Economics Letters 151 (2017) 119–122

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

# Heteroscedasticity-robust model screening: A useful toolkit for model averaging in big data analytics[☆]

Tian Xie

*Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, Xiamen, Fujian 361005, China*
*Department of Finance, School of Economics, Xiamen University, Xiamen, Fujian 361005, China*
*MOE Key Lab of Econometrics, Xiamen University, Ministry of Education, Xiamen, Fujian 361005, China*
*Fujian Key Lab of Statistical Sciences, Xiamen University, Xiamen, Fujian 361005, China*

## HIGHLIGHTS

- Propose a new heteroscedasticity-robust model screening (HRMS) method.
- Show that HRMS has good performance in simulation.
- Demonstrate that HRMS is computationally efficient.
- Show that HRMS can lead to large gains in box office prediction accuracy.

ARTICLE INFO

*Article history:*
Received 11 August 2016
Received in revised form 17 October 2016
Accepted 10 December 2016
Available online 16 December 2016

*JEL classification:*
C52
C53
D03
M21

*Keywords:*
Model screening
Model averaging
Big data analytics

## ABSTRACT

Frequentist model averaging has been demonstrated as an efficient tool to deal with model uncertainty in big data analysis. In contrast with a conventional data set, the number of regressors in a big data set is usually quite large, which leads to a exponential number of potential candidate models. In this paper, we propose a heteroscedasticity-robust model screening (HRMS) method that constructs a candidate model set through an iterative procedure. Our simulation results and empirical exercise with big data analytics demonstrate the superiority of our HRMS method over existing methods.

## 1. Introduction

The term big data is now commonly used in popular press in part, due to excitement in industry about using social media data to predict people's reactions to new products including movies. A big challenge researchers in this area face is determining what explanatory variables to extract from approximately 350 million tweets and 6 billion Facebook messages per day, and how to use them in forecasting exercises. As demonstrated in Lehrer and Xie (2016), frequentist model averaging is an efficient tool to deal with this uncertainty in big data analytics.

A model averaging estimator obtains a weighted average of estimates from a set of candidate models through numerical optimization routines. The performance of a model averaging estimator crucially depends on the candidate model set, as demonstrated in Hansen (2007). In practice, one possible approach is to construct the candidate model set using a full permutation of all regressors. One obvious drawback is that the total number of candidate models increases exponentially with the number of regressors. As shown in many works including Wan et al. (2010) and Xie (2015), keeping the total number of candidate models small or slowing its convergence to infinity is a necessary condition to maintain the asymptotic optimality of model averaging estimators. While most existing theoretical works assumed a given candidate model set, a recent paper by Zhang et al. (forthcoming) established the asymptotic optimality of model averaging estimators with screened candidate models.

[☆] The paper is partially supported by the Fundamental Research Funds for the Central Universities 20720171002.

*E-mail address:* xietian001@hotmail.com.

http://dx.doi.org/10.1016/j.econlet.2016.12.019
0165-1765/© 2016 Published by Elsevier B.V.

Inspired by the forward (FW) method of Claeskens et al. (2006), we propose a new heteroscedasticity-robust model screening (HRMS) technique that constructs a candidate model set through an iterative procedure which adds one regressor at a time and ends with a sequence of nested candidate models. As demonstrated in practice, for example, Liang et al. (2011), Lehrer and Xie (2016), and the results in our Appendix B, it is quite common that a handful of models can assume more than 95% of the total weights and a considerable proportion of the candidate models have near-zero weights. These models can be discarded without harming the result.

This paper continues with a detailed introduction of our HRMS method in Section 2. Section 3 studies the finite sample performance of our HRMS method by comparing it with existing model screening methods using Monte Carlo simulations. In Section 4, we examine to what extent can data on aggregate product sentiment obtained from messages in Twitterverse can improve business decisions via demand forecasting.

## 2. Heteroscedasticity-robust model screening

Our setup is similar to that of Wan et al. (2010). We observe a random sample of $(y_i, x_i)$ for $i = 1, \ldots, n$, in which $y_i$ is a scalar and $x_i = (x_{i1}, x_{i2}, \ldots)$ is countably infinite. We consider the following data generating process (DGP):

$$y_i = \mu_i + e_i, \quad \mu_i = \sum_{j=1}^{\infty} \beta_j x_{ij}, \quad \mathbb{E}(e_i|x_i) = 0 \tag{1}$$

for $i = 1, \ldots, n$ and where $\mu_i$ can be considered as the conditional mean $\mu_i = \mu(x_i) = \mathbb{E}(y_i|x_i)$. We allow the error term to be heteroscedastic by setting the error term variance as $\sigma_i^2 = \mathbb{E}(e_i^2|x_i)$.

In the model averaging literature, we usually assume that there exists a total of $M$ candidate models approximating the DGP in Eq. (1):

$$y_i = \sum_{j=1}^{k^m} \beta_j^m x_{ij}^m + b_i^m + e_i,$$

for $m = 1, \ldots, M$, where $x_{ij}^m$ for $j = 1, \ldots, k^m$ denotes the regressors, $\beta_j^m$ denotes the coefficients, and $b_i^m \equiv \mu_i - \sum_{j=1}^{k^m} \beta_j^m x_{ij}^m$ is the modeling bias. The $M$ candidate models form a model set $\mathcal{M}^K$ that consists of $K$ regressors. Each candidate model contains $k^m \leq K$ regressors.

