Prediction Consistency of Forward Iterated Regression and Selection Technique

Shikai Luo and Subhashis Ghosal
Department of Statistics, North Carolina State University

Abstract

Recently, Hwang *et al.* (2009) proposed a variable selection method for high dimensional linear regression by attaching a penalty term with the loss function in a forward selection scheme, and regulate stopping by the amount of improvement in the loss function. The method appears to possess excellent prediction and variable selection property, but thus far no theoretical result on the procedure is given. In this article, we show that the procedure is prediction consistent.

Keywords: Forward selection; High dimension; Sparsity; Linear Regression; Penalization; Prediction consistency; Variable selection

1 Introduction and Main Result

Regression models with a large number of predictors are commonly used nowadays. Sparsity is typically present in such situations, that is only a few predictors are actually active, allowing meaningful prediction through some variable selection technique. Regularization methods are often used for the problem such as the LASSO [Tibshirani (1996)] or its numerous variants which use appropriate penalty functions to get a sparse minimizer of the penalized sum of squares in a linear regression model. Another approach to variable selection which has existed long in practice is that of forward selection. In this approach, the variable whose inclusion makes the sum of squares the smallest is selected, and variables not yet added in the model are sequentially added. The procedure continues until a certain number of steps which is typically found by tuning it with the help of a validating set of data. Consistency of forward selection in terms of average prediction error is given in the monograph Bühlmann and van de Geer (2011).

Recently, Hwang *et al.* (2009) proposed a variable selection technique in linear regression which is a hybrid of regularization method and forward selection. The procedure, called the forward iterative regression and shrinkage technique (FIRST), reduces p -dimensional optimization problems to p one-dimensional optimization problems each of which admits an analytical solution, where p stands for the ambient dimension of the predictor. The reduction is very useful in avoiding computer memory and storage problems for high dimensional predictors and parallel computing can be easily implemented. Compared with an ordinary forward selection procedure which is often viewed as “greedy” because it tries to minimize model fitting error at each step and hence often allows irrelevant predictors in the model, the additional penalty in FIRST filters out many irrelevant predictors easily. FIRST typically leads to substantially sparser models than LASSO without compromising the predictive power. However, convergence properties of FIRST have not been investigated. In this paper, we study consistency of FIRST in estimating the regression function, also called prediction consistency.

We consider the linear model $Y_i = \beta^T X_i + \varepsilon_i$, $i = 1, \dots, n$, where the predictors are deterministic and the regression errors are independent and identically distributed (i.i.d.) with mean zero and have subgaussian tails. Let $Y = (Y_1, \dots, Y_n)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ and $\psi_j = (X_{1j}, X_{2j}, \dots, X_{nj})^T$, $j = 1, 2, \dots, p$, where X_{ij} stands for the j th component of X_i . Let β^0 be the true value of β and let $\hat{\beta}$ be the estimated value of it. Then the quality of the estimation procedure can be assured by the convergence of the average squared distance between the estimated and the true value of the regression function at the observed values of the predictor to zero in probability, that is, $\|X_i^T(\hat{\beta} - \beta^0)\|^2/n \rightarrow 0$ in probability. We study this “prediction consistency” of FIRST in the true high dimensional setting: dimension $p = p_n \rightarrow \infty$ as $n \rightarrow \infty$ possibly at a faster rate than n . We shall assume that the very mild growth restriction of the dimension $\log p = o(n)$, the covariates are centered and scaled:

$$\sum_{i=1}^n X_{ij} = 0, \quad n^{-1} \sum_{i=1}^n X_{ij}^2 = 1 \text{ for all } j = 1, 2, \dots, p, \quad (1)$$

the true model is sparse with respect to the ℓ_1 -norm:

$$\|\beta^0\|_1 = \sum_{j=1}^p |\beta_j^0| = o\left(\sqrt{\frac{n}{\log p}}\right), \quad (2)$$

and the regression function is uniformly bounded in scaled ℓ_2 -norm in the

sense that

$$\|X\beta^0\|_2^2/n \leq C < \infty \text{ for all } n. \quad (3)$$

Before proceeding further, we recall the formal definition of FIRST. Let tuning parameters $\lambda, \eta > 0$ be given. Set $Y_{i,1} = Y_i$, $\hat{f}^{(1)} = 0$ and for $k = 1, 2, \dots$, repeat computing

