Variable Selection and Forecasting in High Dimensional Linear Regressions with Parameter Instability*

Alexander Chudik†
Federal Reserve Bank of Dallas

M. Hashem Pesaran
University of Southern California, USA and Trinity College, Cambridge, UK

Mahrad Sharifvaghefi
University of Pittsburgh

April 23, 2021

Abstract

This paper is concerned with the problem of variable selection and forecasting in the presence of parameter instability. There are a number of approaches proposed for forecasting in the presence of time-varying parameters, including the use of rolling windows and exponential down-weighting. However, these studies start with a given model specification and do not consider the problem of variable selection, which is complicated by time variations in the effects of signals on target variables. In this study we investigate whether or not we should use weighted observations at the variable selection stage in the presence of parameter instability, particularly when the number of potential covariates is large. Amongst the extant variable selection approaches we focus on the recently developed One Covariate at a time Multiple Testing (OCMT) method. This procedure allows a natural distinction between the selection and forecasting stages. We establish three main theorems on selection, estimation post selection, and in-sample fit. These theorems provide justification for using the full (not down-weighted) sample at the selection stage of OCMT and down-weighting of observations only at the forecasting stage (if needed). The benefits of the proposed method are illustrated by empirical applications to forecasting monthly stock market returns and quarterly output growths.

Keywords: Time-varying parameters, high-dimensional, multiple testing, variable selection, Lasso, one covariate at a time multiple testing (OCMT), forecasting, monthly returns, Dow Jones
JEL Classifications: C22, C52, C53, C55

*We are grateful to George Kapetanios and Ron Smith for constructive comments and suggestions. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Dallas or the Federal Reserve System. This research was supported in part through computational resources provided by the Big-Tex High Performance Computing Group at the Federal Reserve Bank of Dallas. This paper in part was written when Sharifvaghefi was a doctoral student at the University of Southern California (USC). Sharifvaghefi gradefully acknowledges financial support from the Center for Applied Financial Economics at USC.

†Corresponding author. Federal Reserve Bank of Dallas, Research Department, 2200 N Pearl St, Dallas, TX 75201. Email: alexander.chudik@gmail.com.

1 Introduction

"*When you have eliminated the impossible, whatever remains, however improbable, must be the truth*" Sir Arthur Conan Doyle, *The Sign of the Four* (1890)

There is mounting evidence that models fitted to many statistical relationships are subject to breaks. In an extensive early study, [25] find that a large majority of time series regressions in economics are subject to breaks. [4] consider parameter instability to be one of the main sources of forecast failure. This problem has been addressed at the estimation stage given a set of selected regressors. However, the issue of variable selection in the presence of time-varying parameters is still largely underdeveloped. In this study, we investigate whether or not we should use weighted observations at the variable selection stage in the presence of parameter instability, particularly when the number of potential covariates is large. We provide theoretical arguments in favor of using the full (unweighted) sample at the selection stage, and suggest that one should only consider weighting the observations post selection, at the estimation and forecasting stages.

Studies on breaks at the estimation stage usually assume a number of different model specifications that allow for parameter instability. Typical solutions are either to use rolling windows or exponential down-weighting. For instance, [22], [18] and [11] consider the choice of an observation window, and [10] and [19], respectively consider exponential and non-exponential down-weighting of the observations. There are also Bayesian approaches to prediction that allow for a possibility of breaks over the forecast horizon, e.g. [2], [15], and [17]. [23] provides a review of the literature on forecasting under instability. There are also related time varying parameter (TVP) and regime switching models that are used for forecasting. See, for example, [9] and [5]. All these studies take the model specification as given and then consider different ways of modeling and allowing for parameter instability. But, to the best of our knowledge, none of these studies considers the problem of variable selection in the presence of parameter instability.

In the absence of instability, it is optimal to weigh the observations equally for both variable selection and estimation purposes. Yet, in the presence of instability, the literature does not discuss whether or not weighted observations should be used at the variable selection stage, particularly when the number of potential covariates is large. There are a number of recent studies that consider predicting stock returns using penalized regression, especially the Least Absolute Shrinkage and Selection Operator (Lasso) initially proposed by [26] - for example, [1] and [14]. But they do not allow for parameter instability at the Lasso stage and suggest recursive application of Lasso using rolling windows. [16] have proposed a Lasso procedure that allows for a threshold effect. [12] have proposed a time-varying Lasso procedure, where all the parameters of the model vary locally. These are interesting extensions of Lasso, but are likely to suffer from the over-fitting problem, and could be sensitive to how cross validation is carried out. Also recently, [29] propose an interesting boosting procedure for the estimation of high-dimensional models with locally time varying parameters. It is important to note that, in the case of both penalized regression and boosting procedures, variable selection and estimation are carried out in one stage.

[3] propose an alternative procedure called **o**ne **c**ovariate at a time **m**ultiple **t**esting (OCMT). In the absence of parameter instability, the authors establish that the suggested procedure asymptotically selects all the relevant covariates and none of the pure noise covariates under general assumptions. Moreover, they show that the estimation errors of coefficients and prediction loss converge to zero. Finally, their Monte Carlo studies indicate that OCMT tends to perform better than penalized regression or boosting procedures under various designs. [24] has recently generalized the OCMT procedure to allow the covariates under consideration to be highly correlated, while penalized regression methods require the covariates to be weakly correlated (see e.g. [30]). One clear advantage of OCMT is its natural separation of the two problems of variable selection and estimation/forecasting. One can, therefore, decide whether to use the weighted observations at the variable selection stage or

not. In this paper we argue that in the presence of parameter instability full (unweighted) sample should be used for variable selection using the OCMT procedure. Forecasting can then be carried out conditional on the selected variables, using available techniques such as rolling or exponential downweighting techniques post selection. Also existing theoretical results from the forecasting literature can be applied to the post OCMT selected model to test for breaks and decide on the optimal choice of the estimation window or down-weighting among the remaining true covariates.

We provide three main theorems to back up our proposed selection/forecasting strategy. Under certain fairly general regularity conditions we show that the probability of selecting the true approximating model that contains all the signals and none of the noise variable tends to unity as the number of time series observations (T) and the number of covariates under consideration (N) tend to infinity. We also establish that least squares estimates of the coefficients of the selected covariates will tend to zero unless they are (true) signals. Lastly, we show that the mean square error of the selected model achieves the oracle rate for regression models with time-varying coefficients. These theoretical findings provide a formal justification for application of statistical techniques from the time-varying parameters literature to the post OCMT selected model.

Finally, we consider two empirical applications: forecasting monthly returns of stocks in Dow Jones and output growths across 33 countries, to illustrate the benefits of the OCMT procedure with full unweighted sample at the selection stage. Our results consistently suggest that using down-weighted observations at the selection stage of the OCMT procedure worsens forecast accuracy in terms of mean square forecast error and mean directional forecast accuracy. Moreover, our results suggest that overall OCMT with no down-weighting at the selection stage outperforms penalized regression methods, such as Lasso and/or Adaptive Lasso, which are prone to the over-fitting problem.

The rest of the paper is organized as follows: Section 2 sets out the model specification.

Section 3 explains the basic idea behind using the OCMT procedure with no down-weighting for variable selection in the presence parameter instability. Section 4 discusses the technical assumptions and the asymptotic properties of the OCMT procedure under parameter instability. Section 5 presents the empirical applications, and Section 6 concludes. Mathematical proofs are provided in the appendix.

Notations: Generic finite positive constants are denoted by C_i for $i = 1, 2, \cdots$. $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_F$ denote the spectral and Frobenius norms of matrix $\mathbf{A}$, respectively. $\lambda_i(\mathbf{A})$ denotes the i^{th} eigenvalue of a square matrix $\mathbf{A}$. $\|\mathbf{x}\|$ denotes the ℓ_2 norm of vector $\mathbf{x}$. If $\{f_n\}_{n=1}^{\infty}$ and $\{g_n\}_{n=1}^{\infty}$ are both positive sequences of real numbers, then $f_n = \ominus(g_n)$ if there exist $n_0 \geq 1$ and positive constants C_0 and C_1, such that $\inf_{n \geq n_0}(f_n/g_n) \geq C_0$ and $\sup_{n \geq n_0}(f_n/g_n) \leq C_1$.

2 Model specification under parameter instability

Consider the following data generating process (DGP) for the target variable, y_t, in terms of the signal variables (x_{it}, for $i = 1, 2, ..., k$)

$$y_t = \mathbf{z}_t'\mathbf{a}_t + \sum_{i=1}^{k} \beta_{it} x_{it} + u_t, \text{ for } t = 1, 2, \cdots, T \tag{1}$$

with time-varying parameters, $\mathbf{a}_t = (\mathrm{a}_{1t}, \mathrm{a}_{2t}, \cdots, \mathrm{a}_{mt})'$ and $\{\beta_{it},\, i = 1, 2, ..., k\}$, where $\mathbf{z}_t$ is an $m \times 1$ vector of pre-selected covariates, and u_t is an error term. Since the parameters are time-varying we refer to the covariate i as " *signal"* if the average value of its coefficient, $\bar{\beta}_{i,T} = T^{-1}\sum_{t=1}^{T} \mathbb{E}(\beta_{it})$, does not tend to zero very fast, namely such that $\bar{\beta}_{i,T} = \ominus(T^{-\vartheta_i})$ for some $0 \leq \vartheta_i < 1/2$.

Parameters can vary continuously following a stochastic process as in the standard random coefficient model, $\beta_{it} = \beta_i + \sigma_{it}\xi_{it}$, or could be fixed and change at discrete intervals: $\beta_{it} = \beta_i^{[s]}$, if $t \in [T_{s-1}, T_s)$ for $s = 1, 2, \cdots, S$, where $T_0 = 1$ and $T_S = T$. The vector $\mathbf{z_t}$ can contains deterministic components such as a constant, dummy variables, and a deterministic time trend as well as stochastic variables including observed common factors. The problem is that both the structure of the breaks and the identity of the k signals are unknown. The task

facing the investigator is to select the signals from a set of covariates under consideration, $\mathcal{S}_{Nt} = \{x_{1t}, x_{2t}, \cdots, x_{Nt}\}$, known as the active set, with N, the number of covariates in the active set, possibly much larger than T, the number of data points available for estimation prior to forecasting. We assume the coefficients ($\mathbf{a}_t$, and β_{it}, for $i = 1, 2, ..., k$) are independently distributed of the pre-selected covariates ($\mathbf{z}_t$) and all the covariates in the active set $\mathcal{S}_{Nt}$.

The application of penalized regression techniques to variable selection is theoretically justified under two key parameter stability assumptions: the stability of β_{it} and the stability of the correlation matrix of the covariates in the active set. Under these assumptions, the application of the penalized regression to the active set can proceed using the full sample without down-weighting or separating the variable selection from the forecasting stage. However, in the presence parameter instability, it is not clear how the use of penalized regressions could be justified. The problem has been recognized in the empirical literature focusing on slowly varying parameters and/or the use of rolling windows without making a distinction between variable selection and forecasting. It is also worth highlighting that in this paper, we relax the assumption of fixed correlation among the covariates in the active set, which is very common in the penalized regression studies, and allow for time-varying correlations.

