Statistica Sinica **28** (2018), 449-469
doi:https://doi.org/10.5705/ss.202016.0122

SEQUENTIAL MODEL AVERAGING FOR HIGH DIMENSIONAL LINEAR REGRESSION MODELS

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: In high-dimensional data analysis, we propose a sequential model averaging (SMA) method to make accurate and stable predictions. Specifically, we introduce a hybrid approach that combines a sequential screening process with a model averaging algorithm, where the weight of each model is determined by its Bayesian information (BIC) score (Schwarz (1978); Chen and Chen (2008)). The sequential technique makes SMA computationally feasible with high-dimensional data, because the averaging process assures the prediction's accuracy and stability. Results show that SMA not only yields a good model, but also mitigates overfitting. We demonstrate that SMA provides consistent estimators for the regression coefficients and yields reliable predictions under mild conditions. Simulations and empirical examples are presented to illustrate the usefulness of the proposed method.

Key words and phrases: Forward regression, sequential model averaging, sequential screening, univariate model averaging.

1. Introduction

In regression analysis, parameter estimation and variable selection play important roles in the process of making accurate and reliable predictions. To this end, various shrinkage methods have been proposed under a fixed dimension setting; see, for example, the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996), smoothly clipped absolute deviation operator (SCAD) of Fan and Li (2001), adaptive LASSO of Zou (2006) and Zhang and Lu (2007). These methods are further extended to the case of a diverging number of parameters under the constraint that the predictor dimension (p) is no larger than the sample size (n); see, for example, Fan and Peng (2004), Huang, Ma and Zhang (2007), and Zou and Zhang (2009). For the case where the number of predictors (p) exceeds the sample size (n), several variable screening methods have recently been developed. These methods for ultra-high dimensional models (say

$\log(p) = O(n^a)$ for some $a > 0$), include sure independence screening (Fan and Lv (2008); Fan and Song (2010))(SIS), forward regression (Wang (2009))(FR) and distance correlation learning (Li, Zhong and Zhu (2012)). These screening methods can reduce the data dimension from ultra-high to low, or even fixed, so that classical methods are applicable.

Shrinkage and/or screening methods are useful for high-dimensional data analysis. By identifying sparse solutions for the regression coefficients, model interpretability and forecasting accuracy can be improved, provided the true sparse structure is correctly identified. Otherwise, estimation and prediction results can be biased. Hence, one typically requires that either the sample size or the signal-to-noise ratio be sufficiently large. This is particularly true if the predictor dimension is high; see, for example, the simulation experiments reported in Zhang and Lu (2007), Fan and Lv (2008), Wang (2009), Zou and Zhang (2009), and Fan and Song (2010). In practice, sample size may be limited and the signal-to-noise ratio weak due to complex data generating mechanisms. Then, estimation accuracy is unreliable for essentially any variable selection method (Shao (1997); Yang (2005); Leeb and Pötscher (2008)) in finite samples, and the resulting forecasts can be unstable and inaccurate.

Model averaging approaches are commonly used to improve predictive performance. Instead of employing a single selected best model to make predictions, these techniques average all possible candidate models with suitable weights. Such methods include, but are not limited to, Akaike information criterion (AIC) model averaging (Akaike (1979); Burnham and Anderson (2002)), Bayesian information criterion (BIC) model averaging (Buckland, Burnham and Augustin (1997); Hoeting et al. (1999)), Mallows C_p model averaging (Hansen (2007); Wan, Zhang and Zou (2010)) and Jackknife model averaging (Racine and Hansen (2012); Zhang, Wan and Zou (2013); Ando and Li (2014)).

Commonly, averaging methods are designed for predictor dimensions no bigger than the sample size, as in low-dimensional data with $p = o(n)$, and cannot be directly applied to a model with ultra-high dimensional predictors. In practice, classical results about risk efficiency (see, e.g., Li (1987)) may not be valid because of the difficulty in finding an optimal rate of convergence that can serve as a universal lower bound for the lowest risk among all weight choices; see Theorem 1 of Ando and Li (2014) and the accompanying discussion. To resolve this, Ando and Li (2014) suggested sorting the predictors into groups according to their marginal correlation with the response, then averaging over this small number of groups. Weights are determined by a delete-one cross-validation procedure (we refer to it as MCV for the model-averaging CV). The authors demonstrated that MCV is computationally feasible for ultra-high dimensional predictors and established its risk efficiency.

Thus, LASSO, SIS, and FR can handle variable selection and parameter estimation simultaneously, while they focus on searching for a single best model to improve prediction accuracy and model interpretability and discount some model certainty. These methods require the true model to be sparse in order to attain model selection consistency. With sequential screening, SIS and FR can cope with high-dimensional data. Model averaging accounts for model uncertainty and makes better predictions, but the commonly used averaging methods are not directly applicable to high-dimensional data when p is much larger than n. Accordingly, we propose a sequential model averaging (SMA) approach for predictions that combines a sequential screening process and a model averaging algorithm. This approach leverages the computational convenience of screening procedures (Fan and Lv (2008); Wang (2009)) and the forecasting reliability of model averaging methods (Claeskens and Hjort (2008)). Furthermore, it does not require sparsity of regression coefficients.

SMA is implemented in a sequential manner so that in each step the candidate models with size one are considered. This is computationally feasible even when the predictor dimension is ultra-high. The response vector in each step of SMA is updated by the residual calculated from the previous step. The larger weights determined by BIC scores (Schwarz (1978);Chen and Chen (2008)) can be sequentially assigned to more relevant predictors. Accordingly, SMA conducts both variable screening and model averaging. It yields consistent estimators of regression coefficients even when the predictor dimension is much larger than the sample size, even if it is ultra-high dimensional.

The rest of this article is organized as follows. Section 2 introduces SMA and then investigates its theoretical properties. Applications to simulations and data are reported in Section 3. The article concludes with a short discussion in Section 4. Technical details are relegated to the supplementary materials.

2. Sequential Model Averaging

In this section, we review the classical Bayesian model averaging procedure, and then extend it to high-dimensional data. Since it is not an optimal procedure, we incorporate sequential screening approach into the univariate model averaging process, which results in the sequential model averaging algorithm. Then we study its theoretical properties.

2.1. Model averaging

Let (Y_i, X_i) for $i = 1, \cdots, n$ be n independent and identically distributed random vectors, where $Y_i \in \mathbb{R}^1$ $(1 \le i \le n)$ is the response collected from the i-th subject and $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor. We take the Y_i and covariates X_i to be standardized; $E(Y_i) = E(X_{ij}) = 0$ and $\mathrm{var}(Y_i) = \mathrm{var}(X_{ij}) = 1$ for all $1 \le j \le p$. The correlation of X_{ij_1} and X_{ij_2} is labeled as $\sigma_{j_1 j_2}$ for any $j_1 \ne j_2$. We consider a response vector $\mathbb{Y} = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^n$ and a design matrix $\mathbb{X} = (X_1, \cdots, X_n)^\top \in \mathbb{R}^{n \times p}$. For j with $1 \le j \le p$, $\mathbb{X}_j = (X_{1j}, \cdots, X_{nj})^\top \in \mathbb{R}^n$ denotes the j-th column of $\mathbb{X}$. Let $\mathcal{M} = \{j_1, \cdots, j_d\}$ and $\mathcal{M}_F = \{1, 2, \cdots, p\}$ represent the candidate model with explanatory variables $X_{ij_1}, \cdots, X_{ij_d}$ and the full model, respectively. Accordingly, $\mathbb{X}_{(\mathcal{M})} = (\mathbb{X}_j : j \in \mathcal{M}) \in \mathbb{R}^{n \times |\mathcal{M}|}$ is the design matrix associated with model $\mathcal{M}$, where $|\mathcal{M}|$ denotes the size of the candidate model.

