Federal Reserve Bank of Dallas
Globalization and Monetary Policy Institute
Working Paper No. 290
http://www.dallasfed.org/assets/documents/institute/wpapers/2016/0290.pdf

A One-Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models*

A. Chudik
Federal Reserve Bank of Dallas

G. Kapetanios
King's College, London

M. Hashem Pesaran
Department of Economics & USC Dornsife INET, University of Southern California, USA
and Trinity College, Cambridge, UK

November 2016

Abstract

Model specification and selection are recurring themes in econometric analysis. Both topics become considerably more complicated in the case of large-dimensional data sets where the set of specification possibilities can become quite large. In the context of linear regression models, penalised regression has become the de facto benchmark technique used to trade off parsimony and fit when the number of possible covariates is large, often much larger than the number of available observations. However, issues such as the choice of a penalty function and tuning parameters associated with the use of penalized regressions remain contentious. In this paper, we provide an alternative approach that considers the statistical significance of the individual covariates one at a time, whilst taking full account of the multiple testing nature of the inferential problem involved. We refer to the proposed method as One Covariate at a Time Multiple Testing (OCMT) procedure. The OCMT provides an alternative to penalised regression methods: It is based on statistical inference and is therefore easier to interpret and relate to the classical statistical analysis, it allows working under more general assumptions, it is faster, and performs well in small samples for almost all of the different sets of experiments considered in this paper. We provide extensive theoretical and Monte Carlo results in support of adding the proposed OCMT model selection procedure to the toolbox of applied researchers. The usefulness of OCMT is also illustrated by an empirical application to forecasting U.S. output growth and inflation.

JEL codes: C52, C55

* Alexander Chudik, Research Department, Federal Reserve Bank of Dallas, 2200 N. Pearl Street, Dallas, TX 75201. 214-922-5769. alexander.chudik@dal.frb.org. G. Kapetanios, King's College London, 1.56 Franklin-Wilkins Building, Waterloo, United Kingdom. 020-78828698. george.kapetanios@kcl.ac.uk. M. Hashem Pesaran, Department of Economics, University of Southern California, 3620 South Vermont Avenue, Kaprielian Hall 300, Los Angeles, CA 90089-0253. pesaran@usc.edu. We have benefited from helpful comments by anonymous referees, Jinchi Lv and Yingying Fan, and seminar participants at the Bank of England, Federal Reserve Bank of Dallas, Third IAAE Conference at the University of Milano-Bicocca, Michigan State University, University of Kent, University of Messina, University of Michigan, and University of Warwick. Earlier version of this paper was circulated under the title "Big Data Analytics: A New Perspective." The views in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Dallas or the Federal Reserve System.

1 Introduction

The problem of correctly specifying a model has been a recurring theme in econometrics. There are a number of competing approaches such as those based on specification testing or the use of information criteria that have been exhaustively analysed in a, hitherto, standard framework where the number of observations is considerably larger than the number of potential model candidates.

However, recently, increased focus has been placed on settings where the latter number is either similar or exceeds the number of observations. Model selection and estimation in a high-dimensional regression setting has largely settled around a set of methods collectively known as penalised (or regularised) regression. Penalised regression is an extension of multiple regression where the vector of regression coefficients, $\boldsymbol{\beta}$, of a regression of y_t on $\boldsymbol{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$ is estimated by $\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta}[\sum_{t=1}^{T}(y_t - \boldsymbol{x}'_{nt}\boldsymbol{\beta})^2 + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})]$. $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ is a penalty function that penalises the complexity of $\boldsymbol{\beta}$, while $\boldsymbol{\lambda}$ is a vector of tuning parameters to be set by the researcher. A wide variety of penalty functions have been considered in the literature, yielding a wide range of penalised regression methods. Chief among them is Lasso, where $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ is chosen to be proportional to the L_1 norm of $\boldsymbol{\beta}$. This has subsequently been generalised to the analysis of functions involving L_q, $0 \leq q \leq 2$, norms. While these techniques have found considerable use in econometrics[1], their theoretical properties have been mainly analysed in the statistical literature starting with the seminal work of Tibshirani (1996) and followed up with important contributions by Zhou and Hastie (2005), Lv and Fan (2009), Efron, Hastie, Johnstone, and Tibshirani (2004), Bickel, Ritov, and Tsybakov (2009), Candes and Tao (2007), Zhang (2010), Fan and Li (2001), Antoniadis and Fan (2001), Fan and Lv (2013) and Fan and Tang (2013). Despite considerable advances made in the theory and practice of penalised regressions, there are still a number of open questions. These include the choice of the penalty function and tuning parameters. The latter seems particularly crucial given the fact that no fully satisfactory method has, hitherto, been proposed in the literature, and the tuning parameters are typically chosen by cross validation. A number of contributions, notably by Fan and Li (2001) and Zhang (2010), have considered the use of nonconvex penalty functions with some success. However, the use of nonconvex penalties introduce numerical challenges and can be unstable and time consuming to implement.

As an alternative to penalised regression, a number of researchers have developed methods that focus on the predictive power of individual regressors instead of considering all the n covariates together. This has led to a variety of alternative specification methods sometimes referred to collectively as "greedy methods". In such settings, regressors are chosen sequentially based on their individual ability to explain the dependent variable. Perhaps the most

[1] A general discussion of high-dimensional data and their use in microeconomic analysis can be found in Belloni, Chernozhukov, and Hansen (2014a).

widely known of such methods, developed in the machine learning literature, is "boosting" whose statistical properties have received considerable attention (Friedman, Hastie, and Tibshirani (2000), Friedman (2001) and Buhlmann (2006)). Other machine learning approaches, such as regression trees, and step-wise regressions, are also widely used, but they lack rigorous theoretical underpinnings.

A further approach that has a number of common elements with our proposal and combines penalised regression with greedy methods is sure screening. It has been put forward by Fan and Lv (2008), and, independently by Huang, J. Horowitz, and Ma (2008), and analysed further by Fan and Song (2010) and Fan, Samworth, and Wu (2009), among others. This approach considers marginal correlations between each of the potential regressors and y_t, and selects either a fixed proportion of the regressors based on a ranking of the absolute correlations, or those regressors whose absolute correlation with y_t exceeds a threshold. The latter variant requires selecting a threshold and so the former variant is used in practice. As this approach is mainly an initial screening device, it may select too many regressors but enables dimension reduction in the case of ultra large datasets. As a result, a second step is usually considered where penalised regression is applied to the regressors selected at the first stage.

The present paper contributes to this general specification literature by proposing a new model selection approach for high-dimensional datasets. The main idea is to test the statistical significance of the net contribution of each potential covariate to y_t separately, whilst taking full and rigorous account of the multiple testing nature of the problem under consideration. The general case requires iterating this process by testing the statistical contribution of covariates that have not been previously selected (again one at a time) to the unexplained part of y_t. In a final step, all statistically significant covariates are included as joint determinants of y_t in a multiple regression setting. Whilst the initial regressions of our procedure are common to boosting and to the screening approach of Fan and Lv (2008), the multiple testing and iterative elements provide a powerful stopping rule without needing to resort to model selection or penalised regression subsequently.

We use ideas from the multiple testing literature to control the probability of selecting the true model, the false positive rate and the false discovery rate. We refer to the proposed method as One Covariate at a Time Multiple Testing (OCMT) procedure. In addition to its theoretical properties which we shall discuss below, OCMT is computationally simple and fast even for extremely large datasets. The method provides an alternative in selecting regressors that are correlated with the true unknown conditional mean of the target variable and, as a result, it also has good estimation properties for the unknown coefficient vector. Like penalised regressions, the proposed method is applicable when the underlying regression model is sparse. Further, it does not require the $\boldsymbol{x}_{nt}$ to have a sparse covariance matrix, and is applicable even if the covariance matrix of the noise variables (to be defined below) is not sparse. Of course, since OCMT is a model selection device, well known impossibility results for the uniform validity of post-selection estimators, such as those obtained in Fan and Pötscher (2006) and Fan and Pötscher (2008), apply.

We provide theoretical results for the proposed OCMT procedure under mild assumptions. In particular, we do not assume either a fixed design or time series independence for $\boldsymbol{x}_{nt}$ but consider a martingale difference condition. While the martingale difference condition is our maintained assumption, we also provide theoretical arguments that allow the covariates to follow mixing processes. We report results on the true positive rate, the false positive rate, the false discovery rate, and the norms of the coefficient estimate as well as the in-sample regression error. We do not report any optimality results for our method. Further, we compare the small sample properties of our proposed method with three penalised regressions and boosting techniques using a large number of Monte Carlo experiments under different data generating schemes, and obtain encouraging results.

The paper is structured as follows: Section 2 provides the setup of the problem. Section 3 introduces the new method. Its theoretical and small sample properties are analysed in Sections 4 and 5, respectively. Section 6 presents a forecasting empirical illustration of the proposed method. Section 7 concludes and technical proofs are relegated to appendices. Two online supplements provide additional theoretical results, a complete set of Monte Carlo results for all the experiments conducted, and additional empirical findings.

Notations: Generic positive finite constants are denoted by C_i for $i = 0, 1, 2, ...$. They can take different values at different instances. Let $\mathbf{a} = (a_1, a_2, ..., a_n)'$ and $\mathbf{A} = (a_{ij})$ be an $n \times 1$ vector and an $n \times m$ matrix, respectively. Then, $\|\mathbf{a}\| = (\Sigma_{i=1}^n a_i^2)^{1/2}$ and $\|\mathbf{a}\|_1 = \Sigma_{i=1}^n |a_i|$ are the Euclidean (L_2) norm and L_1 norm of $\mathbf{a}$, respectively. $\|\boldsymbol{A}\|_F = [Tr(\boldsymbol{A}\boldsymbol{A}')]^{1/2}$ is the Frobenius norm of $\mathbf{A}$. $\boldsymbol{\tau}_T$ is a $T \times 1$ vector of ones, $\boldsymbol{\tau}_T = (1, 1, ..., 1)'$. If $\{f_n\}_{n=1}^{\infty}$ is any real sequence and $\{g_n\}_{n=1}^{\infty}$ is a sequences of positive real numbers, then $f_n = O(g_n)$, if there exists a positive finite constant C_0 such that $|f_n|/g_n \leq C_0$ for all n. $f_n = o(g_n)$ 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 exists $N_0 \geq 1$ and positive finite constants C_0 and C_1, such that $\inf_{n \geq N_0}(f_n/g_n) \geq C_0$, and $\sup_{n \geq N_0}(f_n/g_n) \leq C_1$. $\to_p$ denotes convergence in probability as $n, T \to \infty$.

2 The Variable Selection Problem

Suppose that the target variable, y_t, is generated from the following data generating process (DGP)

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

where u_t is an error term whose properties will be specified below, and $0 < |\beta_i| \leq C < \infty$, for $i = 1, 2, ..., k$, $k > 0$ is fixed. In matrix notation, we have

$$\boldsymbol{y} = a\boldsymbol{\tau}_T + \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{u}, \tag{2}$$

where $\boldsymbol{\tau}_T$ is a $T \times 1$ vector of ones, $\boldsymbol{X}_k = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)$ is the $T \times k$ matrix of observations on the covariates, $\boldsymbol{\beta}_k = (\beta_1, \beta_2, ..., \beta_k)'$ is the $k \times 1$ vector of associated slope coefficients and $\boldsymbol{u} = (u_1, u_2, ..., u_T)'$ is $T \times 1$ vector of errors.