In theory, the model set $\mathcal{M}^K$ is assumed to be predetermined. However, in practice we need to construct $\mathcal{M}^K$ from different combinations of the regressors. One popular approach is to consider a full permutation of all $K$ regressors that generates $M_{\text{full}} = 2^K - 1$ candidate models. However, this approach is not appropriate for large $K$.

In this section, we propose an HRMS method that can efficiently restrict the number of candidate models. Our HRMS method is an iterative procedure that adds one variable at a time and ends with a sequence of nested candidate models. We summarize our HRMS method in the following.

(i) We pick an initial model, denoted by $\mathbb{M}_{(0)}$, which can be a null model that includes no variables, or a model consisting of certain regressors (variables) of interest.

(ii) We add each of the $S_{(0)}$ remaining regressors one at a time to $\mathbb{M}_{(0)}$. This generates $S_{(0)}$ candidate models. Then, we examine each candidate model by the following heteroscedasticity-robust criterion:

$$\text{HRMS}(s) = \left\| y - P^s y \right\|^2 + 2\sum_{i=1}^{n} \left(e_i\right)^2 p_{ii}^s \quad \text{for } s = 1, \ldots, S_{(0)}, \tag{2}$$

where $P^s$ stands for the projection matrix of the regressors, $e_i$ is the $i$th element of the error term that needs to be approximated by a least squares residual, and $p_{ii}^s$ represents the $i$th diagonal term in $P^s$.

(iii) We select the model that yields the lowest value for criterion (2), denoted by $\mathbb{M}_{(1)}$, and treat it as the initial model of the next round.

(iv) We repeat steps (ii)–(iii) iteratively until we reach the pool model that consists of all $K$ variables. We construct our candidate model set $\mathcal{M}^K$ including all selected models, $\mathbb{M}_{(0)}$ (if not null), and the pool model.

The HRMS method adds one and only one variable to the previous step's model each time. Therefore, if there are $K$ variables in total and our initial model $\mathbb{M}_{(0)}$ includes $K_0$ variables, then we end up with only $(K - K_0 + 1)$ nested models, which is much smaller than $M_{\text{full}}$, especially for large $K$.

Our HRMS method can be easily extended to models with homoscedastic error terms by replacing the heteroscedasticity-robust criterion in Eq. (2) with the following HEMS criterion:

$$\text{HEMS}(s) = \left\| y - P^s y \right\|^2 + 2\sigma^2 k^s \qquad \text{for } s = 1, \ldots, S_{(0)},$$

where $\sigma^2$ represents the variance of the error term and $k^s$ is the number of regressors. Since most big data sets exhibit strong heteroscedasticity, we concentrate on HRMS in this paper.

Our HRMS method is inspired by the FW method of Claeskens et al. (2006). Zhang et al. (2012) extended the original FW method to FW-AIC by using the AIC of Akaike (1973) as the selection criterion. As demonstrated in Lehrer and Xie (2016), a simplified version of the automatic GETS approach by Campos et al. (2003) can be used for the model screening process. The ARMS method of Yuan and Yang (2005) also explores the full model set and selects the top $M'$ models according to AIC scores. See Appendix C for more details on the model screening methods.

## 3. Simulation

In this section, we conduct Monte Carlo simulations to investigate the performance of our HRMS method and compare it with those of the GETS, ARMS, and FW-AIC methods. Similar to Liu and Okui (2013) and Zhao et al. (2016), we consider the DGP

$$y_t = \mu_t + e_t = \sum_{j=1}^{\infty} \beta_j x_{jt} + e_t$$

for $t = 1, \ldots, n$. The coefficients are generated by $\beta_j = cj^{-1}$, where $c$ is a parameter we control such that $R^2 = c^2/(1 + c^2)$ varies in $\{0.1, \ldots, 0.9\}$. We set $x_{1t} = 1$ and other $x_{jt}$ follows $\text{N}(0, 1)$ independently. Since the infinite series of $x_{jt}$ is infeasible in practice, we truncate the process at $j_{\max} = 10{,}000$. The error term $e_t$ follows $\text{N}(0, x_{2t}^2)$. We consider 4 different sample sizes where $n = 100, 200, 300$, and $400$. We assume that we can only observe the first 20 regressors. A full permutation of the $K = 20$ regressors leads to 1,048,575 candidate models (the null model is ignored).

We construct four candidate model sets: $\mathcal{M}_{\text{GETS}}^K$, $\mathcal{M}_{\text{ARMS}}^K$, $\mathcal{M}_{\text{FW-AIC}}^K$, and $\mathcal{M}_{\text{HRMS}}^K$, using the four model screening methods implied in the subscripts. The pre-determined parameters for the GETS and the ARMS methods are $p = 0.1$ and $M' = 20$, respectively, and the initial model $\mathbb{M}_{(0)}$ is set as null for both the FW-AIC and the HRMS methods. Then, we evaluate the performance of the four methods by comparing their risks, such that

$$\text{Risk}_i \equiv \frac{1}{n} \sum_{t=1}^{n} \left( \hat{\mu}_t(\mathcal{M}_i^K) - \mu_t \right)^2$$

for $i =$ GETS, ARMS, FW-AIC, and HRMS,