Sequential Model Averaging for High Dimensional Linear Regression Models
Supplementary Materials

Wei Lan, Yingying Ma, Junlong Zhao, Hansheng Wang and Chih-Ling Tsai

Southwestern University of Finance and Economics, Beihang University, Beijing Normal University, Peking University and University of California, Davis

Abstract

This document includes five sections. Section 1 presents simulation studies that compare SMA with SIS, FR, MCV, and Bstg for less sparse regression models. These numerical results are given in Tables S1 and S2 for 10 and 20 true relevant predictors, respectively. Section 2 compares SMA with the Bayesian model averaging method for Examples 1-4, and the simulation results are given in Tables S3 and S4. Section 3 investigates the average number of steps reached via our proposed stopping rule. Sections 4 and 5 provide useful lemmas and theoretical proofs, respectively.

Section 1: Simulation Results For Less Sparse Models

We adopt and modify the simulation settings of Examples 1 and 2, respectively, from the manuscript. Specifically, we generate data from the model $Y_i = X_i^\top \beta + \sigma \varepsilon_i$, where ε_i is generated from a standard normal distribution for $i = 1, \cdots, n$, and σ is selected to generate a theoretical $R^2 = \text{var}(X_i^\top \beta)/\{\text{var}(X_i^\top \beta) + \sigma^2\} = 20\%$. The detailed structures of X_i and β in these two examples are illustrated below.

Example S1: We adapt this example from Fan and Lv (2008) and let d_0 be the size of the true model. In addition, for each i, the j-th covariates X_{ij} $(1 \leq j \leq p)$ are independently generated from $N(0,1)$. The r-th ($1 \le r \le d_0$) nonzero true coefficient of β is set equal to $(-1)^{u_r}(a_r + |v_r|)/10$, $a_r = 4\log(n)n^{-1/2}$, where u_r is a binary random variable with $P(u_r = 1) = 0.5$ and v_r is generated from a standard normal distribution.

Example S2: This example is modified from Tibshirani (1996). Specifically, the covariate vector X_i is generated from a multivariate normal distribution with mean zero and $\text{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$ for $1 \le j_1, j_2 \le p$. In addition, the true non-zero coefficients are set to be $\beta_{0j} = (-1)^j \times 0.5$ for any $1 \le j \le d_0$. Accordingly, $\beta_{0j} = 0$ for any $j > d_0$.

The simulation results for $d_0 = 10$ and 20 based on 1,000 realizations are presented in Tables S1 and S2, respectively. Both tables indicate that SMA performs well in comparison with SIS, FR, MCV and Bstg.

Section 2: Comparison of SMA versus Bayesian Model Averaging

In this section, we compare SMA with the Bayesian model averaging method. It is worth noting that, when p is ultra-high, the size of possible candidates models is 2^p. Hence, this method is computationally infeasible. To this end, we propose the following approach.

We randomly select a sub-model with size d_0, and then evaluate its BIC score (Chen and Chen, 2008). We next repeat the same procedure 1,000 times, which yields 1,000 sub-models, denoted them by $\mathcal{S}^{(1)}, \cdots, \mathcal{S}^{(1000)}$. Afterwards, we record their associated BIC scores $\text{BIC}(\mathcal{S}^{(1)}), \cdots, \text{BIC}(\mathcal{S}^{(1000)})$ and calculate the average. For making comparisons, we name this method feasible Bayesian model averaging (FB). Tables S3 and S4 present simulation results of SMA and FB, based on 1,000 realizations with $d_0 = n/4$, for Examples 1–4 in the manuscript. Both tables show that SMA is mostly superior to FB in terms of all three measures, AOR, SD, and

WP across all four examples.

Section 3: Average Number of Steps Reached via our Stopping Rule

The aim of this section is to study whether our proposed stopping criterion can be reached within a very limited number of steps when the number of true relevant predictors is small. To this end, we examine the average number of steps reached via our stopping rule for Examples 1–2 with the sample sizes $n = 100$, 200 and 300. Table S5 indicates that the resulting average numbers can be considerably larger than their associated true model sizes $d_0 = 5$ and 3. Hence, in practice, SMA may take more steps but yields better results. In addition, we have conducted simulation studies for Examples S1 and S2 with $d_0 = 10$ and $d_0 = 20$, respectively. The results yield similar findings in Table S6.

To assess whether our proposed stopping rule is sufficient for good prediction, we consider a simple simulation example below. Let $K = 50$ be the maximal number of steps during the sequential process of SMA, and define SMA^k be the k-th sequential step for $k = 1, \cdots, K$. We then evaluate the averaged out-of-sample R^2 values (AOR) for SMA^k with $k = 1, \cdots, K$. For the sake of illustration, we only consider Example S1 with $n = 100$ and $p = 100$. Figure S1 shows that AOR can increase very quickly when k is small, and it tends to be flat when k is larger than 25. Since the average stopping step is 24 for $n = 100$ and $p = 100$ (see Table S5), we conclude that the updates for up to 24 of SMA are sufficient for this example.

To see whether β_{kj} quickly become ignorable as k grows large. We also include the plot of $\max_j |\widehat{\beta}_{kj}|$ (we name it beta) against the number of steps k for $k = 1, \cdots, K$. Figure S2 indicates that $\max_j |\widehat{\beta}_{kj}|$ deceases to 0 quite slowly after $k > 1$. Hence, $\widehat{\beta}_{kj}$ is unlikely to become ignorable within a very limited number of sequential steps.

Section 4: Useful Lemmas

Before providing the theoretical proofs of Theorems 1–5, we present the following four useful lemmas. Lemma 1 can be shown in a manner similar to Proposition 1 of Jiang (2013). Lemma 2 can be verified by using the Bonferroni inequality; see, for example, Lemma A.3 in Bickel and Levina (2008). Lemma 3 is slightly modified from Lemma 1 of Wang (2009) and its proof is quite similar to that of Wang. Accordingly, we only present the detailed proof of Lemma 4.

Lemma 1. *Under Conditions (C2) and (C3), we have that, for any* $0 < \xi < 2$ *and* $j \in \{1, \cdots, p\}$,

$$\max_j P\big(|\widehat{\rho}_j - \rho_j| > \xi\big) \leq d_5 \exp\big(-d_6 n \xi^2\big),$$

where d_5 *and* d_6 *are finite constants and they are a function of* C_1, C_2, *and* d_1 *only,* C_1 *and* C_2 *are defined in Condition (C2), and* d_1 *is defined in Condition (C3).*

Lemma 2. *Under Conditions (C1)–(C3), we have that* $\max\limits_{1\leq j\leq p} |\mathbf{X}_j^\top \mathbf{X}_j/n - 1| \to_p 0$ *and* $(\log p)^{-1/2} \max\limits_{1\leq j\leq p} |n^{-1/2}\varepsilon^\top \mathbf{X}_j| = O_p(1)$ *as* $n \to \infty$.

