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.

---

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 ith 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, $\mathbf{a}_t = (a_{1t}, a_{2t}, \cdots, a_{mt})'$ 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 $\mathbf{z}_t$:

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

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

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

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

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

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,

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

$$= \mathbf{a}' T^{-1} \sum_{t=1}^{T} \mathbb{E}(\tilde{x}_{it} \mathbf{z}_t) + T^{-1} \sum_{t=1}^{T} \sum_{j=1}^{k} \mathbb{E}(\beta_{jt}) \mathbb{E}(\tilde{x}_{it} x_{jt}) + T^{-1} \sum_{t=1}^{T} \mathbb{E}(\tilde{x}_{it} u_t)$$

$$= \sum_{j=1}^{k} \left[T^{-1} \sum_{t=1}^{T} \mathbb{E}(\beta_{jt}) \mathbb{E}(\tilde{x}_{it} x_{jt}) \right] + T^{-1} \sum_{t=1}^{T} \mathbb{E}(\tilde{x}_{it} u_t).$$

Therefore, the average net effect can be written simply as

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

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 $\theta_i = \sum_{j=1}^{k} \mathbb{E}(\beta_j) \sigma_{ij}(\mathbf{z})$.

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 $\phi_{i,T} = \left[T^{-1} \sum_{t=1}^{T} \mathbb{E}\left(\tilde{x}_{it}^2\right) \right]^{-1} \left[T^{-1} \sum_{t=1}^{T} \mathbb{E}\left(\tilde{x}_{it}\tilde{y}_t\right) \right] = [\bar{\sigma}_{ii,T}(\mathbf{z})]^{-1} \bar{\theta}_{i,T}$, 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

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

---

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

where $\hat{\phi}_{i,T} = (\mathbf{x}'_i \mathbf{M}_{\mathbf{z}} \mathbf{x}_i)^{-1} (\mathbf{x}'_i \mathbf{M}_{\mathbf{z}} \mathbf{y})$ is the Ordinary Least Square (OLS) estimator of $\phi_{i,T}$, $\hat{\sigma}^2_i = \hat{\boldsymbol{\eta}}'_i \hat{\boldsymbol{\eta}}_i / T$, 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: $\mathcal{F}_t^q = \sigma(\mathbf{q}_t, \mathbf{q}_{t-1}, \cdots)$, $\mathcal{F}_t^a = \sigma(\mathbf{a}_t, \mathbf{a}_{t-1}, \cdots)$, $\mathcal{F}_{jt}^\beta = \sigma(\beta_{jt}, \beta_{j,t-1}, \cdots)$ 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 $\mathcal{F}_t = \mathcal{F}_t^q \cup \mathcal{F}_t^a \cup \mathcal{F}_t^\beta \cup \mathcal{F}_t^u$.

Assumption 1 (Martingale difference processes)

(a) $\mathbb{E}\left[\mathbf{q}_t \mathbf{q}_t' - \mathbb{E}(\mathbf{q}_t \mathbf{q}_t')|\mathcal{F}_{t-1}\right] = 0$ 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) $\mathbb{E}\left[\mathbf{a}_{\ell t} - \mathbb{E}(\mathbf{a}_{\ell t})|\mathcal{F}_{t-1}\right] = 0$ 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) $\sup_{\ell,t} \Pr(|\mathbf{a}_{\ell t}| > \alpha) \leq C_0 \exp(-C_1 \alpha^s)$, *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) $a_{\ell t}$, $\ell = 1, 2, \cdots, m$, are independent of $q_{jt'}, j = 1, \cdots, m + N$, and $u_{t'}$ for all t and t'.

(c) $\mathbb{E}(a_{\ell t}) = a_\ell$ 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 $z_{\ell t}$, $a_{\ell t}$, 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 $z_{\ell t}$, $a_{\ell t}$, x_{it}, β_{it} and u_t.

Assumption 2 imposes the variables $z_{\ell t}$, $a_{\ell t}$, 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 $a_{\ell t}$ 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 $\hat{\boldsymbol{\gamma}}_T$ 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 $0 \le d < \frac{1}{2}$. Consider the LS estimator of $\boldsymbol{\gamma}_T = (\gamma_1, \gamma_2, \cdots, \gamma_{\hat{k}_T})'$, 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

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

(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

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

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 $0 \le d < \frac{1}{2}$.

(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 $\sigma_{ijt,\beta} \ne 0$, 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 $q_{lt} \equiv e_{ltA}^2 - e_{ltB}^2$ be the difference in the squared forecasting errors of procedures A and B, for the target variable y_{lt} ($l = 1, 2, ..., L$) and $t = 1, 2, ..., T_l^f$, where T_l^f 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 $\varepsilon_{lt} \sim \mathcal{N}(0, \sigma_l^2)$. Then under the null hypothesis of $H_0 : \alpha_l = 0$ for all l we have

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

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

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_l^2.

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

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

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

---

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-weighting	MDFA			PT tests		
		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-weighting	MDFA			PT tests		
		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 $\{\hat{\mathcal{J}}_i$ 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 $B > 0$,

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

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

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

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,

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

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 $0 \le d < 1/2$. 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.

Online Theory Supplement to
"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

This online theory supplement has two sections. Section S-1 provides the main lemmas needed for the proofs of Theorems 1-3 in Appendix A of the paper. Section S-2 contains the complementary lemmas needed for the proofs of the main lemmas in the previous section.

Notations: Generic finite positive constants are denoted by C_i for $i = 1, 2, \cdots$ and c. They can take different values in different instances. $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_\infty$ and $\|\mathbf{A}\|_1$ denote the spectral, Frobenius, row, and column norms of matrix $\mathbf{A}$, respectively. $\lambda_i(\mathbf{A})$ denotes the ith eigenvalue of a square matrix A. $\|\mathbf{x}\|$ denotes the ℓ_2 norm of vector $\mathbf{x}$. If $\{f_n\}_{n=1}^\infty$ is any real sequence and $\{g_n\}_{n=1}^\infty$ is a sequence of positive real numbers, then $f_n = O(g_n)$, if there exists a positive constant C_0 and n_0 such that $|f_n|/g_n \leq C_0$ for all $n > n_0$. $f_n = o(g_n)$ if $f_n/g_n \to 0$ as $n \to \infty$. 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 $\sup_{n \geq n_0} (f_n/g_n) \leq C_1$. respectively. If $\{f_n\}_{n=1}^\infty$ is a sequence of random variables and $\{g_n\}_{n=1}^\infty$ is a sequence of positive real numbers, then $f_n = O_p(g_n)$, if for any $\varepsilon > 0$, there exists a positive constant B_ε and n_ε such that $\Pr(|f_n| > g_n B_\varepsilon) < \varepsilon$ for all $n > n_\varepsilon$.

S-1   Main lemmas

Lemma S-1.1 *Let y_t be a target variable generated by equation (1), $\mathbf{z}_t = (z_{1t}, z_{2t}, \cdots, z_{mt})'$ be the $m \times 1$ vector of conditioning covariates in DGP(1) and x_{it} be a covariate in the active*

set $\mathcal{S}_{Nt} = \{x_{1t}, x_{2t}, \cdots, x_{Nt}\}$. Under Assumptions 1, 3, and 4 we have

$$\mathbb{E}\left[y_t x_{it} - \mathbb{E}(y_t x_{it})|\mathcal{F}_{t-1}\right] = 0,$$

for $i = 1, 2, \cdots, N$,

$$\mathbb{E}\left[y_t z_{\ell t} - \mathbb{E}(y_t z_{\ell t})|\mathcal{F}_{t-1}\right] = 0,$$

for $\ell = 1, 2, \cdots, m$, and

$$\mathbb{E}\left[y_t^2 - \mathbb{E}(y_t^2)|\mathcal{F}_{t-1}\right] = 0.$$

Proof. Note that y_t can be written as

$$y_t = \mathbf{z}_t'\mathbf{a}_t + \mathbf{x}_{k,t}'\boldsymbol{\beta}_t + u_t = \sum_{\ell=1}^{m} a_{\ell t} z_{\ell t} + \sum_{j=1}^{k} \beta_{jt} x_{jt} + u_t,$$

where $\mathbf{x}_{k,t} = (x_{1t}, x_{2t}, \cdots, x_{kt})'$, and $\boldsymbol{\beta}_t = (\beta_{1t}, \beta_{2t}, \cdots, \beta_{kt})'$. Moreover, By Assumption 4, $a_{\ell t}$ is independent of $x_{it'}$ and $z_{\ell't'}$ for all i, ℓ', and t'. Hence, for $i = 1, 2, \cdots, N$, we have

$$\mathbb{E}(y_t x_{it}|\mathcal{F}_{t-1}) = \sum_{\ell=1}^{m} \mathbb{E}(a_{\ell t}|\mathcal{F}_{t-1})\mathbb{E}(z_{\ell t} x_{it}|\mathcal{F}_{t-1}) + \sum_{j=1}^{k} \mathbb{E}(\beta_{jt}|\mathcal{F}_{t-1})\mathbb{E}(x_{jt} x_{it}|\mathcal{F}_{t-1}) + \mathbb{E}(u_t x_{it}|\mathcal{F}_{t-1}).$$

By Assumption 1, we have $\mathbb{E}(a_{\ell t}|\mathcal{F}_{t-1}) = \mathbb{E}(a_{\ell t})$, $\mathbb{E}(z_{\ell t} x_{it}|\mathcal{F}_{t-1}) = \mathbb{E}(z_{\ell t} x_{it})$, $\mathbb{E}(\beta_{jt}|\mathcal{F}_{t-1}) = \mathbb{E}(\beta_{jt})$, $\mathbb{E}(x_{jt} x_{it}|\mathcal{F}_{t-1}) = \mathbb{E}(x_{jt} x_{it})$, and $\mathbb{E}(u_t x_{it}|\mathcal{F}_{t-1}) = \mathbb{E}(u_t x_{it})$. Therefore,

$$\mathbb{E}(y_t x_{it}|\mathcal{F}_{t-1}) = \sum_{\ell=1}^{m} \mathbb{E}(a_{\ell t})\mathbb{E}(z_{\ell t} x_{it}) + \sum_{j=1}^{k} \mathbb{E}(\beta_{jt})\mathbb{E}(x_{jt} x_{it}) + \mathbb{E}(u_t x_{it}) = \mathbb{E}(y_t x_{it}).$$

Similarly, we can show that for $\ell = 1, 2, \cdots, m$,

$$\mathbb{E}(y_t z_{\ell t}|\mathcal{F}_{t-1}) = \sum_{\ell'=1}^{m} \mathbb{E}(a_{\ell' t}|\mathcal{F}_{t-1})\mathbb{E}(z_{\ell' t} z_{\ell t}|\mathcal{F}_{t-1}) + \sum_{j=1}^{k} \mathbb{E}(\beta_{jt}|\mathcal{F}_{t-1})\mathbb{E}(x_{jt} z_{\ell t}|\mathcal{F}_{t-1}) + \mathbb{E}(u_t z_{\ell t}|\mathcal{F}_{t-1})$$
$$= \sum_{\ell'=1}^{m} \mathbb{E}(a_{\ell' t})\mathbb{E}(z_{\ell' t} z_{\ell t}) + \sum_{j=1}^{k} \mathbb{E}(\beta_{jt})\mathbb{E}(x_{jt} z_{\ell t}) + \mathbb{E}(u_t z_{\ell t}) = \mathbb{E}(y_t z_{\ell t}).$$

Also to establish the last result, we can write y_t as $y_t = \mathbf{q}_t'\boldsymbol{\delta}_t + u_t$, where $\mathbf{q}_t = (\mathbf{z}_t', \mathbf{x}_{k,t}')'$ and $\boldsymbol{\delta}_t = (\mathbf{a}_t', \boldsymbol{\beta}_t')'$. We have,

$$\mathbb{E}(y_t^2|\mathcal{F}_{t-1}) = \mathbb{E}(\boldsymbol{\delta}_t'|\mathcal{F}_{t-1})\mathbb{E}(\mathbf{q}_t\mathbf{q}_t'|\mathcal{F}_{t-1})\mathbb{E}(\boldsymbol{\delta}_t|\mathcal{F}_{t-1}) + \mathbb{E}(u_t^2|\mathcal{F}_{t-1}) + 2\mathbb{E}(\boldsymbol{\delta}_t'|\mathcal{F}_{t-1})\mathbb{E}(\mathbf{q}_t u_t|\mathcal{F}_{t-1})$$
$$= \mathbb{E}(\boldsymbol{\delta}_t')\mathbb{E}(\mathbf{q}_t\mathbf{q}_t')\mathbb{E}(\boldsymbol{\delta}_t) + \mathbb{E}(u_t^2) + 2\mathbb{E}(\boldsymbol{\delta}_t')\ \mathbb{E}(\mathbf{q}_t u_t) = \mathbb{E}(y_t^2).$$

∎

Lemma S-1.2 *Let y_t be a target variable generated by equation (1). Under Assumptions 2-4, for any value of $\alpha > 0$, there exist some positive constants C_0 and C_1 such that*

$$\sup_t \Pr(|y_t| > \alpha) \leq C_0 \exp(C_1 \alpha^{s/2})$$

Proof. Note that

$$|y_t| \le \sum_{\ell=1}^{m} |\mathsf{a}_{\ell t} z_{\ell t}| + \sum_{j=1}^{k} |\beta_{jt} x_{jt}| + |u_t|.$$

Therefore,

$$\Pr(|y_t| > \alpha) \le \Pr(\sum_{\ell=1}^{m} |\mathsf{a}_{\ell t} z_{\ell t}| + \sum_{j=1}^{k} |\beta_{jt} x_{jt}| + |u_t| > \alpha),$$

and by Lemma S-2.3 for any $0 < \pi_i < 1$, $i = 1, 2, \cdots, k+m+1$, with $\sum_{i=1}^{k+m+1} \pi_j = 1$, we can further write

$$\Pr(|y_t| > \alpha) \le \sum_{\ell=1}^{m} \Pr(|\mathsf{a}_{\ell t} z_{\ell t}| > \pi_\ell \alpha) + \sum_{j=1}^{k} \Pr(|\beta_{jt} x_{jt}| > \pi_j \alpha) + \Pr(|u_t| > \pi_{k+m+1} \alpha).$$

Moreover, by Lemma S-2.4, we have

$$\Pr(|\mathsf{a}_{\ell t} z_{\ell t}| > \pi_\ell \alpha) \le \Pr[|z_{\ell t}| > (\pi_\ell \alpha)^{1/2}] + \Pr[|\mathsf{a}_{\ell t}| > (\pi_\ell \alpha)^{1/2}],$$
$$\Pr(|\beta_{jt} x_{jt}| > \pi_j \alpha) \le \Pr[|x_{jt}| > (\pi_j \alpha)^{1/2}] + \Pr[|\beta_{jt}| > (\pi_i \alpha)^{1/2}],$$

and hence

$$\Pr(|y_t| > \alpha) \le \sum_{\ell=1}^{m} \Pr[|z_{\ell t}| > (\pi_\ell \alpha)^{1/2}] + \sum_{\ell=1}^{m} \Pr[|\mathsf{a}_{\ell t}| > (\pi_\ell \alpha)^{1/2}] +$$
$$\sum_{j=1}^{k} \Pr[|x_{jt}| > (\pi_j \alpha)^{1/2}] + \sum_{j=1}^{k} \Pr[|\beta_{jt}| > (\pi_j \alpha)^{1/2}] + \Pr(|u_t| > \pi_{k+1} \alpha),$$

Therefore, under Assumptions 2-4, we can conclude that for any value of $\alpha > 0$, there exist some positive constants C_0 and C_1 such that

$$\sup_t \Pr(|y_t| > \alpha) \le C_0 \exp(C_1 \alpha^{s/2}).$$

∎

Lemma S-1.3 *Let x_{it} be a covariate in the active set, $\mathcal{S}_{Nt} = \{x_{1t}, x_{2t}, \cdots, x_{Nt}\}$ and $\mathbf{z}_t = (z_{1t}, z_{2t}, \cdots, z_{mt})'$ be the $m \times 1$ vector of conditioning covariates in the DGP, given by (1). Define the projection regression of x_{it} on $\mathbf{z}_t$ as*

$$x_{it} = \bar{\boldsymbol{\psi}}_i' \mathbf{z}_t + \tilde{x}_{it},$$

where $\bar{\boldsymbol{\psi}}_i = (\psi_1, \cdots, \psi_m)'$ is the $m \times 1$ vector of projection coefficients which is equal to $[T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t \mathbf{z}_t')^{-1}][T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t x_{it})]$. Under Assumptions 1, 2, and 4, there exist some finite positive constants C_0, C_1 and C_2 such that if $0 < \lambda \le (s+2)/(s+4)$, then

$$\Pr(|\mathbf{x}_i' \mathbf{M}_z \mathbf{x}_j - \mathbb{E}(\tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_j)| > \zeta_T) \le \exp(-C_0 T^{-1} \zeta_T^2) + \exp(-C_1 T^{C_2})$$

and if $\lambda > (s+2)/(s+4)$, then

$$\Pr(|\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_j - \mathbb{E}(\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_j)| > \zeta_T) \leq \exp(-C_0\zeta_T^{s/(s+1)}) + \exp(-C_1 T^{C_2})$$

for all i and j, where $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \tilde{x}_{i2}, \cdots, \tilde{x}_{iT})'$, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iT})'$, and $\mathbf{M}_z = \mathbf{I} - T^{-1}\mathbf{Z}\hat{\boldsymbol{\Sigma}}_{zz}^{-1}\mathbf{Z}'$ with $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T)'$ and $\hat{\boldsymbol{\Sigma}}_{zz} = T^{-1}\sum_{t=1}^{T}(\mathbf{z}_t\mathbf{z}_t')$.

Proof. By Assumption 1 we have

$$\mathbb{E}\left[z_{\ell t}z_{\ell' t} - \mathbb{E}(z_{\ell t}z_{\ell' t})|\mathcal{F}_{t-1}\right] = 0.$$

for $\ell, \ell' = 1, 2, \cdots, m$,

$$\mathbb{E}\left[x_{it}x_{jt} - \mathbb{E}(x_{it}x_{jt})|\mathcal{F}_{t-1}\right] = 0,$$

for $i, j = 1, 2, \cdots, N$, and

$$\mathbb{E}\left[z_{\ell t}x_{it} - \mathbb{E}(z_{\ell t}x_{it})|\mathcal{F}_{t-1}\right] = 0,$$

for $\ell = 1, 2, \cdots, m$, $i = 1, 2, \cdots, N$. Moreover, by Assumption 2, for all i, ℓ, and t, x_{it}, and $z_{\ell t}$ have exponential decaying probability tails. Additionally, by Assumption 4 the number of pre-selected covariates m is finite. Therefore by Lemma S-2.20, we can conclude that there exist sufficiently large positive constants C_0, C_1, and C_2 such that if $0 < \lambda \leq (s+2)/(s+4)$,

$$\Pr(|\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_j - \mathbb{E}(\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_j)| > \zeta_T) \leq \exp(-C_0 T^{-1}\zeta_T^2) + \exp(-C_1 T^{C_2})$$

and if $\lambda > (s+2)/(s+4)$

$$\Pr(|\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_j - \mathbb{E}(\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_j)| > \zeta_T) \leq \exp(-C_0\zeta_T^{s/(s+1)}) + \exp(-C_1 T^{C_2})$$

for all i and j. ∎

Lemma S-1.4 *Let y_t be a target variable generated by the DGP given by (1), $\mathbf{z}_t = (z_{1t}, z_{2t}, \cdots, z_{mt})'$ be the $m \times 1$ vector of conditioning covariates in DGP(1) and x_{it} be a covariate in the active set, $\mathcal{S}_{Nt} = \{x_{1t}, x_{2t}, \cdots, x_{Nt}\}$. Define the projection regression of x_{it} on $\mathbf{z}_t$ as*

$$x_{it} = \mathbf{z}_t'\bar{\boldsymbol{\psi}}_{i,T} + \tilde{x}_{it},$$

where $\bar{\boldsymbol{\psi}}_{i,T} = (\psi_{1i,T}, \cdots, \psi_{mi,T})'$ is the $m \times 1$ vector of projection coefficients which is equal to $\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{z}_t\mathbf{z}_t')\right]^{-1}[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{z}_t x_{it})]$. Additionally define the projection regression of

y_t on $\mathbf{z}_t$ as

$$y_t = \mathbf{z}_t' \bar{\boldsymbol{\psi}}_{y,T} + \tilde{y}_t,$$

where $\bar{\boldsymbol{\psi}}_{y,T} = (\psi_{1y,T}, \cdots, \psi_{my,T})'$ is equal to $\left[T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t \mathbf{z}_t') \right]^{-1} [T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t y_t)]$. Under Assumptions 1- 4, if $0 < \lambda \leq (s+2)/(s+4)$,

$$\Pr(|\mathbf{x}_i' \mathbf{M}_z \mathbf{y} - \theta_{i,T}| > \zeta_T) \leq \exp(-C_0 T^{-1} \zeta_T^2) + \exp(-C_1 T^{C_2}),$$

and if $\lambda > (s+2)/(s+4)$

$$\Pr(|\mathbf{x}_i' \mathbf{M}_z \mathbf{y} - \theta_{i,T}| > \zeta_T) \leq \exp(-C_0 \zeta_T^{s/(s+1)}) + \exp(-C_1 T^{C_2}),$$

for all $i = 1, 2, \cdots, N$; where $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iT})'$, $\mathbf{y} = (y_1, y_2, \cdots, y_T)'$, $\theta_{i,T} = T\bar{\theta}_{i,T} = \mathbb{E}(\tilde{\mathbf{x}}_i' \tilde{\mathbf{y}})$, $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \tilde{x}_{i2}, \cdots, \tilde{x}_{iT})'$, $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_T)'$, $\mathbf{M}_z = \mathbf{I} - T^{-1} \mathbf{Z} \hat{\boldsymbol{\Sigma}}_{zz}^{-1} \mathbf{Z}'$, $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T)'$ and $\hat{\boldsymbol{\Sigma}}_{zz} = T^{-1} \sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t'$.

