0|, \text{ for some } j \leq q_n\}$ and $A_2 =$

$\{\frac{\lambda_n}{n}|\eta_{2,j}| \geq |\beta_j^0|, \text{ for some } j \leq q_n\}$. Thus, to prove (3.4), it is enough to show that $P(A_j) \to 0$ as $n \to \infty$ for $j = 1, 2$.

For $P(A_1)$, we obtain

$$\begin{aligned} P(A_1) &\leq P\Big(\frac{1}{n}\max_{1\leq j\leq q_n}|\eta_{1,j}| \geq \frac{b_n}{2}\Big) \\ &\leq P\Big(\frac{1}{n}\max_{1\leq j\leq q_n}|\eta_{1,j}^{0,\infty}| \geq \frac{b_n}{4}\Big) + P\Big(\frac{1}{n}\max_{1\leq j\leq q_n}|\eta_{1,j} - \eta_{1,j}^0| \geq \frac{b_n}{8}\Big) \\ &\quad + P\Big(\frac{1}{n}\max_{1\leq j\leq q_n}|\eta_{1,j}^0 - \eta_{1,j}^{0,\infty}| \geq \frac{b_n}{8}\Big), \end{aligned} \tag{3.5}$$

where $\eta_{1,j}^0 = e_{n,j}'(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^0)^2\boldsymbol{\varepsilon}_n^0$ and $\eta_{1,j}^{0,\infty} = e_{n,j}'(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\ \mathbf{X}_n(1)'\ (W_n^0)^2\boldsymbol{\varepsilon}_{n,\infty}^0$, and $\boldsymbol{\varepsilon}_{n,\infty}^0$ is the restriction of the true error $\boldsymbol{\varepsilon}_\infty^0 = (\varepsilon_1^0, \varepsilon_2^0, \ldots)'$ in Model (2.1).

Regarding the first term of (3.5), Assumptions (A6), (A8) and (A9) imply that $\|W_n^0\|_2 \leq \sigma_{min}^{-1}$ and $\|\Gamma_n(1)\|_2 \leq \lambda_{0,max}$. Note that$\lambda_{1,min} < \lambda(\widetilde{\Gamma}_n^0(1))$ and $0 \leq \lambda(\frac{\gamma_n}{n}\Sigma_2(1))$, then $\lambda_{1,min} < \lambda(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))$. That is, $\lambda((\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}) \leq \lambda_{1,min}^{-1}$ and hence $\|(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\| \leq \lambda_{1,min}^{-1}$. Thus we arrive at

$$\begin{aligned} &\Big\|\frac{1}{\sqrt{n}}e_{n,j}'\big(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\mathbf{X}_n(1)'(W_n^0)^2\Big\|_2 \\ &\leq \Big\|\big(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\Big\|_2\Big\|\frac{1}{\sqrt{n}}\mathbf{X}_n(1)\Big\|_2\|(W_n^0)^2\|_2 \\ &\leq \lambda_{1,min}^{-1}\sqrt{\lambda_{0,max}}\sigma_{min}^{-2}. \end{aligned} \tag{3.6}$$

This implies that, as $n \to \infty$,

$$P\Big(\Big\|\frac{1}{\sqrt{n}}e_{n,j}'\big(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}X_n(1)'(W_n^0)^2\Big\|_2 \leq \lambda_{1,min}^{-1}\sqrt{\lambda_{0,max}}\sigma_{min}^{-2}\Big) \to 1.$$

This, together with Lemma 1(i) of [6] and Assumption (A10), yields that

$$\begin{aligned} \Big\|\frac{1}{\sqrt{n}}\eta_{1,j}^{0,\infty}\Big\|_{\psi_d} &= \Big\|\frac{1}{\sqrt{n}}e_{n,j}'(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}X_n(1)'(W_n^0)^2\varepsilon_{n,\infty}^0\Big\|_{\psi_d} \\ &\leq C(\log n)^{I\{d=1\}}. \end{aligned} \tag{3.7}$$

Combining this with Equation (16) of [17], we obtain

$$P\Big(\frac{1}{n}\max_{1\leq j\leq q_n}|\eta_{1,j}^{0,\infty}| \geq \frac{b_n}{4}\Big) \leq \psi_d^{-1}\Big(\frac{b_n\sqrt{n}}{4C(\log(1+q_n))^{1/d}(\log n)^{I\{d=1\}}}\Big). \tag{3.8}$$

Now it follows from Assumption (A11) that

$$P\Big(\frac{1}{n}\max_{1\leq j\leq q_n}|\eta_{1,j}^{0,\infty}| \geq \frac{b_n}{4}\Big) \to 0.$$

For the second term of (3.5), Assumptions (A8) and (A9) ensure that $\|W_n^{[k-1]}\|_2 = \mathcal{O}_p(1)$ and $\|(W_n^0)^2 - (W_n^{[k-1]})^2\|_2 = \mathcal{O}_p(\frac{h_n}{\sqrt{n}})$. Furthermore, we notice that $\|\boldsymbol{\varepsilon}_n^0\|_2 \leq$

