NBER WORKING PAPER SERIES

THE BIGGER PICTURE: COMBINING ECONOMETRICS WITH ANALYTICS IMPROVE FORECASTS OF MOVIE SUCCESS

Steven F. Lehrer
Tian Xie

Working Paper 24755
http://www.nber.org/papers/w24755

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2018

We wish to thank Chris Hansen, seminar participants at the Young Econometricians around Pacific (YEAP) 2017 annual conference, the Canadian Econometrics Study Group (CESG) 2017 annual conference, Carleton University, Chinese Academy of Sciences, Northeastern University, Renmin University, Xiamen University, and Zhejiang University for helpful comments and suggestions. Xie's research is supported by the Natural Science Foundation of China (71701175), the Chinese Ministry of Education Project of Humanities and Social Sciences (17YJC790174), the Natural Science Foundation of Fujian Province of China (2018J01116), the Fundamental Research Funds for the Central Universities in China (20720171002, 20720171076, and 20720181050), and Educational and Scientific Research Program for Young and Middleaged Instructor of Fujian Province (JAS170018). Lehrer wishes to thank SSHRC for research support. The usual caveat applies. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Bigger Picture: Combining Econometrics with Analytics Improve Forecasts of Movie Success

Steven F. Lehrer and Tian Xie

NBER Working Paper No. 24755
June 2018
JEL No. C52,C53,C55

ABSTRACT

There exists significant hype regarding how much machine learning and incorporating social media data can improve forecast accuracy in commercial applications. To assess if the hype is warranted, we use data from the film industry in simulation experiments that contrast econometric approaches with tools from the predictive analytics literature. Further, we propose new strategies that combine elements from each literature in a bid to capture richer patterns of heterogeneity in the underlying relationship governing revenue. Our results demonstrate the importance of social media data and value from hybrid strategies that combine econometrics and machine learning when conducting forecasts with new big data sources. Specifically, while recursive partitioning strategies greatly outperform dimension reduction strategies and traditional econometric approaches in forecast accuracy, there are further significant gains from using hybrid approaches. Further, Monte Carlo experiments demonstrate that these benefits arise from the significant heterogeneity in how social media measures and other film characteristics influence box office outcomes.

Steven F. Lehrer
School of Policy Studies
and Department of Economics
Queen's University
Kingston, ON K7L 3N6
CANADA
and NBER
lehrers@queensu.ca

Tian Xie
Wang Yanan Institute for Studies in Economics
Department of Finance, MOE Key Lab of Econometric
Xiamen, Fujian 361005, China
xietian001@hotmail.com

1 Introduction

Many speculate that in the near future, movie studios will find that predictive analytics may play just as large of a role as either the producer, director, and/or stars of the film when determining if it will be a success. Currently, predictive analytics that incorporate social media data are being predominately used for demand forecasting exercises in the film industry. Improved forecasts are valuable since they not only could increase capital investments by reducing investor uncertainty of the box office consequences, but also help marketing teams tailor effective advertising campaigns. However, there remains both skepticism as to whether social media data truly adds value to these forecasting exercises, and debate if the hype of its potential is truly warranted.

In this paper we introduce new strategies for predictive analytics that are contrasted with existing tools from both the econometrics and machine learning literature to first give guidance on how to improve forecast accuracy in applications within the film industry.[1] Second, we consider the value of different measures of social media data in our application. We examine if there is additional value from analyzing individual social media messages for trends in sentiment that are beyond simply collecting the amount of social media chatter surrounding individual movies.

This paper contributes to a burgeoning literature in the emerging fields of data science and analytics that focuses on developing methods to improve empirical practice including forecast accuracy.[2] Motivating our strategies is that in many forecasting exercises involving social media data we would anticipate heteroskedasticity for at least two reasons. First, the characteristics of individuals attracted to different films will differ sharply leading the data to appear as if coming from different distributions. Second, online respondents may have greater unobserved variability in their opinions of different films.

[1] As discussed in the next section, traditional econometric strategies generally differ from machine learning approaches by first writing an explicit model for how covariates influence outcomes.

[2] For example, among other developments, Vasilios, Theophilos, and Periklis (2015) examine the forecasting accuracy of machine learning techniques on forecasting daily and monthly exchange rates, Wager and Athey (2017) propose variants of random forests to estimate causal effects, and Ban, Karoui, and Lim (2018) adopted machine learning methods for portfolio optimization.

Econometricians have long been well aware that heteroskedasticity of data impacts the predictive ability of many traditional estimators. With recursive partitioning strategies such as regression trees that underlie ensemble methods such as random forests or bagging, researchers do not specify a structure for the model to forecast the mean, but implicitly assume homogeneous variance across the entire explanatory-variable space.[3]

We propose two new empirical strategies for this setting. First, we develop computationally efficient methods to implement model averaging estimators with heteroskedastic data. Specifically, we extend the theoretical results of Zhang, Yu, Zou, and Liang (2016) and prove the asymptotic optimality of Mallows-type model averaging estimators using a set of screened candidate models.[4] Second, we propose a hybrid strategy that uses recursive partitioning methods to develop subgroups and then undertake model averaging within these groups to generate forecasts. Traditionally, forecasts from regression trees assume homogeneity in outcomes within individual leafs. By allowing for model uncertainty in the leaves, richer forms of heterogeneity in the relationships between independent variables and outcomes within each subgroup in a leaf is allowed.

To examine the empirical performance of these alternative approaches we first extend the prediction exercise in Lehrer and Xie (2017) by removing the sampling criteria based on the budget. All movies released over a three-year period ranging from art-house to blockbuster are now included. By relaxing this restriction, the data exhibits strong heteroskedasticity,[5] which likely arises since different films appeal to populations drawn

[3]More generally, both OLS, regression trees and Lasso methods rely on the unweighted sum of squares criterion (SSR), which implicitly assumes homoskedastic errors. It is well known that when this condition is violated and heteroskedasticity is present, the standard errors are biased influencing statistical inference procedures. Further, the objective function ensures that areas of high variability will contribute more to minimizing the unweighted SSR, and will therefore play a larger role when making predictions at the mean. As such, predictions for low-variance areas are expected to be less accurate relative to high variance areas. This is why heteroskedasticity might affect predictions at the mean, since the implicit weights to the data are determined by the local variance. Recent developments continue to use the SSR as a loss function but can generally accommodate richer forms of heterogeneity relative to parametric econometric models by accounting for limited forms of parameter heterogeneity.

[4]Specifically, we consider three model averaging estimators: the MMA estimator of Hansen (2007), the PMA estimator of Xie (2015), and the HRC$_p$ of Liu and Okui (2013).

[5]Results from Breusch-Pagan test are presented in appendix E.1. We should stress that a reason one needs to account for heteroskedasticity is parameter heterogeneity, which is a form of an omitted variables

from different distributions.

Our results provide new insights on the trade-offs researchers face when choosing a forecasting method. Recursive partitioning strategies including regression trees, bagging and random forests yield on average a 30-40% gains in forecast accuracy relative to econometric approaches that either use a model selection criteria or model averaging approach. These large gains from statistical learning methods even relative to econometric estimators and penalization methods that implicitly account for heteroskedastic data, demonstrate the restrictiveness of linear parametric econometric models. These models remain popular in econometrics since as Manski (2004) writes "statisticians studying estimation have long made progress by restricting attention to tractable classes of estimators; for example, linear unbiased or asymptotic normal ones".

Second, we find additional gains of roughly 10% in forecast accuracy from our proposed strategy that allows for model uncertainty in each leaf of a regression tree relative to popular recursive partitioning algorithms such as random forest and bagging.[6] Monte Carlo experiments clarify why these gains arise in our empirical application. We find hybrid strategies are quite useful in settings where heteroskedasticy arises due to significant parameter heterogeneity, perhaps due to jumps or threshold effects, or simply neglected parameter heterogeneity in the underlying behavioral relationships. In this setting, hybrid strategies can explain a portion of the significant amount of heterogeneity in outcomes within each leaf of a bagging tree. In contrast, when heteroskedasticy is due to random factors, we do not observe significant benefits from the hybrid strategies.

Third, our analysis finds tremendous value from incorporating social media data in forecasting exercises. Econometric approaches show that the inclusion of social media

problem. However, the link between neglected parameter heterogeneity and heteroskedasticity are not well known among practitioners, but can be easily explained with the following example. If regression coefficients vary across films (perhaps the role of Twitter volume on box office revenue differs for a blockbuster science fiction film relative to an art house drama), then the variance of the error term varies too for a fixed-coefficient model. This link between neglected heterogeneity and heteroskedasticity has implications for specification tests and Chesher (1984) demonstrates that the well-known information matrix (IM) test due to White (1982) can be interpreted as a test against random parameter variation.

[6]Note, our analysis finds that adding model averaging post variable selection by penalization methods or using a model screening approach leads to small gains relative to traditional econometric approaches.

data leads to large gains in forecast accuracy. Calculations of variable importance from recursive partitioning methods show that measures of social media message volume account for 6 of the 10 most influential variables when forecasting either box office or retail movie unit sales revenue.

This paper is organized as follows. In the next section, we first review traditional econometric and machine learning strategies to conduct forecasting and then propose two new strategies to aid managerial decision making. The strategies are designed to be computationally efficient and can accommodate more general forms of heterogeneity than traditional forecasting methods. The data used and design of the simulation experiments that compares forecasting methods is outlined in section 3. Section 4 presents and discusses our main results that show the value of social media data and combining machine learning with econometrics when undertaking forecasts. Further, we conduct additional Monte Carlo experiments to elucidate why an understanding of the source of heteroskedasticity is useful when selecting a forecasting method. We summarize our findings and conclude in the final section.

2 Empirical Tools for Forecasting

Forecasting involves a choice of a method to identify the underlying factors that might influence the variable being predicted. Econometric approaches begin by considering a linear parametric form for the data generating process (DGP) of this variable as

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

for $i = 1, ..., n$ and μ_i can be considered as the conditional mean $\mu_i = \mu(x_i) = \mathbb{E}(y_i|x_i)$ that is converging in mean square.[7] The error term can be heteroskedastic, where $\sigma_i^2 = \mathbb{E}(e_i^2|x_i)$ denote the conditional variance that depends on x_i. Since the DGP in equation

[7] Convergence in mean square implies that $\mathbb{E}(\mu_i - \sum_{j=1}^{k} \beta_j x_{ij})^2 \to 0$ as $k \to \infty$.

(1) is unknown, econometricians often approximate it with a set of M candidate models:

$$y_i = \sum_{j=1}^{k^m} \beta_j^m x_{ij}^m + u_i, \tag{2}$$

for $m = 1, ..., M$, where x_{ij}^m for $j = 1, ..., k^m$ denotes the regressors, β_j^m denotes the coefficients. The residual now contains both the original error term and a modeling bias term denoted as $b_i^m \equiv \mu_i - \sum_{j=1}^{k^m} \beta_j^m x_{ij}^m$. In practice, econometricians often use specification tests such as the Akaike information criterion to determine a single preferred model.

In the machine learning literature, many popular algorithms select variables rather than models. For example, recursive partitioning methods such as classification and regression decision trees (CART) use a fast divide and conquer greedy algorithm that recursively partitions the data into smaller subsets.[8] A node τ containing n_τ observations with mean outcome $\bar{y}(\tau)$ can only by split by one selected variable into two leaves, denoted as τ_L and τ_R. The split is made at the variable where $\Delta = \text{SSR}(\tau) - \text{SSR}(\tau_L) - \text{SSR}(\tau_R)$, reaches its global maximum;[9] where the within-node sum of squares is $\text{SSR}(\tau) = \sum_i^{n_\tau} (y_i - \bar{y}_\tau)^2$. This splitting process continues at each new node until the $\bar{y}(\tau)$ at nodes can no longer be split since it will not add any additional value to the prediction. Forecasts at each final leaf l are the fitted value from a regression model of

$$y_i = a + u_i, \quad i \in l, \tag{3}$$

where u_i is the error term and a stands for a constant term. The least square estimate of $\hat{a} = \bar{y}_{i \in l}$. In other words, after partitioning the dataset into numerous final leaf nodes, the

[8] Regression trees are applied to real number predicted outcomes and differ sharply from econometric strategies by not linearizing the relationship in (1). These strategies aim to estimate $y = f(x)$ while trying to avoid overfitting. Further, smoothness conditions are not required in contrast to many nonparametric approaches in econometrics. Work on tree-based regression models traces back to Morgan and Sonquist (1963) and within machine learning, most research efforts concentrate on classification (or decision) trees (Hunt, Martin, and Stone, 1966, Quinlan, 1986) and work on regression trees started with RETIS (Karalic and Cestnik, 1991) and M5 (Quinlan, 1992).

[9] Implicitly it is assumed that there are no unobservables relevant to the estimation. That said, the standard methodology to induce regression trees is based on the minimization of the squared error.

forecast assumes any heterogeneity in outcomes within each subgroup is random. From the perspective of the econometrician, this can appear unsatisfying.

In addition, Hastie, Tibshirani, and Friedman (2009) discuss that individual regression trees are not powerful predictors relative to ensemble methods since they exhibit large variance. Ensemble methods that combine estimates from multiple models or trees exist in both the machine learning and econometrics literature. The model averaging literature in econometrics assumes that a weighted average of M linear candidate models can approximate the DGP in equation (1).[10] This strategy can also be motivated by the researcher being uncertain about the appropriate specification. Since M models approximate the DGP as given by $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{e}$, where $\boldsymbol{y} = [y_1, ..., y_M]^\top$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_M]^\top$, $\boldsymbol{e} = [e_1, ..., e_M]^\top$ and we define the variable $\boldsymbol{w} = [w_1, w_2, ..., w_M]^\top$ as a weight vector in the unit simplex in $\mathbb{R}^M$,

$$\mathcal{H} \equiv \left\{ w_m \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\}. \tag{4}$$

There are numerous optimization routines used to estimate these weights and each aims to strike a balance between model performance and complexity of the individual models.[11] Once the optimal weights are obtained, the forecast from the model averaging estimator of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}}(\boldsymbol{w}) = \sum_{m=1}^{M} w_m \hat{\boldsymbol{\mu}}^m = \sum_{m=1}^{M} w_m \boldsymbol{P}^m \boldsymbol{y} = \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}. \tag{5}$$

This forecast is a weighted average of the forecasts of the individual candidate models.

Within machine learning, bootstrap aggregating decision trees (aka bagging) proposed in Breiman (1996) and random forest developed in Breiman (2001) are randomization-based ensemble methods that draw a parallel to model averaging.[12] In bagging, trees are

[10] That is, define the estimator of the m^{th} individual model is given as $\hat{\boldsymbol{\mu}}^m = \boldsymbol{X}^m \left(\boldsymbol{X}^{m\top}\boldsymbol{X}^m\right)^{-1}\boldsymbol{X}^{m\top}\boldsymbol{y} = \boldsymbol{P}^m\boldsymbol{y}$, where $\boldsymbol{X}^m$ is a full rank $n \times k^m$ matrix of independent variables with $(i,j)^{th}$ element being x_{ij}^m and $\boldsymbol{P}^m = \boldsymbol{X}^m(\boldsymbol{X}^{m\top}\boldsymbol{X}^m)^{-1}\boldsymbol{X}^{m\top}$. Similarly, the residual is $\hat{\boldsymbol{e}}^m = \boldsymbol{y} - \hat{\boldsymbol{\mu}}^m = (\boldsymbol{I}_n - \boldsymbol{P}^m)\boldsymbol{y}$ for all m.

[11] Several of these methods are considered in our empirical exercise and described in the appendix.

[12] The main idea is to introduce random perturbations into the learning procedure by growing multiple different decision trees from a single learning set and then an aggregation technique is used to combine the predictions from all these trees.These perturbations help remedy the fact that a single tree may suffer from

built on random bootstrap copies of the original data, producing multiple different trees. Bagging differs from random forest only in the set of explanatory factors being considered in each tree. That is, rather than consider which among the full set of explanatory variables leads to the best split at a node of the tree, random forests only consider a random subset of the predictor variables for the best split. With both strategies, the final forecast is obtained as an equal weight average of the individual tree forecasts.

Forecasts from recursive partitioning and model averaging methods are computationally expensive but differ in three important ways. The first difference relates to how the DGP in equation (1) is approximated and both bagging and random forest do not make any assumptions about the probabilistic structure of the data. The remaining two differences relate to how predictions are weighted across the different models/trees. Optimal weights across models are calculated using equation (4) from predictions using the full sample in model averaging strategies. The weight of each leaf in the tree forecast is simply determined by the sample proportion in each leaf. Second, final predictions from regression trees rule out any model uncertainty in each final leaf $\bar{y}(\tau)$ of the tree.

This lack of heterogeneity and computational considerations motivate our two proposed extensions for forecasting with social media data. The first extension considers an improved method to select candidate models for model averaging estimators. The second extension proposes a hybrid strategy that combines recursive partitioning with model averaging to allow for heterogeneity in forecasts when the final leaf subgroup consists of observations that differ in some observed covariates.

2.1 A New Strategy for Model Screening

The empirical performance of any model averaging estimator crucially depends on the candidate model set. Let $\mathcal{M}$ denote the candidate model set before screening. In practice, one possible approach to construct the candidate model set is to consider a full permu-

high variance and display poor forecast accuracy. See appendix A for more details.

tation of all regressors. One obvious drawback of this approach is that the total number of candidate models increases exponentially with the number of regressors. As shown in Wan, Zhang, and Zou (2010), Xie (2015), Zhang, Zou, and Carroll (2015), among others, by either keeping the total number of candidate models to be small or letting the total number of candidate models converge to infinity slow enough, provides a necessary condition to maintain the asymptotic optimality of model averaging estimators.[13] While most existing research assumes a pre-determined candidate model set, a recent paper by Zhang, Yu, Zou, and Liang (2016) established the asymptotic optimality of Kullback-Leibler (KL) loss based model averaging estimators with screened candidate models. Following this insight, we define $\tilde{\mathcal{M}}$ to be the candidate model set following model screening, in which $\tilde{\mathcal{M}} \subseteq \mathcal{M}$. The weight vector space solved via an optimization routine under $\tilde{\mathcal{M}}$ can be written as

$$\tilde{\mathcal{H}} = \left\{ \boldsymbol{w} \in [0,1]^M : \sum_{m \in \tilde{\mathcal{M}}} w_m = 1 \text{ and } \sum_{m \notin \tilde{\mathcal{M}}} w_m = 0 \right\}. \tag{6}$$

Note that the weight vector under $\tilde{\mathcal{M}}$ is still $M \times 1$, however, models that do not belong in $\tilde{\mathcal{M}}$ are assigned zero weight.

We define the average squared loss as $L(\boldsymbol{w}) = (\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu})^T(\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu})$ where $\hat{\boldsymbol{\mu}}(\boldsymbol{w})$ is defined in (A21). We present the following set of assumptions

Assumption 1 *We assume that there exist a non-negative series of v_n and a weight series of $\boldsymbol{w}_n \in \mathcal{H}$ such that*

(i) $v_n \equiv L(\boldsymbol{w}_n) - \inf_{\boldsymbol{w} \in \mathcal{H}} L(\boldsymbol{w})$,

(ii) $\xi_n^{-1} v_n \to 0$,

(iii) $Pr(\boldsymbol{w}_n \in \tilde{\mathcal{H}}) \to 1$ *as* $n \to \infty$,

where $\tilde{\mathcal{H}}$ is defined in (6) and ξ_n is the (lowest) modified model risk defined in equation (A6).

[13] Moreover, Hansen (2014) and Zhang, Ullah, and Zhao (2016) point out that to satisfy the conditions on the global dominance of averaging estimators over the unrestricted least-squares estimator, the number of candidate models should be limited by screening and every possible model should not be estimated.

Assumption 1(i) is the definition of v_n, which is the distance between a model risk by $\boldsymbol{w}_n$ and the lowest possible model risk. Assumption 1(ii) is a convergence condition. It requires that ξ_n goes to infinity faster than v_n. The final item of Assumption 1 implies the validity of our selected model screening techniques. When the sample size goes to infinity, the chance that the model screening techniques accidentally omit at least one useful model goes to 0. This condition is easily satisfied by imposing mild screening conditions, while keeping the candidate models in $\tilde{\mathcal{M}}$ to be as many as allowed.

The following theorem establishes the asymptotic optimality of Mallows-type model averaging estimators under screened model set.

Theorem 1 *Let Assumption 1 be satisfied, then under the conditions that sustain the asymptotic optimality of Mallows-type model averaging estimators under given (unscreened) candidate model set, we have*

$$\frac{L(\tilde{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{H}} L(\boldsymbol{w})} \xrightarrow{p} 1, \tag{7}$$

as $n \to \infty$.

The proof appears in appendix C. Theorem 1 states that using screened model set $\tilde{\mathcal{M}}$, the model averaging estimator $\tilde{\boldsymbol{w}}$ is asymptotically optimal in the sense of achieving the lowest possible mean squared error (model risk); even compared to a model averaging estimator that used all potential candidate models in its set.

2.2 New Hybrid Approaches: Model Averaging Learning Methods

Building off the idea of Belloni and Chernozhukov (2013) who suggest using OLS estimates after variable selection by the Lasso,[14] Lehrer and Xie (2017) proposed model averaging from models constructed with variables selected by the Lasso. We suggest that at each tree leaf in the forest, there may be a sequence of $m = 1, ..., M$ linear candidate models, in which regressors of each model m is a subset of the regressors belonging to

[14]Penalization methods such as the Lasso have objective functions designed to reduce the dimensionality of explanatory variables.

that tree leaf. The regressors $\boldsymbol{X}_{i\in l}^{m}$ for each candidate model within each tree leaf is constructed such that the number of regressors $k_l^m \ll n_l$ for all m. Using these candidate models, we perform model averaging estimation and obtain

$$\underset{(K\times 1)}{\hat{\boldsymbol{\beta}}_l(\boldsymbol{w})} = \sum_{m=1}^{M} w^m \underset{(K\times 1)}{\tilde{\boldsymbol{\beta}}_l^m}, \tag{8}$$

which is a weighted averaged of the "stretched" estimated coefficient $\tilde{\boldsymbol{\beta}}_l^m$ for each candidate model m. Note that the $K \times 1$ sparse coefficient $\tilde{\boldsymbol{\beta}}_l^m$ is constructed from the $k_l^m \times 1$ least squares coefficient $\hat{\boldsymbol{\beta}}_l^m$ by filling the extra $K - k_l^m$ elements with 0s. The forecast for all observations can then be obtained as

$$\hat{y}_{t\in l} = \boldsymbol{X}_{t\in l}^{p} \hat{\boldsymbol{\beta}}_l(\boldsymbol{w}). \tag{9}$$

This strategy preserves the original classification process and within each leaf allows observations that differ in characteristics to generate different forecasts $\hat{y}_{t\in l}$.