Lemma 3. *Under conditions (C2), (C5) and (C6), as* $n \to \infty$, *we have*

$$2\tau_{\min} < \min_{|\mathcal{M}|\leq|\mathcal{M}_n|} \lambda_{\min}\{n^{-1}\mathbf{X}_{(\mathcal{M})}^\top \mathbf{X}_{(\mathcal{M})}\} \leq \max_{|\mathcal{M}|\leq|\mathcal{M}_n|} \lambda_{\max}\{n^{-1}\mathbf{X}_{(\mathcal{M})}^\top \mathbf{X}_{(\mathcal{M})}\} < \tau_{\max}/2.$$

Lemma 4. *Assume that Conditions (C1)–(C4) and the assumption in Theorem 1 hold. We then have that, for any finite* $k < \infty$, *(i.)* $\sqrt{n}(\widehat{\rho}_{kj} - \rho_{kj}) = O_p(1)$ $(j = 1, \cdots, p)$; *(ii.)* $\max_j |\widehat{\rho}_{kj}^2 - \rho_{kj}^2| \to 0$, *where* ρ_{kj} *is the population version of* $\widehat{\rho}_{kj}$ *and it is defined in Remark 1.*

Proof of Lemma 4. The result of $k = 1$ can be directly obtained from Lemma 1 and the Bonferroni inequality. By induction, we can show that it holds for general k. For the sake of simplicity, we only demonstrate that the result is valid for $k = 2$ by assuming that it holds when $k = 1$. Recall that $\beta^{(1)} = (w_{11}\beta_{11}, \cdots, w_{1p}\beta_{1p})^\top \in \mathbb{R}^p$, $\widehat{\beta}^{(1)} = (\widehat{w}_{11}\widehat{\beta}_{11}, \cdots, \widehat{w}_{1p}\widehat{\beta}_{1p})^\top \in \mathbb{R}^p$, and $\widetilde{\mathbf{Y}}_2 = \mathbf{Y}_1 - \mathbf{X}\beta^{(1)}$, where w_{1j} and $\widehat{w}_{1j}$ are defined in Section 2.3. Then, for any $j = 1, \cdots, p$, we have that

$$\widehat{\rho}_{2j} = \big\|\mathbf{Y}_2\big\|^{-1}\big\|\mathbf{X}_j\big\|^{-1}\mathbf{Y}_2^\top\mathbf{X}_j$$

$$= \big\|\widetilde{\mathbf{Y}}_2 + \mathbf{X}(\beta^{(1)} - \widehat{\beta}^{(1)})\big\|^{-1}\big\|\mathbf{X}_j\big\|^{-1}\Big\{\widetilde{\mathbf{Y}}_2^\top\mathbf{X}_j + \big(\beta^{(1)} - \widehat{\beta}^{(1)}\big)^\top\mathbf{X}^\top\mathbf{X}_j\Big\}.$$

By Conditions (C1)–(C3) and Lemmas 1–2, one can easily verify that, for every fixed j, $\|\widetilde{Y}_2\|^{-1}\|\mathbf{X}_j\|^{-1}(\widetilde{Y}_2^\top\mathbf{X}_j)$ is $\sqrt{n}$-consistent of ρ_{2j} and it is uniformly consistent of ρ_{2j} over j, where $\rho_{2j} = \mathrm{corr}(\widetilde{\mathbf{Y}}_2, \mathbf{X}_j)$ is defined in Remark 1. Hence, to prove Lemma 4, it suffices to demonstrate the following two results:

$$\big\|\mathbf{X}(\beta^{(1)} - \widehat{\beta}^{(1)})\big\| = O_p(1) \tag{0.1}$$

$$\text{and} \quad \max_j \big|n^{-1/2}(\beta^{(1)} - \widehat{\beta}^{(1)})^\top\mathbf{X}^\top\mathbf{X}_j\big| = O_p(1). \tag{0.2}$$

We next prove them via the following two separate steps, respectively.

STEP I. Let $\omega_{1(1)} \geq \cdots \geq \omega_{1(p)}$ be the ordered statistics of $\{\omega_{1j} : 1 \leq j \leq p\}$ and let $\widehat{\omega}_{1(j)}$ be their corresponding estimators for $j = 1, \cdots, p$. By the assumption of Theorem 1 that $\rho_{1(1)} - \rho_{1(2)} > d_2$ for some positive constant $d_2 > 0$, the techniques used in the proof of Theorem 1 and the result that $\max_j |\widehat{\rho}_{1j}^2 - \rho_{1j}^2| \to 0$, we have that there exists some constant $\zeta < 1$ such that

$$\widehat{\omega}_{1(1)} \to 1, \ \widehat{\omega}_{1(j)} \leq \zeta^n, \ \omega_{1(1)} \to 1, \text{ and } \omega_{1(j)} \leq \zeta^n.$$

Define $\widehat{\sigma}_{j_1j_2} = \|\mathbf{X}_{j_1}\|^{-1}\|\mathbf{X}_{j_2}\|^{-1}\mathbf{X}_{j_1}^\top\mathbf{X}_{j_2}$ as the sample counterpart of $\sigma_{j_1j_2}$. Then, we have

$$\max_j \big|n^{-1/2}(\beta^{(1)} - \widehat{\beta}^{(1)})^\top\mathbf{X}^\top\mathbf{X}_j\big| \leq \max_{j_1,j_2} |\widehat{\sigma}_{j_1j_2}| n^{1/2}|\beta^{(1)} - \widehat{\beta}^{(1)}|_1,$$

where $|\beta^{(1)} - \widehat{\beta}^{(1)}|_1 = \sum_j |\beta_j^{(1)} - \widehat{\beta}_j^{(1)}|$. After algebraic simplification, we obtain $n^{1/2}|\beta^{(1)} - \widehat{\beta}^{(1)}|_1 \leq n^{1/2}|\beta_{(1)}^{(1)} - \widehat{\beta}_{(1)}^{(1)}| + n^{1/2}(p-1)\zeta^n$. In addition, Condition (C1) implies that $n^{1/2}(p-1)\zeta^n \to 0$. Hence, $n^{1/2}|\beta^{(1)} - \widehat{\beta}^{(1)}|_1 = O_p(1)$. Moreover, Conditions (C1)–(C3), together with the Bonferroni inequality and Lemma 1, lead to, for any arbitrary positive constant $\gamma > 0$,

$$P\Big(\max_{j_1,j_2} |\widehat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| > \gamma\Big) \leq \sum_{j_1,j_2} P\Big(|\widehat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| > \gamma\Big)$$

$$\leq p^2 d_5 \exp\big(-d_6 n\gamma^2\big) \leq d_5 \exp\big(\nu n^\alpha - d_6 n\gamma^2\Big\},$$

where d_5 and d_6 are defined in Lemma 1 and $\alpha < 1$ is assumed in Condition (C1). Since $d_6 n\gamma^2$ dominates νn^α, we obtain that $\max_{j_1,j_2} |\widehat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| \to 0$. Accordingly, $\max_{j_1,j_2} |\widehat{\sigma}_{j_1j_2}| = O_p(1)$. This, in conjunction with $n^{1/2}|\beta^{(1)} - \widehat{\beta}^{(1)}|_1 = O_p(1)$, yields $\max_j \big|n^{-1/2}(\beta^{(1)} - \widehat{\beta}^{(1)})^\top \mathbf{X}^\top \mathbf{X}_j\big| = O_p(1)$, which completes the proof of (0.2).

STEP II. Note that

$$\big\|\mathbf{X}(\beta^{(1)} - \widehat{\beta}^{(1)})\big\|^2 = (\beta^{(1)} - \widehat{\beta}^{(1)})^\top \mathbf{X}^\top \mathbf{X}(\beta^{(1)} - \widehat{\beta}^{(1)})$$

$$= n\sum_{j_1,j_2} \widehat{\sigma}_{j_1j_2}(\beta_{j_1}^{(1)} - \widehat{\beta}_{j_1}^{(1)})(\beta_{j_2}^{(1)} - \widehat{\beta}_{j_2}^{(1)}) \leq \max_{j_1,j_2} |\widehat{\sigma}_{j_1j_2}| \times n|(\beta^{(1)} - \widehat{\beta}^{(1)})|_1^2.$$

Then, using the results $\big|(\beta^{(1)} - \widehat{\beta}^{(1)})\big|_1^2 = O_p(n^{-1})$ and $\max_{j_1,j_2} |\widehat{\sigma}_{j_1j_2}| = O_p(1)$ obtained in STEP I, we have that $\big\|\mathbf{X}(\beta^{(1)} - \widehat{\beta}^{(1)})\big\|^2 = O_p(1)$, which completes the proof of (0.1).