$\|\boldsymbol{\varepsilon}_n^0 - \boldsymbol{\varepsilon}_{n,\infty}^0\|_2 + \|\boldsymbol{\varepsilon}_{n,\infty}^0\|_2$, while $\|\boldsymbol{\varepsilon}_n^0 - \boldsymbol{\varepsilon}_{n,\infty}^0\|_2 \xrightarrow{\mathcal{P}} 0$, and Assumption (A1) and the weak law of large numbers yield that $\|\boldsymbol{\varepsilon}_{n,\infty}^0\|_2 = \mathcal{O}_p(\sqrt{n})$. This bound implies that $\|\boldsymbol{\varepsilon}_n^0\|_2 = \mathcal{O}_p(\sqrt{n})$.

On the other hand, since

$$\begin{aligned}
&\Big\|(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)) - (\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1))\Big\|_2 = \|\widetilde{\Gamma}_n^0(1) - \widetilde{\Gamma}_n^{[k]}(1)\|_2 \\
&= \|\Gamma_n(1)\|_2\|(W_n^0)^2 - (W_n^{[k-1]})^2\|_2 = \mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big),
\end{aligned}$$

we obtain

$$\begin{aligned}
\|A^{-1} - (A+B)^{-1}\|_2 \ &\leq \ \|A^{-1} - (A+B)^{-1} + A^{-1}BA^{-1}\|_2 + \|A^{-1}BA^{-1}\|_2 \\
&\leq \ O_p(\|B\|_2) + \|A^{-1}\|_2^2\|B\|_2 = \mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big),
\end{aligned}$$

where $A = \widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)$ and $B = (\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)) - (\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))$. That is,

$$\Big\|\Big(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1} - \Big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}\Big\| = \mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big). \tag{3.9}$$

We conclude that, for all $1 \leq j \leq q_n$,

$$\begin{aligned}
&|\eta_{1,j} - \eta_{1,j}^0| \\
&= \Big|e_{n,j}'\Big(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}\mathbf{X}_n(1)'((W_n^0)^2 - (W_n^{[k-1]})^2)\boldsymbol{\varepsilon}_n^0 \\
&\quad + e_{n,j}'\Big((\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)) - (\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1))\Big)\mathbf{X}_n(1)'(W_n^{[k-1]})^2\boldsymbol{\varepsilon}_n^0\Big| \\
&\leq \|n\Gamma_n(1)\|_2^{1/2}\|\boldsymbol{\varepsilon}_n^0\|_2\Big\{\|(W_n^0)^2 - (W_n^{[k-1]})^2\|_2\Big\|(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\Big\|_2 \\
&\quad + \|(W_n^{[k-1]})^2\|_2\Big\|(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1} - (\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\Big\|_2\Big\} \\
&= \mathcal{O}_p(\sqrt{n})\mathcal{O}_p(\sqrt{n})\mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big)\mathcal{O}_p(1) \\
&= \mathcal{O}_p(h_n\sqrt{n}).
\end{aligned}$$

Thus it follows from Assumption (A11) that $P(\frac{1}{n}\max_{1\leq j\leq q_n}|\eta_{1,j} - \eta_{1,j}^0| \geq \frac{b_n}{8}) \leq P(\frac{h_n}{\sqrt{n}b_n} \geq C) \to 0$ as $n \to \infty$.

We proceed to deal with the third term of (3.5). By (3.6),

$$\begin{aligned}
\frac{1}{\sqrt{n}}|\eta_{1,j}^0 - \eta_{1,j}^{0,\infty}| \ &= \ \frac{1}{\sqrt{n}}\Big|e_{n,j}'\big(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\mathbf{X}_n(1)'(W_n^0)^2(\boldsymbol{\varepsilon}_n^0 - \boldsymbol{\varepsilon}_{n,\infty}^0)\Big| \\
&\leq \ \lambda_{1,min}^{-1}\sqrt{\lambda_{0,max}}\,\sigma_{min}^{-2}\|\boldsymbol{\varepsilon}_n^0 - \boldsymbol{\varepsilon}_{n,\infty}^0\|_2 \xrightarrow{\mathcal{P}} 0.
\end{aligned}$$

Hence, by Assumption (A11), we have

$$P\Big(\frac{1}{n}\max_{1\leq j\leq q_n}|\eta_{1,j}^0 - \eta_{1,j}^{0,\infty}| \geq \frac{b_n}{8}\Big) \leq P\Big(\frac{1}{\sqrt{n}b_n}\max_{1\leq j\leq q_n}\frac{1}{\sqrt{n}}|\eta_{1,j}^0 - \eta_{1,j}^{0,\infty}| \geq \frac{1}{8}\Big) \to 0.$$