(i) the one-dimensional lasso estimates

$$\hat{\beta}_{j,k}^L = \begin{cases} \hat{\beta}_{j,k} - \lambda/2, & \text{if } \hat{\beta}_{j,k} > \lambda/2, \\ 0, & \text{if } |\hat{\beta}_{j,k}| \leq \lambda/2, \\ \hat{\beta}_{j,k} + \lambda/2, & \text{if } \hat{\beta}_{j,k} < -\lambda/2, \end{cases}$$

where $\hat{\beta}_{j,k} = \sum_{i=1}^n Y_{i,k} X_{ij}$, $j = 1, \dots, p$,

(ii) select $\hat{j}_k = \arg \min \{j : \sum_{i=1}^n (Y_{i,k} - \hat{\beta}_{j,k}^L X_{i,j})^2\}$

(iii) update $Y_{i,k+1} = Y_{i,k} - \hat{\beta}_{\hat{j}_k,k}^L X_{i,\hat{j}_k}$, $\hat{f}^{(k+1)} = \hat{f}^{(k)} + \psi_{\hat{j}_k} \hat{\beta}_{\hat{j}_k,k}^L$, k to $k+1$

until $\|Y - \hat{f}^{(k-1)}\|^2 - \|Y - \hat{f}^{(k)}\|^2 < \eta$. Upon stopping, the predictor for Y is taken as $\hat{f}^{(m-1)}$ which is a sparse linear combination of $(X_1, \dots, X_p)$ and the corresponding coefficient vector is defined as the FIRST estimate $\hat{\beta}$.

To ensure convergence properties of FIRST, the tuning parameters must depend on n : the penalty must asymptotically vanish and the improvement threshold should become more stringent, i.e., $\lambda = \lambda_n \rightarrow 0$ and $\eta = \eta_n \rightarrow 0$ at an appropriate speed. The following gives the main result of this paper.

Theorem 1. *For the linear model with i.i.d. subgaussian errors, under conditions (1), (2) and (3), if*

$$\lambda_n = O\left(\sqrt{\log p/n}\right), \quad \sqrt{\log p/n} = o(\eta_n), \quad \eta_n \rightarrow 0, \quad (4)$$

then FIRST is prediction consistent, i.e., $\|X(\hat{\beta} - \beta^0)\|^2/n \rightarrow 0$ in probability.

2 Proof of the Result

Let m_n be the stopping time for FIRST. Note that FIRST stops at a given stage k only if either for all j , $|\hat{\beta}_{j,k}| < \lambda_n/2$ or the improvement in the loss function at the $(k+1)$ stage is less than η_n . Since $\max |\hat{\beta}_{j,k}|$ is positive for a fixed k , and $\lambda_n \rightarrow 0$ and $\eta_n \rightarrow 0$ as $n \rightarrow \infty$, the first relations must be false for sufficiently large n , and then the second is so as well since the

improvement will be positive. Thus $m_n > k$ for sufficiently large n , and since k is arbitrary fixed integer, this means that we must have $m_n \rightarrow \infty$.

The rest of the proof closely follows the arguments given in the proof of Theorem 12.2 of Bühlmann and van de Geer (2011) on the prediction consistency of ordinary forward selection, and hence we emphasize on the main differences presenting the rest of the arguments briefly but largely self-contained. The main idea is that if $\lambda_n \rightarrow 0$ sufficiently fast, FIRST should behave like ordinary forward selection, which corresponds to the special case $\lambda_n = 0$ for all n .

Define the scaled Euclidean inner product as $\langle u, v \rangle_n = n^{-1} \sum_{i=1}^n u_i v_i$ and corresponding squared norm $\|u\|_n^2 = n^{-1} \sum_{i=1}^n u_i^2$.