Proof. Note that by Assumption 1 and Lemma S-1.1, for all i and ℓ, cross products of x_{it}, $z_{\ell t}$ and y_t minus their expected values are martingale difference processes with respect to filtration $\mathcal{F}_{t-1}$. Moreover, by Assumption 2 and Lemma S-1.2, for all i, ℓ, and t, x_{it}, $z_{\ell t}$ and y_t have exponential decaying probability tails. Additionally, by Assumption 4 the number of pre-selected covariates m is finite. Therefore by Lemma S-2.20, we can conclude that there exist sufficiently large positive constants C_0, C_1, and C_2 such that if $0 < \lambda \leq (s+2)/(s+4)$, then

$$\Pr(|\mathbf{x}_i' \mathbf{M}_z \mathbf{y} - \theta_{i,T}| > \zeta_T) \leq \exp(-C_0 T^{-1} \zeta_T^2) + \exp(-C_1 T^{C_2}),$$

and if $\lambda > (s+2)/(s+4)$, then

$$\Pr(|\mathbf{x}_i' \mathbf{M}_z \mathbf{y} - \theta_{i,T}| > \zeta_T) \leq \exp(-C_0 \zeta_T^{s/(s+1)}) + \exp(-C_1 T^{C_2}),$$

for all $i = 1, 2, \cdots, N$. ∎

Lemma S-1.5 *Let y_t be a target variable generated by equation (1), $\mathbf{z}_t$ be the $m \times 1$ vector of conditioning covariates in DGP(1) and x_{it} be a covariate in the active set, $\mathcal{S}_{Nt} = \{x_{1t}, x_{2t}, \cdots, x_{Nt}\}$. Define the projection regression of y_t on $\mathbf{q}_t \equiv (\mathbf{z}_t', x_{it})'$ as*

$$y_t = \bar{\boldsymbol{\phi}}_{i,T}' \mathbf{q}_t + \eta_{it},$$

where $\bar{\boldsymbol{\phi}}_{i,T} \equiv \left[T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{q}_t \mathbf{q}_t') \right]^{-1} [T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{q}_t y_t)]$ is the projection coefficients. Under Assumptions 1-4, there exist sufficiently large positive constants C_0, C_1 and C_2 such that if

S.5

$0 < \lambda \leq (s+2)/(s+4)$, *then*

$$\Pr\left[|\boldsymbol{\eta}_i'\mathbf{M}_q \ \boldsymbol{\eta}_i - \mathbb{E}(\boldsymbol{\eta}_i'\boldsymbol{\eta}_i)| > \zeta_T\right] \leq \exp(-C_0 T^{-1}\zeta_T^2) + \exp(-C_1 T^{C_2}),$$

and if $\lambda > (s+2)/(s+4)$, *then*

$$\Pr\left[|\boldsymbol{\eta}_i'\mathbf{M}_q \ \boldsymbol{\eta}_i - \mathbb{E}(\boldsymbol{\eta}_i'\boldsymbol{\eta}_i)| > \zeta_T\right] \leq \exp(-C_0 \zeta_T^{s/(s+1)}) + \exp(-C_1 T^{C_2}),$$

for all $i = 1, 2, \cdots, N$; *where* $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2}, \cdots, \eta_{iT})'$, $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$, *and* $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_T)'$.

Proof. Note that $\boldsymbol{\eta}_i'\mathbf{M}_q\boldsymbol{\eta}_i = \mathbf{y}'\mathbf{M}_q\mathbf{y}$ where $\mathbf{y} = (y_1, y_2, \cdots, y_T)'$. By Lemma S-1.1 we have

$$\mathbb{E}\left[y_t x_{it} - \mathbb{E}(y_t x_{it})|\mathcal{F}_{t-1}\right] = 0,$$

for $i = 1, 2, \cdots, N$,

$$\mathbb{E}\left[y_t z_{\ell t} - \mathbb{E}(y_t z_{\ell t})|\mathcal{F}_{t-1}\right] = 0,$$

for $\ell = 1, 2, \cdots, m$, and

$$\mathbb{E}\left[y_t^2 - \mathbb{E}(y_t^2)|\mathcal{F}_{t-1}\right] = 0.$$

Moreover, by Assumption 2 and Lemma S-1.2, for all i, ℓ, and t, x_{it}, $z_{\ell t}$ and y_t have exponential decaying probability tails. Additionally, by Assumption 4 the number of pre-selected covariates m is finite. Therefore by Lemma S-2.20, we can conclude that there exist sufficiently large positive constants C_0, C_1, and C_2 such that if $0 < \lambda \leq (s+2)/(s+4)$, then

$$\Pr\left[|\boldsymbol{\eta}_i'\mathbf{M}_q \ \boldsymbol{\eta}_i - \mathbb{E}(\boldsymbol{\eta}_i'\boldsymbol{\eta}_i)| > \zeta_T\right] \leq \exp(-C_0 T^{-1}\zeta_T^2) + \exp(-C_1 T^{C_2}),$$

and if $\lambda > (s+2)/(s+4)$, then

$$\Pr\left[|\boldsymbol{\eta}_i'\mathbf{M}_q \ \boldsymbol{\eta}_i - \mathbb{E}(\boldsymbol{\eta}_i'\boldsymbol{\eta}_i)| > \zeta_T\right] \leq \exp(-C_0 \zeta_T^{s/(s+1)}) + \exp(-C_1 T^{C_2}),$$

for all $i = 1, 2, \cdots, N$. ∎

Lemma S-1.6 *Let* y_t *be a target variable generated by equation (1),* $\mathbf{z}_t$ *be the* $m \times 1$ *vector of conditioning covariates in DGP(1) and* x_{it} *be a covariate in the active set* $\mathcal{S}_{Nt} = \{x_{1t}, x_{2t}, \cdots, x_{Nt}\}$. *Define the projection regression of* x_{it} *on* $\mathbf{z}_t$ *as*

$$x_{it} = \mathbf{z}_t'\bar{\boldsymbol{\psi}}_{i,T} + \tilde{x}_{it},$$

where $\bar{\boldsymbol{\psi}}_{i,T} = (\psi_{1i,T}, \cdots, \psi_{mi,T})'$ is the $m \times 1$ vector of projection coefficients which is equal to $[T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t \mathbf{z}_t')^{-1}][T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t x_{it})]$. Additionally define the projection regression of y_t on $\mathbf{z}_t$ as

$$y_t = \mathbf{z}_t' \bar{\boldsymbol{\psi}}_{y,T} + \tilde{y}_t,$$

where $\bar{\boldsymbol{\psi}}_{y,T} = (\psi_{1y,T}, \cdots, \psi_{my,T})'$ is equal to $\left[T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t \mathbf{z}_t')\right]^{-1} [T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t y_t)]$. Lastly, define the projection regression of y_t on $\mathbf{q}_t \equiv (\mathbf{z}_t', x_{it})'$ as

$$y_t = \bar{\boldsymbol{\phi}}_{i,T}' \mathbf{q}_t + \eta_{it},$$

where $\bar{\boldsymbol{\phi}}_{i,T} \equiv \left[T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{q}_t \mathbf{q}_t')\right]^{-1} [T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{q}_t y_t)]$ is the vector of projection coefficients. Consider

$$t_{i,T} = \frac{T^{-1/2} \mathbf{x}_i' \mathbf{M}_z \mathbf{y}}{\sqrt{T^{-1} \boldsymbol{\eta}_i' \mathbf{M}_q \boldsymbol{\eta}_i} \sqrt{T^{-1} \mathbf{x}_i' \mathbf{M}_z \mathbf{x}_i}},$$

for all $i = 1, 2, \cdots, N$; where $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iT})'$, $\mathbf{y} = (y_1, y_2, \cdots, y_T)'$, $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2}, \cdots, \eta_{iT})'$, $\mathbf{M}_z = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T)'$, $\mathbf{M}_q = \mathbf{I} - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$, $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_T)'$. Under Assumptions 1-4, there exist sufficiently large positive constants C_0, C_1 and C_2 such that

$$\Pr\left[|t_{i,T}| > c_p(N, \delta)|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] \leq \exp\left[-C_0 c_p^2(N, \delta)\right] + \exp(-C_1 T^{C_2}), \text{ for } \epsilon_i > \frac{1}{2}$$

where $c_p(N, \delta)$ is defined by (8), $\theta_{i,T} = T \bar{\theta}_{i,T} = \mathbb{E}(\tilde{\mathbf{x}}_i' \tilde{\mathbf{y}})$, $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \tilde{x}_{i2}, \cdots, \tilde{x}_{iT})'$, and $\tilde{\mathbf{y}} = (\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_T)'$. Moreover, if $c_p(N, \delta) = o(T^{1/2-\vartheta-c})$ for any $0 \leq \vartheta < 1/2$ and a finite positive constant c, then there exist some finite positive constants C_0 and C_1 such that,

$$\Pr\left[|t_{i,T}| > c_p(N, \delta)|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] \geq 1 - \exp(-C_0 T^{C_1}), \text{ for } 0 \leq \vartheta_i < \tfrac{1}{2}.$$

Proof. Let $\sigma_{\eta_i}^2 = \mathbb{E}(T^{-1} \boldsymbol{\eta}_i' \boldsymbol{\eta}_i)$, and $\sigma_{\tilde{x}_i}^2 = \mathbb{E}(T^{-1} \tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_i)$. We have $|t_{i,T}| = \mathcal{A}_{iT} \mathcal{B}_{iT}$, where,

$$\mathcal{A}_{iT} = \frac{|T^{-1/2} \mathbf{x}_i \mathbf{M}_z \mathbf{y}|}{\sigma_{\eta_i} \sigma_{\tilde{x}_i}},$$

and

$$\mathcal{B}_{iT} = \frac{\sigma_{\eta_i} \sigma_{\tilde{x}_i}}{\sqrt{T^{-1} \boldsymbol{\eta}_i' \mathbf{M}_q \boldsymbol{\eta}_i} \sqrt{T^{-1} \mathbf{x}_i' \mathbf{M}_z \mathbf{x}_i}}.$$

In the first case where $\theta_{i,T} = \ominus(T^{1-\epsilon_i})$ for some $\epsilon_i > 1/2$, by using Lemma S-2.4 we have

$$\Pr\left[|t_{i,T}| > c_p(n,\delta)|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] \leq \Pr\left[\mathcal{A}_{iT} > c_p(N,\delta)/(1+d_T)|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] + \Pr\left[\mathcal{B}_{iT} > 1 + d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right],$$

where $d_T \to 0$ as $T \to \infty$. By using Lemma S-2.6,

$$\Pr\left[\mathcal{B}_{iT} > 1 + d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right]$$

$$= \Pr\left(|\frac{\sigma_{\eta_i}\sigma_{\tilde{x}_i}}{\sqrt{T^{-1}\boldsymbol{\eta}_i'\mathbf{M}_q\boldsymbol{\eta}_i}\sqrt{T^{-1}\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_i}} - 1| > d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right)$$

$$\leq \Pr\left(|\frac{(T^{-1}\boldsymbol{\eta}_i'\mathbf{M}_q\boldsymbol{\eta}_i)(T^{-1}\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_i)}{\sigma_{\eta_i}^2\sigma_{\tilde{x}_i}^2} - 1| > d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right)$$

$$= \Pr\left[\mathcal{M}_{iT} + \mathcal{R}_{iT} + \mathcal{M}_{iT}\mathcal{R}_{iT} > d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right]$$

where $\mathcal{R}_{iT} = |(T^{-1}\boldsymbol{\eta}_i'\mathbf{M}_q\boldsymbol{\eta}_i)/\sigma_{\eta_i}^2 - 1|$ and $\mathcal{M}_{iT} = |(T^{-1}\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_i)/\sigma_{\tilde{x}_i}^2 - 1|$. By using Lemmas S-2.3 and S-2.4 , for any values of $0 < \pi_i < 1$ with $\sum_{i=1}^3 \pi_i = 1$ and a strictly positive constant, c, we have

$$\Pr\left[\mathcal{B}_{iT} > 1 + d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right]$$

$$\leq \Pr\left[\mathcal{M}_{iT} > \pi_1 d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] + \Pr\left[\mathcal{R}_{iT} > \pi_2 d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] +$$

$$\Pr\left[\mathcal{M}_{iT} > \tfrac{\pi_3}{c}d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] + \Pr\left[\mathcal{R}_{iT} > c|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right].$$

First, consider $\Pr\left[\mathcal{M}_{iT} > \pi_1 d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right]$, and note that

$$\Pr\left[\mathcal{M}_{iT} > \pi_1 d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] = \Pr\left[|\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_i - \mathbb{E}(\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i)| > \pi_1\sigma_{\tilde{x}_i}^2 T d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right].$$

Therefore, by Lemma S-1.3, there exist some constants C_0 and C_1 such that,

$$\Pr\left[\mathcal{M}_{iT} > \pi_1 d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] \leq \exp(-C_0 T^{C_1}).$$

Similarly,

$$\Pr\left[\mathcal{M}_{iT} > \tfrac{\pi_3}{c}d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] \leq \exp(-C_0 T^{C_1}).$$

Also note that

$$\Pr\left[\mathcal{R}_{iT} > \pi_2 d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] = \Pr\left[|\boldsymbol{\eta}_i'\mathbf{M}_q\boldsymbol{\eta}_i - \mathbb{E}(\boldsymbol{\eta}_i'\boldsymbol{\eta}_i)| > \pi_2\sigma_{\eta_i}^2 T d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right].$$

Therefore, by Lemma S-1.5, there exist some constants C_0 and C_1 such that,

$$\Pr\left[\mathcal{R}_{iT} > \pi_2 d_T|\theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] \leq \exp(-C_0 T^{C_1}).$$

Similarly,

$$\Pr\left[\mathcal{R}_{iT} > c | \theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] \leq \exp(-C_0 T^{C_1}).$$

Therefore, we can conclude that there exist some constants C_0 and C_1 such that,

$$\Pr\left[\mathcal{B}_{iT} > 1 + d_T | \theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] \leq \exp(-C_0 T^{C_1})$$

Now consider $\Pr\left[\mathcal{A}_{iT} > c_p(N,\delta)/(1+d_T) | \theta_{i,T} = \ominus(T^{1-\epsilon_i})\right]$, which is equal to

$$\Pr\left(\frac{|\mathbf{x}_i' \mathbf{M}_z\, \mathbf{y} - \theta_{i,T} + \theta_{i,T}|}{\sigma_{\eta_i}\sigma_{\tilde{x}_i}} > T^{1/2}\frac{c_p(N,\delta)}{1+d_T} | \theta_{i,T} = \ominus(T^{1-\epsilon_i})\right)$$

$$\leq \Pr\left(|\mathbf{x}_i' \mathbf{M}_z \mathbf{y} - \theta_{i,T}| > \frac{\sigma_{\eta_i}\sigma_{\tilde{x}_i}}{1+d_T} T^{1/2} c_p(N,\delta) - |\theta_{i,T}| | \theta_{i,T} = \ominus(T^{1-\epsilon_i})\right).$$

Note that since $\epsilon_i > 1/2$ the first term on the right hand side of the inequality dominate the second one. Moreover, Since $c_p(N,\delta) = o(T^\lambda)$ for all values of $\lambda > 0$, by Lemma S-1.4, there exists a finite positive constant C_0 such that

$$\Pr\left[|\mathbf{x}_i' \mathbf{M}_z \mathbf{y}| > k_1 T^{1/2} c_p(N,\delta) | \theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] \leq \exp\left[-C_0 c_p^2(N,\delta)\right],$$

where $k_1 = \frac{\sigma_{\eta_i}\sigma_{\tilde{x}_i}}{1+d_T}$.

Given the probability upper bound for $\mathcal{A}_{iT}$ and $\mathcal{B}_{iT}$, we can conclude that there exist some finite positive constants C_0, C_1 and C_2 such that

$$\Pr\left[|t_{i,T}| > c_p(N,\delta) | \theta_{i,T} = \ominus(T^{1-\epsilon_i})\right] \leq \exp\left[-C_0 c_p^2(N,\delta)\right] + \exp(-C_1 T^{C_2}).$$

Let's consider the next case where $\theta_{i,T} = \ominus(T^{1-\vartheta_i})$ for some $0 \leq \vartheta_i < 1/2$. We know that

$$\Pr\left[|t_{i,T}| > c_p(N,\delta) | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] = 1 - \Pr\left[t_{i,T} < c_p(N,\delta) | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right].$$

By Lemma S-2.8,

$$\Pr\left[|t_{i,T}| < c_p(N,\delta) | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] \leq \Pr\left[\mathcal{A}_{iT} < \sqrt{1+d_T} c_p(N,\delta) | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] +$$

$$\Pr\left[\mathcal{B}_{iT} < 1/\sqrt{1+d_T} | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right].$$

Since $\theta_{i,T} = \ominus(T^{1-\vartheta_i})$, for some $0 \leq \vartheta_i < 1/2$ and $c_p(N,\delta) = o(T^{1/2-\vartheta-c})$, for any $0 \leq \vartheta <$

$1/2$, $|\theta_{i,T}| - \sigma_{\eta_i}\sigma_{\tilde{x}_i}[(1+d_T)T]^{1/2}c_p(N,\delta) = \ominus(T^{1-\vartheta_i}) > 0$ and by Lemma S-2.5, we have

$$\Pr\left[\mathcal{A}_{iT} < \sqrt{1+d_T}c_p(N,\delta)|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right]$$

$$= \Pr\left[\frac{|T^{-1/2}\mathbf{x}_i'\mathbf{M}_z\mathbf{y} - T^{-1/2}\theta_{i,T} + T^{-1/2}\theta_{i,T}|}{\sigma_{\eta_i}\sigma_{\tilde{x}_i}} < \sqrt{1+d_T}c_p(N,\delta)|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right]$$

$$\leq \Pr\left[|\mathbf{x}_i'\mathbf{M}_z\mathbf{y} - \theta_{i,T}| > |\theta_{i,T}| - \sigma_{\eta_i}\sigma_{\tilde{x}_i}[(1+d_T)T]^{1/2}c_p(N,\delta)|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right].$$

Therefore, by Lemma S-1.4, there exist some finite positive constants C_0 and C_1 such that,

$$\Pr\left[|\mathbf{x}_i'\mathbf{M}_z\mathbf{y} - \theta_{i,T}| > |\theta_{i,T}| - \sigma_{\eta_i}\sigma_{\tilde{x}_i}[(1+d_T)T]^{1/2}c_p(N,\delta)|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] \leq \exp(-C_0 T^{C_1}),$$

and therefore

$$\Pr\left[\mathcal{A}_{iT} < \sqrt{1+d_T}c_p(N,\delta)|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] \leq \exp(-C_0 T^{C_1}).$$

Now let consider the probability of $\mathcal{B}_{iT}$,

$$\Pr\left(\mathcal{B}_{iT} < 1/\sqrt{1+d_T}|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right)$$

$$= \Pr\left(\frac{\sigma_{\eta_i}\sigma_{\tilde{x}_i}}{\sqrt{T^{-1}\boldsymbol{\eta}_i'\mathbf{M}_q\boldsymbol{\eta}_i}\sqrt{T^{-1}\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_i}} < \frac{1}{\sqrt{1+d_T}}|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right)$$

$$= \Pr\left(\frac{(T^{-1}\boldsymbol{\eta}_i'\mathbf{M}_q\boldsymbol{\eta}_i)(T^{-1}\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_i)}{\sigma_{\eta_i}^2\sigma_{\tilde{x}_i}^2} > 1 + d_T|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right)$$

$$\leq \Pr(\mathcal{M}_{iT} + \mathcal{R}_{iT} + \mathcal{M}_{iT}\mathcal{R}_{iT} > d_T|\theta_{i,T} = \ominus(T^{1-\vartheta_i})),$$

where $\mathcal{R}_{iT} = |(T^{-1}\boldsymbol{\eta}_i'\mathbf{M}_q\boldsymbol{\eta}_i)/\sigma_{\eta_i}^2 - 1|$ and $\mathcal{M}_{iT} = |(T^{-1}\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_i)/\sigma_{\tilde{x}_i}^2 - 1|$. By using Lemmas S-2.3 and S-2.4 , for any values of $0 < \pi_i < 1$ with $\sum_{i=1}^3 \pi_i = 1$ and a positive constant, c, we have

$$\Pr\left[\mathcal{B}_{iT} < 1/\sqrt{1+d_T}|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right]$$

$$\leq \Pr\left[\mathcal{M}_{iT} > \pi_1 d_T|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] + \Pr\left[\mathcal{R}_{iT} > \pi_2 d_T|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] +$$

$$\Pr\left[\mathcal{M}_{iT} > \frac{\pi_3}{c}d_T|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] + \Pr\left[\mathcal{R}_{iT} > c|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right].$$

Let's first consider the $\Pr\left[\mathcal{M}_{iT} > \pi_1 d_T|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right]$. Note that

$$\Pr\left[\mathcal{M}_{iT} > \pi_1 d_T|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] = \Pr\left[|\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_i - \mathbb{E}(\boldsymbol{\nu}_i'\boldsymbol{\nu}_i)| > \pi_1 \sigma_{\tilde{x}_i}^2 T d_T|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right].$$

So, by Lemma S-1.3, we know that there exist some constants C_0 and C_1 such that,

$$\Pr\left[\mathcal{M}_{iT} > \pi_1 d_T|\theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] \leq \exp(-C_0 T^{C_1}).$$

Similarly,

$$\Pr\left[\mathcal{M}_{iT} > \tfrac{\pi_3}{c} d_T | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] \le \exp(-C_0 T^{C_1}).$$

Also note that

$$\Pr\left[\mathcal{R}_{iT} > \pi_2 d_T | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] = \Pr\left[|\boldsymbol{\eta}_i' \mathbf{M}_q \boldsymbol{\eta}_i - \mathbb{E}(\boldsymbol{\eta}_i' \boldsymbol{\eta}_i)| > \pi_2 \sigma_{\eta_i}^2 T d_T | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right].$$

Therefore, by Lemma S-1.5, there exist some constants C_0 and C_1 such that,

$$\Pr(\mathcal{R}_{iT} > \pi_2 d_T | \theta_{i,T} \ne 0) \le \exp(-C_0 T^{C_1}).$$

Similarly,

$$\Pr(\mathcal{R}_{iT} > c | \theta_{i,T} \ne 0) \le \exp(-C_0 T^{C_1}).$$

Therefore, we can conclude that there exist some constants C_0 and C_1 such that,

$$\Pr\left[\mathcal{B}_{iT} < 1/\sqrt{1 + d_T} | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] \le \exp(-C_0 T^{C_1}).$$