Section 5: Proofs of Theorems 1–5

Proof of Theorem 1: By theorem's assumption, we know that $\rho_{(1)}^2 - \rho_{(2)}^2 > d_2$. We then examine the difference between the corresponding estimates. That is,

$$\begin{aligned}
\widehat{\rho}^2_{(1)} - \widehat{\rho}^2_{(2)} &= \widehat{\rho}^2_{(1)} - \rho^2_{(1)} + \rho^2_{(1)} - \rho^2_{(2)} + \rho^2_{(2)} - \widehat{\rho}^2_{(2)} \\
&\geq \rho^2_{(1)} - \rho^2_{(2)} - |\rho^2_{(1)} - \widehat{\rho}^2_{(1)}| - |\rho^2_{(2)} - \widehat{\rho}^2_{(2)}| \\
&\geq \rho^2_{(1)} - \rho^2_{(2)} - 2\max_j |\rho^2_j - \widehat{\rho}^2_j| \\
&\geq d_2 - 2\max_j |\rho^2_j - \widehat{\rho}^2_j|, \qquad (0.3)
\end{aligned}$$

where the last inequality is due to fact that $\rho^2_{(1)} - \rho^2_{(2)} > d_2$. Furthermore, by Condition (C1), Bonferroni's inequality and Lemma 1, we have that

$$P\Big(\max_j |\widehat{\rho}^2_j - \rho^2_j| > \xi\Big) \leq \sum_j P\Big(|\widehat{\rho}^2_j - \rho^2_j| > \xi\Big) \leq \sum_j P\Big(|\widehat{\rho}_j - \rho_j| > \xi/2\Big)$$

$$\leq pd_5 \exp\big(-d_6 n\xi^2/4\big) \leq d_5 \exp\big(\log p - d_6 n\xi^2/4\big) \to 0.$$

This, together with (0.3), leads to, with probability tending to one, $\widehat{\rho}^2_{(1)} - \widehat{\rho}^2_{(2)} > d_2/2$.

Next, by definition, $\text{BIC}_{\mathcal{M}} = n\log\|\mathbf{Y} - \mathbf{X}_{(\mathcal{M})}\widehat{\beta}_{(\mathcal{M})}\|^2 + |\mathcal{M}| \times (\log n + 2\log p) = n\log\{\mathbf{Y}^\top(I - H_{\mathcal{M}})\mathbf{Y}\} + |\mathcal{M}|(\log n + 2\log p)$, where

$$H_{\mathcal{M}} = \mathbf{X}_{(\mathcal{M})}(\mathbf{X}^\top_{(\mathcal{M})}\mathbf{X}_{(\mathcal{M})})^{-1}\mathbf{X}^\top_{(\mathcal{M})}.$$

For the sake of convenience, let $\text{BIC}_{(1)}$ represent the BIC score associated with $\widehat{\rho}^2_{(1)}$.

We then have

$$
\begin{aligned}
w_{\max}^{U} &= \exp\Big(-\frac{1}{2}\mathrm{BIC}_{(1)}\Big)\left[\sum_{|\mathcal{M}_*|\le 1}\exp\Big(-\frac{1}{2}\mathrm{BIC}_{\mathcal{M}_*}\Big)\right]^{-1} \\
&= \Big(1-\widehat{\rho}_{(1)}^{2}\Big)^{-n/2}\left\{\sum_{k=1}^{p}\Big(1-\widehat{\rho}_{(k)}^{2}\Big)^{-n/2}+\sqrt{n}p\right\}^{-1} \\
&= \left\{\sum_{k\neq 1}\Big(1-\widehat{\rho}_{(1)}^{2}\Big)^{n/2}\Big(1-\widehat{\rho}_{(k)}^{2}\Big)^{-n/2}+\sqrt{n}p\Big(1-\widehat{\rho}_{(1)}^{2}\Big)^{n/2}+1\right\}^{-1}. \quad (0.4)
\end{aligned}
$$

It is noteworthy that $\sum_{k\neq 1}(1-\widehat{\rho}_{(1)}^{2})^{n/2}(1-\widehat{\rho}_{(k)}^{2})^{-n/2} \le p(1-\widehat{\rho}_{(1)}^{2})^{n/2}(1-\widehat{\rho}_{(2)}^{2})^{-n/2} = \exp\big[\log p + n/2\{\log(1-\widehat{\rho}_{(1)}^{2}) - \log(1-\widehat{\rho}_{(2)}^{2})\}\big]$. This, in conjunction with the result proved earlier that $\widehat{\rho}_{(1)}^{2} - \widehat{\rho}_{(2)}^{2} > d_2/2$ with probability tending to 1 and Condition (C3), implies that the right-hand side of the above inequality can be further bounded above by the following quantity

$$
\begin{aligned}
&\exp\Big\{\log p + 2^{-1}n\Big[\log(1-\widehat{\rho}_{(1)}^{2}) - \log(1-\widehat{\rho}_{(1)}^{2}+d_2/2)\Big]\Big\} \\
=\;& \exp\left\{\log p + 2^{-1}n\log\left(\frac{1-\widehat{\rho}_{(1)}^{2}}{1-\widehat{\rho}_{(1)}^{2}+d_2/2}\right)\right\} \\
=\;& \exp\left\{\log p + 2^{-1}n\log\left(1-\frac{d_2/2}{1-\widehat{\rho}_{(1)}^{2}+d_2/2}\right)\right\} \\
<\;& \exp\left\{\log p + 2^{-1}n\log\left(1-\frac{d_2/2}{1-d_1+d_2/2}\right)\right\}.
\end{aligned}
$$

By Condition (C1), we know immediately that the right-hand side of the above inequality is $o_p(1)$. Analogously, we can prove that $\sqrt{n}p(1-\widehat{\rho}_{(1)}^{2})^{n/2} \to_p 0$. As a result, we have $w_{\max}^{U} \to_p 1$, which completes the proof.

Proof of Theorem 2: By the definitions of $\mathbf{Y}_k$, $\widehat{w}_{kj}$, and H_j, we know that $\mathbf{Y}_{k+1} = (I - \sum_{j=1}^{p}\widehat{w}_{kj}H_j)\mathbf{Y}_k$. Then, define $\widetilde{\mathbf{X}}_j = \mathbf{X}_j/\|\mathbf{X}_j\|$, which immediately leads to $H_j =$