Model averaging bagging (MAB) applies this process to each of the B samples used to construct a bagging tree. The final MAB forecast remains the equal weight average of the B model averaged tree forecasts. Model averaging random forest (MARF) operates similarly with the exception that only k predictors out of the total K predictors are considered for the split at each node. With fewer predictors, the candidate model set for each leaf does not potentially consider each of the K regressors as in MAB, but rather is constructed with the k regressors used to split the nodes that generated this leaf l.[15] This restriction affects how $\hat{\boldsymbol{\beta}}_l(\boldsymbol{w})$ is calculated as it is averaged only over those leafs where it was randomly selected.

[15] In a forecast exercise, the predicting observations $\boldsymbol{X}_t^p$ with $t = 1, 2, ..., T$ are dropped down the regression tree. For each $\boldsymbol{X}_t^p$, after several steps of classification, we end up with one particular tree leaf l. We denote the predicting observations that are classified in tree leaf l as $\boldsymbol{X}_{t\in l}^p$. If the full sample contains n observations, the tree leaf l contains a subset $n_l < n$ of the full sample of y, denoted as y_i with $i \in l$. Also, the sum of all n_l for each tree leaf equals n. The mean of $y_{i\in l}$ is calculated, denoted as $\bar{y}_{i\in l}$. The value $\bar{y}_{i\in l}$ is the forecast estimate of $\boldsymbol{X}_{t\in l}^p$. It is quite possible that different predicting observations $\boldsymbol{X}_t^p$ and $\boldsymbol{X}_s^p$ with $t \neq s$ will end up with the same tree leaf, therefore, generates identical forecasts.

3 Data and Empirical Exercise

We collected data on the universe of movies released in North America between October 1, 2010 and June 30, 2013. We extend the analysis in Lehrer and Xie (2017) that concentrated solely on movies with budgets ranging from 20 to 100 million dollars and consider the full suite of films released during this period.[16] With the assistance of the IHS film consulting unit the characteristics of each film were characterized by a series of indicator variables to describe the film's genre,[17] the rating of a film's content provided by the Motion Picture Association of America's system,[18] film budget excluding advertising and both the pre-determined number of weeks and screens the film studio forecasted the specific film will be in theatres measured approximately six weeks prior to opening. In our analysis, we examine the initial demand by using the actual opening weekend box office and total sales of both DVD and Blu-Rays upon initial release.

To measure purchasing intentions from the universe of Twitter messages (on average, approximately 350 million tweets per day) we consider two measures. First, the sentiment specific to a particular film is calculated using an algorithm based on Hannak et al. (2012) that involves textual analysis of movie titles and movie key words.[19] In each Twitter message that mentions a specific film title or key word, sentiment is calculated by examining the emotion words and icons that are captured within.[20] The sentiment index for a film is the average of the sentiment of the scored words in all of the messages associated with a specific film. Second, we calculate the total unweighted volume of Twitter

[16] Movies with budgets above 100 million dollars are usually regarded as "Blockbusters" and many "Arthouse" movies usually have budgets below 20 million dollars.

[17] In total, we have 14 genres: Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, Horror, Mystery, Romance, Sci-Fi, and Thriller.

[18] Specifically, films in our sample were assigned ratings of PG, PG13, and R. There are very few movies in our data set that were given a G rating.

[19] This algorithm developed by Janys Analytics for IHS-Markit was also used for the initial reported measures of the Wall Street Journal-IHS U.S. Sentiment Index

[20] In total, each of 75,065 unique emotion words and icons that appeared in at least 20 tweets between January 1st, 2009 to September 1st, 2009 is given a specific value that is determined using emotional valence. Note that Twitter messages were capped at 140 characters throughout this period. These messages often contain acronyms and Twitter specific syntax such as hashtags that may present challenges to traditional sentiment inference algorithms.

messages for each specific film. We consider volume separate from sentiment in our analyses since the latter may capture perceptions of quality, whereas volume may just proxy for popularity.[21]

Across all the films in our sample, there is a total of 4,155,688 messages to be assessed. There is a large amount of time-varying fluctuations in both the number of, and sentiment within the Twitter messages regarding each film. Some of this variation reflects responses to the release of different marketing campaigns designed to both build awareness and increase anticipation of each film. Thus, in our application we define measures from social media data over different time periods. That is, suppose the movie release date is T, we separately calculate sentiment in ranges of days within the window corresponding to 4 weeks prior to and subsequent the release date.[22]

Summary statistics are presented in table 1. The mean budget of films is respectively approximately 61 and 63 million for the open box office and retail unit sales outcome. On average, these films were in the theatre for 14 weeks and played on roughly 3000 screens. Not surprisingly, given trends in advertising, the volume of Tweets increases sharply close to the release date and peaks that day. Following a film's release we find a steady decline in the amount of social web activity corresponding to a film.

3.1 Simulation Experiment Design

To examine the importance of incorporating data from the social web either using traditional estimators or an approach from the machine learning literature, we follow Hansen and Racine (2012) and conduct the following experiment to assess the relative prediction

[21] We consider both measures since prior work by Liu (2006) and Chintagunta, Gopinath, and Venkataraman (2010) suggest that sentiment in reviews affect subsequent box office revenue. Similarly, Xiong and Bharadwaj (2014) finds that pre-launch blog volume reflects the enthusiasts' interest, excitement and expectations about the new product and Gopinath, Chintagunta, and Venkataraman (2013) study the effects of blogs and advertising on local-market movie box office performance.

[22] For a typical range, T–a/–b, it stands for a days before date T (release date) to b days before date T. We use the sentiment data before the release date in equations that forecast the opening weekend box office. After all, reverse causality issues would exist if we include sentiment data after the release date. Similarly, T+c/+d means c days to d days after date T, which are additionally used for forecasting the retail unit sales.

Table 1: Summary Statistics

Variable	Open Box Office		Retail Unit Sales	
	Mean	Std. Dev.	Mean	Std. Dev.
Genre				
Action	0.3202	0.4679	0.3357	0.4739
Adventure	0.2416	0.4292	0.2378	0.4272
Animation	0.0843	0.2786	0.0909	0.2885
Biography	0.0393	0.1949	0.0420	0.2012
Comedy	0.3652	0.4828	0.3776	0.4865
Crime	0.1966	0.3986	0.1818	0.3871
Drama	0.3483	0.4778	0.3706	0.4847
Family	0.0562	0.2309	0.0629	0.2437
Fantasy	0.1011	0.3023	0.0909	0.2885
Horror	0.1180	0.3235	0.1049	0.3075
Mystery	0.0899	0.2868	0.0909	0.2885
Romance	0.1124	0.3167	0.0979	0.2982
Sci-Fi	0.1124	0.3167	0.1119	0.3163
Thriller	0.2416	0.4292	0.2517	0.4355
Rating				
PG	0.1461	0.3542	0.1608	0.3687
PG13	0.4213	0.4952	0.4126	0.4940
R	0.4270	0.4960	0.4196	0.4952
Core Parameters				
Budget (in million)	60.9152	56.9417	63.1287	56.5959
Weeks	13.9446	5.4486	14.4056	5.7522
Screens (in thousand)	2.9143	0.8344	2.9124	0.8498
Sentiment				
T-21/-27	73.5896	3.2758	73.4497	3.5597
T-14/-20	73.6999	3.0847	73.7530	3.0907
T-7/-13	73.8865	2.6937	73.9411	2.6163
T-4/-6	73.9027	2.7239	73.8931	2.8637
T-1/-3	73.8678	2.8676	73.7937	3.0508
T+0			73.8662	3.0887
T+1/+7			73.8241	3.1037
T+8/+14			73.4367	3.8272
T+15/+21			73.7001	3.3454
T+22/+28			74.0090	2.7392
Volume				
T-21/-27	0.1336	0.6790	0.1499	0.7564
T-14/-20	0.1599	0.6649	0.1781	0.7404
T-7/-13	0.1918	0.6647	0.2071	0.7377
T-4/-6	0.2324	0.8400	0.2494	0.9304
T-1/-3	0.4553	0.9592	0.4952	1.0538
T+0			1.5233	3.2849
T+1/+7			0.6586	1.1838
T+8/+14			0.3059	0.8290
T+15/+21			0.2180	0.7314
T+22/+28			0.1660	0.7204

efficiency of different estimators with different sets of covariates. The estimation strategies that we contrast can be grouped into the following categories (i) traditional econometric approaches, (ii) model screening approaches, (iii) machine learning approaches, and (iv) newly proposed methods that combine econometrics with machine learning algorithms to capture richer patterns of heterogeneity. Table 2 lists each estimator analyzed in the exercise and the online appendix provides further details on their implementation.

The experiment shuffles the original data with sample n, into a training set of n_T and an evaluation set of size $n_E = n - n_T$. Using the training set, we obtain the estimates from each strategy and then forecast the outcomes for the evaluation set. With these forecasts, we evaluate each of the forecasting strategies by calculating mean squared forecast error (MSFE) and mean absolute forecast error (MAFE):

$$\begin{aligned} \text{MSFE} &= \frac{1}{n_E}(y_E - x_E\hat{\beta}_T)^\top(y_E - x_E\hat{\beta}_T), \\ \text{MAFE} &= \frac{1}{n_E}\left|y_E - x_E\hat{\beta}_T\right|^\top \iota_E, \end{aligned}$$

where $(y_E,\, x_E)$ is the evaluation set, n_E is the number of observations of the evaluation set, $\hat{\beta}_T$ is the estimated coefficients by a particular model based on the training set, and ι_E is a $n_E \times 1$ vector of ones. In total, this exercise is carried out 10,001 times for different sizes of the evaluation set, $n_E = 10, 20, 30, 40$.

In total, there are $2^{23} = 8,388,608$ and $2^{29} = 536,870,912$ potential candidate models for open box office and movie unit sales respectively. This presents computational challenges for the HRC_p and other model averaging estimators. Thus, we conducted the following model screening procedure based on the GETS method to reduce the set of potential candidate models for model selection and model averaging methods. First, based on the OLS results presented in table A1, we restrict that each potential model contains a constant term and 7 (11) relatively significant parameters for open box office (movie unit sales). Second, to control the total number of potential models, a simplified version of the automatic general-to-specific approach of Campos, Hendry, and Krolzig (2003) is used for

Table 2: List of Estimators Evaluated in the Prediction Error Experiments

Panel A: Econometric Methods		
(1)	GUM	A general unrestricted model that utilize all the independent variables described above
(2)	MTV	A general unrestricted model that does not incorporate the Twitter based sentiment and volume variables
(3)	GETS	A model developed using the general to specific method of Hendry and Nielsen (2007)
(4)	AIC	A model selected using the Akaike Information Criterion method
(5)	PMA	The model selected using the prediction model averaging proposed by Xie (2015)
(6)	HPMA	The model selected using a heteroskedasticity-robust version of the PMA method discussed in appendix D
(7)	JMA	The model selected by the jackknife model averaging (Hansen and Racine, 2012)
(8)	HRC_p	The model selected by hetero-robust C_p (Liu and Okui, 2013)
(9)	$\text{OLS}_{10,12,15}$	The OLS post Lasso estimator of Belloni and Chernozhukov (2013) with 10, 12, and 15 explanatory variables selected by the Lasso
(10)	$\text{HRC}^p_{10,12,15}$	The HRC_p model averaging post Lasso estimation strategy with 10, 12, and 15 explanatory variables selected by the Lasso
Panel B: Model Screening		
(1)	GETS_s	Three threshold p-values are selected, as $p = 0.24, 0.28$, and 0.32 for open box office, and $p = 0.30, 0.34$, and 0.38 for movie unit sales
(2)	ARMSH	The modified hetero-robust adaptive regression by mixing with model screening method of Yuan and Yang (2005)
(3)	HRMS	The hetero-robust model screening of Xie (2017)
(4)	Double-Lasso	We set all tuning parameters in the two steps as equal, and we control the tuning parameter so as to select a total of 10, 12, and 15 parameters
(5)	Benchmark	The GETS method we used in previous experiments, that is, $p = 0.3$ and 0.35 for open box office and movie unit sales, respectively
Panel C: Machine Learning Methods		
(1)	RT	Regression tree of Breiman, Friedman, and Stone (1984)
(2)	BAG	Bootstrap aggregation of Breiman (1996) with $B = 1000$ bootstrap samples and all of the K^{total} covariates
(3)	RF	Random forest of Breiman (2001) with $B = 1000$ bootstrap samples and $q = \lfloor 1/3K^{total} \rfloor$ covariates
Panel D: Hybrid Methods		
(1)	MAB	Hybrid applying the PMA method on subgroups created by BAG, $B = 1000$ bootstrap samples and all of the K^{total} covariates
(2)	MARF	Hybrid applying the PMA method on subgroups created by RF, $B = 1000$ bootstrap samples and $q = \lfloor 1/3K^{total} \rfloor$ covariates

model screening.[23] While this restriction that rules out many potential candidate model may appear severe, it has been found in numerous applications including Lehrer and Xie (2017), that only a handful of models account for more than 95% of the total weight of the model averaging estimate.[24]

4 Empirical Results

The two panels of table 3 report the median MSFE and MAFE from the prediction error exercise outlined in the preceding section for the 10 different econometric strategies listed in panel A of table 2. Each row of the table considers a different size for the evaluation set and to ease interpretation all MSFEs and MAFEs are normalized by the MSFE and MAFE of the HRC^p. Panel A of table 3 presents results for forecasting open box office and panel B demonstrates results corresponding to forecasting retail movie unit sales. Notice that for open box office, all remaining entries for MSFE are larger than one, indicating inferior performance of the respective estimator relative to HRC^p. In general, the three model averaging approaches and the model selected by AIC perform nearly as well as HRC^p. For movie unit sales, HPMA yields the best results in the majority of experiments. However, the gains from using HPMA in place of PMA appear quite small.

The results in table 3 also stress the importance of social media data for forecast accuracy. Models that ignore social media data (MTV) perform poorly relative to all other strategies. Additional experiments makes clear that both social media measures are

[23]This approach explores through the whole set of potential models and examine each model using the following rule: we first estimate the p-values for testing each parameter in the model to 0. If the maximum of these p-values exceeds our benchmark value, we exclude the corresponding model. In this way, we are deleting models with weak parameters from our model set. We set the benchmark value to equal to 0.3 and 0.35 for open box office and movie unit sales respectively, which is a very mild restriction. These pre-selection restrictions lead us to retain 105 and 115 potential models for open box office and retail movie unit sales respectively. Note, we did investigate the robustness of our results to alternative benchmark values and in each case the results presented in the next section are quite similar.

[24]See appendix E.5 for a detailed discussion of the model averaging weights and top 5 models for both open box office and movie unit sales in our experiment.

needed.[25] In contrast to Lehrer and Xie (2017) we find that the post-Lasso methods listed in table 2,[26] including the double-Lasso method, OLS post Lasso and model averaging post Lasso perform poorly relative to HRC^p in this application.

Table 4 considers the performance of alternative model screening strategies listed in panel B of table 2 relative to HRC^p. We observe small gains in forecast accuracy from model screening relative to the benchmark HRC^p. The hetero-robust methods yields slightly better results than homo-efficient methods for forecasts of box office opening. In contrast, when forecasting retail movie unit sales, the homo-efficient ARMS demonstrates better results than the other screening methods.[27] Taking these findings together with the results contrasting PMA to HPMA table 3 illustrate that there are small gains in practice from using econometric approaches that accommodate heteroskedasticity.[28]

Table 5 demonstrates that are very large gains in prediction efficiency of the recursive partitioning algorithms relative to the benchmark HRC^p. For both outcomes when n_E is small, machine learning methods have dominating performance over the HRC$_p$. Popular approaches such as bagging and random forest greatly outperform the benchmark. However, our proposed MAB has the best performance when evaluating by MSFEs and adding model averaging tends to lead to gains of 10% between bagging and MAB.[29] While regression tree yields the lowest relative MAFE, random forest methods, both conventional and model averaging, have moderate performance in all cases. Note that as n_E increases, all

[25] In appendices E.3 and E.4.1 and E.6, we carried out additional prediction experiments to evaluate the forecast accuracy of alternative strategies with only a single social media measure. In each case, the evidence demonstrates markedly lower degrees of forecast accuracy relative to the corresponding exercise with two measures, thereby providing robust evidence of the need to account for both sentiment and volume.

[26] The post Lasso strategy can be viewed as a model screening method since it limits the number of explanatory variables and hence dimensionality of the candidate models. Full details on how these estimators are implemented is available in appendix B.4.

[27] Interestingly as presented in appendix E.7, the ARMS and ARMSH approaches select nearly identical weights and models.

[28] In appendix E.4, we use the Monte Carlo design introduced in section 4.2 to additionally evaluate whether the source of heteroskedasticity can explain some of these surprising results. This includes (i) the difference in the performance between PMA and HRC^p in table 3 when forecasting retail movie unit sales, and (ii) the relative improved performance of ARMS presented in table 4.

[29] In appendix E.9, we present results from the SPA test of Hansen (2005) that provide significant evidence of the superior predictive ability of the MAB method over all the other considered.

Table 3: Results of Relative Prediction Efficiency by MSFE and MAFE

n_E	GUM	MTV	GETS	AIC	PMA	HPMA	JMA	OLS_{10}	OLS_{12}	OLS_{15}	HRC^p_{10}	HRC^p_{12}	HRC^p_{15}	HRC^p
Panel A: Open Box Office														
Mean Squared Forecast Error (MSFE)														
10	1.1035	2.3032	1.2357	1.0274	1.0022	1.0018	1.0274	1.1223	1.1390	1.1208	1.1205	1.1335	1.1068	**1.0000**
20	1.1328	2.5704	1.2208	1.0246	1.0030	1.0028	1.0221	1.1634	1.1757	1.0833	1.1638	1.1717	1.0863	**1.0000**
30	1.1561	2.5402	1.2305	1.0253	1.0022	1.0012	1.0153	1.2067	1.2284	1.0769	1.2021	1.2209	1.0807	**1.0000**
40	1.1892	2.4835	1.2198	1.0215	1.0018	1.0016	1.0054	1.2160	1.2338	1.0580	1.2161	1.2314	1.0556	**1.0000**
Mean Absolute Forecast Error (MAFE)														
10	1.0597	1.5235	1.1300	1.0194	1.0012	**0.9999**	1.0060	1.0515	1.0589	1.0669	1.0543	1.0591	1.0691	1.0000
20	1.0751	1.5317	1.1258	1.0174	1.0013	**0.9998**	1.0084	1.0543	1.0604	1.0602	1.0562	1.0611	1.0615	1.0000
30	1.0814	1.5251	1.1373	1.0168	1.0026	1.0003	1.0137	1.0571	1.0681	1.0548	1.0588	1.0658	1.0564	**1.0000**
40	1.0929	1.5275	1.1376	1.0207	1.0002	1.0013	1.0038	1.0551	1.0665	1.0564	1.0560	1.0666	1.0555	**1.0000**
Panel B: Movie Unit Sales														
Mean Squared Forecast Error (MSFE)														
10	1.4183	2.4468	1.5231	1.0499	1.0013	1.0019	1.0183	1.3730	1.3481	1.3531	1.3524	1.3302	1.3449	**1.0000**
20	1.5010	2.2299	1.5895	1.0514	**0.9979**	0.9998	1.0263	1.3951	1.2665	1.2617	1.3695	1.2546	1.2498	1.0000
30	1.6988	2.1005	1.5836	1.0455	**0.9943**	0.9981	1.0218	1.3341	1.2393	1.2071	1.3104	1.2348	1.2047	1.0000
40	1.8518	1.9312	1.5235	1.0444	**0.9964**	1.0013	1.0227	1.2205	1.1579	1.1364	1.1947	1.1420	1.1252	1.0000
Mean Absolute Forecast Error (MAFE)														
10	1.1507	1.5950	1.2693	1.0296	1.0015	1.0016	1.0149	1.2354	1.2284	1.1634	1.2297	1.2211	1.1599	**1.0000**
20	1.1863	1.5342	1.2792	1.0266	1.0007	1.0009	1.0146	1.2047	1.1852	1.1365	1.1980	1.1772	1.1310	**1.0000**
30	1.2333	1.5388	1.2886	1.0312	1.0024	1.0013	1.0144	1.1904	1.1735	1.1165	1.1791	1.1642	1.1137	**1.0000**
40	1.2828	1.4793	1.2861	1.0244	**0.9983**	1.0009	1.0157	1.1551	1.1435	1.0952	1.1458	1.1365	1.0900	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column.