For any candidate model $\mathcal{M} \subset \mathcal{M}_F$ with $|\mathcal{M}| \le n$, we establish the relationship between the response and explanatory variables via a linear regression, $\mathbb{Y} = \mathbb{X}_{(\mathcal{M})}\beta_{(\mathcal{M})} + \varepsilon$, where $\beta_{(\mathcal{M})} = (\beta_j : j \in \mathcal{M}) \in \mathbb{R}^{|\mathcal{M}|}$, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^\top$, and the ε_is are i.i.d. normal with mean zero and finite variance σ^2. The ordinary least squares (OLS) estimator of $\beta_{(\mathcal{M})}$ is $\hat\beta_{(\mathcal{M})} = (\mathbb{X}_{(\mathcal{M})}^\top \mathbb{X}_{(\mathcal{M})})^{-1}(\mathbb{X}_{(\mathcal{M})}^\top \mathbb{Y}) \in \mathbb{R}^{|\mathcal{M}|}$. Given candidate model fittings, one can employ Schwarz's (1978) Bayesian information criterion (BIC)

$$\mathrm{BIC}^*_{\mathcal{M}} = n \log \|\mathbb{Y} - \mathbb{X}_{(\mathcal{M})}\hat\beta_{(\mathcal{M})}\|^2 + |\mathcal{M}| \times \log n, \tag{2.1}$$

to select the best model, where $\|\cdot\|$ stands for the usual L_2 norm. Based on BIC scores, we can assign each candidate model with an appropriate weight, leading to a BIC model averaging estimator

$$\hat\beta_B = \sum_{\mathcal{M} \subset \mathcal{M}_F} w_{\mathcal{M}} \hat\beta_{\mathcal{M}}.$$

Specifically, $w_{\mathcal{M}} = \exp(-\mathrm{BIC}^*_{\mathcal{M}}/2)\big\{\sum_{\mathcal{M}_* \subset \mathcal{M}_F} \exp(-\mathrm{BIC}^*_{\mathcal{M}_*}/2)\big\}^{-1}$ and $\hat\beta_{\mathcal{M}}$ is a p-dimensional vector such that the coefficients associated with $\mathcal{M}$ are given by $\hat\beta_{(\mathcal{M})}$, with the rest set to 0. More detailed discussion of the BIC model averaging approach can be found in Hoeting et al. (1999), Claeskens and Hjort (2008), and Hastie, Tibshirani and Friedman (2009).

An important feature of the BIC model averaging estimator is its stability, due to utilizing the information of every candidate model. Many variable selection methods that aim for a single best model are unstable, particularly those designed for consistent model selection (Shao (1997); Yang (2005); Leeb and Pötscher (2008)). Although BIC model averaging has nice properties, Burnham and Anderson (2002) found it difficult to implement when the number of variables is large, as the number of candidate models increases exponentially as the predictor dimensions grows large. As pointed out by Chen and Chen (2008), the classical Bayesian information criterion (2.1) is too liberal for model selection when p is large. They proposed the ultra-high dimensional Bayesian information criterion

$$\mathrm{BIC}_{\mathcal{M}} = n \log \|\mathbb{Y} - \mathbb{X}_{(\mathcal{M})}\hat{\beta}_{(\mathcal{M})}\|^2 + |\mathcal{M}| \times (\log n + 2 \log p). \tag{2.2}$$

Under some mild assumptions, they showed that this BIC criterion is selection consistent for high-dimensional data. This motivates us to employ it in our averaging estimators.

2.2. Univariate model averaging

As traditional model averaging approach for high dimensional data can be computationally intractable, we consider only candidate models with $|\mathcal{M}| = 1$. We define the Univariate Model Averaging (UMA) estimator as

$$\hat{\beta}_U = \sum_{|\mathcal{M}|\le 1} w_{\mathcal{M}}^U \hat{\beta}_{\mathcal{M}}, \tag{2.3}$$

where $w_{\mathcal{M}}^U = \exp(-\mathrm{BIC}_{\mathcal{M}}/2)\big\{\sum_{|\mathcal{M}_*|\le 1} \exp(-\mathrm{BIC}_{\mathcal{M}_*}/2)\big\}^{-1}$. Since we only average over candidate models of size one, there are only p univariate regression models to be estimated. The UMA estimator is a smooth function of the data, and thus, it is stable. We adopt Chen and Chen's (2008) BIC score designed for high-dimensional regression models to construct the weight for each candidate model in (2.3).

To demonstrate theoretical properties of UMA, we need some notation and conditions. Take $\rho_j = \mathrm{cov}(X_{ij}, Y_i)$, so that $\rho_{(1)}^2 \ge \rho_{(2)}^2 \ge \cdots \ge \rho_{(p)}^2$. We use $\hat{\rho}_{(j)}^2$ and $\hat{\rho}_j^2$ to represent their corresponding estimators. We have the following conditions.

(C1) There are constants $\nu > 0$ and $0 \le \alpha < 1$ such that $\log p \le \nu n^{\alpha}$;

(C2) There exist some positive constants C_1 and C_2, free of n and p, such that for any positive constant $\delta > 0$, $P(|X_{ij}| > \delta) \le C_1 \exp(-C_2\delta^2)$ and $P(|\varepsilon_i| > \delta) \le C_1 \exp(-C_2\delta^2)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, p$;

(C3) There is a constant $d_1 > 0$ such that $\max_{j_1 \ne j_2}\{|\sigma_{j_1 j_2}|, \rho_{(1)}^2\} < d_1 < 1$.

Condition (C1) allows the predictor dimension p to be ultra-high (Fan and Lv

(2008)). Condition (C2) is satisfied if (X_i, ε_i) is multivariate normal; see, for example, Li, Zhong and Zhu (2012) and Wang (2012). Condition (C3) requires that no predictor be perfectly correlated with the response vector, and that no two predictors can be perfectly correlated with each other (Kalisch and Bühlmann (2007)). These conditions are mild and make our results easily applicable.

Theorem 1. *If* (C1)–(C3) *hold and* $\rho^2_{(1)} - \rho^2_{(2)} > d_2 > 0$ *for some constant* d_2*, then* $w^U_{\max} \rightarrow_p 1$ *as* $n \rightarrow \infty$*, where* $w^U_{\max}$ *is the weight assigned by UMA to the predictor that has the largest absolute coefficient of correlation with the response.*

Accordingly, the information used for estimation and prediction is mainly from the largest correlated predictor. The assumption that $\rho^2_{(1)} - \rho^2_{(2)} > d_2 > 0$ for some constant d_2 can be appropriately modified. For example, if $\rho^2_{(1)} = \rho^2_{(2)} + o(1) > \rho^2_{(3)}$, then the UMA approach assigns nearly equal weights to the first two predictors that have the largest absolute correlations with the response, while others have negligible effect. The resulting prediction via UMA may of cause not be accurate since it only uses a small portion of the available information.

2.3. Sequential model averaging

To improve forecasting accuracy, one has to take into account the information from other relevant predictors. This can be done by enlarging the candidate model size from 1 to 2, in which case the number of candidate models is p^2. With high-dimensional data, the difference between p and p^2 may carry a heavy computational burden. This motivates us to utilize forward regression to update the coefficient estimate sequentially.

Specifically, we sequentially update the coefficient estimate of the UMA algorithm at each step. While updating the response vector to the residual calculated from the previous step. Thus, the effects of heavily weighted predictors in the previous steps can be substantially reduced, allowing other relevant predictors to contribute more to subsequent parameter estimates. This hybrid approach between a sequential method and an averaging method retains the estimation stability of UMA, and achieves computational feasibility. We refer to this procedure as sequential model averaging (SMA).