Then (3.5) implies that $P(A_1) \to 0$ as $n \to \infty$. In order to prove $P(A_2) \to 0$ as $n \to \infty$, we examine the bound of $\|(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\|_2$. By (3.9) and Weyl's perturbation theorem for eigenvalues of the matrices, for all $1 \le j \le q_n$,

$$\begin{aligned}&\Big|\lambda_j\Big(\big(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\Big) - \lambda_j\Big(\big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\Big)\Big| \\ &\le \Big\|\big(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1} - \big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\Big\|_2 = \mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big).\end{aligned}$$

Therefore $\|(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\|_2 \le \lambda_{1,min}^{-1} + C$ with probability arbitrarily close to 1 for sufficiently large n. It follows from Assumptions (A4), (A6) and (A11) that

$$\begin{aligned}P(A_2) &\le P\Big(\frac{\lambda_n}{n}\max_{1\le j\le q_n}|\eta_{2,j}| \ge b_n\Big) \\ &\le P\Big(\frac{\lambda_n}{n}\Big\|\big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\Big\|_2\|s_n^0(1)\|_2 \ge b_n\Big) \\ &\le P\Big(\frac{\lambda_n\sqrt{q_n}}{nb_n^2} \ge C\Big) \to 0\end{aligned}$$

due to the fact that $\|s_n^0(1)\| \le \|\Sigma_1(1)\|_2\|sign(\boldsymbol{\beta}_n^0(1))\|_2 \le \frac{b_1\sqrt{q_n}}{b_n} = \mathcal{O}_p(\frac{\sqrt{q_n}}{b_n})$.

This completes the proof of (3.4). We now turn to show that

$$\lim_{n\to\infty} P(\delta_n^{[k]} \ne_s \boldsymbol{\beta}_n^{[k]}) = 0. \tag{3.10}$$

Observe that

$$\begin{aligned}&\big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1} \\ &= \big(\widetilde{\Gamma}_n^{[k]}(1)\big)^{-1} - \big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\frac{\gamma_n}{n}\Sigma_2(1)\big(\widetilde{\Gamma}_n^{[k]}(1)\big)^{-1}.\end{aligned} \tag{3.11}$$

Then, by Assumptions (A4) and (A11),

$$\begin{aligned}\|\boldsymbol{\beta}_n^{[k]} - \delta_n^{[k]}\|_2 &= \Big\|\Big[\big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\widetilde{\Gamma}_n^{[k]}(1) - I_{q_n}\Big]\boldsymbol{\beta}_n^0(1)\Big\|_2 \\ &= \Big\| - \big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\frac{\gamma_n}{n}\Sigma_2(1)\boldsymbol{\beta}_n^0(1)\Big\|_2 \\ &\le \frac{\gamma_n}{n}\Big\|\big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\Big\|_2\|\Sigma_2(1)\|_2\|\boldsymbol{\beta}_n^0(1)\|_2 \\ &= \mathcal{O}_p\Big(\frac{\gamma_n}{nb_n}\Big) \xrightarrow{\mathcal{P}} 0.\end{aligned}$$

This implies (3.10). Combining (3.4) and (3.10) leads to

$$\lim_{n\to\infty} P(\boldsymbol{\beta}_n^0 \ne_s \boldsymbol{\beta}_n^{[k]}) = 0. \tag{3.12}$$

Hence, to prove the sign consistency of the iteratively reweighted adaptive elastic net estimator, it suffices to show that, as $n \to \infty$, $\boldsymbol{\beta}_n^{[k]}$ satisfies the KKT conditions (3.1) and (3.2), so that $\boldsymbol{\beta}_n^{[k]}$ is indeed the solution of (2.4).

The above arguments for proving (3.4) and (3.10) imply that

$$\|\boldsymbol{\beta}_n^{[k]} - \boldsymbol{\beta}_n^0(1)\|_2 = \mathcal{O}_p\Big(\frac{\gamma_n}{nb_n} + \frac{(\log n)^{I\{d=1\}} + h_n}{\sqrt{n}} + \frac{\lambda_n\sqrt{q_n}}{nb_n}\Big). \tag{3.13}$$