So, overall we conclude that

$$\Pr\left[|t_{i,T}| > c_p(N, \delta) | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right]$$
$$= 1 - \Pr\left[t_{i,T} < c_p(N, \delta) | \theta_{i,T} = \ominus(T^{1-\vartheta_i})\right] \ge 1 - \exp(-C_0 T^{C_1}).$$

∎

Lemma S-1.7 *Consider the following data generating process (DGP) for y_t:*

$$y_t = \sum_{i=1}^{k} x_{it} \beta_{it} + u_t \text{ for } t = 1, 2, \cdots, T. \tag{S.1}$$

Estimate the following regression

$$y_t = \sum_{i=1}^{k} x_{it} \phi_i + \sum_{j=1}^{l_T} x_{k+j,t} \delta_j + \eta_t = \mathbf{q}_t' \boldsymbol{\phi} + \mathbf{s}_t' \boldsymbol{\delta}_T + \eta_t, \tag{S.2}$$

by least squares (LS), where $\mathbf{q}_t = (x_{1t}, x_{2t}, \cdots, x_{kt})'$, $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_k)'$, $\mathbf{s}_t = (x_{k+1,t}, x_{k+2,t}, \cdots, x_{k+l_T,t})'$, and $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_{l_T})'$. The LS estimator of $\boldsymbol{\gamma}_T = (\boldsymbol{\phi}', \boldsymbol{\delta}_T')'$ is

$$\hat{\boldsymbol{\gamma}}_T = \left(T^{-1} \mathbf{W}' \mathbf{W}\right)^{-1} \left(T^{-1} \mathbf{W}' \mathbf{y}\right), \tag{S.3}$$

where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T)'$, $\mathbf{w}_t = (\mathbf{q}_t', \mathbf{s}_t')'$ *and* $\mathbf{y} = (y_1, y_2, \cdots, y_T)'$. *The model error is*

$$\hat{\boldsymbol{\eta}} = \mathbf{y} - \mathbf{W}\hat{\boldsymbol{\gamma}}_T. \tag{S.4}$$

Suppose that $\lambda_{\min}[T^{-1}\mathbb{E}(\mathbf{W}'\mathbf{W})] > c > 0$, *and* $l_T = \ominus(T^d)$, *where* $0 \leq d < \frac{1}{2}$. *Moreover suppose that Assumptions 1-4 holds. Now,*

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

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

where $\boldsymbol{\gamma}_T^* = (\boldsymbol{\beta}', \mathbf{0}_{l_T}')'$ *and* $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_k)'$. *If Assumption 6 also holds, then*

$$T^{-1}\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(T^{-1}\sum_{t=1}^{T}\sigma_{ijt,x}\sigma_{ijt,\beta}\right) + \bar{\sigma}_{u,T}^2 + O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\frac{l_T}{T}\right), \tag{S.6}$$

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

(ii) *If* $\mathbb{E}(\mathbf{w}_t\mathbf{w}_t')$ *is time invariant, then*

$$\|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^\diamond\| = O_p\left(T^{\frac{d-1}{2}}\right), \tag{S.7}$$

where $\boldsymbol{\gamma}_T^\diamond = (\bar{\boldsymbol{\beta}}_T', \mathbf{0}_{l_T}')'$, $\bar{\boldsymbol{\beta}}_T = (\bar{\beta}_{1T}, \bar{\beta}_{2T}, \cdots, \bar{\beta}_{kT})'$, *and* $\bar{\beta}_{iT} = T^{-1}\sum_{t=1}^{T}\mathbb{E}(\beta_{it})$. *If Assumption 6 also holds, then*

$$T^{-1}\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(T^{-1}\sum_{t=1}^{T}\sigma_{ijt,x}\sigma_{ijt,\beta}^*\right) + \bar{\sigma}_{u,T}^2 + O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\frac{l_T}{T}\right), \tag{S.8}$$

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

Proof. In the first scenario, where $\mathbb{E}(\beta_{it}) = \beta_i$ for all t, we can write (S.1) as

$$y_t = \sum_{i=1}^{k}x_{it}\beta_i + \sum_{i=1}^{k}x_{it}(\beta_{it} - \beta_i) + u_t = \sum_{i=1}^{k}x_{it}\beta_i + \sum_{i=1}^{k}r_{it} + u_t = \mathbf{q}_t'\boldsymbol{\beta} + \mathbf{r}_t'\boldsymbol{\tau} + u_t,$$

where $r_{it} = x_{it}(\beta_{it} - \beta_i)$, $\mathbf{r}_t = (r_{1t}, r_{2t}, \cdots, r_{kt})'$, and $\boldsymbol{\tau}$ is a $k \times 1$ vector of ones. We can further write the DGP in a following matrix format,

$$\mathbf{y} = \mathbf{Q}\boldsymbol{\beta} + \mathbf{R}\boldsymbol{\tau} + \mathbf{u}, \tag{S.9}$$

where $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_T)'$, $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_T)'$ and $\mathbf{u} = (u_1, u_2, \cdots, u_T)'$. By substitut-

ing (S.9) into (S.3), we obtain

$$\hat{\boldsymbol{\gamma}}_T = \left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\mathbf{Q}\boldsymbol{\beta}\right)+\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right)+\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\mathbf{u}\right),$$
(S.10)

where $\mathbf{W} = (\mathbf{Q}, \mathbf{S})$, and $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_T)'$. Since $\boldsymbol{\gamma}_T^* = (\boldsymbol{\beta}', \mathbf{0}'_{l_T})'$, $\mathbf{Q}\boldsymbol{\beta} = \mathbf{Q}\boldsymbol{\beta} + \mathbf{S}\mathbf{0}_{l_T} = \mathbf{W}\boldsymbol{\gamma}_T^*$, which in turn allows us to write the above result as:

$$\hat{\boldsymbol{\gamma}}_T = \left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\ \boldsymbol{\gamma}_T^*+\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right)+\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\mathbf{u}\right),$$

and hence

$$\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^* = \left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right) + \left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\mathbf{u}\right).$$
(S.11)

We can further write

$$
\begin{aligned}
\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^* = &\left\{\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\}\left\{T^{-1}\left[(\mathbf{W}'\mathbf{R}\boldsymbol{\tau}) - \mathbb{E}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right)\right]\right\} + \\
&\left\{\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\}\left[T^{-1}\mathbb{E}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right)\right] + \\
&\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\left\{T^{-1}\left[(\mathbf{W}'\mathbf{R}\ \boldsymbol{\tau}) - \mathbb{E}\left(\mathbf{W}'\mathbf{R}\ \boldsymbol{\tau}\right)\right]\right\} + \\
&\left\{\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\}\left\{T^{-1}\left[(\mathbf{W}'\mathbf{u}) - \mathbb{E}\left(\mathbf{W}'\mathbf{u}\right)\right]\right\} + \\
&\left\{\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\}\left[T^{-1}\mathbb{E}\left(\mathbf{W}'\mathbf{u}\right)\right] + \\
&\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\left\{T^{-1}\left[(\mathbf{W}'\mathbf{u}) - \mathbb{E}\left(\mathbf{W}'\mathbf{u}\right)\right]\right\}.
\end{aligned}
$$

Hence, by the sub-additive property of norms and Lemma S-2.9, we have

$$
\begin{aligned}
\|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| \leq &\left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F \left\|T^{-1}\left[(\ \mathbf{W}'\mathbf{R}\boldsymbol{\tau}) - \mathbb{E}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right)\right]\right\| + \\
&\left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F \left\|T^{-1}\mathbb{E}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right)\right\| + \\
&\left\|\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_2 \left\|T^{-1}\left[(\ \mathbf{W}'\mathbf{R}\boldsymbol{\tau}) - \mathbb{E}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right)\right]\right\| + \\
&\left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F \left\|T^{-1}\left[(\mathbf{W}'\mathbf{u}) - \mathbb{E}\left(\mathbf{W}'\mathbf{u}\right)\right]\right\|_F + \\
&\left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F \left\|T^{-1}\mathbb{E}\left(\mathbf{W}'\mathbf{u}\right)\right\| + \\
&\left\|\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_2 \left\|T^{-1}\left[(\ \mathbf{W}'\mathbf{u}) - \mathbb{E}\left(\mathbf{W}'\ \mathbf{u}\right)\right]\right\|.
\end{aligned}
$$

Since, by Assumption 3, β_{it} for $i = 1, 2, \cdots, k$ are distributed independently of $\mathbf{w}_t$ for $t = 1, 2, \cdots, T$,

$$T^{-1}\mathbb{E}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right) = \sum_{i=1}^{k}\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{w}_t r_{it})\right] = \sum_{i=1}^{k}\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{w}_t x_{it}(\beta_{it} - \beta_i))\right]$$

$$= \sum_{i=1}^{k}\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{w}_t x_{it})\mathbb{E}(\beta_{it} - \beta_i)\right] = \mathbf{0}.$$

Also,

$$T^{-1}\mathbb{E}\left(\mathbf{W}'\mathbf{u}\right) = T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{w}_t u_t) = \mathbf{0}.$$

Hence,

$$\|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\| \leq \left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F \|T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\| +$$

$$\left\|\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_2 \|T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\| +$$

$$\left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F \|T^{-1}\mathbf{W}'\mathbf{u}\| +$$

$$\left\|\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_2 \|T^{-1}\mathbf{W}'\mathbf{u}\|.$$

Since Assumptions 1 and 2 imply that $\mathbf{W}$, and $\mathbf{u}$ satisfy condition (i) and (ii) of Lemma S-2.12, by Lemmas S-2.12 and S-2.13, we have

$$\|T^{-1}\mathbf{W}'\mathbf{u}\| = O_p\left(\sqrt{\frac{l_T}{T}}\right).$$

Similarly,

$$\left\|T^{-1}\left[(\mathbf{W}'\mathbf{W}) - \mathbb{E}\left(\mathbf{W}'\mathbf{W}\right)\right]\right\|_F = O_p\left(\frac{l_T}{\sqrt{T}}\right),$$

and since $l_T = \ominus(T^d)$ with $0 \leq d < 1/2$, by Lemma S-2.14,

$$\left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F = O_p\left(\frac{l_T}{\sqrt{T}}\right).$$

Now consider $\|T^{-1}\mathbf{W}'\mathbf{R}\ \boldsymbol{\tau}\|$. Note that the row j and column i of $l_T \times p$ matrix $T^{-1}\mathbf{W}'\mathbf{R}$ is equal to $T^{-1}\sum_{t=1}^{T}w_{jt}r_{it}$. Hence the jth element of $l_T \times 1$ vector $T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau}$ is equal $T^{-1}\sum_{i=1}^{k}\sum_{t=1}^{T}w_{jt}r_{it}$. In other words, $T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau} = T^{-1}\sum_{i=1}^{k}\sum_{t=1}^{T}\mathbf{w}_t r_{it}$. Therefore, (re-

calling that $r_{it} = x_{it} (\beta_{it} - \beta_i))$

$$\left\|T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right\|^2 = \left\|T^{-1}\sum_{i=1}^{k}\sum_{t=1}^{T}(\mathbf{w}_t r_{it})\right\|^2 \leq \sum_{i=1}^{k}\left\|T^{-1}\sum_{t=1}^{T}\mathbf{w}_t x_{it}(\beta_{it} - \beta_i)\right\|^2$$

$$=T^{-2}\sum_{i=1}^{k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\mathbf{w}_t'\mathbf{w}_{t'}x_{it}x_{it'}(\beta_{it} - \beta_i)(\beta_{it'} - \beta_i)$$

$$=T^{-2}\sum_{i=1}^{k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{\ell=1}^{k+l_T}w_{\ell t}w_{\ell t'}x_{it}x_{it'}(\beta_{it} - \beta_i)(\beta_{it'} - \beta_i).$$

Since, by Assumption 1, β_{it} for $i = 1, 2, \cdots, k$ are distributed independently of $\mathbf{w}_t$ for $t = 1, 2, \cdots, T$, we can further write,

$$\mathbb{E}\left\|T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right\|^2 \leq T^{-2}\sum_{i=1}^{k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{\ell=1}^{k+\ell_T}\mathbb{E}\left(w_{\ell t}w_{\ell t'}x_{it}x_{it'}\right)\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{it'} - \beta_i)\right]$$

$$\leq T^{-2}\sum_{i=1}^{k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{\ell=1}^{k+\ell_T}\left|\mathbb{E}\left(w_{\ell t}w_{\ell t'}x_{it}x_{it'}\right)\right| \times \left|\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{it'} - \beta_i)\right]\right|$$

$$\leq T^{-2}(k + \ell_T)\,sup_{i,\ell,t,t'}\left|\mathbb{E}\left(w_{\ell t}w_{\ell t'}x_{it}x_{it'}\right)\right|\sum_{i=1}^{k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\left|\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{it'} - \beta_i)\right]\right|$$

Since $\mathbf{W}$ satisfy condition (i) of Lemma S-2.12, we have $sup_{i,\ell,t,t'}\left|\mathbb{E}(w_{\ell t}w_{\ell t'}x_{it}x_{it'})\right| < C < \infty$. Also, note that for any $t' < t$,

$$\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{it'} - \beta_i)\right] = \mathbb{E}\left[(\beta_{it'} - \beta_i)\,\mathbb{E}\left(\beta_{it} - \beta_i|\mathcal{F}_{t-1}\right)\right],$$

and by Assumption 1, $\mathbb{E}\left(\beta_{it} - \beta_i|\mathcal{F}_{t-1}\right) = 0$. Therefore,

$$\sum_{t=1}^{T}\sum_{t'=1}^{T}\left|\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{it'} - \beta_i)\right]\right| = \sum_{t=1}^{T}\left|\mathbb{E}\left[(\beta_{it} - \beta_i)^2\right]\right| + 2\sum_{t=2}^{T}\sum_{t'=1}^{t}\left|\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{it'} - \beta_i)\right]\right|$$

$$= \sum_{t=1}^{T}\left|\mathbb{E}\left[(\beta_{it} - \beta_i)^2\right]\right| = O(T).$$

Since, by Assumption 3, k is also a finite fixed integer, we conclude that

$$\mathbb{E}\left\|T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau}\right\|^2 = O\left(\frac{l_T}{T}\right),$$

and hence, by Lemma S-2.13,

$$\left\| T^{-1}\mathbf{W}'\mathbf{R}\boldsymbol{\tau} \right\| = O_p\left(\sqrt{\frac{l_T}{T}} \right).$$

So, we can conclude that

$$\left\| \hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^* \right\| = O_p\left(\sqrt{\frac{l_T}{T}} \right),$$

as required.

In the next step, consider the mean square error of the model, $T^{-1}\hat{\boldsymbol{\eta}}_T'\hat{\boldsymbol{\eta}}_T$. By substituting y from (S.9) into equation (S.4) for the model error, we have

$$\hat{\boldsymbol{\eta}} = \mathbf{y} - \mathbf{W}\hat{\boldsymbol{\gamma}}_T = \mathbf{Q}\boldsymbol{\beta} + \mathbf{R}\boldsymbol{\tau} + \mathbf{u} - \mathbf{W}\hat{\boldsymbol{\gamma}}_T.$$

Since $\mathbf{Q}\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma}_T^*$, where $\boldsymbol{\gamma}_T^* = (\boldsymbol{\beta}', \mathbf{0}_{l_T}')'$, we can further write,

$$\hat{\boldsymbol{\eta}} = \mathbf{R}\boldsymbol{\tau} + \mathbf{u} - \mathbf{W}\left(\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\right).$$

Therefore,

$$
\begin{aligned}
T^{-1}\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} &= T^{-1}\left[\mathbf{R}\boldsymbol{\tau} + \mathbf{u} - \mathbf{W}\left(\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\right)\right]'\left[\mathbf{R}\boldsymbol{\tau} + \mathbf{u} - \mathbf{W}\left(\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\right)\right] \\
&= T^{-1}\left(\mathbf{R}\boldsymbol{\tau} + \mathbf{u}\right)'\left(\mathbf{R}\boldsymbol{\tau} + \mathbf{u}\right) + T^{-1}\left[\mathbf{W}\left(\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\right)\right]'\left[\mathbf{W}\left(\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\right)\right] - \\
&\quad 2T^{-1}\left[\mathbf{W}\left(\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\right)\right]'\left(\mathbf{R}\boldsymbol{\tau} + \mathbf{u}\right) \\
&= T^{-1}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau} + \mathbf{u}'\mathbf{u}\right) + 2T^{-1}\boldsymbol{\tau}'\mathbf{R}'\mathbf{u} + \left(\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\right)'\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\left(\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\right) - \\
&\quad 2\left(\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*\right)'\left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right].
\end{aligned}
$$

By substituting for $\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^*$ from (S.11), we get

$$
\begin{aligned}
T^{-1}\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} &= T^{-1}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau} + \mathbf{u}'\mathbf{u}\right) + 2T^{-1}\boldsymbol{\tau}'\mathbf{R}'\mathbf{u} + \\
&\quad \left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right]'\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right] - \\
&\quad 2\left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right]'\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right] \\
&= T^{-1}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau} + \mathbf{u}'\mathbf{u}\right) + 2T^{-1}\boldsymbol{\tau}'\mathbf{R}'\mathbf{u} - \\
&\quad \left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right]'\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right].
\end{aligned}
$$

we can further write

$$T^{-1}\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = T^{-1}\mathbb{E}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau} + \mathbf{u}'\mathbf{u}\right) + T^{-1}\left\{\left[\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\ \boldsymbol{\tau} - \mathbb{E}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau}\right)\right] + \left[\mathbf{u}'\mathbf{u} - \mathbb{E}\left(\mathbf{u}'\mathbf{u}\right)\right]\right\} +$$
$$2T^{-1}\boldsymbol{\tau}'\mathbf{R}'\mathbf{u} - \left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right]'\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right] -$$
$$\left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right]'\left\{\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\}\left[T^{-1}\left(\mathbf{W}'\mathbf{R}\boldsymbol{\tau} + \mathbf{W}'\mathbf{u}\right)\right].$$

Therefore,

$$T^{-1}\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} - T^{-1}\mathbb{E}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau} + \mathbf{u}'\mathbf{u}\right) \leq$$
$$T^{-1}\left[\boldsymbol{\tau}'\mathbf{R}'\ \mathbf{R}\boldsymbol{\tau} - \mathbb{E}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau}\right)\right] + T^{-1}\left[\mathbf{u}'\mathbf{u} - \mathbb{E}\left(\mathbf{u}'\mathbf{u}\right)\right] +$$
$$2T^{-1}\boldsymbol{\tau}'\mathbf{R}'\mathbf{u} + \left\|T^{-1}\mathbf{W}'\left(\mathbf{R}\boldsymbol{\tau} + \mathbf{u}\right)\right\|^2 \left\|\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_2 + \tag{S.12}$$
$$\left\|T^{-1}\mathbf{W}'\left(\mathbf{R}\boldsymbol{\tau} + \mathbf{u}\right)\right\|^2 \left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F.$$

First, consider $T^{-1}\left[\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau} - \mathbb{E}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau}\right)\right]$. Note that

$$\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau} = \boldsymbol{\tau}'\left(\sum_{t=1}^{T}\mathbf{r}_t\mathbf{r}_t'\right)\boldsymbol{\tau} = \sum_{t=1}^{T}\left(\ \boldsymbol{\tau}'\mathbf{r}_t\right)\left(\mathbf{r}_t'\boldsymbol{\tau}\right) = \sum_{t=1}^{T}\left(\sum_{i=1}^{k}r_{it}\right)\left(\sum_{j=1}^{k}r_{jt}\right) = \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{t=1}^{T}r_{it}r_{jt}.$$

Recalling that $r_{it} = x_{it}(\beta_{it} - \beta_i)$, and hence,

$$T^{-1}\left[\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau} - \mathbb{E}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau}\right)\right] = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(T^{-1}\sum_{t=1}^{T}a_{ij,t}\right),$$

where

$$a_{ij,t} = x_{it}x_{jt}(\beta_{it} - \beta_i)(\beta_{jt} - \beta_j) - \mathbb{E}\left(x_{it}x_{jt}\right)\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{jt} - \beta_j)\right].$$

Now consider $\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}a_{ij,t}\right)^2$ and note that

$$\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}a_{ij,t}\right)^2 = T^{-2}\sum_{t=1}^{T}\mathbb{E}\left(a_{ij,t}^2\right) + 2T^{-2}\sum_{t=2}^{T}\sum_{t'=1}^{t}\mathbb{E}\left(a_{ij,t}a_{ij,t'}\right)$$
$$= T^{-2}\sum_{t=1}^{T}\mathbb{E}\left(a_{ij,t}^2\right) + 2T^{-2}\sum_{t=2}^{T}\sum_{t'=1}^{t}\mathbb{E}\left[a_{ij,t'}\mathbb{E}\left(a_{ij,t}|\mathcal{F}_{t-1}\right)\right].$$

But, by Assumptions 1, 3, and 6,

$$\mathbb{E}\left(a_{ij,t}|\mathcal{F}_{t-1}\right) = \mathbb{E}\left(x_{it}x_{jt}|\mathcal{F}_{t-1}\right)\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{jt} - \beta_j)|\mathcal{F}_{t-1}\right] - \mathbb{E}\left(x_{it}x_{jt}\right)\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{jt} - \beta_j)\right]$$
$$= \mathbb{E}\left(x_{it}x_{jt}\right)\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{jt} - \beta_j)\right] - \mathbb{E}\left(x_{it}x_{jt}\right)\mathbb{E}\left[(\beta_{it} - \beta_i)(\beta_{jt} - \beta_j)\right] = 0.$$

Therefore,

$$\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}a_{ij,t}\right)^{2} = T^{-2}\sum_{t=1}^{T}\mathbb{E}\left(a_{ij,t}^{2}\right) = O\left(\frac{1}{T}\right),$$

and by Lemma S-2.13 we conclude that

$$\left|T^{-1}\sum_{t=1}^{T}a_{ij,t}\right| = O_{p}\left(\frac{1}{\sqrt{T}}\right).$$