In this paper we follow [3] and consider the application of the OCMT procedure for variable selection stage using the full unweighted sample, and provide theoretical arguments to justify such an approach. We first recall that OCMT's variable selection is based on the net effect of x_{it} on y_t conditional $\mathbf{z}_t$. However, when the regression coefficients and/or the correlations across the covariates in the active set are time-varying, the net effects will also be time-varying and we need to base our selection on average net effects. Also, we need to filter out the effects of the pre-selected covariates, $\mathbf{z}_t$, from x_{it} and y_t, before defining average net effects. To this end consider the following auxiliary regressions of x_{it} and y_t on $\mathbf{z}_t$:

$$\tilde{y}_t = y_t - \mathbf{z}_t'\bar{\boldsymbol{\psi}}_{y,T}, \text{ and } \tilde{x}_{it} = x_{it} - \mathbf{z}_t'\bar{\boldsymbol{\psi}}_{i,T} \tag{2}$$

where $\bar{\boldsymbol{\psi}}_{y,T}$ and $\bar{\boldsymbol{\psi}}_{i,T}$ are the $m \times 1$ vectors of projection coefficients defined by $\bar{\boldsymbol{\psi}}_{y,T} \equiv \left(T^{-1}\sum_{t=1}^{T}\mathbb{E}\left(\mathbf{z}_t\mathbf{z}_t'\right)\right)^{-1}\left(T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{z}_t y_t)\right)$ and $\bar{\boldsymbol{\psi}}_{i,T} \equiv \left(T^{-1}\sum_{t=1}^{T}\mathbb{E}\left(\mathbf{z}_t\mathbf{z}_t'\right)\right)^{-1}\left(T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{z}_t x_{it})\right)$.

Given the filtered series, $\tilde{x}_{it}$ and $\tilde{y}_t$, we now define the average net effect of covariate x_{it} on y_t, conditional on $\mathbf{z}_t$, as

$$\bar{\theta}_{i,T} = T^{-1}\textstyle\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}\tilde{y}_t). \tag{3}$$

Substituting for $\tilde{y}_t = y_t - \mathbf{z}_t'\,\bar{\boldsymbol{\psi}}_{y,T}$ in the above and noting that $\bar{\theta}_{i,T}$ is a given constant, then

$$\bar{\theta}_{i,T} = T^{-1}\textstyle\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}y_t) - \bar{\boldsymbol{\psi}}_{y,T}'\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}\mathbf{z}_t)\right].$$

Also,

$$\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}\mathbf{z}_t) = \sum_{t=1}^{T}\mathbb{E}(x_{it}\mathbf{z}_t) - \left[\sum_{t=1}^{T}\mathbb{E}(\mathbf{z}_t\mathbf{z}_t')\right]\bar{\boldsymbol{\psi}}_{i,T} = \sum_{t=1}^{T}\mathbb{E}(x_{it}\mathbf{z}_t) - \sum_{t=1}^{T}\mathbb{E}(x_{it}\mathbf{z}_t) = \mathbf{0}. \tag{4}$$

Hence, $\bar{\theta}_{i,T} = T^{-1}\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}y_t)$. Now by substituting y_t from (1) we can further write,

$$\begin{aligned}
\bar{\theta}_{i,T} &= T^{-1}\textstyle\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}y_t) = T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[\tilde{x}_{it}\left(\mathbf{z}_t'\mathbf{a}_t + \sum_{j=1}^{k}\beta_{jt}x_{jt} + u_t\right)\right] \\
&= \mathbf{a}'T^{-1}\textstyle\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}\mathbf{z}_t) + T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{k}\mathbb{E}(\beta_{jt})\mathbb{E}(\tilde{x}_{it}x_{jt}) + T^{-1}\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}u_t) \\
&= \textstyle\sum_{j=1}^{k}\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\beta_{jt})\mathbb{E}(\tilde{x}_{it}x_{jt})\right] + T^{-1}\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}u_t).
\end{aligned}$$

Therefore, the average net effect can be written simply as

$$\bar{\theta}_{i,T} = \textstyle\sum_{j=1}^{k}\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\beta_{jt})\sigma_{ij,t}(\mathbf{z})\right] + \bar{\sigma}_{iu,T}(\mathbf{z}), \tag{5}$$

where $\sigma_{ij,t}(\mathbf{z}) = \mathbb{E}(\tilde{x}_{it}x_{jt})$, and $\bar{\sigma}_{iu,T}(\mathbf{z}) = T^{-1}\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}u_t)$. Furthermore, $\bar{\sigma}_{iu,T}(\mathbf{z}) = T^{-1}\sum_{t=1}^{T}\mathbb{E}(x_{it}u_t) - \bar{\boldsymbol{\psi}}_{i,T}'\left(T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{z}_t u_t)\right)$, which will be identically zero if the covariates and the conditioning variables are weakly exogenous with respect to u_t. In what follows we allow for a mild degree of correlation between $(x_{it}, \mathbf{z}_t)$ and u_t by assuming that $\bar{\sigma}_{iu,T}(\mathbf{z}) = \ominus(T^{-\epsilon_i})$, for some $\epsilon_i > 1/2$. It is also easily seen that when the parameters and the cross covariate covariances are time-invariant the above average net effect reduces to $\theta_i = \Sigma_{j=1}^{k}\mathbb{E}(\beta_j)\sigma_{ij}(\mathbf{z})$.

Given the average net effect of x_{it} on y_t, the covariates in the active set can be categorized into three groups: *signals*, *pseudo-signals* and *noise variables*. As mentioned before, *signals* are those covariates with average value of their coefficient, namely $\bar{\beta}_{iT} = T^{-1}\sum_{t=1}^{T}\mathbb{E}(\beta_{it})$, which does not approach zero too fast, namely $\bar{\beta}_{i,T} = \ominus(T^{-\vartheta_i})$, for some $0 \leq \vartheta_i < 1/2$. *Pseudo-signals* are the covariates that do not enter the DGP but have average net effects, $\bar{\theta}_{i,T}$, that do not converge to zero sufficiently fast, namely $\bar{\theta}_{i,T} = \ominus(T^{-\vartheta_i})$, for some $0 \leq \vartheta_i < 1/2$. Finally, *noise variables* are those covariates that do not enter the DGP and at the same time have either zero or sufficiently small average net effects in the sense that $\bar{\theta}_{i,T} = \ominus(T^{-\epsilon_i})$, for some $\epsilon_i > 1/2$.

In what follows, we first describe the OCMT procedure and then discuss the conditions under which the approximating model that includes all the signals and none of the noise variables is selected by OCMT.

3 Parameter instability and OCMT

The OCMT procedure considers the following N regressions of y_t on each of the covariates in the active set $\mathcal{S}_{Nt}$ one at a time, conditional on $\mathbf{z}_t$:

$$y_t = \boldsymbol{\varrho}'_{i,T}\mathbf{z}_t + \phi_{i,T}x_{it} + \eta_{it}, \text{ for } t = 1, 2, \cdots, T;\ i = 1, 2, ..., N, \tag{6}$$

where $\phi_{i,T} = \left[T^{-1}\sum_{t=1}^{T}\mathbb{E}\left(\tilde{x}_{it}^2\right)\right]^{-1}\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}\left(\tilde{x}_{it}\tilde{y}_t\right)\right] = \left[\bar{\sigma}_{ii,T}(\mathbf{z})\right]^{-1}\bar{\theta}_{i,T}$, with $\bar{\sigma}_{ii,T}(\mathbf{z}) = T^{-1}\sum_{t=1}^{T}\sigma_{ii,t}(\mathbf{z})$. [3] assume parameter stability, and set $\beta_{it} = \beta_i$ for all t, where β_i is deterministic, and assume zero conditional correlation between the signals and the error term, namely $\sigma_{iu,t} = 0$ for all t. Under parameter stability the average net effects can be simplified to the net effects defined by $\theta_{i,T} = \sum_{j=1}^{k}\beta_j\bar{\sigma}_{ij,T}(\mathbf{z})$, for $i = 1, 2, \cdots, N$, where $\bar{\sigma}_{ij,T}(\mathbf{z}) = T^{-1}\sum_{t=1}^{T}\sigma_{ij,t}(\mathbf{z})$. Hence,

$$\phi_{i,T} = \frac{\theta_{i,T}}{\bar{\sigma}_{ii,T}(\mathbf{z})} = \frac{\sum_{j=1}^{k}\beta_j\bar{\sigma}_{ij,T}(\mathbf{z})}{\bar{\sigma}_{ii,T}(\mathbf{z})}. \tag{7}$$

However, in the more general set up of DGP (1), the net effect of x_{it} on y_t is time-varying. Therefore, by running one at a time regressions of y_t on each of the covariates: x_{it}, $i = 1, 2, \cdots, N$, we focus on average net effect of x_{it} on y_t, defined over the full sample, denoted by $\bar{\theta}_{i,T}$ and given by (5).

Due to non-zero correlations between the covariates, knowing whether $\bar{\theta}_{i,T}$ is zero or not does not necessarily allow us to establish whether $\bar{\beta}_{i,T}$ is sufficiently close to zero or not. There are four possibilities:

(I) *Signals*	$\bar{\beta}_{i,T} = \ominus(T^{-\vartheta_i})$ and $\bar{\theta}_{i,T} = \ominus(T^{-\vartheta_i})$
(II) *Hidden Signals*	$\bar{\beta}_{i,T} = \ominus(T^{-\vartheta_i})$ and $\bar{\theta}_{i,T} = \ominus(T^{-\epsilon_i})$
(III) *Pseudo-signals*	$\bar{\beta}_{i,T} = \ominus(T^{-\epsilon_i})$ and $\bar{\theta}_{i,T} = \ominus(T^{-\vartheta_i})$
(IV) *Noise variables*	$\bar{\beta}_{i,T} = \ominus(T^{-\epsilon_i})$ and $\bar{\theta}_{i,T} = \ominus(T^{-\epsilon_i})$

for some $0 \leq \vartheta_i < 1/2$, and $\epsilon_i > 1/2$. Notice, if the covariate x_{it} is a noise variable, then $\bar{\theta}_{i,T}$, the average net effect of x_{it} on y_t, converges to zero very fast. Therefore, down-weighting of observations at the variable selection stage is likely to be inefficient for eliminating the noise variables. Moreover, for a signal to remain hidden, we need the terms of higher order, $\ominus(T^{-\vartheta_j})$ with $0 \leq \vartheta_i < 1/2$, to *exactly* cancel out such that $\theta_{i,T}$ becomes a lower order, i.e. $\ominus(T^{-\varepsilon_i})$, that tends to zero at a sufficiently fast rate (with $\epsilon_i > 1/2$). This combination of events seem quite unlikely, and to simplify the theoretical derivations in what follows we abstract from such a possibility and assume that there are no hidden signals and only consider the first stage of the OCMT procedure for variable selection.[1]

The OCMT procedure

1. For $i = 1, 2, \cdots, N$, regress $\mathbf{y} = (y_1, y_1, ..., y_T)'$ on $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T)'$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iT})'$; $\mathbf{y} = \mathbf{Z}\boldsymbol{\varrho}_{i,T} + \phi_{i,T}\mathbf{x}_i + \boldsymbol{\eta}_i$; and compute the t-ratio of $\phi_{i,T}$, given by

$$t_{i,T} = \frac{\hat{\phi}_{i,T}}{s.e.\left(\hat{\phi}_{i,T}\right)} = \frac{\mathbf{x}_i'\mathbf{M}_{\mathbf{z}}\mathbf{y}}{\hat{\sigma}_i\sqrt{\mathbf{x}_i'\mathbf{M}_{\mathbf{z}}\mathbf{x}_i}},$$

[1] To allow for hidden signals, [3] extend the OCMT method to have multiple stages.

where $\hat{\phi}_{i,T} = \left(\mathbf{x}_i'\mathbf{M}_{\mathbf{z}}\mathbf{x}_i\right)^{-1}\left(\mathbf{x}_i'\mathbf{M}_{\mathbf{z}}\mathbf{y}\right)$ is the Ordinary Least Square (OLS) estimator of $\phi_{i,T}$, $\hat{\sigma}_i^2 = \hat{\boldsymbol{\eta}}_i'\hat{\boldsymbol{\eta}}_i/T$, and $\hat{\boldsymbol{\eta}}_i$ is a $T \times 1$ vector of regression residuals.