From (3.3), (3.11)-(3.13) and Assumption (A11), for $1 \le j \le q_n$,

$$\begin{aligned}
&\mathbf{X}_j^{0'}(W_n^{[k-1]})^2(\mathbf{Y}_n - \mathbf{X}_n\boldsymbol{\beta}_n^{[k]}(1)) - \gamma_n u_{n,j}\beta_{n,j}^{[k]} \\
&= \mathbf{X}_j^{0'}(W_n^{[k-1]})^2\boldsymbol{\varepsilon}_n^0 + \mathbf{X}_j^{0'}(W_n^{[k-1]})^2\mathbf{X}_n(1)\big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\frac{\gamma_n}{n}\Sigma_2(1)\boldsymbol{\beta}_n^0(1) \\
&\quad -\mathbf{X}_j^{0'}(W_n^{[k-1]})^2\mathbf{X}_n(1)\frac{1}{n}\Big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}\mathbf{X}_n(1)'(W_n^{[k-1]})^2\boldsymbol{\varepsilon}_n^0 \\
&\quad +\mathbf{X}_j^{0'}(W_n^{[k-1]})^2\mathbf{X}_n(1)\frac{\lambda_n}{2n}\Big(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1)\Big)^{-1}s_n^0(1) - \gamma_n u_{n,j}\beta_{n,j}^{[k]} \\
&= \gamma_n u_{n,j}\beta_j + \frac{\lambda_n}{2}v_{n,j}sign(\beta_j) - \gamma_n u_{n,j}\beta_{n,j}^{[k]} + \mathcal{O}_p\Big(\frac{\gamma_n}{\sqrt{n}b_n}\Big) \\
&= \frac{\lambda_n}{2}v_{n,j}sign(\beta_{n,j}^{[k]}) + \mathcal{O}_p\Big(\frac{\gamma_n}{\sqrt{n}b_n} + \frac{\gamma_n\big((\log n)^{I\{d=1\}} + h_n\big)}{\sqrt{n}b_n} + \frac{\gamma_n\lambda_n\sqrt{q_n}}{n{b_n}^2}\Big).
\end{aligned}$$

This means that $\boldsymbol{\beta}_n^{[k]}$ satisfies the first KKT condition (3.1) as $n \to \infty$.

Let $\eta_{3,j} = \mathbf{X}_j^{0'}(W_n^{[k-1]})^2\big[I_n - \frac{1}{n}\mathbf{X}_n(1)(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^{[k-1]})^2\big]\boldsymbol{\varepsilon}_n^0$ and $\eta_{4,j} = \frac{\lambda_n}{2n}\mathbf{X}_j^{0'}(W_n^{[k-1]})^2\mathbf{X}_n(1)(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}s_n^0(1) + \mathbf{X}_j^{0'}(W_n^{[k-1]})^2\mathbf{X}_n(1)$ $\big[(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\,\widetilde{\Gamma}_n^{[k]}(1) - I_{q_n}\big]\boldsymbol{\beta}_n^0(1)$. Denote $A_3 = \Big\{|\eta_{3,j}| \ge \frac{\lambda_n}{4}v_{n,j}$, for some $j > q_n\Big\}$ and $A_4 = \Big\{|\eta_{4,j}| \ge \frac{\lambda_n}{4}v_{n,j},\ \text{ for some }\ j > q_n\Big\}$.

Then, to show that $\boldsymbol{\beta}_n^{[k]}$ satisfies the second KKT condition (3.2), we only need to prove that $P\big(|\eta_{3,j} - \eta_{4,j}| < \frac{\lambda_n}{2}v_{n,j}\big) \to 0$ as $n \to \infty$ for any $q_n < j \le p_n$. So it is enough to show that $P(A_j) \to 0$ as $n \to \infty$ for $j = 3, 4$.

Let $\eta_{3,j}^0 = \mathbf{X}_j^{0'}(W_n^0)^2\big[I_n - \frac{1}{n}\mathbf{X}_n(1)(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^0)^2\big]\boldsymbol{\varepsilon}_n^0$ and $\eta_{3,j}^{0,\infty} = \mathbf{X}_j^{0'}(W_n^0)^2\ \big[I_n - \frac{1}{n}\mathbf{X}_n(1)(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^0)^2\big]\boldsymbol{\varepsilon}_{n,\infty}^0$. Then

$$\begin{aligned}
P(A_3) \ \le\ & P\Big(\max_{q_n<j\le p_n}|\eta_{3,j}^{0,\infty}| \ge \frac{\lambda_n r_n}{8}\Big) + P\Big(\max_{q_n<j\le p_n}|\eta_{3,j} - \eta_{3,j}^0| \ge \frac{\lambda_n r_n}{16}\Big) \\
&+P\Big(\max_{q_n<j\le p_n}|\eta_{3,j}^0 - \eta_{3,j}^{0,\infty}| \ge \frac{\lambda_n r_n}{16}\Big) \\
&+P\Big(\max_{q_n<j\le p_n}|\beta_{j,init1}|^{\tau_1} \ge \frac{1}{r_n}\Big),
\end{aligned} \tag{3.14}$$

where $\beta_{j,init1}$ is the jth element of $\boldsymbol{\beta}_{n,init1}$.

For estimating the first term of (3.14), let $H_{n,j}^0 = \mathbf{X}_j^{0'}(W_n^0)^2\big[I_n - \frac{1}{n}\mathbf{X}_n(1)(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\ \mathbf{X}_n(1)'(W_n^0)^2\big]$. Thus we have $\eta_{3,j}^{0,\infty} = H_{n,j}^0\boldsymbol{\varepsilon}_{n,\infty}^0$. Note that

$$\begin{aligned}
\|H_{n,j}^0\|_2 \ &\le\ \|\mathbf{X}_j^0\|_2\|(W_n^0)^2\|_2\Big[1 + \Big\|\frac{1}{n}\mathbf{X}_n(1)(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'\Big\|_2\|(W_n^0)^2\|_2\Big] \\
&=\ \mathcal{O}_p(\sqrt{n}).
\end{aligned}$$