Note that before the procedure stops,

$$\begin{aligned} \hat{j}_k &= \arg \min_j \|Y - \hat{f}^{(k-1)} - \hat{\beta}_{j,k}^L \psi_j\|_n^2 \\ &= \arg \min_j \left(\|Y - \hat{f}^{(k-1)}\|_n^2 - |\langle Y - \hat{f}^{(k-1)}, \psi_j \rangle_n|^2 + \lambda_n^2/4 \right) \\ &= \arg \max_j |\langle Y - \hat{f}^{(k-1)}, \psi_j \rangle_n| \end{aligned}$$

because $\|Y - \hat{f}^{(k-1)} - \hat{\beta}_{j,k}^L \psi_j\|_n^2$ is given by

$$\begin{cases} \|Y - \hat{f}^{(k-1)} - (\langle Y - \hat{f}^{(k-1)}, \psi_j \rangle - \lambda_n/2) \psi_j\|_n^2, & \text{if } \hat{\beta}_{j,k} \geq \lambda_n/2 \\ \|Y - \hat{f}^{(k-1)} - (\langle Y - \hat{f}^{(k-1)}, \psi_j \rangle + \lambda_n/2) \psi_j\|_n^2, & \text{if } \hat{\beta}_{j,k} \leq -\lambda_n/2, \\ \|Y - \hat{f}^{(k-1)}\|_n^2, & \text{if } |\hat{\beta}_{j,k}| \leq \lambda_n/2. \end{cases}$$

Define $\hat{R}^k f^0 = f^0 - \hat{f}^{(k)}$ and we consider the norm of $\hat{R}^k f^0$. If $\hat{\beta}_{\hat{j}_k,k} \geq \lambda_n/2$, we have that

$$\hat{R}^k f^0 = \hat{R}^{k-1} f^0 - \langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n \psi_{\hat{j}_k} + \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n \psi_{\hat{j}_k} - \lambda_n \psi_{\hat{j}_k} / 2,$$

and hence

$$\begin{aligned} \|\hat{R}^k f^0\|_n^2 &= \|\hat{R}^{k-1} f^0\|_n^2 + \langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n^2 + \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n^2 + \lambda_n^2/4 \\ &\quad - 2\langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n^2 + 2\langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n - \lambda_n \langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n \\ &\quad - 2\langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n + \lambda_n \langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n - \lambda_n \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n \\ &= \|\hat{R}^{k-1} f^0\|_n^2 - \langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n^2 + \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n^2 - \lambda_n \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n + \lambda_n^2/4. \end{aligned}$$

If $\hat{\beta}_{\hat{j}_k,k} \leq -\lambda_n/2$, similar arguments show that

$$\hat{R}^k f^0 = \hat{R}^{k-1} f^0 - \langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n \psi_{\hat{j}_k} + \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n \psi_{\hat{j}_k} + \lambda_n \psi_{\hat{j}_k} / 2,$$

and

$$\|\hat{R}^k f^0\|_n^2 = \|\hat{R}^{k-1} f^0\|_n^2 - \langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n^2 + \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n^2 + \lambda_n \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n + \lambda_n^2/4.$$

If $|\hat{\beta}_{\hat{j}_k, k}| < \lambda_n/2$ we have that

$$\hat{R}^k f^0 = \hat{R}^{k-1} f^0, \quad \|\hat{R}^k f^0\|_n^2 = \|\hat{R}^{k-1} f^0\|_n^2.$$

Furthermore, when $|\hat{\beta}_{\hat{j}_k, k}| < \lambda_n/2$, we have that

$$|\hat{\beta}_{\hat{j}_k, k}| = |\langle \hat{R}^{k-1} f^0 + \varepsilon, \psi_{\hat{j}_k} \rangle_n| = |\langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n + \langle \varepsilon, \psi_{\hat{j}_k} \rangle_n| < \lambda_n/2$$

which leads to $|\langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n| < |\langle \varepsilon, \psi_{\hat{j}_k} \rangle_n| + \lambda_n/2 \leq \Delta_n + \lambda_n/2$.

Thus, for all three cases, we have

$$\|\hat{R}^k f^0\|_n^2 \leq \|\hat{R}^{k-1} f^0\|_n^2 - \langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n^2 + (\Delta_n + \lambda_n/2)^2.$$

Define $a_k = \|\hat{R}^k f^0\|_n^2$, $\Delta_n^* = \Delta_n + \lambda_n/2$ and $\Delta_n = \max_{j=1,2,\dots,p} |\langle \varepsilon, \psi_j \rangle_n|$. By a standard maximal inequality for subgaussian random variables we have that $\Delta_n = O_P(\sqrt{\log p/n})$, and hence $\Delta_n^* = O_P(\sqrt{\log p/n})$ as well for the choice λ_n given by (4). Therefore it follows that

$$a_k \leq a_{k-1} - |\langle \hat{R}^{k-1} f^0, \psi_{\hat{j}_k} \rangle_n|^2 + \Delta_n^{*2}. \quad (5)$$

Expanding the square in the expression for a_k , elementary algebra gives $a_k \leq a_{k-1} - d_k^2 + 2d_k \Delta_n^*$ where $d_k = |\langle Y - \hat{f}^{(k-1)}, \psi_{\hat{j}_k} \rangle_n|$.