Assume that the SMA algorithm consists of K sequential steps; the selection of K is discussed in Remark 4 at the end of Subsection 2.5. Let $\mathbb{Y}_1 = \mathbb{Y}$ be the initial response vector, and $\mathbb{Y}_k$ be the response vector used in the k-th step, $1 \leq k \leq K$. For the given response $\mathbb{Y}_k$ and explanatory variables $\mathbb{X}_j$ ($j = 1, \cdots, p$), we fit the univariate regression model and obtain the OLS estimator

$\hat{\beta}_{kj} = (\mathbb{X}_j^\top \mathbb{X}_j)^{-1}(\mathbb{X}_j^\top \mathbb{Y}_k)$. Thus, $\hat{\beta}_{kj}$ is the OLS estimator obtained in the k-th sequential step via the j-th explanatory variable only. By (2.2), the corresponding high-dimensional BIC score is given by

$$\begin{aligned} \mathrm{BIC}_{kj} &= n \log \|\mathbb{Y}_k - \mathbb{X}_j \hat{\beta}_{kj}\|^2 + \log n + 2 \log p \\ &= n \log \left\{ \|\mathbb{Y}_k\|^2 (1 - \hat{\rho}_{kj}^2) \right\} + \log n + 2 \log p, \end{aligned}$$

where $\hat{\rho}_{kj}^2 = \|\mathbb{Y}_k\|^{-2} \mathbb{Y}_k^\top H_j \mathbb{Y}_k$ is the squared correlation coefficient between $\mathbb{X}_j$ and $\mathbb{Y}_k$, and $H_j = \|\mathbb{X}_j\|^{-2} \mathbb{X}_j \mathbb{X}_j^\top$. We also fit the null model, which leads to a residual sum of squares, $\|\mathbb{Y}_k\|^2$, and the resulting BIC score $\mathrm{BIC}_{k0} = n \log(\|\mathbb{Y}_k\|^2)$. Including the null model at every step is crucial for the SMA procedure. By doing so, as long as the weight of the null model is less than 1, some information of the response can still be explained by the covariates. We adopt the idea of classical BIC model averaging algorithm and define the averaging weight for each candidate model j, $0 \le j \le p$, as

$$\hat{w}_{kj} = \exp\Big(-\frac{1}{2} \mathrm{BIC}_{kj} \Big) \left[\sum_{j'=0}^{p} \exp\Big(-\frac{1}{2} \mathrm{BIC}_{kj'} \Big) \right]^{-1}.$$

After algebraic simplification, one can verify that

$$\begin{aligned} \hat{w}_{k0} &= \sqrt{n} p \Bigg[\sum_{j'=1}^{p} \Big(1 - \hat{\rho}_{kj'}^2 \Big)^{-n/2} + \sqrt{n} p \Bigg]^{-1}, \\ \text{and} \quad \hat{w}_{kj} &= \Big(1 - \hat{\rho}_{kj}^2 \Big)^{-n/2} \Bigg[\sum_{j'=1}^{p} \Big(1 - \hat{\rho}_{kj'}^2 \Big)^{-n/2} + \sqrt{n} p \Bigg]^{-1} \quad \text{for} \quad 1 \le j \le p. \end{aligned}$$

This leads to a coefficient vector $\hat{\beta}^{(k)} = (\hat{w}_{k1} \hat{\beta}_{k1}, \cdots, \hat{w}_{kp} \hat{\beta}_{kp})^\top \in \mathbb{R}^p$. Subsequently, the response $\mathbb{Y}_k$ is updated to $\mathbb{Y}_{k+1} = \mathbb{Y}_k - \mathbb{X} \hat{\beta}^{(k)}$. After completing K iteration steps, we obtain the SMA estimator, $\hat{\beta}^K = \sum_{k=1}^{K} \hat{\beta}^{(k)}$. Thus, for (X^*, Y^*), an independent copy of (X_i, Y_i) for some $1 \le i \le n$, we can predict the value of Y^* by $\hat{Y}^* = X^{*\top} \hat{\beta}^K$.

Remark 1. The weight $\hat{\omega}_{kj}$ for the j-th predictor in the k-th iteration step is data driven and closely related to $\hat{\rho}_{kj}$. To study the properties of $\hat{\omega}_{kj}$, we need to control the magnitude of $\hat{\rho}_{kj}$ for every k. We define ρ_{kj}, the population version of $\hat{\rho}_{kj}$, in a sequential manner. When $k = 1$, we have $\rho_{1j} = \rho_j$, and then take

$$\begin{aligned} w_{10} &= \sqrt{n} p \Bigg[\sum_{j'=1}^{p} \Big(1 - \rho_{1j'}^2 \Big)^{-n/2} + \sqrt{n} p \Bigg]^{-1}, \\ \text{and} \quad w_{1j} &= \Big(1 - \rho_{1j}^2 \Big)^{-n/2} \Bigg[\sum_{j'=1}^{p} \Big(1 - \rho_{1j'}^2 \Big)^{-n/2} + \sqrt{n} p \Bigg]^{-1} \quad \text{for} \quad 1 \le j \le p. \end{aligned}$$

Table 1. Simulation results of comparisons for Examples 1 and 2.

			AOR (%)					SD (%)					WP(%)			
Example	n	p	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg
1	100	100	2.03	1.50	2.79	4.52	6.17	7.39	4.40	4.07	3.87	3.72	76.00	91.80	86.70	75.60
		1,000	−4.16	0.05	0.43	1.06	2.43	9.75	4.14	1.74	4.02	3.73	77.00	79.60	76.20	65.90
		10,000	−11.01	−0.13	−0.04	0.18	0.57	9.71	3.25	1.10	3.21	2.74	87.90	52.10	50.30	54.60
	200	100	8.05	7.13	4.24	8.43	10.96	4.56	5.20	4.38	3.44	3.29	84.20	89.50	95.50	72.60
		1,000	5.83	4.42	1.66	6.12	7.12	5.65	5.13	1.21	4.01	3.92	66.00	87.20	96.20	63.40
		10,000	2.18	2.34	−0.02	2.29	4.23	8.06	4.53	1.17	3.86	3.96	61.50	84.90	92.80	55.70
	300	100	11.47	11.46	4.62	10.56	13.48	4.14	4.43	4.03	3.12	2.72	78.30	77.20	99.50	73.00
		1,000	9.25	8.65	2.63	8.33	10.41	4.29	5.03	1.34	3.58	3.49	72.10	79.30	99.40	69.30
		10,000	7.26	6.05	0.17	6.34	7.63	4.88	5.22	1.29	4.07	3.96	61.50	81.20	98.60	57.60
2	100	100	8.76	6.14	3.85	8.44	9.96	7.27	6.86	4.77	5.21	4.30	57.90	79.40	94.30	75.60
		1,000	3.52	3.08	0.69	5.21	6.13	11.70	6.42	1.98	5.99	4.93	51.30	75.30	89.50	68.50
		10,000	−4.26	0.94	−0.02	2.14	3.08	14.02	5.63	1.10	5.46	4.62	69.10	71.40	73.90	58.40
	200	100	12.90	12.83	4.64	12.88	14.55	2.85	3.67	4.48	3.02	2.83	83.20	83.90	99.60	80.40
		1,000	11.99	11.02	1.81	10.89	12.24	4.18	5.04	1.44	3.87	3.65	64.20	72.40	99.70	68.20
		10,000	10.85	9.10	0.02	9.77	10.29	6.13	6.25	1.24	4.90	4.50	45.40	62.00	98.50	58.10
	300	100	14.05	14.90	6.11	13.58	16.29	2.27	2.68	3.85	2.01	2.20	91.60	84.00	99.90	79.40
		1,000	13.34	13.54	2.58	12.89	14.63	2.11	2.63	1.69	2.56	2.31	88.20	86.70	100.00	69.00
		10,000	12.83	12.58	0.16	12.77	13.22	2.49	3.20	1.38	2.32	2.72	70.30	72.60	100.00	60.40

