Testing-Based Forward Model Selection

Kozbur, Damian

Abstract: This paper defines and studies a variable selection procedure called Testing-Based Forward Model Selection. The procedure inductively selects covariates which increase predictive accuracy into a working statistical regression model until a stopping criterion is met. The stopping criteria and selection criteria are defined using statistical hypothesis tests. The paper explicitly describes a testing procedure in the context of high-dimensional linear regression with heteroskedastic disturbances. Finally, a simulation study examines finite sample performance of the proposed procedure and shows that it behaves favorably in high-dimensional sparse settings in terms of prediction error and size of selected model.

DOI: https://doi.org/10.1257/aer.p20171039

Testing-Based Forward Model Selection

Damian Kozbur [1]

This paper studies an algorithm called Testing-Based Forward Model Selection. Forward selection algorithms are model selection procedures that inductively select covariates which increase predictive accuracy into a working statistical regression model until a stopping criterion is met.

Deciding which covariate provides the best additional predictive capability is complicated in finite samples by the fact that outcomes are observed with noise or are partly idiosyncratic. In linear regression, a covariate associated to a positive increment of in-sample R-squared upon inclusion may not add any predictive power out-of-sample. Statistical hypothesis tests offer one way to determine whether a covariate of interest likely improves out-of-sample prediction. Furthermore, in many econometric applications, the classical assumption of iid data is inappropriate. An example of this is the presence of heteroskedastic disturbances. In such settings, higher R-squared resulting from inclusion of one variable relative to another need not imply that the first variable is a better choice. The availability of hypothesis tests for diverse classes of problems and settings motivates us to introduce a testing-based model selection strategy.

In this paper, there is particular interest in the application of model selection involving high-dimensional data, characterised as data with a large number of covariates relative to the sample size. High-dimensional data arise through a combination of two ways. The data may be intrinsically high dimensional with many different characteristics per observation. Alternatively, even when the number of available variables is relatively small, researchers are often faced with a large set of potential covariates formed by different ways of interacting and transforming the original variables.

Dealing with a high-dimensional dataset necessarily involves dimension reduction or regularization. Without dimension reduction or regularization, any statistical model will overfit a high-dimensional dataset. Understanding the behavior of Testing-Based Forward Model Selection in this context is of interest since it potentially offers a completely data-driven way to regularize high dimensional models.

There are several earlier theoretical analyses of forward selection, though none use of testing as a criteria for stopping. [7] gives performance bounds under a β-min condition, restricting the minimum magnitude of nonzero coefficients. [8] and [5] prove performance bounds under a strong irrepresentability condition. [3] prove bounds on the relative performance in population R-squared of the a forward selection based model (relative to infeasible R-squared) when the number of variables allowed for selection is fixed.

This paper defines Testing-Based Forward Model Selection and specifies a testing procedure in the context of high-dimensional linear regression with heteroskedastic disturbances. This paper then presents a simulation study which examines finite sample performance of testing-based forward model selection.

In economic applications, models learned using formal model selection are often used in subsequent estimation steps. One example is the selection of instrumental variables for later use in a first stage regression (see [1]). Another example is the selection of control variables into a conditioning set (see [2], [6]). Such applications require a model selection procedure with a hybrid objective: (1) produce a good fit, and (2) return a sparse set of covariates. The results given here address both objectives by studying resulting mean squared prediction errors and size of selected models.

[1]I gratefully acknowledge helpful discussion with Christian Hansen, Tim Conley, Attendants at the ETH Zürich Seminar für Statistik and ETH Zürich Center for Law and Economics Seminars, and support of the ETH Fellowship program.

I. Framework

Consider random variables $\{y_i\}_{i=1}^n \in (\mathbb{R})^n$ and a set of covariates $\{x_i\}_{i=1}^n \in (\mathbb{R}^p)^n$ which satisfy

$$\text{(I.1)} \quad y_i = x_i'\theta^* + \varepsilon_i, \quad |\text{supp}(\theta^*)| \leqslant s < n.$$

Interest lies in estimating θ^* with an estimate $\widehat{\theta}$ which is (1) sparse and (2) gives good predictions $x_i'\widehat{\theta}$ for $x_i'\theta_0$.

Consider a family of quadratic loss functions indexed by θ:

$$\text{(I.2)} \quad \ell_\theta(\{(x_i, y_i)\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n (y_i - x_i'\theta)^2.$$

Define also $\mathcal{E}(\theta) = \text{E}\ell_\theta - \text{E}\ell_{\theta^*}$, $\mathcal{E}(S) = \inf_{\theta:\text{supp}(\theta)\subset S} \mathcal{E}(\theta)$.

The goal is to select $\widehat{S}$ by a forward selection procedure which involves the use of statistical hypothesis tests. For any S define the incremental loss from the jth covariate by

$$\text{(I.3)} \qquad \Delta_j\mathcal{E}(S) = \mathcal{E}(S \cup \{j\}) - \mathcal{E}(S).$$

Consider a set of tests which will guide the forward selection process:

$$\text{(I.4)} \qquad T_{jS\alpha} \in \{0, 1\}$$

associated to

$$\text{(I.5)} \qquad H_0 : \Delta_j\mathcal{E}(S) = 0$$

with level $\alpha > 0$. Rejection ($T_{jS\alpha} = 1$) occurs for large values of a test statistic W_{jS}.