The identity of the covariates, x_{it}, for $i = 1, 2, ..., k$, also referred to as the "signal" variables, is not known to the investigator who faces the task of identifying them from a large set of n covariates, denoted as $\mathcal{S}_{nt} = \{x_{it}, i = 1, 2, ..., n\}$, with n being potentially larger than T. We assume that the signal variables x_{it}, for $i = 1, 2, ..., k$, belong to $\mathcal{S}_{nt}$, and without loss of generality suppose that they are arranged as the first k variables of $\mathcal{S}_{nt}$. We refer to the remaining $n-k$ regressors in $\mathcal{S}_{nt}$ as 'noise' variables, defined by $\beta_i = 0$ for $i = k+1, k+2, ..., n$. In addition to the constant term, other deterministic terms can also be easily incorporated in (1), without any significant complications. It is further assumed that the following exact sparsity condition holds: $\sum_{i=1}^{n} I\left(\beta_i \neq 0\right) = k$, where k is bounded but otherwise unknown, and $I\left(A\right)$ is an indicator function which takes the value of unity if A holds and zero otherwise. In the presence of n potential covariates, the DGP can be written equivalently as

$$y_t = a + \sum_{i=1}^{n} I(\beta_i \neq 0)\beta_i x_{it} + u_t. \tag{3}$$

Our variable selection approach focusses on the overall or net impact of x_{it} (if any) on y_t rather than the marginal effects defined by $I(\beta_i \neq 0)\beta_i$. As noted by Pesaran and Smith (2014), the mean net impact of x_{it} on y_t is given by

$$\theta_{i,T} = \sum_{j=1}^{n} I(\beta_j \neq 0)\beta_j \sigma_{ij,T} = \sum_{j=1}^{k} \beta_j \sigma_{ij,T}, \tag{4}$$

where $\sigma_{ij,T} = E\left(T^{-1}\boldsymbol{x}_i'\boldsymbol{M}_\tau \boldsymbol{x}_j\right)$, and $\boldsymbol{M}_\tau = \boldsymbol{I}_T - \boldsymbol{\tau}_T\boldsymbol{\tau}_T'/T$. To simplify the notations we suppress the T subscript and use θ_i and σ_{ij} below. The parameter θ_i plays a crucial role in our proposed approach. Ideally, we would like to be able to base our selection decision directly on β_i and its estimate. But when n is large such a strategy is not feasible. Instead, we propose to base inference on θ_i and then decide if such an inference can help in deciding whether or not $\beta_i = 0$. It is important to stress that knowing θ_i does not imply we can determine β_i. But it is possible to identify conditions under which knowing $\theta_i = 0$ or $\theta_i \neq 0$ will help identify whether $\beta_i = 0$ or not. Due to the correlation between variables, nonzero β_i does not necessarily imply nonzero θ_i and we have the following four possibilities:

	$\theta_i \neq 0$	$\theta_i = 0$
$\beta_i \neq 0$	(I) Signal net effect is nonzero	(II) Signal net effect is zero
$\beta_i = 0$	(III) Noise net effect is nonzero	(IV) Noise net effect is zero

.

The first and the last case where $\theta_i \neq 0$ if and only if $\beta_i \neq 0$ is ideal. But there is also a possibility of the second case where $\theta_i = 0$ and $\beta_i \neq 0$ and the third case where $\theta_i \neq 0$ and $\beta_i = 0$. These cases will also be considered in our analysis. The specificity of zero signal net effects (case II) makes it somewhat less plausible than the other scenario, since it requires that $\beta_i = -\sum_{j=1, j\neq i}^{k} \beta_j \sigma_{ii}^{-1} \sigma_{ij}$. On the other hand, the third case of noise variables with nonzero net effect is quite likely.

For future reference we also define a conditional net impact coefficient

$$\theta_{i,T}(\boldsymbol{z}) = \sum_{j=1}^{k} \beta_j \sigma_{ij,T}(\boldsymbol{z}), \tag{5}$$

where $\sigma_{ij,T}(\boldsymbol{z}) = E\left(T^{-1}\boldsymbol{x}_i'\boldsymbol{M}_z\boldsymbol{x}_j\right)$, $\boldsymbol{M}_z = \boldsymbol{I}_T - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_T)'$, and $\boldsymbol{z}_t$ is a vector of variables that includes the constant and a subset of $\mathcal{S}_{nt}$. We suppress the T subscript and use $\theta_i(\boldsymbol{z})$ and $\sigma_{ij}(\boldsymbol{z})$ below. For the noise variables, we require their net effects on the target variable to be controlled, which can be formalized by imposing bounds on $\sum_{j=k+1}^{n} |\theta_j|$. Such bounds can be specified in different ways. The first and main assumption is that there exist possibly a further k^* variables which have $\beta_i = 0$ but are correlated with the signals. We shall refer to them as "pseudo-signal" variables since they are correlated with the signal variables and can be mistaken as possible determinants of y_t. Without loss of generality, these will be ordered so as to follow the k signal variables, so that the first $k + k^*$ variables in $\mathcal{S}_{nt}$ are signal/pseudo-signal variables. We define $\boldsymbol{X}_{k^*}^* = (\mathbf{x}_{k+1}, \mathbf{x}_{k+2}, ..., \mathbf{x}_{k+k^*})$. The remaining $n - k - k^*$ variables will be assumed to have $\theta_i = 0$ and be uncorrelated with the signals. They will be referred to as "pure noise" or simply "noise" variables. We assume that k is an unknown fixed constant, but allow k^* to rise with n such that $k^*/n \to 0$, and $k^*/T \to 0$, at a sufficiently slow rate. Specifically, we allow $k^* = \ominus(n^\epsilon)$ for some appropriately bounded $\epsilon \geq 0$. We expect ϵ to be small when the correlation between the signal variables and the remaining covariates is sparse. In future discussions, we shall refer to the set of models that contain the true signal variables as well as one or more of the pseudo-signal variables as the pseudo-true model. We make the following assumption concerning the signal and pseudo-signal variables.

Assumption 1 *Let $\boldsymbol{X}_{k,k^*} = (\boldsymbol{X}_k, \boldsymbol{X}_{k^*}^*)$, where $\boldsymbol{X}_k = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)$, and $\boldsymbol{X}_{k^*}^* = (\mathbf{x}_{k+1}, \mathbf{x}_{k+2}, ..., \mathbf{x}_{k+k^*})$ are $T \times k$ and $T \times k^*$ observation matrices on signal and noise variables, and suppose that there exists T_0 such that for all $T > T_0$, $\left(T^{-1}\boldsymbol{X}_{k,k^*}'\boldsymbol{X}_{k,k^*}\right)^{-1}$ is nonsingular with its smallest eigenvalue uniformly bounded away from 0, and $\boldsymbol{\Sigma}_{k,k^*} = E\left(T^{-1}\boldsymbol{X}_{k,k^*}'\boldsymbol{X}_{k,k^*}\right)$ is nonsingular for all T.*

Our secondary maintained assumptions are somewhat more general and, accordingly, lead to fewer and weaker results. A first specification assumes that there exists an ordering (possibly unknown) such that $\theta_i = C_i \varrho^i$, $|\varrho| < 1$, $i = 1, 2, ..., n$. A second specification modifies the decay rate and assumes that $\theta_i = C_i i^{-\gamma}$, for some $\gamma > 0$. In both specifications $\max_{1\leq i\leq n} |C_i| < C < \infty$. These specifications allow for various decays in the way noise variables are correlated with the signals. These cases are of technical interest and cover the autoregressive type designs considered in the literature in order to model the correlations across the covariates. See, for example, Zhang (2010) and Belloni, Chernozhukov, and Hansen (2014b).

3 An Iterated Multiple Testing Approach

The standard approach to dealing with the problem of identifying the signal variables from the noise variables is to use penalised regression techniques such as the Lasso. In what follows, we propose an alternative iterative approach which is inspired by the multiple testing literature, although here we focus on controlling the probability of selecting the true model, the false positive rate and the false discovery rate, rather than controlling the size of the union of the multiple tests that are being carried out. We refer to this procedure as One Covariate at a Time Multiple Testing (OCMT). The need for an iterative scheme arises due to the possibility of hidden signal discussed in the previous section that arises when $\theta_i = 0$ even though $\beta_i \neq 0$. We call such signal variables hidden signals.

Suppose we have T observations on y_t and the n covariates, x_{it}, for $i = 1, 2, ..., n$; $t = 1, 2, ..., T$. In the first stage we consider the n bivariate regressions of y_t on a constant and x_{it}, for $i = 1, 2, ..., n$,

$$y_t = c_{i,(1)} + \phi_{i,(1)} x_{it} + e_{it,(1)},\ t = 1, 2, ..., T, \tag{6}$$

where $\phi_{i,(1)} = \theta_i/\sigma_{ii}$ and θ_i is defined in (4). Denote the t-ratio of $\phi_{i,(1)}$ in this regression by $t_{\hat{\phi}_{T,i,(1)}}$, and note that

$$t_{\hat{\phi}_{i,(1)}} = \frac{\hat{\phi}_{T,i,(1)}}{s.e.\left(\hat{\phi}_{T,i,(1)}\right)} = \frac{T^{-1/2} \boldsymbol{x}_i' \boldsymbol{M}_{(0)} \boldsymbol{y}}{\hat{\sigma}_{i,(1)} \sqrt{\boldsymbol{x}_i' \boldsymbol{M}_{(0)} \boldsymbol{x}_i}}, \tag{7}$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iT})'$, $\boldsymbol{y} = (y_1, y_2, ..., y_T)'$, $\hat{\phi}_{T,i,(1)} = \left(\boldsymbol{x}_i' \boldsymbol{M}_{(0)} \boldsymbol{x}_i\right)^{-1} \boldsymbol{x}_i' \boldsymbol{M}_{(0)} \boldsymbol{y}$, $\hat{\sigma}^2_{i,(1)} = \boldsymbol{e}'_{i,(1)} \boldsymbol{e}_{i,(1)}/T$, $\boldsymbol{e}_{i,(1)} = \boldsymbol{M}_{i,(0)} \boldsymbol{y}$, $\boldsymbol{M}_{i,(0)} = \boldsymbol{I}_T - \boldsymbol{X}_{i,(0)} (\boldsymbol{X}'_{i,(0)} \boldsymbol{X}_{i,(0)})^{-1} \boldsymbol{X}'_{i,(0)}$, $\boldsymbol{X}_{i,(0)} = (\boldsymbol{x}_i, \boldsymbol{\tau}_T)$, $\boldsymbol{M}_{(0)} = \boldsymbol{I}_T - \boldsymbol{\tau}_T \boldsymbol{\tau}'_T/T$, and $\boldsymbol{\tau}_T$ is a $T \times 1$ vector of ones. $\hat{\phi}_{T,i,(1)}$ denotes the OLS estimator of $\phi_{i,(1)}$. In future, if there is no confusion we will suppress the T subscript to simplify notation. The first stage multiple testing estimator of $I(\beta_i \neq 0)$ is given by $\widehat{I_{(1)}(\beta_i \neq 0)} = I\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n, \delta)\right]$, for $i = 1, 2, ..., n$, where $c_p(n, \delta)$ is a 'critical value function' defined by

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

where $\Phi^{-1}(.)$ is the inverse of standard normal distribution function, $f(n, \delta) = cn^{\delta}$ for some positive constants δ and c, and p $(0 < p < 1)$ is the nominal size of the individual tests to be set by the investigator. We will refer to δ as the critical value exponent.

The choice of the critical value function, $c_p(n,\delta)$, is important since it allows the investigator to relate the size and power of the selection procedure to the inferential problem in classical statistics, with the modification that p (type I error) is now scaled by a function of the number of covariates under consideration. As we shall see, the OCMT procedure applies irrespective of whether n is small or large relative to T, so long as $T = \ominus(n^{\kappa_1})$, for any finite $\kappa_1 > 0$. This follows from result (i) of Lemma 2, which establishes that $c_p^2(n,\delta) = O[\delta \ln(n)]$. It is also helpful to bear in mind that, using (ii) of Lemma 2,

$$\exp\left[-\frac{\varkappa c_p^2(n,\delta)}{2}\right] = \ominus\left(n^{-\delta\varkappa}\right), \tag{9}$$

and $c_p(n,\delta) = o\left(T^{C_0}\right)$, for all $C_0 > 0$, assuming there exists $\kappa_1 > 0$, such that $T = \ominus(n^{\kappa_1})$.

If other deterministic terms, besides the constant, were considered they could be included in the definition of the orthogonal projection matrix $\boldsymbol{M}_{(0)}$ that filters out these effects. Similarly, if some variables were *a priori* known to be signals, then they could also be included in the definition of $\boldsymbol{M}_{(0)}$. The multiple testing method can easily accommodate both possibilities, while alternative approaches, such as Lasso, may not readily allow for such conditioning.