Table 4: Comparing Hetero-robust and Homo-efficient Model Screening Methods

n_E	GETS						ARMS						HRMS	HEMS	Benchmark
	Hetero-robust			Homo-efficient			Hetero-robust			Homo-efficient					
	$(p = 0.24$	0.28	0.32)	$(p = 0.24$	0.28	$0.32)^{\dagger}$	$(L = 100$	50	25)	$(L = 100$	50	25)			
	$(p = 0.30$	0.34	0.38)	$(p = 0.30$	0.34	$0.38)^{\ddagger}$									
Panel A: Open Box Office															
							Mean Squared Forecast Error (MSFE)								
10	0.9992	1.0040	0.9999	0.9954	0.9989	1.0021	0.9825	0.9813	**0.9751**	0.9820	0.9834	0.9926	1.0121	1.0172	1.0000
20	**0.9878**	1.0005	0.9996	0.9809	0.9971	1.0190	0.9994	0.9971	0.9908	1.0005	1.0000	0.9951	1.0143	1.0136	1.0000
30	**0.9927**	0.9991	1.0007	0.9939	1.0019	0.9997	0.9947	0.9929	1.0006	0.9987	1.0015	0.9998	1.0466	1.0283	1.0000
40	0.9921	0.9983	1.0025	0.9671	0.9990	1.0075	1.0045	0.9874	**0.9842**	1.0010	1.0094	1.0066	1.0449	1.0296	1.0000
							Mean Absolute Forecast Error (MAFE)								
10	1.0019	1.0034	1.0025	**0.9809**	1.0114	1.0037	0.9890	0.9930	1.0002	0.9904	0.9875	1.0008	1.0135	1.0143	1.0000
20	0.9955	0.9994	0.9986	0.9932	0.9978	1.0118	0.9944	0.9968	0.9956	0.9898	0.9894	**0.9863**	1.0042	1.0000	1.0000
30	0.9992	1.0015	1.0011	**0.9814**	1.0124	0.9881	0.9990	0.9976	1.0022	0.9988	0.9966	0.9972	1.0098	1.0059	1.0000
40	0.9974	1.0031	1.0020	0.9912	1.0113	0.9930	0.9954	**0.9886**	0.9930	0.9950	0.9938	0.9914	1.0172	1.0072	1.0000
Panel B: Movie Unit Sales															
							Mean Squared Forecast Error (MSFE)								
10	1.0370	1.0008	0.9940	1.0338	0.9799	0.9880	0.9620	0.9577	0.9598	0.9613	0.9504	**0.9328**	1.0481	1.0380	1.0000
20	1.0388	1.0002	0.9912	1.0374	1.0033	1.0097	0.9675	0.9713	0.9682	0.9482	0.9318	**0.9271**	1.1770	1.1245	1.0000
30	1.0309	1.0003	0.9913	1.0290	1.0010	1.0019	0.9765	0.9811	0.9843	0.9471	0.9394	**0.9344**	1.1491	1.1072	1.0000
40	1.0113	0.9977	0.9985	1.0063	1.0023	1.0004	0.9600	0.9519	0.9615	0.9316	0.9370	**0.9202**	1.2418	1.1842	1.0000
							Mean Absolute Forecast Error (MAFE)								
10	1.0122	0.9988	0.9923	1.0036	0.9881	0.9926	0.9778	0.9728	**0.9681**	0.9819	0.9828	0.9773	1.0242	1.0067	1.0000
20	1.0215	1.0001	0.9953	1.0059	1.0025	0.9859	0.9818	0.9818	0.9808	0.9814	0.9809	**0.9766**	1.0544	1.0340	1.0000
30	1.0203	1.0000	0.9966	1.0038	1.0000	1.0026	1.0014	0.9952	0.9919	0.9915	0.9920	**0.9866**	1.0637	1.0477	1.0000
40	1.0134	0.9997	0.9956	1.0213	1.0079	1.0011	0.9809	0.9742	0.9722	0.9725	0.9714	**0.9689**	1.0787	1.0664	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^{p} method presented in the last column.

learning methods observe decreases in performance. Last, note that the large gains in performance of all strategies in table 5 relative to the results presented in tables 3 and 4.

A potential explanation for the improved performance of statistical learning approaches relative to all of the econometric strategies is that the full suite of predictors is considered. Recall, that due to computational challenges we undertook model screening to reduce the number of candidate models for model averaging estimators and by so doing reduced the number of predictors. In appendix E.8, we reconsider table 5 where we restrict the set of predictors to be identical for the recursive partitioning strategies as the model screening and model averaging approaches. We continue to find large gains in forecast accuracy from random forest and bagging relative to the econometric approaches. This suggests that the gains in forecast accuracy are not from allowing a larger dimension of predictor variables, but rather likely are obtained by relaxing the linearity assumption imposed by the econometric estimator considered when constructing candidate models.

4.1 Relative Importance of the Factors

While recursive partitioning algorithms were developed to make predictions and not understand the underlying process of how predictors influence outcomes, strategies have since been developed to identify which predictor variables are the most important in making forecasts.[30] The importance of each predictor variable is first computed at the tree level, and the scores are averaged across all trees to obtain the final, global importance score for the variable.[31] The most important variables are the ones leading to the

[30] Variable importance is often computed by applied researchers but the theoretical properties and statistical mechanisms of these algorithms are not well studied. To the best of our knowledge, Ishwaran (2007) presents the sole theoretical study of tree-based variable importance measures.

[31] With bagging and random forests, each tree is grown with its respective randomly drawn bootstrap sample and the excluded data from the Out-Of-Bag sample (OOB) for that tree. The OOB sample can be used to evaluate the tree without the risk of overfitting since the observations did not build the tree. To determine importance, a given predictor is randomly permuted in the OOB sample and the prediction error of the tree on the modified OOB sample is compared with the prediction error of the tree in the untouched OOB sample. This process is repeated for both each tree and each predictor variable. The average of this gap in prediction errors across all OOB samples provides an estimate of the overall decrease in accuracy that the permutation of removing a specific predictor induced.

Table 5: Results of Relative Prediction Efficiency Between Machine Learning and Model Averaging Learning

n_E	Reg.Tree	Bagging	Random Forest			MAB	Random Forest			Benchmark
			RF_{10}	RF_{15}	RF_{20}		$MARF_{10}$	$MARF_{15}$	$MARF_{20}$	
Panel A: Open Box Office										
					Mean Squared Forecast Error (MSFE)					
10	0.6285	0.6724	0.6786	0.6683	0.6591	**0.6045**	0.6761	0.6745	0.6542	1.0000
20	0.8852	0.9145	0.8892	0.8821	0.8652	**0.8098**	0.9047	0.8861	0.8901	1.0000
30	0.9927	0.9622	0.9660	0.9267	0.9125	**0.9123**	0.9728	0.9487	0.9428	1.0000
40	1.2060	1.0225	1.0615	1.0130	1.0032	1.0281	1.0869	1.0391	1.0377	**1.0000**
					Mean Absolute Forecast Error (MAFE)					
10	0.6600	0.7246	0.7577	0.7418	0.7327	**0.6538**	0.7449	0.7291	0.7158	1.0000
20	0.7494	0.8024	0.8214	0.8005	0.7912	**0.7296**	0.8124	0.7956	0.7895	1.0000
30	0.7780	0.8079	0.8408	0.8163	0.8100	**0.7541**	0.8433	0.8137	0.8090	1.0000
40	0.8276	0.8328	0.8689	0.8417	0.8337	**0.7890**	0.8688	0.8440	0.8435	1.0000
Panel B: Movie Unit Sales										
					Mean Squared Forecast Error (MSFE)					
10	0.9732	0.8243	0.9967	0.9338	0.8646	**0.8114**	1.0026	0.9298	0.8686	1.0000
20	1.0625	0.9322	1.1735	1.0631	1.0197	**0.8914**	1.1720	1.0647	1.0244	1.0000
30	1.1463	0.9598	1.2066	1.0852	1.0592	**0.9270**	1.2086	1.0890	1.0622	1.0000
40	1.1714	1.0027	1.2310	1.1179	1.0755	**0.9734**	1.2284	1.1180	1.0777	1.0000
					Mean Absolute Forecast Error (MAFE)					
10	0.8075	0.8407	0.9471	0.8978	0.8688	**0.7740**	0.9461	0.8984	0.8683	1.0000
20	0.8722	0.8759	0.9896	0.9388	0.9145	**0.8192**	0.9891	0.9397	0.9144	1.0000
30	0.8945	0.8921	1.0009	0.9515	0.9298	**0.8269**	1.0003	0.9507	0.9314	1.0000
40	0.9143	0.9109	1.0162	0.9700	0.9418	**0.8470**	1.0166	0.9691	0.9413	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^{p} method presented in the last column.

greatest losses in accuracy.

We calculate variable importance scores using the MAB and MARF strategies where we include and exclude the social media variables as predictors.[32] Table 6 reports the top 10 most important predictors for open box office and movie unit sales in panels A and B, respectively. The results with both strategies reinforce the importance of social media data and volume related variables are found to have a greater influence on revenue outcomes than sentiment measures. Further, and perhaps unsurprising is the predetermined budget and screens as well as weeks in theatre are important predictors. Taken together, these results suggest that the amount of social media buzz is more important than the emotional content when forecasting revenue outcomes.

Table 6: Relative Importance of the Predictors

Ranking	With Twitter Variables		Without Twitter Variables	
	MAB	MARF	MAB	MARF
Panel A: Open Box Office				
1	Screens	Screens	Screens	Screens
2	Budget	Budget	Rating: R	Budget
3	Volume: T-1/-3	Volume: T-1/-3	Genre: Horror	Genre: Horror
4	Volume: T-4/-6	Volume: T-4/-6	Genre: Adventure	Weeks
5	Volume: T-7/-13	Volume: T-7/-13	Budget	Genre: Adventure
6	Volume: T-21/-27	Volume: T-14/-20	Rating: PG	Genre: Fantasy
7	Volume: T-14/-20	Genre: Adventure	Genre: Comedy	Rating: PG13
8	Sentiment: T-1/-3	Volume: T-21/-27	Genre: Animation	Rating: R
9	Weeks	Weeks	Rating: PG13	Genre: Comedy
10	Rating: R	Genre: Horror	Genre: Fantasy	Rating: PG
Panel B: Movie Unit Sales				
1	Screens	Screens	Screens	Screens
2	Budget	Budget	Weeks	Budget
3	Weeks	Weeks	Budget	Weeks
4	Volume: T+0	Volume: T+0	Genre: Comedy	Genre: Fantasy
5	Volume: T+8/+14	Volume: T+8/+14	Rating: R	Genre: Adventure
6	Volume: T+15/+21	Volume: T+1/+7	Genre: Horror	Rating: R
7	Volume: T-21/-27	Volume: T-1/-3	Genre: Fantasy	Genre: Drama
8	Volume: T+22/+28	Volume: T+15/+21	Rating: PG	Genre: Family
9	Volume: T+1/+7	Volume: T-4/-6	Genre: Thriller	Genre: Comedy
10	Volume: T-1/-3	Volume: T-21/-27	Genre: Adventure	Genre: Animation

Note: This table presents the rank order of the importance of the predictors for film revenue by the respective machine learning.

[32] We consider both MAB and MARF since Strobl et al. (2008) showed that using mean decreased accuracy in variable importance with random forests is biased and could overestimate the importance of correlated variables. This bias exists if random forest did not select the correct covariate, but rather chose a highly correlated counterpart in a bootstrapped sample. This bias should not exist with bagging strategies that use all available predictors. However, it should also be noted that the finding in Strobl et al. (2008) were not replicated in Genuer, Poggi, and Tuleau-Malot (2010).

To examine whether sentiment plays a larger role for small budget films that may benefit more from word of mouth or critical reviews, we calculated variable importance scores for films located in different budget quartile. The results are presented in table 7. Notice that constructed buzz measures are highly important for large budget films, but the messages are important for many films ranked in lower quartiles.

While the Lasso can be used to select variables to include in a regression model it does not rank them. In table 8, we report the numbers of Twitter sentiment and volume variables selected by Lasso in various samples. The results show that the Lasso favors the inclusion of sentiment variables in almost all subsamples. This difference in the importance of social media variables selected may explain the uneven prediction performance of Lasso-related estimators in tables 3 and 4. In summary, the evidence in this study continues to point to the inclusion of both social media measures and different forecasting strategies yield different rankings of the importance of each measure.

4.2 Monte Carlo Evidence: Contrasting Random Heteroskedasticity from Parameter Heterogeneity

To provide insights on when allowing for model uncertainty may improve forecasts from recursive partioning strategies, we conduct the following Monte Carlo study. Similar to Liu and Okui (2013),[33] we consider the following DGP

$$y_t = \mu_t + e_t = \sum_{j=1}^{\infty} (\beta_j + r \cdot \sigma_t) x_{jt} + e_t \tag{10}$$

for $t = 1, ..., n$. The coefficients are generated by $\beta_j = cj^{-1}$, where c is a parameter that we control, such that $R^2 = c^2/(1 + c^2)$ that varies in $\{0.1, ..., 0.9\}$. The parameter σ_t is drawn from a $N(0, 1)$ and introduces potential heterogeneity (depends on values of the scale variable r) to the model. We set $x_{1t} = 1$ and other x_{jt}s follow $\mathrm{N}(0, 1)$. Since the infinite

[33]The simulation design aims to mimic a big data environment, where the number of explanatory variables is large.

Table 7: Heterogeneity in the Relative Importance of Predictors by Film Budget

Ranking	1^{st} Quartile		2^{nd} Quartile		3^{rd} Quartile		4^{th} Quartile	
	MAB	MARF	MAB	MARF	MAB	MARF	MAB	MARF
Panel A: Open Box Office								
1	Screens	Screens	Screens	Screens	Screens	Screens	VOL: T-7/-13	VOL: T-7/-13
2	VOL: T-1/-3	VOL: T-14/-20	Weeks	Weeks	VOL: T-7/-13	VOL: T-7/-13	Screens	VOL: T-1/-3
3	Genre: Horror	VOL: T-1/-3	VOL: T-21/-27	VOL: T-21/-27	VOL: T-4/-6	VOL: T-4/-6	VOL: T-14/-20	VOL: T-4/-6
4	VOL: T-14/-20	VOL: T-7/-13	SEN: T-14/-20	Genre: Thriller	VOL: T-21/-27	VOL: T-21/-27	VOL: T-1/-3	Screens
5	VOL: T-7/-13	Genre: Horror	Rating: PG	VOL: T-14/-20	Weeks	Weeks	VOL: T-4/-6	VOL: T-14/-20
6	Genre: Thriller	VOL: T-21/-27	Genre: Crime	SEN: T-7/-13	VOL: T-1/-3	VOL: T-1/-3	Budget	VOL: T-21/-27
7	SEN: T-21/-27	VOL: T-4/-6	SEN: T-21/-27	SEN: T-4/-6	VOL: T-14/-20	VOL: T-14/-20	VOL: T-21/-27	Budget
8	Genre: Comedy	SEN: T-14/-20	Genre: Drama	SEN: T-1/-3	Rating: PG	SEN: T-4/-6	SEN: T-14/-20	SEN: T-14/-20
9	Weeks	Genre: Drama	VOL: T-14/-20	SEN: T-14/-20	Genre: Sci-Fi	Budget	Genre: Adventure	SEN: T-1/-3
10	SEN: T-14/-20	Weeks	Genre: Thriller	SEN: T-21/-27	Genre: Family	SEN: T-1/-3	SEN: T-1/-3	Rating: PG13
Panel B: Movie Unit Sales								
1	Screens	Screens	Screens	Weeks	VOL: T+8/+14	VOL: T+8/+14	Screens	Screens
2	SEN: T+22/+28	SEN: T+22/+28	Weeks	Screens	Weeks	Weeks	VOL: T-21/-27	VOL: T-21/-27
3	Weeks	VOL: T-4/-6	Genre: Family	VOL: T-21/-27	VOL: T+0	VOL: T+1/+7	VOL: T+8/+14	VOL: T+8/+14
4	VOL: T+15/+21	SEN: T+1/+7	VOL: T+1/+7	SEN: T-7/-13	VOL: T+1/+7	VOL: T+0	VOL: T-4/-6	VOL: T-4/-6
5	SEN: T-7/-13	VOL: T+1/+7	Genre: Mystery	SEN: T-1/-3	VOL: T-7/-13	VOL: T-7/-13	VOL: T-14/-20	VOL: T-7/-13
6	VOL: T-4/-6	VOL: T-14/-20	SEN: T-4/-6	VOL: T-14/-20	VOL: T+15/+21	VOL: T+15/+21	VOL: T-1/-3	VOL: T-14/-20
7	SEN: T+1/+7	VOL: T+8/+14	Genre: Drama	SEN: T-4/-6	VOL: T-21/-27	VOL: T-4/-6	VOL: T-7/-13	VOL: T-1/-3
8	SEN: T-4/-6	VOL: T+15/+21	Constant	SEN: T+1/+7	Screens	Screens	Genre: Animation	VOL: T+1/+7
9	VOL: T-14/-20	SEN: T-4/-6	Genre: Adventure	Genre: Family	VOL: T-1/-3	VOL: T-1/-3	VOL: T+1/+7	VOL: T+0
10	VOL: T-7/-13	SEN: T-7/-13	Genre: Animation	Genre: Drama	VOL: T-4/-6	VOL: T+22/+28	VOL: T+0	Genre: Animation

Note: This table presents the rank order of the importance of the predictors for film revenue by the respective machine learning in each budget subsample.

Table 8: Describing the Selected Parameters by OLS-post-Lasso

Method	1^{st} Quartile		2^{nd} Quartile		3^{rd} Quartile		4^{th} Quartile		Full Sample	
	Sentiment	Volume	Sentiment	Volume	Sentiment	Volume	Sentiment	Volume	Sentiment	Volume
Panel A: Open Box Office										
OLS_{10}	5	1	3	2	4	1	5	2	6	2
OLS_{12}	6	1	5	2	4	1	6	2	7	2
OLS_{15}	7	2	6	2	5	1	6	2	8	3
Panel B: Movie Unit Sales										
OLS_{10}	7	1	8	1	8	2	6	2	7	2
OLS_{12}	9	1	9	1	9	2	8	2	8	2
OLS_{15}	11	1	10	1	10	2	11	2	10	2

Note: Each entry in the table lists the number of respective social media variables chosen as one of the first 10 predictors among all variables in different budget subsamples 10, 12, or 15.

series of x_{jt} is infeasible in practice, we truncate the process at $j_{\max} = 10,000$ without violating our assumption on the model set-up.[34] We assume that the whole 10,000 x_{jt}s set is not entirely feasible and we can only observe the first 20 regressors. Two scenarios designed to represent pure random heteroskedasticity and heteroskedasticity that arises due to neglected parameter heterogeneity are considered. Formally,

1. **Random Heteroskedasticity**: we set the parameter $r = 0$, eliminating heterogeneity and pure random heteroskedasticity is created by drawing $e_t \sim N(0, x_{2t}^2)$.

2. **Parameter Heterogeneity**: heterogeneity in β for each observation is created by setting $r = 1/10$ and drawing $e_t \sim N(0,1)$.[35]

With this DGP, we compare the performance of conventional learning methods and model averaging learning methods using their risks.[36] We assume that the first $K = 5$

[34]Note that we can ignore variables with close-to-0 coefficients, as they have little influence on the dependent variable. Such is the case for x_{jt} with $j > j_{\max}$.

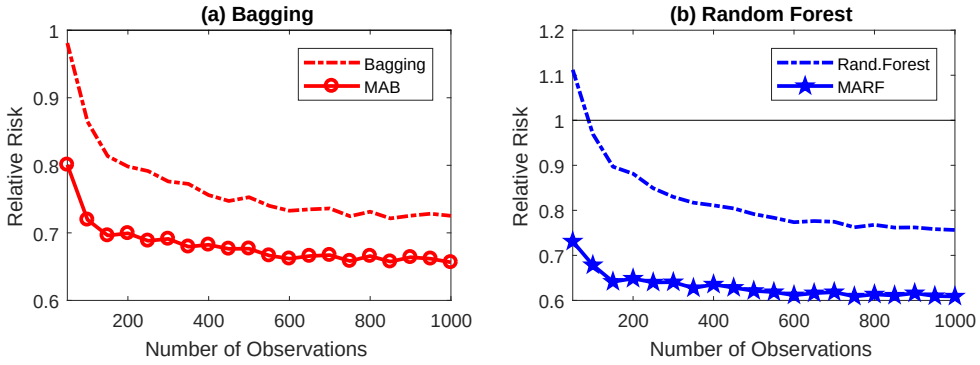
[35]Our results are robust to alternative values of r.

[36]Specifically, $\text{Risk}_i \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_i^L - \mu_i\right)^2$, where μ_i is the true fitted value (feasible in simulation) and $\hat{\mu}_i^L$ is the fitted value obtained by a specific learning method for for L = Regression Tree, Bagging, MAB, Random Forest, and MARF. For each sample size, we compute the risk for all methods and average across 1,000 simulation draws. For bagging and random forest, we set the total number of bootstraps as $B = 20$. For random forest, we randomly draw 2 regressors out of 5 to split each node. The same settings apply to the model averaging learning methods. For all model averaging learning methods, the candidate model set for each leaf contains all feasible combinations of the regressors. To ease interpretation, we normalize all risks by the risk of the generalized unrestricted model.

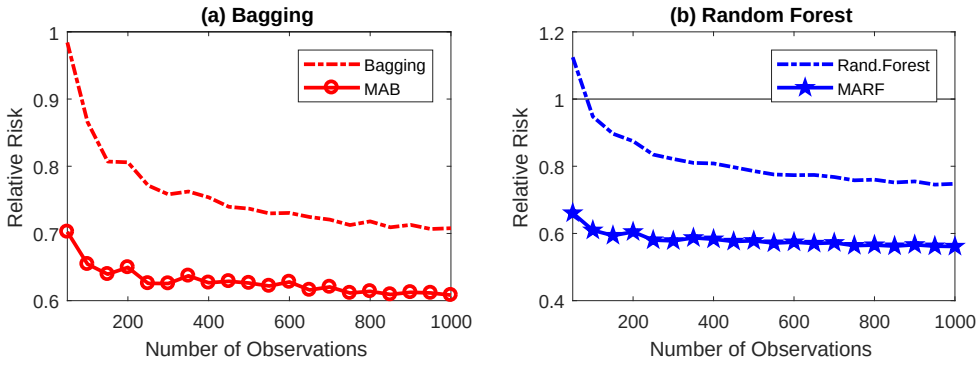
regressors are observed in both scenarios and fix the control parameter $c = 2$ when generating the true coefficients. Figure 1 panels A and B present results respectively for the random heteroskedasticity and parameter heterogeneity scenario. In each figure, the number of observations is presented on the horizontal axis, the relative risk is displayed on the vertical axis and dash-dotted (solid) lines respectively represent bagging and random forest (the model averaging counterpart). The results indicate that: i) the model averaging learning method performs much better than their respective conventional learning method in all values of n; ii) as sample sizes increase, all methods tend to yield smaller risks; and iii) MARF has the best relative performance in all cases. Overall, we observe smaller relative risks in the parameter heterogeneity scenario.