Using the fact that $n^{-1}\|\mathbb{X}_j\|^2 \rightarrow_p 1$ and $n^{-1}\|\mathbb{Y}\|^2 \rightarrow_p 1$, we obtain that $\hat{\beta}_{1j} \rightarrow_p \rho_j$. As a consequence, the exact response is taken as $\widetilde{\mathbb{Y}}_2 = \mathbb{Y}_1 - \mathbb{X}\beta^{(1)}$, where $\beta^{(1)} = (w_{11}\beta_{11}, \cdots, w_{1p}\beta_{1p})^\top$ with $\beta_{1j} = \rho_{1j}$. At the k-th sequential step, if $\widetilde{\mathbb{Y}}_{k+1}$ is well defined, we then define $\rho_{(k+1)j} = \mathrm{corr}(\widetilde{\mathbb{Y}}_{k+1}, \mathbb{X}_j)$ and $\beta_{(k+1)j} = \mathrm{cov}(\widetilde{\mathbb{Y}}_{k+1}, \mathbb{X}_j)$; this sequential procedure can be repeated for each k. Consequently, our technical conditions are imposed on ρ_{kj} rather than on $\hat{\rho}_{kj}$; e.g., see Condition (C4).

Remark 2. SMA is closely related to the method of boosting (see Friedman, Hastie and Tibshirani (2000); Bühlmann and Yu (2003, 2006)), but SMA is more stable than boosting; see the simulation results in Tables 1, 2 and 4. Further, the weight assigned to each covariate in the SMA algorithm is solely data driven.

2.4. Fitting capability

In classical sequential learning algorithms, forward regression has the capability to reduce the residual sum of squares monotonically. This allows forward regression to capture the true regression relationship in a very limited number of steps (Wang (2009)). Our results show that SMA possesses this desirable feature.

Theorem 2. *For $k \geq 1$, we have $\|\mathbb{Y}_k\|^2 - \|\mathbb{Y}_{k+1}\|^2 \geq \|\mathbb{Y}_k\|^2 \sum_{j=1}^{p} \hat{w}_{kj}\hat{\rho}_{kj}^2$.*

Thus, SMA reduces the residual sum of squares in each iteration step, similarly to forward regression. To go further, we need the following condition.

(C4) Let $\rho_{k(1)}^2 \geq \cdots \geq \rho_{k(p)}^2$ denote the ordered statistics of $\{\rho_{kj}^2 : 1 \leq j \leq p\}$ for any $k = 1, \cdots, K$. Assume that $\sup_{1\leq k\leq K} \rho_{k(1)}^2 \leq d_3 < 1$ for some fixed constant $d_3 > 0$.

Table 2. Simulation results of comparisons for Examples 3 and 4.

			AOR (%)					SD (%)					WP(%)			
Example	n	p	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg
3	100	100	6.91	5.18	10.84	12.06	12.33	4.90	5.39	4.54	4.43	4.30	94.80	96.40	67.30	74.20
		1,000	3.60	1.17	8.75	11.43	11.81	4.71	4.38	4.47	4.17	4.05	98.30	98.80	78.90	65.70
		10,000	1.22	−1.39	7.24	10.89	11.96	4.87	4.73	4.83	4.07	4.00	99.30	99.70	89.00	58.30
	200	100	10.46	10.26	13.66	12.64	13.79	3.86	4.12	3.02	3.28	3.19	94.80	91.80	53.40	72.30
		1,000	7.18	6.57	11.34	11.78	12.71	3.99	4.05	3.08	3.08	3.30	98.50	96.80	71.70	60.40
		10,000	4.97	4.16	10.07	11.08	12.25	3.58	3.56	3.32	3.12	3.25	99.30	98.60	78.50	56.70
	300	100	13.28	13.31	14.57	12.79	14.68	3.27	3.34	2.60	2.51	2.88	79.00	76.90	53.40	78.00
		1,000	9.86	9.76	12.95	12.38	13.46	3.98	4.13	2.77	3.37	3.30	94.20	91.50	60.00	66.90
		10,000	7.23	6.97	11.42	12.06	12.61	3.76	3.70	2.97	3.45	3.33	97.80	96.30	70.90	60.50
4	100	100	5.15	2.52	7.85	6.88	9.05	3.42	4.09	4.72	2.79	2.86	93.80	97.20	59.50	76.30
		1,000	4.27	0.59	6.92	6.76	7.52	3.82	3.28	4.82	2.80	2.41	87.80	98.80	53.90	67.70
		10,000	3.47	−0.38	5.91	6.46	7.04	4.05	3.19	5.07	2.28	2.13	84.10	99.10	59.40	55.50
	200	100	6.89	8.62	8.97	8.98	10.78	3.05	4.62	3.68	2.67	2.78	97.20	75.10	72.20	78.50
		1,000	6.14	6.16	9.27	7.68	9.13	2.61	4.51	3.51	2.40	2.30	93.40	82.00	48.70	68.00
		10,000	5.53	3.70	8.23	7.25	8.10	2.58	3.83	3.64	2.08	1.88	90.20	91.80	49.60	58.70
	300	100	8.28	12.34	9.99	10.67	12.46	3.30	4.25	3.64	3.02	2.88	96.60	55.50	88.70	88.70
		1,000	7.04	9.68	9.84	9.23	10.27	2.78	4.41	3.14	2.46	2.61	96.10	63.40	57.30	79.20
		10,000	6.47	7.96	9.82	8.45	9.14	2.45	4.36	3.19	2.34	2.16	94.40	68.10	41.90	56.90

Table 3. Simulation results of $\sum_{m=1}^{1,000} \|\hat{\beta}_{[m]} - \beta_0\|/1,000$ for Examples 1-4.

		Example 1	Example 2	Example 3	Example 4
n	p	Mean	Mean	Mean	Mean
200	100	0.225	0.634	0.888	1.246
	1,000	0.274	0.752	0.992	1.387
	10,000	0.303	0.854	1.029	1.446
400	100	0.163	0.444	0.798	1.033
	1,000	0.200	0.546	0.895	1.207
	10,000	0.227	0.623	0.969	1.324
800	100	0.106	0.223	0.662	0.767
	1,000	0.127	0.271	0.731	0.913
	10,000	0.146	0.344	0.810	1.002

Condition (C4) is fairly mild and similar to condition (C3). It requires that no predictor can be perfectly correlated with the response vector at any sequential step.

Theorem 3. *If* (C1), (C2), *and* (C4) *hold, for those values of* k *satisfying* $\rho^2_{k(1)} - \rho^2_{k(2)} > d_4 > 0$ *with some constant* d_4, *we have* $\|\mathbb{Y}_k\|^2 - \|\mathbb{Y}_{k+1}\|^2 \geq \|\mathbb{Y}_k\|^2 \hat{\rho}^2_{k(1)} \hat{\omega}_{k(1)}$, *where* $\hat{\rho}^2_{k(1)}$ *is the estimator of* $\rho^2_{k(1)}$. *In addition,* $\hat{\omega}_{k(1)} \rightarrow_p 1$ *as* $n \rightarrow \infty$.

Remark 3. One can verify that, after fitting $\mathbb{Y}_k$ by $\mathbb{X}_j$ in the k-th step, the incremental reduction in the residual sum of squares is $\|\mathbb{Y}_k\|^2 \hat{\rho}^2_{kj}$. This suggests that the reduction in the residual sum of squares in the k-th step can never exceed $\|\mathbb{Y}_k\|^2 \hat{\rho}^2_{k(1)}$. This, together with Theorem 3, indicates that the amount of reduction achieved by SMA can be arbitrarily close to this upper bound asymp-

Table 4. Data analysis results for 14 different datasets. Each row corresponds to a dataset and the industry IDs 1 = online retailing, 2 = professional training, 3 = online recruiting, 4 = microblogging, 5 = mortgage, 6 = travel planning, and 7 = real-estate advertising. The sample and covariate sizes in boldface indicate $n < p$.