Since by Assumption 3, k is a finite fixed integer, we can further conclude that

$$T^{-1}\left[\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau} - \mathbb{E}\left(\boldsymbol{\tau}'\mathbf{R}'\mathbf{R}\boldsymbol{\tau}\right)\right] = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(T^{-1}\sum_{t=1}^{T}a_{ij,t}\right) = O_{p}\left(\frac{1}{\sqrt{T}}\right). \tag{S.13}$$

Now, consider, $T^{-1}\boldsymbol{\tau}'\mathbf{R}'\mathbf{u}$. Note that

$$T^{-1}\boldsymbol{\tau}'\mathbf{R}'\mathbf{u} = T^{-1}\boldsymbol{\tau}'\left(\sum_{t=1}^{T}\mathbf{r}_{t}u_{t}\right) = T^{-1}\sum_{t=1}^{T}\boldsymbol{\tau}'\mathbf{r}_{t}u_{t} = T^{-1}\sum_{t=1}^{T}\sum_{i=1}^{k}r_{it}u_{t} = \sum_{i=1}^{k}\left(T^{-1}\sum_{t=1}^{T}r_{it}u_{t}\right).$$

We have

$$\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}r_{it}u_{t}\right)^{2} = T^{-2}\sum_{t=1}^{T}\mathbb{E}\left(r_{it}^{2}u_{t}^{2}\right) + 2T^{-2}\sum_{t=2}^{T}\sum_{t'=1}^{t}\mathbb{E}\left(r_{it}r_{it'}u_{t}u_{t'}\right).$$

Since $r_{it} = x_{it}(\beta_{it} - \beta_{i})$, and β_{it} for $i = 1, 2, \cdots, k$ are distributed independently of x_{js}, $j = 1, 2, \cdots, N$, and u_{s} for all t and s, we can further write for any $t' < t$

$$\mathbb{E}\left(r_{it}r_{it'}u_{t}u_{t'}\right) = \mathbb{E}\left(x_{it}u_{t}x_{it'}u_{t'}\right)\mathbb{E}\left[(\beta_{it} - \beta_{i})(\beta_{it'} - \beta_{i})\right]$$
$$= \mathbb{E}\left(x_{it}u_{t}x_{it'}u_{t'}\right)\mathbb{E}\left\{(\beta_{it'} - \beta_{i})\mathbb{E}\left[(\beta_{it} - \beta_{i})|\mathcal{F}_{t-1}\right]\right\}.$$

But, by Assumption 1, $\mathbb{E}\left[(\beta_{it} - \beta_{i})|\ \mathcal{F}_{t-1}\right] = 0$ and thus $\mathbb{E}\left(r_{it}r_{it'}u_{t}u_{t'}\right) = 0$ for any $t' < t$. Therefore,

$$\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}r_{it}u_{t}\right)^{2} = T^{-2}\sum_{t=1}^{T}\mathbb{E}\left(r_{it}^{2}u_{t}^{2}\right) = O\left(\frac{1}{T}\right).$$

Hence, by Lemma S-2.13, $\left|T^{-1}\sum_{t=1}^{T}r_{it}u_{t}\right| = O_{p}\left(\frac{1}{\sqrt{T}}\right)$. Since, by Assumption 3, k is a finite fixed integer, we conclude that

$$T^{-1}\boldsymbol{\tau}'\mathbf{R}'\mathbf{u} = \sum_{i=1}^{k}\left(T^{-1}\sum_{t=1}^{T}r_{it}u_{t}\right) = O_{p}\left(\frac{1}{\sqrt{T}}\right). \tag{S.14}$$

By substituting (S.13) and (S.14) into (S.12), and noting that $\left\| T^{-1}\mathbf{W}'\left(\mathbf{R}\boldsymbol{\tau}+\mathbf{u}\right)\right\|^2 = O_p(l_T/T)$, $\left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}-\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F = O_p(l_T/\sqrt{T})$, and $T^{-1}\left[\mathbf{u}'\mathbf{u}-\mathbb{E}\left(\mathbf{u}'\mathbf{u}\right)\right] = O_p(1/\sqrt{T})$, we conclude that

$$T^{-1}\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(T^{-1}\sum_{t=1}^{T}\sigma_{ijt,x}\sigma_{ijt,\beta}\right) + \bar{\sigma}_{u,T}^2 + O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\frac{l_T}{T}\right),$$

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 $\bar{\sigma}_{u,T}^2 = T^{-1}\mathbb{E}\left(\mathbf{u}'\mathbf{u}\right)$.

In the second scenario, where $\mathbb{E}\left(\mathbf{w}_t\mathbf{w}_t'\right)$ is time invariant, we can write (S.1) as

$$y_t = \sum_{i=1}^{k}x_{it}\bar{\beta}_{iT} + \sum_{i=1}^{k}x_{it}\left(\beta_{it}-\bar{\beta}_{iT}\right) + u_t = \sum_{i=1}^{k}x_{it}\bar{\beta}_{iT} + \sum_{i=1}^{k}h_{it} + u_t = \mathbf{q}_t'\bar{\boldsymbol{\beta}} + \mathbf{h}_t'\boldsymbol{\tau} + u_t,$$

where $h_{it} = x_{it}\left(\beta_{it}-\bar{\beta}_{iT}\right)$, and $\mathbf{h}_t = (h_{1t}, h_{2t}, \cdots, h_{kt})'$. We can further write the DGP in a following matrix format,

$$\mathbf{y} = \mathbf{Q}\bar{\boldsymbol{\beta}} + \mathbf{H}\,\boldsymbol{\tau} + \mathbf{u},$$

where $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_T)'$. Now, by using the similar lines of arguments as in the first scenario, we get

$$\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^{\diamond} = \left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\mathbf{H}\boldsymbol{\tau}\right) + \left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(T^{-1}\mathbf{W}'\,\mathbf{u}\right).$$

Notice that

$$\begin{aligned}
T^{-1}\mathbb{E}\left(\mathbf{W}'\mathbf{H}\boldsymbol{\tau}\right) &= \sum_{i=1}^{k}\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{w}_t h_{it})\right] = \sum_{i=1}^{k}\left\{T^{-1}\sum_{t=1}^{T}\mathbb{E}[\mathbf{w}_t x_{it}(\beta_{it}-\bar{\beta}_{iT})]\right\} \\
&= \sum_{i=1}^{k}\left[T^{-1}\sum_{t=1}^{T}\mathbb{E}(\mathbf{w}_t x_{it})\mathbb{E}(\beta_{it}-\bar{\beta}_{iT})\right] \\
&= \sum_{i=1}^{k}\left[\mathbb{E}(\mathbf{w}_t x_{it})T^{-1}\sum_{t=1}^{T}\mathbb{E}(\beta_{it}-\bar{\beta}_{iT})\right] = \mathbf{0}.
\end{aligned}$$

Hence, we can further use the similar lines of arguments as in the first scenario and conclude

that

$$\left\|\hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^\diamond\right\| \leq \left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F \left\|T^{-1}\mathbf{W}'\mathbf{H}\boldsymbol{\tau}\right\| +$$
$$\left\|\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_2 \left\|T^{-1}\mathbf{W}'\mathbf{H}\boldsymbol{\tau}\right\| +$$
$$\left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F \left\|T^{-1}\mathbf{W}'\mathbf{u}\right\| +$$
$$\left\|\left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_2 \left\|T^{-1}\mathbf{W}'\mathbf{u}\right\|.$$

We know that

$$\left\|T^{-1}\mathbf{W}'\mathbf{u}\right\| = O_p\left(\sqrt{\frac{l_T}{T}}\right),$$

and

$$\left\|\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1} - \left[\mathbb{E}\left(T^{-1}\mathbf{W}'\mathbf{W}\right)\right]^{-1}\right\|_F = O_p\left(\frac{l_T}{\sqrt{T}}\right).$$

Now consider $\left\|T^{-1}\mathbf{W}'\mathbf{H}\,\boldsymbol{\tau}\right\|$. By using the similar lines of arguments as in the first scenario, we have

$$\left\|T^{-1}\mathbf{W}'\mathbf{H}\boldsymbol{\tau}\right\|^2 \leq T^{-2}\sum_{i=1}^{k}\sum_{\ell=1}^{k+l_T}\sum_{t=1}^{T}\sum_{t'=1}^{T} w_{\ell t}w_{\ell t'}x_{it}x_{it'}\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right).$$

Since, by Assumption 3, β_{it} for $i = 1, 2, \cdots, k$ are distributed independently of $\mathbf{w}_t$ for $t = 1, 2, \cdots, T$, we can further write,

$$\mathbb{E}\left\|T^{-1}\mathbf{W}'\mathbf{H}\boldsymbol{\tau}\right\|^2 \leq T^{-2}\sum_{i=1}^{k}\sum_{\ell=1}^{k+l_T}\sum_{t=1}^{T}\sum_{t'=1}^{T}\mathbb{E}\left(w_{\ell t}w_{\ell t'}x_{it}x_{it'}\right)\mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right]$$
$$= T^{-2}\sum_{i=1}^{k}\sum_{\ell=1}^{k+l_T}\sum_{t=1}^{T}\mathbb{E}\left(w_{\ell t}^2 x_{it}^2\right)\mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)^2\right] +$$
$$T^{-2}\sum_{i=1}^{k}\sum_{\ell=1}^{k+l_T}\sum_{t=1}^{T}\sum_{t'\neq t}\mathbb{E}\left(w_{\ell t}w_{\ell t'}x_{it}x_{it'}\right)\mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right].$$

Since, by Assumption 1, $\mathbb{E}\left[w_{\ell t}w_{\ell' t} - \mathbb{E}(w_{\ell t}w_{\ell' t})|\mathcal{F}_{t-1}\right] = 0$ for all ℓ, ℓ' and $t = 1, 2, \cdots, T$, we have for any $t' \neq t$

$$\mathbb{E}\left(w_{\ell t}w_{\ell t'}x_{it}x_{it'}\right) = \mathbb{E}\left(w_{\ell t}x_{it}\right)\mathbb{E}\left(w_{\ell t'}x_{it'}\right).$$

Therefore,

$$\sum_{t=1}^{T} \sum_{t' \neq t} \mathbb{E}\left(w_{\ell t} w_{\ell t'} x_{it} x_{it'}\right) \mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right]$$

$$= \sum_{t=}^{T} \sum_{t' \neq t} \mathbb{E}\left(w_{\ell t} x_{it}\right) \mathbb{E}\left(w_{\ell t'} x_{it'}\right) \mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right].$$

Since $\mathbb{E}\left(\mathbf{w}_t \mathbf{w}_t'\right)$ is time invariant, we can further write

$$\sum_{t=1}^{T} \sum_{t' \neq t} \mathbb{E}\left(w_{\ell t} w_{\ell t'} x_{it} x_{it'}\right) \mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right]$$

$$= \mathbb{E}\left(w_{\ell t} x_{it}\right)^2 \sum_{t=1}^{T} \sum_{t' \neq t} \mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right].$$

Note that, by Assumption 1, for any $t' \neq t$, $\mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right] = \left[\mathbb{E}\left(\beta_{it}\right) - \bar{\beta}_i\right]\left[\mathbb{E}\left(\beta_{it'}\right) - \bar{\beta}_i\right]$.
Therefore

$$\sum_{t=1}^{T} \sum_{t' \neq t} \mathbb{E}\left(w_{\ell t} w_{\ell t'} x_{it} x_{it'}\right) \mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right]$$

$$= \left[\mathbb{E}\left(w_{\ell t} x_{it}\right)\right]^2 \sum_{t=1}^{T} \sum_{t' \neq t} \left[\mathbb{E}\left(\beta_{it}\right) - \bar{\beta}_i\right]\left[\mathbb{E}\left(\beta_{it'}\right) - \bar{\beta}_i\right].$$

We can further write,

$$\sum_{t=1}^{T} \sum_{t' \neq t} \mathbb{E}\left(w_{\ell t} w_{\ell t'} x_{it} x_{it'}\right) \mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right]$$

$$= \left[\mathbb{E}\left(w_{\ell t} x_{it}\right)\right]^2 \left\{\sum_{t=1}^{T} \sum_{t'=1}^{T} \left[\mathbb{E}\left(\beta_{it}\right) - \bar{\beta}_i\right]\left[\mathbb{E}\left(\beta_{it'}\right) - \bar{\beta}_i\right] - \sum_{t=1}^{T} \left[\mathbb{E}\left(\beta_{it}\right) - \bar{\beta}_i\right]^2\right\}$$

$$= \left[\mathbb{E}\left(w_{\ell t} x_{it}\right)\right]^2 \left\{\sum_{t=1}^{T} \left[\mathbb{E}\left(\beta_{it}\right) - \bar{\beta}_i\right]\right\} \left\{\sum_{t'=1}^{T} \left[\mathbb{E}\left(\beta_{it'}\right) - \bar{\beta}_i\right]\right\} -$$

$$\left[\mathbb{E}\left(w_{\ell t} x_{it}\right)\right]^2 \sum_{t=1}^{T} \left[\mathbb{E}\left(\beta_{it}\right) - \bar{\beta}_i\right]^2.$$

But, $\sum_{t=1}^{T} \left[\mathbb{E}\left(\beta_{it}\right) - \bar{\beta}_i\right] = 0$, and therefore,

$$\sum_{t=1}^{T} \sum_{t' \neq t} \mathbb{E}\left(w_{\ell t} w_{\ell t'} x_{it} x_{it'}\right) \mathbb{E}\left[\left(\beta_{it} - \bar{\beta}_i\right)\left(\beta_{it'} - \bar{\beta}_i\right)\right] = - \left[\mathbb{E}\left(w_{\ell t} x_{it}\right)\right]^2 \sum_{t=1}^{T} \left[\mathbb{E}\left(\beta_{it}\right) - \bar{\beta}_i\right]^2.$$

So,

$$\mathbb{E} \left\| T^{-1} \mathbf{W}' \mathbf{H} \boldsymbol{\tau} \right\|^2$$

$$\leq T^{-2} \sum_{i=1}^{p} \sum_{\ell=1}^{p+l_T} \sum_{t=1}^{T} \left\{ \mathbb{E} \left(w_{\ell t}^2 x_{it}^2 \right) \mathbb{E} \left[\left(\beta_{it} - \bar{\beta}_i \right)^2 \right] - \left[\mathbb{E} \left(w_{\ell t} x_{it} \right) \right]^2 \left[\mathbb{E} \left(\beta_{it} \right) - \bar{\beta}_i \right]^2 \right\}$$

$$= O \left(\frac{l_T}{T} \right),$$

and hence, by Lemma S-2.13,

$$\left\| T^{-1} \mathbf{W}' \mathbf{H} \boldsymbol{\tau} \right\| = O_p \left(\sqrt{\frac{l_T}{T}} \right).$$

So, we conclude that

$$\left\| \hat{\boldsymbol{\gamma}}_T - \boldsymbol{\gamma}_T^{\diamond} \right\| = O_p \left(\sqrt{\frac{l_T}{T}} \right).$$

Lastly, consider the model mean square error for the second scenario. Following the same lines of argument as in the first scenario, we can write,

$$T^{-1} \hat{\boldsymbol{\eta}}' \hat{\boldsymbol{\eta}} - T^{-1} \mathbb{E} \left(\boldsymbol{\tau}' \mathbf{H}' \mathbf{H} \boldsymbol{\tau} + \mathbf{u}' \mathbf{u} \right) \leq$$

$$T^{-1} \left[\boldsymbol{\tau}' \mathbf{H}' \ \mathbf{H} \boldsymbol{\tau} - \mathbb{E} \left(\boldsymbol{\tau}' \mathbf{H}' \mathbf{H} \boldsymbol{\tau} \right) \right] + T^{-1} \left[\mathbf{u}' \mathbf{u} - \mathbb{E} \left(\mathbf{u}' \mathbf{u} \right) \right] +$$

$$2 T^{-1} \boldsymbol{\tau}' \mathbf{H}' \mathbf{u} + \left\| T^{-1} \mathbf{W}' \left(\mathbf{H} \boldsymbol{\tau} + \mathbf{u} \right) \right\|^2 \left\| \left[\mathbb{E} \left(T^{-1} \mathbf{W}' \mathbf{W} \right) \right]^{-1} \right\|_2 + \tag{S.15}$$

$$\left\| T^{-1} \mathbf{W}' \left(\mathbf{H} \boldsymbol{\tau} + \mathbf{u} \right) \right\|^2 \left\| \left(T^{-1} \mathbf{W}' \mathbf{W} \right)^{-1} - \left[\mathbb{E} \left(T^{-1} \mathbf{W}' \mathbf{W} \right) \right]^{-1} \right\|_F.$$

First, consider $T^{-1} \left[\boldsymbol{\tau}' \mathbf{H}' \mathbf{H} \boldsymbol{\tau} - \mathbb{E} \left(\boldsymbol{\tau}' \mathbf{H}' \mathbf{H} \boldsymbol{\tau} \right) \right]$. Note that

$$\boldsymbol{\tau}' \mathbf{H}' \mathbf{H} \boldsymbol{\tau} = \boldsymbol{\tau}' \left(\sum_{t=1}^{T} \mathbf{h}_t \mathbf{h}_t' \right) \boldsymbol{\tau} = \sum_{t=1}^{T} \left(\boldsymbol{\tau}' \mathbf{r}_t \right) \left(\mathbf{r}_t' \boldsymbol{\tau} \right) = \sum_{t=1}^{T} \left(\sum_{i=1}^{k} h_{it} \right) \left(\sum_{j=1}^{k} h_{jt} \right) = \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{t=1}^{T} h_{it} h_{jt}.$$

Recalling that $h_{it} = x_{it} (\beta_{it} - \bar{\beta}_{iT})$, and hence,

$$T^{-1} \left[\boldsymbol{\tau}' \mathbf{H}' \mathbf{H} \boldsymbol{\tau} - \mathbb{E} \left(\boldsymbol{\tau}' \mathbf{H}' \mathbf{H} \boldsymbol{\tau} \right) \right] = \sum_{i=1}^{k} \sum_{j=1}^{k} \left(T^{-1} \sum_{t=1}^{T} b_{ij,t} \right),$$

where

$$b_{ij,t} = x_{it} x_{jt} (\beta_{it} - \bar{\beta}_{iT}) (\beta_{jt} - \bar{\beta}_{jT}) - \mathbb{E} \left(x_{it} x_{jt} \right) \mathbb{E} \left[(\beta_{it} - \bar{\beta}_{iT}) (\beta_{jt} - \bar{\beta}_{jT}) \right].$$

Now consider $\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}b_{ij,t}\right)^2$ and note that

$$\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}b_{ij,t}\right)^2 = T^{-2}\sum_{t=1}^{T}\mathbb{E}\left(b_{ij,t}^2\right) + 2T^{-2}\sum_{t=2}^{T}\sum_{t'=1}^{t}\mathbb{E}\left(b_{ij,t}b_{ij,t'}\right)$$

$$= T^{-2}\sum_{t=1}^{T}\mathbb{E}\left(b_{ij,t}^2\right) + 2T^{-2}\sum_{t=2}^{T}\sum_{t'=1}^{t}\mathbb{E}\left[b_{ij,t'}\mathbb{E}\left(b_{ij,t}|\mathcal{F}_{t-1}\right)\right].$$

But, by Assumptions 1, 3, and 6,

$$\mathbb{E}\left(b_{ij,t}|\mathcal{F}_{t-1}\right) = \mathbb{E}\left(x_{it}x_{jt}|\mathcal{F}_{t-1}\right)\mathbb{E}\left[(\beta_{it}-\bar{\beta}_{iT})(\beta_{jt}-\bar{\beta}_{jT})|\mathcal{F}_{t-1}\right] - \mathbb{E}\left(x_{it}x_{jt}\right)\mathbb{E}\left[(\beta_{it}-\bar{\beta}_{iT})(\beta_{jt}-\bar{\beta}_{jT})\right]$$

$$= \mathbb{E}\left(x_{it}x_{jt}\right)\mathbb{E}\left[(\beta_{it}-\bar{\beta}_{iT})(\beta_{jt}-\bar{\beta}_{jT})\right] - \mathbb{E}\left(x_{it}x_{jt}\right)\mathbb{E}\left[(\beta_{it}-\bar{\beta}_{iT})(\beta_{jt}-\bar{\beta}_{jT})\right] = 0.$$

Therefore,

$$\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}b_{ij,t}\right)^2 = T^{-2}\sum_{t=1}^{T}\mathbb{E}\left(b_{ij,t}^2\right) = O\left(\frac{1}{T}\right),$$

and by Lemma S-2.13 we conclude that

$$\left|T^{-1}\sum_{t=1}^{T}b_{ij,t}\right| = O_p\left(\frac{1}{\sqrt{T}}\right).$$

Since by Assumption 3, k is a finite fixed integer, we can further conclude that

$$T^{-1}\left[\boldsymbol{\tau}'\mathbf{H}'\mathbf{H}\boldsymbol{\tau} - \mathbb{E}\left(\boldsymbol{\tau}'\mathbf{H}'\mathbf{H}\boldsymbol{\tau}\right)\right] = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(T^{-1}\sum_{t=1}^{T}b_{ij,t}\right) = O_p\left(\frac{1}{\sqrt{T}}\right). \qquad (S.16)$$

Now, consider, $T^{-1}\boldsymbol{\tau}'\mathbf{H}'\mathbf{u}$. Note that

$$T^{-1}\boldsymbol{\tau}'\mathbf{H}'\mathbf{u} = T^{-1}\boldsymbol{\tau}'\left(\sum_{t=1}^{T}\mathbf{h}_t u_t\right) = T^{-1}\sum_{t=1}^{T}\boldsymbol{\tau}'\mathbf{h}_t u_t = T^{-1}\sum_{t=1}^{T}\sum_{i=1}^{k}h_{it}u_t = \sum_{i=1}^{k}\left(T^{-1}\sum_{t=1}^{T}h_{it}u_t\right).$$

We have

$$\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}h_{it}u_t\right)^2 = T^{-2}\sum_{t=1}^{T}\mathbb{E}\left[(h_{it}u_t)^2\right] + T^{-2}\sum_{t=1}^{T}\sum_{t'\neq t}\mathbb{E}\left(h_{it}h_{it'}u_t u_{t'}\right).$$

Since $h_{it} = x_{it}(\beta_{it}-\bar{\beta}_{iT})$, and β_{it} for $i = 1, 2, \cdots, k$ are distributed independently of x_{js},