Figure 1: Relative Performance of Conventional and Model Averaging Learning

A. Random Heteroskedasticity

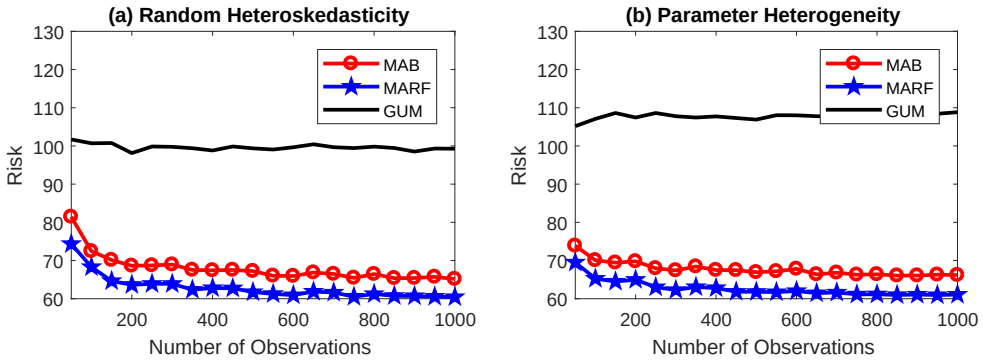


B. Parameter Heterogeneity



Since the results in figure 1 panel A are relative to generalized unrestricted model, we next present absolute risks for all model averaging learning methods along with the risks of the generalized unrestricted model in figure 2. Figure 2(a) and (b) presents results for the absolute risks under random heteroskedasticity and parameter heterogeneity, respectively. In each figure, MAB, MARF, and GUM are presented by circle-, and star-solid lines, respectively. The ranking of the methods is identical and GUM yields significantly higher risks in the parameter heterogeneity scenario. This suggests that conventional regressions suffer from efficiency loss in the presence of heterogeneity. Yet the statistical learning methods are immune to heterogeneity, since it has been acknowledged and treated during the classification process.

Figure 2: Risk Comparison under Different Scenarios



In summary, the results from the Monte Carlo experiments suggest that in settings where there may be significant parameter heterogeneity, perhaps due to jumps or threshold effects, or simply parameter heterogeneity in the underlying behavioral relationships, hybrid strategies may be quite useful. Econometric strategies that use of mean or average marginal effects simply do not allow for good forecasts when there is large heterogeneity in effects both within and across subgroups. Intuitively, this additional heterogeneity shifts to the residual, creating new outliers that change the effective weighting on different observations. In contrast, the recursive partitioning methods provide equal weights to observation within each constructed leaf subgroup, thereby ruling out heterogeneity

within groups.[37]

5 Conclusion

Several high profile economists have recently speculated that machine learning may soon transform applied econometrics in a manner similar to how economic imperialism[38] changed the evolution of research in several other social science disciplines.[39] There is also much excitement about using new sources of data from social media products in forecasting exercises. Using data from the film industry, we present evidence that this excitement for both machine learning and social media data is warranted. Specifically, recursive partitioning strategies greatly outperform dimension reduction strategies and traditional econometrics approaches in forecast accuracy. Incorporating social media data in forecasting exercises increases accuracy sharply, in part since recursive partitioning methods find that 6 of the 10 most influential variables when forecasting either box office or retail movie unit sales outcomes are from this new source.

Despite these enthusiastic findings for machine learning and social media data, we also find that even even in the era of big data, heteroskedastic data will continue to present challenges for forecasters. On the one hand, our investigation cast doubt that there are significant gains from modifying traditional econometric approaches, penalization methods or model screening methods to account for heteroskedasticity.

[37]The theoretical benefits related to most model screening methods are related to efficiency and there appears to be benefits from using machine learning approaches to shrink the number of potential models since recursive partitioning models divide the data repeatedly based on identifying differences. Model screening approaches and model averaging or Lasso methods that additionally consider heteroskedasticity do not seem to perform differently whatever the source of heteroskedasticity, and in practice yield minimal gains to approaches that treat the data as homoskedastic.

[38]Economic imperialism refers to economic analysis of seemingly non-economic aspects of life, such as politics, sociology, religion, etc. It has been asserted that these and a focus on economic efficiency have been ignored in other social sciences and "allowed economics to invade intellectual territory that was previously deemed to be outside the discipline's realm." See Mäki (2009) for a detailed discussion.

[39]Recent articles by Varian (2014), Bajari, Nekipelov, Ryan, and Yang (2015), and Athey and Imbens (2015) have discussed the possibility that computer science based analytics tools such as machine learning would make conventional statistical and econometric techniques, such as regression, obsolete.

On the other hand, we propose a hybrid strategy that applies model averaging to observations in each leaf subgroup created by either bagging or random forest and find that it leads to significant gains in forecast accuracy. These gains exist over econometric strategies and popular machine learning strategies such as random forest. To shed light on these additional gains, Monte Carlo evidence indicates that when neglected parameter heterogeneity is the underlying rationale for heteroskedasticity, gains from allowing for model uncertainty with recursive partioning are obtained. Future work is needed to not only understand the properties of hybrid strategies that combine machine learning with econometrics as well as developing tests that can detect the source of heteroskedasticity in settings with many covariates, to help guide practitioners choice of tools to undertake forecasts with social media data.

References

AKAIKE, H. (1973): "Information theory and an extension of the maximum likelihood principle," *Second International Symposium on Information Theory*, pp. 267–281.

AMEMIYA, T. (1980): "Selection of Regressors," *International Economic Review*, 21(2), 331–354.

ATHEY, S., AND G. W. IMBENS (2015): "Machine Learning for Estimating Heretogeneous Casual Effects," *Working Paper*.

BAJARI, P., D. NEKIPELOV, S. P. RYAN, AND M. YANG (2015): "Machine Learning Methods for Demand Estimation," *American Economic Review*, 105(5), 481–485.

BAN, G.-Y., N. E. KAROUI, AND A. E. B. LIM (2018): "Machine Learning and Portfolio Optimization," *Management Science*, 64(3), 1136–1154.

BARNARD, G. A. (1963): "New methods of quality control," *Journal of Royal Statistical Society Series: A*, 126, 255.

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(6), 2369–2429.

BELLONI, A., AND V. CHERNOZHUKOV (2013): "Least Squares after Model Selection in High-Dimensional Sparse Models," *Bernoulli*, 19(2), 521–547.

BELLONI, A., V. CHERNOZHUKOV, AND L. WANG (2011): "Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98(4), 791–806.

——— (2014): "Pivotal Estimation via Square-root Lasso in Nonparametric Regression," *Annals of Statistics*, 42(2), 757–788.

BOLLEN, J., H. MAO, AND X. ZHENG (2011): "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2(1), 1–8.

BREIMAN, L. (1996): "Bagging Predictors," *Machine Learning*, 26, 123–140.

——— (2001): "Random Forests," *Machine Learning*, 45, 5–32.

BREIMAN, L., J. FRIEDMAN, AND C. J. STONE (1984): *Classification and Regression Trees.* Chapman and Hall/CRC.

BREUSCH, T. S., AND A. R. PAGAN (1979): "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, 47(5), 1287–1294.

BUCKLAND, S. T., K. P. BURNHAM, AND N. H. AUGUSTIN (1997): "Model Selection: An Integral Part of Inference," *Biometrics*, 53(2), pp. 603–618.

CAMPOS, J., D. F. HENDRY, AND H.-M. KROLZIG (2003): "Consistent Model Selection by an Automatic Gets Approach," *Oxford Bulletin of Economics and Statistics*, 65(s1), 803–819.

CHERNOZHUKOV, V., C. HANSEN, AND Y. LIAO (2016): "A Lava Attack on the Recovery of Sums of Sparse and Dense Signals," *Annals of Statistics*, forthcoming.

CHESHER, A. (1984): "Testing for Neglected Heterogeneity," *Econometrica*, 52(4), 865–872.

CHINTAGUNTA, P. K., S. GOPINATH, AND S. VENKATARAMAN (2010): "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets," *Marketing Science*, 29(5), 944–957.

CLAESKENS, G., AND N. L. HJORT (2003): "The Focused Information Criterion," *Journal of the American Statistical Association*, 98(464), 900–945.

DURLAUF, S. N., S. NAVARRO, AND D. A. RIVERS (2016): "Model uncertainty and the effect of shall-issue right-to-carry laws on crime," *European Economic Review*, 81, 32 – 67.

FIRPO, S., N. M. FORTIN, AND T. LEMIEUX (2009): "Unconditional Quantile Regressions," *Econometrica*, 77(3), 953–973.

GENUER, R., J.-M. POGGI, AND C. TULEAU-MALOT (2010): "Variable Selection Using Random Forests," *Pattern Recognition Letters*, 31(14), 2225 – 2236.

GOPINATH, S., P. K. CHINTAGUNTA, AND S. VENKATARAMAN (2013): "Blogs, Advertising, and Local-Market Movie Box Office Performance," *Management Science*, 59(12), 2635–2654.

HANNAK, A., E. ANDERSON, L. F. BARRETT, S. LEHMANN, A. MISLOVE, AND M. RIEDEWALD (2012): "Tweetin ' in the Rain: Exploring Societal-scale Effects of Weather on Mood," *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pp. 479–482.

HANSEN, B. (2014): "Model Averaging, Asymptotic Risk, and Regressor Groups," *Quantitative Economics*, 5, 495–530.

HANSEN, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75(4), 1175–1189.

HANSEN, B. E., AND J. S. RACINE (2012): "Jackknife Model Averaging," *Journal of Econometrics*, 167(1), 38–46.

HANSEN, P. R. (2005): "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23(4), 365–380.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

HENDRY, D. F., AND B. NIELSEN (2007): *Econometric Modeling: A Likelihood Approach*, chap. 19, pp. 286–301. Princeton University Press.

HOETING, J. A., D. MADIAN, A. E. RAFTERY, AND C. T. VOLINSKY (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14(4), 382–417.

HUNT, E., J. MARTIN, AND P. STONE (1966): *Experiments in Induction*. Academic Press, New York.

ISHWARAN, H. (2007): "Variable Importance in Binary Regression Trees and Forests," *Electronic Journal of Statistics*, 1, 519–537.

KARABULUT, Y. (2013): "Can Facebook Predict Stock Market Activity?," *Working Paper*.

KARALIC, A., AND B. CESTNIK (1991): "The Bayesian Approach to Tree-structured Regression," *Proceedings of Information Technology Interfaces 091*.

LEAMER, E. (1978): *Specification Searches*. Wiley, New York.

LEHRER, S. F., AND T. XIE (2017): "Box Office Buzz: Does Social Media Data Steal the Show from Model Uncertainty When Forecasting for Hollywood?," *The Review of Economics and Statistics*, 99(5), 749–755.

LIANG, H., G. ZOU, A. T. K. WAN, AND X. ZHANG (2011): "Optimal Weight Choice for Frequentist Model Average Estimators," *Journal of the American Statistical Association*, 106(495), 1053–1066.

LIU, C.-A. (2015): "Distribution Theory of the Least Squares Averaging Estimator," *Journal of Econometrics*, 186, 142–159.

LIU, Q., AND R. OKUI (2013): "Heteroskedasticity-robust C_p Model Averaging," *The Econometrics Journal*, 16, 463–472.

LIU, Y. (2006): "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing*, 70(3), 74–89.

MÄKI, U. (2009): "Economics Imperialism: Concept and Constraints," *Philosophy of the Social Sciences*, 39(3), 351–380.

MALLOWS, C. L. (1973): "Some Comments on C_p," *Technometrics*, 15(4), 661–675.

MANSKI, C. F. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72(4), 1221–1246.

MISHNE, G., AND N. GLANCE (2006): "Predicting Movie Sales from Blogger Sentiment," *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

MORGAN, J. N., AND J. A. SONQUIST (1963): "Problems in the Analysis of Survey Data, and a Proposal," *Journal of the American Statistical Association*, 58(302), 415–434.

QUINLAN, J. R. (1986): "Induction of Decision Trees," *Machine Learning*, 1(1), 81–106.

——— (1992): "Learning With Continuous Classes," pp. 343–348. World Scientific.

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *The Annals of Statistics*, 6(2), 461–464.

STROBL, C., A.-L. BOULESTEIX, T. KNEIB, T. AUGUSTIN, AND A. ZEILEIS (2008): "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, 9(1), 307.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

VARIAN, H. R. (2014): "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28(2), 3–28.

VASILIOS, P., P. THEOPHILOS, AND G. PERIKLIS (2015): "Forecasting Daily and Monthly Exchange Rates with Machine Learning Techniques," *Journal of Forecasting*, 34(7), 560–573.

WAGER, S., AND S. ATHEY (2017): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, forthcoming.

WAN, A. T., X. ZHANG, AND G. ZOU (2010): "Least squares model averaging by Mallows criterion," *Journal of Econometrics*, 156(2), 277–283.

WHITE, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48(4), 817–838.

——— (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50(1), 1–25.

WHITTLE, P. (1960): "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability and Its Applications*, pp. 302–305.

XIE, T. (2015): "Prediction Model Averaging Estimator," *Economics Letters*, 131, 5–8.

——— (2017): "Heteroscedasticity-robust Model Screening: A Useful Toolkit for Model Averaging in Big Data Analytics," *Economics Letter*, 151, 119–122.

XIONG, G., AND S. BHARADWAJ (2014): "Prerelease Buzz Evolution Patterns and New Product Performance," *Marketing Science*, 33(3), 401–421.

YUAN, Z., AND Y. YANG (2005): "Combining Linear Regression Models: When and How?," *Journal of the American Statistical Association*, 100(472), 1202–1214.

ZHANG, X., A. ULLAH, AND S. ZHAO (2016): "On the dominance of Mallows model averaging estimator over ordinary least squares estimator," *Economics Letters*, 142, 69–73.

ZHANG, X., D. YU, G. ZOU, AND H. LIANG (2016): "Optimal Model Averaging Estimation for Generalized Linear Models and Generalized Linear Mixed-Effects Models," *Journal of the American Statistical Association*, 111(516), 1775–1790.

ZHANG, X., G. ZOU, AND R. J. CARROLL (2015): "Model Averaging Based on Kullback-Leibler Distance," *Statistica Sinica*, 25, 1583.

ZHAO, S., X. ZHANG, AND Y. GAO (2016): "Model Averaging with Averaging Covariance Matrix," *Economics Letters*, 145, 214–217.

APPENDIX

A Review of Popular Machine Learning Tools for Forecasting

Algorithms in machine learning build forecasting models by a series of data-driven decisions that optimize what can be learnt from the data to subsequently make predictions. Proponents of machine learning algorithms point to their improved performance in out of sample forecast exercises and stress the intuition on why they perform well, but do not consider their small sample or asymptotic properties.

The majority of machine learning tools used for forecasting explicitly assume homoskedasticity and ex ante we would expect their performance to deteriorate with heteroskedastic data. In this section we summarize why we make this conjecture with six alternative strategies. First, estimates from the least absolute selection and shrinkage operator (Lasso) are obtained by minimizing the l_1-penalized least squares criterion. The criterion involves the unweighted sum of squares and a penalty to make the model sparse. Further, as some parameter estimates are shrunk relative to traditional OLS estimates, some omitted variable bias may arise.

Breiman, Friedman, and Stone (1984) introduced the classification and regression decision trees (CART). A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute , each branch (or tree leaf) represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules. In the application in the paper, we concentrate on the regression tree case, since our predicted outcomes are real number.

A regression tree (RT) recursively partition data into groups that are as different as possible and fit the mean response for each group as its prediction. The variable and splitting point are chosen to reduce the residual sum of squares (SSR) as much as possible after the split as compared to before the split.[40] That is, similar to stepwise regression the first split is akin to choosing which variable should be first included in the model. With regression trees, splits can continue within each subgroup until some stopping rule is reached. This could lead to overfitting and as such, in practice the full trees are pruned

[40] As mentioned in the main text, in RT, a node τ contains n_τ of observations. Each node can only by split into two leaves, denoted as τ_L and τ_R, each contains subsets of n_L and n_R observations with $n_\tau = n_L + n_R$. Define the within-node sum of squares as $\text{SSR}(\tau) = \sum_i^{n_\tau}(y_i - (\bar{y}_\tau)\tilde{})^2$, where $\bar{y}(\tau)$ is the mean of those cases. We split the n_τ observations of node τ into τ_L and τ_R if the following value reach its global maximum: $\Delta = \text{SSR}(\tau) - \text{SSR}(\tau_L) - \text{SSR}(\tau_R)$. Each tree leaf τ_L or τ_R can be treated as a new node and continue with the splitting process. We start from the top of the tree (full sample) and apply the same approach to all subsequent nodes. Once a tree is constructed, the full sample is split into a number of leaves. Each leaf contains a subset of the full sample and the accumulation of all leaves is the full sample.

using a cost-complexity criterion. This criterion takes into account the amount of squared error explained by each sub-tree plus a penalty chosen by cross-validation for the number of terminal nodes in the sub-tree in an attempt to trade-off tree size and over-fitting.

Forecasts from RT involve calculating the average of the associated observations of the dependent variable in each leaf calculated and treated as the fitted value of the regression tree. Hastie, Tibshirani, and Friedman (2009) provide evidence that in practice, predictions from RT have low bias but large variance. This variance arises due to the instability of RT as very small changes in the observed data can lead to a dramatically different sequence of splits, and hence a different prediction. This instability is due to the hierarchical nature; once a split is made, it is permanent and can never be "unmade" further down in the tree. Variations of RT have been shown to have better predictive abilities and we now briefly outline the procedures of two popular approaches known as bagging and random forest.

Bootstrap aggregating decision trees, or bagging, was proposed by Breiman (1996) to improve the classification by combining classifications of randomly generated training sets. Given a standard data set $\{y_i, \boldsymbol{X}_i\}$ with $i = 1, ..., n$, bagging generates B new training sets $\{y_i, \boldsymbol{X}_i\}^b$ for $b = 1, ..., B$, in which each set is a random sample of size n replacement from the original training set $\{y_i, \boldsymbol{X}_i\}$. By sampling with replacement, some observations may be repeated and for large n the set $\{y_i, \boldsymbol{X}_i\}^b$ is expected to have the fraction $(1 - 1/e) \approx 63.2\%$ of the unique examples of $\{y_i, \boldsymbol{X}_i\}$. Each data set will construct one regression tree that is grown deep and not pruned. In a forecasting exercise, we first obtain forecasts from each tree that similar to RT has a high variance with low bias. The final forecast takes the equal weight averages of these tree forecasts and by averaging across trees, the variability of the prediction declines. Much research has found that bagging, which combines hundreds or thousands of trees, leads to sharp improvements by over a single RT.

A challenge that bagging faces is that each tree is identically distributed and in the presence of a single strong predictor in the data set, all bagged trees will select the strong predictor at the first node of the tree. Thus, all trees will look similar and be correlated. The bias of bagged trees is identical to the bias of the individual trees but the variance declines even when trees are correlated as B increases.

To reduce the chance of getting correlated trees, Breiman (2001) developed the random forest method. Random forest is similar to bagging, as both involve constructing B new trees with bootstrap samples from the original data set. But for random forest, as each tree is constructed, we take a random sample (without replacement) of q predictors out of the total K^{total} ($q < K^{total}$) predictors before each node is split. This process is repeated for each node and the default value for q is $\lfloor 1/3K^{total} \rfloor$. Note that if $q = K^{total}$, random forest is equivalent to bagging. Eventually, we end up with B trees and the final random forest estimate is calculated as the simple average of forecasts from each tree.

Research has found that random forests do a good job at forecasting when the number of relevant variables in the set K is large. After all, if there are many irrelevant variables

the chance of a split on something relevant becomes low. Yet, by randomly selecting predictors they produce trees with much lower degrees of correlation than bagging.

For space considerations, we did not consider four other methods developed in the machine learning literature. Boosted regression trees use a sequential process of fitting regression trees (without bootstrap sampling) to determine the weights of each tree in the forest. This relaxes the equal weight assumption implicit in the final forecast of random forest and bagging, but the method still relies on homoskedasticity in determining the splits at each node. Artificial neural networks and multivariate adaptive regression splines also have algorithms that make decisions assuming homoskedasticity.[41] Strategies based on Bayesian adaptive regression tree require researchers to assign priors including a functional form of the residual. In summary, heteroskedastic data is not considered with many popular tools in the machine learning literature.

B Review of Existing Methods

In this section, we review several existing heteroskedasticity-robust model averaging methods and several the Lasso methods. We summarize the theoretical conclusions and provide details on the computational algorithm used for each method.

B.1 Jackknife Model Averaging

Hansen and Racine (2012) proposed a jackknife model averaging (JMA) estimator for the linear regression model. The model set-up is identical to that provided in section 2. Hansen and Racine (2012) demonstrate the asymptotic optimality of the JMA estimator in the presence of heteroskedasticity and suggest selecting the weights by minimizing a leave-one-out cross-validation criterion

$$\text{JMA}(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{w}^\top \tilde{\boldsymbol{E}}^\top \tilde{\boldsymbol{E}}\boldsymbol{w} \quad \text{with} \quad \hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in \boldsymbol{H}^*}{\arg\min}\ \text{JMA}(\boldsymbol{w}), \tag{A1}$$

where $\tilde{\boldsymbol{E}} = [\tilde{\boldsymbol{e}}^1, ..., \tilde{\boldsymbol{e}}^M]^\top$ is an $n \times M$ matrix of jackknife residuals and $\tilde{\boldsymbol{e}}^m$ stands for the jackknife residuals of model m.

The jackknife residual vector $\tilde{\boldsymbol{e}}^m = \boldsymbol{y} - \tilde{\boldsymbol{\mu}}^m$ for model m requires the estimate of $\tilde{\boldsymbol{\mu}}^m$, where its i^{th} element, $\tilde{\mu}_i^m$, is the least squares estimator $\hat{\mu}_i^m$ computed with the i^{th} observation deleted. In practice, $\tilde{\boldsymbol{e}}^m$ can be conveniently written as $\tilde{\boldsymbol{e}}^m = \boldsymbol{D}^m\hat{\boldsymbol{e}}^m$, where $\hat{\boldsymbol{e}}^m$ is the

[41] Briefly, with artificial neural networks the weights for each node that correspond to different explanatory variables are estimated by minimizing the residual sum of squares; this approach is called back-propagation. With multivariate adaptive regression splines, terms are added to the regression model if they give the largest reduction in the residual sum of squares and to prevent over-fitting a backward deletion process is used to make the model sparse.

least squares residual vector and $\boldsymbol{D}^m$ is the $n \times n$ diagonal matrix with the i^{th} diagonal element equal to $(1 - h_i^m)^{-1}$. The term h_i^m is the i^{th} diagonal element of the projection matrix $\boldsymbol{P}^m$.