The model selection procedure is as follows. Start with an empty model (consisting of no covariates). At each step, if the current model is $\widehat{S}$, select one covariate such that $T_{j\widehat{S}\alpha} = 1$, append it to $\widehat{S}$, and continue to the next step; if no covariates have $T_{j\widehat{S}\alpha} = 1$, then terminate the model selection procedure and return the current model. If at any juncture, there are two indices j, k (or more) such that $T_{jS\alpha} = T_{kS\alpha} = 1$, the selection is made according to the larger value of W_{jS}, W_{kS}.

The algorithm for forward selection given the set of hypothesis tests $\{T_{jS\alpha}, W_{jS}\}$ is given formally by:

Testing-Based Forward Model Selection Algorithm

Initialize. Set: $\widehat{S} = \{\}$
For $1 \leqslant k \leqslant p$**:**
  If: $T_{j\widehat{S}\alpha} = 1$, for some $j \in \{1, ..., p\} \setminus \widehat{S}$,
    Set: $\widehat{j} \in \arg\max\{W_{j\widehat{S}} : T_{j\widehat{S}\alpha} = 1\}$
    Update: $\widehat{S} = \widehat{S} \cup \{\widehat{j}\}$
  Else: Break
Set: $\widehat{\theta} \in \underset{\text{supp}(\theta)\subset\widehat{S}}{\text{argmin}}\, \ell_\theta(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n)$

II. Testing-Based Forward Model Selection in the Case of Heteroskedastic Disturbances

This section provides an example by illustrating an application of model selection in the presence of heteroskedasticity.

Model. For each n consider the following model:

$$\text{(II.1)} \qquad y_i = x_i'\theta^* + \epsilon_i$$

with $x_i \in \mathbb{R}^p$. The disturbances ϵ_i satisfy $\text{E}[\epsilon_i|x_i] = 0$, are independent, but are not necessarily identically distributed. Finally, $s = s(n) := |\text{supp}(\theta^*)| < n$.

Note that for any subset S and any $j \notin S$, that the condition $\Delta_j\mathcal{E}(S) \neq 0$ is equivalent to $[\theta^*_{jS}]_j \neq 0$ where θ^*_{jS} is defined the as optimal coefficient given the model $j \cup S$. Therefore, work with second formulation and consider

$$\text{(II.2)} \qquad H_0 : [\theta^*_{jS}]_j = 0.$$

To construct the tests, begin with least squares estimate of $[\theta^*_{jS}]_j$, which is $[\widehat{\theta}^*_{jS}]_j$. Denote with $\widehat{V}_{jS}$ a heteroskedasticity robust estimate of the variance (the simulation analysis below uses HC1 in [4]). Finally, define the test statistics:

$$\text{(II.3)} \qquad W_{jS} = \widehat{V}_{jS}^{-1/2} \left|[\widehat{\theta}_{jS}]_j\right|.$$

Reject the null H_0 for large values of W_{jS} defined relative to an appropriately chosen threshold. To define the threshold first let $\eta_{jS} := (1\ , -\beta'_{jS})'$ where β_{jS} is the ordinary least squares coefficient from regressing x_{ij} on x_{iS}. Let $|\eta_{jS}|$ denote the vector whose components are the absolute values of η_{jS}. Next, let $\Psi^{\widehat{\epsilon}}$ be defined by $[\Psi^{\widehat{\epsilon}}_{jS}]_{k,l} = \sum_{i=1}^{n} \widehat{\epsilon}^2_{ijS} x_{ik} x_{il}$ for $k, l \in jS$. Then define

$$\widehat{\tau}_{jS} = \frac{|\eta_{jS}|' \mathrm{diag}(\Psi^{\widehat{\epsilon}}_{jS})}{\sqrt{\eta'_{jS} \Psi^{\widehat{\epsilon}}_{jS} \eta_{jS}}}. \tag{II.4}$$

Hypothesis Tests. Set tuning parameters $c_\tau > 1$, $\alpha > 0$. Assign:

$$T_{jS\alpha} = 1 \tag{II.5}$$

whenever

$$W_{js} \geqslant c_\tau \widehat{\tau}_{jS} \Phi^{-1}(1 - \alpha/p). \tag{II.6}$$

The idea behind this formulation is that α should control the family-wise error rate over the entire set of hypothesis tests encountered. α is a tuning parameter which maybe be chosen by the researcher. The term $\Phi^{-1}(1 - \alpha/p)$ can be informally thought of as a Bonferonni correction term which takes into account of the fact that there are p potential covariates. The term $c_\tau \widehat{\tau}_{jS}$ can be informally thought of as a correction term which can account for the fact that the set S is random and can have many potential realizations. In the simulations, $\alpha = .05, c_\tau = 1.01$ are used.