2. Consider the critical value function, $c_p(N,\delta)$, defined by

$$c_p(N,\delta) = \Phi^{-1}\left(1 - p/2N^{\delta}\right), \tag{8}$$

where $\Phi^{-1}(.)$ is the inverse of a standard normal distribution function; δ is a finite positive constant; and p is the nominal size of the tests to be set by the investigator.

3. Given $c_p(N,\delta)$, the selection indicator is given by

$$\hat{\mathcal{J}}_i = I\left[|t_{i,T}| > c_p(N,\delta)\right], \text{ for } i = 1,2,\cdots,N. \tag{9}$$

The covariate x_{it} is selected if $\hat{\mathcal{J}}_i = 1$.

The main goal of OCMT is to use the t-ratio of the estimated $\phi_{i,T}$ to select all the signals and none of the noise variables, the selected model is referred to as an *approximating model* since it can include pseudo-signals. To deal with the multiple testing nature of the problem, the critical value of the tests is chosen to be an appropriately increasing function of N.

4 Asymptotic properties of OCMT under parameter instability

We now provide the theoretical justification for using the OCMT procedure for variable selection in models with time-varying parameters. It is assumed that $m = \dim(\mathbf{z}_t)$ and k, the number of signals, are finite fixed integers. But we allow the number of pseudo-signals, which we denote by k_T^* , to grow at a sufficiently slow rate relative to N and T. Finally, we define the *approximating model* to be a model that contains all the signals, $\{x_{it} : i = 1,2,\cdots,k\}$, and none of the noise variables, $\{x_{it} : k + k_T^* + 1, k + k_T^* + 2, \cdots, N\}$. Clearly, such a model can contain one or more of the pseudo-signals, $\{x_{it} : k+1, k+2, \cdots, k+k_T^*\}$. We start with some technical assumptions in Section 4.1 and then provide the asymptotic properties of the OCMT procedure under parameter instability in Section 4.2.

4.1 Technical assumptions

Let $\mathbf{q}_t = (z_{1t}, z_{2t}, \cdots, z_{mt}, x_{1t}, x_{2t}, \cdots, x_{Nt})'$ be an $(m+N) \times 1$ vector that include all the covariates. In what follows we make use of the following filtrations: $\mathcal{F}_t^q = \sigma(\mathbf{q}_t, \mathbf{q}_{t-1}, \cdots)$, $\mathcal{F}_t^{\mathrm{a}} = \sigma(\mathbf{a}_t, \mathbf{a}_{t-1}, \cdots)$, $\mathcal{F}_{jt}^{\beta} = \sigma(\beta_{jt}, \beta_{j,t-1}, \cdots)$ for $j = 1, 2, \cdots, k$ and $\mathcal{F}_t^u = \sigma(u_t, u_{t-1}, \cdots)$. Moreover, we set $\mathcal{F}_t^{\beta} = \cup_{j=1}^k \mathcal{F}_{jt}^{\beta}$ and $\mathcal{F}_t = \mathcal{F}_t^q \cup \mathcal{F}_t^{\mathrm{a}} \cup \mathcal{F}_t^{\beta} \cup \mathcal{F}_t^u$.

Assumption 1 (Martingale difference processes)

(a) $\mathbb{E}\left[\mathbf{q}_t\mathbf{q}_t' - \mathbb{E}(\mathbf{q}_t\mathbf{q}_t\prime)|\mathcal{F}_{t-1}\right] = 0$ *for* $t = 1, 2, \cdots, T$.

(b) $\mathbb{E}\left[u_t^2 - \mathbb{E}\left(u_t^2\right)|\mathcal{F}_{t-1}\right] = 0$ *for* $t = 1, 2, \cdots, T$.

(c) $\mathbb{E}\left[\mathbf{q}_t u_t - \mathbb{E}(\mathbf{q}_t u_t)|\mathcal{F}_{t-1}\right] = 0$ *for* $t = 1, 2, \cdots, T$.

(d) $\mathbb{E}\left[\mathrm{a}_{\ell t} - \mathbb{E}(\mathrm{a}_{\ell t})|\mathcal{F}_{t-1}\right] = 0$ *for* $\ell = 1, 2, \cdots, m$ *and* $t = 1, 2, \cdots, T$.

(e) $\mathbb{E}\left[\beta_{it} - \mathbb{E}(\beta_{it})|\mathcal{F}_{t-1}\right] = 0$ *for* $i = 1, 2, \cdots, k$ *and* $t = 1, 2, \cdots, T$.

Assumption 2 (Exponential decaying probability tails)

There exist sufficiently large positive constants C_0 *and* C_1*, and* $s > 0$ *such that*

(a) $\sup_{j,t} \Pr(|q_{jt}| > \alpha) \leq C_0 \exp(-C_1\alpha^s)$, *for all* $\alpha > 0$.

(b) $\sup_{\ell,t} \Pr(|\mathrm{a}_{\ell t}| > \alpha) \leq C_0 \exp(-C_1\alpha^s)$, *for all* $\alpha > 0$.

(c) $\sup_{i,t} \Pr(|\beta_{it}| > \alpha) \leq C_0 \exp(-C_1\alpha^s)$, *for all* $\alpha > 0$.

(d) $\sup_t \Pr(|u_t| > \alpha) \leq C_0 \exp(-C_1\alpha^s)$, *for all* $\alpha > 0$.

Assumption 3 (Coefficients of signals)

(a) The number of signals, k*, is a finite fixed integer.*

(b) β_{it}*,* $i = 1, 2, \cdots, k$*, are independent of* $q_{jt'}$*,* $j = 1, \cdots, m+N$*, and* $u_{t'}$ *for all* t *and* t'.

(c) $\bar{\beta}_{i,T} \equiv T^{-1}\sum_{t=1}^T \mathbb{E}(\beta_{it}) = \ominus(T^{-\vartheta_i})$*, for some* $0 \leq \vartheta_i < 1/2$.

Assumption 4 (Coefficients of conditioning variables)

(a) The number of conditioning variables, m, is finite.

(b) $\mathrm{a}_{\ell t}$, $\ell = 1, 2, \cdots, m$, *are independent of* $q_{jt'}$,$j = 1, \cdots, m+N$, *and* $u_{t'}$ *for all* t *and* t'.

(c) $\mathbb{E}(\mathrm{a}_{\ell t}) = \mathrm{a}_{\ell}$ *for* $\ell = 1, 2, \cdots, m$ *and all* t.

Before presenting the theoretical results, we briefly mention pros and cons of our assumptions and compare them with the assumptions typically made in the high-dimensional linear regression and the time-varying parameters literature.

Assumptions 1 allows the variables $z_{\ell t}$, $\mathrm{a}_{\ell t}$, x_{it}, β_{it} and u_t to follow martingale difference processes, which is weaker than the IID assumption typically made in the literature. Following a similar line of argument as in Section 4.2 of [3], we can relax some of these assumptions somewhat to allow for weak serial correlation in $z_{\ell t}$, $\mathrm{a}_{\ell t}$, x_{it}, β_{it} and u_t.

Assumption 2 imposes the variables $z_{\ell t}$, $\mathrm{a}_{\ell t}$, x_{it}, β_{it} and u_t to have exponentially decaying probability tails to ensure all moments exist. This assumption is stronger than those needed in the studies on breaks, but it is required to drive upper and lower probability bounds for selection of the approximating model. It is common in the high-dimensional linear literature to assume some form of exponentially decaying probability bound for the variables. For example, see [31], [8] and [3].

Assumptions 3(a) and 4(a) are required to establish that the target variable, y_t, has the exponentially decaying probability tail of the same order as the other random variables. Assumptions 3(b) and 4(b) ensure the distribution of time-varying parameters $\mathrm{a}_{\ell t}$ and β_{it} to be independent of the observed covariates (x_{it} and $z_{\ell t}$) and u_t, which is a standard assumption in the literature on time-varying parameters. Assumptions 3(c) ensures the average value of the coefficients of the signal variables does not approach zero too fast. It is an identification assumption that allows distinguishing signal from noise variables. Finally, Assumption 4(c) constrains the expected values of coefficients of pre-selected covariates to be time-invariant.

4.2 Theoretical findings

As mentioned in Section 1, the purpose of this paper is to provide the theoretical argument for applying the OCMT procedure with no down-weighting at the variable selection stage in linear high-dimensional settings subject to parameter instability. We now show that under certain conditions discussed in Section 4.1, the OCMT procedure selects the approximating model that contains all the signals; $\{x_{it} : i = 1, 2, \cdots, k\}$; and none of the noise variables; $\{x_{it} : k + k_T^* + 1, k + k_T^* + 2, \cdots, N\}$. The event of choosing the approximating model is defined by

$$\mathcal{A}_0 = \left\{\sum_{i=1}^{k} \hat{\mathcal{J}}_i = k\right\} \cap \left\{\sum_{i=k+k_T^*+1}^{N} \hat{\mathcal{J}}_i = 0\right\}. \tag{10}$$

Note the the approximate model can contain pseudo-signals variables. In what follows, we show that $\Pr(\mathcal{A}_0) \to 1$, as $N, T \to \infty$.

Theorem 1 *Let y_t for $t = 1, 2, \cdots, T$ be generated by (1), and let $T = \ominus(N^{\kappa_1})$ with $\kappa_1 > 0$, and $\mathcal{S}_{Nt} = \{x_{1t}, x_{2t}, \cdots, x_{Nt}\}$ which contains k signals, k_T^* pseudo-signals, and $N - k - k_T^*$ noise variables. Consider the OCMT procedure with the critical value function $c_p(N, \delta)$ given by (8), for some $\delta > 0$. Then under Assumptions 1-4, there exist finite positive constants C_0, and C_1 such that, the probability of selecting the approximating model, $\mathcal{A}_0$, defined by (10), is given by*

$$\Pr(\mathcal{A}_0) = 1 - O(N^{1-2C_0\delta}) - O[\exp(-N^{C_1\kappa_1})]. \tag{11}$$

See Appendix A.1 for a proof.