In a same way as in (3.7) and (3.8), by Assumption (11), we obtain

$$
\begin{aligned}
& P\Big(\max_{q_n<j\le p_n}|\eta_{3,j}^{0,\infty}|\ge\frac{\lambda_n r_n}{8}\Big)\\
& \le(\psi_d^{-1}\Big(\frac{\lambda_n r_n}{c_7\sqrt{n}(\log(1+p_n-q_n))^{1/d}(\log n)^{I\{d=1\}}}\Big)\Big)\\
& \longrightarrow 0. \qquad (3.15)
\end{aligned}
$$

Since

$$
\begin{aligned}
& |\eta_{3,j}-\eta_{3,j}^0|\\
& =\Big|\mathbf{X}_j^{0\prime}\big\{(W_n^0)^2[I_n-\frac{1}{n}\mathbf{X}_n(1)(\widetilde{\Gamma}_n^0(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^0)^2]\\
& \quad -(W_n^{[k-1]})^2[I_n-\frac{1}{n}\mathbf{X}_n(1)(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^{[k-1]})^2]\big\}\boldsymbol{\varepsilon}_n^0\Big|\\
& \le\|\mathbf{X}_j^0\|_2\|(W_n^0)^2-(W_n^{[k-1]})^2\|_2\|\boldsymbol{\varepsilon}_n^0\|_2+\|\mathbf{X}_j^0\|_2\|G_n\|_2\|\boldsymbol{\varepsilon}_n^0\|_2,
\end{aligned}
$$

where $G_n=\frac{1}{n}(W_n^0)^2\mathbf{X}_n(1)(\widetilde{\Gamma}_n^0(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^0)^2-\frac{1}{n}(W_n^{[k-1]})^2\mathbf{X}_n(1)$ $(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^{[k-1]})^2$, and

$$
\begin{aligned}
& \|G_n\|_2\\
& \le\|(W_n^0)^2-(W_n^{[k-1]})^2\|_2\|\Gamma_n(1)\|_2^2\|(\widetilde{\Gamma}_n^0(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\|_2\|(W_n^0)^2\|_2\\
& +\|(W_n^0)^2\|_2\|\Gamma_n(1)\|_2^2\|(\widetilde{\Gamma}_n^0(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}-(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\|_2\|(W_n^{[k-1]})^2\|_2\\
& +\|(W_n^0)^2-(W_n^{[k-1]})^2\|_2\|\Gamma_n(1)\|_2^2\|(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\|_2\|(W_n^{[k-1]})^2\|_2\\
& =\mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big)\mathcal{O}_p(1)\mathcal{O}_p(1)\mathcal{O}_p(1)=\mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big),
\end{aligned}
$$

then we have

$$
|\eta_{3,j}-\eta_{3,j}^0|=\mathcal{O}_p(\sqrt{n})\mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big)\mathcal{O}_p(\sqrt{n})+\mathcal{O}_p(\sqrt{n})\mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big)\mathcal{O}_p(\sqrt{n})=\mathcal{O}_p(h_n\sqrt{n}).
$$

This, together with Assumption (A11), yields that

$$
P\Big(\max_{q_n<j\le p_n}|\eta_{3,j}^0-\eta_{3,j}|\ge\frac{\lambda_n r_n}{16}\Big)\le P\Big(\frac{h_n\sqrt{n}}{\lambda_n r_n}\ge C\Big)\to 0. \qquad (3.16)
$$

Moreover, since $\frac{1}{\sqrt{n}}|\eta_{3,j}^0-\eta_{3,j}^{0,\infty}|\le\frac{1}{\sqrt{n}}\|H_{n,j}^0\|_2\|\boldsymbol{\varepsilon}_n^0-\boldsymbol{\varepsilon}_{n,\infty}^0\|_2=\mathcal{O}_p(1)$, it follows from Assumption (A11) that

$$
P\Big(\max_{q_n<j\le p_n}|\eta_{3,j}^0-\eta_{3,j}^{0,\infty}|\ge\frac{\lambda_n r_n}{16}\Big)\le P(\frac{\sqrt{n}}{\lambda_n r_n}\ge C\Big)\to 0. \qquad (3.17)
$$

By (3.14)-(3.17) and Assumption (A5), we arrive at $P(A_3)\to 0$ as $n\to\infty$.