$\widetilde{\mathbf{X}}_j\widetilde{\mathbf{X}}_j^\top$. After algebraic simplification, we obtain that

$$\begin{aligned}
\|\mathbf{Y}_{k+1}\|^2 &= \|\mathbf{Y}_k\|^2 - 2\sum_{j=1}^p \widehat{w}_{kj}\mathbf{Y}_k^\top H_j \mathbf{Y}_k + \|\sum_{j=1}^p \widehat{w}_{kj} H_j \mathbf{Y}_k\|^2 \\
&= \|\mathbf{Y}_k\|^2 - 2\|\mathbf{Y}_k\|^2 \sum_{j=1}^p \widehat{w}_{kj}\widehat{\rho}_{kj}^2 + \|\sum_{j=1}^p \widehat{w}_{kj}\widetilde{\mathbf{X}}_j(\widetilde{\mathbf{X}}_j^\top \mathbf{Y}_k)\|^2 \\
&= \|\mathbf{Y}_k\|^2 \left(1 - 2\sum_{j=1}^p \widehat{w}_{kj}\widehat{\rho}_{kj}^2 + \|\sum_{j=1}^p \widehat{w}_{kj}\widetilde{\mathbf{X}}_j\widehat{\rho}_{kj}\|^2\right) \\
&= \|\mathbf{Y}_k\|^2 \left(1 - 2\sum_{j=1}^p \widehat{w}_{kj}\widehat{\rho}_{kj}^2 + \sum_{1\le j_1,j_2\le p} \widehat{w}_{kj_1}\widehat{w}_{kj_2}\widehat{\rho}_{kj_1}\widehat{\rho}_{kj_2}\widetilde{\sigma}_{j_1j_2}\right), \quad (0.5)
\end{aligned}$$

where $\widetilde{\sigma}_{j_1,j_2} = \widetilde{\mathbf{X}}_{j_1}^\top \widetilde{\mathbf{X}}_{j_2}$. Note that $|\widetilde{\sigma}_{j_1j_2}| \le 1$ for any $1 \le j_1, j_2 \le p$ since $\|\widetilde{\mathbf{X}}_j\| = 1$. Thus, the right-hand side of equation (0.5) can be bounded above by

$$\begin{aligned}
&\|\mathbf{Y}_k\|^2 \left\{1 - 2\sum_{j=1}^p \widehat{w}_{kj}\widehat{\rho}_{kj}^2 + \sum_{1\le j_1,j_2\le p} \widehat{w}_{kj_1}\widehat{w}_{kj_2}|\widehat{\rho}_{kj_1}\widehat{\rho}_{kj_2}|\right\} \\
= \ &\|\mathbf{Y}_k\|^2 \left\{1 - 2\sum_{j=1}^p \widehat{w}_{kj}\widehat{\rho}_{kj}^2 + \Big(\sum_{j=1}^p \widehat{w}_{kj}|\widehat{\rho}_{kj}|\Big)^2\right\}.
\end{aligned}$$

This suggests that

$$\|\mathbf{Y}_k\|^2 - \|\mathbf{Y}_{k+1}\|^2 \ge \|\mathbf{Y}_k\|^2 \left\{2\sum_{j=1}^p \widehat{w}_{kj}\widehat{\rho}_{kj}^2 - \Big(\sum_{j=1}^p \widehat{w}_{kj}|\widehat{\rho}_{kj}|\Big)^2\right\}. \quad (0.6)$$

By Cauchy's Inequality, we have $(\sum_{j=1}^p \widehat{w}_{kj})(\sum_{j=1}^p \widehat{w}_{kj}\widehat{\rho}_{kj}^2) \ge (\sum_{j=1}^p \widehat{w}_{kj}|\widehat{\rho}_{kj}|)^2$, which implies that $\sum_{j=1}^p \widehat{w}_{kj}\widehat{\rho}_{kj}^2 \ge (\sum_{j=1}^p \widehat{w}_{kj}|\widehat{\rho}_{kj}|)^2$. This, together with (0.6), yields $\|\mathbf{Y}_k\|^2 - \|\mathbf{Y}_{k+1}\|^2 \ge \|\mathbf{Y}_k\|^2 \sum_{j=1}^p \widehat{w}_{kj}\widehat{\rho}_{kj}^2$. This completes the proof.

Proof of Theorem 3: Using the result of Theorem 2, we have that $\|\mathbf{Y}_k\|^2 - \|\mathbf{Y}_{k+1}\|^2 \ge \|\mathbf{Y}_k\|^2 \widehat{w}_{k(1)}\widehat{\rho}_{k(1)}^2$. Hence, to prove this theorem, it suffices to show that

$\widehat{w}_{k(1)} = 1 + o_p(1)$. By definition,

$$\begin{aligned}
\widehat{w}_{k(1)} &= (1-\widehat{\rho}^2_{k(1)})^{-n/2}\left\{\sum_{j=1}^{p}(1-\widehat{\rho}^2_{kj})^{-n/2}+\sqrt{n}p\right\}^{-1}\\
&= \left\{\sum_{j\neq(1)}(1-\widehat{\rho}^2_{k(1)})^{n/2}(1-\widehat{\rho}^2_{kj})^{-n/2}+\sqrt{n}p(1-\widehat{\rho}^2_{k(1)})^{n/2}+1\right\}^{-1}. \qquad (0.7)
\end{aligned}$$

It is noteworthy that the first term on the right-hand side of (0.7) satisfies

$$\sum_{j\neq(1)}^{p}\left(\frac{1-\widehat{\rho}^2_{k(1)}}{1-\widehat{\rho}^2_{kj}}\right)^{n/2} \leq p\left(\frac{1-\widehat{\rho}^2_{k(1)}}{1-\widehat{\rho}^2_{k(2)}}\right)^{n/2}. \qquad (0.8)$$

In addition, the right-hand side of the above inequality equals

$$\exp\left[2^{-1}\left\{2\log p + n\log(1-\widehat{\rho}^2_{k(1)}) - n\log(1-\widehat{\rho}^2_{k(2)})\right\}\right]. \qquad (0.9)$$

Moreover, by Lemma 4 and using similar arguments for obtaining (0.2), we have that $|\widehat{\rho}^2_{k(1)} - \widehat{\rho}^2_{k(2)}| \geq d_4 - 2\max_j|\widehat{\rho}^2_{kj} - \rho^2_{kj}| \geq d_4/2$ with probability approaching 1. Subsequently, by Conditions (C1) and (C4), (0.9) can be further asymptotically bounded by

$$\begin{aligned}
&\exp\Big\{\log p + 2^{-1}n\Big[\log(1-\widehat{\rho}^2_{k(1)}) - \log(1-\widehat{\rho}^2_{k(1)} + d_4/2)\Big]\Big\}\\
=\ &\exp\left\{\log p + 2^{-1}n\log\left(\frac{1-\widehat{\rho}^2_{(1)}}{1-\widehat{\rho}^2_{(1)}+d_4/2}\right)\right\}\\
=\ &\exp\left\{\log p + 2^{-1}n\log\left(1-\frac{d_4/2}{1-\widehat{\rho}^2_{(1)}+d_4/2}\right)\right\}\\
<\ &\exp\left\{\log p + 2^{-1}n\log\left(1-\frac{d_4/2}{1-d_3+d_4/2}\right)\right\} \to 0.
\end{aligned}$$

As a result, the right-hand side in (0.8) goes to 0. Similarly, by Condition (C4), we can prove that the second term in (0.7) is of order $o_p(1)$. Consequently, $\widehat{w}_{k(1)} = 1 + o_p(1)$, which completes the proof.