Covariates for which $I_{(1)}\widehat{(\beta_i \neq 0)} = 1$ are selected as signals or pseudo-signals. Denote the number of variables selected in the first stage by $\hat{k}^o_{n,T,(1)}$, the index set of the selected variables by $\mathcal{S}^o_{(1)}$, and the $T \times \hat{k}^o_{n,T,(1)}$ observation matrix of the $\hat{k}^o_{n,T,(1)}$ selected variables by $\boldsymbol{X}^o_{(1)}$. Further, let $\boldsymbol{X}_{(1)} = (\boldsymbol{\tau}_T, \boldsymbol{X}^o_{(1)}) = (\boldsymbol{x}_{(1),1}, ..., \boldsymbol{x}_{(1),T})'$, $\hat{k}_{n,T,(1)} = \hat{k}^o_{n,T,(1)}$, $\mathcal{S}_{(1)} = \mathcal{S}^o_{(1)}$ and $\mathcal{N}_{(1)} = \{1,2,...,n\} \setminus \mathcal{S}_{(1)}$. In stages $j = 2,3,...$, we consider the $n - \hat{k}_{n,T,(j-1)}$ regressions of y_t on the variables in $\boldsymbol{X}_{(j-1)}$ and, one at the time, x_{it} for $i \in \mathcal{N}_{(j-1)}$. We then compute the following t-ratios

$$t_{\hat{\phi}_{T,i,(j)}} = \frac{\hat{\phi}_{T,i,(j)}}{s.e.\left(\hat{\phi}_{T,i,(j)}\right)} = \frac{\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\mathbf{y}}{\hat{\sigma}_{i,(j)}\sqrt{\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{x}_i}}, \text{ for } i \in \mathcal{N}_{(j-1)},\ j = 2,3,..., \tag{10}$$

where $\hat{\phi}_{T,i,(j)} = \hat{\phi}_{i,(j)} = \left(\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{x}_i\right)^{-1}\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{y}$, denotes the estimated conditional net effect of x_{it} on y_t in stage j, $\hat{\sigma}^2_{i,(j)} = T^{-1}\boldsymbol{e}'_{i,(j)}\boldsymbol{e}_{i,(j)}$, $\boldsymbol{M}_{(j-1)} = \boldsymbol{I}_T - \boldsymbol{X}_{(j-1)}(\mathbf{X}'_{(j-1)}\mathbf{X}_{(j-1)})^{-1}\mathbf{X}'_{(j-1)}$, $\mathbf{e}_{i,(j)}$ denotes the residual of the regression of $\boldsymbol{y}$ on $\boldsymbol{X}_{i,(j-1)} = \left(\boldsymbol{x}_i, \boldsymbol{X}_{(j-1)}\right)$. Regressors for which $I_{(j)}\widehat{(\beta_i \neq 0)} = I\left[\left|t_{\hat{\phi}_{T,i,(j)}}\right| > c_p(n,\delta)\right] = 1$, are then added to the set of already selected signal variables from the previous stages. Denote the number of variables selected in stage j by $\hat{k}^o_{n,T,(j)}$, their index set by $\mathcal{S}^o_{(j)}$, and the $T \times \hat{k}^o_{n,T,(j)}$ matrix of the $\hat{k}^o_{n,T,(j)}$ selected variables in stage j by $\mathbf{X}^o_{(j)}$. Also let $\mathbf{X}_{(j)} = (\mathbf{X}_{(j-1)}, \mathbf{X}^o_{(j)}) = (\boldsymbol{x}_{(j),1}, \boldsymbol{x}_{(j),2}, ..., \boldsymbol{x}_{(j),T})'$, $\hat{k}_{n,T,(j)} = \hat{k}_{n,T,(j-1)} + \hat{k}^o_{n,T,(j)}$, $\mathcal{S}_{(j)} = \mathcal{S}_{(j-1)} \cup \mathcal{S}^o_{(j)}$, and $\mathcal{N}_{(j)} = \{1,2,...,n\} \setminus \mathcal{S}_{(j)}$, and then proceed to the next stage by increasing j by one. Note that $\hat{k}_{n,T,(j)}$ is the *total* number of variables selected up to and including stage j, $\hat{\phi}_{T,i,(j)} \rightarrow_p \theta_{i,(j)}/\sigma_{ii}$, where $\theta_{i,(j)}$ is used in the remainder of this paper to denote $\theta_i\left(\boldsymbol{x}_{(j-1)}\right)$, introduced in (5), and note that $\theta_{i,(1)}$ is θ_i. The procedure stops when no regressors are selected at a given stage, say $\hat{j}_{n,T}$, in which case the final number of selected variables will be given by $\hat{k}_{n,T} = \hat{k}_{n,T,(\hat{j}_{n,T}-1)}$.

It is important to characterise the number of stages needed for OCMT. To do this we note that not all signal variables can be hidden and that once one conditions on the set of signal variables that are not hidden, then there exists i such that $\theta_i(\boldsymbol{z}) \neq 0$, while $\theta_i = 0$ and $\beta_i \neq 0$, where $\boldsymbol{z}$ denotes the signal variables that are not hidden.[2] This is proven in Lemma 1. Using this lemma one can successively uncover all hidden signals. We denote by P the number of stages that need to be considered to uncover all hidden signals. Its true population value is denoted by P_0. This is defined as the index of the last stage where OCMT finds further signals (or pseudo-signals), assuming that $\Pr[|t_{\hat{\phi}_{i,(j)}}| > c_p(n,\delta)\,|\theta_{i,(j)} \neq 0] = 1$ and $\Pr[|t_{\hat{\phi}_{i,(j)}}| > c_p(n,\delta)\,|\theta_{i,(j)} = 0] = 0$, for all variables, indexed i and OCMT stages, indexed j. Of course, these probabilities do not take the values 1 and 0 respectively, in small samples, but we will handle this complication later on. Then, the following proposition, proven in subsection A.1 of the Appendix, using Lemma 1, provides an upper limit for P_0.

Proposition 1 *Suppose that y_t, $t = 1, 2, ..., T$, are generated according to (1), with $\beta_i \neq 0$ for $i = 1, 2, ..., k$, and that Assumption 1 holds. Then, there exists j, $1 \leq j \leq k$, for which $\theta_{i,(j)} \neq 0$, and the population value of the number of stages required to select all the signals, denoted as P_0, satisfies $1 \leq P_0 \leq k$.*

Example 1 *As an illustration of Proposition 1 consider the case where $k = 2$, x_{1t} and x_{2t} are signal variables (hence $\beta_1 \neq 0$ and $\beta_2 \neq 0$) and the remaining $n-2$ variables in $\boldsymbol{x}_{nt}$ are noise variables. Then $\theta_1 = \beta_1\sigma_{11} + \beta_2\sigma_{12}$ and $\theta_2 = \beta_2\sigma_{22} + \beta_1\sigma_{12}$, and $\theta_i = 0$, for $i > 2$. Now if $\theta_1 = 0$, then $\beta_1 = -\frac{\beta_2\sigma_{12}}{\sigma_{11}}$ and $\theta_2 = \beta_2\left(\sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right)$ which can only be zero if the two signals are perfectly correlated. This is disallowed by Assumption 1. Furthermore, suppose that x_{2t} is selected in the first stage of OCMT, then it follows that once we condition on x_{2t} the net effect of x_{1t}, denoted by $\theta_{1,(2)}$ will be equal to $\beta_1\sigma_{11}$ which is non-zero by assumption.*

In finite samples, when no variables are selected in stage j, then stage $j-1$ will be denoted by $\hat{P}_{n,T}$, the estimator of P_0. So

$$\hat{P}_{n,T} = \min_j \left\{ j : \sum_{i=1}^{n} I_{(j)}\widehat{(\beta_i \neq 0)} = 0 \right\} - 1, \text{ and } I\widehat{(\beta_i \neq 0)} = \textstyle\sum_{j=1}^{\hat{P}_{n,T}} I_{(j)}\widehat{(\beta_i \neq 0)}. \tag{11}$$

In practice, $\hat{P}_{n,T}$ is likely to be small, since the occurrence of hidden signals (zero signal net effects) is less plausible, and all signals with nonzero θ will be picked up (with probability tending to one) in the first stage. Stopping after the first stage tends to improve the small sample performance of the OCMT approach, investigated in Section 5, only marginally when

[2] Note that $\boldsymbol{z}$ may contain principal components or other estimates of common effects as well as covariates that investigator believes must be included.

no hidden signals are present. Thus, allowing $\hat{P}_{n,T} > 1$, using the stopping rule defined above, does not significantly deteriorate the small sample performance when hidden signal variables are absent, while it picks-up all hidden signal variables with probability tending to one. Since the possibility of hidden signal variables cannot be ruled out in practice, we focus on the iterated version.

In a final step, the regression model is estimated by running the ordinary least squares (OLS) regression of y_t on all selected covariates, namely the regressors x_{it} for which $\widehat{I(\beta_i \neq 0)} = 1$, over all $i = 1, 2, ..., n$. Accordingly, the OCMT estimator of β_i, denoted by $\tilde{\beta}_i$, is then given by

$$\tilde{\beta}_i = \begin{cases} \hat{\beta}_i^{(\hat{k}_{n,T})}, & \text{if } \widehat{I(\beta_i \neq 0)} = 1 \\ 0, & \text{otherwise} \end{cases}, \text{ for } i = 1, 2, ..., n, \tag{12}$$

where $\hat{\beta}_i^{(\hat{k}_{n,T})}$ is the OLS estimator of the coefficient of the i^{th} variable in a regression that includes *all* the covariates for which $\widehat{I(\beta_i \neq 0)} = 1$, and a constant term.

Remark 1 *It is important to emphasise the role played by the critical value exponent, δ, in the OCMT procedure, as a means to ensure that noise variables are not selected. Its value can differ in various OCMT stages and, in fact, we will analyse OCMT under such a setting where one value of δ is used in the first stage, while another (denoted by δ^*) in subsequent stages. In particular, while $\delta > 1$ is a theoretically valid choice for the first stage of OCMT, subsequent stages of the procedure require $\delta^* > 2$ for the full set of our theoretical results to hold. Henceforth, we will assume that $\delta^* > \delta$ to simplify the analysis.*

We investigate the asymptotic properties of the OCMT procedure and the associated OCMT estimators, $\tilde{\beta}_i$, for $i = 1, 2, ..., n$. To this end we consider support recovery statistics used in the Lasso literature, namely the true positive rate, and the false positive rate, defined by

$$TPR_{n,T} = \frac{\sum_{i=1}^{n} I\left[\widehat{I(\beta_i \neq 0)} = 1 \text{ and } \beta_i \neq 0\right]}{\sum_{i=1}^{n} I(\beta_i \neq 0)}, \tag{13}$$

$$FPR_{n,T} = \frac{\sum_{i=1}^{n} I\left[\widehat{I(\beta_i \neq 0)} = 1, \text{ and } \beta_i = 0\right]}{\sum_{i=1}^{n} I(\beta_i = 0)}, \tag{14}$$

and the false discovery rate (if $\sum_{i=1}^{n} \widehat{I(\beta_i \neq 0)} > 0$) defined by[3]

$$FDR_{n,T} = \frac{\sum_{i=1}^{n} I\left[\widehat{I(\beta_i \neq 0)} = 1, \text{ and } \beta_i = 0\right]}{\sum_{i=1}^{n} \widehat{I(\beta_i \neq 0)}}. \tag{15}$$

[3]In cases where $\sum_{i=1}^{n} \widehat{I(\beta_i \neq 0)} = 0$, we set $FDR_{n,T} = 0$. Alternatively, one could re-define $FDR_{n,T}$ by replacing the denominator of (15) by $1 + \sum_{i=1}^{n} \widehat{I(\beta_i \neq 0)}$, without any material difference to the theoretical results.