For A_4, notice that

$$
\begin{aligned}
&|\eta_{4,j}| \\
&\leq \frac{\lambda_n}{2n}\|\mathbf{X}_j^0\|_2\|(W_n^{[k-1]})^2\|_2\|\mathbf{X}_n(1)\|_2\Big\|\big(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\Big\|_2\|s_n^0(1)\|_2 \\
&\quad+\|\mathbf{X}_j^0\|_2\|(W_n^{[k-1]})^2\|_2\|\mathbf{X}_n(1)\|_2\Big\|\big(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1)\big)^{-1}\widetilde{\Gamma}_n^{[k]}(1)-I_{q_n}\Big\|_2\|\boldsymbol{\beta}_n^0(1)\|_2 \\
&=\frac{\lambda_n}{n}\mathcal{O}_p(\sqrt{n})\mathcal{O}_p(1)\mathcal{O}_p(\sqrt{n})\mathcal{O}_p\Big(\frac{\sqrt{q_n}}{b_n}\Big)+\mathcal{O}_p(\sqrt{n})\mathcal{O}_p(1)\mathcal{O}_p(\sqrt{n})\mathcal{O}_p\Big(\frac{\gamma_n}{nb_n}\Big)\mathcal{O}_p(1) \\
&=\mathcal{O}_p\Big(\frac{\lambda_n\sqrt{q_n}}{b_n}\Big)+\mathcal{O}_p\Big(\frac{\gamma_n}{b_n}\Big).
\end{aligned}
$$

Then Assumption (A11) implies that

$$P\Big(\max_{q_n<j\leq p_n}|\eta_{4,j}|\geq\frac{\lambda_n r_n}{4}\Big)\leq P\Big(\frac{\sqrt{q_n}}{b_n r_n}\geq C\Big)+P\Big(\frac{\gamma_n}{b_n\lambda_n r_n}\geq C\Big)\to 0.$$

Assumption (A5) yields that

$$P(A_4)\leq P\Big(\max_{q_n<j\leq p_n}|\eta_{4,j}|\geq\frac{\lambda_n r_n}{4}\Big)+P\Big(\max_{q_n<j\leq p_n}|\beta_{j,init1}|^{\tau_1}\geq\frac{1}{r_n}\Big)\to 0$$

as $n\to\infty$.

This concludes the proof of the sign consistency of the estimator $\boldsymbol{\beta}_n^{[k]}$. Next we proceed to show the asymptotic normality of $\boldsymbol{\beta}_n^{[k]}$.

(II) The asymptotic normality

From (3.3), we have

$$
\begin{aligned}
&\frac{\sqrt{n}}{s_n(k)}\boldsymbol{\xi}_n'(\boldsymbol{\beta}_n^{[k]}(1)-\boldsymbol{\beta}_n^0(1)) \\
&=\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n'(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^{[k-1]})^2\boldsymbol{\varepsilon}_n^0 \\
&\quad-\frac{\lambda_n}{2\sqrt{n}s_n(k)}\boldsymbol{\xi}_n'(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}s_n^0(1) \\
&\quad+\frac{\sqrt{n}}{s_n(k)}\boldsymbol{\xi}_n'[(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\widetilde{\Gamma}_n^{[k]}(1)-I_{q_n}]\boldsymbol{\beta}_n^0(1).
\end{aligned}
\tag{3.18}
$$

For the first term of (3.18), similarly to the proof of part (I), we have the decomposition

$$(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^{[k-1]})^2=B_1+B_2+B_3,$$

where

$$
\begin{aligned}
B_1 &= (\widetilde{\Gamma}_n^0(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'(W_n^0)^2, \\
B_2 &= [(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}-(\widetilde{\Gamma}_n^0(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}]\mathbf{X}_n(1)'(W_n^0)^2, \\
B_3 &= (\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_n(1)'((W_n^{[k-1]})^2-(W_n^0)^2).
\end{aligned}
$$

Note that

$$\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n' B_1 \boldsymbol{\varepsilon}_n^0 = \frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n' B_1 \boldsymbol{\varepsilon}_{n,\infty}^0 + \frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n' B_1 (\boldsymbol{\varepsilon}_n^0 - \boldsymbol{\varepsilon}_{n,\infty}^0),$$

while

$$\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n' B_1 \boldsymbol{\varepsilon}_{n,\infty}^0 = \sum_{t=1}^{n} a_t Z_t$$

with $a_t = \frac{1}{\sqrt{n}s_n(k)\sigma_t}\boldsymbol{\xi}_n'(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\mathbf{X}_{t,n}(1)$. It is easy to see that $\{a_t Z_t, \mathcal{F}_{n,t}, 1 \le t \le n\}$ is a martingale difference array, where $\mathcal{F}_{n,t} = \sigma\{Z_{t_1}, X_{t_2,\infty}, 1 \le t_2 \le n, 1 \le t_1 \le t\}$ is the σ-field. Moreover, $E(\sum_{t=1}^n a_t Z_t) = 0$ and $E(\sum_{t=1}^n a_t Z_t)^2 = E(Z_t^2)E(\sum_{t=1}^n a_t^2) = 1$.