Proof of Theorem 4: By the definitions of $\mathbf{Y}_k$, $\widehat{w}_{kj}$, and $\widehat{\rho}^2_{kj}$, and the equation (0.5) in the proof of Theorem 2, we obtain the following relationship:

$$\begin{aligned} \|\mathbf{Y}_k\|^2 - \|\mathbf{Y}_{k+1}\|^2 &= 2\|\mathbf{Y}_k\|^2 \left(\sum_{j=1}^{p} \widehat{w}_{kj} \widehat{\rho}^2_{kj} \right) - \|\mathbf{Y}_k\|^2 \cdot \left\| \sum_{j=1}^{p} \widehat{w}_{kj} \widetilde{\mathbf{X}}_j \widehat{\rho}_{kj} \right\|^2 \\ &\leq 2\|\mathbf{Y}_k\|^2 \left(\sum_{j=1}^{p} \widehat{w}_{kj} \widehat{\rho}^2_{kj} \right). \end{aligned} \qquad (0.10)$$

Since $\widehat{\rho}^2_{kj} \leq \widehat{\rho}^2_{k(1)}$ for $j = 1, \cdots, p$ and $\sum_{j=1}^{p} \widehat{w}_{kj} = 1 - \widehat{w}_{k0}$, the right-hand side of (0.10) is bounded by $2(1 - \widehat{w}_{k0})\|\mathbf{Y}_k\|^2 \widehat{\rho}^2_{k(1)}$. It is noteworthy that

$$\widehat{w}_{k0} = \sqrt{n}p \left\{ \sum_{j=1}^{p} (1 - \widehat{\rho}^2_{kj})^{-n/2} + \sqrt{n}p \right\}^{-1} = \left\{ \sum_{j=1}^{p} (1 - \widehat{\rho}^2_{kj})^{-n/2} (\sqrt{n}p)^{-1} + 1 \right\}^{-1}.$$

In addition, by theorem's assumption that $\rho^2_{k(1)} = O(n^{-1})$ and using the results of Lemma 4, we have $(1 - \widehat{\rho}^2_{kj})^{-n/2} = O_p(1)$ for any j. As a result, $n^{-1/2}p^{-1}\sum_{j=1}^{p}(1 - \widehat{\rho}^2_{kj})^{-n/2} \to_p 0$, which implies that $\widehat{w}_{k0} = 1 + o_p(1)$. Consequently, we can obtain that $(\|\mathbf{Y}_k\|^2 - \|\mathbf{Y}_{k+1}\|^2)/\|\mathbf{Y}_k\|^2 = o_p(\widehat{\rho}^2_{k(1)})$, which completes the proof.

Proof of Theorem 5: To prove the theorem, we consider two steps: (i.) demonstrate that $\widehat{\rho}_{K(1)} \to_p 0$ as $\min\{K, n\} \to \infty$; (ii.) use the result from (i) to show that $\|\widehat{\beta}^K - \beta_0\| \to_p 0$ as $\min\{K, n\} \to \infty$.

STEP I. We prove this step by the contradiction approach. Suppose there exists a sequence of $K_n \to \infty$ and a positive constant $c \in (0, 1)$ such that $\widehat{\rho}^2_{K_n(1)} > c$ for any $K_n > 0$. Accordingly,

$$\widehat{\omega}_{K_n 0}/\widehat{\omega}_{K_n(1)} = \sqrt{n}p\{1 - \widehat{\rho}^2_{K_n(1)}\}^{n/2} \leq \sqrt{n}p(1 - c)^{n/2}.$$

Since $\widehat{\omega}_{K_n(1)} \le 1$, we further have $\widehat{\omega}_{K_n0} \le \sqrt{n}p(1-c)^{n/2}$. This leads to

$$\sum_{j=1}^{p} \widehat{\omega}_{K_nj} \ge 1 - \sqrt{n}p(1-c)^{n/2}. \tag{0.11}$$

Next, let $j_n^* = \min\big\{j : \widehat{\rho}^2_{K_n(1)} - \widehat{\rho}^2_{K_n(j)} > c/2\big\}$. Then, for any $j \ge j_n^*$, we have

$$\begin{aligned}\widehat{\omega}_{K_n(j)}/\widehat{\omega}_{K_n(1)} &= \{1-\widehat{\rho}^2_{K_n(1)}\}^{n/2}\{1-\widehat{\rho}^2_{K_n(j)}\}^{-n/2}\\ &= \Big(\frac{1-\widehat{\rho}^2_{K_n(1)}}{1-\widehat{\rho}^2_{K_n(j)}}\Big)^{n/2} \le (1-c/2)^{n/2}.\end{aligned}$$

As a result, $\sum_{j=j_n^*}^{p} \widehat{\omega}_{K_n(j)} \le p(1-c/2)^{n/2}$. This, together with (0.11), leads to $\sum_{j=1}^{j_n^*-1} \widehat{\omega}_{K_n(j)} \ge 1 - \sqrt{n}p(1-c)^{n/2} - p(1-c/2)^{n/2}$. In addition, by Theorem 2 and the assumption of $var(Y_i) = 1$, we have that $n^{-1}\|\mathbf{Y}_K\|^2 > n^{-1}\|\mathbf{Y}_{K+1}\|^2$ and $n^{-1}\|\mathbf{Y}_1\|^2 < \infty$, respectively, as $K \to \infty$. This indicates that $n^{-1}\|\mathbf{Y}_K\|^2$ is a bounded decreasing sequence. Hence, there exists a positive constant c^* such that $n^{-1}\|\mathbf{Y}_K\|^2 \to_p c^*$, which implies that $n^{-1}\|\mathbf{Y}_K\|^2 - n^{-1}\|\mathbf{Y}_{K+1}\|^2 \to 0$ as K goes to infinity. Subsequently, by Theorem 2, we further have $\sum_{j=1}^{p} \widehat{\omega}_{K_nj}\widehat{\rho}^2_{K_nj} \to 0$. Since $\sum_{j=1}^{p} \widehat{\omega}_{K_nj}\widehat{\rho}^2_{K_nj} \ge \sum_{j=1}^{j_n^*-1} \widehat{\omega}_{K_n(j)}\widehat{\rho}^2_{K_n(j)}$, we finally obtain that

$$\sum_{j=1}^{j_n^*-1} \widehat{\omega}_{K_n(j)}\widehat{\rho}^2_{K_n(j)} \to_p 0. \tag{0.12}$$

On the other hand, by the definition of j_n^*, we have $\widehat{\rho}^2_{K_n(j)} \ge \widehat{\rho}^2_{K_n(1)} - c/2 \ge c/2$, for any $1 \le j \le j_n^* - 1$. Accordingly, $\sum_{j=1}^{j_n^*-1} \widehat{\omega}_{K_n(j)}\widehat{\rho}^2_{K_n(j)} \ge c/2\sum_{j=1}^{j_n^*-1} \widehat{\omega}_{K_n(j)} \ge c\big\{1 - \sqrt{n}p(1-c)^{n/2} - p(1-c/2)^{n/2}\big\}/2$. Using Condition (C1), one can easily verify that

$$1 - \sqrt{n}p(1-c)^{n/2} - p(1-c/2)^{n/2} \to 1$$

as $n \to \infty$. Consequently, we obtain that

$$\sum_{j=1}^{j_n^*-1} \widehat{\omega}_{K_n(j)} \widehat{\rho}^2_{K_n(j)} \geq c\big\{1 - \sqrt{n}p(1-c)^{n/2} - p(1-c/2)^{n/2}\big\}/2 \to c/2 \neq 0,$$

which contradicts (0.12). Thus, we have shown that $\widehat{\rho}_{K(1)} \to_p 0$ as $\min\{K, n\} \to \infty$.