We also consider the residual norm of the selected model, defined by

$$F_{\tilde{u}} = T^{-1} \sum_{t=1}^{T} \tilde{u}_t^2, \tag{16}$$

and the coefficient norm of the selected model, defined by

$$F_{\tilde{\boldsymbol{\beta}}} = ||\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n || = \left[\sum_{i=1}^{n} \left(\tilde{\beta}_i - \beta_i \right)^2 \right]^{1/2}, \tag{17}$$

where $\tilde{u}_t = y_t - \hat{a} - \tilde{\boldsymbol{\beta}}_n' \boldsymbol{x}_{nt}$, $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_n)'$, $\tilde{\beta}_i$, for $i = 1, 2, ..., n$ are given by (12), $\boldsymbol{\beta}_n = (\beta_1, \beta_2, ..., \beta_n)'$, and $\hat{a}$ represents the OLS estimator of the constant term in the final regression.

We consider the following assumptions:

Assumption 2 *The error term, u_t, in DGP (1) is a martingale difference process with respect to $\mathcal{F}_{t-1}^u = \sigma(u_{t-1}, u_{t-2}, ...,)$, with zero mean and a constant variance, $0 < \sigma^2 < C < \infty$. Each of the n covariates considered by the researcher, collected in the set $\mathcal{S}_{nt} = \{x_{1t}, x_{2t}, ..., x_{nt}\}$, is independently distributed of the errors $u_{t'}$, for all t and t'.*

Assumption 3 *Let $\mathcal{F}_{it}^x = \sigma(x_{it}, x_{i,t-1}, ...)$, where x_{it}, for $i = 1, 2, ..., n$, is the i-th covariate in the set $\mathcal{S}_{nt}$ considered by the researcher. Define $\mathcal{F}_t^{xn} = \cup_{j=k+k^*+1}^{n} \mathcal{F}_{jt}^x$, $\mathcal{F}_t^{xo} = \cup_{i=1}^{k+k^*} \mathcal{F}_{jt}^x$, and $\mathcal{F}_t^x = \mathcal{F}_t^{xn} \cup \mathcal{F}_t^{xo}$. Then, x_{it}, $i = 1, 2, ..., n$, are martingale difference processes with respect to $\mathcal{F}_{t-1}^x$. x_{it} is independent of $x_{jt'}$ for $i = 1, 2, ..., k + k^*$, $j = k + k^* + 1, k + k^* + 2, ..., n$, and for all t and t', and $E\left[x_{it}x_{jt} - E(x_{it}x_{jt}) \left| \mathcal{F}_{t-1}^x \right.\right] = 0$, for $i, j = 1, 2, ..., n$, and all t.*

Assumption 4 *There exist sufficiently large positive constants C_0, C_1, C_2 and C_3 and $s_x, s_u > 0$ such that the covariates $\mathcal{S}_{nt} = \{x_{1t}, x_{2t}, ..., x_{nt}\}$ satisfy*

$$\sup_{i,t} \Pr(|x_{it}| > \alpha) \leq C_0 \exp(-C_1 \alpha^{s_x}), \text{ for all } \alpha > 0, \tag{18}$$

and the errors, u_t, in DGP (1) satisfy

$$\sup_{t} \Pr(|u_t| > \alpha) \leq C_2 \exp(-C_3 \alpha^{s_u}), \text{ for all } \alpha > 0. \tag{19}$$

Assumption 5 *Consider the pair $\{x_t, \boldsymbol{q}_{\cdot t}\}$, for $t = 1, 2, ..., T$, where $\boldsymbol{q}_{\cdot t} = (q_{1,t}, q_{2,t}, ..., q_{l_T,t})'$ is an $l_T \times 1$ vector containing a constant and a subset of $\mathcal{S}_{nt}$, and x_t is a generic element of $\mathcal{S}_{nt}$ that does not belong to $\boldsymbol{q}_{\cdot t}$. It is assumed that $E(\boldsymbol{q}_{\cdot t} x_t)$ and $\boldsymbol{\Sigma}_{qq} = E(\boldsymbol{q}_{\cdot t} \boldsymbol{q}_{\cdot t}')$ exist and $\boldsymbol{\Sigma}_{qq}$ is invertible. Define $\boldsymbol{\gamma}_{qx,T} = \boldsymbol{\Sigma}_{qq}^{-1} \left[T^{-1} \sum_{t=1}^{T} E(\boldsymbol{q}_{\cdot t} x_t) \right]$ and*

$$u_{x,t,T} =: u_{x,t} = x_t - \boldsymbol{\gamma}_{qx,T}' \boldsymbol{q}_{\cdot t}. \tag{20}$$

All elements of the vector of projection coefficients, $\boldsymbol{\gamma}_{qx,T}$, are uniformly bounded and only a finite number of the elements of $\boldsymbol{\gamma}_{qx,T}$ are different from zero.

Assumption 6 *The number of the true regressors in DGP (1), k, is finite, and their slope coefficients could change with T, such that for $i = 1, 2, ..., k$, $\beta_{i,T} = \ominus\left(T^{-\vartheta}\right)$, for some $0 \leq \vartheta < 1/2$.*

The above assumption allows for the possibility of weak signals whose coefficients, $\beta_{i,T}$, for $i = 1, 2, ..., k$, decline with the sample size, T, at a sufficiently slow rate. But to simplify the notations subscript T is dropped subsequently, and it is understood that the slope and net effect coefficients can change with the sample size according to Assumption 6. Given the DGP (1), it is helpful to write the conditional net effect coefficient as

$$\theta_{i,(j)} = \sum_{\ell=1}^{k} \beta_\ell \sigma_{i\ell}\left(\boldsymbol{x}_{(j-1)}\right) = E\left(T^{-1}\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{X}_k\boldsymbol{\beta}_k\right) = E\left(T^{-1}\boldsymbol{x}_i'\boldsymbol{M}_{(j-1)}\boldsymbol{y}\right). \tag{21}$$

Under Assumption 6, and given that $\sigma_{i\ell}\left(\boldsymbol{x}_{(j-1)}\right)$ is bounded, $\theta_{i,(j)}$ are, for a suitable j, either bounded away from 0, or declining to 0 but not faster than the rate $\ominus\left(T^{-\vartheta}\right)$ for some $0 \leq \vartheta < 1/2$ introduced in Assumption 6. Using $\theta_{i,(j)}$, we can refine our concept of pseudo-signal variables as variables with $\theta_{i,(j)} = \ominus\left(T^{-\vartheta}\right)$ for $i = k+1, k+2, ..., k+k^*$, some $0 \leq \vartheta < 1/2$ and some $1 \leq j \leq P_0$.

Before presenting our theoretical results we provide some remarks on the pros and cons of our assumptions as compared to the ones typically assumed in the penalised and boosting literature. The signal and pure noise variables are allowed to be correlated amongst themselves; namely, no restrictions are imposed on σ_{ij} for $i, j = 1, 2, ..., k$, and on σ_{ij} for $i, j = k+k^*+1, k+k^*+2, ..., n$. Also, signal and pseudo-signal variables are allowed to be correlated; namely, σ_{ij} could be non-zero for $i, j = 1, 2, ..., k+k^*$. Therefore, signal and pseudo-signal variables as well as pure noise variables can contain common factors. But under Assumption 3, $E\left[x_{it} - E\left(x_{it}\right)|x_{jt}\right] = 0$ for $i = 1, 2, ..., k$ and $j = k+k^*+1, k+k^*+2, ..., n$. This implies that, if there are common factors, they cannot be shared between signal/pseudo-signal variables and noise variables, although one can condition on such factors, as we do in our empirical illustration.[4]

The exponential bounds in Assumption 4 are sufficient for the existence of all moments of covariates, x_{it}, and errors, u_t. It is very common in the literature to assume some form of exponentially declining bound for probability tails for u_t and x_{it} where appropriate. Such an assumption can take the simplified form of assuming normality, as in, e.g., Zheng, Fan, and Lv (2014).

[4] Note that our theory allows for conditioning on observed common factors. But when factors are unobserved they need to be replaced by their estimates using, for example, principal components. A formal argument that the associated estimation error is asymptotically negligible involves additional technical complications, and requires deriving exponential inequalities for the quantities analysed in Theorem 1 of Bai and Ng (2002) and Lemma A1 of Bai and Ng (2006), and then assuming that $\sqrt{T}/n \rightarrow 0$ as $n, T \rightarrow \infty$. While such a derivation is clearly feasible under appropriate regularity conditions, a formal analysis is beyond the scope of the present paper.

Assumption 6 is a set of regularity conditions. It allows for small β_i and $\theta_{i,(j)}$, for a suitable j, as long as they are not too small - i.e. they can tend to zero but at a rate slower than $T^{-1/2}$. Remark 3 discusses further how this condition enters the theoretical results. Assumption 5 is a technical condition that is required for some results derived in the Appendix, which consider a more general multiple regression context where subsets of regressors in $\boldsymbol{x}_{nt}$ are included in the regression equation. If $\mathbf{Q} = (\boldsymbol{q}_{\cdot 1}, \boldsymbol{q}_{\cdot 2}, ..., \boldsymbol{q}_{\cdot T})' = \boldsymbol{\tau}_T = (1, 1, ..., 1)'$, then Assumption 5 is trivially satisfied given the rest of the assumptions. Then, $\boldsymbol{\gamma}_{qx,T} = \mu_{x,T} = \frac{1}{T}\sum_{t=1}^{T} E(x_t)$ and $u_{x,t,T} = x_t - \mu_{x,T}$.

It is important to place our assumptions in the context of the existing literature. In many analyses of alternative methods, such as penalised regression, it is usual to assume that the covariates, $\boldsymbol{x}_{nt}$, are either deterministic or stochastic but distributed as IID random variables. (See, for example, Buhlmann and van de Geer (2011) or Zheng, Fan, and Lv (2014) for recent contributions). Our martingale difference assumption relaxes the IID assumption somewhat. Further relaxation of this assumption is discussed in Section 4.

Regarding our assumptions on the correlation between signal and pseudo-signal covariates, we allow for noise variables to have a common factor, and do not require the covariance matrix of $\boldsymbol{x}_{nt}$ to be sparse. To identify the signal variables we do need to assume the sparsity of correlation between the signal and non-signal variables as captured by the presence of k^* pseudo-signal variables. The OCMT approach can identify the k signal and up to k^* pseudo-signal variables with a probability tending towards 1. The selected regressors are then considered in a multiple regression and the relevant regression coefficients are estimated consistently, under mild restrictions on k^* such as $k^* = o(T^{1/4})$. In contrast, a number of crucial issues arise in the context of Lasso, or more generally when L_q penalty functions with $0 \leq q \leq 1$ are used. Firstly, it is customary to assume a framework of fixed-design regressor matrices, where in many cases a generalisation to stochastic regressors is not straightforward, requiring conditions such as the spark condition of Donoho and Elad (2003) and Zheng, Fan, and Lv (2014). Secondly, a frequent condition for Lasso to be a valid variable selection method is the irrepresentable condition which bounds the maximum of all regression coefficients, in regression of any noise or pseudo-signal variable on the signal variables, to be less than one in the case of normalised regressor variables. See, for example, Section 7.5 of Buhlmann and van de Geer (2011).

A further issue relates to the fact that most results for penalised regressions essentially take as given the knowledge of the tuning parameter associated with the penalty function, in order to obtain oracle results. In practice, cross-validation is recommended to determine this parameter but theoretical results on the properties of such cross-validation schemes are rare. Finally, it is worth commenting on the assumptions underlying boosting as presented in Buhlmann (2006). There, it is assumed that the regressors are iid and bounded while few restrictions are placed on their correlation structure. Nevertheless, it is important to note that the aim of boosting in that paper is to obtain a good approximation to the regression function and not to select the true regressors.

4 Main Theoretical Results

We now present the main theoretical results using lemmas established in the Appendix. The key is Lemma 10, which provides sharp bounds for $\Pr\left[\left|t_{\hat{\phi}_{i,(j)}}\right| > c_p\left(n,\delta\right)|\theta_{i,(j)} \neq 0\right]$. Since we wish to allow for the possibility that $\theta_i = 0$ if $\beta_i \neq 0$, the results in the appendix are obtained for t-ratios in multiple regression contexts where subsets of regressors in $\boldsymbol{x}_{nt}$ are included in the regression equation. It is instructive to initially consider the properties of the first step of the iterative OCMT as it is simpler and covers the dominant case where $\theta_i \neq 0$ if $\beta_i \neq 0$. Our results will consequently and formally be generalised for the full iterative method. We present results for $TPR_{n,T}$, $FPR_{n,T}$, $FDR_{n,T}$, the probability of selecting the pseudo-true model and parameter estimate error and regression error norms. Below we sketch the results we obtain using the first step of OCMT as a vehicle, for ease of exposition, while the formal analysis is provided in Theorems 1 and 2 and proven in Section A.2 of the Appendix.