$j = 1, 2, \cdots, N$, and u_s for all t and s, we can further write for any $t' \neq t$

$$\mathbb{E}\left(h_{it}h_{it'}u_tu_{t'}\right) = \mathbb{E}\left(x_{it}u_tx_{it'}u_{t'}\right)\mathbb{E}\left[(\beta_{it} - \bar{\beta}_{iT})(\beta_{it'} - \bar{\beta}_{iT})\right]$$

But, by Assumption 1, $\mathbb{E}\left[x_{it}u_t - \mathbb{E}(x_{it}u_t)|\mathcal{F}_{t-1}\right] = 0$ and we also have $\mathbb{E}(x_{it}u_t) = 0$ for $i = 1, 2, \cdots, k$ and thus for any $t' \neq t$ we have

$$\mathbb{E}\left(x_{it}u_tx_{it'}u_{t'}\right) = \mathbb{E}\left(x_{it}u_t\right)\mathbb{E}\left(x_{it'}u_{t'}\right) = 0.$$

Therefore,

$$\mathbb{E}\left(T^{-1}\sum_{t=1}^{T}h_{it}u_t\right)^2 = T^{-2}\sum_{t=1}^{T}\mathbb{E}\left[(h_{it}u_t)^2\right] = O\left(\frac{1}{T}\right).$$

Hence, by Lemma S-2.13, $\left|T^{-1}\sum_{t=1}^{T}h_{it}u_t\right| = O_p\left(\frac{1}{\sqrt{T}}\right)$. Since, by Assumption 3, k is a finite fixed integer, we conclude that

$$T^{-1}\boldsymbol{\tau}'\mathbf{H}'\mathbf{u} = \sum_{i=1}^{k}\left(T^{-1}\sum_{t=1}^{T}h_{it}u_t\right) = O_p\left(\frac{1}{\sqrt{T}}\right). \tag{S.17}$$

By substituting (S.16) and (S.17) into (S.15), and noting that $\|T^{-1}\mathbf{W}'(\mathbf{H}\boldsymbol{\tau} + \mathbf{u})\|^2 = O_p(l_T/T)$, $\left\|(T^{-1}\mathbf{W}'\mathbf{W})^{-1} - [\mathbb{E}(T^{-1}\mathbf{W}'\mathbf{W})]^{-1}\right\|_F = O_p(l_T/\sqrt{T})$, and $T^{-1}[\mathbf{u}'\mathbf{u} - \mathbb{E}(\mathbf{u}'\mathbf{u})] = O_p(1/\sqrt{T})$, we conclude that

$$T^{-1}\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(T^{-1}\sum_{t=1}^{T}\sigma_{ijt,x}\sigma_{ijt,\beta}^*\right) + \bar{\sigma}_{u,T}^2 + O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\frac{l_T}{T}\right),$$

where $\sigma_{ijt,\beta}^* = \mathbb{E}\left[(\beta_{it} - \bar{\beta}_{i,T})(\beta_{jt} - \bar{\beta}_{j,T})\right]$, $\bar{\beta}_{iT} = T^{-1}\sum_{t=1}^{T}\mathbb{E}(\beta_{it})$, and $\bar{\sigma}_{u,T}^2 = T^{-1}\mathbb{E}(\mathbf{u}'\mathbf{u})$. ∎

S-2 Complementary lemmas

Lemma S-2.1 *Let z_t be a martingale difference process with respect to $\mathcal{F}_{t-1}^z = \sigma(z_{t-1}, z_{t-2}, \cdots)$, and suppose that there exist some finite positive constants C_0 and C_1, and $s > 0$ such that*

$$\sup_t \Pr(|z_t| > \alpha) \leq C_0 \exp(-C_1\alpha^s), \quad \text{for all } \alpha > 0.$$

Let also $\sigma_{zt}^2 = \mathbb{E}(z_t^2|\mathcal{F}_{t-1}^z)$ and $\bar{\sigma}_{z,T}^2 = T^{-1}\sum_{t=1}^T \sigma_{zt}^2$. Suppose that $\zeta_T = \ominus(T^\lambda)$, for some $0 < \lambda \leq (s+1)/(s+2)$. Then for any π in the range $0 < \pi < 1$, we have,

$$\Pr\left(|\textstyle\sum_{t=1}^T z_t| > \zeta_T\right) \leq \exp\left[\frac{-(1-\pi)^2\zeta_T^2}{2T\bar{\sigma}_{z,T}^2}\right].$$

if $\lambda > (s+1)/(s+2)$, then for some finite positive constant C_2,

$$\Pr\left(|\textstyle\sum_{t=1}^T z_t| > \zeta_T\right) \leq \exp\left(-C_2\zeta_T^{s/(s+1)}\right).$$

Proof. The results follow from Lemma A3 of Chudik et al. (2018) Online Theory Supplement. ∎

Lemma S-2.2 *Let*

$$c_p(n, \delta) = \Phi^{-1}\left(1 - \frac{p}{2f(n, \delta)}\right), \tag{S.18}$$

where $\Phi^{-1}(.)$ is the inverse of standard normal distribution function, p $(0 < p < 1)$ is the nominal size of a test, and $f(n, \delta) = cn^\delta$ for some positive constants δ and c. Moreover, let $a > 0$ and $0 < b < 1$. Then (I) $c_p(n, \delta) = O\left[\sqrt{\delta\ln(n)}\right]$ and (II) $n^a\exp\left[-bc_p^2(n, \delta)\right] = \ominus(n^{a-2b\delta})$.

Proof. The results follow from Lemma 3 of Bailey et al. (2019) Supplementary Appendix A. ∎

Lemma S-2.3 *Let x_i, for $i = 1, 2, \cdots, n$, be random variables. Then for any constants π_i, for $i = 1, 2, \cdots, n$, satisfying $0 < \pi_i < 1$ and $\sum_{i=1}^n \pi_i = 1$, we have*

$$\Pr(\textstyle\sum_{i=1}^n |x_i| > C_0) \leq \sum_{i=1}^n \Pr(|x_i| > \pi_i C_0),$$

where C_0 is a finite positive constant.

Proof. The result follows from Lemma A11 of Chudik et al. (2018) Online Theory Supplement. ∎

Lemma S-2.4 *Let x, y and z be random variables. Then for any finite positive constants C_0, C_1, and C_2, we have*

$$\Pr(|x| \times |y| > C_0) \leq \Pr(|x| > C_0/C_1) + \Pr(|y| > C_1),$$

and

$$\Pr(|x| \times |y| \times |z| > C_0) \leq \Pr(|x| > C_0/(C_1 C_2)) + \Pr(|y| > C_1) + \Pr(|z| > C_2).$$

Proof. The results follow from Lemma A11 of Chudik et al. (2018) Online Theory Supplement. ∎

Lemma S-2.5 *Let x be a random variable. Then for some finite constants B, and C, with $|B| \geq C > 0$, we have*

$$\Pr(|x + B| \leq C) \leq \Pr(|x| > |B| - C).$$

Proof. The results follow from Lemma A12 of Chudik et al. (2018) Online Theory Supplement. ∎

Lemma S-2.6 *Let x_T to be a random variable. Then for a deterministic sequence, $\alpha_T > 0$, with $\alpha_T \to 0$ as $T \to \infty$, there exists $T_0 > 0$ such that for all $T > T_0$ we have*

$$\Pr\left(\left|\frac{1}{\sqrt{x_T}} - 1\right| > \alpha_T\right) \leq \Pr(|x_T - 1| < \alpha_T).$$

Proof. The results follow from Lemma A13 of Chudik et al. (2018) Online Theory Supplement. ∎

Lemma S-2.7 *Consider random variables x_t and z_t with the exponentially bounded probability tail distributions such that*

$$\sup_t \Pr(|x_t| > \alpha) \leq C_0 \exp(-C_1 \alpha^{s_x}), \text{ for all } \alpha > 0,$$

$$\sup_t \Pr(|z_t| > \alpha) \leq C_0 \exp(-C_1 \alpha^{s_z}), \text{ for all } \alpha > 0,$$

where C_0, and C_1 are some finite positive constants, $s_x > 0$, and $s_z > 0$. Then

$$\sup_t \Pr(|x_t z_t| > \alpha) \leq C_0 \exp(-C_1 \alpha^{s/2}), \text{ for all } \alpha > 0,$$

where $s = \min\{s_x, s_z\}$.

Proof. By using Lemma S-2.4, for all $\alpha > 0$,

$$\Pr(|x_t z_t| > \alpha) \leq \Pr(|x_t| > \alpha^{1/2}) + \Pr(|z_t| > \alpha^{1/2})$$

So,

$$\sup_t \Pr(|x_t z_t| > \alpha) \leq \sup_t \Pr(|x_t| > \alpha^{1/2}) + \sup_t \Pr(|z_t| > \alpha^{1/2})$$

$$\leq C_0 \exp(-C_1 \alpha^{s_x/2}) + C_0 \exp(-C_1 \alpha^{s_z/2})$$

$$\leq C_0 \exp(-C_1 \alpha^{s/2})$$

where $s = \min\{s_x, s_z\}$. ∎

Lemma S-2.8 *Let x, y and z be random variables. Then for some finite positive constants C_0, and C_1, we have*

$$\Pr(|x| \times |y| < C_0) \leq \Pr(|x| < C_0/C_1) + \Pr(|y| < C_1),$$

Proof. Define events $\mathfrak{A} = \{|x| \times |y| < C_0\}$, $\mathfrak{B} = \{|x| < C_0/C_1\}$ and $\mathfrak{C} = \{|y| < C_1\}$. Then $\mathfrak{A} \in \mathfrak{B} \cup \mathfrak{C}$. Therefore, $\Pr(\mathfrak{A}) \leq \Pr(\mathfrak{B} \cup \mathfrak{C})$. But $\Pr(\mathfrak{B} \cup \mathfrak{C}) \leq \Pr(\mathfrak{B}) + \Pr(\mathfrak{C})$ and hence $\Pr(\mathfrak{A}) \leq \Pr(\mathfrak{B}) + \Pr(\mathfrak{C})$. ∎

Lemma S-2.9 *Let $\mathbf{A}$ and $\mathbf{B}$ be $n \times p$ and $p \times m$ matrices respectively, then*

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2, \text{ and } \|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_F. \tag{S.19}$$

Proof. $\|\mathbf{AB}\|_F^2 = \mathrm{tr}(\mathbf{ABB'A'}) = \mathrm{tr}[\mathbf{A}(\mathbf{BB'})\mathbf{A'}]$, and by result (12) of Lü tkepohl (1996, p.44),

$$\mathrm{tr}\,[\mathbf{A}(\mathbf{BB'})\mathbf{A'}] \leq \lambda_{\max}(\mathbf{BB'})\mathrm{tr}(\mathbf{AA'}) = \|\mathbf{A}\|_F^2 \|\,\mathbf{B}\|_2^2,$$

where $\lambda_{\max}(\mathbf{BB'})$ is the largest eigenvalue of $\mathbf{BB'}$. Therefore, $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$, as required. Similarly,

$$\|\mathbf{AB}\|_F^2 = \mathrm{tr}(\mathbf{B'A'AB}) = \mathrm{tr}[\mathbf{B'}(\mathbf{A'A})\mathbf{B}] \leq \lambda_{\max}(\mathbf{A'A})\mathrm{tr}(\mathbf{B'B}) = \|\,\mathbf{A}\|_2^2 \|\mathbf{B}\|_F^2,$$

and hence

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_F.$$

∎

Lemma S-2.10 *Let $\mathbf{A} = (a_{ij})_{n \times m}$ where $\sup_{ij} |a_{ij}| < C < \infty$, then*

$$\|\mathbf{A}\|_2 = O\left(\sqrt{nm}\right). \tag{S.20}$$

Proof. This result follows, since $\|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_\infty \|\mathbf{A}\|_1}$, $\|\mathbf{A}\|_\infty = O(m)$ and $\|\mathbf{A}\|_1 = O(n)$. ∎

Lemma S-2.11 *Consider two $N \times N$ nonsingular matrices $\mathbf{A}$ and $\mathbf{B}$ such that*

$$\|\mathbf{B}^{-1}\|_2 \|\mathbf{A} - \mathbf{B}\|_F < 1.$$

Then

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_F \leq \frac{\|\mathbf{B}^{-1}\|_2^2 \|\mathbf{A} - \mathbf{B}\|_F}{1 - \|\mathbf{B}^{-1}\|_2 \|\mathbf{A} - \mathbf{B}\|_F}.$$

Proof. By Lemma S-2.9,

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_F = \|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}\|_F \leq \|\mathbf{A}^{-1}\|_2\|\mathbf{B} - \mathbf{A}\|_F\|\mathbf{B}^{-1}\|_2$$

Note that

$$\|\mathbf{A}^{-1}\|_2 = \|\mathbf{A}^{-1} - \mathbf{B}^{-1} + \mathbf{B}^{-1}\|_2 \leq \|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_2 + \|\mathbf{B}^{-1}\|_2$$
$$\leq \|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_F + \|\mathbf{B}^{-1}\|_2,$$

and therefore,

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_F \leq (\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_F + \|\mathbf{B}^{-1}\|_2)\|\mathbf{B} - \mathbf{A}\|_F\|\mathbf{B}^{-1}\|_2.$$

Hence,

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_F(1 - \|\mathbf{B}^{-1}\|_2\|\mathbf{B} - \mathbf{A}\|_F) \leq \|\mathbf{B}^{-1}\|_2^2\|\mathbf{B} - \mathbf{A}\|_F.$$

Since $\|\mathbf{B}^{-1}\|_2\|\mathbf{B} - \mathbf{A}\|_F < 1$, we can further write,

$$\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\|_F \leq \frac{\|\mathbf{B}^{-1}\|_2^2\|\mathbf{A} - \mathbf{B}\|_F}{1 - \|\mathbf{B}^{-1}\|_2\|\mathbf{A} - \mathbf{B}\|_F}.$$

∎

Lemma S-2.12 *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be* $T \times N_x$ *and* $T \times N_y$ *matrices of observations on random variables* x_{it} *and* y_{jt}, *for* $i = 1, 2, \cdots, N_x$, $j = 1, 2, \cdots, N_y$ *and* $t = 1, 2, \cdots, T$, *respectively. Denote*

$$w_{ij,t} = x_{it}y_{jt} - \mathbb{E}(x_{it}y_{jt}), \ \textit{for all } i, j \textit{ and } t.$$

Suppose that

(i) $\sup_{i,t} \mathbb{E}|x_{it}|^4 < C$, $\sup_{j,t} \mathbb{E}|y_{jt}|^4 < C$, *and*

(ii) $\sup_{i,j} \left[\sum_{t=1}^{T} \sum_{t'=1}^{T} \mathbb{E}(w_{ij,t}w_{ij,t'}) \right] = O(T).$

Then,

$$\mathbb{E}\left\|T^{-1}\left[\mathbf{X}'\mathbf{Y} - \mathbb{E}(\mathbf{X}'\mathbf{Y})\right]\right\|_F^2 = O\left(\frac{N_x N_y}{T}\right). \tag{S.21}$$

Proof. The results follow from Lemma A18 of Chudik et al. (2018) Online Theory Supplement. ∎

Lemma S-2.13 Let $\mathbf{X} = (x_{ij})_{T \times N_x}$ and $\mathbf{Y} = (y_{ij})_{T \times N_y}$ be matrices of random variables, respectively. Suppose that,

$$\mathbb{E} \left\| T^{-1} \left[\mathbf{X}'\mathbf{Y} - \mathbb{E}(\mathbf{X}'\mathbf{Y}) \right] \right\|_F^2 = O(a_T), \tag{S.22}$$

where $a_T > 0$. Then

$$\left\| T^{-1} \left[\mathbf{X}'\mathbf{Y} - \mathbb{E}(\mathbf{X}'\mathbf{Y}) \right] \right\|_F = O_p(\sqrt{a_T}). \tag{S.23}$$

Proof. For any $B > 0$, by the Markov's inequality

$$\Pr \left(\left\| T^{-1} \left[\mathbf{X}'\mathbf{Y} - \mathbb{E}(\mathbf{X}'\mathbf{Y}) \right] \right\|_F > B\sqrt{a_T} \right) \leq \frac{\mathbb{E} \left\| T^{-1} \left[\mathbf{X}'\mathbf{Y} - \mathbb{E}(\mathbf{X}'\mathbf{Y}) \right] \right\|_F^2}{a_T B^2}$$

Since $\mathbb{E} \left\| T^{-1} \left[\mathbf{X}'\mathbf{Y} - \mathbb{E}(\mathbf{X}'\mathbf{Y}) \right] \right\|_F^2 = O(a_T)$, there exist C and T_0 such that for all $T > T_0$

$$\mathbb{E} \left\| T^{-1} \left[\mathbf{X}'\mathbf{Y} - \mathbb{E}(\mathbf{X}'\mathbf{Y}) \right] \right\|_F^2 \leq Ca_T.$$

Hence, for any $\varepsilon > 0$, there exist $B_\varepsilon = \sqrt{\frac{C}{\varepsilon}}$ and $T_\varepsilon = T_0$, such that for all $T > T_\varepsilon$

$$\Pr \left(\left\| T^{-1} \left[\mathbf{X}'\mathbf{Y} - \mathbb{E}(\mathbf{X}'\mathbf{Y}) \right] \right\|_F > B_\varepsilon \sqrt{a_T} \right) \leq \varepsilon.$$

Therefore,

$$\left\| T^{-1} \left[\mathbf{X}'\mathbf{Y} - \mathbb{E}(\mathbf{X}'\mathbf{Y}) \right] \right\|_F = O_p \left(\sqrt{a_T} \right).$$

∎

Lemma S-2.14 Let $\mathbf{\Sigma}_T$ be a positive definite matrix and $\hat{\mathbf{\Sigma}}_T$ be its corresponding estimator. Suppose that $\lambda_{\min} (\mathbf{\Sigma}_T) > c > 0$, and

$$\mathbb{E} \left\| \hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T \right\|_F^2 = O(a_T) \tag{S.24}$$

where $a_T > 0$, and $a_T = o(1)$. Then

$$\left\| \hat{\mathbf{\Sigma}}_T^{-1} - \mathbf{\Sigma}_T^{-1} \right\|_F = O_p(\sqrt{a_T}) \tag{S.25}$$

Proof. Let $\mathcal{A}_T = \left\{ \left\| \mathbf{\Sigma}_T^{-1} \right\|_2 \left\| \hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T \right\|_F < 1 \right\}$, $\mathcal{B}_T = \left\{ \left\| \hat{\mathbf{\Sigma}}_T^{-1} - \mathbf{\Sigma}_T^{-1} \right\|_F > B\sqrt{a_T} \right\}$ and $\mathcal{D}_T = \left\{ \frac{\left\| \mathbf{\Sigma}_T^{-1} \right\|_2^2 \left\| \hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T \right\|_F}{\left(1 - \left\| \mathbf{\Sigma}_T^{-1} \right\|_2 \left\| \hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T \right\|_F \right)} > B\sqrt{a_T} \right\}$ where $B > 0$ is an arbitrary constant. If $\mathcal{A}_T$

holds, by Lemma S-2.11,

$$\left\|\hat{\mathbf{\Sigma}}_T^{-1} - \mathbf{\Sigma}_T^{-1}\right\|_F \leq \frac{\left\|\mathbf{\Sigma}_T^{-1}\right\|_2^2 \left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F}{1 - \left\|\mathbf{\Sigma}_T^{-1}\right\|_2 \left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F}.$$

Hence $\mathcal{B}_T \cap \mathcal{A}_T \subseteq \mathcal{D}_T$. Therefore

$$\Pr(\mathcal{B}_T \cap \mathcal{A}_T) \leq \Pr\left(\frac{\left\|\mathbf{\Sigma}_T^{-1}\right\|_2^2 \left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F}{\left(1 - \left\|\mathbf{\Sigma}_T^{-1}\right\|_2 \left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F\right)} > B\sqrt{a_T}\right)$$

$$= \Pr\left(\left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F > \frac{B\sqrt{a_T}}{\left\|\mathbf{\Sigma}_T^{-1}\right\|_2 \left(\left\|\mathbf{\Sigma}_T^{-1}\right\|_2 + B\sqrt{a_T}\right)}\right)$$

By the Markov's inequality, we can further conclude that

$$\Pr(\mathcal{B}_T \cap \mathcal{A}_T) \leq \frac{\mathbb{E}\left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F^2}{a_T} \times \frac{\left\|\mathbf{\Sigma}_T^{-1}\right\|_2^2 \left(\left\|\mathbf{\Sigma}_T^{-1}\right\|_2 + B\sqrt{a_T}\right)^2}{B^2}.$$

Since by assumption $\mathbb{E}\left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F^2 = O(a_T)$, there exist C and $T_0 > 0$ such that for all $T > T_0$,

$$\mathbb{E}\left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F^2 \leq Ca_T.$$

Therefore, for all $T > T_0$,

$$\Pr(\mathcal{B}_T \cap \mathcal{A}_T) \leq \frac{C\left\|\mathbf{\Sigma}_T^{-1}\right\|_2^2 \left(\left\|\mathbf{\Sigma}_T^{-1}\right\|_2 + B\sqrt{a_T}\right)^2}{B^2}.$$

Moreover,

$$\Pr(\mathcal{A}_T^c) = \Pr\left(\left\|\mathbf{\Sigma}_T^{-1}\right\|_2 \left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F \geq 1\right) = \Pr\left(\left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F \geq \frac{1}{\left\|\mathbf{\Sigma}_T^{-1}\right\|_2}\right).$$

By the Markov's inequality, we can further write

$$\Pr(\mathcal{A}_T^c) \leq \left\|\mathbf{\Sigma}_T^{-1}\right\|_2^2 \times \mathbb{E}\left\|\hat{\mathbf{\Sigma}}_T - \mathbf{\Sigma}_T\right\|_F^2,$$

and hence, for all $T > T_0$,

$$\Pr(\mathcal{A}_T^c) \leq C\left\|\mathbf{\Sigma}_T^{-1}\right\|_2^2 a_T.$$

Note that

$$\Pr(\mathcal{B}_T) = \Pr\left(\mathcal{B}_T \cap \mathcal{A}_T\right) + \Pr(\mathcal{B}_T | \mathcal{A}_T^c)\Pr(\mathcal{A}_T^c),$$

and since $\Pr(\mathcal{B}_T \cap \mathcal{A}_T) \leq \Pr(\mathcal{D}_T)$ and $\Pr(\mathcal{B}_T | \mathcal{A}_T^c) \leq 1$, we have

$$\Pr(\mathcal{B}_T) \leq \Pr(\mathcal{B}_T \cap \mathcal{A}_T) + \Pr(\mathcal{A}_T^c).$$