STEP II. By the definition of $\widehat{\rho}^2_{K(1)}$, we have

$$\begin{aligned} \widehat{\rho}^2_{K(1)} &= \max_{1\leq j\leq p} (\mathbf{X}_j^\top \mathbf{Y}_K)^2 \|\mathbf{X}_j\|^{-2} \|\mathbf{Y}_K\|^{-2} \\ &\geq \max_{1\leq j\leq p} (n^{-1}\mathbf{X}_j^\top \mathbf{Y}_K)^2 \big\{\max_{1\leq j\leq p} n^{-1}\|\mathbf{X}_j\|^2\big\}^{-1} \big\{n^{-1}\|\mathbf{Y}_K\|^2\big\}^{-1} \\ &\geq \max_{1\leq j\leq p} (n^{-1}\mathbf{X}_j^\top \mathbf{Y}_K)^2 \big\{\max_{1\leq j\leq p} n^{-1}\|\mathbf{X}_j\|^2\big\}^{-1} \big\{n^{-1}\|\mathbf{Y}\|^2\big\}^{-1}, \qquad (0.13) \end{aligned}$$

where the last inequality is due to the fact that $\|\mathbf{Y}_K\|^2 \leq \|\mathbf{Y}\|^2$ for any $K \geq 1$; see Theorem 2. Using the assumptions of $\mathrm{var}(X_{ij}) = \mathrm{var}(Y_i) = 1$ for $1 \leq j \leq p$ and Lemma 2, we obtain that $n^{-1}\|\mathbf{Y}\|^2 \leq 2$ and $\max_j n^{-1}\|\mathbf{X}_j\|^2 \leq 2$ with probability tending to 1. As a result, the right-hand side of (0.13) can be further bounded away from 0; i.e.,

$$\widehat{\rho}^2_{K(1)} \geq 4^{-1} \max_{1\leq j\leq p} (n^{-1}\mathbf{X}_j^\top \mathbf{Y}_K)^2 \qquad (0.14)$$

with probability approaching 1, and it is uniform for any K.

Next, define $\Delta = \beta_0 - \widehat{\beta}^K$, which leads to $\mathbf{Y}_K = \mathbf{Y} - \mathbf{X}\widehat{\beta}^K = \mathbf{X}\Delta + \varepsilon$. By triangle inequality, we obtain

$$|\mathcal{M}_n|^{1/2} \max_{1\leq j\leq p} n^{-1}|\mathbf{X}_j^\top \mathbf{X}\Delta| \leq |\mathcal{M}_n|^{1/2} \max_{1\leq j\leq p} n^{-1}|\mathbf{X}_j^\top \mathbf{Y}_K|$$

$$+|\mathcal{M}_n|^{1/2} \max_{1\leq j\leq p} n^{-1}|\mathbf{X}_j^\top \varepsilon|.$$

Using equation (0.14) and the result, $\widehat{\rho}_{K(1)} \to_p 0$ as $\min\{K, n\} \to \infty$, proved in Step I, we have $|\mathcal{M}_n|^{1/2} \max_{1\leq j\leq p} n^{-1}|\mathbf{X}_j^\top \mathbf{Y}_K| \to_p 0$. In addition, Lemma 2 implies that

$|\mathcal{M}_n|^{1/2}\max_{1\le j\le p} n^{-1}|\mathbf{X}_j^\top\varepsilon| \to_p 0$. Accordingly,

$$|\mathcal{M}_n|^{1/2}\max_{1\le j\le p} n^{-1}|\mathbf{X}_j^\top\mathbf{X}\Delta| \to_p 0. \tag{0.15}$$

Moreover, employing Cauchy-Schwarz inequality and the stated assumption, together with the fact that $\|\mathbf{Y}_k\|^2 \le \|Y\|^2 < 2$ and $\widehat{\rho}_{kj}^2 \le 1$ uniformly for any $k = 1, \cdots, K$, we have

$$\begin{aligned}
|\mathcal{M}_n^c|\|\widehat{\beta}^K_{\mathcal{M}_n^c}\|^2 &= |\mathcal{M}_n^c|\sum_{j\in\mathcal{M}_n^c}\big(\sum_{k=1}^K \widehat{\omega}_{kj}\widehat{\beta}_{kj}\big)^2 \le |\mathcal{M}_n^c|K\sum_{j\in\mathcal{M}_n^c}\sum_{k=1}^K \widehat{\omega}_{kj}^2\widehat{\beta}_{kj}^2\\
&= |\mathcal{M}_n^c|K\sum_{j\in\mathcal{M}_n^c}\sum_{k=1}^K \widehat{\omega}_{kj}^2\|\mathbf{Y}_k\|^2\|\mathbf{X}_j\|^{-2}\widehat{\rho}_{kj}^2\\
&\le \big\{\min_j\|\mathbf{X}_j\|\big\}^{-2}\|\mathbf{Y}\|^2|\mathcal{M}_n^c|K^2\sup_{k>1}\sum_{j\in\mathcal{M}_n^c}\widehat{\omega}_{kj}^2 \to 0, \qquad (0.16)
\end{aligned}$$

with probability approaching to 1. As a result, we can have $\lambda_{\max}(\mathbf{X}_{(\mathcal{M}_n^c)}^\top\mathbf{X}_{(\mathcal{M}_n^c)})\|\widehat{\beta}^K_{\mathcal{M}_n^c}\|^2 \le tr(\mathbf{X}_{(\mathcal{M}_n^c)}^\top\mathbf{X}_{(\mathcal{M}_n^c)})\|\widehat{\beta}^K_{\mathcal{M}_n^c}\|^2 = O(|\mathcal{M}_n^c|\|\widehat{\beta}^K_{\mathcal{M}_n^c}\|^2) \to 0$. This, together with Condition (C5.2), further implies that $\lambda_{\max}(\mathbf{X}_{(\mathcal{M}_n^c)}^\top\mathbf{X}_{(\mathcal{M}_n^c)})\|\Delta_{(\mathcal{M}_n^c)}\|^2 \to 0$. Subsequently, by the Cauchy-Schwarz inequality again, Lemma 2, and Condition (C5.2), we obtain that

$$\begin{aligned}
|\mathcal{M}_n|\big\{n^{-1}\max_{1\le j\le p}|\mathbf{X}_j^\top\mathbf{X}_{(\mathcal{M}_n^c)}\Delta_{(\mathcal{M}_n^c)}|\big\}^2 &\le |\mathcal{M}_n|\big\{n^{-1}\max_{1\le j\le p}\|\mathbf{X}_j\|^2\big\}\big\{n^{-1}\|\mathbf{X}_{(\mathcal{M}_n^c)}\Delta_{(\mathcal{M}_n^c)}\|^2\big\}\\
&\le 2|\mathcal{M}_n|\lambda_{\max}\big\{n^{-1}\mathbf{X}_{(\mathcal{M}_n^c)}^\top\mathbf{X}_{(\mathcal{M}_n^c)}\big\}\|\Delta_{(\mathcal{M}_n^c)}\|^2 \to_p 0.
\end{aligned}$$

This, in conjunction with (0.15), implies that

$$|\mathcal{M}_n|^{1/2}\max_{1\le j\le p} n^{-1}|\mathbf{X}_j^\top\mathbf{X}_{(\mathcal{M}_n)}\Delta_{(\mathcal{M}_n)}| = o_p(1). \tag{0.17}$$