We first examine $TPR_{n,T}$ defined by (13), under the assumption that $\theta_i \neq 0$ if $\beta_i \neq 0$. Note that

$$TPR_{n,T} = \frac{\sum_{i=1}^{n} I\left[I\left(\widehat{\beta_i \neq 0}\right) = 1 \text{ and } \beta_i \neq 0\right]}{\sum_{i=1}^{n} I(\beta_i \neq 0)} = \frac{\sum_{i=1}^{k} I\left[I\left(\widehat{\beta_i \neq 0}\right) = 1 \text{ and } \beta_i \neq 0\right]}{k}.$$

Since the elements in the above summations are 0 or 1, then taking expectations we have

$$\frac{\sum_{i=1}^{k} E\left\{I\left[I_{(1)}\left(\widehat{\beta_i \neq 0}\right) = 1 \text{ and } \theta_i \neq 0\right]\right\}}{k} = \frac{\sum_{i=1}^{k} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p\left(n,\delta\right)|\theta_i \neq 0\right]}{k}.$$

Suppose there exists $\kappa_1 > 0$ such that $T = \ominus\left(n^{\kappa_1}\right)$. Using (A.108) of Lemma 10, where the matrix $\mathbf{Q}$, referred to in the statement of the Lemma, is set equal to $\boldsymbol{\tau}_T$, and noting that $c_p\left(n,\delta\right)$ is given by (8), $1-\Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p\left(n,\delta\right)|\theta_i \neq 0\right] = O\left[\exp\left(-C_2 T^{C_3}\right)\right] = O\left[\exp\left(-C_2 n^{C_3\kappa_1}\right)\right]$, for some $C_2, C_3 > 0$, where as defined by (21), $\theta_i = \theta_{i,(1)} = E\left(\boldsymbol{x}_i' \boldsymbol{M}_\tau \boldsymbol{y}/T\right)$. Using $P\left(\mathcal{A}\right) = 1 - P\left(\mathcal{A}^c\right)$, where $\mathcal{A}^c$ denotes the complement of event $\mathcal{A}$, we obtain

$$\Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| \leq c_p\left(n,\delta\right)|\theta_i \neq 0\right] = O\left[\exp\left(-C_2 T^{C_3}\right)\right], \tag{22}$$

and noting that $\theta_i \neq 0$ for all signals $i = 1, 2, ..., k$, then under Assumption 6 we have

$$k^{-1}\sum_{i=1}^{k} \Pr\left(\left|t_{\hat{\phi}_{i,(1)}}\right| \leq c_p\left(n,\delta\right)|\beta_i \neq 0\right) = k^{-1}\sum_{i=1}^{k} O\left[\exp\left(-C_2 T^{C_3}\right)\right]. \tag{23}$$

Consider now $FPR_{n,T}$ defined by (14). Again, note that the elements of $FPR_{n,T}$ are either 0 or 1 and hence $|FPR_{n,T}| = FPR_{n,T}$. Taking expectations of (14), and assuming

$\theta_i = \ominus\left(T^{-\vartheta}\right)$, for $i = k+1, k+2, ..., k+k$, and some $0 \leq \vartheta < 1/2$, we have

$$\frac{\sum_{i=k+1}^{n} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,\delta) \,|\beta_i = 0\right]}{n-k} = \frac{\left[\begin{array}{c} \sum_{i=k+1}^{k+k^*} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,\delta) \,|\theta_i \neq 0\right] + \\ \sum_{i=k+k^*+1}^{n} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,\delta) \,|\theta_i = 0\right] \end{array}\right]}{n-k},$$

where, as before, $\theta_i = \theta_{i,(1)} = E\left(\boldsymbol{x}_i'\boldsymbol{M}_\tau \boldsymbol{y}/T\right)$ (see (21)). Using (A.108) of Lemma 10 and assuming there exists $\kappa_1 > 0$ such that $T = \ominus\left(n^{\kappa_1}\right)$, we have $k^* - \sum_{i=k+1}^{k+k^*} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,\delta) \,|\theta_i \neq 0\right] = O\left[\exp\left(-C_2 T^{C_3}\right)\right]$, for some finite positive constants C_2 and C_3. Moreover, (A.107) of Lemma 10, which holds uniformly over i, given the uniformity of (18) and (19) of Assumption 4, implies that for any $0 < \varkappa < 1$ there exist finite positive constants C_0 and C_1 such that

$$\sum_{i=k+k^*+1}^{n} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,\delta) \,|\theta_i = 0\right] \leq \sum_{i=k+k^*+1}^{n} \left\{\exp\left[\frac{-\varkappa c_p^2(n,\delta)}{2}\right] + \exp\left(-C_0 T^{C_1}\right)\right\}. \tag{24}$$

Using these results we obtain

$$\frac{\sum_{i=k+1}^{n} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,\delta) \,|\beta_i = 0\right]}{n-k} = \left(\frac{k^*}{n-k}\right) + O\left\{\exp\left[-\frac{\varkappa c_p^2(n,\delta)}{2}\right]\right\} + O\left[\exp(-C_0 T^{C_1})\right] + O\left[(n-k)^{-1}\exp\left(-C_2 T^{C_3}\right)\right]. \tag{25}$$

Next, we consider the probability of choosing the pseudo-true model. We denote a selected regression model as a pseudo-true model if it contains the (true) regressors x_{it}, $i = 1, 2, ..., k$, and none of the noise variables, x_{it}, $i = k+k^*+1, k+k^*+2, ..., n$. The models in the set may contain one or more of the pseudo-signal variables, x_{it}, $i = k+1, k+2, ..., k+k^*$. We refer to all such regressions as the set of pseudo-true models. So, the event of choosing the pseudo-true model is given by

$$\mathcal{A}_0 = \left\{\sum_{i=1}^{k} I\left(\widehat{\beta_i \neq 0}\right) = k\right\} \cap \left\{\sum_{i=k+k^*+1}^{n} I\left(\widehat{\beta_i \neq 0}\right) = 0\right\}. \tag{26}$$

Theorem 1 states that, under certain conditions, $\Pr\left(\mathcal{A}_0\right) \to 1$. The above discussion relates mainly to the first step of OCMT. The results for the general case are given in the following theorem, proven in Subsection A.2.1 of the Appendix. Given our relative n/T rate assumption, all rate results in our analysis are reported in terms of n for presentational consistency and ease of comprehension. They could, of course, be reported in terms of T instead.

Theorem 1 *Consider the DGP (1) with k signal variables, k^* pseudo-signal variables, and $n-k-k^*$ noise variables, and suppose that Assumptions 1-4 and 6 hold, Assumption 5 holds for all pairs $(x_{it}, \boldsymbol{X}_{(j-1)})$, $i \in \mathcal{N}_{(j-1)}$, $j = 1, 2, ...,$ where j denotes the stage of the OCMT procedure, and $\boldsymbol{X}_{(j-1)}$, and $\mathcal{N}_{(j-1)}$ are defined in Section 3. $c_p(n,\delta)$ is given by (8) with*

$0 < p < 1$ and let $f(n,\delta) = cn^{\delta}$, for the first stage of OCMT and $f(n,\delta^*) = cn^{\delta^*}$, for subsequent stages, for some $c > 0$, $\delta^* > \delta > 0$. $n, T \to \infty$, such that $T = \ominus(n^{\kappa_1})$, for some $\kappa_1 > 0$, and $k^* = \ominus(n^{\epsilon})$ for some positive $\epsilon < \min\{1, \kappa_1/3\}$. Then, for any $0 < \varkappa < 1$, and for some constant $C_0 > 0$,

(a) *the probability that the number of stages in the OCMT procedure, $\hat{P}_{n,T}$, defined by (11), exceeds k is given by*

$$\Pr\left(\hat{P}_{n,T} > k\right) = O\left(n^{1-\varkappa\delta^*}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[\exp\left(-n^{C_0\kappa_1}\right)\right], \tag{27}$$

(b) *the probability of selecting the pseudo-true model, $\mathcal{A}_0$, defined by (26), is given by*

$$\Pr\left(\mathcal{A}_0\right) = 1 + O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[\exp\left(-n^{C_0\kappa_1}\right)\right], \tag{28}$$

(c) *for the True Positive Rate, $TPR_{n,T}$, defined by (13), we have*

$$E\left|TPR_{n,T}\right| = 1 + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[\exp\left(-n^{C_0\kappa_1}\right)\right], \tag{29}$$

and if $\delta > 1-\kappa_1/3$, then $TPR_{n,T} \to_p 1$; for the False Positive Rate, $FPR_{n,T}$, defined by (14), we have

$$E\left|FPR_{n,T}\right| = \frac{k^*}{n-k} + O\left(n^{-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left(n^{\epsilon-1}\right) + O\left[\exp\left(-n^{C_0\kappa_1}\right)\right], \tag{30}$$

and if $\delta > \min\{0, 1-\kappa_1/3\}$, and $\delta^* > 1$, then $FPR_{n,T} \to_p 0$. For the False Discovery Rate, $FDR_{n,T}$, defined in (15), we have

$$FDR_{n,T} \to_p \frac{k^*}{k^*+k}, \tag{31}$$

if $\sum_{i=1}^{n} I\left(\widehat{\beta_i \neq 0}\right) > 0$, $\delta > \max\{1, 2-\kappa_1/3\}$, $\delta^* > 2$, and $\theta_{i,(j)} = \ominus\left(T^{-\vartheta}\right)$ for $i = k+1, k+2, ..., k+k^*$, some $0 \leq \vartheta < 1/2$ and some $1 \leq j \leq P_0$.

(d) *For the residual norm of the selected model, $F_{\tilde{u}}$, defined by (16), we have*

$$E\left(F_{\tilde{u}}\right) \to \sigma^2, \text{ if } \delta > 1 \text{ and } \delta^* > 2. \tag{32}$$

Remark 2 *Although our proof requires that $0 < \varkappa < 1$, in practice it sufficient to set $\varkappa$ to be arbitrarily close to, but less than, unity. Also, κ_1 can be arbitrarily small which allows n to rise much faster than T. The condition $0 \leq \epsilon < \min\{1, \kappa_1/3\}$ ensures that $k^*/n \to 0$ and $k^* = o(T^{1/3})$. Finally, it is clear from (28) that if $\delta > 1$ and $\delta^* > 2$, $\Pr(\mathcal{A}_0) \to 1$, as n and $T \to \infty$.*

Remark 3 *Assumption 6 allows for weak signals. In particular, we allow slope coefficients of order* $\ominus\left(T^{-\vartheta}\right)$, *for some* $0 \leq \vartheta < 1/2$. *Then, by (A.113) and (A.114) of Lemma 10, it is seen that such weak signals can be picked up at no cost, in terms of rates, with respect to the exponential inequalities that underlie all the theoretical results. In particular, the power of the OCMT procedure in selecting the signal* x_{it} *rises with* $\sqrt{T}\left|\theta_{i,(j)}\right|/\sigma_{e_i,(T)}\sigma_{x_i,(T)}$, *so long as* $\frac{c_p(n,\delta)}{\sqrt{T}\left|\theta_{i,(j)}\right|} \to 0$, *as* n *and* $T \to \infty$, *where* $\sigma_{e_i,(T)}$ *and* $\sigma_{x_i,(T)}$ *are defined by (A.105), replacing* $\mathbf{e}$, $\mathbf{x}$, *and* $\mathbf{Q}$ *by* $\mathbf{e}_i$, $\mathbf{x}_i$, *and* $\boldsymbol{M}_{(j-1)}$, *respectively. When this ratio is low, a large* T *will be required for the OCMT approach to select the* i^{th} *signal. This condition is similar to the so-called 'beta-min' condition assumed in the penalised regression literature. (See, for example, Section 7.4 of Buhlmann and van de Geer (2011) for a discussion.)*

Remark 4 *OCMT selects signals as well as pseudo-signals with nonzero net effect coefficients, hence the probability limit of* $FDR_{n,T}$ *can be nonzero when pseudo-signals are present* ($k^* \neq 0$). *If FDR per se was the main objective of the analysis, then, a post-OCMT selection, using, for example, the Schwarz information criterion, could be considered to separate the signals from the pseudo-signals. However, when the norm of slope coefficients or the in-sample fit of the model is of main concern, then, under appropriate conditions on the rate at which* k^* *expands with* n, *the inclusion of pseudo-signals is asymptotically innocuous, as shown in Theorem 2 below.*

Consider now the coefficient norm of the selected model, $F_{\tilde{\boldsymbol{\beta}}}$, defined in (17). We assume the following additional regularity condition.