Hansen and Racine (2012) assume $\boldsymbol{H}^*$ to be a discrete set of $\left\{0, \frac{1}{N}, \frac{2}{N}, ..., 1\right\}$ for some positive integer N. Obtaining $\boldsymbol{w}$ following equation (A1) with condition $\boldsymbol{w} \in \boldsymbol{H}^*$, is a quadratic optimization process. Note that while there is a difference between our continuous $\mathcal{H}$ set defined in equation (A20) and $\mathcal{H}^*$, this should be neglectable in practice since N can take any value.

B.2 Heteroskedasticity-Robust C_p Model Averaging

Liu and Okui (2013) also use the same model set-up to propose the heteroskedasticity-robust C_p (HRC$_p$) model averaging estimator for linear regression models with heteroskedastic errors. They demonstrate the asymptotic optimality of the HRC$_p$ estimator when the error term exhibits heteroskedasticity. Liu and Okui (2013) propose computing the weights by the following feasible HRC$_p$ criterion

$$\text{HRC}_p(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}\|^2 + 2\sum_{i=1}^{n} \hat{e}_i^2 p_{ii}(\boldsymbol{w}) \tag{A2}$$

with $\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{H}}{\arg\min}\ \text{HRC}_p(\boldsymbol{w})$. Obtaining $\boldsymbol{w}$ following (A2) with condition $\boldsymbol{w} \in \mathcal{H}$ is a quadratic optimization process.

Equation (A2) includes a preliminary estimate $\hat{e}_i$ that must be obtained prior to estimation. Liu and Okui (2013) discuss several ways to obtain $\hat{e}_i$ in practice. When the models are nested, Liu and Okui (2013) suggest using the residuals from the largest model. When the models are non-nested, they recommended constructing a model that contains all the regressors in the potential models and use the predicted residuals from the estimated model. In addition, a degree-of-freedom correction on $\hat{e}_i$ is reccomended to improve finite-sample properties. For example, when the m^{th} model is used to obtain $\hat{e}_i$, we can use

$$\hat{\boldsymbol{e}} = \sqrt{n/(n-k^m)}(\boldsymbol{I} - \boldsymbol{P}^m)\boldsymbol{y}$$

instead of $(\boldsymbol{I} - \boldsymbol{P}^m)\boldsymbol{y}$ to generate the preliminary estimate $\hat{e}_i$.

B.3 Iterative HRC$_p$ Model Averaging

Liu and Okui (2013) also consider an iterative procedure in the presence of too many regressors, a common feature of big data sources. The procedure takes the following steps

1. Begin with an initial estimate $\hat{\sigma}_i$ using one selected model (Liu and Okui (2013) recommended using the largest model). This initial estimate can always be written as $\hat{\sigma}_i(\hat{\boldsymbol{w}}^0)$, with $\boldsymbol{w}^0$ being a special weight vector such that the selected model is assigned weight 1 and 0s for all other models.

2. Plug $\hat{\sigma}_i(\hat{\boldsymbol{w}}^0)$ in the HRC$_p$ criterion function defined in equation (A2) and obtain the next round $\hat{\boldsymbol{w}}^1$.

3. Using $\hat{\boldsymbol{w}}^1$, we obtain the average residual $\hat{e}_i(\hat{\boldsymbol{w}}^1)$ and hence $\hat{\sigma}_i(\boldsymbol{w}^1)$. We then use $\hat{\sigma}_i(\boldsymbol{w}^1)$ to generate the next round weight vector.

4. Repeat steps (2) and (3) until weight vector $\hat{w}^j$ is obtained that satisfies $\left|\widehat{\text{HRC}_p}(\hat{\boldsymbol{w}}^j) - \widehat{\text{HRC}_p}(\hat{\boldsymbol{w}}^{j-1})\right| \leq \varphi$, where φ is a predetermined tolerance level (usually a small number).

A problem with this iterative process is that it can be computationally demanding, since multiple steps of quadratic optimization are required. To overcome this problem, we can either choose a relatively large φ or fix the total number of iterations.

B.4 Lasso, Post Model Selection by Lasso, and Double Lasso

Consider the linear regression model:

$$y_i = \boldsymbol{x}_{0i}^\top \boldsymbol{\beta}_0 + \sum_{j=1}^{p} x_{ji}\beta_j + u_i$$

for $i = 1, ..., n$, where $\boldsymbol{x}_{0i}$ is $k_0 \times 1$ and x_{ji} is scalar for $j \geq 1$. Let

$$\begin{aligned} \boldsymbol{\beta} &= \left[\boldsymbol{\beta}_0^\top, \beta_1, ..., \beta_p\right]^\top \\ \boldsymbol{x}_i &= \left[\boldsymbol{x}_0^\top, x_{1i}, ..., x_{pi}\right]^\top \end{aligned}$$

and define the matrices $\boldsymbol{X}$ and $\boldsymbol{y}$ by stacking observations. The OLS estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$. Consider a constrained least-squares estimate $\tilde{\boldsymbol{\beta}}$ subject to the constraint $\beta_1 = \beta_2 = ... = 0$. The Lasso estimator shrinks $\hat{\boldsymbol{\beta}}$ towards $\tilde{\boldsymbol{\beta}}$ by solving

$$\hat{\boldsymbol{\beta}}^L = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p}|\beta_j|, \tag{A3}$$

where λ is the tuning parameter that controls the penalty term. In practice, researchers either assign λ to take on a specific value or use k-fold cross-validation to determine the

optimal λ. A common choice is to pick λ to minimize 5-fold cross-validation. In general, the benefits from applying the Lasso in place of OLS exist in settings where either the number of regressors exceeds the number of observations since it involves shrinkage, or in settings where the number of parameters is not small relative to the sample size and some form of regularization is necessary.

The drawback of k-fold cross-validation is its lack of computational efficiency. For example, using five-fold cross-validation, the Lasso computation procedure will need to be carried out over 200 times. This computational inefficiency becomes especially significant when either the sample size is large or the number of variables is large. Thus, we follow Belloni and Chernozhukov (2013) and ex ante pick the number of explanatory variables that will not have their coefficient shrunk to zero, a form of post model selection by Lasso.

The double-lasso regression is similar to the post model selection by Lasso. The goal is to identify covariates for inclusion in two steps, finding those that predict the dependent variable and those that predict the independent variable of interest. Without loss of generality, we focus on the case with a single focal independent variable of interest, x_{0i}, and we want to know how it relates to dependent variable y_i. The double-Lasso variable selection procedure can be carried out as follows:

Step 1. Fit a lasso regression predicting the dependent variable, and keeping track of the variables with non-zero estimated coefficients:

$$y_i = c_1 + \sum_{j=1}^{p} x_{ji}\beta_j + u_i,$$

where c_1 is a constant.

Step 2. Fit a lasso regression predicting the focal independent variable, keeping track of the variables with non-zero estimated coefficients:

$$x_{0i} = c_2 + \sum_{j=1}^{p} x_{ji}\beta_j + u_i,$$

where c_2 is a constant. If x_{0i} is an effectively randomized treatment, no covariates should be selected in this step.

Step 3. Fit a linear regression of the dependent variable on the focal independent variable, including the covariates selected in either of the first two steps:

$$y_i = c_3 + x_{0i}\beta_0 + \sum_{k\in A} x_{ki}\beta_k + u_i,$$

where c_3 is a constant, A is the union of the variables estimated to have non-zero coefficients in Steps 1 and 2.

C More Details on the Econometric Theory

In this section, we prove the asymptotic optimality of Mallows-type model averaging estimator under the constraint of screened model set. Our proof is inspired by the work of Zhang, Zou, and Carroll (2015) who demonstrated the asymptotic optimality of Kullback-Leibler (KL) type model averaging estimators under screened model set. We extend their results, allowing their findings to be applied to a broader set of model averaging estimators.

We first lay out the following conditions that have been verified in the existing literature such as White (1982).

Condition 1 *We have* $\|\boldsymbol{X}^\top \boldsymbol{\mu}_0\| = O(n)$ *and* $\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\| = O_p(n^{1/2})$.

Note that our proof is built upon the conditions that sustain the asymptotic optimality of Mallows-type model averaging estimators under given unscreened candidate model set. For example, see either equations (7) and (8) in Wan, Zhang, and Zou (2010), or assumptions 1 to 3 in Xie (2015), or assumptions 2.1 to 2.7 in Liu and Okui (2013). Condition 2 corresponds to these suppositions and would change slightly as we adopt different model averaging estimators.

Condition 2 *Conditions that sustain the asymptotic optimality of Mallows-type model averaging estimators (homoskedasticity or heteroskedasticity-robust) under given unscreened candidate model set in the original paper.*

For each approximation model m, we can define its mean squared error as

$$L(\boldsymbol{\beta}_m) \equiv (\boldsymbol{\mu}(\boldsymbol{\beta}_m) - \boldsymbol{\mu}_0)^\top (\boldsymbol{\mu}(\boldsymbol{\beta}_m) - \boldsymbol{\mu}_0), \tag{A4}$$

where $\boldsymbol{\mu}_0$ is the true value and $\boldsymbol{\mu}(\boldsymbol{\beta}_m) = \boldsymbol{X}\boldsymbol{\beta}_m$. Note that in our definition, all $\boldsymbol{\beta}_m$ for $m = 1, ..., M$ are $k \times 1$ vector, in which certain coefficients are set to 0 if the associated independent variables are not included in model m. Let $\boldsymbol{\beta}_m^*$ be the coefficient that minimizes equation (A4) such that $\boldsymbol{\beta}_m^* = \arg\min L(\boldsymbol{\beta}_m)$. The coefficient vector $\boldsymbol{\beta}_m^*$ minimizes the mean squared error of model m with respect to the true prediction value $\boldsymbol{\mu}_0$, which is different from $\hat{\boldsymbol{\beta}}_m$ that minimizes the sum squared residual (SSR) of model m.

We define the following averaged coefficients

$$\hat{\boldsymbol{\beta}}(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w_m \hat{\boldsymbol{\beta}}_m \quad \text{and} \quad \boldsymbol{\beta}^*(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w_m \boldsymbol{\beta}_m^*$$

Since $\boldsymbol{\mu}(\boldsymbol{w}) = \sum_{m=1}^{M} w_m \boldsymbol{X}\hat{\boldsymbol{\beta}}_m = \boldsymbol{X}\sum_{m=1}^{M} w_m \hat{\boldsymbol{\beta}}_m = \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{w})$, we define $\boldsymbol{\mu}^*(\boldsymbol{w}) \equiv \boldsymbol{X}\boldsymbol{\beta}^*(\boldsymbol{w})$ and the associated mean squared error can be written as

$$L^*(\boldsymbol{w}) = \big(\boldsymbol{\mu}^*(\boldsymbol{w}) - \boldsymbol{\mu}_0\big)^\top \big(\boldsymbol{\mu}^*(\boldsymbol{w}) - \boldsymbol{\mu}_0\big). \tag{A5}$$

We then define the ξ_n as

$$\xi_n = \inf_{\boldsymbol{w}\in\mathcal{H}} L^*(\boldsymbol{w}), \tag{A6}$$

which is the lowest possible value of $L^*(\boldsymbol{w})$ under set $\mathcal{H}$.

Although the mean squared error $L^*(\boldsymbol{w})$ is based on a different averaged coefficients $\hat{\boldsymbol{\beta}}^*(\boldsymbol{w})$, it is closely related to the $L(\boldsymbol{w})$ defined in (A24).

Lemma 1 *Given Conditions 1 and 2, we have*

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \frac{|L(\boldsymbol{w}) - L^*(\boldsymbol{w})|}{L^*(\boldsymbol{w})} = o_p(1), \tag{A7}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \frac{|C(\boldsymbol{w}) - \sum_{i=1}^{n}\sigma_i^2 - L^*(\boldsymbol{w})|}{L^*(\boldsymbol{w})} = o_p(1). \tag{A8}$$

Proof of Lemma 1 In line with the Theorem 3.2 of White (1982), under standard regularity conditions, it is straightforward to show that $\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m^* = O_p(n^{-1/2})$. Therefore,

$$\hat{\boldsymbol{\beta}}(\boldsymbol{w}) - \boldsymbol{\beta}^*(\boldsymbol{w}) = \sum_{m=1}^{M} w_m\big(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m^*\big) = O_p(n^{-1/2}) \tag{A9}$$

holds uniformly for $\boldsymbol{w} \in \mathcal{H}$.

By Taylor expansion and Condition 1,

$$\begin{aligned} L^*(\boldsymbol{w}) &= L(\boldsymbol{w}) + 2\boldsymbol{X}^\top\big(\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{w}) - \boldsymbol{\mu}_0\big)\big(\boldsymbol{\beta}^*(\boldsymbol{w}) - \hat{\boldsymbol{\beta}}(\boldsymbol{w})\big) + o_p(1) \\ &= L(\boldsymbol{w}) + O_p(n^{1/2}) + o_p(1), \end{aligned}$$

which implies $\sup_{\boldsymbol{w}\in\mathcal{H}} |L(\boldsymbol{w}) - L^*(\boldsymbol{w})| \le O_p(n^{1/2})$. Since $\sup_{\boldsymbol{w}\in\mathcal{H}} |L(\boldsymbol{w}) - L^*(\boldsymbol{w})|$ is in a smaller order than $L^*(\boldsymbol{w})$, we obtain (A7).

Moreover, for Mallows-type criterion, we have

$$\begin{aligned} C(\boldsymbol{w}) &= \big(\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\big)^\top\big(\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\big) + 2\sigma^2 k \\ &= L(\boldsymbol{w}) + \boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^\top\big(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\big) + 2\sigma^2 k \\ &= L^*(\boldsymbol{w}) + \big(L(\boldsymbol{w}) - L^*(\boldsymbol{w})\big) + \boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^\top\boldsymbol{X}\big(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}(\boldsymbol{w})\big) + 2\sigma^2 k. \end{aligned}$$

Therefore, by Condition 2,

$$\sup_{\boldsymbol{w}\in\mathcal{H}} |C(\boldsymbol{w}) - L^*(\boldsymbol{w})| \leq \sup_{\boldsymbol{w}\in\mathcal{H}} |L(\boldsymbol{w}) - L^*(\boldsymbol{w})| + 2\sup_{\boldsymbol{w}\in\mathcal{H}} |\boldsymbol{\epsilon}^\top \boldsymbol{X}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}(\boldsymbol{w}))| + \sum_{i=1}^{n} \sigma_i^2 + o_p(1).$$

Note that the term $\sum_{i=1}^n \sigma_i^2$ can be simplified as $n\sigma^2$ if we assume homoskedasticity. Following Condition 1 and results in (A7), we have $\sup_{\boldsymbol{w}\in\mathcal{H}} |C(\boldsymbol{w}) - \sum_{i=1}^n \sigma_i^2 - L^*(\boldsymbol{w})| \leq O_p(n^{1/2})$. Hence, we obtain (A8) and complete the proof. ∎

Once Lemma 1 is established, we can prove Theorem 1 with the following steps.

Proof of Theorem 1 Our proof follows Zhang, Yu, Zou, and Liang (2016). Define $a(\boldsymbol{w}) = C(\boldsymbol{w}) - \sum_{i=1}^n \sigma_i^2 - L(\boldsymbol{w})$. As demonstrated in Lemma 1, Assumption 1, and Conditions 1 and 2, it is straightforward to show that, as $n \to \infty$

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \left| \frac{a(\boldsymbol{w})}{L^*(\boldsymbol{w})} \right| \xrightarrow{p} 0, \tag{A10}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \left| \frac{v_n}{L^*(\boldsymbol{w})} \right| \xrightarrow{p} 0, \tag{A11}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \left| \frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w})} \right| \xrightarrow{p} 1. \tag{A12}$$

Therefore,

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \left| \frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w}) - v_n} \right| \leq \left\{ 1 - \sup_{\boldsymbol{w}\in\mathcal{H}} \left| \frac{L(\boldsymbol{w}) - L^*(\boldsymbol{w})|}{L^*(\boldsymbol{w})} \right| - \sup_{\boldsymbol{w}\in\mathcal{H}} \left| \frac{v_n}{L^*(\boldsymbol{w})} \right| \right\}^{-1} \xrightarrow{p} 0, \tag{A13}$$

as $n \to \infty$. Then, we expand equation (7) of Theorem 1 as

$$\begin{aligned}
& \Pr\left\{ \left| \frac{\inf_{\boldsymbol{w}\in\mathcal{H}} L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})} - 1 \right| > \delta \right\} \\
= \; & \Pr\left\{ \left| \frac{\inf_{\boldsymbol{w}\in\tilde{\mathcal{H}}} \left(L(\boldsymbol{w}) + a(\boldsymbol{w})\right) - a(\tilde{\boldsymbol{w}}) - \inf_{\boldsymbol{w}\in\mathcal{H}} L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})} \right| > \delta \right\} \\
= \; & \Pr\left\{ \left| \frac{\inf_{\boldsymbol{w}\in\tilde{\mathcal{H}}} \left(L(\boldsymbol{w}) + a(\boldsymbol{w})\right) - a(\tilde{\boldsymbol{w}}) - \inf_{\boldsymbol{w}\in\mathcal{H}} L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})} \right| > \delta, \boldsymbol{w}_n \in \tilde{\mathcal{H}} \right\} \\
& \quad + \Pr\left\{ \left| \frac{\inf_{\boldsymbol{w}\in\tilde{\mathcal{H}}} \left(L(\boldsymbol{w}) + a(\boldsymbol{w})\right) - a(\tilde{\boldsymbol{w}}) - \inf_{\boldsymbol{w}\in\mathcal{H}} L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})} \right| > \delta, \boldsymbol{w}_n \notin \tilde{\mathcal{H}} \right\}
\end{aligned} \tag{A14}$$

By definitions of conditional and joint probabilities, we have

RHS of equation (A14)

$$
\begin{aligned}
&\leq \Pr\left\{\left|\frac{\inf_{w\in\tilde{\mathcal{H}}}\left(L(\boldsymbol{w})+a(\boldsymbol{w})\right)-a(\tilde{\boldsymbol{w}})-\inf_{w\in\mathcal{H}}L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\middle|\boldsymbol{w}_n\in\tilde{\mathcal{H}}\right\}\Pr(\boldsymbol{w}_n\in\tilde{\mathcal{H}}) \\
&\qquad +\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}}) \\
&\leq \Pr\left\{\left|\frac{L(\boldsymbol{w}_n)+a(\boldsymbol{w}_n)-a(\tilde{\boldsymbol{w}})-\inf_{w\in\mathcal{H}}L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\middle|\boldsymbol{w}_n\in\tilde{\mathcal{H}}\right\}\Pr(\boldsymbol{w}_n\in\tilde{\mathcal{H}})+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}}) \\
&\leq \Pr\left\{\left|\frac{L(\boldsymbol{w}_n)+a(\boldsymbol{w}_n)-a(\tilde{\boldsymbol{w}})-\inf_{w\in\mathcal{H}}L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}}). \qquad \text{(A15)}
\end{aligned}
$$

Following the definition of v_n defined in Assumption 1(i), we have

$$
\begin{aligned}
&\text{RHS of equation (A15)} \\
&= \Pr\left\{\left|\frac{v_n+a(\boldsymbol{w}_n)-a(\tilde{\boldsymbol{w}})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}}) \\
&\leq \Pr\left\{\left|\frac{v_n}{L(\tilde{\boldsymbol{w}})}\right|+\left|\frac{a(\boldsymbol{w}_n)}{L(\tilde{\boldsymbol{w}})}\right|+\left|\frac{a(\tilde{\boldsymbol{w}})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}}) \\
&\leq \Pr\left\{\sup_{w\in\mathcal{H}}\left|\frac{v_n}{L(\boldsymbol{w})}\right|+\left|\frac{a(\boldsymbol{w}_n)}{\inf_{w\in\mathcal{H}}L(\boldsymbol{w})}\right|+\sup_{w\in\mathcal{H}}\left|\frac{a(\boldsymbol{w})}{L(\boldsymbol{w})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}}) \\
&\leq \Pr\left\{\sup_{w\in\mathcal{H}}\left|\frac{v_n}{L^*(\boldsymbol{w})}\right|\sup_{w\in\mathcal{H}}\left|\frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w})}\right|+\sup_{w\in\mathcal{H}}\left|\frac{a(\boldsymbol{w})}{L^*(\boldsymbol{w})}\right|\sup_{w\in\mathcal{H}}\left|\frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w})-v_n}\right|\right. \\
&\qquad \left.+\sup_{w\in\mathcal{H}}\left|\frac{a(\boldsymbol{w})}{L^*(\boldsymbol{w})}\right|\sup_{w\in\mathcal{H}}\left|\frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}}). \qquad \text{(A16)}
\end{aligned}
$$

According to Conditions (A10), (A11), (A12), (A13), and Assumption 1(iii), we obtain that the RHS of equation (A16) converge to 0 as $n\to\infty$. This completes the proof. ∎

D Heteroskedasticity-robust Prediction Model Averaging (HPMA) Method

Our setup is similar to both Wan, Zhang, and Zou (2010) and Liu and Okui (2013) by allowing the candidate models to be non-nonested. We observe a random sample (y_i, x_i) for $i = 1, ..., n$, in which y_i is a scalar and $x_i = (x_{i1}, x_{i2}, ...)$ is countably infinite. We consider the following data generating process (DGP)

$$
y_i = \mu_i + e_i, \quad \mu_i = \sum_{j=1}^{\infty}\beta_j x_{ij}, \quad \mathbb{E}(e_i|x_i) = 0 \qquad \text{(A17)}
$$

for $i = 1, ..., n$ and μ_i can be considered as the conditional mean $\mu_i = \mu(x_i) = \mathbb{E}(y_i|x_i)$ that is converging in mean square.[42] We assume the error term to be heteroskedastic by letting $\sigma_i^2 = \mathbb{E}(e_i^2|x_i)$ denotes the conditional variance which is allowed to depend on x_i.