For an arbitrary vector $q = (q_1, \cdots, q_d)^\top \in \mathbb{R}^d$, define $\|q\|_1 = \sum|q_j|$ to be its L_1

norm. Then, by Condition (C6) and Lemma 3, we have that

$$\begin{aligned}
\|\Delta_{(\mathcal{M}_n)}\|^2 &\leq \tau_{\min}^{-1}\Delta_{(\mathcal{M}_n)}^{\top}\big\{n^{-1}\mathbf{X}_{(\mathcal{M}_n)}^{\top}\mathbf{X}_{(\mathcal{M}_n)}\big\}\Delta_{(\mathcal{M}_n)} \\
&\leq \tau_{\min}^{-1}\|\Delta_{(\mathcal{M}_n)}\|_1\max_j\big\{n^{-1}|\mathbf{X}_j^{\top}\mathbf{X}_{(\mathcal{M}_n)}\Delta_{(\mathcal{M}_n)}|\big\} \\
&\leq \tau_{\min}^{-1}|\mathcal{M}_n|^{1/2}\|\Delta_{(\mathcal{M}_n)}\|\max_j\big\{n^{-1}|\mathbf{X}_j^{\top}\mathbf{X}_{(\mathcal{M}_n)}\Delta_{(\mathcal{M}_n)}|\big\}.
\end{aligned}$$

This, in conjunction with (0.17), leads to

$$\|\Delta_{(\mathcal{M}_n)}\| \leq \tau_{\min}^{-1}|\mathcal{M}_n|^{1/2}\max_j\{n^{-1}|\mathbf{X}_j^{\top}\mathbf{X}_{(\mathcal{M}_n)}\Delta_{(\mathcal{M}_n)}|\} \to_p 0.$$

Furthermore, by the result $\lambda_{\max}(\mathbf{X}_{(\mathcal{M}_n^c)}^{\top}\mathbf{X}_{(\mathcal{M}_n^c)})\|\Delta_{(\mathcal{M}_n^c)}\|^2 \to 0$, we have that $\|\Delta\| = \|\widehat{\beta}^K - \beta_0\| \to_p 0$, which completes the proof.

REFERENCES

Bickel, P. J. and Levina, E. (2008), Covariance regularization by thresholding, *The Annals of Statistics* **36**, 2577–2604.

Chen, J. and Chen, Z. (2008), Extended Bayesian information criterion for model selection with large model spaces, *Biometrika* **95**, 759–771.

Fan, J. and Lv, J. (2008), Sure independence screening for ultra-high dimensional feature space (with discussion), *Journal of the Royal Statistical Society, Series B* **70**, 849–911.

Jiang, B. Y. (2013), Covariance selection by thresholding the sample correlation matrix, *Statistics and Probability Letters* **83**, 2492–2498.

Tibshirani, R. J. (1996), Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Wang, H. (2009), Forward regression for ultra-high dimensional variable screening, *Journal of the American Statistical Association* **104**, 1512–1524.

Table S1: Simulation results for Examples S1 and S2 with $d_0 = 10$.

			AOR (%)					SD (%)					WP(%)			
Example	n	p	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg
S1	100	100	-2.26	-0.43	1.24	3.45	3.85	6.21	3.23	3.30	2.24	3.07	89.40	94.80	92.40	67.50
		1000	-7.34	-2.67	-1.33	-0.23	0.78	5.89	3.77	3.48	3.23	3.19	86.80	90.50	88.10	77.60
		10000	-15.07	-6.47	-4.89	-3.56	-0.56	6.02	3.34	3.78	4.01	2.98	88.50	86.30	78.30	68.50
	200	100	2.78	1.84	2.91	6.71	7.10	3.77	3.38	2.39	2.50	2.70	91.50	99.00	92.50	63.50
		1000	-3.27	-3.38	-0.34	1.77	2.38	3.90	3.87	2.75	2.89	3.02	92.60	98.40	86.40	58.30
		10000	-6.29	-5.33	-2.57	0.12	0.87	5.07	4.06	3.02	2.77	3.68	94.50	90.80	82.70	60.30
	300	100	4.89	3.66	3.89	7.45	9.01	4.02	3.20	2.88	2.27	2.55	84.30	90.20	89.10	68.10
		1000	0.57	-0.22	0.01	3.29	5.20	3.89	3.32	2.76	2.19	2.49	85.20	88.10	82.60	65.20
		10000	-2.47	-2.58	-2.12	0.28	2.12	4.56	3.29	2.39	3.02	2.89	84.80	85.80	80.90	70.60
S2	100	100	-5.78	-5.42	-0.55	0.01	0.05	4.42	2.48	2.18	1.92	2.38	95.00	92.50	70.50	60.30
		1000	-9.68	-8.05	-3.23	-2.87	-2.55	5.89	3.87	3.02	3.02	3.33	92.30	87.30	76.50	68.50
		10000	-16.85	-15.28	-6.49	-5.78	-5.20	6.59	4.80	4.08	4.76	4.01	94.70	88.50	77.20	57.10
	200	100	-1.24	-0.98	0.47	1.69	1.46	2.84	1.89	1.55	1.28	1.44	90.00	78.90	72.10	44.50
		1000	-4.29	-3.28	-0.39	0.35	0.46	3.07	2.21	1.89	1.68	1.98	92.30	74.30	70.60	54.70
		10000	-8.56	-6.48	-3.67	-2.67	-2.34	3.21	2.67	2.01	1.87	2.02	94.20	75.40	68.30	53.80
	300	100	1.27	1.35	3.47	4.89	5.04	3.89	2.89	2.22	1.79	1.58	94.30	82.00	68.10	55.30
		1000	-1.28	-0.88	0.66	1.02	1.12	4.08	3.45	3.02	2.77	2.88	90.20	78.20	70.50	56.10
		10000	-5.89	-4.37	-1.22	-1.06	-0.67	4.18	3.89	4.07	1.90	2.76	84.10	74.50	67.70	54.10

Table S2: Simulation results for Examples S1 and S2 with $d_0 = 20$.

			AOR (%)					SD (%)					WP(%)			
Example	n	p	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg
S1	100	100	-3.85	-2.65	0.62	1.87	2.68	4.75	3.72	2.41	2.24	2.77	95.50	92.00	80.30	71.50
		1000	-9.24	-6.62	-3.30	-1.23	-0.26	4.17	3.28	2.19	2.78	3.02	94.20	90.00	79.50	67.40
		10000	-18.12	-10.47	-5.29	-4.33	-2.55	4.13	3.67	2.78	3.08	2.89	92.10	87.50	77.10	70.10
	200	100	-0.07	0.46	1.29	4.19	4.79	2.97	3.23	3.02	2.55	2.64	85.10	79.50	74.50	58.60
		1000	-4.78	-4.15	-0.25	0.23	0.46	3.50	3.32	2.86	2.44	2.87	84.00	75.80	70.40	56.10
		10000	-8.20	-7.22	-4.21	-2.47	-2.04	3.68	3.29	3.07	2.56	3.08	88.10	78.40	71.60	60.00
	300	100	1.06	1.22	2.57	4.89	5.06	4.06	3.78	3.45	2.98	3.23	90.50	86.50	80.80	54.60
		1000	-2.56	-2.15	-1.45	0.46	0.97	4.46	3.45	2.49	3.04	2.77	88.70	85.10	78.50	55.10
		10000	-5.42	-4.88	-3.27	-1.24	-0.78	4.21	4.29	2.67	3.89	3.08	89.20	83.40	80.20	58.30
S2	100	100	-6.76	-5.78	-1.42	-0.71	-0.23	4.23	3.32	2.34	2.09	2.51	98.50	92.30	80.20	53.30
		1000	-11.27	-10.16	-5.26	-4.02	-3.76	4.44	4.01	3.41	2.89	3.35	97.60	90.50	77.20	52.60
		10000	-20.75	-18.62	-9.22	-7.29	-7.00	4.58	5.08	3.78	3.28	4.22	95.40	89.60	74.30	55.80
	200	100	-2.16	-1.56	0.01	0.67	0.89	3.04	2.87	2.98	2.17	1.67	97.00	87.40	70.50	55.30
		1000	-6.31	-5.39	-0.39	0.35	0.46	3.07	2.21	1.89	1.68	1.98	92.30	74.30	70.60	54.70
		10000	-14.44	-12.47	-7.33	-6.23	-5.89	3.19	3.07	2.55	2.43	2.22	95.60	77.40	70.30	55.90
	300	100	-0.56	0.78	1.28	1.66	1.89	4.07	3.56	2.78	1.99	1.98	95.40	78.90	70.10	53.20
		1000	-3.45	-2.89	-0.76	0.08	0.22	4.78	3.22	2.98	2.02	3.02	94.30	80.50	70.50	58.90
		10000	-7.59	-6.86	-3.24	-3.04	-2.79	4.46	3.10	3.79	1.95	2.36	90.70	84.20	68.80	60.00

Table S3: Simulation results for comparing SMA with FB in Examples 1 and 2.