In addition, Assumption (A7) implies that $1/s_n(k) \le 1/\sqrt{\lambda_{2,min}}$. Then it follows from Assumptions (A3), (A8), (A9) and (A11) that

$$\begin{aligned}\max_{1\le t\le n}|a_t| &\le \frac{1}{\sqrt{n}s_n(k)}\|\boldsymbol{\xi}_n\|_2\|(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\|_2 \max_{1\le t\le n}\Big\|\frac{1}{\sigma_t}\mathbf{X}_{t,n}(1)\Big\|_2 \\ &\le \frac{C}{\sqrt{n}}\|\mathbf{X}_{t,n}(1)\|_2 = \mathcal{O}_p(\frac{\sqrt{q_n}v_n}{\sqrt{n}}) \xrightarrow{\mathcal{P}} 0.\end{aligned}$$

So the conditional Lindeberg condition is satisfied and the martingale central limit theorem (see, e.g. Theorem 2 of [1]) yields that

$$\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n' B_1 \boldsymbol{\varepsilon}_{n,\infty}^0 \xrightarrow{\mathcal{D}} Z. \tag{3.19}$$

On the other hand,

$$\begin{aligned}&\Big|\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n' B_1(\boldsymbol{\varepsilon}_n^0 - \boldsymbol{\varepsilon}_{n,\infty}^0)\Big| \\ &\le \frac{1}{\sqrt{n}s_n(k)}\|\boldsymbol{\xi}_n\|_2\|(\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\|_2\|\mathbf{X}_n(1)\|_2\|(W_n^0)^2\|_2\|\boldsymbol{\varepsilon}_n^0 - \boldsymbol{\varepsilon}_{n,\infty}^0\|_2 \\ &\le C\|\boldsymbol{\varepsilon}_n^0 - \boldsymbol{\varepsilon}_{n,\infty}^0\|_2 \xrightarrow{\mathcal{P}} 0.\end{aligned}$$

By Slutsky's Theorem,

$$\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n' B_1 \boldsymbol{\varepsilon}_{n,\infty}^0 \xrightarrow{\mathcal{D}} Z. \tag{3.20}$$

For B_2, we know that

$$\begin{aligned}&\Big|\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n' B_2 \boldsymbol{\varepsilon}_n^0\Big| \\ &\le \frac{1}{\sqrt{n}s_n(k)}\Big\|(\widetilde{\Gamma}_n^{[k]}(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1} - (\widetilde{\Gamma}_n^0(1) + \frac{\gamma_n}{n}\Sigma_2(1))^{-1}\Big\|_2\|\boldsymbol{\xi}_n\|_2 \\ &\quad \|\mathbf{X}_n(1)'(W_n^0)^2\boldsymbol{\varepsilon}_n^0\|_2,\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{X}_n(1)'(W_n^0)^2\boldsymbol{\varepsilon}_n^0\|_2 &\le \|\mathbf{X}_n(1)'(W_n^0)^2(\boldsymbol{\varepsilon}_n^0-\boldsymbol{\varepsilon}_{n,\infty}^0)\|_2+\|\mathbf{X}_n(1)'(W_n^0)^2\boldsymbol{\varepsilon}_{n,\infty}^0\|_2\\ &\le \|\mathbf{X}_n(1)\|_2\|(W_n^0)^2\|_2\|\boldsymbol{\varepsilon}_n^0-\boldsymbol{\varepsilon}_{n,\infty}^0\|_2+\|\mathbf{X}_n(1)'(W_n^0)^2\boldsymbol{\varepsilon}_{n,\infty}^0\|_2\\ &\le \mathcal{O}_p(\sqrt{n})\mathcal{O}_p(1)\mathcal{O}_p(1)+\|\mathbf{X}_n(1)'(W_n^0)^2\boldsymbol{\varepsilon}_{n,\infty}^0\|_2\\ &= \mathcal{O}_p(\sqrt{n})+\|\mathbf{X}_n(1)'(W_n^0)^2\boldsymbol{\varepsilon}_{n,\infty}^0\|_2.\end{aligned}$$

Markov's inequality and Assumptions (A1), (A2) and (A8) give

$$\begin{aligned}P\Big(\frac{1}{q_n n}\|\mathbf{X}_n(1)'(W_n^0)^2\boldsymbol{\varepsilon}_{n,\infty}^0\|_2^2>C\Big) &\le \frac{1}{Cq_n n}\sum_{i=1}^{q_n}E\Big(\sum_{t=1}^{n}X_{t,i}\frac{Z_t}{\sigma_t}\Big)^2\\ &\le \frac{1}{Cq_n n}\sum_{i=1}^{q_n}E\Big(\sum_{t=1}^{n}X_{t,i}\frac{Z_t}{\sigma_{min}}\Big)^2\le\frac{1}{C\sigma_{min}^2}.\end{aligned}$$

This means that $\|\mathbf{X}_n(1)'(W_n^0)^2\boldsymbol{\varepsilon}_{n,\infty}^0\|_2=\mathcal{O}_p(\sqrt{q_n n})$. These bounds together with Assumption (A11) imply that

$$\Big|\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n'B_2\boldsymbol{\varepsilon}_n^0\Big|=\mathcal{O}_p\Big(\frac{1}{\sqrt{n}}\Big)\mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big)\mathcal{O}_p(\sqrt{q_n n})\xrightarrow{\mathcal{P}}0. \tag{3.21}$$