Now we consider a set of M candidate models. We allow the M models to be non-nested. The m^{th} candidate model that approximates the DGP in equation (A17) is

$$y_i = \sum_{j=1}^{k^m} \beta_j^m x_{ij}^m + b_i^m + e_i, \tag{A18}$$

for $m = 1, ..., M$, where x_{ij}^m for $j = 1, ..., k^m$ denotes the regressors, β_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.

Define $\boldsymbol{y} = [y_1, ..., y_n]^\top$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_n]^\top$, and $\boldsymbol{e} = [e_1, ..., e_n]^\top$. The DGP in equation (A17) can be presented by $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{e}$. Let $\boldsymbol{X}^m$ be a full rank $n \times k^m$ matrix of independent variables with $(i, j)^{th}$ element being x_{ij}^m. The estimator of $\boldsymbol{\mu}$ from the m^{th} model is

$$\hat{\boldsymbol{\mu}}^m = \boldsymbol{X}^m \left(\boldsymbol{X}^{m\top} \boldsymbol{X}^m\right)^{-1} \boldsymbol{X}^{m\top} \boldsymbol{y} = \boldsymbol{P}^m \boldsymbol{y},$$

where $\boldsymbol{P}^m = \boldsymbol{X}^m (\boldsymbol{X}^{m\top} \boldsymbol{X}^m)^{-1} \boldsymbol{X}^{m\top}$ for all M. Similarly, the residual is $\hat{\boldsymbol{e}}^m = \boldsymbol{y} - \hat{\boldsymbol{\mu}}^m = (\boldsymbol{I}_n - \boldsymbol{P}^m)\boldsymbol{y}$ for all m. Since $\boldsymbol{P}^m$ is $n \times n$ for each m, we follow standard model averaging procedure and construct an averaged projection matrix $\boldsymbol{P}(\boldsymbol{w})$:

$$\boldsymbol{P}(\boldsymbol{w}) = \sum_{m=1}^{M} w^m \boldsymbol{P}^m, \tag{A19}$$

where $\boldsymbol{P}(\boldsymbol{w})$ is a weighted average of all potential $\boldsymbol{P}^m$. Due to its structure, $\boldsymbol{P}(\boldsymbol{w})$ is symmetric but not idempotent. The variable $\boldsymbol{w} = [w^1, w^2, ..., w^M]^\top$ is a weight vector we defined in the unit simplex in $\mathbb{R}^M$,

$$\mathcal{H} \equiv \left\{ \boldsymbol{w} \in [0, 1]^M : \sum_{m=1}^{M} w^m = 1 \right\}. \tag{A20}$$

Then, the model averaging estimator of μ is

$$\hat{\boldsymbol{\mu}}(\boldsymbol{w}) = \sum_{m=1}^{M} w^m \hat{\boldsymbol{\mu}}^m = \sum_{m=1}^{M} w^m \boldsymbol{P}^m \boldsymbol{y} = \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}. \tag{A21}$$

[42] Convergence in mean square implies that $\mathbb{E}(\mu_i - \sum_{j=1}^{k} \beta_j x_{ij})^2 \to 0$ as $k \to \infty$.

Similarly, we define the averaged residual as

$$\hat{e}(\boldsymbol{w}) = \sum_{m=1}^{M} w^m \hat{\boldsymbol{e}}^m = (\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{w}))\boldsymbol{y}. \tag{A22}$$

The performance of a model averaging estimator crucially depends on its choice of the weight vector w. Xie (2015) proposed a predictive model averaging (PMA) method that selects w through a convex optimization of a PMA criterion function of Amemiya (1980). One merit of the PMA method is that no preliminary estimates are required. The limitation of the PMA method is that the error term is required to be homoskedastic.

In the spirit of Liu and Okui (2013), we extend the PMA method to a heteroskedastic-robust predictive model averaging (HPMA) method with the following criterion function

$$\text{HPMA}(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}\|^2 + 2\sum_{i=1}^{n} \left(\hat{\boldsymbol{e}}_i(\boldsymbol{w})\right)^2 p_{ii}(\boldsymbol{w}), \tag{A23}$$

where $\boldsymbol{P}(\boldsymbol{w})$ is defined in (A19), $\hat{\boldsymbol{e}}_i(\boldsymbol{w})$ is the i^{th} element in $\hat{\boldsymbol{e}}(\boldsymbol{w})$ defined in equation (A22), $p_{ii}(\boldsymbol{w})$ is the i^{th} diagonal term in $\boldsymbol{P}(\boldsymbol{w})$. We estimate the weighting vector following

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{H}}{\arg\min}\, \text{HPMA}(\boldsymbol{w}).$$

Similar to PMA, obtaining $\hat{\boldsymbol{w}}$ from HPMA with restrictions $\boldsymbol{w} \in \mathcal{H}$ is a convex optimization process.

D.1 Asymptotic Optimality

In this subsection, we investigate the asymptotic properties of the HPMA estimator of w. We demonstrate that the proposed HPMA estimator is asymptotically optimal, in the sense of achieving the lowest possible mean squared error.

Let the average squared error loss and the corresponding l_2 type risk be

$$L(\boldsymbol{w}) = (\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu})^\top (\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu}), \tag{A24}$$

$$R(\boldsymbol{w}) = \mathbb{E}L(\boldsymbol{w}), \tag{A25}$$

where $\hat{\boldsymbol{\mu}}(\boldsymbol{w})$ is defined in equation (A21). To prove the optimality of HPMA, we assume the following regularity conditions similar to those demonstrated in Liu and Okui (2013),

Assumption A1 *There exists* $\epsilon > 0$ *such that* $\min_{1\le i\le n} \sigma_i^2 > \epsilon$.

Assumption A2 $\mathbb{E}(e_i^{4G}|\boldsymbol{x}_i) \le \kappa < \infty$ *for some integer* $1 \le G < \infty$ *and for some* κ.

Assumption A3 $M\xi^{-2G}\sum_{m=1}^{M}\left(R(\boldsymbol{w}_m^0)\right)^G \to 0$ *as* $n \to \infty$, *where* $\xi \equiv \inf_{w\in\mathcal{H}} R(\boldsymbol{w})$ *and* $\boldsymbol{w}_m^0$ *is a vector whose* m^{th} *element is 1 and all other elements are 0s.*

Assumption A4 $\max_{1\le m\le M}\max_{1\le i\le n} p_{ii}^m = O(n^{-1/2})$, p_{ii}^m *is the* i^{th} *diagonal element of* P^m.

Assumptions A1-A4 correspond to Assumptions 2.1-2.4 in Liu and Okui (2013). Assumptions A1 and A2 establish bounds on the error terms and conditional moments. Assumptions A3 is a convergence condition that requires ξ goes to infinity faster than M and $\max_m R(\boldsymbol{w}_m^0)$. Assumptions A4 is a standard convergence condition on projection matrices.

Assumption A5 $\max_{1\le m\le M}\xi^{-1}\tilde{p}\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^m)\boldsymbol{\mu} \xrightarrow{p} 0$ *and* $\max_{1\le m\le M} M^2\xi^{-2G}\tilde{p}^{2G}\left(\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^m)\boldsymbol{\mu}\right)^G \xrightarrow{p} 0$, *where* $\tilde{p} \equiv \sup_{w\in\mathcal{H}}\max_{1\le i\le n}(p_{ii}(\boldsymbol{w}))$.

Assumption A6 $\max_{1\le m\le M}\xi^{-1}\tilde{p}\boldsymbol{e}^\top\boldsymbol{P}^m\boldsymbol{e} \xrightarrow{p} 0$, $\max_{1\le m\le M}\xi^{-1}\tilde{p}tr(\boldsymbol{P}^m\boldsymbol{\Omega}) \xrightarrow{p} 0$, *and* $\max_{1\le m\le M} M^2\xi^{-2G}\tilde{p}^{2G}\left(tr(\boldsymbol{P}^m)\right)^G \xrightarrow{p} 0$, *where* $\tilde{p}$ *is defined in Assumption A5 and* $\boldsymbol{\Omega}$ *is an* $n\times n$ *diagonal matrix with* σ_i^2 *being its* i^{th} *diagonal element. .*

Assumption A5 requires that the bias from the worst potential model is small and Assumption A6 states that the associated variance be small. Similar requirements can be found in Wan, Zhang, and Zou (2010), which implies that some pre-selection procedures are always needed not just for the sake of computational efficiency, but also to maintain asymptotic optimality.[43] Finally, we demonstrate the optimality of HPMA estimator in the following Theorem.

Theorem 2 *Let Assumptions A1-A6 hold, as* $n \to \infty$, *we have*

$$\frac{L(\hat{\boldsymbol{w}})}{\inf_{w\in\mathcal{H}} L(\boldsymbol{w})} \xrightarrow{p} 1, \tag{A26}$$

where $L(\boldsymbol{w})$ *is defined in equation* (A24) *and* $\hat{w}$ *is the HPMA estimator.*

[43] Frequentist model averaging usually involves a constraint optimization (quadratic, convex, etc.) process that can be quite computationally demanding when the set of approximation models is large. A pre-selection procedure can reduce the total number of models by removing some poorly constructed models following certain criteria, therefore, improves computation efficiency. On the other hand, conditions like Assumptions A5 and A6 are frequently used (Wan, Zhang, and Zou (2010), Liu and Okui (2013), Xie (2015), etc) in demonstrating asymptotic optimality. As argued in Wan, Zhang, and Zou (2010), a necessary condition for Assumptions A5 and A6 type conditions to hold is removing some poorly constructed models (by a pre-selection procedure) before commencing the model averaging process. See Xie (2017) for a detailed discussion of various pre-selection methods for frequentist model averaging.

Proof of Theorem 2 Our proof follows Liu and Okui (2013) and Xie (2015). Let $\bar{\boldsymbol{P}}(\boldsymbol{w})$ be a diagonal matrix whose i^{th} diagonal element is $p_{ii}(\boldsymbol{w})$. Let $\hat{e}_i(\boldsymbol{w})$ as the i^{th} element of $\hat{\boldsymbol{e}}(\boldsymbol{w})$. Because:

$$\begin{aligned}\widehat{\text{HPMA}}(\boldsymbol{w}) &= (\boldsymbol{y}-\hat{\boldsymbol{\mu}}(\boldsymbol{w}))^\top(\boldsymbol{y}-\hat{\boldsymbol{\mu}}(\boldsymbol{w}))+2\sum_{i=1}^{n}\hat{e}_i^2(\boldsymbol{w})p_{ii}(\boldsymbol{w})\\ &= \text{HRC}_p(\boldsymbol{w})+2\left(\sum_{i=1}^{n}\hat{e}_i^2(\boldsymbol{w})p_{ii}(\boldsymbol{w})-\text{tr}(\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w}))\right).\end{aligned}$$

where $\text{HPC}_p(\boldsymbol{w})$ takes another form of the heteroskedasticity-robust model averaging method Liu and Okui (2013) proposed in (A2) such that

$$\text{HPC}_p(\boldsymbol{w}) = \|\boldsymbol{y}-\boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}\|^2+2\text{tr}\big(\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w})\big), \tag{A27}$$

where $\boldsymbol{\Omega}$ is an $n\times n$ diagonal matrix with σ_i^2 being its i^{th} diagonal element.

Theorem 1 of Liu and Okui (2013) showed that under Assumptions A1 to A3

$$\sup_{\boldsymbol{w}\in\mathcal{H}}\Big\{\text{HRC}_p(\boldsymbol{w})\Big/R(\boldsymbol{w})\Big\}\xrightarrow{p}0.$$

Therefore, we just need to prove that

$$\sup_{w\in\mathcal{H}}\left\{\left|\sum_{i=1}^{n}\hat{e}_i^2(\boldsymbol{w})p_{ii}(\boldsymbol{w})-\text{tr}(\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w}))\right|\Big/R(\boldsymbol{w})\right\}\xrightarrow{p}0 \tag{A28}$$

LHS of equation (A28) can be rewritten as

$$\begin{aligned}&\sup_{w\in\mathcal{H}}\left\{\left|\sum_{i=1}^{n}\hat{e}_i^2(\boldsymbol{w})p_{ii}(\boldsymbol{w})-\text{tr}(\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w}))\right|\Big/R(\boldsymbol{w})\right\}\\ \leq\ &\sup_{w\in\mathcal{H}}|\hat{\boldsymbol{e}}(\boldsymbol{w})^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\hat{\boldsymbol{e}}(\boldsymbol{w})-\mathbb{E}(\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e})|/\xi\\ \leq\ &\sup_{w\in\mathcal{H}}\{|\hat{\boldsymbol{e}}(\boldsymbol{w})^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\hat{\boldsymbol{e}}-\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}|+|\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}-\mathbb{E}(\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e})|\}/\xi.\end{aligned} \tag{A29}$$

where $\boldsymbol{e}(\boldsymbol{w})$ is defined in (A22), $\bar{\boldsymbol{P}}(\boldsymbol{w})$ is an $n\times n$ diagonal matrix with $p_{ii}(\boldsymbol{w})$ being its i^{th} diagonal element, and ξ is defined in Assumption A3. The first term in (A29) is

$$\begin{aligned}&\hat{\boldsymbol{e}}(\boldsymbol{w})^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\hat{\boldsymbol{e}}(\boldsymbol{w})-\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}\\ =\ &\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\bar{\boldsymbol{P}}(\boldsymbol{w})(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\boldsymbol{\mu}+2\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\bar{\boldsymbol{P}}(\boldsymbol{w})(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\boldsymbol{e}\\ &+\boldsymbol{e}^\top(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\bar{\boldsymbol{P}}(\boldsymbol{w})(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\boldsymbol{e}-\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}.\end{aligned}$$

We have

$$\sup_{w\in\mathcal{H}} \boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I}-\boldsymbol{P}(w))\boldsymbol{\mu}/\xi \le \tilde{p}\max_{1\le m\le M}\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^m)\boldsymbol{\mu}/\xi \xrightarrow{p} 0 \tag{A30}$$

by Assumption A6. Next, we consider the term

$$\boldsymbol{e}^\top(\boldsymbol{I}-\boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I}-\boldsymbol{P}(w))\boldsymbol{e}-\boldsymbol{e}^\top\bar{\boldsymbol{P}}(w)\boldsymbol{e} = -2\boldsymbol{e}^\top\boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e}+\boldsymbol{e}^\top\boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{P}(w)\boldsymbol{e},$$

where

$$\sup_{w\in\mathcal{H}} \boldsymbol{e}^\top\boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{P}(w)\boldsymbol{e}/\xi \le \tilde{p}\max_{1\le n\le M}\boldsymbol{e}^\top\boldsymbol{P}^m\boldsymbol{e}/\xi \xrightarrow{p} 0. \tag{A31}$$

by Assumption A6. For the term $\boldsymbol{e}^\top\boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e}$, we note that

$$\begin{aligned}
\mathbb{E}(\boldsymbol{e}^\top\boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e}) &= \mathbb{E}\left(\boldsymbol{e}^\top\sum_{m=1}^M w^m\boldsymbol{P}^m\bar{\boldsymbol{P}}(w)\boldsymbol{e}\right) = \sum_{m=1}^M\mathbb{E}(\boldsymbol{e}^\top w^m\boldsymbol{P}^m\bar{\boldsymbol{P}}(w)\boldsymbol{e}) \\
&= \sum_{m=1}^M\mathbb{E}(w^m\boldsymbol{e}^\top P^m\bar{\boldsymbol{P}}(w)\boldsymbol{e}) = \sum_{m=1}^M\mathbb{E}(w^m tr(\boldsymbol{P}^m\bar{\boldsymbol{P}}(w)\boldsymbol{e}\boldsymbol{e}^\top) \\
&= \sum_{m=1}^M w^m tr(\bar{\boldsymbol{P}}(w)\boldsymbol{P}^m\boldsymbol{\Omega}).
\end{aligned}$$

Therefore,

$$\sup_{w\in\mathcal{H}}\mathbb{E}(\boldsymbol{e}^\top\boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e})/\xi \le \max_{1\le m\le M}\xi^{-1}\tilde{p}tr(\boldsymbol{P}^m\boldsymbol{\Omega}) \xrightarrow{p} 0$$

by Assumption A6. Moreover, using Chebyshev's inequality and Theorem 2 of Whittle (1960), for any $\delta>0$, we have

$$\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{H}}\left|(\boldsymbol{e}^\top\boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e})-\mathbb{E}(\boldsymbol{e}^\top\boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e})\right|/\xi>\delta\right\} \\
\le\ & \sum_{l=1}^M\sum_{m=1}^M\mathbb{E}\left\{\frac{[(\boldsymbol{e}^\top\boldsymbol{P}^{(l)}\bar{\boldsymbol{P}}(w_m^0)\boldsymbol{e})-\mathbb{E}(\boldsymbol{e}^\top\boldsymbol{P}^{(l)}\bar{\boldsymbol{P}}(w_m^0)\boldsymbol{e})]^{2G}}{\delta^{2G}\xi^{2G}}\right\} \\
\le\ & \delta^{-2G}\xi^{-2G}\sum_{l=1}^M\sum_{m=1}^M C_1\left\{\sum_{i=1}^n\sum_{j=1}^n(p_{ij}^{(l)})^2p_{ii}^2(w_m^0)[\mathbb{E}(e_i^{4G})]^{1/2G}[\mathbb{E}(e_i^{4G})]^{1/2G}\right\}^G \\
\le\ & C_1\max_{1\le j\le n}\mathbb{E}(e_i^{4G})\delta^{-2G}\xi^{-2G}\tilde{p}^{2G}\sum_{l=1}^M\sum_{m=1}^M\left\{\sum_{i=1}^n\sum_{i=1}^n(p_{ij}^{(l)})^2\right\}^G \\
=\ & C_2\max_{1\le l\le M}\delta^{-2G}\xi^{-2G}M^2\tilde{p}^{2G}[\mathrm{tr}(\boldsymbol{P}^{(l)})]^G\to 0
\end{aligned}$$

by Assumption A6, where C_1 is a constant and $C_2\equiv C_1\max_{1\le i\le n}\mathbb{E}(e_i^{4G})$ is a bounded

constant according to Assumption A2. It follows that

$$\sup_{w\in\mathcal{H}}(\boldsymbol{e}^\top \boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e})/\xi = o_p(1). \tag{A32}$$

Noting that $\mathbb{E}[\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I}-\boldsymbol{P}(w))\boldsymbol{e}] = 0$, we again use Chebyshev's inequality and Theorem 2 of Whittle (1960) to show that

$$\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{H}}\left|\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I}-\boldsymbol{P}(w))\boldsymbol{e}\right|/\xi > \delta\right\} \\
\leq\ & \sum_{l=1}^{M}\sum_{m=1}^{M}\mathbb{E}\left\{\frac{[\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(l)})\bar{\boldsymbol{P}}(w_m^0)(\boldsymbol{I}-\boldsymbol{P}^{(l)})\boldsymbol{e}]^{2G}}{\delta^{2G}\xi^{2G}}\right\} \\
\leq\ & \delta^{-2G}\xi^{-2G}M\sum_{m=1}^{M}C_3\left\{\sum_{i=1}^{n}\gamma_{im}^2[\mathbb{E}(e_i^{2G})]^{1/G}\right\}^G,
\end{aligned}$$

where γ_{im} is the i^{th} element of $\max_{1\leq l\leq M}\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(l)})\bar{\boldsymbol{P}}(w_m^0)(\boldsymbol{I}-\boldsymbol{P}^{(l)})$, and C_3 is a constant. We now have that

$$\delta^{-2G}\xi^{-2G}M\sum_{m=1}^{M}C_3\left\{\sum_{i=1}^{n}\gamma_{im}^2[\mathbb{E}(e_i^{2G})]^{1/G}\right\}^G \leq \delta^{-2G}\xi^{-2G}M\sum_{m=1}^{M}C_4\left\{\sum_{i=1}^{n}\gamma_{im}^2\right\}^G,$$

where $C_4 \equiv C_3\max_{1\leq i\leq n}\mathbb{E}(e_i^{2G})$ is a bounded constant according to Assumption A2 and

$$\begin{aligned}
\sum_{i=1}^{n}\gamma_{im}^2 &= \max_{1\leq l\leq M}\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(l)})\bar{\boldsymbol{P}}(w_m^0)(\boldsymbol{I}-\boldsymbol{P}^{(l)})(\boldsymbol{I}-\boldsymbol{P}^{(l)})\bar{\boldsymbol{P}}(w_m^0)(\boldsymbol{I}-\boldsymbol{P}^{(l)})\boldsymbol{\mu} \\
&\leq \max_{1\leq l\leq M}(\tilde{p})^2\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(l)})\boldsymbol{\mu}.
\end{aligned}$$

Therefore, it holds that

$$\delta^{-2G}\xi^{-2G}\sum_{m=1}^{M}C_4\left\{\sum_{i=1}^{n}\gamma_{jm}^2\right\}^G \leq \max_{1\leq l\leq M}\delta^{-2G}\xi^{-2G}C_4\tilde{p}^{2G}M^2\left\{\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^m)\boldsymbol{\mu}\right\}^G \to 0$$

by Assumption A5. Therefore, we have

$$\sup_{w\in\mathcal{H}}\left|\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I}-\boldsymbol{P}(w))\boldsymbol{e}\right|/\xi \xrightarrow{p} 0. \tag{A33}$$

By (A30), (A31), (A32), and (A33), we have that the first term in (A29)

$$\sup_{w\in\mathcal{H}}\left|\hat{\boldsymbol{e}}(w)^\top\bar{\boldsymbol{P}}(w)\hat{\boldsymbol{e}}(w) - \boldsymbol{e}^\top\bar{\boldsymbol{P}}(w)\boldsymbol{e}\right|/\xi \xrightarrow{p} 0. \tag{A34}$$