Assumption 7 *Let* $\mathbf{S}$ *denote the* $T \times l_T$ *observation matrix on the* l_T *regressors selected at any one of the* $\hat{P}_{n,T}$ *stages of the OCMT procedure. Then,*

1. *Let* $\boldsymbol{\Sigma}_{ss} = E\left(\mathbf{S}'\mathbf{S}/T\right)$ *with eigenvalues denoted by* $\mu_1 \leq \mu_2 \leq ... \leq \mu_{l_T}$. *Let* $\mu_i = O\left(l_T\right)$, $i = l_T - M + 1, l_T - M + 2, ..., l_T$, *for some finite* M, *and* $\sup_{1 \leq i \leq l_T - M} \mu_i < C_0 < \infty$, *for some* $C_0 > 0$. *In addition,* $\inf_{1 \leq i < l_T} \mu_i > C_1 > 0$, *for some* $C_1 > 0$.
2. $E\left[\left(1 - \|\boldsymbol{\Sigma}_{ss}^{-1}\|_F \left\|\hat{\boldsymbol{\Sigma}}_{ss} - \boldsymbol{\Sigma}_{ss}\right\|_F\right)^{-4}\right] = O\left(1\right)$, *where* $\hat{\boldsymbol{\Sigma}}_{ss} = \mathbf{S}'\mathbf{S}/T$.

Theorem 2 *Consider the DGP defined by (1), and the coefficient norm of the selected model,* $F_{\tilde{\boldsymbol{\beta}}}$ *defined in (17). Suppose that Assumptions 1-4 and 6-7 hold, Assumption 5 holds for the pairs* $(x_{it}, \boldsymbol{X}_{(j-1)})$, $i \in \mathcal{N}_{(j-1)}$, $j = 1, 2, ...$, *where* j *denotes the stage of the OCMT procedure, and* $\boldsymbol{X}_{(j-1)}$, $\mathcal{N}_{(j-1)}$ *are defined in Section 3, and* k^* *(the number of pseudo signals) is of order* $\ominus\left(n^{\epsilon}\right)$ *for some positive* ϵ. *Let* $c_p\left(n, \delta\right)$ *defined by (8),* $0 < p < 1$ *and let* $f\left(n, \delta\right) = cn^{\delta}$, *for the first stage of OCMT and* $f\left(n, \delta^*\right) = cn^{\delta^*}$, *for subsequent stages, for some* $c > 0$, $\delta^* > \delta > 1$ *and* $\delta^* > 2$. *Denote the maximum number of selected regressors that is allowed to enter the* *final stage regression by* $l_{\max}$ *and suppose that* $l_{\max} = \ominus(n^{\kappa_2})$, *for some* $\kappa_2 > 0$. *Let* $\tilde{\boldsymbol{\beta}}_n$ *be the estimator of* $\boldsymbol{\beta}_n = (\beta_1, \beta_2, ..., \beta_n)'$ *in the final regression with at most* $l_{\max}$ *regressors. In addition,* $T = \ominus(n^{\kappa_1})$, *for some* $\kappa_1 > 0$. *Assume that* $\epsilon < \min\{\kappa_2, \kappa_1/3\}$. *Then, for any* $0 < \varkappa < 1$, *and some constant* $C_0 > 0$, *we have*

$$E\left(F_{\tilde{\boldsymbol{\beta}}}\right) = O\left(n^{2\epsilon-\kappa_1/2}\right) + O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left(n^{1-\delta\varkappa+2\kappa_2-\kappa_1/2}\right) + O\left(n^{2-\delta^*\varkappa+2\kappa_2-\kappa_1/2}\right) + O\left[\exp\left(-n^{C_0\kappa_1}\right)\right]. \quad (33)$$

As can be seen from the above theorem, (33) requires stronger conditions than those needed for the proof of the earlier results in Theorem 1. In particular, the two conditions in Assumption 7 are needed for controlling the rate of convergence of the inverse of sample covariance matrix of the selected regressors. The first condition relates to the eigenvalues of the population covariance of the selected regressors, denoted by $\boldsymbol{\Sigma}_{ss}$, and aims to control the rate at which $\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F$ grows. The second condition bounds the expectation of $(1 - \|\boldsymbol{\Sigma}_{ss}^{-1}\|_F \, ||\hat{\boldsymbol{\Sigma}}_{ss} - \boldsymbol{\Sigma}_{ss}||_F)^{-4}$, which is needed for our derivations. Under our conditions on the number of selected regressors, $\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F \, E(||\hat{\boldsymbol{\Sigma}}_{ss} - \boldsymbol{\Sigma}_{ss}||_F) = o(1)$, but this is not sufficient for $E[(1-\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F \, ||\hat{\boldsymbol{\Sigma}}_{ss}-\boldsymbol{\Sigma}_{ss}||_F)^{-4}] = O(1)$, so an extra technical assumption is needed. Note that $E(F_{\tilde{\boldsymbol{\beta}}})$ is related to, and has the same rate as, the RMSE of $\tilde{\boldsymbol{\beta}}_n$. It is possible to easily obtain a rate for $E(F_{\tilde{\boldsymbol{\beta}}}^2)$, i.e. the MSE of $\tilde{\boldsymbol{\beta}}_n$, which is the square of the rate given in (33). We focus on $E(F_{\tilde{\boldsymbol{\beta}}})$ to avoid more complex regularity conditions than those given in Assumption 7.

It is important to provide intuition on why we can get a consistency result for the coefficient norm of the selected model even though the selection process includes pseudo-signal variables. There are two reasons for this. First, since OCMT procedure selects all the signals with probability approaching one as $n, T \to \infty$, then the coefficients of the additionally selected regressors (whether pseudo-signal or noise) will tend to zero with T. Second, restricting k^* implies that the inclusion of pseudo-signal variables can be accommodated since their estimated coefficients will tend to zero and the variance of these estimated coefficients will be controlled. Some noise variables may also be selected, but we restrict the overall number of regressors that enter the final regression by using a bound, $l_{\max}$. This bound applies only at the *final* regression stage after the OCMT selection procedure. In the unlikely event that $\hat{k}_{n,T} + 1 > l_{\max}$, $\hat{k}_{n,T} - l_{\max} - 1$ variables are dropped ex post. The proof of Theorem 2 does not depend on which of the variables are dropped. In practice, this could be done by dropping selected regressors with the lowest t-statistics, in absolute value, over all OCMT stages. The bound is assumed, to allow consideration of smaller values of δ. This follows if we note that κ_2 can be set to 1 which would imply that the restriction is not binding but, then, larger values of δ would be required for norm consistency. The Monte Carlo evidence in this paper suggests that the number of noise variables selected is well controlled by multiple testing and there is no need to impose a bound in small samples. It is also worth noting that the result (32) on the residual norm of the selected model does not require Assumption 7. This is because fitted values are defined even if the sample covariance of the selected regressors is not invertible.

In the case when the net effect coefficients of signal variables in the first stage of OCMT satisfy $\theta_i = \ominus\left(T^{-\vartheta}\right)$, if $\beta_i \neq 0$, for some $0 \leq \vartheta < 1/2$ and for $i = 1, 2, ..., k$, then $P_0 = 1$, and further iterations ($j > 1$) of the OCMT will not be required. Consequently, the results of Theorems 1 and 2 can be simplified and obtained under a less restrictive set of conditions. Under $P_0 = 1$, and assuming that the conditions of Theorem 1 hold, with the exception of the condition on ϵ which could lie in $[0, 1)$, we obtain the following results, established in Section A.2.4 in the Appendix. The probability of selecting the pseudo-true model is given by

$$\Pr\left(\mathcal{A}_0\right) = 1 + O\left(n^{1-\delta\varkappa}\right) + O\left[n \exp\left(-n^{C_0}\right)\right], \tag{34}$$

and $\Pr\left(\mathcal{A}_0\right) \rightarrow_p 1$, if $\delta > 1$. For the support recovery statistics, we have

$$E\left|TPR_{n,T}\right| = 1 + O\left[\exp\left(-n^{C_0}\right)\right], \text{ and} \tag{35}$$

$$E\left|FPR_{n,T}\right| = \frac{k^*}{n-k} + O\left(n^{-\delta\varkappa}\right) + O\left(n^{\epsilon-1}\right) + O\left[\exp(-n^{C_0})\right]. \tag{36}$$

Hence, if $\delta > 0$, then $TPR_{n,T} \rightarrow_p 1$, and $FPR_{n,T} \rightarrow_p 0$. In addition, if $\sum_{i=1}^{n} I\left(\widehat{\beta_i \neq 0}\right) > 0$, $\delta > 1$, and $\theta_i = \ominus\left(T^{-\vartheta}\right)$, for $i = k+1, k+2, ..., k+k^*$, and some $0 \leq \vartheta < 1/2$, then the result on the false discovery rate, (31), hold. The result on the residual norm of the selected model, (32), also hold, if $\delta > 1$. Further, if the conditions of Theorem 2 hold with the exception of the condition on ϵ, which now could lie in $[0, 1)$, we have

$$E\left\|F_{\tilde{\boldsymbol{\beta}}}\right\| = O\left(n^{2\epsilon-\kappa_1/2}\right) + O\left(n^{1+2\kappa_2-\kappa_1/2-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta}\right) + O\left[\exp\left(-n^{C_0}\right)\right]. \tag{37}$$

Theorems 1 and 2, and the rest of the results above relate to the first maintained assumption about the pseudo-signal variables where at most k^* of them have non-zero $\theta_{i,(j)}$ for some j. This result can be extended to the case where potentially all variables have non-zero θ_i, as long as θ_i's are absolutely summable. Two leading cases considered in the literature are to assume that there exists a (possibly unknown) ordering such that

$$\theta_i = C_i \varrho^i, \text{ for } i = 1, 2, ..., n, \text{ and } |\varrho| < 1, \tag{38}$$

for a given set of constants, C_i, with $\sup_i |C_i| < \infty$, or

$$\theta_i = C_i i^{-\gamma}, \text{for } i = 1, 2, ..., n, \text{ and for some } \gamma > 0. \tag{39}$$

The assumption that there is only a finite number of variables for which $\beta_i \neq 0$, is retained. The rationale for hidden signals is less clear for these cases, since rather than a discrete separation between variables with zero and non-zero θ_i, we consider a form of continuum that unites these two classes of variables. Essentially, we have no separation in terms of signal variables (or pseudo-signal variables) and noise variables, since there are no noise variables. Therefore, the relevance of the iterative OCMT scheme is less clear. As a result, we focus on the first stage of OCMT ($j = 1$) and provide some results for the settings implied by (38) and (39).

Theorem 3 *Consider the DGP defined by (1), suppose that Assumptions 1-4 and 6 hold, Assumption 5 holds for the pairs* $(x_{it}, 1)$, $i = 1, 2, ..., n$, *and condition (38) holds. Moreover, let* $c_p(n, \delta)$ *be given by (8) with* $0 < p < 1$ *and* $f(n, \delta) = cn^{\delta}$, *for some* $c, \delta > 0$, *and suppose there exists* $\kappa_1 > 0$ *such that* $T = \ominus(n^{\kappa_1})$. *Consider the variables selected at the first stage of the OCMT procedure. Then, for all* $\zeta > 0$, *we have* $E|FPR_{n,T}| = o(n^{\zeta-1}) + O\left[\exp(-n^{C_0})\right]$, *for some finite positive constant* C_0, *where* $FPR_{n,T}$ *is defined by (14). If condition (39) holds instead of condition (38), then, assuming* $\gamma > \frac{1}{2}\kappa_1^2$, *we have* $FPR_{n,T} \to_p 0$.