III. Simulation

In this section a simulation study evaluates the performance of the Testing-Based Forward Model Selection in finite samples. The estimates are compared to those of Lasso and Post-Lasso since these are popular and important generic high dimensional estimation strategies.

Consider the following data generating process:

$$\begin{aligned} &y_i = x_i'\theta^* + \epsilon_i, \theta^*_j = b^{j-1}\mathbf{1}_{j \leqslant s} \\ &x_{ij} \sim N(0,1), \rho(x_{ij}, x_{ik}) = 0.5^{|j-k|} \\ &\epsilon_i \sim \sigma_i N(0,1). \end{aligned} \tag{III.1}$$

Replicate all simulations with parameter choices $b \in \{0.5, -0.5\}, n = 100, p = 120, s = 6$. The high-dimensional ($n < p$) setting with sparsity ($s < n$) setting is where Lasso and Testing-Based Forward Model Selection are expected to perform well. In the heteroskedastic simulation, set $\sigma_i = \exp(0.5 \sum_{j=1}^{p} 0.75^{(p-j)} x_{ij})$. Otherwise, set $\sigma_i = 0.5$. The study proceeds with 1000 replications for each design. The results are presented in Table 1.

For the forward selection estimator, we use tuning parameters $c_\tau = 1.01$, $\alpha = .05$. The resulting estimator is called Forward. To construct a Lasso and Post-Lasso estimate, use the implementation found in [1] since it is designed to handle heteroskedasticity. [1] require two tuning parameters which are directly analogous to c_τ and α, so again use $c_\tau = 1.01$ and $\alpha = 0.05$. Finally, consider an oracle estimator of least squares on the true support.

TABLE 1
Testing-Based Forward Model Selection

Simulation Results: 1000 Replications
$n = 100,\ p = 120,\ s = 6$

	MPEN	MSSS	MPEN	MSSS
	Homoskedastic		Heteroskedastic	
	Panel A. $\theta_j = 0.5^{j-1}\mathbf{1}_{j \leqslant s}$			
Forward	0.27	2.57	0.88	0.93
Lasso	207	2.05	1.67	6.42
Post-Lasso	0.33	2.05	1.58	6.42
Oracle	0.12	6.00	0.76	6.00
	Panel B. $\theta_j = (-0.5)^{j-1}\mathbf{1}_{j \leqslant s}$			
Forward	0.26	1.74	0.72	0.45
Lasso	0.63	0.91	1.17	5.00
Post-Lasso	0.44	0.91	1.60	5.00
Oracle	0.12	6.00	0.75	6.00

Note: Mean prediction error norm (MPEN) and mean size of selected set (MSSS) for several estimators.

Table 1 prints mean prediction error norm (PEN = $(\sum_{i=1}^{n}(x_i'\theta^* - x_i\widehat{\theta}))^2/n)^{1/2})$, and mean size of selected set (MSSS). There are important instances when the forward selection estimators consistently out-

perform the Lasso-based estimators. Forward selection estimates tend to do better relative to Post-Lasso in the presence of heteroskedasticity. In addition, Post-Lasso gives very poor estimates when $b = -0.5$ while the forward selection estimators perform well (relative to Oracle) suggesting that the performance of these estimators depends on the configuration of the signal.

IV. Conclusion

This paper develops theory for Testing-Based Forward Model Selection in linear regression problems with heteroskedasticity. The proposed procedure is compared to Lasso and Post- Lasso in a simulation studies which finds that it shows favorable performance.

REFERENCES

[1] **Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen.** 2012. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." *Econometrica*, 80: 2369–2429. Arxiv, 2010.

[2] **Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen.** 2014. "Inference on Treatment Effects after Selection Amongst High-Dimensional Controls with an Application to Abortion on Crime." *Review of Economic Studies*, 81(2): 608–650.

[3] **Das, Abhimanyu, and David Kempe.** 2011. "Submodular meets Spectral: Greedy Algorithms for Subset Selection, Sparse Approximation and Dictionary Selection." 1057–1064. New York, NY, USA:ACM.

[4] **MacKinnon, J. G., and H. White.** 1985. "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties." *Journal of Econometrics*, 29: 305–325.

[5] **Tropp, Joel A.** 2004. "Greed is good: algorithmic results for sparse approximation." *Information Theory, IEEE Transactions on*, 50(10): 2231–2242.

[6] **van de Geer, Sara, Peter Bhlmann, Yaacov Ritov, and Ruben Dezeure.** 2014. "On asymptotically optimal confidence regions and tests for high-dimensional models." *Ann. Statist.*, 42(3): 1166–1202.

[7] **Wang, Hansheng.** 2009. "Forward Regression for Ultra-High Dimensional Variable Screening." *Journal of the American Statistical Association*, 104:488: 1512–1524.

[8] **Zhang, Tong.** 2009. "On the Consistency of Feature Selection using Greedy Least Squares." *Journal of Machine Learning*, 10: 555–568.