Industry	Size		AOR (%)					SD (%)					WP(%)			
ID	n	p	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg	SMA	SIS	FR	MCV	Bstg
1	**150**	**296**	3.35	−0.15	4.11	7.65	12.75	5.78	2.82	11.43	4.34	4.21	97.80	98.90	91.70	85.60
	421	367	1.02	−0.04	−17.69	1.25	2.98	2.82	0.71	64.80	2.65	2.50	84.80	93.30	76.80	94.50
	438	377	9.61	2.08	−15.65	9.64	10.45	3.08	7.63	53.79	2.88	3.18	80.80	96.30	97.10	58.70
	497	490	8.87	2.33	−9.09	8.46	14.35	4.50	5.01	49.39	4.98	5.05	87.60	94.20	87.60	83.60
	1,438	979	6.87	1.51	−11.58	9.32	11.29	8.24	4.03	38.94	3.02	2.97	96.80	97.30	92.50	78.40
	1,277	808	7.12	1.02	0.25	3.78	13.23	4.63	2.71	40.78	1.90	2.80	91.20	98.70	73.30	92.70
	1,438	979	9.38	1.90	−18.18	5.90	11.00	1.51	3.98	76.96	2.03	0.97	92.20	93.70	96.70	95.40
2	788	487	12.48	3.55	−18.52	11.87	13.33	1.98	5.85	65.24	2.05	1.81	83.20	94.60	94.20	75.40
3	883	517	3.61	2.42	−39.05	7.90	13.97	78.87	4.87	199.32	1.66	1.68	81.20	96.10	96.70	89.50
	1,352	820	11.44	2.95	−3.59	15.28	18.46	1.36	5.37	67.59	1.90	1.74	99.80	99.70	94.90	79.40
4	**154**	**275**	40.24	18.46	6.83	38.49	42.21	32.05	24.90	25.86	15.57	13.46	73.20	88.80	96.00	68.00
5	**257**	**312**	−0.32	−0.42	−7.75	1.23	2.41	3.22	1.27	29.06	2.56	2.93	83.90	88.30	68.60	89.40
6	1,534	788	23.76	7.26	3.07	18.78	26.92	1.75	10.37	37.01	1.44	1.57	97.30	99.30	85.60	92.00
7	1,545	695	11.49	3.26	−17.31	11.67	13.28	1.93	5.47	54.33	1.49	1.66	98.60	98.00	93.90	76.30

totically since $\hat{\omega}_{k(1)} \to_p 1$ as $n \to \infty$ for any given k. This upper bound can be achieved by setting $\hat{\omega}_{k(1)} = 1$ for any $k \geq 1$. Accordingly, this yields the FR procedure. Although FR achieves the upper bound, it suffers from a nontrivial overfitting effect.

2.5. Overfitting resistance

By Theorem 3, we know that SMA has good fitting capability, but only under the assumption that $\rho^2_{k(1)}$ is not too small. If $\rho^2_{k(1)}$ is very small, we believe that further reduction in the residual sum of squares is not desirable. It would be primarily due to overfitting. This is a serious drawback, from which forward regression suffers (Wang (2009)), but the overfitting effect suffered by SMA is considerably weaker.

Theorem 4. *If* (C1)–(C4) *hold, for those values of* k *satisfying* $\rho^2_{k(1)} = O(n^{-1})$, *we have that* $\big(\|\mathbb{Y}_k\|^2 - \|\mathbb{Y}_{k+1}\|^2\big)\|\mathbb{Y}_k\|^{-2} \leq 2(1 - \hat{w}_{k0})\hat{\rho}^2_{k(1)}$, *where* $\hat{\rho}^2_{k(1)}$ *is the estimator of* $\rho^2_{k(1)}$. *In addition,* $\hat{w}_{k0} \to_p 1$ *as* $n \to \infty$.

If $\mathbb{X}_j$ is the overfitted variable, after fitting $\mathbb{Y}_k$ by $\mathbb{X}_j$ in the k-th step, the resulting overfitting effect is $(\|\mathbb{Y}_k\|^2 - \|\mathbb{Y}_{k+1}\|^2)\|\mathbb{Y}_k\|^{-2} = \hat{\rho}^2_{kj}$. This could be as large as $\hat{\rho}^2_{k(1)}$ if $\mathbb{X}_j$ is the variable most correlated with $\mathbb{Y}_k$. Such an outcome is indeed the case with forward regression. However, by Theorem 4, the overfitting effect suffered by SMA is a smaller order of $\hat{\rho}^2_{k(1)}$ as $\hat{w}_{k0} \to_p 1$. This suggests that the SMA algorithm suffers considerably less overfitting than the forward regression procedure.

Remark 4. There is a close relationship between the reduction of the residual sum of squares and $\hat{w}_{k0}$. For example, if there is still one predictor that contributes valuable information to $\mathbb{Y}_k$ in the k-th sequential step, then the resulting $\hat{w}_{k(1)}$ tends to 1 and $\hat{w}_{k0}$ shrinks to 0. Accordingly, the reduction error is large; see Theorem 3. Otherwise, $\hat{w}_{k0}$ tends to 1 and the reduction error grows small; see Theorem 4. As we can demonstrate that as k gets large, the value of $\hat{w}_{k0}$ increases, the SMA algorithm could be stopped if $(\hat{w}_{(k+1)0} - \hat{w}_{k0})/\hat{w}_{k0} < \delta$ for some small $\delta > 0$. In our numerical experiments, when we set $\delta = 0.001$, the resulting performance is satisfactory.

2.6. Estimation consistency

The accuracy of forecasts relies on parameter estimates, so we study the asymptotic property of $\hat{\beta}^K$. To this end, we assume that the data is generated from a linear regression model, $\mathbb{Y} = \mathbb{X}^\top \beta_0 + \varepsilon$, where $\beta_0 = (\beta_{01}, \cdots, \beta_{0p})^\top \in \mathbb{R}^p$ is the true regression coefficient vector and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^\top \in \mathbb{R}^n$ is the noise vector. As before, we use $\mathcal{M} = \{j_1, \cdots, j_d\}$ to represent any candidate model. And, for any $\mathcal{M}$, we use $\mathcal{M}^c = \{1, 2, \cdots, p\} \backslash \mathcal{M}$ to denote its complement. Let $\beta_{0(\mathcal{M})} = (\beta_{0j} : j \in \mathcal{M}) \in \mathbb{R}^{|\mathcal{M}|}$ be the subvector of β_0 associated to $\mathcal{M}$. Analogously, we define the subvectors $X_{i(\mathcal{M})}$, $\hat{\beta}^K_{(\mathcal{M})}$, and $\hat{\beta}^K_{(\mathcal{M}^c)}$ as well as the submatrices $\mathbb{X}_{(\mathcal{M})}$ and $\mathbb{X}_{(\mathcal{M}^c)}$. Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively, be the largest and smallest eigenvalues of any semi-positive definite matrix A. We need the following conditions.

(C5) There exists a sequence of $\mathcal{M}_n$ such that

(C5.1) $|\mathcal{M}_n| \to \infty$ and $|\mathcal{M}_n|/n \to 0$ as $n \to \infty$;

(C5.2) $|\mathcal{M}_n^c| \|\beta_{0(\mathcal{M}_n^c)}\|^2 \to 0$ as $K \to \infty$ and $n \to \infty$;

(C6) As $n \to \infty$, there exist two constants $0 < \tau_{\min} < \tau_{\max} < \infty$ such that $\tau_{\min} < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) < \tau_{\max}$, where $\Sigma = \mathrm{cov}(X_i)$.