An important assumption made so far is that noise variables are martingale difference processes which could be quite restrictive in the case of time series applications. This assumption can be relaxed. In particular, under the less restrictive assumption that noise variables are exponentially mixing, it can be shown that all the theoretical results derived above hold. Details are provided in Section B of the online theory Supplement. A further extension involves relaxing the martingale difference assumption for the signal and pseudo-signal covariates. If we are willing to assume that either u_t is normally distributed or the covariates are deterministic, then a number of results become available. The relevant lemmas for the deterministic case are presented in Section D of the online theory Supplement. Alternatively, signal/pseudo-signal regressors can be assumed to be exponentially mixing. In this general case, some results can still be obtained. These are described in Section B of the online theory Supplement.

5 A Monte Carlo Study

We employ five different Monte Carlo (MC) designs that differ in the extent of correlation across covariates, in the way $\theta_{i,(j)}$ and β_i are related, and in the size of the β_i coefficients.

5.1 Data-generating processes (DGPs)

In all five designs described below, we consider several options in generating the covariates. We allow the covariates to be serially correlated and consider different degrees of correlations across them. In addition, we also consider experiments with Gaussian and non-Gaussian errors.

5.1.1 Design I (zero correlations between signal and noise variables)

There are no pseudo-signal variables and all signal variables have $\theta_i \neq 0$. y_t is generated as:

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \varsigma u_t, \tag{40}$$

where $u_t \sim IIDN\,(0,1)$ in the Gaussian case, and $u_t = \left[\chi_t^2(2) - 2\right]/2$ in the non-Gaussian case, in which $\chi_t^2(2)$ are independent draws from a χ^2-distribution with 2 degrees of freedom, for $t = 1, 2, ..., T$. We set $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ and consider the following alternatives ways of generating $\boldsymbol{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$:

DGP-I(a) Temporally uncorrelated and weakly collinear regressors: signal variables are generated as $x_{it} = (\varepsilon_{it} + \nu g_t)/\sqrt{1+\nu^2}$, for $i = 1, 2, 3, 4$, and noise variables are generated as $x_{5t} = \varepsilon_{5t}$, $x_{it} = (\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2}$, for $i > 5$, where g_t and ε_{it} are independent draws either from $N(0,1)$ or from $\left[\chi_t^2(2) - 2\right]/2$, for $t = 1, 2, ..., T$, and $i = 1, 2, ..., n$. We set $\nu = 1$, which implies 50% pair-wise correlation among the signal variables.

DGP-I(b) Temporally correlated and weakly collinear regressors: Regressors are generated as in DGP-I(a), but with $\varepsilon_{it} = \rho_i \varepsilon_{i,t-1} + \sqrt{1-\rho_i^2}\, e_{it}$, in which $e_{it} \sim IIDN\,(0,1)$ or $IID\left[\chi_t^2(2) - 2\right]/2$. We set $\rho_i = 0.5$ for all i.

DGP-I(c) Strongly collinear noise variables due to a persistent unobserved common factor: signal variables are generated as $x_{it} = (\varepsilon_{it} + g_t)/\sqrt{2}$, for $i = 1, 2, 3, 4$, and noise variables are generated as $x_{5t} = (\varepsilon_{5t} + b_i f_t)/\sqrt{3}$ and $x_{it} = \left[(\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2} + b_i f_t\right]/\sqrt{3}$, for $i > 5$, where $b_i \sim IIDN\,(1,1)$, $f_t = 0.95 f_{t-1} + \sqrt{1-0.95^2}\, v_t$, and v_t, g_t and ε_{it} are independent draws from $N\,(0,1)$ or $\left[\chi_t^2(2) - 2\right]/2$.

DGP-I(d) Low or high pair-wise correlation of signal variables: Regressors are generated as in DGP-I(a), but we set $\nu = \sqrt{\omega/(1-\omega)}$, for $\omega = 0.2$ (low pair-wise correlation) and 0.8 (high pair-wise correlation). This ensures that average correlation among the signal variables is ω.

5.1.2 Design II (non-zero correlations between signal and noise variables)

We allow for pseudo-signal variables ($k^* > 0$). The DGP is given by (40) and $\boldsymbol{x}_{nt}$ is generated as:

DGP-II(a) Two pseudo-signal variables: signal variables are generated as $x_{it} = (\varepsilon_{it} + g_t)/\sqrt{2}$, for $i = 1, 2, 3, 4$, pseudo-signal are generated as $x_{5t} = \varepsilon_{5t} + \kappa x_{1t}$, and $x_{6t} = \varepsilon_{6t} + \kappa x_{2t}$, and pure noise variables are generated as $x_{it} = (\varepsilon_{i-1,t} + \varepsilon_{it})/\sqrt{2}$, for $i > 6$, where, as before, g_t, and ε_{it} are independent draws from $N\,(0,1)$ or $\left[\chi_t^2(2) - 2\right]/2$. We set $\kappa = 1.33$ (to achieve 80% correlation between the signal and the pseudo-signal variables).

DGP-II(b) All noise variables collinear with signals: $\boldsymbol{x}_{nt} \sim IID\,(\mathbf{0}, \boldsymbol{\Sigma}_x)$ with the elements of $\boldsymbol{\Sigma}_x$ given by $0.5^{|i-j|}$, $1 \leq i, j \leq n$. We generate $\boldsymbol{x}_{nt}$ with Gaussian and non-Gaussian innovations. In particular, $\boldsymbol{x}_{nt} = \boldsymbol{\Sigma}_x^{1/2}\varepsilon_t$, where $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t}, ..., \varepsilon_{nT})'$, and ε_{it} are generated as independent draws from $N(0,1)$ or $[\chi_t^2(2) - 2]/2$.

5.1.3 Design III (zero net signal effects)

We consider designs that allow for some signal variables to have zero θ. y_t is generated by (40), $\boldsymbol{x}_{nt}$ is generated as in DGP-I(a), and the slope coefficients for the signal variables in (40) are selected so that $\theta_4 = 0$:

DGP-III The fourth signal variables has zero net effect: we set $\beta_1 = \beta_2 = \beta_3 = 1$ and $\beta_4 = -1.5$ This implies $\theta_i \neq 0$ for $i = 1, 2, 3$ and $\theta_i = 0$ for $i \geq 4$.

5.1.4 Design IV (zero net signal effects and pseudo-signal variables)

We allow for signal variables with zero θ as well as pseudo-signal variables with non-zero θ's.

DGP-IV(a) We generate $\boldsymbol{x}_{nt}$ in the same way as in DGP-II(a) which features two pseudo-signal variables. We generate slope coefficients β_i as in DGP-III to ensure $\theta_i \neq 0$ for $i = 1, 2, 3$, and $\theta_i = 0$ for $i = 4$.

DGP-IV(b) We generate $\boldsymbol{x}_{nt}$ in the same way as in DGP-II(b), where all noise variables are collinear with signals. We set $\beta_1 = -0.875$ and $\beta_2 = \beta_3 = \beta_4 = 1$. This implies $\theta_i = 0$ for $i = 1$ and $\theta_i > 0$ for all $i > 1$.

5.1.5 Design V (Many signal variables)

For this design the DGP (**DGP-V**) is given by

$$y_t = \sum_{i=1}^{n} \left(\frac{1}{i}\right)^2 x_{it} + \varsigma u_t, \tag{41}$$

where $\boldsymbol{x}_{nt}$ are generated as in design DGP-II(b), and u_t is generated in the same way as before. This design is inspired by the literature on approximately sparse models (Belloni, Chernozhukov, and Hansen (2014b)).

Autoregressive processes are generated with zero starting values and 100 burn-in periods. ς is set so that $R^2 = 30\%$, 50% or 70% (on average). The sample combinations, $n = (100, 200, 300)$ and $T = (100, 300, 500)$ are considered, and all experiments are carried out using $R_{MC} = 2,000$ replications.

5.2 Variable selection methods

We consider six variable selection procedures, namely OCMT, Lasso, Adaptive Lasso (A-Lasso), Hard thresholding, Sica, and boosting. The OCMT method is implemented as outlined in Section 3, where $c_p(n, \delta)$ is defined by (8) with $f(n, \delta) = n^{\delta}$ in the first stage and $f(n, \delta^*) = n^{\delta^*}$ in the subsequent stages. We use $p = 0.01$, and in line with the theoretical derivations we set $\delta = 1$ and $\delta^* = 2$. An online Supplement provides results for other choices of $p \in \{0.01, 0.05, 0.1\}$ and $(\delta, \delta^*) \in \{(1, 1.5), (1, 2)\}$.[5] It turns out that the choice of p is of second order importance. Penalised regressions are implemented using the same set of possible values for the penalisation parameter λ as in Zheng, Fan, and Lv (2014), and following the literature λ is selected using 10-fold cross-validation. All methods are described in detail in an online Supplement.

5.3 Monte Carlo results

Here we focus on the relative performance of Lasso, adaptive Lasso and OCMT methods, and provide the full set of results for all experiments and all six variable selection procedures in an online Supplement. We evaluate the small sample performance of individual methods, using the true positive rate (TPR) defined by (13), the false positive rate (FPR) defined by (14), the false discovery rate (FDR) defined by (15), the out-of-sample root mean square forecast error (RMSFE), and the root mean square error of $\tilde{\boldsymbol{\beta}}$ (RMSE$_{\tilde{\boldsymbol{\beta}}}$).[6] We find that no method uniformly outperforms in the set of experiments we consider. This is true for the full set of methods (OCMT, Lasso, adaptive Lasso, Hard thresholding, Sica and Boosting) reported in the Supplement. As a way of highlighting this point, in Table 1 we report results for DGP-I(d) with $\omega = 0.2$ and $R^2 = 30\%$, where Lasso clearly outperforms OCMT for $T = 100$ (the upper left panel), and for DGP-III with $R^2 = 70\%$, where OCMT clearly dominates Lasso (the right panel). For example, for $n = T = 100$, the RMSE$_{\tilde{\boldsymbol{\beta}}}$ of OCMT is about 60% larger than that of Lasso in the case of DGP-I(d), whereas for DGP-III the RMSE$_{\tilde{\boldsymbol{\beta}}}$ of Lasso is about three times as large as that of the OCMT. Adaptive Lasso has better FPR and FDR performance than Lasso, but worse TPR, RMSFE and RMSE$_{\tilde{\boldsymbol{\beta}}}$ performance. It is also interesting to point out that the relative performance of the Lasso, adaptive Lasso and OCMT methods could crucially depend on the sample size, especially the time dimension. For example, when T is increased from 100 in the upper panel of the table to $T = 300$ in the lower panel, RMSE$_{\tilde{\boldsymbol{\beta}}}$ of OCMT dominates the Lasso and adaptive Lasso in both DGPs. It is clear that the performance of individual methods can be quite different for individual experiments, and an average relative assessment of these methods seems to be in order.

Tables 2-4 report averaged summary statistics across the three choices of R^2 (30%, 50%, 70%) for each of the DGPs. Lasso's TPR is in the majority of experiments larger than OCMT's, but so is the FPR and FDR as Lasso tends to overestimate the number of signal variables, which is well known the literature. Adaptive Lasso in turn achieves better FPR and FDR outcomes compared with Lasso, but the performance of adaptive Lasso is worse

[5] Monte Carlo findings for the first stage of the OCMT procedure are available upon request.