If $b_0 = \|\beta^0\|$ and $b_k = b_{k-1} + d_k$, then by following the arguments of Bühlmann and van de Geer (2011), we have

$$\max_{j=1,\dots,p} |\langle \hat{R}^{k-1} f^0, \psi_j \rangle_n| \geq a_{k-1}/b_{k-1}. \quad (6)$$

If now $\max\{|\langle \hat{R}^{k-1} f^0, \psi_j \rangle_n| : j = 1, \dots, p\} \geq 2\Delta_n^*/\kappa$ for some $0 < \kappa < 1/2$, then by second assertion of Lemma 12.1 of Bühlmann and van de Geer (2011) with Δ_n replaced by Δ_n^* , we have that

$$d_k \geq (1 - \kappa/2) \max_j |\langle \hat{R}^{k-1} f^0, \psi_j \rangle_n| \geq (1 - \kappa/2) a_{k-1}/b_{k-1}. \quad (7)$$

Also if $\max\{|\langle \hat{R}^{k-1} f^0, \psi_j \rangle_n| : j = 1, \dots, p\} \geq 2\kappa^{-1}(1 - \kappa/2)^{-1}\Delta_n^*$, then applying Lemma 12.2 of Bühlmann and van de Geer (2011) with Δ_n replaced

by Δ_n^* , we obtain $a_k \leq a_{k-1}[1 - (1 - \kappa)(1 - \kappa/2)^2 a_{k-1}/b_{k-1}^2]$. Since the sequence b_k is increasing, this gives

$$a_k b_k^{-2} \leq a_{k-1} b_{k-1}^{-2} (1 - C_\kappa^2 a_{k-1} b_{k-1}^{-2}), \quad (8)$$

where $C_\kappa = \sqrt{1 - \kappa}(1 - \kappa/2)$. Furthermore, $a_0 b_0^{-2} \leq \|\beta_n^0\|_1^{-2}$ using the scaling $\|f^0\|_n^2 \leq 1$.

Now on the event

$$B_n(m) = \bigcap_{k=1}^m \left\{ \max_{j=1, \dots, p} |\langle \hat{R}^{k-1} f^0, \psi_j \rangle_n| \geq 2\kappa^{-1} (1 - \kappa/2)^{-1} \Delta_n^* \right\},$$

using (7), (8) and Lemma 12.3 of Bühlmann and van de Geer (2011), we obtain

$$a_m b_m^{-2} \leq \|\beta^0\|_1^{-2} (1 + C_\kappa^2 m)^{-1}. \quad (9)$$

Also by (7) and Lemma 12.2 of Bühlmann and van de Geer (2011), on $B_n(m)$

$$\begin{aligned} a_m \leq a_{m-1} - (1 - \kappa) d_m^2 &\leq a_{m-1} - (1 - \kappa)(1 - \kappa/2) d_m a_{m-1} / b_{m-1} \\ &= a_{m-1} \left(1 - \frac{D_\kappa d_m}{b_{m-1}} \right), \end{aligned} \quad (10)$$

where $D_\kappa = (1 - \kappa)(1 - \kappa/2)$. Writing $b_m = b_{m-1}(1 + d_m/b_{m-1})$ and using $(1 + u)^\alpha \leq 1 + \alpha u$, $0 \leq \alpha \leq 1$, $u \geq 0$, (10) leads to

$$a_m b_m^{D_\kappa} \leq a_{m-1} \left(1 - \frac{D_\kappa d_m}{b_{m-1}} \right) b_{m-1}^{D_\kappa} \left(1 + \frac{D_\kappa d_m}{b_{m-1}} \right) \leq a_{m-1} b_{m-1}^{D_\kappa}$$

and by repeated applications we get the bound $a_m b_m^{D_\kappa} \leq a_0 b_0^{D_\kappa} \leq \|\beta^0\|_1^{D_\kappa}$ since $b_0 = \|\beta^0\|_1$ and $a_0 \leq 1$. Combining this with (9) this gives

$$a_m^{2+D_\kappa} = (a_m b_m^{-2})^{D_\kappa} (a_m b_m^{D_\kappa})^2 \leq (1 + C_\kappa^2 m)^{-D_\kappa}.$$