Condition (C5.1) specifies a sequence of models $\mathcal{M}_n$, whose size diverges. This allows the regression coefficients to be estimated with less bias, but is not possible if its size is larger than the sample size. Condition (C5.2) requires that $\|\beta_{0(\mathcal{M}_n^c)}\| \to 0$ at a sufficiently fast rate, $\beta_{0(\mathcal{M}_n)}$ should be a sufficiently good approximation to β_0 and be estimable, since $|\mathcal{M}_n|/n \to 0$. This condition is typically satisfied if the regression coefficient β is sparse (see, for example, Fan and Lv (2008); Wang (2009)). Finally, Condition (C6) is a Sparse Riesz type condition as used by Zhang and Huang (2008), Wang (2009), Fan and Peng (2004), and Zou and Zhang (2009).

If $\rho_{k(1)}$ is relatively large for any $k > 1$, then Theorem 1 implies that SMA assigns almost all of the weights to only a few of the largest correlated predictors. Thus, the rest of the predictors are suppressed to 0 exponentially. If $\rho_{k(1)}$ is relatively small for any $k > 1$, then Theorem 4 indicates that $\hat{\omega}_{k0} \rightarrow_p 1$ and $\sum_{j=1}^{p} \hat{\omega}_{kj} = 1 - \hat{\omega}_{k0} \rightarrow_p 0$. Accordingly, the weight $\hat{\omega}_{kj}$ go to 0 at a sufficiently fast rate for some j. In this case, we expect to have $|\mathcal{M}_n^c| \sum_{j \in \mathcal{M}_n^c} \hat{\omega}_{kj}^2 \rightarrow 0$ for any $k > 1$ as $n \rightarrow \infty$.

Theorem 5. *If Conditions* (C1), (C2), *and* (C5) *hold, and* $|\mathcal{M}_n^c| \sup_{k>1} \sum_{j \in \mathcal{M}_n^c} \hat{\omega}_{kj}^2 \rightarrow 0$ *as* $n \rightarrow \infty$, *then* $\|\hat{\beta}^K - \beta_0\| \rightarrow_p 0$ *as* $\min\{K, n\} \rightarrow \infty$.

This indicates that SMA yields a consistent estimator of β_0. Accordingly, $\hat{Y}^* = X^{*\top}\hat{\beta}^K$ is a consistent estimator of $X^{*\top}\beta_0$ for the given X^*. So $E(\|X^{*\top}\hat{\beta}^K - X^{*\top}\beta_0\|^2)$ tends to 0 under Conditions (C1), (C2), (C5), and (C6).

3. Simulation Studies

3.1. Simulation examples and settings

We consider simulation examples based on the linear regression model $Y_i = X_i^\top \beta + \sigma\varepsilon_i$, where ε_i is generated from a standard normal distribution, for $i = 1, \cdots, n$. Our findings in data examples (see Table 4) suggest that σ be 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 four examples are given below.

Example 1. We adapt this example from Fan and Lv (2008) assuming that the size of the true model is $d_0 = 5$. For each i, the j-th covariates X_{ij} $(1 \le j \le p)$ were independently generated from $N(0, 1)$. The r-th $(1 \le r \le d_0)$ nonzero true coefficient of β was set equal to $(-1)^{u_r}(a_r + |v_r|)/10$, $a_r = 4\log(n)n^{-1/2}$, where u_r was a binary random variable with $P(u_r = 1) = 0.5$ and v_r was generated from a standard normal.

Example 2. This example is modified from Tibshirani (1996). The covariate vector X_i was generated from a multivariate normal with mean zero and $\text{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$ for $1 \le j_1, j_2 \le p$. Three true non-zero coefficients were set as $\beta_{01} = -0.5$, $\beta_{04} = 1$, and $\beta_{07} = 0.5$, with $\beta_{0j} = 0$ for any $j \notin \{1, 4, 7\}$.

Example 3. This example is adapted from Fan and Lv (2008), where the covariate X_i was generated from a normal distribution with mean zero and $\text{cov}(X_i) = \Sigma = (\sigma_{j_1 j_2}) \in \mathbb{R}^{p \times p}$, where $\sigma_{j_1 j_2} = 0.5$ for $j_1 \neq j_2$. The true non-zero coefficients were set as $\beta_{01} = 0.5$, $\beta_{02} = 0.5$, and $\beta_{03} = 0.5$, while $\beta_{0j} = 0$ for any $3 < j \leq p$.

Example 4. This example is modified from Wang (2009). The first 5 covariates were taken as $X_{ij} = (U_{ij} + V_{ij})/\sqrt{2}$, for $j = 1, \cdots, 5$, and $X_{ij} = (U_{ij} + \sum_{j'=1}^{5} U_{ij'})/2$ for $5 < j \leq p$. In addition, $U_i = (U_{ij}) \in \mathbb{R}^p$ and $V_i = (V_{ij}) \in \mathbb{R}^p$ were generated from $N(0, I_p)$, where I_p is an identify matrix of dimension p. The true non-zero coefficients were set as $\beta_{01} = 0.2$, $\beta_{02} = 0.4$, $\beta_{03} = 0.6$, $\beta_{04} = 0.8$, and $\beta_{05} = 1$ with $\beta_{0j} = 0$ for any $5 < j \leq p$.

In sum, the explanatory variables in Examples 1, 2, and 3, are independent, autocorrelated, and uniform correlated, respectively. The setting in Example 4 was $p - 5$ irrelevant covariates X_{ij}, $j > 5$, that have non-zero correlations with the response variable.

For each simulation, we considered three sample sizes (n = 100, 200, and 300) and three covariate dimensions (p = 100, 1,000, and 10,000), which results in 9 different (n, p) combinations. For each (n, p) combination, a total of $M =$ 1,000 realizations were conducted, with the number of sequential steps K selected according to the method proposed in Remark 4. We denote the data generated in the m-th simulation replication as $(\mathbb{Y}_{[m]}, \mathbb{X}_{[m]})$ with $\mathbb{Y}_{[m]} \in \mathbb{R}^n$ and $\mathbb{X}_{[m]} \in \mathbb{R}^{n \times p}$. Based on this data, we subsequently obtained the SMA estimator, denoted as $\hat{\beta}_{[m]}$.

To evaluate the forecasting performance, we generated the independent testing dataset, denoted as $(\mathbb{Y}^*_{[m]}, \mathbb{X}^*_{[m]})$, where $\mathbb{Y}^*_{[m]} \in \mathbb{R}^{n^*}$ and $\mathbb{X}^*_{[m]} \in \mathbb{R}^{n^* \times p}$ with $n^* = 2{,}000$. We then employed the out-of-sample R^2, $\mathrm{OR}_{[m]} = 1 - \|X^*_{[m]}\hat{\beta}_{[m]} - Y^*_{[m]}\|^2 \|Y^*_{[m]}\|^{-2}$, to measure performance. For the sake of comparison, similar quantities were also computed for SIS (Fan and Lv (2008)), FR (Wang (2009)), MCV (Ando and Li (2014)), and sparse L2-Boosting (Bühlmann and Yu (2006)) denoted by Bstg. The methods of SIS and FR were used to generate solution paths, from which an optimal model was selected according to the BIC of (2.2). To avoid unnecessary bias, we adopted the method of Fan and Li (2001) and Leng, Lin and Wahba (2006) by applying the OLS estimates obtained from the selected model to make out-of-sample forecasts. The number of models and the number of regressors of the MCV method were optimized through cross-validation, as suggested by Ando and Li (2014). Several known regularization methods, such as the LASSO of Tibshirani (1996), SCAD of Fan and Li (2001), and MCP of Zhang (2010), are not presented here since all these methods have been demonstrated to be comparable to the method of MCV for high-dimensional data predictions (see Ando and Li (2014)). We only report the results of MCV in our simulation studies. We quantified their predictability via the corresponding averaged out-of-sample R^2 values (AOR) as AOR $= M^{-1}\sum_{m=1}^{M}\mathrm{OR}_{[m]}$. We measured its forecasting stability by the corresponding standard deviation (SD). To compare SMA with its specific competitor, we considered a measure, called Winning Probability (WP), as

$$\mathrm{WP} = \frac{1}{M}\sum_{m=1}^{M} I\Big(\mathrm{OR}_{[m]} > \mathrm{OR}^*_{[m]}\Big),$$

where $\mathrm{OR}^*_{[m]}$ represents the OR value for one particular competitor (e.g., SIS, FR, MCV and Bstg) in the m-th simulation replication.