It is interesting that the asymptotic results regarding the probability of selecting the approximating model are unaffected by parameter instability, so long as the average net effects of the signals are non-zero or tend to zero sufficiently slowly in T, as defined formally by Assumption 3. In the next step, we focus on estimation of the coefficients of the selected model. To simply the exposition assume that there are no pre-selected covariates, in which

case, the DGP (1) simplifies to

$$y_t = \sum_{i=1}^{k} \beta_{it} x_{it} + u_t = \boldsymbol{\beta}_t' \mathbf{x}_{kt} + u_t, \text{ for } t = 1, 2, \cdots, T, \tag{12}$$

where $\mathbf{x}_{kt} = (x_{1t}, x_{2t}, \cdots, x_{kt})'$ and $\boldsymbol{\beta}_t = (\beta_{1t}, \beta_{2t}, \cdots, \beta_{kt})'$. For the next set of results the following additional assumption is also needed.

Assumption 5 (Eigenvalues) *Let* $\mathbf{x}_{kk_T^*,t} = \left(x_{1t}, x_{2t}, \cdots, x_{kt}, x_{k+1,t}, x_{k+2,t}, \cdots, x_{k_T^* t}\right)'$ *be a* $(k + k_T^*) \times 1$ *vector of signals and pseudo-signals, then* $\lambda_{\min}\left[T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{x}_{kk_T^*,t} \mathbf{x}_{kk_T^*,t}')\right] > c > 0$.

This assumption ensures that the post OCMT selected model can be estimated and the associated regressions coefficients can be consistently estimated subject to certain regularity conditions to be discussed above. This assumption rules out perfect multicollinearity among the signals and selected pseudo signals. It also requires that $k_T^* < T$, which will be met for sufficiently large T, under our assumption, namely that $k_T^*/T \to 0$, as $T \to \infty$, at a sufficiently fast rate.

The post OCMT selected model can be written as

$$y_t = \sum_{i=1}^{N} \hat{\mathcal{J}}_i x_{it} b_i + \eta_t \tag{13}$$

where $\hat{\mathcal{J}}_i = I\left[|t_{i,T}| > c_p(N, \delta)\right]$, defined by (9). Also $\sum_{i=1}^{N} \hat{\mathcal{J}}_i = \hat{k}_T$, where $\hat{k}_T$ is the number of covariates selected by OCMT. By Theorem 1 the probability that the selected model contains the signals tends to unity as $T \to \infty$. We can further write

$$y_t = \sum_{i=1}^{N} \hat{\mathcal{J}}_i x_{it} b_i + \eta_t = \sum_{\ell=1}^{\hat{k}_T} \gamma_\ell w_{\ell t} + \eta_t, \tag{14}$$

where $\mathbf{w}_t = \left(w_{1t}, w_{2t}, \cdots, w_{\hat{k}_T t}\right)'$. The least squares (LS) estimator of selected coefficients, $\boldsymbol{\gamma}_T = \left(\gamma_1, \gamma_2, \cdots, \gamma_{\hat{k}_T}\right)'$, is given by

$$\hat{\boldsymbol{\gamma}}_T = \left(T^{-1} \sum_{t=1}^{T} \mathbf{w}_t \mathbf{w}_t'\right)^{-1} \left(T^{-1} \sum_{t=1}^{T} \mathbf{w}_t y_t\right), \tag{15}$$

In establishing the rate of convergence of $\hat{\gamma}_T$ we distinguish between two cases: when the

vector of signals, $\mathbf{x}_{k,t} = (x_{1t}, x_{2t}, \cdots, x_{kt})$ is included in $\mathbf{w}_t$ as a subset, and when this is not the case. But we know by Theorem 1 the probability of the latter tends to zero at a sufficiently fast rate. The following theorem provides the conditions under which the estimator of the coefficients of the selected pseudo-signals and signals tend to their mean values, defined formally below.

Theorem 2 *Let the DGP for y_t, $t = 1, 2, \cdots, T$ be given by (12) and write down the regression model selected by the OCMT procedure as (14). Suppose that Assumptions 1-5 hold and the number of pseudo-signals, k_T^*, grow with T such that $k_T^* = \ominus(T^d)$ with $0 \leq d < \frac{1}{2}$. Consider the LS estimator of $\boldsymbol{\gamma}_T = \left(\gamma_1, \gamma_2, \cdots, \gamma_{\hat{k}_T}\right)'$, given by (15).*

(i) If $\mathbb{E}(\beta_{it}) = \beta_i$ for all t, then,

$$\|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| = O_p\left(T^{\frac{d-1}{2}}\right), \tag{16}$$

where $\boldsymbol{\gamma}_T^* = (\gamma_1^*, \gamma_2^*, \cdots, \gamma_{\hat{k}_T}^*)'$, and

$$\begin{cases} \gamma_\ell^* \in \boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_k)', & \text{if } w_{\ell t} \in \mathbf{x}_{kt} \\ \gamma_\ell^* = 0, & \text{otherwise.} \end{cases} \tag{17}$$

(ii) If $\mathbb{E}\left(\mathbf{x}_{kk_T^*,t}\mathbf{x}'_{kk_T^*,t}\right) = \boldsymbol{\Sigma}$ is a fixed matrix, then,

$$\|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^\diamond\| = O_p\left(T^{\frac{d-1}{2}}\right), \tag{18}$$

where $\boldsymbol{\gamma}_T^\diamond = (\gamma_{1T}^\diamond, \gamma_{2T}^\diamond, \cdots, \gamma_{\hat{k}_T,T}^\diamond)'$, and

$$\begin{cases} \gamma_{\ell,T}^\diamond \in \bar{\boldsymbol{\beta}}_T = (\bar{\beta}_{1T}, \bar{\beta}_{2T}, \cdots, \bar{\beta}_{kT})', & \text{if } w_{\ell t} \in \mathbf{x}_{kt} \\ \gamma_{\ell,T}^\diamond = 0, & \text{otherwise,} \end{cases} \tag{19}$$

and $\bar{\beta}_{iT} = T^{-1}\sum_{t=1}^T \mathbb{E}(\beta_{it})$, $i = 1, 2, \cdots, k$.

See Appendix A.2 for a proof.

Remark 1 *The above theorem builds on Theorem 1 and establishes that in the post OCMT selected model only signals will end up having non-zero limiting values, as N and $T \to \infty$, so*

long as $0 \leq d < 1/2$ *and* δ *is sufficiently large.* d *controls the rate at which number of pseudo-signals is allowed to rise with* T*. The latter condition rules out the possibility of true and pseudo-signals sharing the same unobserved common factors. To deal with such a possibility, following [24], one can first filter out the common factors using principle components (PC) and then apply the OCMT procedure to the least squares residuals of the regressions of the covariates on one or more of their top PCs.*

Remark 2 *The conditions of Theorem 2 are met in the case of random coefficient models where* $\beta_{it} = \beta_i + \sigma_{it}\xi_{it}$*, and* ξ_{it} *are distributed independently of the signals (and of the pre-selected covariates, if any), and the LS estimator of* $\boldsymbol{\gamma}_T^*$ *is consistent, so long as* $0 \leq d < 1/2$*. Interestingly, if signal and pseudo-signal variables are generated by a stationary process, and hence they satisfy condition (ii) of Theorem 2, then we can extend the random coefficient model to have time-variant means, and still estimate* $\boldsymbol{\gamma}_T^*$ *consistently by LS.*

Lastly, we provide our finding about the mean square error (MSE) of the selected model estimated by LS. Here, we need one more assumption as described below.

Assumption 6 (Cross product variations in signal coefficients)
$\mathbb{E}\left[\beta_{it}\beta_{jt} - \mathbb{E}(\beta_{it}\beta_{jt})|\mathcal{F}_{t-1}\right] = 0$ *for* $i = 1, 2, \cdots, k$*,* $j = 1, 2, \cdots, k$*, and* $t = 1, 2, \cdots, T$*.*

Remark 3 *Assumption 6 ensures that the MSE of the oracle model that contains only the signals, exists. This assumption can be relaxed to allow for weak time dependence in* $x_{it}\beta_{it}$*.*

The in-sample error of post OCMT selected model can be written as

$$\hat{\eta}_t = y_t - \sum_{\ell=1}^{\hat{k}_T} \hat{\gamma}_\ell w_{\ell t}. \tag{20}$$

The following theorem establishes the limiting property of MSE of the selected model, given by $T^{-1}\sum_{t=1}^{T}\hat{\eta}_t^2$.

Theorem 3 *Let the DGP for* y_t*,* $t = 1, 2, \cdots, T$ *be given by (12) and write down the regression model selected by the OCMT procedure as (14). The error of the selected model,*

estimated by LS, is given by (20). Suppose that Assumptions 1-6 hold and the number of pseudo-signals, k_T^*, grow with T such that $k_T^* = \ominus(T^d)$ with $0 \leq d < \frac{1}{2}$.

(i) If $\mathbb{E}(\beta_{it}) = \beta_i$ for all t, then

$$T^{-1}\sum_{t=1}^{T}\hat{\eta}_t^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(T^{-1}\sum_{t=1}^{T}\sigma_{ijt,x}\sigma_{ijt,\beta}\right)+\bar{\sigma}_{u,T}^2+O_p\left(T^{-\frac{1}{2}}\right)+O_p\left(T^{d-1}\right), \quad (21)$$

where $\sigma_{ijt,x} = \mathbb{E}\left(x_{it}x_{jt}\right)$, $\sigma_{ijt,\beta} = \mathbb{E}\left[(\beta_{it}-\beta_i)(\beta_{jt}-\beta_j)\right]$, and $\bar{\sigma}_{u,T}^2 = T^{-1}\mathbb{E}\left(\mathbf{u}'\mathbf{u}\right)$.

(ii) If $\mathbb{E}\left(\mathbf{x}_{kk_T^*,t}\mathbf{x}'_{kk_T^*,t}\right) = \mathbf{\Sigma}$ is a fixed matrix, then,

$$T^{-1}\sum_{t=1}^{T}\hat{\eta}_t^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(T^{-1}\sum_{t=1}^{T}\sigma_{ijt,x}\sigma^*_{ijt,\beta}\right)+\bar{\sigma}_{u,T}^2+O_p\left(T^{-\frac{1}{2}}\right)+O_p\left(T^{d-1}\right), \quad (22)$$

where $\sigma^*_{ijt,\beta} = \mathbb{E}\left[(\beta_{it}-\bar{\beta}_{i,T})(\beta_{jt}-\bar{\beta}_{j,T})\right]$.

See Appendix A.3 for a proof.