Similarly, for the second term in (A29), using Chebyshev's inequality and Theorem 2 of Whittle (1960), for any $\delta > 0$, we have

$$
\begin{aligned}
& \Pr\left\{ \sup_{w\in\mathcal{H}} \left| e^\top \bar{P}(w) e - \mathbb{E}\left(e^\top \bar{P}(w) e \right) \right| / \xi > \delta \right\} \\
\leq & \sum_{m=1}^{M} \mathbb{E}\left\{ \frac{[e^\top \bar{P}(w_m^0) e - \mathbb{E}\left(e^\top \bar{P}(w_m^0) e \right)]^{2G}}{\delta^{2G} \xi^{2G}} \right\} \\
\leq & \delta^{-2G} \xi^{-2G} \sum_{m=1}^{M} C_5 \left\{ \sum_{i=1}^{n} p_{ii}^2(w_m^0) [\mathbb{E}(e_i^{4G})]^{1/G} \right\}^G \\
\leq & C_6 \max_{1\leq j\leq n} \mathbb{E}(e_i^{4G}) \delta^{-2G} \xi^{-2G} \sum_{m=1}^{M} \left\{ \sum_{i=1}^{n} p_{ii}^2(w_m^0) \right\}^G \\
\leq & C_6 \delta^{-2G} \xi^{-2G} \sum_{m=1}^{M} [\mathrm{tr}[\left(P(w_m^0) \right)^2]]^G \\
= & C_6 \delta^{-2G} \xi^{-2G} \left(\inf_{1\leq i\leq M} \sigma_i^2 \right)^{-G} \sum_{m=1}^{M} [\mathrm{tr}[\inf_{1\leq i\leq M} \sigma_i^2 \left(P(w_m^0) \right)^2]]^G \\
\leq & C_6 \delta^{-2G} \xi^{-2G} \left(\inf_{1\leq i\leq M} \sigma_i^2 \right)^{-G} \sum_{m=1}^{M} [\mathrm{tr}[\Omega \left(P(w_m^0) \right)^2]]^G \\
= & C_6 \delta^{-2G} \xi^{-2G} \left(\inf_{1\leq i\leq M} \sigma_i^2 \right)^{-G} \sum_{m=1}^{M} [\mathrm{tr}[\Omega P(w_m^0)]]^G \\
\leq & C_6 \delta^{-2G} \xi^{-2G} \sum_{m=1}^{M} \left[R(w_m^0) \right]^G \to 0,
\end{aligned}
$$

where C_5 is a constant and $C_6 \equiv C_5 \max_{1\leq i\leq n} \mathbb{E}(e_i^{4G})$ is a bounded constant according to Assumption A2. The last inequality is due to

$$
\begin{aligned}
R(w_m^0) &= \mathbb{E}(L\left(w_m^0 \right)) = \mathbb{E}\left[\left(P^m y - \mu \right)^\top \left(P^m y - \mu \right) \right] \\
&= \mathbb{E}\left[\left(P^m \left(\mu + e \right) - \mu \right)^\top \left(P^m \left(\mu + e \right) - \mu \right) \right] \\
&= \mathbb{E}\left[\left(\left(P^m - I \right) \mu - P^m e \right)^\top \left(\left(P^m - I \right) \mu - P^m e \right) \right] \\
&= \mu^\top \left(P^m - I \right)^\top \left(P^m - I \right) \mu - 2\mathbb{E}\left[\mu^\top \left(P^m - I \right)^\top P^m e \right] + \mathbb{E}\left[e^\top P^m e \right] \\
&= \mu^\top \left(P^m - I \right)^\top \left(P^m - I \right) \mu + \mathrm{tr}[\Omega P^m] \\
&= \mu^\top \left(P^m - I \right)^\top \left(P^m - I \right) \mu + \mathrm{tr}[\Omega P(w_m^0)] \\
&\geq \mathrm{tr}[\Omega P(w_m^0)],
\end{aligned}
$$

where $\boldsymbol{P}(\boldsymbol{w}_m^0) = \boldsymbol{P}^m$ and the expectation is conditional on X. Therefore, we have

$$\sup_{w \in \mathcal{H}} \left| \boldsymbol{e}^\top \bar{\boldsymbol{P}}(\boldsymbol{w}) \boldsymbol{e} - \mathbb{E}\left(\boldsymbol{e}^\top \bar{\boldsymbol{P}}(\boldsymbol{w}) \boldsymbol{e} \right) \right| / \xi \xrightarrow{p} 0. \tag{A35}$$

Results of (A34) and (A35) imply that condition (A28) hold. This completes the proof.

E More Empirical Results

This appendix consists of numerous subsections that provide further analyses and robustness checks of our main findings. OLS estimates of the GUM model are provided in subsection E.1. Breusch-pagan tests are provided in Table A1 show strong evidence of heteroskesdasticity for both open box office and movie unit sales.

The first piece of evidence pertaining to using two social media measures versus one is obtained by comparing estimates across tables in subsection E.3 (see tables A3 and A4). In subsection E.2 and E.6 we provide evidence of the relative prediction efficiency for double Lasso and Lasso based Strategies respectively. As observed in tables A2 and A8 – A10, the benchmark HRC^p outperforms all Lasso based methods considered. Finally, the evidence contrasting tables A8 – A10 present further evidence for why two social media measures are preferred to either one.

In subsection E.4, a Monte Carlo study is used to shed further light on the relative performance of ARMS and ARMSH under different scenarios related to what is the source of heteroskedasticity. Related, in subsection E.7 we present additional analyses that contrasts which models (and their contents) are selected by ARMS to ARMSH. These sections explain when differences between these methods could occur and why in our application, there were many similarities. Related to E.7, in subsection E.5 we present weights of, and contents of the top 5 models selected by the HRC^p estimator. These results continue to show that in practice, the model averaging estimator gives lots of weight to very few of all the potential models and is consistent with other applications of these methods including in policy oriented applications such as crime deterrence (Durlauf, Navarro, and Rivers, 2016).

In subsection E.8, we provide evidence that even when we restrict machine learning strategies to use the identical set of predictors as model screening choices made for model averaging that recursive partitioning methods yield more accurate forecasts. This shows that much of the gains we observed in our application come from the restrictiveness of the linear model and that additional gains can still be obtained by allowing for model uncertainty and considering that the data is heteroskedastic. Subsection E.9 provides formal evidence that the proposed MAB method significantly outperforms other forecasting strategies considered in the main text (tables 3 – 5).

Last, subsection E.10 considers adding a model averaging flavor to a single regression tree (MART). For space considerations, we did not include this in the main text since as seen in the single figure A2 presented in subsection E.10, the MART method is outperformed by both MAB and MARF by a large margin in both heteroskedasticity scenarios. Thus, similar to the discussion in the statistical learning literature that forecasts from RT are unreliable and both bagging and random forest present improvement, we advocate only adding model averaging to strategies that used bagging or random forest to create subgroups.

E.1 OLS Estimates of the GUM Model

Table A1: OLS Estimates of the Generalized Unrestricted Model

Variable	Open Box Office		Movie Unit Sales	
	Coefficient	Robust S.E.	Coefficient	Robust S.E.
Genre				
Action	-1.6895	3.0838	-0.0622	0.1194
Adventure	4.6542	3.7732	-0.0967	0.1588
Animation	-1.9354	5.6046	0.8167*	0.3609
Biography	0.1229	4.2324	-0.0109	0.2015
Comedy	-0.9595	3.7382	-0.1287	0.1859
Crime	2.6461	2.7335	-0.0931	0.1052
Drama	-1.7884	3.6083	0.0139	0.1092
Family	2.6236	6.7679	-0.4118	0.3503
Fantasy	12.8881*	4.9159	0.5634	0.3937
Horror	3.0486	2.4376	-0.3655*	0.1441
Mystery	3.3377	2.4852	0.1414	0.1243
Romance	-2.5919	3.3696	-0.0986	0.0921
Sci-Fi	-0.3705	2.6569	0.0336	0.1391
Thriller	0.8643	2.9379	0.0306	0.1301
Rating				
PG	2.8901	5.4757	-0.6290	0.4196
PG13	1.8691	6.8517	-0.8369	0.5112
R	2.6378	6.6841	-0.7490	0.4964
Core Parameters				
Budget	0.1182*	0.0399	0.0035*	0.0016
Weeks	0.3738	0.2768	0.0447*	0.0109
Screens	6.1694*	1.3899	0.3215*	0.0526
Sentiment				
T-21/-27	-0.1570	0.6610	-0.0148	0.0241
T-14/-20	-0.9835	0.9393	-0.0040	0.0304
T-7/-13	-1.2435	1.0695	0.1802	0.1104
T-4/-6	0.2277	1.1775	-0.1708*	0.0842
T-1/-3	2.5070*	0.7509	-0.0422	0.0839
T+0			0.2172*	0.0864
T+1/+7			-0.0927*	0.0399
T+8/+14			0.0212	0.0234
T+15/+21			0.0085	0.0291
T+22/+28			-0.0808	0.1072
Volume				
T-21/-27	-97.5186*	31.6624	-1.6863	0.9608
T-14/-20	19.4109	38.6929	0.0724	1.1598
T-7/-13	-45.2885	30.9011	-1.8770	1.1417
T-4/-6	86.2881*	27.2008	2.5302*	0.7184
T-1/-3	18.9664*	5.1687	-1.2437*	0.4167
T+0			0.4423*	0.1064
T+1/+7			-0.2006	0.2404
T+8/+14			1.1195	0.9779
T+15/+21			0.4945	0.6281
T+22/+28			-0.3414	0.3104
Breusch-Pagan Statistic	**249.9485**		**207.3698**	
Breusch-Pagan p-value	**$<$0.0001**		**$<$0.0001**	
R-square	**0.7973**		**0.8016**	

Note: * indicates the associated variable is significant at 5% level.

E.2 Performance of Double-Lasso Strategy in Simulation Experiment

Table A2: Comparing Hetero-robust and Homo-efficient Model Screening Methods

n_E	OLS_{10}	OLS_{12}	OLS_{15}	HRC^p_{10}	HRC^p_{12}	HRC^p_{15}	Benchmark
Panel A: Open Box Office							
Mean Squared Forecast Error (MSFE)							
10	1.4388	1.5229	1.1787	1.4181	1.5075	1.1564	**1.0000**
20	1.6213	1.6090	1.2135	1.5898	1.5814	1.1854	**1.0000**
30	1.7625	1.6869	1.2597	1.7322	1.6714	1.2344	**1.0000**
40	1.8172	1.7028	1.2622	1.7745	1.6768	1.2548	**1.0000**
Mean Absolute Forecast Error (MAFE)							
10	1.2064	1.2131	1.0778	1.1962	1.2054	1.0680	**1.0000**
20	1.2356	1.2208	1.0880	1.2262	1.2173	1.0841	**1.0000**
30	1.2420	1.2273	1.0882	1.2331	1.2192	1.0833	**1.0000**
40	1.2475	1.2330	1.0845	1.2360	1.2187	1.0766	**1.0000**
Panel B: Movie Unit Sales							
Mean Squared Forecast Error (MSFE)							
10	1.3855	1.4254	1.4699	1.3645	1.3892	1.4364	**1.0000**
20	1.3562	1.3960	1.4022	1.3321	1.3651	1.3730	**1.0000**
30	1.2831	1.3096	1.3088	1.2733	1.2909	1.2821	**1.0000**
40	1.1793	1.2094	1.2499	1.1573	1.1807	1.2210	**1.0000**
Mean Absolute Forecast Error (MAFE)							
10	1.2604	1.2731	1.2840	1.2514	1.2616	1.2683	**1.0000**
20	1.2345	1.2541	1.2626	1.2273	1.2365	1.2472	**1.0000**
30	1.2014	1.2190	1.2314	1.1920	1.2053	1.2169	**1.0000**
40	1.1682	1.1878	1.2051	1.1565	1.1706	1.1880	**1.0000**

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column.

E.3 Additional Evidence on the Importance of Social Media Data

Table A3: OLS Estimates of Models with Sentiment Only

Variable	Open Box		Movie Unit	
	Coefficient	Robust S.E.	Coefficient	Robust S.E.
Genre				
Action	-11.8297	5.1756	-0.5991	0.2118
Adventure	1.8903	9.0801	-0.2221	0.3721
Animation	-8.6157	7.2188	0.3618	0.3987
Biography	-10.6777	7.3202	-0.3815	0.3079
Comedy	-6.1906	4.4094	-0.3875	0.2136
Crime	5.6338	3.7323	0.1658	0.1751
Drama	-4.3020	4.9879	-0.2661	0.1924
Family	-1.3797	8.0709	-0.3123	0.3741
Fantasy	19.2129	10.5968	0.8570	0.4906
Horror	-0.8574	4.6042	-0.6504	0.2190
Mystery	-4.1597	3.1965	-0.1284	0.1412
Romance	-1.3851	4.4953	0.1232	0.1784
Sci-Fi	0.6611	6.1694	0.1187	0.2989
Thriller	1.4062	5.2588	0.0971	0.2140
Rating				
PG	7.6872	7.0093	-1.0293	0.4639
PG13	21.6049	10.7996	-0.5286	0.5447
R	19.5326	10.5227	-0.5796	0.5433
Core Parameters				
Budget	0.1525	0.0827	0.0064	0.0033
Weeks	1.3267	0.5057	0.0943	0.0204
Screens	13.8708	2.9586	0.5949	0.1233
Sentiment				
T-21/-27	0.9289	0.7021	-0.0195	0.0292
T-14/-20	-0.7583	0.7503	0.0373	0.0366
T-7/-13	-1.1656	1.6137	0.3103	0.1303
T-4/-6	0.9664	2.1090	-0.0694	0.1113
T-1/-3	-0.1460	1.1729	-0.0401	0.1418
T+0			0.1238	0.1668
T+1/+7			-0.1016	0.0603
T+8/+14			0.0649	0.0372
T+15/+21			-0.0992	0.0411
T+22/+28			-0.1859	0.1286
R-square	**0.5322**		**0.6488**	

Note: * indicates the associated variable is significant at 5% level.

Table A4: OLS Estimates of Models with Volume Only

Variable	Open Box		Movie Unit	
	Coefficient	Robust S.E.	Coefficient	Robust S.E.
Genre				
Action	-1.7845	3.0495	-0.1049	0.1163
Adventure	4.8425	3.7630	0.0347	0.1508
Animation	-3.8178	5.2420	0.6189	0.3508
Biography	0.5099	4.4590	-0.1050	0.2038
Comedy	-0.5934	3.8404	-0.1896	0.1556
Crime	3.1958	2.6371	0.0043	0.0961
Drama	-1.9479	3.5767	-0.0280	0.1078
Family	3.6903	6.5546	-0.3090	0.3424
Fantasy	13.3327	4.9812	0.5544	0.3864
Horror	3.6698	2.5120	-0.2299	0.1305
Mystery	2.6945	2.5712	-0.0145	0.1100
Romance	-2.5929	3.4036	-0.0859	0.0909
Sci-Fi	-0.5145	2.7094	0.0015	0.1279
Thriller	0.6968	3.0682	-0.0407	0.1181
Rating				
PG	1.8990	5.2023	-0.3739	0.3662
PG13	1.6943	6.7034	-0.5650	0.4418
R	2.3396	6.4815	-0.5206	0.4475
Core Parameters				
Budget	0.1142	0.0396	0.0029	0.0016
Weeks	0.4335	0.2705	0.0424	0.0114
Screens	6.9067	1.4856	0.3422	0.0557
Volume				
T-21/-27	-97.6733	30.6043	-1.5188	0.9072
T-14/-20	21.1375	36.7023	-0.0649	1.1053
T-7/-13	-39.7233	31.2763	-1.6555	1.1440
T-4/-6	81.3088	27.3566	2.1988	0.6776
T-1/-3	18.1939	4.9561	-1.4011	0.3762
T+0			0.4675	0.1007
T+1/+7			-0.2659	0.2455
T+8/+14			1.6392	0.8910
T+15/+21			0.2306	0.5984
T+22/+28			-0.2764	0.3631
R-square	**0.8224**		**0.8445**	

Note: * indicates the associated variable is significant at 5% level.

E.4 Using Monte Carlo Study to Understand How Different Sources for Heteroskedasticity Affect Strategies

We found in forecasts of retail movie unit sales that the difference in the performance between PMA and HRC^p in table 3 in conjunction with the relative improved performance of ARMS presented in table 4 to be surprising. A potential explanation for these findings is the source of heteroskedasticity in the data. We examine the performance of five different model screening methods that are implied in the subscripts of the following model sets: $\mathcal{M}^K_{\text{GETS}}$, $\mathcal{M}^K_{\text{Lasso}}$, $\mathcal{M}^K_{\text{ARMS}}$, $\mathcal{M}^K_{\text{ARMSH}}$, and $\mathcal{M}^K_{\text{HRMS}}$.[44] Using data generated by the Monte Carlo design described in section 4.2, we compare the risks of each method:

$$\text{Risk}_i \equiv \frac{1}{n}\sum_{t=1}^{n}\left(\hat{\mu}_t(\mathcal{M}_i^K) - \mu_t\right)^2 \qquad \text{for } i = \text{GETS, Lasso, ARMS, ARMSH, and HRMS,}$$

where μ_t is the true fitted value (feasible in simulation) and $\hat{\mu}_t(\mathcal{M}_i^K)$ is the average fitted value obtained by HRC^p using specific candidate model set. Four different sample sizes ($n = 100, 200, 300$, and 400) are considered and the risk for each method - sample size pair is averaged across 10,000 simulation draws.

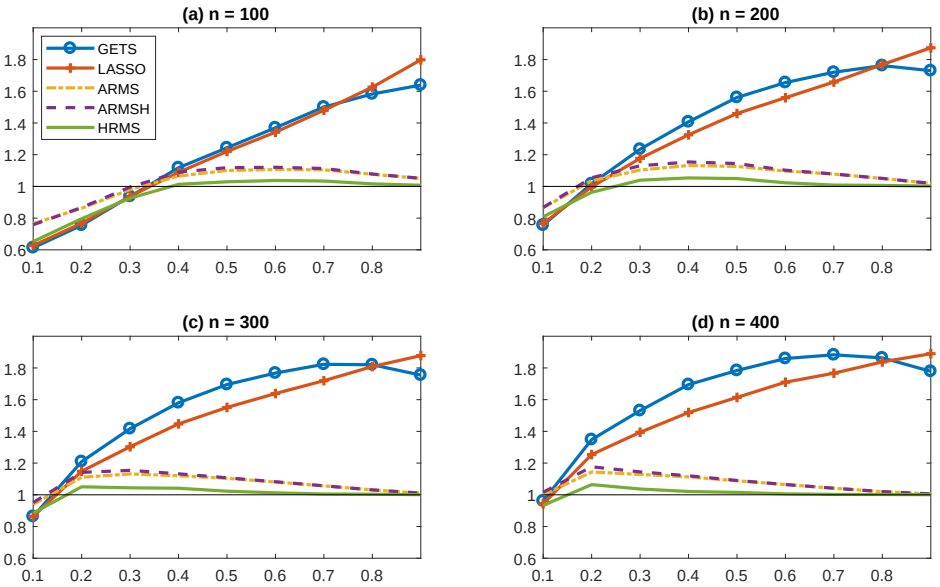
Figures A1 presents the results from this exercise where we normalize the risks by the risk of the infeasible optimal model. Each line presents the relative risks of each model screening method associated with R^2 from 0.1 to 0.9, respectively. Each sub-panel (a) to (d) presents the results for different sample sizes.

In virtually every panel of figures A1, HRMS has the best performance. In the random heteroskedasticity scenario, GETS and Lasso perform well only when R^2 is low. As R^2 increases, the relative improved performance of ARMS, ARMSH, and HRMS emerges. The performance of both ARMS and ARMSH more closely mimics HRMS at larger sample sizes. However, in simulations where heteroskedasticity arises due to neglected parameter heterogeneity both GETS and Lasso perform poorly, particularly when there is strong correlation among the regressors. The performance of both screening methods is relatively poorer when either the sample size or R^2 increases. In contrast, ARMS and ARMSH yield consistently better results that are similar with increasing n and R^2. Note that for both cases, ARMS and ARMSH yield quite similar results. The results in figures A1 point out that the performance of both GETS and Lasso rely heavily on homoskedasticity.

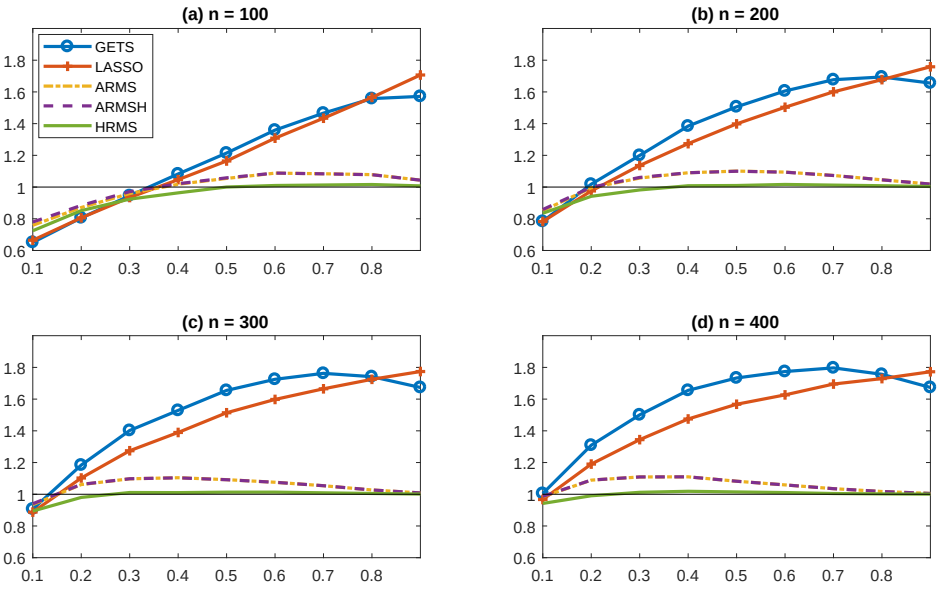
[44] A full permutation of the $K = 20$ regressors leads to a total of 1,048,575 candidate models (the null model is ignored). In our experiments, the pre-determined parameters for GETS and ARMS(H) are $p = 0.1$ and $M' = 20$ respectively, whereas we manipulate the tuning parameter for Lasso and select 5 predictors. We construct $2^5 - 1 = 31$ models based on permutation of the selected parameters.