3.2. Comparisons of SMA versus alternatives

Tables 1 and 2 present simulation results for Examples 1-2 and Examples 3-4, respectively. We find that the performance of SMA is often considerably better than that of SIS in terms of AOR and SD; see columns 4 and 9 in Tables 1 and 2. This finding is not surprising since SIS employs the marginal correlation between one covariate and the response to justify its relevance, and the useful information contained in other covariates is ignored. In contrast, SMA takes the approach of removing the contribution from previously selected covariates, which can yield superior performance.

Since FR is only a variable screening algorithm, the resulting estimate is a non-smooth and non-continuous function of data, with unsatisfactory forecasting stability. This is particularly true for small sample sizes and high-dimensional cases; see, for example, the case with $(n, p) = (100, 10{,}000)$ in Tables 1 and 2.

The AORs in Examples 1 and 2 of Table 1 indicate that SMA outperforms MCV. Although the SMA SDs are larger than those of MCV when $p = 1{,}000$ and $p = 10{,}000$, the overall measure MP shows that SMA is superior to MCV. The performance of the MCV is largely due to leave-one-out cross-validation, which exhibits deficiency in model selection and predictive ability (e.g., see Shao (1993)). Analogous results with less superiority of SMA versus MCV can be found in Example 3. Example 4 is a challenging for the task of discovering relevant predictors. In this example, the overall MP measure shows that SMA is better than, or comparable to, MCV, aside from the case ($n = 300$ and $p = 10{,}000$).

The performance of SMA is slightly better than that of boosting in terms of forecasting accuracy. This finding is expected since SMA is more stable than boosting. Overall, SMA is superior to boosting in these examples.

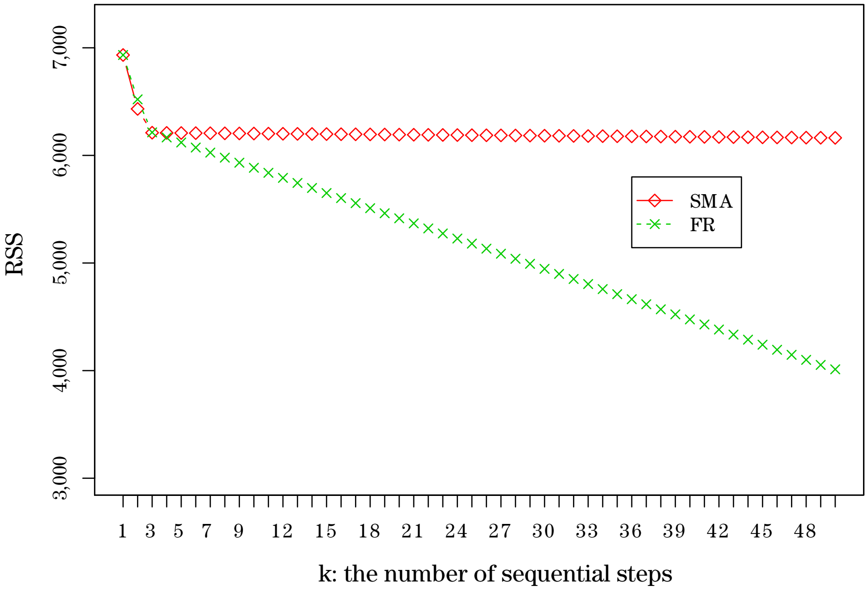

Figure 1. Residual sum of squares for SMA and FR at the k-th sequential step ($1 \leq k \leq 50$) in Example 2 with $n = 1{,}000$ and $p = 4{,}000$.

As suggested by an anonymous reviewer, we studied the computational cost by assessing the execution times from programming in R with an Intel (R) Core (TM) CPU (2.20 GHz). We found that SMA was computationally friendly. For instance, in Example 1 with $n = 100$ and $p = 1{,}000$, SMA took about 0.033 (× 1,000) seconds to finish the computations, while SIS, FR, MCV, and Bstg took about 0.015 (× 1,000), 0.025 (× 1,000), 0.062 (× 1,000) and 0.875 (× 1,000) seconds, respectively, to finish the computations. Here, SMA is slightly inferior to SIS and FR, while superior to MCV and Bstg in terms of computational cost.

The four simulation examples indicate that SMA is a computationally effective procedure and almost always has the best AOR values, and its SDs are competitive. Here the WP measures demonstrate that SMA is generally superior to its alternatives and can yield accurate and stable forecasts.

3.3. Finite sample properties of SMA

According to Theorems 3 and 4, it is not surprising that SMA performs well in these studies. To further illustrate those theoretical properties in finite samples, we repeated Example 2 with a relatively large sample size of $n = 1{,}000$, with $p = 4{,}000$ and $K = 50$ to be the maximal number of steps in the sequential SMA process. In each step, we computed the residual sum of squares (RSS),

$\|\mathbb{Y}_k\|^2$ for $k = 1, \cdots, 50$. For the sake of illustration, we also include the FR procedure. Figure 1 depicts the RSS versus k for SMA and FR. It shows that the RSS values of SMA and FR are almost identical in the first three steps. Since the number of relevant variables in Example 2 is 3, Figure 1 shows that SMA has the same capability as FR to fit the model by including important variables in a very limited number of steps; see Wang (2009). This result corroborates Theorem 3. Starting from step 4, the RSS values of SMA decrease much slower than those of FR. This indicates that, if useful information in covariates has all been exhausted, SMA's resistance to overfitting is considerably stronger than that of FR. This finding supports Theorem 4.

Motivated by an anonymous referee's suggestion, we also examined the average number of steps reached via our stopping rule for Examples 1–2 with the sample sizes $n = 100$, 200 and 300. We found 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 yield better results.

We revisited all four examples to assess the finite sample performance of Theorem 5. To better illustrate the asymptotic property of $\hat{\beta}$, we increased the sample sizes to $n = 200$, 400, and 800. In the m-th replication of $M = 1{,}000$ realizations, we denote the resulting SMA estimate and its error measure by $\hat{\beta}_{[m]}$ and $\|\hat{\beta}_{[m]} - \beta_0\|$, respectively. Table 3 reports the mean values of error measures for Examples 1–4. For the fixed sample size n, the mean value of estimation error steadily increases as p increases. This is expected, since a bigger model usually yields larger errors. In contrast, for fixed p, the mean value of the estimation error steadily decreases as n increases, which supports Theorem 5.

Due to an anonymous referee's suggestion, we further compared SMA with SIS, FR, MCV, and Bstg when $d_0 = 10$ and $d_0 = 20$. In addition, we compare SMA with the modified version of the traditional Bayesian model averaging method. A detailed description of these simulation settings and the results are given in the Supplemental Materials. The numerical results demonstrate that SMA is mostly superior to other methods. In conclusion, all simulations presented in the paper and its supplemental materials show that SMA performs well in both estimation and forecasting.

4. Data Analysis

4.1. Background

Due the rapid development of the search engine market, many companies want to expand their product exposure by purchasing advertisements that appear on the desired pages of search engines of such as Google and Baidu. With paid search advertising, a company purchases specific keyphrases and creates an advertisement that is displayed alongside organic (non-sponsored) web search results when a consumer searches for those keywords. Past industry experience suggests that paid search advertising is extremely effective.