Therefore, for all $T > T_0$,

$$\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}_T^{-1} - \boldsymbol{\Sigma}_T^{-1}\right\|_F > B\sqrt{a_T}\right) \leq \frac{C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2\left(\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2 + B\sqrt{a_T}\right)^2}{B^2} + C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2 a_T.$$

Now, for a given $\varepsilon > 0$, we are interested to find $B_\varepsilon > 0$ and $T_\varepsilon > 0$ such that for all $T > T_\varepsilon$,

$$\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}_T^{-1} - \boldsymbol{\Sigma}_T^{-1}\right\|_F > B_\varepsilon\sqrt{a_T}\right) \leq \varepsilon.$$

To do so, we first find a value of B such that

$$\frac{C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2\left(\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2 + B\sqrt{a_T}\right)^2}{B^2} + C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2 a_T = \varepsilon.$$

By multiplying both sides of the above equality by B^2 and bringing all the equations to the left hand side we have

$$\left(\varepsilon - 2C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2 a_T\right)B^2 - 2C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^3\sqrt{a_T}B - C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^4 = 0.$$

By solving the above quadratic equation of B we have

$$B^* = \frac{2C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^3\sqrt{a_T} \pm \sqrt{4C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^4\varepsilon - 4C^2\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^6 a_T}}{2\left(\varepsilon - 2C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2 a_T\right)}$$

$$= \frac{\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2\left(\sqrt{a_T} \pm \sqrt{\frac{\varepsilon}{C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2} - a_T}\right)}{\frac{\varepsilon}{C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2} - 2a_T}$$

Notice that $a_T \to 0$ as $T \to \infty$, therefore for large enough T^* we have both $\frac{\varepsilon}{C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2} - 2a_T$ and $\frac{\varepsilon}{C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2} - a_T$ being greater than zero for all $T > T^*$. Now, by setting $T_\varepsilon = \max\{T^*, T_0\}$ and

$$B_\varepsilon = \frac{\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2\left(\sqrt{a_T} + \sqrt{\frac{\varepsilon}{C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2} - a_T}\right)}{\frac{\varepsilon}{C\left\|\boldsymbol{\Sigma}_T^{-1}\right\|_2^2} - 2a_T} > 0,$$

we achieve our goal that for all $T > T_\varepsilon$,

$$\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}_T^{-1} - \boldsymbol{\Sigma}_T^{-1}\right\|_F > B_\varepsilon\sqrt{a_T}\right) \le \varepsilon.$$

∎

Remark 7 *By using Lemma S-2.11 we achieve the probability convergence order for $\left\|\hat{\boldsymbol{\Sigma}}_T^{-1} - \boldsymbol{\Sigma}_T^{-1}\right\|_F$ that is sharper than the one shown in the proof Lemma A21 of Chudik et al. (2018) (see equations (B.103) and (B.105) of Chudik et al. (2018) Online Theory Supplement).*

Lemma S-2.15 *Let z_{ij} be a random variable for $i = 1, 2, \cdots, N$, and $j = 1, 2, \cdots, N$. Then, for any $d_T > 0$,*

$$\Pr(N^{-2} \textstyle\sum_{i=1}^{N} \sum_{j=1}^{N} |z_{ij}| > d_T) \le N^2 \sup_{i,j} \Pr(|z_{ij}| > d_T)$$

Proof. We know that $N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} |z_{ij}| \le \sup_{i,j} |z_{ij}|$. Therefore,

$$\begin{aligned}
\Pr(N^{-2} \textstyle\sum_{i=1}^{N} \sum_{j=1}^{N} |z_{ij}| > d_T) &\le \Pr(\sup_{i,j} |z_{ij}| > d_T)\\
&\le \Pr[\cup_{i=1}^{N} \cup_{j=1}^{N} (|z_{ij}| > d_T)] \le \textstyle\sum_{i=1}^{N} \sum_{j=1}^{N} \Pr(|z_{ij}| > d_T)\\
&\le N^2 \sup_{i,j} \Pr(|z_{ij}| > d_T).
\end{aligned}$$

∎

Lemma S-2.16 *Let $\hat{\boldsymbol{\Sigma}}$ be an estimator of a $N \times N$ symmetric invertible matrix $\boldsymbol{\Sigma}$. Suppose that there exits a finite positive constant C_0, such that*

$$\sup_{i,j} \Pr(|\hat{\sigma}_{ij} - \sigma_{ij}| > d_T) \le \exp(-C_0 T d_T^2), \text{ for any } d_T > 0,$$

where σ_{ij} and $\hat{\sigma}_{ij}$ are the elements of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ respectively. Then, for any $b_T > 0$,

$$\Pr(\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_F > b_T) \le N^2 \exp\left[-C_0 \frac{T b_T^2}{N^2 \|\boldsymbol{\Sigma}^{-1}\|_2^2 (\|\boldsymbol{\Sigma}^{-1}\|_2 + b_T)^2}\right] + $$
$$N^2 \exp\left(-C_0 \frac{T}{N^2 \|\boldsymbol{\Sigma}^{-1}\|_2^2}\right).$$

Proof. Let $\mathcal{A}_N = \{\|\boldsymbol{\Sigma}^{-1}\|_2 \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F \le 1\}$ and $\mathcal{B}_N = \{\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_F > b_T\}$, and note that by Lemma S-2.11 if $\mathcal{A}_N$ holds we have

$$\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_F \le \frac{\|\boldsymbol{\Sigma}^{-1}\|_2^2 \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F}{1 - \|\boldsymbol{\Sigma}^{-1}\|_2 \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_F}.$$

Hence

$$\Pr(\mathcal{B}_N|\mathcal{A}_N) \le \Pr\left(\frac{\|\mathbf{\Sigma}^{-1}\|_2^2\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F}{1 - \|\mathbf{\Sigma}^{-1}\|_2\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F} > b_T\right)$$

$$= \Pr\left[\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F > \frac{b_T}{\|\mathbf{\Sigma}^{-1}\|_2(\|\mathbf{\Sigma}^{-1}\|_2 + b_T)}\right]$$

Note that $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F = \left(\sum_{i=1}^N \sum_{j=1}^N (\hat{\sigma}_{ij} - \sigma_{ij})^2\right)^{1/2}$. Therefore,

$$\Pr(\mathcal{B}_N|\mathcal{A}_N) \le \Pr\left[\left(\sum_{i=1}^N \sum_{j=1}^N (\hat{\sigma}_{ij} - \sigma_{ij})^2\right)^{1/2} > \frac{b_T}{\|\mathbf{\Sigma}^{-1}\|_2(\|\mathbf{\Sigma}^{-1}\|_2 + b_T)}\right]$$

$$= \Pr\left[\sum_{i=1}^N \sum_{j=1}^N (\hat{\sigma}_{ij} - \sigma_{ij})^2 > \frac{b_T^2}{\|\mathbf{\Sigma}^{-1}\|_2^2(\|\mathbf{\Sigma}^{-1}\|_2 + b_T)^2}\right]$$

By Lemma S-2.15, we can further write,

$$\Pr(\mathcal{B}_N|\mathcal{A}_N) \le N^2 \sup_{i,j} \Pr\left[(\hat{\sigma}_{ij} - \sigma_{ij})^2 > \frac{b_T^2}{N^2\|\mathbf{\Sigma}^{-1}\|_2^2(\|\mathbf{\Sigma}^{-1}\|_2 + b_T)^2}\right]$$

$$= N^2 \sup_{i,j} \Pr\left[|\hat{\sigma}_{ij} - \sigma_{ij}| > \frac{b_T}{N\|\mathbf{\Sigma}^{-1}\|_2(\|\mathbf{\Sigma}^{-1}\|_2 + b_T)}\right]$$

$$\le N^2 \exp\left[-C_0 \frac{Tb_T^2}{N^2\|\mathbf{\Sigma}^{-1}\|_2^2(\|\mathbf{\Sigma}^{-1}\|_2 + b_T)^2}\right]$$

Furthermore,

$$\Pr(\mathcal{A}_N^c) = \Pr(\|\mathbf{\Sigma}^{-1}\|_2\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F > 1)$$

$$= \Pr(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F > \|\mathbf{\Sigma}^{-1}\|_2^{-1})$$

$$= \Pr\left[\left(\sum_{i=1}^N \sum_{j=1}^N (\hat{\sigma}_{ij} - \sigma_{ij})^2\right)^{1/2} > \|\mathbf{\Sigma}^{-1}\|_2^{-1}\right]$$

$$= \Pr\left[\sum_{i=1}^N \sum_{j=1}^N (\hat{\sigma}_{ij} - \sigma_{ij})^2 > \|\mathbf{\Sigma}^{-1}\|_2^{-2}\right]$$

$$\le N^2 \sup_{i,j} \Pr\left[(\hat{\sigma}_{ij} - \sigma_{ij})^2 > \frac{1}{N^2\|\mathbf{\Sigma}^{-1}\|_2^2}\right]$$

$$\le N^2 \sup_{i,j} \Pr\left[|\hat{\sigma}_{ij} - \sigma_{ij}| > \frac{1}{N\|\mathbf{\Sigma}^{-1}\|_2}\right]$$

$$\le N^2 \exp\left[-C_0 \frac{T}{N^2\|\mathbf{\Sigma}^{-1}\|_2^2}\right]$$

Note that

$$\Pr(\mathcal{B}_N) = \Pr(\mathcal{B}_N|\mathcal{A}_N)\Pr(\mathcal{A}_N) + \Pr(\mathcal{B}_N|\mathcal{A}_N^c)\Pr(\mathcal{A}_N^c),$$

and since $\Pr(\mathcal{A}_N)$ and $\Pr(\mathcal{B}_N|\mathcal{A}_N^c)$ are less than equal to one, we have

$$\Pr(\mathcal{B}_N) \leq \Pr(\mathcal{B}_N|\mathcal{A}_N) + \Pr(\mathcal{A}_N^c).$$

Therefore,

$$\Pr(\mathcal{B}_{NT}) \leq N^2 \exp\left[-C_0 \frac{Tb_T^2}{N^2\|\mathbf{\Sigma}^{-1}\|_2^2(\|\mathbf{\Sigma}^{-1}\|_2 + b_T)^2}\right] + N^2 \exp\left[-C_0 \frac{T}{N^2\|\mathbf{\Sigma}^{-1}\|_2^2}\right].$$

∎

Lemma S-2.17 *Let $\hat{\mathbf{\Sigma}}$ be an estimator of a $N \times N$ symmetric invertible matrix $\mathbf{\Sigma}$. Suppose that there exits a finite positive constant C_0, such that*

$$\sup_{i,j} \Pr(|\hat{\sigma}_{ij} - \sigma_{ij}| > d_T) \leq \exp\left[-C_0(Td_T)^{s/s+2}\right], \text{ for any } d_T > 0,$$

where σ_{ij} and $\hat{\sigma}_{ij}$ are the elements of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ respectively. Then, for any $b_T > 0$,

$$\Pr(\|\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\|_F > b_T) \leq N^2 \exp\left[-C_0 \frac{(Tb_T)^{s/s+2}}{N^{s/s+2}\|\mathbf{\Sigma}^{-1}\|_2^{s/s+2}(\|\mathbf{\Sigma}^{-1}\|_2 + b_T)^{s/s+2}}\right] +$$
$$N^2 \exp\left(-C_0 \frac{T^{s/s+2}}{N^{s/s+2}\|\mathbf{\Sigma}^{-1}\|_2^{s/s+2}}\right).$$

Proof. The proof is similar to the proof of Lemma S-2.16. ∎

Lemma S-2.18 *Let $\{x_{it}\}_{t=1}^T$ for $i = 1, 2, \cdots, N$ and $\{z_{jt}\}_{t=1}^T$ for $j = 1, 2, \cdots, m$ be time-series processes. Also let $\mathcal{F}_{it}^x = \sigma(x_{it}, x_{i,t-1}, \cdots)$ for $i = 1, 2, \cdots, N$, $\mathcal{F}_{jt}^z = \sigma(z_{jt}, z_{j,t-1}, \cdots)$ for $j = 1, 2, \cdots, m$, $\mathcal{F}_t^x = \cup_{i=1}^N \mathcal{F}_{it}^x$, $\mathcal{F}_t^z = \cup_{j=1}^m \mathcal{F}_{jt}^z$, and $\mathcal{F}_t = \mathcal{F}_t^x \cup \mathcal{F}_t^z$. Define the projection regression of x_{it} on $\mathbf{z}_t = (z_{1t}, z_{2t}, \cdots, z_{m,t})'$ as*

$$x_{it} = \mathbf{z}_t'\boldsymbol{\psi}_{i,T} + \nu_{it}$$

where $\boldsymbol{\psi}_{i,T} = (\psi_{1i,T}, \psi_{2i,T}, \cdots, \psi_{mi,T})'$ is the $m \times 1$ vector of projection coefficients which is equal to $\left[T^{-1}\sum_{t=1}^T \mathbb{E}\left(\mathbf{z}_t\mathbf{z}_t'\right)\right]^{-1}\left[T^{-1}\sum_{t=1}^T \mathbb{E}(\mathbf{z}_t x_{it})\right]$. Suppose, $\mathbb{E}[x_{it}x_{i't} - \mathbb{E}(x_{it}x_{i't})|\mathcal{F}_{t-1}] = 0$ for all $i, i' = 1, 2, \cdots, N$, $\mathbb{E}[z_{jt}z_{j't} - \mathbb{E}(z_{jt}z_{j't})|\mathcal{F}_{t-1}] = 0$ for all $j, j' = 1, 2, \cdots, m$, and $\mathbb{E}[z_{jt}x_{it} - \mathbb{E}(z_{jt}x_{it})|\mathcal{F}_{t-1}] = 0$ for all $j = 1, 2, \cdots, m$ and for all $i = 1, 2, \cdots, N$. Then

$$\mathbb{E}\left[\nu_{it}\nu_{i't} - \mathbb{E}(\nu_{it}\nu_{i't})|\mathcal{F}_{t-1}\right] = 0,$$

for all $j, j' = 1, 2, \cdots, N$,

$$\mathbb{E}\left[\nu_{it} z_{jt} - \mathbb{E}(\nu_{it} z_{jt}) | \mathcal{F}_{t-1}\right] = 0,$$

for all $i = 1, 2, \cdots, N$ and $j = 1, 2, \cdots, m$, and

$$T^{-1} \sum_{t=1}^{T} \mathbb{E}(\nu_{it} z_{jt}) = 0,$$

for all $i = 1, 2, \cdots, N$ and $j = 1, 2, \cdots, m$.

Proof.

$$
\begin{aligned}
\mathbb{E}(\nu_{it} \nu_{i't} | \mathcal{F}_{t-1}) &= \mathbb{E}(x_{it} x_{i't} | \mathcal{F}_{t-1}) - \mathbb{E}(x_{it} \mathbf{z}'_t | \mathcal{F}_{t-1}) \boldsymbol{\psi}_{i',T} - \\
&\quad \mathbb{E}(x_{i't} \mathbf{z}'_t | \mathcal{F}_{t-1}) \boldsymbol{\psi}_{i,T} + \boldsymbol{\psi}'_{i,T} \mathbb{E}(\mathbf{z}_t \mathbf{z}'_t | \mathcal{F}_{t-1}) \boldsymbol{\psi}_{i',T} \\
&= \mathbb{E}(x_{it} x_{i't}) - \mathbb{E}(x_{it} \mathbf{z}'_t) \boldsymbol{\psi}_{i',T} - \mathbb{E}(x_{i't} \mathbf{z}'_t) \boldsymbol{\psi}_{i,T} + \\
&\quad \boldsymbol{\psi}'_{i,T} \mathbb{E}(\mathbf{z}_t \mathbf{z}'_t) \boldsymbol{\psi}_{i',T} = \mathbb{E}(\nu_{it} \nu_{i't}).
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(\nu_{it} z_{jt} | \mathcal{F}_{t-1}) &= \mathbb{E}(x_{it} z_{jt} | \mathcal{F}_{t-1}) - \mathbb{E}(\mathbf{z}'_t z_{jt} | \mathcal{F}_{t-1}) \boldsymbol{\psi}_{i,T} \\
&= \mathbb{E}(x_{it} z_{jt}) - \mathbb{E}(\mathbf{z}'_t z_{jt}) \boldsymbol{\psi}_{i,T} = \mathbb{E}(\nu_{it} z_{it}).
\end{aligned}
$$

$$
\begin{aligned}
T^{-1} \sum_{t=1}^{T} \mathbb{E}(\nu_{it} \mathbf{z}_t) &= T^{-1} \sum_{t=1}^{T} \mathbb{E}(x_{it} \mathbf{z}_t) - \boldsymbol{\psi}'^{-1}_{i,T} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t \mathbf{z}'_t)] \\
&= T^{-1} \sum_{t=1}^{T} \mathbb{E}(x_{it} \mathbf{z}_t) - T^{-1} \sum_{t=1}^{T} \mathbb{E}(x_{it} \mathbf{z}_t) = \mathbf{0}.
\end{aligned}
$$

∎

Lemma S-2.19 *Let $\{x_{it}\}_{t=1}^{T}$ for $i = 1, 2, \cdots, N$ and $\{z_{jt}\}_{t=1}^{T}$ for $j = 1, 2, \cdots, m$ be time-series processes. Define the projection regression of x_{it} on $\mathbf{z}_t = (z_{1t}, z_{2t}, \cdots, z_{m,t})'$ as*

$$x_{it} = \mathbf{z}'_t \boldsymbol{\psi}_{i,T} + \nu_{it}$$

where $\boldsymbol{\psi}_{i,T} = (\psi_{1i,T}, \psi_{2i,T}, \cdots, \psi_{mi,T})'$ is the $m \times 1$ vector of projection coefficients which is equal to $\left[T^{-1} \sum_{t=1}^{T} \mathbb{E}\left(\mathbf{z}_t \mathbf{z}'_t\right)\right]^{-1} [T^{-1} \sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t x_{it})]$. Suppose that only a finite number of elements in $\boldsymbol{\psi}_{i,T}$ is different from zero for all $i = 1, 2, \cdots, N$ and there exist sufficiently large positive constants C_0 and C_1, and $s > 0$ such that

(i) $\sup_{j,t} \Pr(|z_{jt}| > \alpha) \leq C_0 \exp(-C_1 \alpha^s)$, for all $\alpha > 0$, and

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

Then, there exist sufficiently large positive constants C_0 and C_1, and $s > 0$ such that

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

Proof. Without loss of generality assume that the first finite ℓ elements of $\psi_{i,T}$ are different from zero and write

$$x_{it} = \sum_{j=1}^{\ell} \psi_{ji,T} z_{jt} + \nu_{it}.$$

Now, note that

$$\Pr(|\nu_{it}| > \alpha) \leq \Pr\left(|x_{it}| + \sum_{j=1}^{\ell}|\psi_{ji,T} z_{jt}| > \alpha\right),$$

and hence by Lemma S-2.3, for any $0 < \pi_j < 1$, $j = 1, 2, \cdots, \ell+1$ we have,

$$\begin{aligned}
\Pr(|\nu_{it}| > \alpha) &\leq \sum_{j=1}^{\ell} \Pr\left(|\psi_{ji,T} z_{jt}| > \pi_j \alpha\right) + \Pr\left(|x_{it}| > \pi_{\ell+1}\alpha\right) \\
&= \sum_{j=1}^{\ell} \Pr\left(|z_{jt}| > |\psi_{ji,T}|^{-1}\pi_j \alpha\right) + \Pr\left(|x_{it}| > \pi_{\ell+1}\alpha\right) \\
&\leq \ell \sup_{j,t} \Pr\left(|z_{jt}| > |\psi_T^*|^{-1}\pi^*\alpha\right) + \sup_{i,t} \Pr\left(|x_{it}| > \pi^*\alpha\right).
\end{aligned}$$

where $\psi_T^* = sup_{i,j}\{\psi_{ji,T}\}$ and $\pi^* = \inf_{j \in 1,2,\cdots,\ell+1}\{\pi_j\}$. Therefore, by the exponential decaying probability tail assumptions for x_{it} and z_{jt} we have

$$\Pr(|\nu_{it}| > \alpha) \leq \ell C_0 \exp(-C_1 \alpha^s) + C_0 \exp(-C_1 \alpha^s),$$

and hence there exist sufficiently large positive constants C_0 and C_1, and $s > 0$ such that

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

∎

Lemma S-2.20 *Let $\{x_{it}\}_{t=1}^{T}$ for $i = 1, 2, \cdots, N$ and $\{z_{\ell t}\}_{t=1}^{T}$ for $\ell = 1, 2, \cdots, m$ be time-series processes and $m = \ominus(T^d)$. Also let $\mathcal{F}_{it}^x = \sigma(x_{it}, x_{i,t-1}, \cdots)$ for $i = 1, 2, \cdots, N$, $\mathcal{F}_{\ell t}^z = \sigma(z_{\ell t}, z_{\ell,t-1}, \cdots)$ for $\ell = 1, 2, \cdots, m$, $\mathcal{F}_t^x = \cup_{i=1}^{N}\mathcal{F}_{it}^x$, $\mathcal{F}_t^z = \cup_{\ell=1}^{m}\mathcal{F}_{\ell t}^z$, and $\mathcal{F}_t = \mathcal{F}_t^x \cup \mathcal{F}_t^z$. Define the projection regression of x_{it} on $\mathbf{z}_t = (z_{1t}, z_{2t}, \cdots, z_{m,t})'$ as*

$$x_{it} = \mathbf{z}_t' \boldsymbol{\psi}_{i,T} + \nu_{it}$$

where $\boldsymbol{\psi}_{i,T} = (\psi_{1i,T}, \psi_{2i,T}, \cdots, \psi_{mi,T})'$ is the $m \times 1$ vector of projection coefficients which is equal to $\left[T^{-1}\sum_{t=1}^{T} \mathbb{E}\left(\mathbf{z}_t \mathbf{z}_t'\right)\right]^{-1} [T^{-1}\sum_{t=1}^{T} \mathbb{E}(\mathbf{z}_t x_{it})]$. Suppose, $\mathbb{E}[x_{it}x_{jt} - \mathbb{E}(x_{it}x_{jt})|\mathcal{F}_{t-1}] = 0$ for all $i, j = 1, 2, \cdots, N$, $\mathbb{E}[z_{\ell t}z_{\ell' t} - \mathbb{E}(z_{\ell t}z_{\ell t})|\mathcal{F}_{t-1}] = 0$ for all $\ell, \ell' = 1, 2, \cdots, m$, and $\mathbb{E}[z_{\ell t}x_{it} - \mathbb{E}(z_{\ell t}x_{it})|\mathcal{F}_{t-1}] = 0$ for all $\ell = 1, 2, \cdots, m$ and for all $i = 1, 2, \cdots, N$. Additionally, assume that only a finite number of elements in $\boldsymbol{\psi}_{i,T}$ is different from zero for all $i = 1, 2, \cdots, N$ and there exist sufficiently large positive constants C_0 and C_1, and $s > 0$ such that

(i) $\sup_{j,t} \Pr(|z_{\ell t}| > \alpha) \leq C_0 \exp(-C_1 \alpha^s), \ for \ all \ \alpha > 0$, and

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

Then, there exist some finite positive constants C_0, C_1 and C_2 such that if $d < \lambda \leq (s+2)/(s+4)$,

$$\Pr(|\mathbf{x}_i' \mathbf{M}_z \mathbf{x}_j - \mathbb{E}(\boldsymbol{\nu}_i' \boldsymbol{\nu}_j)| > \zeta_T) \leq \exp(-C_0 T^{-1} \zeta_T^2) + \exp(-C_1 T^{C_2})$$

and if $\lambda > (s+2)/(s+4)$

$$\Pr(|\mathbf{x}_i' \mathbf{M}_z \mathbf{x}_j - \mathbb{E}(\boldsymbol{\nu}_i' \boldsymbol{\nu}_j)| > \zeta_T) \leq \exp(-C_0 \zeta_T^{s/(s+1)}) + \exp(-C_1 T^{C_2})$$

for all $i, j = 1, 2, \cdots, N$, where $\boldsymbol{\nu}_i = (\nu_{i1}, \nu_{i2}, \cdots, \nu_{iT})'$, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iT})'$, and $\mathbf{M}_z = \mathbf{I} - T^{-1} \mathbf{Z} \hat{\boldsymbol{\Sigma}}_{zz}^{-1} \mathbf{Z}'$ with $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T)'$ and $\hat{\boldsymbol{\Sigma}}_{zz} = T^{-1} \sum_{t=1}^{T} (\mathbf{z}_t \mathbf{z}_t')$.