Clearly $C_\kappa^{-2D_\kappa} = ((1 - \kappa)(1 - \kappa/2)^2)^{-(1-\kappa)(1-\kappa/2)} \leq 2$ as $0 < \kappa < 1/2$. and hence $a_m^{2+D_\kappa} \leq 2m^{-D_\kappa}$.

On the other hand, on the event $B_n(m)^c$, $\max\{|\langle \hat{R}^{k^*-1} f^0, \psi_j \rangle_n| : j = 1, \dots, p\} < 2\kappa^{-1}(1 - \kappa/2)^{-1} \Delta_n^*$ for some $k^* \leq m$. By (5),

$$\|\hat{R}^m f^0\|_n^2 \leq \|\hat{R}^{m-1} f^0\|_n^2 + \Delta_n^{*2} \leq \dots \leq \|\hat{R}^{k^*-1} f^0\|_n^2 + (m - k^*) \Delta_n^{*2}.$$

Furthermore, using (6) and the definition of the sequence b_k ,

$$\begin{aligned} \|\hat{R}^{k^*} f^0\|_n^2 &\leq \max_j |\langle \hat{R}^{k^*-1} f^0, \psi_j \rangle_n| b_{k^*-1} \\ &\leq \max_j |\langle \hat{R}^{k^*-1} f^0, \psi_j \rangle_n| \left(\|\beta^0\|_1 + \sum_{r=0}^{k^*-1} \|Y - f^{(r)}\|_n \right) \\ &\leq \max_j |\langle \hat{R}^{k^*-1} f^0, \psi_j \rangle_n| (\|\beta^0\|_1 + k^* \|Y\|_n). \end{aligned}$$

and consequently, by the definition of k^* and because $k^* \leq m$, the bound above becomes:

$$\|\hat{R}^m f^0\|_n^2 \leq 2\kappa^{-1}(1 - \kappa/2)^{-1} \Delta_n^* (\|\beta_n^0\|_1 + m\|Y\|_n) + m\Delta_n^{*2}.$$

Combining with the estimate $2m^{-D\kappa/(2+D\kappa)}$ on the event $B_n(m)$, we get all over the sample space the common estimate

$$\|\hat{R}^m f^0\|_n^2 \leq \max\{2m^{-\frac{D\kappa}{2+D\kappa}}, 2\kappa^{-1}(1 - \kappa/2)^{-1} \Delta_n^* (\|\beta_n^0\|_1 + m\|Y\|_n) + m\Delta_n^{*2}\}.$$

Evaluating the bound at $m = m_n$, the stopping time for FIRST, the first term inside the maximum goes to zero in probability as $n \rightarrow \infty$ because $m_n \rightarrow \infty$. Now $E\|Y\|_n^2 = \sigma^2 + \|X\beta^0\|^2/n$ is bounded by assumption (3), so that $\|Y\|_n$ is stochastically bounded. Since $\Delta_n^* = O_P(\sqrt{\log p/n})$, $\Delta_n^* \|\beta_n^0\|_1 = o_P(1)$ by (2). Thus it remains to show that $m_n \Delta_n^* = o_P(1)$. Note that prior to stopping in FIRST, every step must improve loss function by at least η_n and the maximum improvement in aggregate cannot exceed $\|Y\|_n^2 = O_P(1)$, so $m_n \eta_n \leq \|Y\|_n^2$ which leads to $m_n \leq \|Y\|_n^2/\eta_n = O_P(\eta_n^{-1})$. Thus $m_n \Delta_n^* \rightarrow 0$ in probability in view of assumption (4). The proof is now complete.

Acknowledgment: Research of the second author is partially supported by National Security Agency grant number H98230-12-1-0219.

References

- Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.
- Hwang, W. Y., Zhang, H. H. and Ghosal, S. (2009). FIRST: Combining forward iterative selection and shrinkage in high dimensional sparse linear regression. *Statistics and Its Interface*, **2**, 341–348.
- Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.