Remark 4 *The condition $d < \frac{1}{2}$ in Theorem 3 ensures that the number of pseudo-signals grows sufficiently slowly relative to T, and as a result, $T^{1-d} < T^{-\frac{1}{2}}$ and hence from equations (21) and (22), we can conclude that the MSE of the Post OCMT selected model convergences at the same rate of $T^{-\frac{1}{2}}$ under both scenarios (i) and (ii).*

Remark 5 *Results (21) and (22) show that the MSE of the selected model depends on (i) pure uncertainty due to the unobserved error term, u_t, of the DGP, as given by the term $\bar{\sigma}_{u,T}^2$, (ii) the traditional $O_p\left(T^{-1/2}\right)$ sampling uncertainty, which dominates the additional $O_p\left(T^{d-1}\right)$ uncertainty due to inclusion of $k_T^* = \ominus(T^d)$ pseudo-signals, and (iii) an additional term that depends on the product of $\sigma_{ijt,x}$ and $\sigma_{ijt,\beta}$ (or $\sigma^*_{ijt,\beta}$), which represents the cost (in terms of fit) of ignoring the time variation in the coefficients of the signals, β_{it}, $i = 1, 2, \cdots, k$. This cost is larger when the time variation in the coefficients of the signals (as measured by $\sigma_{ijt,\beta}$ or $\sigma^*_{ijt,\beta}$) is larger, and, for a given $\sigma_{ijt,\beta} \neq 0$, it increases with $\sigma_{ijt,x}$. This finding for the in-sample fit is similar to the results for mean square forecast errors (MSFE) in the presence of breaks in the literature, such as Proposition 2 of [22] or equation (20) of [19], where the main focus is to minimize the MSFE by mitigating the cost from the time*

variation in parameters at the expense of increased sampling uncertainty by weighting the observations, such as the use of optimal estimation windows or exponential downweighting of observations.

Remark 6 *The above three theorems require the exponent δ in the critical value function, (8), to be sufficiently large such that $\delta > \frac{1}{2C_0}$, for some positive constant C_0. The extensive Monte Carlo Studies in [3] suggest that setting $\delta = 1$ preforms well in practice.*

5 Empirical applications

The rest of the paper considers a number of empirical applications whereby the forecast performance of the proposed OCMT approach with no down-weighting at the selection stage is compared with those of Lasso and Adaptive Lasso. In particular, we consider the following two applications:[2]

- Forecasting monthly rate of price changes for 28 (out of 30) stocks in Dow Jones using a relatively large number of financial, economic, as well as technical indicators.
- Forecasting quarterly output growth rates across 33 countries using macro and financial variables.

In each application, we first compare the performance of OCMT with and without down-weighted observations at the selection stage. We then consider the comparative performance of OCMT (with variable selection carried out without down-weighting) relative to Lasso and Adaptive Lasso, with and without down-weighting. For down-weighting we make use of exponentially down-weighted observations, namely $\hat{x}_{it}(\lambda) = \lambda^{T-t}x_{it}$, and $\hat{y}_t(\lambda) = \lambda^{T-t}y_t$, where y_t is the target variable to be forecasted, x_{it}, for $i = 1, 2, ..., N$ are the covariates in the active set, and λ is the exponential decay coefficient. We consider two sets of values for the degree of exponential decay, λ: (1) Light down-weighting with $\lambda = 0.975, 0.98, 0.985, 0.99, 0.995, 1$, and

[2]We also consider forecasting Euro Area quarterly output growth using the European Central Bank (ECB) survey of professional forecasters as our third application. The results of this application can be found in Section S-4 of the online empirical supplement.

(2) Heavy down-weighting with $\lambda = 0.95, 0.96, 0.97, 0.98, 0.99, 1$. For each of the above two sets of exponential down-weighting schemes we focus on simple average forecasts computed over the individual forecasts obtained for each value of λ in the set under consideration.

For forecast evaluation we consider Mean Squared Forecasting Error (MSFE) and Mean Directional Forecast Accuracy (MDFA), together with related pooled versions of Diebold-Mariano (DM), and Pesaran-Timmermann (PT) test statistics. A panel version of [6] test is proposed by [20]. Let $q_{lt} \equiv e^2_{ltA} - e^2_{ltB}$ be the difference in the squared forecasting errors of procedures A and B, for the target variable y_{lt} $(l = 1, 2, ..., L)$ and $t = 1, 2, ..., T^f_l$, where T^f_l is the number of forecasts for target variable l (could be one or multiple step ahead) under consideration. Suppose $q_{lt} = \alpha_l + \varepsilon_{lt}$ with $\varepsilon_{lt} \sim \mathcal{N}(0, \sigma^2_l)$. Then under the null hypothesis of $H_0 : \alpha_l = 0$ for all l we have

$$\overline{DM} = \frac{\bar{q}}{\sqrt{V(\bar{q})}} \overset{a}{\sim} \mathcal{N}(0,1), \text{ for } T_{Lf} \to \infty, \text{ where } T_{Lf} = \sum_{l=1}^{L} T^f_l, \bar{q} = T^{-1}_{Lf} \sum_{l=1}^{L} \sum_{t=1}^{T^f_l} q_{lt}, \text{ and}$$

$$V(\bar{q}) = \frac{1}{T^2_{Lf}} \sum_{l=1}^{L} T^f_l \hat{\sigma}^2_l, \text{ with } \hat{\sigma}^2_l = \frac{1}{T^f_l} \sum_{t=1}^{T^f_l} (q_{lt} - \bar{q}_l)^2 \text{ and } \bar{q}_l = \frac{1}{T^f_l} \sum_{t=1}^{T^f_l} q_{lt}.$$

Note that $V(\bar{q})$ needs to be modified in the case of multiple-step ahead forecast errors, due to the serial correlation that results in the forecast errors from the use of over-lapping observations. There is no adjustment needed for one-step ahead forecasting, since it is reasonable to assume that in this case the loss differentials are serially uncorrelated. However, to handle possible serial correlation for h-step ahead forecasting with $h > 1$, we can modify the panel DM test by using the Newey-West type estimator of σ^2_l.

The $MDFA$ statistic compares the accuracy of forecasts in predicting the direction (sign) of the target variable, and is computed as

$$MDFA = 100 \left\{ \frac{1}{T_{Lf}} \sum_{l=l}^{L} \sum_{t=1}^{T^f_l} \mathbf{1}[\text{sgn}(y_{lt} y^f_{lt}) > 0] \right\},$$

where $\mathbf{1}(w > 0)$ is the indicator function takes the value of 1 when $w > 0$ and zero otherwise,

sgn(w) is the sign function, y_{lt} is the actual value of dependent variable at time t and y_{lt}^f is its corresponding predicted value. To evaluate statistical significance of the directional forecasts for each method, we also report a pooled version of the test suggested by [21]:

$$PT = \frac{\hat{P} - \hat{P}^*}{\sqrt{\hat{V}(\hat{P}) - \hat{V}(\hat{P}^*)}},$$

where $\hat{P}$ is the estimator of the probability of correctly predicting the sign of y_{lt}, computed by

$$\hat{P} = \frac{1}{T_{Lf}} \sum_{l=1}^{L} \sum_{t=1}^{T_l^f} \mathbf{1}[\ \text{sgn}(y_{lt} y_{lt}^f) > 0], \text{ and } \hat{P}^* = \bar{d}_y \bar{d}_{y^f} + (1 - \bar{d}_y)(1 - \bar{d}_{y^f}), \text{ with}$$

$$\bar{d}_y = \frac{1}{T_{Lf}} \sum_{l=1}^{L} \sum_{t=1}^{T_l^f} \mathbf{1}[\text{sgn}(y_{lt}) > 0], \text{ and } \bar{d}_{y^f} = \frac{1}{T_{Lf}} \sum_{l=1}^{L} \sum_{t=1}^{T_l^f} \mathbf{1}[\text{sgn}(y_{lt}^f) > 0].$$

Finally, $\hat{V}(\hat{P}) = T_{Lf}^{-1} \hat{P}^*(1 - \hat{P}^*)$, and

$$\hat{V}(\hat{P}^*) = \frac{1}{T_{Lf}}(2\bar{d}_y - 1)^2 \bar{d}_{y^f}(1 - \bar{d}_{y^f}) + \frac{1}{T_{Lf}}(2\bar{d}_y^f - 1)^2 \bar{d}_y(1 - \bar{d}_y) + \frac{4}{T_{Lf}^2} \bar{d}_y \bar{d}_{y^f}(1 - \bar{d}_y)(1 - \bar{d}_{y^f}).$$

The last term of $\hat{V}(\hat{P}^*)$ is negligible and can be ignored. Under the null hypothesis, that prediction and realization are independently distributed, PT is asymptotically distributed as a standard normal distribution.

5.1 Forecasting monthly returns of stocks in Dow Jones

In this application the focus is on forecasting one-month ahead stock returns, defined as monthly change in natural logarithm of stock prices. We consider stocks that were part of the Dow Jones index in 2017m12, and have non-zero prices for at least 120 consecutive data points (10 years) over the period 1980m1 and 2017m12. We ended up forecasting 28 blue chip stocks.[3] Daily close prices for all the stocks are obtained from Data Stream. For stock i, the price at the last trading day of each month is used to construct the corresponding monthly

[3]Visa and DwoDuPont are excluded since they have less than 10 years of historical price data.

stock prices, P_{it}. Finally, monthly returns are computed by $r_{i,t+1} = 100 \ln(P_{i,t+1}/P_{it})$, for $i = 1, 2, ..., 28$. For all 28 stocks we use an expanding window starting with the observations for the first 10 years ($T = 120$). The active set for predicting $r_{i,t+1}$ consists of 40 financial, economic, and technical variables.[4] The full list and the description of the indicators considered can be found in Section S-1 of online empirical supplement.

Overall we computed 8,659 monthly forecasts for the 28 target stocks. The results are summarized as average forecast performances across the different variable selection procedures. Table 1 reports the effects of down-weighting at the selection stage of the OCMT procedure. It is clear that down-weighting worsens the predictive accuracy of OCMT. From the Panel DM tests, we can also see that down-weighting at the selection stage worsens the forecasts significantly. Panel DM test statistics is -5.606 (-11.352) for light (heavy) versus no down-weighing at the selection stage. Moreover, Table 2 shows that the OCMT procedure with no down-weighting at the selection stage dominates Lasso and Adaptive Lasso in terms of MSFE and the differences are statistically highly significant.

Further, OCMT outperforms Lasso and Adaptive Lasso in terms of Mean Directional Forecast Accuracy (MDFA), measured as the percent number of correctly signed one-month ahead forecasts across all the 28 stocks over the period 1990m2-2017m12. See Table 3. As can be seen from this table, OCMT with no down-weighting performs the best; correctly predicting the direction of 56.057% of 8,659 forecasts, as compared to 55.33%, which we obtain for Lasso and Adaptive Lasso forecast, at best. This difference is highly significant considering the very large number of forecasts involved. It is also of interest that the better of performance of OCMT is achieved with a much fewer number of selected covariates as compared to Lasso and Adaptive Lasso. As can be seen from the last column of Table 3, Lasso and Adaptive Lasso on average select many more covariates than OCMT (1-3 variables as compared to 0.072 for OCMT).

So far we have focussed on average performance across all the 28 stocks. Table 4 provides

[4]All regressions include the intercept as the only conditioning (pre-selected) variable.

the summary results for individual stocks, showing the relative performance of OCMT in terms of the number of stocks, using MSFE and MDFA criteria. The results show that OCMT performs better than Lasso and Adaptive Lasso in the majority of the stocks in terms of MSFE and MDFA. OCMT outperforms Lasso in 23 out of 28 stocks in terms of MSFE, under no down-weighting, and almost universally when Lasso or Adaptive Lasso are implemented with down-weighting. Similar results are obtained when we consider MDFA criteria, although the differences in performance are somewhat less pronounced. Overall, we can conclude that the better average performance of OCMT (documented in Tables 2 and 3) is not driven by a few stocks and holds more generally.