In practice, there exist many keyphrases with similar semantic meanings. The only difference between these keyphrases is their textual formulation. Consider, for example, if a consumer wants to search for information about a property mortgage in the largest Chinese city, Shanghai. The customer can search for either "Home Mortgage Loan in Shanghai" (directly translated from Chinese) or "Mortgage Loan for Buying a Home in Shanghai" (among other options). These have exactly the same meaning, but differ slightly in their formulation because the latter includes the word "buying." Past experience suggests that different textual formulations can generate a dramatically different number of impressions, even if their intended semantic meaning is perfectly identical. Since the number of impressions reflects the number of customers who are looking for a particular keyphrase, it directly reflects the size of the potential market represented by the keyphrase. Consequently, it is of importance to understand the relationship between a keyphrase's textual formulation and the number of impressions it generates. Hence, it is critical for practitioners to have a statistical model that can predict the number of impressions accurately and stably by simply using the information contained in a keyphrase's textual formulation. This motivates us to apply SMA and its alternatives to such problems.

4.2. Data description

We considered 14 datasets collected by one of the largest search engine marketing agencies in mainland China. The response of interest is the number of impressions (after taking the log-transformation). The covariates we collected included each keyphrase's textual information. We created a high-dimensional covariate vector, in which each component is a binary variable indicating the presence or absence of a particular keyword. For example, we used i to represent a particular keyphrase (e.g., "Home Mortgage Loan in Shanghai"). We then defined a binary variable $X_{ij} = 1$ if the j-th keyword (say "Shanghai") appeared in the keyphrase i. On the other hand, if the keyphrase is "Home Mortgage Loan," which does not contain the keyword "Shanghai", we defined $X_{ij} = 0$. Because the number of keywords is large, the dimension of the binary vector X_i is usually high. For the sake of completeness, the number of keywords contained in the keyphrase is also included as a covariate. The response variable and covariates were standardized.

4.3. Performance comparison

Using this setting, we compared the SMA approach with its alternatives using 14 different datasets. Each dataset corresponds to one particular row in Table 4. The 14 datasets can be roughly classified into seven different online industries: online retailing, professional training, online recruiting, microblogging, mortgage lending, travel planning, and real-estate advertising. For convenience, we labeled each dataset with an industry ID; see the first column in Table 4. For evaluation purposes, each dataset was randomly split into two subsets of equal size. One subset serveed as the training sample, while the other one was used for testing. Each experiment was randomly replicated 1,000 times. Table 4 shows that the forecasting results were qualitatively similar to that of the simulation studies. Due to SMA's competitive performance in terms of both AOR and SD, most of the WP values are well above 70%. Then, SMA can more effectively and accurately predict impressions of keyphrases than the other four methods.

To make SMA practically useful, we propose a keyphrase index (KI). Specifically, we randomly split each dataset into the training sample and the testing sample. For a given dataset, let N be the total possible number of random splittings and denote splittings by $l = 1, \cdots, N$. For the l-th random splitting, we rank keyphrases according to their predicted impressions in the testing sample, with $r_l^{(k)}$ the rank of the k-th keyphrase in the l-th random splitting. The KI of the k-th keyphrase is taken as $\mathrm{KI}_k = N^{-1} \sum_{l=1}^{N} r_l^{(k)}$. Based on this index, the top three keyphrases in the mortgage industry example with $N = 1{,}000$ were (1) Shanghai real estate mortgage, (2) Shanghai housing mortgage, and (3) Shanghai bank mortgage. Based on the KI, vendors can purchase the most relevant keyphrases to meet their advertising goals subject to budget constraints.

5. Concluding Remarks

For high-dimensional data analysis, we propose a sequential model averaging approach to forecasting. Since it combines sequential screening and model averaging, SMA yields accurate and stable predictions. Although we only present empirical studies for internet advertising, SMA is applicable to such fields with high-dimensional data, as biological science, engineering, finance, marketing, medicine, physics, social science, etc. Replacing the OLS estimator in the SMA algorithm by a robust estimator is an interesting avenue for future research. We believe that extending this work to generalized linear models (McCullagh and Nelder (1989)) and semiparametric models (Fan and Gijbels (1996); Härdle, Liang and Gao (2000)) would further facilitate the use of SMA.

Supplementary Materials

The online supplemental materials present simulation studies that compare SMA with SIS, FR, MCV, and Bstg for less sparse regression models, compare SMA with the Bayesian model averaging method for Examples 1-4, and investigates the average number of steps reached via our proposed stopping rule. Proofs are presented in this material.

Acknowledgment

This research was supported by National Natural Science Foundation of China (NSFC, 71532001, 11525101, 11401482, 71420107025, 11471030), China's National Key Research Special Program Grant 2016YFC0207700, Fundamental Research Funds for the Central Universities. The authors are grateful to the Editor and reviewers for their constructive comments and suggestions, which have led to a significant improvement of this paper.

References

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* **66**, 237–242.

Ando, T. and Li, K. C. (2014). A model-averaging approach for high dimensional regression. *Journal of the American Statistical Association* **109**, 254–265.

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

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association* **98**, 324–339.

Bühlmann, P. and Yu, B. (2006). Sparse boosting. *Journal of Machine Learning Research* **7**, 1001–1024.

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics* **53**, 603–618.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edition.* Springer: New York.

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

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* New York: Cambridge University Press.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall: New York.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

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.

Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *The Annals of Statistics* **32**, 928–961.

Fan, J. and Song, R. (2010). Sure independent screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

Friedman J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussions). *The Annals of Statistics* **28**, 337–407.

Hansen, B. (2007). Least squares model averaging. *Econometrica* **75**, 1175-1189.

Härdle, W., Liang, H. and Gao, J. (2000). *Partially Linear Models.* Heidelberg: Springer.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning, 2nd Ed.* New York: Springer.

Hoeting, J., Madigan, D., Raftery, A. and Volinsky, C. (1999). Bayesian model averaging. *Statistical Science* **14**, 382–401.

Huang, J., Ma, S. and Zhang, C. H. (2007). Adaptive LASSO for sparse high dimensional regression. *Statistica Sinica* **18**, 1603–1618.

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

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* **8**, 613–636.

Leeb, H. and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* **142**, 201–211.

Leng, C., Lin, Y. and Wahba, G. (2006). A note on lasso and related procedures in model selection. *Statistica Sinica* **16**, 1273–1284.

Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics* **15**, 958–975.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129–1139.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, 2nd Ed.* New York: Chapman and Hall.

Racine, J. and Hansen, B. (2012). Jackknife model averaging. *Journal of Econometrics* **167**, 38–46.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486–494.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221–264.

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

Wan, A., Zhang, X. and Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics* **156**, 277–283.

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

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika* **99**, 15–28.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937–950.

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

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazard model. *Biometrika* **94**, 691–703.

Zhang, X., Zou, G. and Wan, A. (2013). Jackknife model averaging under a general covariance structure. *Journal of Econometrics* **172**, 82–94.

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

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics* **37**, 1733–1751.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

School of Statistics and Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, Sichuan, 611130, P. R. China.

E-mail: lanwei@swufe.edu.cn

School of Economics and Management, Beihang University, Beijing, 100871, P.R. China.

E-mail: mayingying_11@163.com

School of Statistics, Beijing Normal University, Beijing, 100871, P.R. China.

E-mail: zhaojunlong928@126.com

Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing, 100871, P.R. China.

E-mail: hansheng@gsm.pku.edu.cn

Graduate School of Management, University of California, Davis, CA 95616-8609, USA.

E-mail: cltucd@gmail.com

(Received March 2016; accepted January 2017)