Proof.

$$\Pr\left[|\mathbf{x}_i' \mathbf{M}_z \mathbf{x}_j - \mathbb{E}(\boldsymbol{\nu}_i' \boldsymbol{\nu}_j)| > \zeta_T\right] = \Pr\left[|\boldsymbol{\nu}_i' \mathbf{M}_z \boldsymbol{\nu}_j - \mathbb{E}(\boldsymbol{\nu}_i' \boldsymbol{\nu}_j)| > \zeta_T\right]$$
$$= \Pr\left[|\boldsymbol{\nu}_i' \boldsymbol{\nu}_j - \mathbb{E}(\boldsymbol{\nu}_i' \boldsymbol{\nu}_j) - T^{-1} \boldsymbol{\nu}_i' \mathbf{Z} \boldsymbol{\Sigma}_{zz}^{-1} \mathbf{Z}' \boldsymbol{\nu}_j - T^{-1} \boldsymbol{\nu}_i' \mathbf{Z}(\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}) \mathbf{Z}' \boldsymbol{\nu}_j| > \zeta_T\right]$$

where $\boldsymbol{\Sigma}_{zz} = \mathbb{E}[T^{-1} \sum_{t=1}^{T} (\mathbf{z}_t \mathbf{z}_t')]$. By Lemma S-2.3, we can further write

$$\Pr\left[|\mathbf{x}_i' \mathbf{M}_z \mathbf{x}_j - \mathbb{E}(\boldsymbol{\nu}_i' \boldsymbol{\nu}_j)| > \zeta_T\right]$$
$$\leq \Pr\left[|\boldsymbol{\nu}_i' \boldsymbol{\nu}_j - \mathbb{E}(\boldsymbol{\nu}_i' \boldsymbol{\nu}_j)| > \pi_1 \zeta_T\right] + \Pr(|T^{-1} \boldsymbol{\nu}_i' \mathbf{Z} \boldsymbol{\Sigma}_{zz}^{-1} \mathbf{Z}' \boldsymbol{\nu}_j| > \pi_2 \zeta_T) +$$
$$\Pr\left[|T^{-1} \boldsymbol{\nu}_i' \mathbf{Z}(\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}) \mathbf{Z}' \boldsymbol{\nu}_j|) > \pi_3 \zeta_T\right].$$

where $0 < \pi_i < 1$ and $\sum_{i=1}^{3} \pi_i = 1$. By Lemma S-2.9,

$$\Pr(|T^{-1} \boldsymbol{\nu}_i' \mathbf{Z} \boldsymbol{\Sigma}_{zz}^{-1} \mathbf{Z}' \boldsymbol{\nu}_j| > \pi_2 \zeta_T) \leq \Pr(\|\boldsymbol{\nu}_i' \mathbf{Z}\|_F \|\boldsymbol{\Sigma}_{zz}^{-1}\|_2 \|\mathbf{Z}' \boldsymbol{\nu}_j\|_F > \pi_2 \zeta_T T),$$

and again by Lemma S-2.4, we have

$$\Pr(|T^{-1} \boldsymbol{\nu}_i' \mathbf{Z} \boldsymbol{\Sigma}_{zz}^{-1} \mathbf{Z}' \boldsymbol{\nu}_j| > \pi_2 \zeta_T)$$
$$\leq \Pr(\|\boldsymbol{\nu}_i' \mathbf{Z}\|_F > \|\boldsymbol{\Sigma}_{zz}^{-1}\|_2^{-1/2} \pi_2^{1/2} \zeta_T^{1/2} T^{1/2}) + \Pr(\|\mathbf{Z}' \boldsymbol{\nu}_j\|_F > \|\boldsymbol{\Sigma}_{zz}^{-1}\|_2^{-1/2} \pi_2^{1/2} \zeta_T^{1/2} T^{1/2}).$$

Similarly, we can show that

$$\Pr(|T^{-1} \boldsymbol{\nu}_i' \mathbf{Z}(\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}) \mathbf{Z}' \boldsymbol{\nu}_j| > \pi_3 \zeta_T)$$
$$\leq \Pr(\|\boldsymbol{\nu}_i' \mathbf{Z}\|_F \|\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}\|_F \|\mathbf{Z}' \boldsymbol{\nu}_j\|_F > \pi_3 \zeta_T T)$$
$$\leq \Pr(\|\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}\|_F > \delta_T^{-1} \zeta_T) + \Pr(\|\boldsymbol{\nu}_i' \mathbf{Z}\|_F > \pi_3^{1/2} \delta_T^{1/2} T^{1/2})$$
$$+ \Pr(\|\mathbf{Z}' \boldsymbol{\nu}_j\|_F > \pi_3^{1/2} \delta_T^{1/2} T^{1/2})$$

where $\delta_T = \ominus(T^\alpha)$ with $0 < \alpha < \lambda$.

Note that $\Pr(\|\mathbf{Z}'\boldsymbol{\nu}_i\|_F > c) = \Pr(\|\mathbf{Z}'\boldsymbol{\nu}_i\|_F^2 > c^2) = \Pr[\sum_{\ell=1}^{m}(\sum_{t=1}^{T} \nu_{it} z_{\ell t})^2 > c^2]$, where c is a positive constant. So, by Lemma S-2.3, we have

$$\Pr(\|\mathbf{Z}'\boldsymbol{\nu}_i\|_F > c) \le \sum_{\ell=1}^{m} \Pr[(\sum_{t=1}^{T} \nu_{it} z_{\ell t})^2 > m^{-1} c^2]$$

Hence, $\Pr(\|\mathbf{Z}'\boldsymbol{\nu}_i\|_F > c) \le \sum_{\ell=1}^{m} \Pr(|\sum_{t=1}^{T} \nu_{it} z_{\ell t}| > m^{-1/2} c)$. Also, by Lemma S-2.18 we have $\sum_{t=1}^{T} \mathbb{E}(\nu_{it} z_{\ell t}) = 0$ and hence we can further write

$$\Pr(\|\mathbf{Z}'\boldsymbol{\nu}_i\|_F > c) \le \sum_{\ell=1}^{m} \Pr\{|\sum_{t=1}^{T}[\nu_{it} z_{\ell t} - \mathbb{E}(\nu_{it} z_{\ell t})]| > m^{-1/2} c\}.$$

Note that $\|\boldsymbol{\Sigma}_{zz}^{-1}\|_2$ is equal to the largest eigenvalue of $\boldsymbol{\Sigma}_{zz}^{-1}$ and it is a finite positive constant. So, there exists a positive constant $C > 0$ such that,

$$\begin{aligned}
\Pr(|\mathbf{x}_i' \mathbf{M}_z \mathbf{x}_j - \mathbb{E}(\boldsymbol{\nu}_i' \boldsymbol{\nu}_j)| &> \zeta_T) \\
&\le \Pr\{|\sum_{t=1}^{T}[\nu_{it}\nu_{jt} - \mathbb{E}(\nu_{it}\nu_{jt})]| > CT^\lambda\} + \\
&\quad \sum_{\ell=1}^{m} \Pr\{|\sum_{t=1}^{T}[\nu_{it} z_{\ell t} - \mathbb{E}(\nu_{it} z_{\ell t})]| > CT^{1/2+\lambda/2-d/2}\} + \\
&\quad \sum_{\ell=1}^{m} \Pr\{|\sum_{t=1}^{T}[\nu_{jt} z_{\ell t} - \mathbb{E}(\nu_{jt} z_{\ell t})]| > CT^{1/2+\lambda/2-d/2}\} + \\
&\quad \sum_{\ell=1}^{m} \Pr\{|\sum_{t=1}^{T}[\nu_{it} z_{\ell t} - \mathbb{E}(\nu_{it} z_{\ell t})]| > CT^{1/2+\alpha/2-d/2}\} + \\
&\quad \sum_{\ell=1}^{m} \Pr\{|\sum_{t=1}^{T}[\nu_{jt} z_{\ell t} - \mathbb{E}(\nu_{jt} z_{\ell t})]| > CT^{1/2+\alpha/2-d/2}\} + \\
&\quad \Pr(\|\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}\|_F > \delta_T^{-1} \zeta_T)
\end{aligned}$$

Let

$$\kappa_{T,i}(h, d) = \sum_{\ell=1}^{m} \Pr\{|\sum_{t=1}^{T}[\nu_{it} z_{\ell t} - \mathbb{E}(\nu_{it} z_{\ell t})]| > CT^{1/2+\kappa/2-d/2}\}, \text{ for } h = \lambda, \alpha,$$

and $i = 1, 2, ..., N$. By Lemmas S-2.7, S-2.18, and S-2.19, we have $\nu_{it}\nu_{jt} - \mathbb{E}(\nu_{it}\nu_{jt})$ and $\nu_{it} z_{\ell t} - \mathbb{E}(\nu_{it} z_{\ell t})$ are martingale difference processes with exponentially bounded probability tail, $\frac{s}{2}$. So, depending on the value of exponentially bounded probability tail parameter, from Lemma S-2.1, we know that either

$$\kappa_{T,i}(h, d) \le m \exp\left[-\ominus\left(T^{h-d}\right)\right]$$

or

$$\kappa_{T,i}(h, d) \le m \exp\left[-\ominus\left(T^{s(1/2+h/2-d/2)/(s+2)}\right)\right],$$

for $h = \lambda, \alpha$. Also, depending on the value of exponentially bounded probability tail param-

eter, from Lemmas S-2.16 and S-2.17 we have,

$$\Pr(\|\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}\|_F > \delta_T^{-1}\zeta_T) \le m^2 \exp\left[-C_0 \frac{T\delta_T^{-2}\zeta_T^2}{m^2\|\boldsymbol{\Sigma}_{zz}^{-1}\|_2^2(\|\boldsymbol{\Sigma}_{zz}^{-1}\|_2 + \delta_T^{-1}\zeta_T)^2}\right] +$$
$$m^2 \exp\left(-C_0 \frac{T}{m^2\|\boldsymbol{\Sigma}_{zz}^{-1}\|_2^2}\right),$$

or

$$\Pr(\|\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}\|_F > \delta_T^{-1}\zeta_T) \le m^2 \exp\left[-C_0 \frac{(T\delta_T^{-1}\zeta_T)^{s/s+2}}{m^{s/s+2}\| \boldsymbol{\Sigma}_{zz}^{-1}\|_2^{s/s+2}(\|\boldsymbol{\Sigma}_{zz}^{-1}\|_2 + \delta_T^{-1}\zeta_T)^{s/s+2}}\right] +$$
$$m^2 \exp\left(-C_0 \frac{T^{s/s+2}}{m^{s/s+2}\|\boldsymbol{\Sigma}_{zz}^{-1}\|_2^{s/s+2}}\right).$$

Therefore,

$$\Pr(\|\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}\|_F > \delta_T^{-1}\zeta_T)$$
$$\le m \exp[-\ominus (T^{\max\{1-2d+2(\lambda-\alpha),1-2d+\lambda-\alpha,1-2d\}})] +$$
$$m \exp[-\ominus (T^{1-2d})],$$

or,

$$\Pr(\|\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}\|_F > \delta_T^{-1}\zeta_T)$$
$$\le m \exp[-\ominus (T^{s(\max\{1-d+\lambda-\alpha,1-d\})/(s+2)})] +$$
$$m \exp[-\ominus (T^{s(1-d)/(s+2)})].$$

Setting $d < 1/2$, $\alpha = 1/2$, and $\lambda > d$, we have all the terms going to zero as $T \to \infty$ and there exist some finite positive constants C_1 and C_2 such that

$$\kappa_{T,i}(\lambda, d) \le \exp\left(-C_1 T^{C_2}\right), \ \kappa_{T,i}(\alpha, d) \le \exp\left(-C_1 T^{C_2}\right),$$

and

$$\Pr(\|\hat{\boldsymbol{\Sigma}}_{zz}^{-1} - \boldsymbol{\Sigma}_{zz}^{-1}\|_F > \delta_T^{-1}\zeta_T) \le \exp(-C_1 T^{C_2}).$$

Hence, if $d < \lambda \le (s+2)/(s+4)$,

$$\Pr(|\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_j - \mathbb{E}(\boldsymbol{\nu}_i'\boldsymbol{\nu}_j)| > \zeta_T) \le \exp(-C_0 T^{-1}\zeta_T^2) + \exp(-C_1 T^{C_2}),$$

and if $\lambda > (s+2)/(s+4)$,

$$\Pr(|\mathbf{x}_i'\mathbf{M}_z\mathbf{x}_j - \mathbb{E}(\boldsymbol{\nu}_i'\boldsymbol{\nu}_j)| > \zeta_T) \le \exp(-C_0 \zeta_T^{s/(s+1)}) + \exp(-C_1 T^{C_2}),$$

where C_0, C_1 and C_2 are some finite positive constants. ∎

References

Bailey, N., Pesaran, M. H., and Smith, L. V. (2019). A multiple testing approach to the regularisation of large sample correlation matrices. *Journal of Econometrics*, 208(2): 507-534. https://doi.org/10.1016/j.jeconom.2018.10.006

Chudik, A., Kapetanios, G., and Pesaran, M. H. (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica*, 86(4): 1479-1512. https://doi.org/10.3982/ECTA14176

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1-22. https://doi.org/10.18637/jss.v033.i01

Lütkepohl, H. (1996). *Handbook of Matrices*. John Wiley & Sons, West Sussex, UK. ISBN-10: 9780471970156

Online Empirical Supplement to
"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

This online empirical supplement has four sections. Section S-1 provides the full list and description of technical indicators considered in the stock market application. Section S-2 provides the list of variables in the conditioning and active sets in the application on forecasting output growth rates across 33 countries. Section S-3 explains the algorithms used for implementing Lasso, Adaptive Lasso and Cross-validation. Last section focuses on the third application, forecasting Euro Area quarterly output growth using the European Central Bank (ECB) survey of professional forecasters. The section starts with description of the data and then discusses the results.

S-1   Technical and financial indicators

Our choice of the technical trading indicators is based on the extensive literature on system trading, reviewed by [27] and [13]. Most of the technical indicators are based on historical daily high, low and adjusted close prices, which we denote by $H_{it}(\tau)$, $L_{it}(\tau)$, and $P_{it}(\tau)$, respectively. These prices refer to stock i in month t, for day τ. Moreover, let D_t^i be the number of trading days, and denote by $D_{l_t}^i$ the last trading day of stock i in month t. For each stock i, monthly high, low and close prices are set to the last trading day of the month, namely $H_{it}(D_{l_t}^i)$, $L_{it}(D_{l_t}^i)$ and $P_{it}(D_{l_t}^i)$, or H_{it}, L_{it}, and P_{it}, for simplicity. The logarithms of these are denoted by h_{it}, l_{it}, and p_{it}, respectively.

The 28 stocks considered in our study are allocated to 19 sectoral groups according to

Industry Classification Benchmark.[6] The group membership of stock i is denoted by the set $\mathfrak{g}_i$, which includes all S&P 500 stocks in stock i^{th} group, and $|\mathfrak{g}_i|$ is the number of stocks in the group.

The technical and financial indicators considered are:

1. Return of Stock i (r_{it}): $r_{it} = 100(p_{it} - p_{i,t-1})$.

2. The Group Average Return of Stock i $(\bar{r}_{it}^g)$: $\bar{r}_{it}^g = |\mathfrak{g}_i|^{-1} \sum_{j \in \mathfrak{g}_i} r_{jt}$.

3. Moving Average Stock Return of order s $(mar_{it}(s))$: This indicator, which is also known as s-day momentum (see, for example, 13), is defined as

   $$mar_{it}(s) = \text{MA}(r_{it}, s),$$

   where $\text{MA}(x_{it}, s)$ is Moving Average of a time-series process x_{it} with degree of smoothness s which can be written as

   $$\text{MA}(x_{it}, s) = s^{-1} \sum_{\ell=1}^{s} x_{i,t-\ell}.$$

4. Return Gap $(gr_{it}(s))$: This indicator represents a belief in mean reversion that prices will eventually return to their means (for further details see 13).

   $$gr_{it}(s) = r_{it} - \text{MA}(r_{it}, s).$$

5. Price Gap $(gp_{it}(s))$: $gp_{it}(s) = 100 \left[p_{it} - \text{MA}(p_{it}, s) \right].$

6. Realized Volatility (RV_{it}): $RV_{it} = \sqrt{\sum_{\tau=1}^{D_t^i} \left(R_{it}(\tau) - \bar{R}_{it} \right)^2}$, where

   $$R_{it}(\tau) = 100 \left[P_{it}(\tau)/P_{it}(\tau-1) - 1 \right], \text{ and } \bar{R}_{it} = \sum_{\tau=1}^{D_t^i} R_{it}(\tau)/D_t^i.$$

7. Group Realized Volatility (RV_{it}^g): $RV_{it}^g = \sqrt{|\mathfrak{g}|^{-1} \sum_{i \in \mathfrak{g}} RV_{it}^2}$.

8. Moving Average Realized Volatility $(mav_{it}(s))$: "Signals are generated when a price change is accompanied by an unusually large move relative to average volatility" [13].

---

[6]The 19 groups are as follows: Oil & Gas, Chemicals, Basic Resources, Construction & Materials, Industrial Goods & Services, Automobiles & Parts, Food & Beverage, Personal & Household Goods, Health Care, Retail, Media, Travel & Leisure, Telecommunications, Utilities, Banks, Insurance, Real Estate, Financial Services, and Technology.