5.2 Forecasting quarterly output growth rates across 33 countries

We consider one and two years ahead predictions of output growth for 33 countries (20 advanced and 13 emerging). We use quarterly data from 1979Q2 to 2016Q4 taken from the GVAR dataset.[5] We predict $\Delta_4 y_{it} = y_{it} - y_{i,t-4}$, and $\Delta_8 y_{it} = y_{it} - y_{i,t-8}$, where y_{it}, is the log of real output for country i. We adopt the following direct forecasting equations:

$$\Delta_h y_{i,t+h} = y_{i,t+h} - y_{it} = \alpha_{ih} + \lambda_{ih}\Delta_1 y_{it} + \boldsymbol{\beta}'_{ih}\mathbf{x}_{it} + u_{iht},$$

where we consider $h = 4$ (one-year-ahead forecasts) and $h = 8$ (two-years-ahead forecasts). Given the known persistence in output growth, in addition to the intercept in the present application we also condition on the most recent lagged output growth, denoted by $\Delta_1 y_{it} = y_{it} - y_{i,t-1}$, and confine the variable selection to list of variables set out in Table S.2 in the online empirical supplement. Overall, we consider a maximum of 15 covariates in the active set covering quarterly changes in domestic variables such as real output growth, real short term interest rate, and long-short interest rate spread and quarterly change in the corresponding foreign variables.

We use expanding samples, starting with the observations on the first 15 years (60 data

[5]The GVAR dataset is available at `https://sites.google.com/site/gvarmodelling/data`.

points), and evaluate the forecasting performance of the three methods over the period 1997Q2 to 2016Q4.

Tables 5 and 6, respectively, report the MSFE of OCMT for one-year and two-year ahead forecasts of output growth, with and without down-weighting at the selection stage. Consistent with the previous two applications, down-weighting at the selection stage worsens the forecasting accuracy. Moreover, in Tables 7 and 8, we can see that OCMT (without down-weighting at the selection stage) outperforms Lasso and Adaptive Lasso in two-year ahead forecasting. In the case of one-year ahead forecasts, OCMT and Lasso are very close to each other and both outperform Adaptive Lasso. Table 9 summarizes country-specific MSFE and DM findings for OCMT relative to Lasso and Adaptive Lasso. The results show OCMT under-performs Lasso in more than half of the countries for one-year ahead horizon, but outperforms Lasso and Adaptive Lasso in more than 70 percent of the countries in the case of two-year ahead forecasts. It is worth noting that while Lasso generally outperforms OCMT in the case of one-year ahead forecasts, overall its performance is not significantly better. See Panel DM test of Table 7. On the other hand we can see from Table 8 that overall OCMT significantly outperforms Lasso in the case of the two-year ahead forecasts.

Finally in Tables 10 and 11 we reports MDFA and PT test statistics for OCMT, Lasso and Adaptive Lasso. Overall, OCMT has a slightly higher MDFA and hence predicts the direction of real output growth better than Lasso and Adaptive Lasso in most cases. The PT test statistics suggest that while all the methods perform well in forecasting the direction of one-year ahead real output growth, none of the methods considered are successful at predicting the direction of two-year ahead output growth.

It is also worth noting that as with the previous applications, OCMT selects very few variables from the active set (0.1 on average for both horizons, with the maximum number of selected variables being 2 for $h = 4$ and 8). On the other hand, Lasso on average selects 2.7 variables from the active set for $h = 4$, and 1 variable on average for $h = 8$. Maximum

number of variables selected by Lasso is 9 and 13 for $h = 4$, 8, respectively (out of possible 15). Again as to be expected, Adaptive Lasso selects a fewer number of variables as compared to Lasso (2.3 and 0.8 on average for $h = 4, 8$, respectively), but this does not lead to a better forecast performance in comparison with Lasso.

In conclusion, down-weighting at both selection and forecasting stages deteriorates OCMT's MSFE for both one–year and two-years ahead forecast horizons, as compared to down-weighting only at the forecasting stage. Moreover, light down-weighting at the forecasting stage improves forecasting performance for both horizons. Statistically significant evidence of forecasting skill is found for OCMT relative to Lasso only in the case of two-years ahead forecasts. However, it is interesting that none of the big data methods can significantly beat the simple (light down-weighted) AR(1) baseline model.

6 Conclusion

The penalized regression approach has become the de facto benchmark in the literature in the context of linear regression models without breaks. These studies (with few exceptions, including 12) do not consider the problem of variable selection when breaks are present. Recently, [3] proposed OCMT as an alternative procedure to penalized regression. One clear advantage of the OCMT procedure is the fact that the problem of variable selection is separated from the forecasting stage, which is in contrast to the penalized regression techniques where the variable selection and estimation are carried out simultaneously. Using OCMT one can decide whether to use the weighted observations at the variable selection stage or not, without preempting a different down-weighting procedure at the forecasting stage.

We have provided theoretical arguments for using the full (not down-weighted) sample at the selection stage of OCMT, and down-weighted observations (if needed) at the forecasting stage of OCMT. The benefits of the proposed method are illustrated by a number of empirical

applications to forecasting output growth and stock market returns. Our results consistently suggests that using down-weighted observations at the selection stage of OCMT deteriorate the forecasting accuracy in terms of mean square forecast error and mean directional forecast accuracy. Moreover, our results suggest that overall OCMT with no down-weighting at the selection stage outperforms penalized regression methods, i.e. Lasso and Adaptive Lasso, which are prone to over-fitting.

Table 1: Mean square forecast error (MSFE) and panel DM test of OCMT of one-month ahead monthly return forecasts across the 28 stocks in Dow Jones index between 1990m2 and 2017m12 (8659 forecasts)

	Down-weighting at†			
	Selection stage	Forecasting stage		MSFE
(M1)	no	no		61.231
	Light Down-weighting, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$			
(M2)	no	yes		61.641
(M3)	yes	yes		68.131
	Heavy Down-weighting, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$			
(M4)	no	yes		62.163
(M5)	yes	yes		86.073
	Pair-wise panel DM tests			
	Light down-weighting			Heavy down-weighting
	(M2)	(M3)		(M4) (M5)
(M1)	-1.528	-5.643	(M1)	-2.459 -11.381
(M2)	-	-5.606	(M4)	- -11.352

Notes: The active set consists of 40 covariates. The conditioning set only contains an intercept.
†For each of the two sets of exponential down-weighting (light/heavy) forecasts of the target variable are computed as the simple average of the forecasts obtained using the down-weighting coefficient, λ, in the "light" or the "heavy" down-weighting set under consideration. See footnote to Table S.3.

Table 2: Mean square forecast error (MSFE) and panel DM test of OCMT versus Lasso and Adaptive Lasso of one-month ahead monthly return forecasts across the 28 stocks in Dow Jones index between 1990m2 and 2017m12 (8659 forecasts)

	MSFE under different down-weighting scenarios		
	No down-weighting	Light down-weighting†	Heavy down-weighting‡
OCMT	61.231	61.641	62.163
Lasso	61.849	63.201	69.145
A-Lasso	63.069	65.017	72.038

	Selected pair-wise panel DM tests					
	No down-weighting		Light down-weighting		Heavy down-weighting	
	Lasso	A-Lasso	Lasso	A-Lasso	Lasso	A-Lasso
OCMT	-1.533	-4.934	-2.956	-6.025	-7.676	-10.261
Lasso	-	-4.661	-	-6.885	-	-9.569

Notes: The active set consists of 40 covariates. The conditioning set contains only the intercept.
† Light down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$.
‡ Heavy down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$.

Table 3: Mean directional forecast accuracy (MDFA) and the average number of selected variables ($\hat{k}$) of OCMT, Lasso and Adaptive Lasso of one-month ahead monthly return forecasts across the 28 stocks in Dow Jones index between 1990m2 and 2017m12 (8659 forecasts).

	Down-weighting	MDFA	$\hat{k}$
OCMT	No	56.057	0.072
	Light†	55.330	0.072
	Heavy‡	54.302	0.072
Lasso	No	55.364	1.659
	Light	54.221	2.133
	Heavy	53.205	3.794
Adaptive Lasso	No	54.648	1.312
	Light	53.840	1.623
	Heavy	52.951	2.855

Notes: The active set consists of 40 variables. The conditioning set contains an intercept.
† Light down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$.
‡ Heavy down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$.

Table 4: The number of stocks out of the 28 stocks in Dow Jones index where OCMT outperforms/underperforms Lasso, and Adaptive Lasso in terms of mean square forecast error (MSFE), panel DM test and mean directional forecast accuracy (MDFA) between 1990m2 and 2017m12 (8659 forecasts).

		MSFE			
	Down-weighting	OCMT outperforms	OCMT significantly outperforms	OCMT underperforms	OCMT significantly underperforms
Lasso	No	23	4	5	2
	Light†	25	5	3	0
	Heavy‡	26	14	2	0
A-Lasso	No	24	9	4	2
	Light	27	10	1	0
	Heavy	28	24	0	0
		MDFA			
	Down-weighting	OCMT outperforms	OCMT underperforms		
Lasso	No	14	6		
	Light	24	4		
	Heavy	17	10		
A-Lasso	No	18	4		
	Light	21	3		
	Heavy	19	7		

Notes: The active set consists of 40 variables. The conditioning set only contains an intercept.
† Light down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$.
‡ Heavy down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$.

Table 5: Mean square forecast error (MSFE) and panel DM test of OCMT of one-year ahead output growth forecasts across 33 countries over the period 1997Q2-2016Q4 (2607 forecasts)

	Down-weighting at†		MSFE ($\times 10^4$)		
	Selection stage	Forecasting stage	All	Advanced	Emerging
(M1)	no	no	11.246	7.277	17.354
Light down-weighting, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$					
(M2)	no	yes	10.836	6.913	16.871
(M3)	yes	yes	10.919	6.787	17.275
Heavy down-weighting, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$					
(M4)	no	yes	11.064	7.187	17.028
(M5)	yes	yes	11.314	6.906	18.094

	Pair-wise panel DM tests (all countries)				
	Light down-weighting			Heavy down-weighting	
	(M2)	(M3)		(M4)	(M5)
(M1)	2.394	1.662	(M1)	0.668	-0.204
(M2)	-	-0.780	(M4)	-	-1.320

Notes: There are up to 15 macro and financial variables in the active set.
†For each of the two sets of exponential down-weighting (light/heavy) forecasts of the target variable are computed as the simple average of the forecasts obtained using the down-weighting coefficient, λ, in the "light" or the "heavy" down-weighting set under consideration.

Table 6: Mean square forecast error (MSFE) and panel DM test of OCMT of two-year ahead output growth forecasts across 33 countries over the period 1997Q2-2016Q4 (2343 forecasts)

	Down-weighting at†		MSFE ($\times 10^4$)		
	Selection stage	Forecasting stage	All	Advanced	Emerging
(M1)	no	no	9.921	7.355	13.867
Light down-weighting, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$					
(M2)	no	yes	9.487	6.874	13.505
(M3)	yes	yes	9.549	6.848	13.704
Heavy down-weighting, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$					
(M4)	no	yes	9.734	7.027	13.898
(M5)	yes	yes	10.389	7.277	15.177

	Pair-wise panel DM test (all countries)				
	Light down-weighting			Heavy down-weighting	
	(M2)	(M3)		(M4)	(M5)
(M1)	3.667	2.827	(M1)	0.943	-1.664
(M2)	-	-1.009	(M4)	-	-3.498

Notes: There are up to 15 macro and financial variables in the active set.
†For each of the two sets of exponential down-weighting (light/heavy) forecasts of the target variable are computed as the simple average of the forecasts obtained using the down-weighting coefficient, λ, in the "light" or the "heavy" down-weighting set under consideration..