Along similar lines for B_2, we obtain

$$\begin{aligned}&\Big|\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n'B_3\boldsymbol{\varepsilon}_n^0\Big|\\ &\le\frac{1}{\sqrt{n}s_n(k)}\|(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\|_2\|\boldsymbol{\xi}_n\|_2\|\mathbf{X}_n(1)'((W_n^0)^2-(W_n^{[k-1]})^2)\boldsymbol{\varepsilon}_n^0\|_2,\end{aligned}$$

and

$$\begin{aligned}\|\mathbf{X}_n(1)'((W_n^0)^2-(W_n^{[k-1]})^2)\boldsymbol{\varepsilon}_n^0\|_2\le\|\mathbf{X}_n(1)'((W_n^0)^2-(W_n^{[k-1]})^2)(\boldsymbol{\varepsilon}_n^0-\boldsymbol{\varepsilon}_{n,\infty}^0)\|_2\\ +\|\mathbf{X}_n(1)'((W_n^0)^2-(W_n^{[k-1]})^2)\boldsymbol{\varepsilon}_{n,\infty}^0\|_2.\end{aligned}$$

Moreover,

$$\begin{aligned}&\|\mathbf{X}_n(1)'((W_n^0)^2-(W_n^{[k-1]})^2)(\boldsymbol{\varepsilon}_n^0-\boldsymbol{\varepsilon}_{n,\infty}^0)\|_2\\ &\le\|\mathbf{X}_n(1)\|_2\|(W_n^0)^2-(W_n^{[k-1]})^2\|_2\|\boldsymbol{\varepsilon}_n^0-\boldsymbol{\varepsilon}_{n,\infty}^0\|_2\le\mathcal{O}_p(\sqrt{n})\mathcal{O}_p\Big(\frac{h_n}{\sqrt{n}}\Big)\mathcal{O}_p(1)\\ &=\mathcal{O}_p(h_n).\end{aligned}$$

From Markov's inequality and Assumptions (A1), (A2) and (A9)

$$\begin{aligned}&P\Big(\frac{1}{q_n h_n^2}\|\mathbf{X}_n(1)'((W_n^0)^2-(W_n^{[k-1]})^2)\boldsymbol{\varepsilon}_{n,\infty}^0\|_2^2>C\Big)\\ &\le\frac{1}{Cq_n h_n^2}\sum_{i=1}^{q_n}E\Big(\sum_{t=1}^{n}X_{t,i}\Big(\frac{1}{\sigma_t^2}-\frac{1}{\hat{\sigma}_t^{[k-1]}}\Big)\varepsilon_t\Big)^2\le\frac{1}{Cq_n h_n^2}\frac{h_n^2}{n}\sum_{i=1}^{q_n}E\Big(\sum_{t=1}^{n}X_{t,i}\varepsilon_t\Big)^2.\end{aligned}$$

That is $\|\mathbf{X}_n(1)'((W_n^0)^2-(W_n^{[k-1]})^2)\boldsymbol{\varepsilon}_{n,\infty}^0\|_2=\mathcal{O}_p(\sqrt{q_n}h_n)$. Therefore we have

$$\Big|\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n'B_3\boldsymbol{\varepsilon}_n^0\Big|=\mathcal{O}_p\Big(\frac{1}{\sqrt{n}}\Big)\mathcal{O}_p(1)\mathcal{O}_p(\sqrt{q_n}h_n)\xrightarrow{\mathcal{P}}0. \tag{3.22}$$

By (3.20)-(3.22) and Slutsky's Theorem,

$$\frac{1}{\sqrt{n}s_n(k)}\boldsymbol{\xi}_n'(B_1+B_2+B_3)\boldsymbol{\varepsilon}_n^0 \xrightarrow{\mathcal{D}} Z. \tag{3.23}$$

Now it suffices to show that the last two terms of (3.18) converge to zero in probability. By Assumption (A11),

$$\begin{aligned}
&\Big|\frac{\lambda_n}{2\sqrt{n}s_n(k)}\boldsymbol{\xi}_n'(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}s_n^0(1)\Big| \\
&\leq \frac{\lambda_n}{2\sqrt{n}s_n(k)}\|\boldsymbol{\xi}_n\|_2\|(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\|_2\|s_n^0(1)\|_2 \\
&= \mathcal{O}_p\Big(\frac{\lambda_n\sqrt{q_n}}{\sqrt{n}b_n}\Big) \xrightarrow{\mathcal{P}} 0.
\end{aligned}$$

For the last term of (3.18), by (3.11), we obtain

$$\Big|\frac{\sqrt{n}}{s_n(k)}\boldsymbol{\xi}_n'\Big[(\widetilde{\Gamma}_n^{[k]}(1)+\frac{\gamma_n}{n}\Sigma_2(1))^{-1}\widetilde{\Gamma}_n^{[k]}(1)-I_{q_n}\Big]\boldsymbol{\beta}_n^0(1)\Big| = \mathcal{O}_p\Big(\frac{\gamma_n}{\sqrt{n}b_n}\Big) \xrightarrow{\mathcal{P}} 0.$$

This completes the proof of Theorem 1. $\square$