			AOR (%)		SD (%)		WP(%)
Example	n	p	FB	SMA	FB	SMA	FB
1	100	100	3.42	6.17	3.77	3.72	89.40
		1000	1.77	2.43	3.68	3.73	86.60
		10000	0.04	0.57	2.56	2.74	88.90
	200	100	7.98	10.96	3.08	3.29	85.40
		1000	5.21	7.12	3.45	3.92	82.60
		10000	2.10	4.23	3.56	3.96	87.10
	300	100	10.45	13.48	2.43	2.72	78.40
		1000	7.43	10.41	3.07	3.49	76.00
		10000	5.28	7.63	4.02	3.96	80.30
2	100	100	7.22	9.96	4.20	4.30	82.80
		1000	4.21	6.13	4.33	4.93	77.50
		10000	1.08	3.08	4.43	4.62	78.00
	200	100	11.22	14.55	3.04	2.83	78.20
		1000	8.69	12.24	3.75	3.65	75.10
		10000	6.77	10.29	4.53	4.50	78.90
	300	100	13.29	16.29	1.89	2.20	81.80
		1000	10.12	14.63	2.43	2.31	80.40
		10000	9.24	13.22	2.66	2.72	77.30

Table S4: Simulation results for comparing SMA with FB in Examples 3 and 4.

			AOR (%)		SD (%)		WP(%)
Example	n	p	FB	SMA	FB	SMA	FB
3	100	100	10.45	12.33	4.30	4.20	76.30
		1000	9.77	11.81	3.89	4.05	75.10
		10000	10.23	11.96	4.05	4.00	68.60
	200	100	12.48	13.79	3.23	3.19	65.40
		1000	11.27	12.71	3.45	3.30	70.60
		10000	10.70	12.25	3.02	3.25	66.00
	300	100	12.89	14.68	2.85	2.88	75.40
		1000	10.56	13.46	3.17	3.30	78.20
		10000	10.44	12.61	3.21	3.33	67.10
4	100	100	8.67	9.05	2.78	2.86	65.20
		1000	6.54	7.52	2.32	2.41	63.60
		10000	5.88	7.04	2.09	2.13	58.40
	200	100	9.56	10.78	2.76	2.78	60.60
		1000	10.05	9.13	2.25	2.30	46.80
		10000	9.16	8.10	1.77	1.88	45.90
	300	100	11.65	12.46	2.67	2.88	65.80
		1000	9.78	10.27	2.56	2.61	56.70
		10000	8.12	9.14	2.05	2.16	58.30

Table S5: Simulation results for the average number of steps reached via our stopping rule in Examples 1 and 2 with $d_0 = 5$ and $d_0 = 3$, respectively.

n	p	Example 1	Example 2
100	100	24.2	22.0
	1000	20.4	18.2
	10000	17.8	16.4
200	100	22.1	20.1
	1000	18.3	17.5
	10000	14.4	13.9
300	100	20.6	18.9
	1000	15.3	17.0
	10000	12.8	11.7

Table S6: Simulation results for the average number of steps reached via our stopping rule in Examples S1 and S2 with $d_0 = 10$ and $d_0 = 20$, respectively.

		Example S1		Example S2	
n	p	$d_0 = 10$	$d_0 = 20$	$d_0 = 10$	$d_0 = 20$
100	100	26.0	33.5	32.0	37.0
	1000	22.0	29.0	28.3	31.6
	10000	18.9	26.5	26.2	28.8
200	100	24.6	30.1	29.7	32.5
	1000	19.5	27.5	28.4	30.6
	10000	16.8	27.5	25.8	27.1
300	100	25.3	20.9	14.7	13.0
	1000	18.2	26.5	27.8	29.1
	10000	15.4	27.2	24.0	27.2

Figure S1. AOR of SMA^k versus the number of sequential steps k in Example S1.

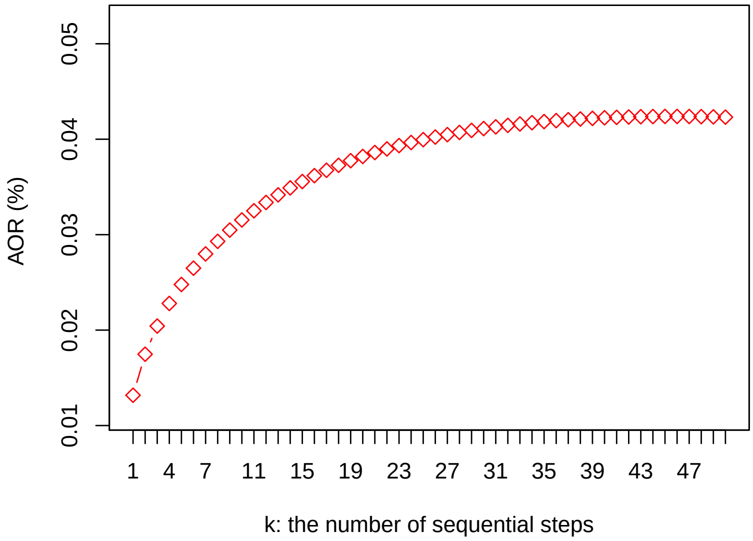
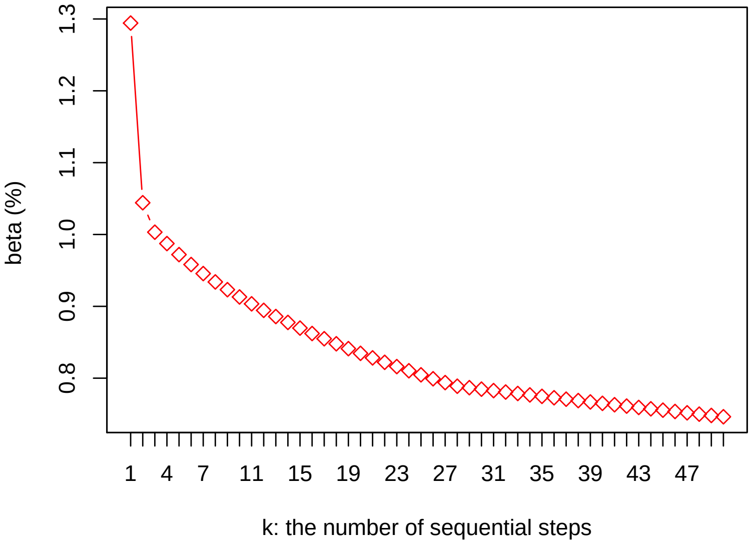


Figure S2. $\max_j |\widehat{\beta}_{kj}|$ of SMA^k versus the number of sequential steps k in Example S1.


beta (%)
k: the number of sequential steps