Table 7: Mean square forecast error (MSFE) and panel DM test of OCMT versus Lasso, and Adaptive Lasso for one-year ahead output growth forecasts across 33 countries over the period1997Q2-2016Q4 (2607 forecasts)

	MSFE under different down-weighting scenarios								
	No down-weighting			Light down-weighting†			Heavy down-weighting‡		
	All	Adv.*	Emer.**	All	Adv.	Emer.	All	Adv.	Emer.
OCMT	11.246	7.277	17.354	10.836	6.913	16.871	11.064	7.187	17.028
Lasso	11.205	6.975	17.714	10.729	6.427	17.347	11.749	7.186	18.769
A-Lasso	11.579	7.128	18.426	11.153	6.548	18.236	12.254	7.482	19.595

	Pair-wise panel DM tests (All countries)					
	No down-weighting		Light down-weighting		Heavy down-weighting	
	Lasso	A-Lasso	Lasso	A-Lasso	Lasso	A-Lasso
OCMT	0.220	-1.079	0.486	-1.007	-1.799	-2.441
Lasso	-	-2.625	-	-3.626	-	-3.157

Notes: There are up to 15 macro and financial covariates in the active set.
† Light down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$.
‡ Heavy down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$.
* Adv. stands for advanced economies.
** Emer. stands for emerging economies.

Table 8: Mean square forecast error (MSFE) and panel DM test of OCMT versus Lasso, and Adaptive Lasso of two-year ahead output growth forecasts across 33 countries over the period1997Q2-2016Q4 (2343 forecasts)

	MSFE under different down-weighting scenarios								
	No down-weighting			Light down-weighting†			Heavy down-weighting‡		
	All	Adv.*	Emer.**	All	Adv.	Emer.	All	Adv.	Emer.
OCMT	9.921	7.355	13.867	9.487	6.874	13.505	9.734	7.027	13.898
Lasso	10.151	7.583	14.103	9.662	7.099	13.605	10.202	7.428	14.469
A-Lasso	10.580	7.899	14.705	10.090	7.493	14.087	11.008	8.195	15.336

	Pair-wise panel DM tests (All countries)					
	No down-weighting		Light down-weighting		Heavy down-weighting	
	Lasso	A-Lasso	Lasso	A-Lasso	Lasso	A-Lasso
OCMT	-2.684	-4.200	-2.137	-4.015	-3.606	-4.789
Lasso	-	-5.000	-	-4.950	-	-4.969

Notes: There are up to 15 macro and financial covariates in the active set.

† Light down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$.

‡ Heavy down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$.

* Adv. stands for advanced economies. ** Emer. stands for emerging economies.

Table 9: The number of countries out of the 33 countries where OCMT outperforms/underperforms Lasso, and Adaptive Lasso in terms of mean square forecast error (MSFE) and panel DM test over the period 1997Q2 -2016Q4

	Down-weighting	OCMT outperforms	OCMT significantly outperforms	OCMT underperforms	OCMT significantly underperforms
One-year-ahead horizon ($h = 4$ quarters)					
Lasso	No	13	0	20	3
	Light†	12	1	21	3
	Heavy‡	17	1	16	3
Adaptive Lasso	No	16	1	17	2
	Light	14	2	19	2
	Heavy	19	1	14	0
Two-years-ahead horizon ($h = 8$ quarters)					
Lasso	No	24	1	9	0
	Light	25	1	8	1
	Heavy	25	1	8	0
Adaptive Lasso	No	25	2	8	0
	Light	28	3	5	1
	Heavy	30	3	3	0

Notes: There are up to 15 macro and financial covariates in the active set.

†Light down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$.

‡ Heavy down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$.

Table 10: Mean directional forecast accuracy (MDFA) and PT test of OCMT, Lasso and Adaptive Lasso for one-year ahead output growth forecasts over the period 1997Q2-2016Q4 (2607 forecasts)

	Down-	MDFA			PT tests		
	weighting	All	Advanced	Emerging	All	Advanced	Emerging
OCMT	No	87.6	87.4	88.0	8.12	7.40	3.48
	Light†	87.4	87.1	87.8	7.36	6.95	2.53
	Heavy‡	86.8	86.3	87.5	6.25	5.93	1.95
Lasso	No	87.0	86.9	87.2	9.64	9.15	3.80
	Light	87.1	87.1	87.1	8.12	8.22	2.26
	Heavy	86.0	85.8	86.4	6.24	6.43	1.40
Adaptive Lasso	No	87.3	87.3	87.2	10.80	9.91	4.75
	Light	86.5	86.6	86.4	8.25	8.36	2.48
	Heavy	85.5	85.3	85.7	6.84	6.92	1.88

Notes: There are up to 15 macro and financial variables in the active set.
† Light down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$.
‡ Heavy down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$.

Table 11: Mean directional forecast accuracy (MDFA) and PT test of OCMT, Lasso and Adaptive Lasso for two-year ahead output growth forecasts over the period 1997Q2-2016Q4 (2343 forecasts)

	Down-	MDFA			PT tests		
	weighting	All	Advanced	Emerging	All	Advanced	Emerging
OCMT	No	88.0	86.7	89.9	0.52	0.00	0.47
	Light†	87.7	86.6	89.3	1.11	0.39	0.94
	Heavy‡	87.0	85.8	88.8	0.50	0.89	0.34
Lasso	No	87.6	86.6	89.2	0.77	0.60	0.66
	Light	87.5	86.3	89.4	0.07	0.79	0.88
	Heavy	86.8	85.5	88.8	1.54	1.87	0.34
Adaptive Lasso	No	87.0	85.6	89.2	0.33	0.13	1.00
	Light	87.1	85.9	88.9	1.03	1.82	1.10
	Heavy	86.2	84.8	88.4	1.53	1.92	0.62

Notes: There are up to 15 macro and financial variables in the active set.
†Light down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$.
‡ Heavy down-weighted forecasts are computed as simple averages of forecasts obtained using the down-weighting coefficient, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$.

A Appendix A: Mathematical Derivations

This appendix provides the proofs of Theorems 1 to 3. The proofs are based on lemmas presented in the online theory supplement. Among these, Lemmas S-1.6 and S-1.7 are key. For each covariate $i = 1, 2, \cdots, N$, Lemmas S-1.6 establishes exponential probability inequalities for the t-ratio multiple tests conditional on the average net effect, $\bar{\theta}_{i,T}$, being either of the order $\ominus(T^{-\varepsilon_i})$ for some $\varepsilon_i > 1/2$, or of the order $\ominus(T^{-\vartheta_i})$, for some $0 \leq \vartheta_i < 1/2$. For DGP given by (12), Lemma S-1.7 provides asymptotic properties of LS estimator of coefficients and MSE of a regression model that includes all the signals and pseudo-signals. This lemma establishes that the coefficients of pseudo-signals estimated by LS converges to zero so long as $k_T^* = \ominus(T^d)$ grows at a slow rate relative to T, i.e. $0 \leq d < 1/2$. This lemma also shows that the MSE of the regression model converges to that of the oracle model, which includes only the signals.

Additional notations and definitions: Throughout this appendix we consider the following events:

$$\mathcal{A}_0 = \mathcal{H} \cap \mathcal{G}, \text{ where } \mathcal{H} = \left\{ \sum_{i=1}^{k} \hat{\mathcal{J}}_i = k \right\} \text{ and } \mathcal{G} = \left\{ \sum_{i=k+k_T^*+1}^{N} \hat{\mathcal{J}}_i = 0 \right\}, \tag{A.1}$$

where $\{\hat{\mathcal{J}}_i \text{ for } i = 1, 2, \cdots, N\}$ are the selection indicators defined by (9). $\mathcal{A}_0$ is the event of selecting the approximating model, defined by $\mathcal{H}$, is the event that all signals are selected, and $\mathcal{G}$ is the event that no noise variable is selected. To simplify the exposition, with slight abuse of notation, we denote the probability of an event $\mathcal{E}$ conditional on $\bar{\theta}_{i,T}$ being of order $\ominus(T^{-a})$ by $\Pr[\mathcal{E}|\bar{\theta}_{i,T} = \ominus(T^{-a})]$,where a is a nonnegative constant.

A.1 Proof of Theorem 1

To establish result (11), first note that $\mathcal{A}_0^c = \mathcal{H}^c \cup \mathcal{G}^c$ and hence ($\mathcal{H}^c$ denotes the complement of $\mathcal{H}$)

$$\Pr(\mathcal{A}_0^c) = \Pr(\mathcal{H}^c) + \Pr(\mathcal{G}^c) - \Pr(\mathcal{H}^c \cap \mathcal{G}^c) \leq \Pr(\mathcal{H}^c) + \Pr(\ \mathcal{G}^c), \tag{A.2}$$

where $\mathcal{H}$ and $\mathcal{G}$ are given by (A.1). We also have $\mathcal{H}^c = \{\sum_{i=1}^{k} \hat{\mathcal{J}}_i < k\}$ and $\mathcal{G}^c = \{\sum_{i=k+k_T^*+1}^{N} \hat{\mathcal{J}}_i > 0\}$. Let's consider $\Pr(\mathcal{H}^c)$ and $\Pr(\ \mathcal{G}^c)$ in turn. We have $\Pr(\mathcal{H}^c) \leq \sum_{i=1}^{k} \Pr(\hat{\mathcal{J}}_i = 0)$. But for any signal

$$\Pr(\hat{\mathcal{J}}_i = 0) = \Pr\left[|t_{i,T}| < c_p(N,\delta)|\bar{\theta}_{i,T} = \ominus(T^{-\vartheta_i})\right] = 1 - \Pr\left[|t_{i,T}| > c_p(N,\delta)|\bar{\theta}_{i,T} = \ominus(T^{-\vartheta_i})\right],$$

where $0 \leq \vartheta_i < 1/2$ and hence by Lemma S-1.6, we can conclude that there exist sufficiently large positive constants C_0 and C_1 such that $\Pr(\hat{\mathcal{J}}_i = 0) = O\left[\exp(-C_0 T^{C_1})\right]$. Since by Assumption 3, the number of signals is finite we can further conclude that

$$\Pr(\mathcal{H}^c) = O\left[\exp(-C_0 T^{C_1})\right] \tag{A.3}$$

for some finite positive constants C_0 and C_1. In the next step note that

$$\Pr(\mathcal{G}^c) = \Pr\left(\sum_{i=k+k_T^*+1}^{N} \hat{\mathcal{J}}_i > 0\right) \leq \sum_{i=k+k_T^*+1}^{N} \Pr\left(\hat{\mathcal{J}}_i = 1\right).$$

But for any noise variable $\Pr(\hat{\mathcal{J}}_i = 1) = \Pr\left[|t_{i,T}| > c_p(N,\delta)|\bar{\theta}_{i,T} = \ominus(T^{-\epsilon_i})\right]$, where $\epsilon_i > 1/2$ and hence by Lemma S-1.6, we can conclude that there exist sufficiently large positive constants C_0, C_1 and C_2 such that $\Pr(\hat{\mathcal{J}}_i = 1) \leq \exp\left[-C_0 c_p^2(N,\delta)\right] + \exp(-C_1 T^{C_2})$. Therefore,

$$\Pr(\mathcal{G}^c) \leq N \exp\left[-C_0 c_p^2(N,\delta)\right] + N \exp(-C_1 T^{C_2}),$$

and by result (II) of Lemma S-2.2 in online theory supplement we can further write

$$\Pr(\mathcal{G}^c) = O(N^{1-2C_0\delta}) + O\left[N \exp(-C_1 T^{C_2})\right]. \tag{A.4}$$

Using (A.3) and (A.4) in (A.2), we obtain $\Pr(\mathcal{A}_0^c) = O(N^{1-2C_0\delta}) + O\left[N \exp(-C_1 T^{C_2})\right]$ and $\Pr(\mathcal{A}_0) = 1 - O(N^{1-2C_0\delta}) - O\left[N \exp(-C_1 T^{C_2})\right]$, which completes the proof.