Figure A1: Comparing Model Screening Methods with Simulated Data

Scenario A. Random Heteroskedasticity



Scenario B. Parameter Heterogeneity



E.4.1 Prediction Comparison Using One Set of Measures

Table A5: Evaluating the Importance of Twitter Variable using MAB

n_E	Include Both	Sentiment Only	Volume Only	Include None	Benchmark
Panel A: Open Box Office					
		Mean Squared Forecast Error (MSFE)			
10	**0.6045**	0.8659	0.6009	1.5271	1.0000
20	**0.8098**	1.2111	0.8242	1.6091	1.0000
30	**0.9123**	1.4463	0.9654	1.8287	1.0000
40	1.0281	1.6934	1.0810	2.1822	**1.0000**
		Mean Absolute Forecast Error (MAFE)			
10	**0.6538**	0.7505	0.6635	0.9881	1.0000
20	**0.7296**	0.8531	0.7428	1.0911	1.0000
30	**0.7541**	0.9057	0.7940	1.2939	1.0000
40	**0.7890**	0.9653	0.8151	1.3988	1.0000
Panel B: Movie Unit Sales					
		Mean Squared Forecast Error (MSFE)			
10	**0.8114**	0.9235	0.8683	1.4882	1.0000
20	**0.8914**	1.0621	0.9038	1.6761	1.0000
30	**0.9270**	1.1196	0.9325	1.7988	1.0000
40	**0.9734**	1.1244	0.9757	1.9982	1.0000
		Mean Absolute Forecast Error (MAFE)			
10	**0.7740**	0.8397	0.7970	1.0981	1.0000
20	**0.8192**	0.8861	0.8287	1.1532	1.0000
30	**0.8269**	0.9052	0.8525	1.2887	1.0000
40	**0.8470**	0.9311	0.8617	1.4109	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column.

E.5 Weights of, and Contents of the Top 5 Models Selected by the HRC^p Estimator

Table A6: Describing the 5 Highest Weight Models: Open Box Office

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**HRC^p**
Weight in HRC^p	0.3862	0.2159	0.1755	0.0945	0.0816	
Genre						
Action				x		x
Adventure	x		x		x	x
Animation						x
Biography						x
Comedy				x		x
Crime	x					x
Drama				x		x
Family						x
Fantasy	x	x	x	x	x	x
Horror	x	x			x	x
Mystery		x	x		x	x
Romance		x	x			x
Sci-Fi						x
Thriller						x
Rating						
PG						x
PG13						x
R						x
Core						
Budget	x	x	x	x	x	x
Weeks	x	x	x	x	x	x
Screens	x	x	x	x	x	x
Sentiment						
T-21/-27						x
T-14/-20	x			x	x	x
T-7/-13		x	x			x
T-4/-6						x
T-1/-3	x	x	x	x	x	x
Volume						
T-21/-27	x	x	x	x	x	x
T-14/-20						x
T-7/-13	x	x	x	x	x	x
T-4/-6	x	x	x	x	x	x
T-1/-3	x	x	x	x	x	x
R^2 w/ SV.	0.8265	0.8249	0.8258	0.8248	0.8259	0.8230
R^2 w/o SV.	0.4836	0.4796	0.4789	0.4911	0.4795	0.7383

Note: x denotes that explanatory variable is included in the particular model, SV denotes social media data and HRC^p refers to a specific model averaging method.

Table A7: Describing the 5 Highest Weight Models: Retail Movie Unit Sales

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**HRC^p**
Weight in HRC^p	0.2977	0.1645	0.1558	0.1447	0.0989	
Genre						
Action						x
Adventure						x
Animation	x	x	x	x	x	x
Biography						x
Comedy			x			x
Crime						x
Drama						x
Family	x	x	x	x	x	x
Fantasy	x	x	x	x	x	x
Horror	x	x	x	x	x	x
Mystery	x	x			x	x
Romance						x
Sci-Fi						x
Thriller		x				x
Rating						
PG	x	x	x	x		x
PG13	x	x	x	x	x	x
R	x	x	x	x		x
Core						
Budget	x	x	x	x	x	x
Weeks	x	x	x	x	x	x
Screens	x	x	x	x	x	x
Sentiment						
T-21/-27						x
T-14/-20			x			x
T-7/-13	x	x	x	x	x	x
T-4/-6	x	x	x	x	x	x
T-1/-3					x	x
T+0	x	x	x	x	x	x
T+1/+7	x	x	x	x		x
T+8/+14		x				x
T+15/+21						x
T+22/+28					x	x
Volume						
T-21/-27	x	x	x	x	x	x
T-14/-20						x
T-7/-13	x	x	x	x	x	x
T-4/-6	x	x	x	x	x	x
T-1/-3	x	x	x	x	x	x
T+0	x	x	x	x	x	x
T+1/+7						x
T+8/+14		x				x
T+15/+21	x		x	x		x
T+22/+28						x
R^2 w/ SV.	0.8512	0.8517	0.8530	0.8503	0.8362	0.8450
R^2 w/o SV.	0.5976	0.6024	0.6027	0.5976	0.5918	0.7002

Note: x denotes that explanatory variable is included in the particular model, SV denotes social media data and HRC^p refers to a specific model averaging method.

E.6 Further Comparison of the Relative Prediction Efficiency for Lasso-based Strategies

Table A8: Further Comparison of the Relative Prediction Efficiency (with Both Sentiment and Volume)

n_E	OLS_{10}	OLS_{11}	OLS_{12}	OLS_{13}	OLS_{14}	OLS_{15}	HRC^p_{10}	HRC^p_{11}	HRC^p_{12}	HRC^p_{13}	HRC^p_{14}	HRC^p_{15}	HRC^p
Panel A: Open Box Office													
							Mean Squared Forecast Error (MSFE)						
10	1.1464	1.1704	1.1671	1.1778	1.1132	1.1221	1.1462	1.1642	1.1647	1.1717	1.1094	1.1203	**1.0000**
20	1.1620	1.1809	1.1803	1.1830	1.0943	1.0992	1.1606	1.1771	1.1797	1.1755	1.0815	1.0826	**1.0000**
30	1.1922	1.2092	1.2067	1.2113	1.0731	1.0696	1.1899	1.2068	1.2037	1.2092	1.0636	1.0624	**1.0000**
40	1.2076	1.2295	1.2174	1.2233	1.0608	1.0633	1.2027	1.2197	1.2141	1.2199	1.0573	1.0537	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.0529	1.0591	1.0669	1.0689	1.0623	1.0632	1.0430	1.0595	1.0576	1.0687	1.0593	1.0594	**1.0000**
20	1.0603	1.0657	1.0692	1.0767	1.0556	1.0549	1.0506	1.0631	1.0689	1.0750	1.0551	1.0546	**1.0000**
30	1.0568	1.0619	1.0669	1.0722	1.0560	1.0558	1.0473	1.0528	1.0576	1.0719	1.0542	1.0538	**1.0000**
40	1.0591	1.0663	1.0673	1.0734	1.0549	1.0537	1.0578	1.0654	1.0641	1.0720	1.0536	1.0530	**1.0000**
Panel B: Movie Unit Sales													
							Mean Squared Forecast Error (MSFE)						
10	1.3737	1.2921	1.3495	1.3456	1.3621	1.3757	1.3558	1.2784	1.3354	1.3434	1.3512	1.3704	**1.0000**
20	1.3756	1.2772	1.2811	1.2457	1.2578	1.2768	1.3448	1.2459	1.2697	1.2432	1.2568	1.2651	**1.0000**
30	1.3001	1.2388	1.2086	1.1616	1.1666	1.1814	1.2728	1.2282	1.2012	1.1530	1.1644	1.1822	**1.0000**
40	1.2306	1.1718	1.1609	1.1135	1.1364	1.1454	1.2069	1.1565	1.1486	1.1093	1.1281	1.1398	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.2303	1.2058	1.2161	1.1581	1.1534	1.1600	1.2229	1.1974	1.2155	1.1575	1.1523	1.1564	**1.0000**
20	1.2096	1.1844	1.1958	1.1386	1.1427	1.1436	1.2036	1.1760	1.1890	1.1369	1.1411	1.1398	**1.0000**
30	1.1887	1.1656	1.1735	1.1182	1.1204	1.1195	1.1794	1.1569	1.1675	1.1161	1.1180	1.1149	**1.0000**
40	1.1704	1.1469	1.1557	1.0989	1.1064	1.1086	1.1600	1.1364	1.1459	1.0959	1.1005	1.1027	**1.0000**

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column.

Table A9: Further Comparison of the Relative Prediction Efficiency (with Sentiment Only)

n_E	OLS_{10}	OLS_{11}	OLS_{12}	OLS_{13}	OLS_{14}	OLS_{15}	HRC^p_{10}	HRC^p_{11}	HRC^p_{12}	HRC^p_{13}	HRC^p_{14}	HRC^p_{15}	HRC^p
Panel A: Open Box Office													
							Mean Squared Forecast Error (MSFE)						
10	1.1111	1.1240	1.1428	1.1403	1.1389	1.1528	1.0865	1.0922	1.1077	1.1022	1.1068	1.1084	**1.0000**
20	1.0836	1.0940	1.1102	1.1121	1.0887	1.0896	1.0802	1.0766	1.0912	1.1010	1.0795	1.0842	**1.0000**
30	1.0648	1.0700	1.0888	1.0871	1.0799	1.0840	1.0641	1.0643	1.0787	1.0809	1.0702	1.0772	**1.0000**
40	1.0732	1.0779	1.1027	1.1099	1.0902	1.0909	1.0727	1.0768	1.0939	1.0916	1.0777	1.0795	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.0305	1.0422	1.0485	1.0528	1.0552	1.0652	1.0302	1.0325	1.0368	1.0281	1.0318	1.0457	**1.0000**
20	1.0314	1.0399	1.0467	1.0535	1.0556	1.0647	1.0276	1.0311	1.0323	1.0369	1.0413	1.0456	**1.0000**
30	1.0303	1.0378	1.0474	1.0522	1.0542	1.0669	1.0256	1.0298	1.0318	1.0364	1.0382	1.0421	**1.0000**
40	1.0355	1.0468	1.0542	1.0615	1.0592	1.0719	1.0281	1.0343	1.0398	1.0402	1.0411	1.0475	**1.0000**
Panel B: Movie Unit Sales													
							Mean Squared Forecast Error (MSFE)						
10	1.0179	1.0323	1.0391	1.0299	1.0494	1.0257	1.0152	1.0037	1.0030	1.0192	1.0151	1.0068	**1.0000**
20	1.0462	1.0589	1.0635	1.0528	1.0639	1.0362	1.0388	1.0515	1.0557	1.0429	1.0509	1.0303	**1.0000**
30	1.0308	1.0406	1.0501	1.0376	1.0445	1.0199	1.0273	1.0296	1.0342	1.0338	1.0328	1.0168	**1.0000**
40	1.0111	1.0214	1.0307	1.0291	1.0309	1.0094	1.0019	1.0204	1.0263	1.0227	1.0233	1.0033	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.0180	1.0219	1.0216	1.0212	1.0394	1.0330	1.0063	1.0101	1.0073	1.0088	1.0192	1.0132	**1.0000**
20	1.0044	1.0132	1.0162	1.0194	1.0366	1.0242	1.0072	1.0049	1.0062	1.0056	1.0166	1.0115	**1.0000**
30	1.0013	1.0100	1.0145	1.0195	1.0327	1.0253	1.0010	1.0072	1.0014	1.0019	1.0148	1.0122	**1.0000**
40	1.0081	1.0042	1.0089	1.0149	1.0300	1.0214	1.0028	1.0032	1.0013	1.0099	1.0052	1.0023	**1.0000**

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column.

Table A10: Further Comparison of the Relative Prediction Efficiency (with Volume Only)

n_E	OLS_{10}	OLS_{11}	OLS_{12}	OLS_{13}	OLS_{14}	OLS_{15}	HRC^p_{10}	HRC^p_{11}	HRC^p_{12}	HRC^p_{13}	HRC^p_{14}	HRC^p_{15}	HRC^p
Panel A: Open Box Office													
							Mean Squared Forecast Error (MSFE)						
10	1.0614	1.0391	1.0312	1.0315	1.0309	1.0380	1.0551	1.0351	1.0297	1.0224	1.0255	1.0362	**1.0000**
20	1.0817	1.0181	1.0074	1.0122	1.0069	1.0137	1.0791	1.0102	0.9984	1.0108	1.0041	1.0121	**1.0000**
30	1.1556	1.0217	1.0176	1.0200	1.0207	1.0263	1.1517	1.0159	1.0107	1.0131	1.0117	1.0205	**1.0000**
40	1.1705	1.0267	1.0179	1.0198	1.0170	1.0271	1.1689	1.0227	1.0104	1.0164	1.0172	1.0199	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.0058	1.0113	1.0109	1.0113	1.0116	1.0115	1.0012	1.0037	1.0067	1.0119	1.0036	1.0018	**1.0000**
20	1.0228	1.0150	1.0160	1.0131	1.0117	1.0163	1.0148	1.0120	1.0069	1.0078	1.0045	1.0137	**1.0000**
30	1.0343	1.0122	1.0147	1.0149	1.0172	1.0212	1.0249	1.0075	1.0091	1.0125	1.0159	1.0158	**1.0000**
40	1.0280	1.0169	1.0194	1.0203	1.0213	1.0247	1.0264	1.0084	1.0104	1.0186	1.0166	1.0196	**1.0000**
Panel B: Movie Unit Sales													
							Mean Squared Forecast Error (MSFE)						
10	1.2868	1.2680	1.2518	1.1204	1.0814	1.0996	1.2614	1.2570	1.2493	1.1113	1.0772	1.0969	**1.0000**
20	1.2641	1.2501	1.2383	1.1429	1.0971	1.0951	1.2537	1.2472	1.2332	1.1340	1.0883	1.0879	**1.0000**
30	1.1739	1.1650	1.1541	1.0774	1.0604	1.0389	1.1700	1.1549	1.1439	1.0704	1.0522	1.0304	**1.0000**
40	1.1208	1.1178	1.1126	1.0543	1.0504	1.0125	1.1103	1.1082	1.1092	1.0474	1.0408	1.0093	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.1268	1.1229	1.1274	1.0750	1.0752	1.0715	1.1236	1.1128	1.1177	1.0668	1.0728	1.0724	**1.0000**
20	1.1125	1.1080	1.1138	1.0688	1.0631	1.0547	1.1096	1.0970	1.1043	1.0610	1.0632	1.0461	**1.0000**
30	1.0874	1.0886	1.0918	1.0492	1.0479	1.0439	1.0803	1.0828	1.0820	1.0490	1.0455	1.0364	**1.0000**
40	1.0784	1.0833	1.0877	1.0487	1.0474	1.0425	1.0768	1.0803	1.0786	1.0463	1.0434	1.0367	**1.0000**

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column.

E.7 Comparing ARMS and ARMSH

From the exercises in the main text, we notice that ARMS and ARMSH provide similar results in many cases. Although ARMSH is hetero-robust, ARMS and ARMSH end up with similar candidate model sets. In the following table A11, we show the 5 highest weight models estimated by HRC^p using candidate model sets screened by ARMS and ARMSH respectively. For each model screening method, an "x" denotes the associated explanatory variable is included in the particular model. Each model screening method contains a candidate model set of 100 selected models. Estimated model weights are presented in the last row for each method.

Table A11: Describing the 5 Highest Weight Models Using Model Sets Screened by ARMS and ARMSH

	ARMS					ARMSH				
	M1	**M2**	**M3**	**M4**	**M5**	**M1**	**M2**	**M3**	**M4**	**M5**
Genre										
Action										
Adventure	x			x	x	x			x	x
Animation										
Biography										
Comedy										
Crime	x					x				
Drama										
Family										
Fantasy	x	x	x	x	x	x	x	x	x	x
Horror	x			x	x	x			x	x
Mystery		x	x	x			x	x		x
Romance		x					x			
Sci-Fi										
Thriller										
Rating										
PG										
PG13										
R										
Core										
Budget	x	x	x	x	x	x	x	x	x	x
Weeks	x	x	x	x	x	x	x	x	x	x
Screens	x	x	x	x	x	x	x	x	x	x
Sentiment										
T-21/-27										
T-14/-20	x		x	x	x	x		x	x	x
T-7/-13		x					x			
T-4/-6										
T-1/-3	x	x	x	x	x	x	x	x	x	x
Volume										
T-21/-27	x	x	x	x	x	x	x	x	x	x
T-14/-20										
T-7/-13	x	x	x	x	x	x	x	x	x	x
T-4/-6	x	x	x	x	x	x	x	x	x	x
T-1/-3	x	x	x	x	x	x	x	x	x	x
Weights	0.4278	0.3914	0.1296	0.0332	0.0155	0.4283	0.4220	0.1038	0.0291	0.0168

Note: x denotes that explanatory variable is included in the particular model. The above exercise is carried out by using the top 100 models screened by ARMS and ARMSH respectively for open box office.

The top 5 models for each method accumulates more than 95% of the total weights. Moreover, we notice that the top 5 models for each method are identical with the same

ranking. This explains why in our prediction experiment, ARMS and ARMSH yield quite similar results in terms of forecast accuracy. In Subsection E.4, we conduct a Monte Carlo study to shed further light on the relative performance of ARMS and ARMSH under different scenarios related to what is the source of heteroskedasticity.

E.8 Performance of Recursive Partitioning Methods Using Identical Variables to Model Screening/Averaging Strategies

In the empirical exercises, we restrict that each potential model contains a constant term and 7 (11) relatively significant parameters for open box office (movie unit sales) based on the OLS results presented in table A1. To examine if our findings are driven by pre-selection, we compare the performance of recursive partitioning methods to econometric strategies using identical set of selected 7 (11) parameters. Results are presented in table A12.

As usual, we report the median MSFE and MAE of different strategies listed in panel A of table A12 for each evaluation set of different sizes $n_E = 10, 20, 30, 40$. Panel A presents results for forecasting open box office and panel B demonstrates results for forecasting movie unit sales. To ease interpretation, in each row of table A12 we normalize the MSFEs and MAFEs, respectively, by the MSFE and MAFE of the HRC^p.

For both panels, table A12 demonstrates that there are very large gains in prediction efficiency of the recursive partitioning algorithms relative to the benchmark HRC^p, although such gains are not as large as those demonstrated in table 5, in which the recursive partitioning methods use all the potential variables available. Take the MSFE results under $n_E = 10$ in panel A for example, Reg.Tree shows approximately 37% increase in prediction efficiency in table 5 and 20% increase in table A12. The results indicate that the pre-selected 7 (11) variables play crucial roles in predicting the open box office (movie unit sales). On the other hand, the other potential variables also jointly provide significant predicting power. In summary, the gains from machine learning strategies that use recursive partitioning over econometric methods is not due to differences in the set of predictors.

E.9 Test for Superior Predictive Ability (SPA) of the MAB Method

In this subsection, we perform the SPA test of Hansen (2005) to examine if the MAB method we proposed demonstrates superior predictive ability over all the other methods listed in this paper. We consider both the squared forecast error (SFE) and the absolute forecast error (AFE) as the quantities for comparing predictive ability. We set the results of MAB as the benchmark.

The null hypothesis of the SPA test states that the average performance of the benchmark is as good as the best average performance across the other competing methods. The

Table A12: Results of Relative Prediction Efficiency between Recursive Partitioning Methods Using Selective Variables and the Benchmark Method

n_E	Reg. Tree	Bagging	Random Forest			Benchmark
			RF_{10}	RF_{15}	RF_{20}	
Panel A: Open Box Office						
Mean Squared Forecast Error (MSFE)						
10	**0.8020**	0.9501	0.8155	0.8542	0.9559	1.0000
20	1.0149	0.9287	0.8560	**0.8540**	0.8940	1.0000
30	1.1125	0.8611	0.8679	**0.8525**	0.9940	1.0000
40	1.3306	1.1571	1.2549	1.1343	1.2340	**1.0000**
Mean Absolute Forecast Error (MAFE)						
10	**0.7794**	0.8487	0.7865	0.7973	0.8641	1.0000
20	0.8079	0.7635	0.7571	**0.7359**	0.7507	1.0000
30	0.8780	**0.8487**	0.8536	0.8670	0.8909	1.0000
40	**0.8501**	0.8539	0.8649	0.8837	0.8914	1.0000
Panel B: Movie Unit Sales						
Mean Squared Forecast Error (MSFE)						
10	0.9236	0.9580	**0.9009**	0.9151	0.9571	1.0000
20	1.0261	0.9600	0.9439	**0.9053**	0.9557	1.0000
30	1.2982	**0.9810**	1.0447	1.0652	1.1236	1.0000
40	1.1213	1.0037	0.9886	**0.9761**	0.9834	1.0000
Mean Absolute Forecast Error (MAFE)						
10	**0.8390**	0.9794	0.9201	0.9525	0.9443	1.0000
20	0.8409	**0.8303**	0.8448	0.8388	0.8563	1.0000
30	0.9485	**0.9103**	0.9431	0.9220	0.9250	1.0000
40	0.8905	0.8367	**0.8332**	0.8456	0.8398	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column.

alternative is that there is at least one competing method has better average performance than the benchmark. We estimate the p-values under the two forecast error quantities for open box office and movie unit sales. Large p-values signify the superior predictive ability of the MAB method over others.

Results for different n_E values are presented in table A13 and all the p-values are larger than 5% implying the superior predictive ability of the MAB method over others. This is particularly true for movie unit sales, in which the p-values are as high as 1 in all cases. The p-values for open box office under SFE are relatively smaller than other cases which coincides with the MSFE results demonstrated in table 5.

E.10 Model Averaging Regression Tree

This subsection considers adding a model averaging flavor to a single regression tree (MART). We duplicate the Monte Carlo simulations in section 4.2 and the MART method is represented by the lines with dots in figure A2. Although MART dominates RT for

Table A13: SPA Test Results of the MAB Method

n_E	Open Box Office		Movie Unit Sales	
	SFE	AFE	SFE	AFE
10	0.2651	1.0000	1.0000	1.0000
20	0.0912	0.8180	1.0000	1.0000
30	0.1770	1.0000	1.0000	1.0000
40	0.0938	1.0000	1.0000	1.0000

both heteroskedasticity scenarios in figures A2(a) and A2(b), it is clear in figures A2(c) and A2(d) that the MART method is outperformed by both MAB and MARF by a large margin in both scenarios. In fact under random heteroskedasticiy MART performs similarly to OLS estimation of GUM. This reinforces our claim that gains to adding model averaging to recursively partitioned subgroups occurs when there is systemic heterogeneity perhaps due to parameter heterogeneity. The MART method only outperforms GUM under parameter heterogeneity.

Figure A2: Risk Comparison under Different Scenarios