The following two indicators are constructed to capture such signals

$$mav_{it}(s) = \text{MA}(RV_{it}, s)$$

9. Realized Volatility Gap $(RVG_{it}(s))$: $RVG_{it}(s) = RV_{it} - \text{MA}(RV_{it}, s)$

10. Percent Price Oscillator$(PPO_{it}(s_1, s_2))$:

$$PPO_{it}(s_1, s_2) = 100 \left(\frac{\text{MA}(P_{it}, s_1) - \text{MA}(P_{it}, s_2)}{\text{MA}(P_{it}, s_2)} \right), \text{ where } s_1 < s_2.$$

11. Relative Strength Indicator (RSI_{it}^s): This is a price momentum indicator developed by [27] to capture overbought and oversold conditions. Let

$$\Delta P_{it}^+ = \Delta P_{it} \text{I}_{\Delta P_{it} > 0}(\Delta P_{it}), \text{ and } \Delta P_{it}^- = \Delta P_{it} \text{I}_{\Delta P_{it} \leq 0}(\Delta P_{it}),$$

where $\Delta P_{it} = P_{it} - P_{i,t-1}$ and $\text{I}_A(x_{it})$ is an indicator function that take a value of one if $x_{it} \in A$ and zero otherwise. Then

$$RS_{it}^s = -\frac{\text{MA}(\Delta P_{it}^+, s)}{\text{MA}(\Delta P_{it}^-, s)}, \text{ and } RSI_{it}^s = 100 \left(1 - \frac{1}{1 + RS_{it}^s} \right).$$

12. Williams R $(WILLR_{it}(s))$: This indicator proposed by [28] to measure buying and selling pressure.

$$WILLR_{it}(s) = -100 \left(\frac{\max\limits_{j \in \{1, \cdots, s\}} (h_{i,t-s+j}) - p_{it}}{\max\limits_{j \in \{1, \cdots, s\}} (h_{i,t-s+j}) - \min\limits_{j \in \{1, \cdots, s\}} (l_{i,t-s+i})} \right).$$

13. Average Directional Movement Index $(ADX_{it}(s))$: This is a filtered momentum indicator by [27]. To compute $ADX_{it}(s)$, we first calculate up-ward directional movement (DM_{it}^+), down-ward directional movement (DM_{it}^-), and true range (TR_{it}) as:

$$DM_{it}^+ = \begin{cases} h_{it} - h_{i,t-1}, & \text{if } h_{it} - h_{i,t-1} > 0 \text{ and } h_{it} - h_{i,t-1} > l_{i,t-1} - l_{it}, \\ 0, & \text{otherwise.} \end{cases}$$

$$DM_{it}^- = \begin{cases} l_{i,t-1} - l_{it}, & if \ l_{i,t-1} - l_{it} > 0 \ and \ l_{i,t-1} - l_{it} > h_{it} - h_{i,t-1}, \\ 0, & otherwise. \end{cases}$$

$$TR_{it} = \max\{h_{it} - l_{it}, |h_{it} - p_{i,t-1}|, |p_{i,t-1} - l_{it}|\}.$$

Then, positive and negative directional indexes denoted by $ID_{it}^+(s)$ and $ID_{it}^-(s)$ re-

spectively, are computed by

$$ID_{it}^+(s) = 100\left(\frac{\mathrm{MA}(DM_{it}^+,s)}{\mathrm{MA}(TR_{it},s)}\right), \text{ and } ID_{it}^-(s) = 100\left(\frac{\mathrm{MA}(DM_{it}^-,s)}{\mathrm{MA}(TR_{it},s)}\right),$$

Finally, directional index $DX_{it}(s)$ and $ADX_{it}(s)$ are computed as

$$DX_{it}(s) = 100\left(\frac{|ID_{it}^+(s) - ID_{it}^-(s)|}{ID_{it}^+(s) + ID_{it}^-(s)}\right), \text{ and } ADX_{it}(s) = \mathrm{MA}(DX_{it}(s),s).$$

14. Percentage Change in Kaufman's Adaptive Moving Average ($\Delta KAMA_{it}(s_1,s_2,m)$): Kaufman's Adaptive Moving Average accounts for market noise or volatility. To compute $\Delta KAMA_{it}(s_1,s_2,m)$, we first need to calculate the Efficiency Ratio (ER_{it}) defined by

$$ER_{it} = 100\left(\frac{|p_{it} - p_{i,t-m}|}{\sum_{j=1}^m |\Delta P_{i,t-m+j}|}\right),$$

where $\Delta P_{it} = P_{it} - P_{i,t-1}$, and then calculate the Smoothing Constant (SC_{it}) which is

$$SC_{it} = \left[ER_{it}\left(\frac{2}{s_1+1} - \frac{2}{s_2+1}\right) + \frac{2}{s_2+1}\right]^2,$$

where $s_1 < m < s_2$. Then, Kaufman's Adaptive Moving Average is computed as

$$\mathrm{KAMA}(P_{it},s_1,s_2,m) = SC_{it}P_{it} + (1 - SC_{it})\mathrm{KAMA}(P_{i,t-1},s_1,s_2,m)$$

where

$$\mathrm{KAMA}(P_{is_2},s_1,s_2,m) = s_2^{-1}\sum_{\kappa=1}^{s_2} P_{i\kappa}.$$

The Percentage Change in Kaufman's Adaptive Moving Average is then computed as

$$\Delta KAMA_{it}(s_1,s_2,m) = 100\left(\frac{\mathrm{KAMA}(P_{it},s_1,s_2,m) - \mathrm{KAMA}(P_{i,t-1},s_1,s_2,m)}{\mathrm{KAMA}(P_{i,t-1},s_1,s_2,m)}\right).$$

For further details see [13].

Other financial indicators

In addition to the above technical indicators, we also make use of daily prices of Brent Crude Oil, S&P 500 index, monthly series on Fama and French market factors, and annualized percentage yield on 3-month, 2-year and 10-year US government bonds. Based on this data,

we have constructed the following variables. These series are denoted by PO_t and $P_{sp,t}$ respectively, and their logs by po_t and $p_{sp,t}$. The list of additional variables are:

1. Return of S&P 500 index ($r_{sp,t}$): $r_{sp,t} = 100(p_{sp,t} - p_{sp,t-1})$, where $p_{sp,t}$ is the log of S&P 500 index at the end of month t.

2. Realized Volatility of S&P 500 index ($RV_{sp,t}$):

$$RV_{sp,t} = \sqrt{\sum_{\tau=1}^{D_t^{sp}} \left(R_{sp,t}(\tau) - \bar{R}_{sp,t}\right)^2},$$

where $\bar{R}_{sp,t} = \sum_{\tau=1}^{D_t^{sp}} R_{it}(\tau)/D_t^{sp}$, $R_{sp,t}(\tau) = 100([P_{sp,t}(\tau)/P_{sp,t}(\tau-1) - 1]$, $P_{sp,t}(\tau)$ is the S&P 500 price index at close of day τ of month t, and D_t^{sp} is the number of days in month t.

3. Percent Rate of Change in Oil Prices (Δpo_t): $\Delta po_t = 100(po_t - po_{t-1})$, where po_t is the log of oil princes at the close of month t.

4. Long Term Interest Rate Spread ($LIRS_t$): The difference between annualized percentage yield on 10-year and 3-month US government bonds.

5. Medium Term Interest Rate Spread ($MIRS_t$): The difference between annualized percentage yield on 10-year and 2-year US government bonds.

6. Short Term Interest Rate Spread ($SIRS_t$): The difference between annualized percentage yield on 2-year and 3-month US government bonds.

7. Small Minus Big Factor (SMB_t): Fama and French Small Minus Big market factor.

8. High Minus Low Factor (HML_t): Fama and French High Minus Low market factor.

A summary of the covariates in the active set used for prediction of monthly stock returns is given in Table S.1.

S-2   List of variables used when forecasting output growths

Variables in the conditioning and active sets for forecasting output growth across 33 countries are listed in Table S.2 below.

Table S.1: Active set for percentage change in equity price forecasting

Target Variable:	r_{it+1} (one-month ahead percentage change in equity price of stock i)
A. Financial Variables:	r_{it}, $\bar{r}_{it}^g$, $r_{sp,t}$, RV_{it}, RV_{it}^g, $RV_{sp,t}$, SMB_t, HML_t.
B. Economic Variables:	Δpo_t, $LIRS_t - LIRS_{t-1}$, $MIRS_t - MIRS_{t-1}$, $SIRS_t - SIRS_{t-1}$.
C. Technical Indicators:	mar_{it}^s for $s = \{3, 6, 12\}$, mav_{it}^s for $s = \{3, 6, 12\}$, gr_{it}^s for $s = \{3, 6, 12\}$, gp_{it}^s for $s = \{3, 6, 12\}$, RVG_{it}^s for $s = \{3, 6, 12\}$, RSI_{it}^s for $s = \{3, 6, 12\}$, ADX_{it}^s for $s = \{3, 6, 12\}$, $WILLR_{it}^s$ for $s = \{3, 6, 12\}$, $PPO_{it}(s_1, s_2)$ for $(s_1, s_2) = \{(3, 6), (6, 12), (3, 12)\}$, $\Delta KAMA_{it}(s_1, s_2, m)$ for $(s_1, s_2, m) = (2, 12, 6)$.

Table S.2: List of variables in the conditioning and active sets for forecasting quarterly output growths across 33 countries

Conditioning set	
c, $\Delta_1 y_{it}$	
Active Set	
(a) Domestic variables, $\ell = 0, 1$.	(b) Foreign counterparts, $\ell = 0, 1$.
$\Delta_1 y_{i,t-1}$	$\Delta_1 y_{i,t-\ell}^*$
$\Delta_1 r_{i,t-\ell} - \Delta_1 \pi_{i,t-\ell}$	$\Delta_1 r_{i,t-\ell}^* - \Delta_1 \pi_{i,t-\ell}^*$
$\Delta_1 r_{i,t-\ell}^L - \Delta_1 r_{i,t-\ell}$	$\Delta_1 r_{i,t-\ell}^{L*} - \Delta_1 r_{i,t-\ell}^*$
$\Delta_1 q_{i,t-\ell} - \Delta_1 \pi_{i,t-\ell}$	$\Delta_1 q_{i,t-\ell}^* - \Delta_1 \pi_{i,t-\ell}^*$
Total number of variables in the active set $\mathbf{x}_t$: $n = 15$ (max)	

S-3   Lasso, Adaptive Lasso and cross-validation algorithms

This section explains how Lasso, K-fold cross-validation and Adaptive Lasso are implemented in this paper. Let $\mathbf{y} = (y_1, y_2, \cdots, y_T)'$ be a $T \times 1$ vector of target variable, and let $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T)'$ be a $T \times m$ matrix of conditioning covariates where $\{\mathbf{z}_t : t = 1, 2, \cdots, T\}$ are $m \times 1$ vectors and let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)'$ be a $T \times N$ matrix of covariates in the active set where $\{\mathbf{x}_t : t = 1, 2, \cdots, T\}$ are $N \times 1$ vectors.

Lasso Procedure

1. *Construct the filtered variables* $\tilde{\mathbf{y}} = \mathbf{M}_z \mathbf{y}$ *and* $\tilde{\mathbf{X}} = \mathbf{M}_z \mathbf{X} = (\tilde{\mathbf{x}}_{1\circ}, \tilde{\mathbf{x}}_{2\circ}, ..., \tilde{\mathbf{x}}_{N\circ})$, *where* $\mathbf{M}_z = \mathbf{I}_T - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, *and* $\tilde{\mathbf{x}}_{i\circ} = (\tilde{x}_{i1}, \tilde{x}_{i2}, \cdots, \tilde{x}_{iT})'$.

2. Normalize each covariate $\tilde{\mathbf{x}}_{io} = (\tilde{x}_{i1}, \tilde{x}_{i2}, \cdots, \tilde{x}_{iT})'$ by its ℓ_2 norm, such that

$$\tilde{\mathbf{x}}_{io}^* = \tilde{\mathbf{x}}_{io}/\|\tilde{\mathbf{x}}_{io}\|_2,$$

where $\|.\|_2$ denotes the ℓ_2 norm of a vector. The corresponding matrix of normalized covariates in the active set is now denoted by $\tilde{\mathbf{X}}^*$.

3. For a given value of $\varphi \geq 0$, find $\hat{\boldsymbol{\gamma}}_x^*(\varphi) \equiv [\hat{\gamma}_{1x}^*(\varphi), \hat{\gamma}_{2x}^*(\varphi), \cdots, \hat{\gamma}_{Nx}^*(\varphi)]'$ such that

$$\hat{\boldsymbol{\gamma}}_x^*(\varphi) = \arg\min_{\boldsymbol{\gamma}_x^*} \left\{ \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}^* \boldsymbol{\gamma}_x^*\|_2^2 + \varphi \|\boldsymbol{\gamma}_x^*\|_1 \right\},$$

where $\|.\|_1$ denotes the ℓ_1 norm of a vector.

4. Divide $\hat{\gamma}_{ix}^*(\varphi)$ for $i = 1, 2, \cdots, N$ by ℓ_2 norm of the $\tilde{\mathbf{x}}_{io}$ to match the original scale of $\tilde{\mathbf{x}}_{io}$, namely set

$$\hat{\gamma}_{ix}(\varphi) = \hat{\gamma}_{ix}^*(\varphi)/\|\tilde{\mathbf{x}}_{io}\|_2,$$

where $\hat{\boldsymbol{\gamma}}_x(\varphi) \equiv [\hat{\gamma}_{1x}(\varphi), \hat{\gamma}_{2x}(\varphi), \cdots, \hat{\gamma}_{Nx}(\varphi)]'$ denotes the vector of scaled coefficients.

5. Compute $\hat{\boldsymbol{\gamma}}_z(\varphi) \equiv [\hat{\gamma}_{1z}(\varphi), \hat{\gamma}_{2z}(\varphi), \cdots, \hat{\gamma}_{mz}(\varphi)]'$ by $\hat{\boldsymbol{\gamma}}_z(\varphi) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}}(\varphi)$ where $\hat{\mathbf{e}}(\varphi) = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\gamma}}_x(\varphi)$.

For a given set of values of φ's, say $\{\varphi_j : j = 1, 2, \cdots, h\}$, the optimal value of φ is chosen by K-fold cross-validation as described below.

K-fold Cross-validation

1. Create a $T \times 1$ vector $\mathbf{w} = (1, 2, \cdots, K, 1, 2, \cdots, K, \cdots)'$ where K is the number of folds.

2. Let $\mathbf{w}^* = (w_1^*, w_2^*, \cdots, w_T^*)'$ be a $T \times 1$ vector generated by randomly permuting the elements of $\mathbf{w}$.

3. Group observations into K folds such that

$$g_k = \{t : t \in \{1, 2, \cdots, T\} \text{ and } w_t^* = k\} \text{ for } k = 1, 2, \cdots, K.$$

4. For a given value of φ_j and each fold $k \in \{1, 2, \cdots, K\}$,

   (a) Remove the observations related to fold k from the set of all observations.

(b) *Given the value of φ_j, use the remaining observations to estimate the coefficients of the model.*

(c) *Use the estimated coefficients to compute predicted values of the target variable for the observations in fold k and hence compute mean square forecast error of fold k denoted by $MSFE_k(\varphi_j)$.*

5. *Compute the average mean square forecast error for a given value of φ_j by*

$$\overline{MSFE}(\varphi_j) = \sum_{k=1}^{K} MSFE_k(\varphi_j)/K.$$

6. *Repeat steps 1 to 5 for all values of $\{\varphi_j : j = 1, 2, \cdots, h\}$.*

7. *Select φ_j with the lowest corresponding average mean square forecast error as the optimal value of φ.*

In this study, following Friedman et al. (2010), we consider a sequence of 100 values of φ's decreasing from $\varphi_{\max}$ to $\varphi_{\min}$ on log scale where $\varphi_{\max} = \max_{i=1,2,\cdots,N} \left\{ |\sum_{t=1}^{T} \tilde{x}_{it}^* \tilde{y}_t| \right\}$ and $\varphi_{\min} = 0.001\varphi_{\max}$. We use 10-fold cross-validation ($K = 10$) to find the optimal value of φ.

Denote $\hat{\boldsymbol{\gamma}}_x \equiv \hat{\boldsymbol{\gamma}}_x(\varphi_{op})$ where φ_{op} is the optimal value of φ obtained by the K-fold cross-validation. Given $\hat{\boldsymbol{\gamma}}_x$, we implement Adaptive Lasso as described below.

Adaptive Lasso

1. *Let $\mathcal{S} = \{i : i \in \{1, 2, \cdots, N\}$ and $\hat{\gamma}_{ix} \neq 0\}$ and $\mathbf{X}_{\mathcal{S}}$ be the $T \times s$ set of covariates in the active set with $\hat{\gamma}_{ix} \neq 0$ (from the Lasso step) where $s = |\mathcal{S}|$. Additionally, denote the corresponding $s \times 1$ vector of non-zero Lasso coefficients by $\hat{\boldsymbol{\gamma}}_{x,\mathcal{S}} = (\hat{\gamma}_{1x,\mathcal{S}}, \hat{\gamma}_{2x,\mathcal{S}}, \cdots, \hat{\gamma}_{sx,\mathcal{S}})'$.*

2. *For a given value of $\psi \geq 0$, find $\hat{\boldsymbol{\delta}}_{x,\mathcal{S}}^*(\psi) \equiv [\hat{\delta}_{1x,\mathcal{S}}^*(\psi), \hat{\delta}_{2x,\mathcal{S}}^*(\psi), \cdots, \hat{\delta}_{sx,\mathcal{S}}^*(\psi)]'$ such that*

$$\hat{\boldsymbol{\delta}}_{x,\mathcal{S}}^*(\psi) = \arg\min_{\boldsymbol{\delta}_{x,\mathcal{S}}^*} \left\{ \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_{\mathcal{S}} diag(\hat{\boldsymbol{\gamma}}_{x,\mathcal{S}}) \boldsymbol{\delta}_{x,\mathcal{S}}^*\|_2^2 + \psi \|\boldsymbol{\delta}_{x,\mathcal{S}}^*\|_1 \right\},$$

where $diag(\hat{\boldsymbol{\gamma}}_{x,\mathcal{S}})$ is an $s \times s$ diagonal matrix with its diagonal elements given by the corresponding elements of $\hat{\boldsymbol{\gamma}}_{x,\mathcal{S}}$.

3. *Post multiply $\hat{\boldsymbol{\delta}}_{x,\mathcal{S}}^*(\psi)$ by $diag(\hat{\boldsymbol{\gamma}}_{x,\mathcal{S}})$ to match the original scale of $\tilde{\mathbf{X}}_{\mathcal{S}}$, such that*

$$\hat{\boldsymbol{\delta}}_{x,\mathcal{S}}(\psi) = diag(\hat{\boldsymbol{\gamma}}_{x,\mathcal{S}}) \hat{\boldsymbol{\delta}}_{x,\mathcal{S}}^*(\psi).$$

S.8

The coefficients of the covariates in the active set that belong to $\mathcal{S}^c$ are set equal to zero. In other words, $\hat{\boldsymbol{\delta}}_{x,\mathcal{S}^c}(\psi) = 0$ for all $\psi \geq 0$.

4. *Compute $\hat{\boldsymbol{\delta}}_z(\psi) \equiv [\hat{\delta}_{1z}(\psi), \hat{\delta}_{2z}(\psi), \cdots, \hat{\delta}_{mz}(\psi)]'$ by $\hat{\boldsymbol{\delta}}_z(\psi) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{e}}(\psi)$ where $\hat{\mathbf{e}}(\psi) = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}_{\mathcal{S}}\hat{\boldsymbol{\delta}}_{x,\mathcal{S}}(\psi)$.*

As in the Lasso step, the optimal value ψ is set using 10-fold cross-validation as described before.[7]

S-4   Forecasting euro area output growths using ECB surveys of professional forecasters

This application considers forecasting one-year ahead Euro Area real output growth using the ECB survey of professional forecasters, recently analyzed by [7]. The dataset consists of quarterly predictions of 25 professional forecasters over the period 1999Q3 to 2014Q1.[8] The predictions of these forecasters are highly correlated suggesting the presence of a common factor across these forecasts. To deal with this issue at the variable selection stage following [24] we also include the simple average of the 25 forecasts in the conditioning set, $\mathbf{z}_t$, as a proxy for the common factor in addition to the intercept. We consider 39 quarterly forecasts (from 2004Q3 and 2014Q1) for forecast evaluation, using expanding samples (weighted and unweighted) from 1999Q3. We also consider two simple baseline forecasts: a simple cross sectional (CS) average of the professional forecasts, and forecasts computed using a regression of output growths on an intercept and the CS average of the professional forecasts.

Table S.3 compares the forecast performance of OCMT with and without down-weighting at the selection and forecasting stages, in terms of MSFE. The results suggest that down-weighting at the selection stage leaves us with larger forecasting errors. The MSFE goes from 3.765 (3.995) to 3.874 (4.672) in case of light (heavy) down-weighting. However, the panel DM tests indicate that the MSFE among different scenarios are not statistically significant, possibly due to the short samples being considered. In Table S.4, we compare OCMT (with no down-weighting at the selection stage) with Lasso and Adaptive Lasso. The results indicate that the OCMT procedure outperforms Lasso and Adaptive Lasso in terms of MSFE when using no down-weighting, light down-weighting, and heavy down-weighting at the forecasting stage. It is worth mentioning that OCMT selects 3 forecasters (Forecaster #21

---

[7]To implement Lasso, Adaptive Lasso and 10-fold cross-validation we take advantage of glmnet package (Matlab version) available at `http://web.stanford.edu/~hastie/glmnet_matlab/`

[8]We are grateful to Frank Diebold for providing us with the data set.

for 2004Q4-2005Q1, Forecaster #7 for 2007Q2-2008Q3, and Forecaster #18 for 2011Q2-2011Q3). This means that over the full evaluating sample, only 0.3 variables are selected by OCMT from the active set on average. In contrast, Lasso selects 12.6 forecasters on average. Each individual forecaster is selected for at least part of the evaluation period. As to be expected, Adaptive Lasso selects a fewer number of forecasters (9.8 on average) as compared to Lasso (12.6 on average), and performs slightly worse.

To summarize, we find that down-weighting at the selection stage of OCMT leads to forecast deterioration (in terms of MSFE). OCMT outperforms Lasso and Adaptive Lasso, but the panel DM tests are not statistically significant. Moreover, none of the considered big data methods can beat the simple baseline models.

Table S.3: Mean square forecast error (MSFE) and panel DM test of OCMT of one-year ahead Euro Area annual real output growth forecasts between 2004Q3 and 2014Q1 (39 forecasts)

	Down-weighting at[†]		MSFE
	Selection stage	Forecasting stage	
(M1)	no	no	3.507
Light down-weighting, $\lambda = \{0.975, 0.98, 0.985, 0.99, 0.995, 1\}$			
(M2)	no	yes	3.765
(M3)	yes	yes	3.874
Heavy down-weighting, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$			
(M4)	no	yes	3.995
(M5)	yes	yes	4.672

Pair-wise panel DM tests					
	Light down-weighting			Heavy down-weighting	
	(M2)	(M3)		(M4)	(M5)
(M1)	-0.737	-0.474	(M1)	-0.656	-0.741
(M2)	-	-0.187	(M5)	-	-0.645

Notes: The active set consists of 25 individual forecasts. The conditioning set consists of an intercept and the cross sectional average of 25 forecasts.
[†]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 S.4: Mean square forecast error (MSFE) and panel DM test of OCMT versus Lasso and Adaptive Lasso of one-year ahead Euro Area annual real output growth forecasts between 2004Q3 and 2014Q1 (39 forecasts)

	MSFE under different down-weighting scenarios		
	No down-weighting	Light down-weighting[†]	Heavy down-weighting[‡]
OCMT	3.507	3.765	3.995
Lasso	5.242	5.116	5.385
A-Lasso	7.559	6.475	6.539
	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	-1.413	-1.544	-0.990	-1.265	-1.070	-1.267
Lasso	-	-1.484	-	-1.589	-	-1.527

Notes: The active set consists of forecasts by 25 individual forecasters. The conditioning set contains an intercept and the cross sectional average of the 25 forecasts.
[†] 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, $\lambda = \{0.95, 0.96, 0.97, 0.98, 0.99, 1\}$.