A.2 Proof of Theorem 2

For any $B > 0$,

$$\begin{aligned}\Pr\left(T^{\frac{1-d}{2}} \|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| > B\right) = &\Pr\left(T^{\frac{1-d}{2}} \|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| > B | \mathcal{A}_0\right) \Pr(\mathcal{A}_0) + \\ &\Pr\left(T^{\frac{1-d}{2}} \|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| > B | \mathcal{A}_0^c\right) \Pr(\mathcal{A}_0^c).\end{aligned}$$

Since $\Pr\left(T^{\frac{1-d}{2}} \|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| > B | \mathcal{A}_0^c\right)$ and $\Pr(\mathcal{A}_0)$ are less than or equal to one, we can further write,

$$\Pr\left(T^{\frac{1-d}{2}} \|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| > B\right) \leq \Pr\left(T^{\frac{1-d}{2}} \|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| > B | \mathcal{A}_0\right) + \Pr(\mathcal{A}_0^c).$$

By conditioning on $\mathcal{A}_0$ the dimension of vector $\hat{\boldsymbol{\gamma}}_T$ is at most equal to $k + k_T^*$ and by assumption $k_T^* = \ominus(T^d)$ where $0 \leq d < 1/2$. Therefore, by Lemma S-1.7 in online theory supplement, conditional on $\mathcal{A}_0$, $\|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\|$ is $O_p\left(T^{\frac{d-1}{2}}\right)$. By Theorem 1, we also have

$\lim_{T\to\infty} \Pr(\mathcal{A}_0^c) = 0$. Hence, for any $\varepsilon > 0$, there exists $B_\varepsilon > 0$ and $T_\varepsilon > 0$ such that

$$\Pr\left(T^{\frac{1-d}{2}} \|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| > B_\varepsilon | \mathcal{A}_0\right) + \Pr(\mathcal{A}_0^c) < \varepsilon \text{ for all } T > T_\varepsilon,$$

Therefore, $\Pr\left(T^{\frac{1-d}{2}} \|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| > B_\varepsilon\right) < \varepsilon$ for all $T > T_\varepsilon$, and we conclude that

$$\|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| = O_P\left(T^{\frac{d-1}{2}}\right), \tag{A.5}$$

as required. Similar lines of arguments can be used to show that if $\mathbb{E}\left(\mathbf{x}_{kk_T^*,t}\mathbf{x}'_{kk_T^*,t}\right) = \boldsymbol{\Sigma}$ is a fixed matrix, then $\|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^\diamond\| = O_P\left(T^{\frac{d-1}{2}}\right)$, which completes the proof.

A.3 Proof of Theorem 3

Let $D_T = T^{-1}\sum_{t=1}^T \hat{\eta}_t^2 - \left[\sum_{i=1}^k \sum_{j=1}^k \left(T^{-1}\sum_{t=1}^T \sigma_{ijt,x}\sigma_{ijt,\beta}\right) + \bar{\sigma}_{u,T}^2\right]$. For any $B > 0$,

$$\Pr\left(T^{\frac{1}{2}}|D_T| > B\right) = \Pr\left(T^{\frac{1}{2}}|D_T| > B|\mathcal{A}_0\right)\Pr(\mathcal{A}_0) + \Pr\left(T^{\frac{1}{2}}|D_T| > B|\mathcal{A}_0^c\right)\Pr(\mathcal{A}_0^c).$$

Since $\Pr\left(T^{\frac{1}{2}}|D_T| > B|\mathcal{A}_0^c\right)$ and $\Pr(\mathcal{A}_0)$ are less than or equal to one, we can further write,

$$\Pr\left(T^{\frac{1}{2}}|D_T| > B\right) \leq \Pr\left(T^{\frac{1}{2}}|D_T| > B|\mathcal{A}_0\right) + \Pr(\mathcal{A}_0^c).$$

By conditioning on $\mathcal{A}_0$ the number of selected covariates is at most equal to $k + k_T^*$ and by assumption $k_T^* = \ominus(T^d)$, where $0 \leq d < 1/2$. Therefore, by Lemma S-1.7 in online theory supplement, conditional on $\mathcal{A}_0$, D_T is $O_p\left(T^{-\frac{1}{2}}\right)$. By Theorem 1, we also have $\lim_{T\to\infty}\Pr(\mathcal{A}_0^c) = 0$. Hence, for any $\varepsilon > 0$, there exists $B_\varepsilon > 0$ and $T_\varepsilon > 0$ such that $\Pr\left(T^{\frac{1}{2}}|D_T| > B_\varepsilon|\mathcal{A}_0\right) + \Pr(\mathcal{A}_0^c) < \varepsilon$, for all $T > T_\varepsilon$. Therefore, $\Pr\left(T^{\frac{1}{2}}|D_T| > B_\varepsilon\right) < \varepsilon$ for all $T > T_\varepsilon$, and we conclude that

$$T^{-1}\sum_{t=1}^T \hat{\eta}_t^2 - \left[\sum_{i=1}^k \sum_{j=1}^k \left(T^{-1}\sum_{t=1}^T \sigma_{ijt,x}\sigma_{ijt,\beta}\right) + \bar{\sigma}_{u,T}^2\right] = O_p\left(T^{-\frac{1}{2}}\right),$$

as required. Following similar lines of argument we get that if $\mathbb{E}\left(\mathbf{x}_{kk_T^*,t}\mathbf{x}'_{kk_T^*,t}\right) = \boldsymbol{\Sigma}$ is a fixed matrix, then,

$$T^{-1}\sum_{t=1}^T \hat{\eta}_t^2 - \left[\sum_{i=1}^k \sum_{j=1}^k \left(T^{-1}\sum_{t=1}^T \sigma_{ijt,x}\sigma_{ijt,\beta}^*\right) + \bar{\sigma}_{u,T}^2\right] = O_p\left(T^{-\frac{1}{2}}\right),$$

which completes the proof.

References

[1] Caner, M. and K. Knight (2013). An alternative to unit root tests: Bridge estimators differentiate between nonstationary versus stationary models and select optimal lag. *Journal of Statistical Planning and Inference 143*, 691–715.

[2] Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics 86*, 221–241.

[3] Chudik, A., G. Kapetanios, and M. H. Pesaran (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica 86*, 1479–1512.

[4] Clements, M. and D. Hendry (1998). *Forecasting Economic Time Series*. Cambridge, England: Cambridge University Press.

[5] Dangl, T. and M. Halling (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics 106*, 157–181.

[6] Diebold, F. X. and R. S. Mariano (2002). Comparing predictive accuracy. *Journal of Business & economic statistics 20*, 134–144.

[7] Diebold, F. X. and M. Shin (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian Lasso and its derivatives. *International Journal of Forecasting 35*, 1679–1691.

[8] Fan, Y., J. Lv, M. Sharifvaghefi, and Y. Uematsu (2020). Ipad: stable interpretable forecasting with knockoffs inference. *Journal of the American Statistical Association 115*, 1822–1834.

[9] Hamilton, J. D. (1988). Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates. *Journal of Economic Dynamics and Control 12*(2-3), 385–423.

[10] Hyndman, R., A. B. Koehler, J. K. Ord, and R. D. Snyder (2008). *Forecasting with Exponential Smoothing : The State Space Approach*. Berlin, Germany: Springer Series in Statistics.

[11] Inoue, A., L. Jin, and B. Rossi (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics 196*, 55–67.

[12] Kapetanios, G. and F. Zikes (2018). Time-varying Lasso. *Economics Letters 169*, 1–6.

[13] Kaufman, P. (2020). *Trading Systems and Methods*. New Jersey, US: John Wiley & Sons.

[14] Koo, B., H. M. Anderson, M. H. Seo, and W. Yao (2020). High-dimensional predictive regression in the presence of cointegration. *Journal of Econometrics, forthcoming*.

[15] Koop, G. and S. Potter (2004). Forecasting in dynamic factor models using Bayesian model averaging. *The Econometrics Journal 7*, 550–565.

[16] Lee, S., M. H. Seo, and Y. Shin (2016). The Lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*, 193–210.

[17] Pesaran, M. H., D. Pettenuzzo, and A. Timmermann (2006). Forecasting time series subject to multiple structural breaks. *The Review of Economic Studies 73*, 1057–1084.

[18] Pesaran, M. H. and A. Pick (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics 29*, 307–318.

[19] Pesaran, M. H., A. Pick, and M. Pranovich (2013). Optimal forecasts in the presence of structural breaks. *Journal of Econometrics 177*, 134–152.

[20] Pesaran, M. H., T. Schuermann, and L. V. Smith (2009). Forecasting economic and financial variables with global VARs. *International journal of forecasting 25*, 642–675.

[21] Pesaran, M. H. and A. Timmermann (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics 10*, 461–465.

[22] Pesaran, M. H. and A. Timmermann (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics 137*, 134–161.

[23] Rossi, B. (2013). Advances in forecasting under instability. In *Handbook of Economic Forecasting*, Volume 2B, Chapter 21, pp. 1203–1324. Elsevier.

[24] Sharifvaghefi, M. (2021). Variable selection in high dimensional linear regression setting with high multicollinearity. *unpublished manuscript, available at: https://sites.google.com/view/mahrad/research*.

[25] Stock, J. and M. Watson (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics 14*, 11–30.

[26] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 267–288.

[27] Wilder, J. W. (1978). *New Concepts in Technical Trading Systems*. North Carolina, US: Trend Research.

[28] Williams, L. R. (1979). *How I Made One Million Dollars ... Last Year ... Trading Commodities*. Place of publication not identified: Windsor Books.

[29] Yousuf, K. and S. Ng (2019). Boosting high dimensional predictive regressions with time varying parameters. *arXiv preprint arXiv:1910.03109*.

[30] Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine learning research 7*, 2541–2563.

[31] Zheng, Z., Y. Fan, and J. Lv (2014). High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*, 627–649.