[6] RMSE$_{\tilde{\boldsymbol{\beta}}}$ is the square root of the trace of the MSE matrix of $\tilde{\beta}$. Additional summary statistics, including the probabilities of selecting the true model, and the statistics summarizing the distribution of the number of selected covariates are reported in the online Supplement.

for TPR, RMSFE and RMSE$_{\tilde{\boldsymbol{\beta}}}$ in these experiments. Lasso and adaptive Lasso are never the best in all support recovery statistics (TPR, FPR and FDR) simultaneously in Tables 2-4, whereas OCMT outperforms in all three dimensions simultaneously in some instances (when $T > 100$). The reported RMSFE averages of Lasso are outperformed by OCMT in all instances in Tables 2-4, by about 0.7% to 5.3%. Findings for RMSE$_{\tilde{\boldsymbol{\beta}}}$ are not uniform with OCMT outperforming Lasso in 40 out of the 45 reported average RMSE$_{\tilde{\boldsymbol{\beta}}}$'s. The reported Lasso's RMSE$_{\tilde{\boldsymbol{\beta}}}$ averages are in the range 86% to 718% of the reported OCMT's averages. As mentioned in Remark 3, the power of the OCMT procedure rises with $\sqrt{T}\left|\theta_{i,(j)}\right|/\sigma_{e_i,(T)}\sigma_{x_i,(T)}$, hence the magnitude of $\theta_{i,(j)}$, T and R^2 are all important for the power of the OCMT. For instance, an increase in the collinearity among signal variables, which results in a larger $\theta_{i,(j)}$, will improve the performance of OCMT, but it will worsen the performance of Lasso, since a higher collinearity of signals diminishes the marginal contribution of signals to the fit of the model. The average number of stages in OCMT procedure, $\hat{P}_{n,T}$, is either close to one or close to two, depending on whether zero net effect signals are present in the design.

It is also interesting to note that the relative performance of OCMT, Lasso and adaptive Lasso methods tends to improve in OCMT's favor with n. For example, for $T = 100$, the relative performance of OCMT and Lasso, based on the average statistics reported in Table 2, increases in OCMT's favor by about 0.8% to 1.9% in the case of RMSFE, and by about 7% to 14% in the case of RMSE$_{\tilde{\boldsymbol{\beta}}}$, when n is increased from 100 to 300.

Moving on to consider the relative performance of adaptive Lasso, we note that it improves greatly upon the FPR and FDR performance of Lasso while still performing less well than OCMT for these statistics, most of the time. The exception is DGP-II where it performs better that both Lasso and OCMT for a considerable number of cases and especially when $T > 100$. The downside to this improvement compared to Lasso, is that Adaptive Lasso performs considerably worse that both Lasso and OCMT in terms of TPR, especially for small T, DGP-I and DGP-II.

Overall, the small sample evidence suggests that the OCMT method is a valuable alternative to penalised regressions, since it can outperform the penalised regressions, that have become the *de facto* benchmark in the literature, in some cases. Another advantage of the OCMT procedure, which could be important in some applications, is that it is very fast to compute, about 10^2 to 10^4 times faster than penalised regression methods.

The findings presented so far relate to experiments with Gaussian innovations and, with the exception of DGP-I(b), serially uncorrelated covariates. The online supplement presents additional experiments to investigate the robustness of the OCMT method to non-Gaussianity and highly serially correlated covariates. The effects of allowing for non-Gaussian innovations seem to be rather marginal. In contrast, the deterioration in performance due to serial correlation of covariates is much larger. This is because longer time series observations are needed to detect spurious correlation when the covariates are highly serially correlated.

6 Empirical Illustration

In this section we present an empirical illustration that highlights the utility of OCMT. In particular, we present a macroeconomic forecasting exercise for US GDP growth and CPI inflation using a large set of macroeconomic variables. The dataset is quarterly and comes from Stock and Watson (2012). We use the smaller dataset considered in Stock and Watson (2012), which contains 109 series. The series are transformed by taking logarithms and/or differencing following Stock and Watson (2012).[7] The transformed series span 1960Q3 to 2008Q4 and are collected in the vector $\boldsymbol{z}_t$. Our estimation period is from 1960Q3 to 1990Q2 (120 periods) while the forecast evaluation period is 1990Q3 to 2008Q4. We produce one step ahead forecasts using five different procedures: (a) $AR(1)$ benchmark; ($AR(1)$), (b) $AR(1)$ augmented with lagged principal components; (factor-augmented $AR(1)$); (c-d) Lasso and adaptive Lasso regressions of the target variable y_t (either US GDP growth or differenced CPI inflation) on y_{t-1}, lagged principal components, and $\boldsymbol{z}_{t-1}$. For Lasso and adaptive Lasso regressions, both the target variable and regressors are demeaned, and the regressors are normalised to have unit variances. (e) OCMT procedure is applied to regressions of y_t conditional on lagged principal components, with y_{t-1}, and elements of $\boldsymbol{z}_{t-1}$ considered one at a time. The procedure is then repeated to convergence after $\hat{P}_{n,T}$ stages defined in (11). Similarly to the MC section, we set $p = 0.01$, and $\delta = 1$ in the first stage of OCMT, and $\delta^* = 2$ in the subsequent stages.[8] In all three data-rich procedures (b) to (e), the principal components are selected in a rolling scheme by the PC_{p_1} Bai and Ng (2002) criterion (with the maximum number of PCs set to 5).

We then use each of the methods by applying a rolling forecasting scheme with a rolling window of 120 observations. It is important to note that all features of our analysis (such as, e.g., lag orders) can be considerably refined. However, our aim is simply to show the potential of OCMT, and not to produce the best forecast for the dependent variables we consider.

We evaluate the forecasting performance of the methods using relative RMSFE where the $AR(1)$ forecast is the benchmark. Relative RMSFE statistics for the whole evaluation period as well as for the pre-crisis subperiod (1990Q3-2007Q2) are reported in Table 5.[9] In the case of GDP growth forecasts, we note that factor-augmented AR, Lasso and OCMT methods perform better than the AR(1) benchmark. OCMT performs the best in the full evaluation sample, whereas Lasso leads in the pre-crisis subsample. Adaptive Lasso is the worst performer. However, the performance of the best methods is very close, especially during the pre-crisis subperiod. Interestingly, the inclusion frequency of lagged dependent

[7] For further details, see the online supplement of Stock and Watson (2012), in particular columns E and T of their Table B.1.

[8] RMSFEs are reasonably robust to the choice of p. Results for $p = 0.05, 0.1$ are reported in the online Supplement.

[9] Diebold-Mariano test statistics, for all pairwise method comparisons, and the variable selection frequencies for both $LASSO$ and $OCMT$ can be found in the online supplement. The RMSFE differences among the best performing methods are not generally statistically significant.

variable using the full evaluation sample is 20% using OCMT, while it is 0% in the case of Lasso. Results are different when inflation is considered. In this case, the inclusion frequency of the lagged dependent variable is 100% in both OCMT and Lasso methods. The differences in RMSFE in the case of inflation are relatively small. For the full evaluation period, OCMT and factor-augmented AR(1) perform about 5% better than the benchmark AR(1) and the Lasso, and about 14% better than the adaptive Lasso. Zooming in on the results for the pre-crisis sub-sample, OCMT, Lasso, and adaptive Lasso underperform the AR(1) benchmark, but the differences in relative performance of OCMT and Lasso methods continue to be rather small. In summary, we see that there is no method that uniformly outperforms all competitor methods and that OCMT is not far behind the best performing method.

7 Conclusion

Model specification and selection are recurring and fundamental topics in econometric analysis. Both problems have become considerably more difficult for large-dimensional datasets where the set of possible specifications rise exponentially with the number of available covariates. In the context of linear regression models, penalised regression has become the *de facto* benchmark method of choice. However, issues such as the choice of penalty function and tuning parameters remains contentious.

In this paper, we provide an alternative approach based on multiple testing that is computationally simple, fast, and effective for sparse regression functions. Extensive theoretical and Monte Carlo results highlight these properties and provide support for adding this method to the toolbox of the applied researcher. In particular, we find that, for moderate values of the R^2 of the true model, with the net effects for the signal variables above some minimum threshold, our proposed method outperforms existing penalised regression methods, whilst at the same time being computationally much faster by some orders of magnitude.

There are a number of avenues for future research. The extension of our set-up to models with weakly exogenous and persistent regressors is clearly important for economic applications. In addition, the possibility of weak and strong common factors affecting both the signal and noise variables is also an important extension of the current set of assumptions. A further possibility is to extend the idea of considering regressors individually to other testing frameworks, such as tests of forecasting ability. It is hoped that the theoretical results and the Monte Carlo evidence presented in this paper provide a basis for such further developments and empirical applications.

Table 1: Monte Carlo findings for two selected experiments

	DGP-I(d) ($\omega = 0.2$, $R^2 = 30\%$)				DGP-III ($R^2 = 70\%$)			
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT
$T = 100$								
$n = 100$								
TPR	1.000	0.874	0.675	0.432	1.000	0.999	0.988	0.993
FPR	0.000	0.068	0.017	0.000	0.000	0.144	0.015	0.000
FDR	0.000	0.559	0.250	0.007	0.000	0.732	0.175	0.004
RMSFE	3.968	4.185	4.213	4.283	1.296	1.456	1.371	1.305
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	0.848	1.982	2.649	3.180	0.142	0.975	0.787	0.306
$n = 200$								
TPR	1.000	0.844	0.662	0.372	1.000	0.998	0.989	0.989
FPR	0.000	0.050	0.016	0.000	0.000	0.103	0.019	0.000
FDR	0.000	0.649	0.368	0.010	0.000	0.797	0.313	0.004
RMSFE	3.968	4.231	4.275	4.318	1.301	1.503	1.396	1.312
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	0.848	2.342	3.366	3.445	0.141	1.185	0.807	0.366
$n = 300$								
TPR	1.000	0.836	0.666	0.335	1.000	0.996	0.981	0.988
FPR	0.000	0.040	0.015	0.000	0.000	0.082	0.019	0.000
FDR	0.000	0.691	0.441	0.012	0.000	0.825	0.390	0.004
RMSFE	3.967	4.267	4.332	4.357	1.300	1.549	1.431	1.314
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	0.851	2.512	3.857	3.589	0.137	1.408	0.996	0.408
$T = 300$								
$n = 100$								
TPR	1.000	0.999	0.962	0.991	1.000	1.000	1.000	1.000
FPR	0.000	0.078	0.009	0.000	0.000	0.152	0.006	0.000
FDR	0.000	0.571	0.123	0.002	0.000	0.755	0.059	0.002
RMSFE	3.903	3.976	3.965	3.907	1.276	1.317	1.283	1.276
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	0.279	0.697	0.830	0.363	0.044	0.231	0.098	0.045
$n = 200$								
TPR	1.000	0.998	0.963	0.987	1.000	1.000	1.000	1.000
FPR	0.000	0.050	0.009	0.000	0.000	0.099	0.011	0.000
FDR	0.000	0.629	0.203	0.003	0.000	0.801	0.130	0.002
RMSFE	3.897	3.984	3.963	3.903	1.276	1.331	1.291	1.276
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	0.275	0.785	0.885	0.398	0.046	0.303	0.132	0.046
$n = 300$								
TPR	1.000	0.999	0.968	0.986	1.000	1.000	1.000	1.000
FPR	0.000	0.038	0.008	0.000	0.000	0.077	0.012	0.000
FDR	0.000	0.657	0.241	0.002	0.000	0.824	0.175	0.003
RMSFE	3.902	4.001	3.976	3.907	1.277	1.339	1.298	1.277
$\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$	0.277	0.841	0.983	0.402	0.045	0.334	0.158	0.046

Notes: This table reports selected experiments using DGP-I(d) and DGP-III, given by (40), with Gaussian innovations and serially uncorrelated covariates. There are $k = 4$ signal variables, and ω is the average pair-wise correlation of the signal variables in DGP-I(d). See Section 5 for further details. TPR (FPR) is the true (false) positive rate. FDR is the false discovery rate. RMSFE is the root mean square forecast error, $\text{RMSE}_{\tilde{\boldsymbol{\beta}}}$ is the root mean square error of $\tilde{\boldsymbol{\beta}}$. Oracle method assumes that the identity of signal variables is known. Lasso is implemented using the same set of possible values for the penalisation parameter λ as in Zheng, Fan, and Lv (2014), and λ is selected using 10-fold cross-validation. Adaptive Lasso method is implemented as described in Section 2.8.4 of Buhlmann and van de Geer (2011) based on the implementation of the Lasso method described above. OCMT results are based on $p = 0.01$, $\delta = 1$ in the first stage, and $\delta^* = 2$ in the subsequent stages of the OCMT procedure. The complete set of findings is reported in an online Supplement.

Table 2: Summary of Monte Carlo results for experiments with Gaussian innovations, serially uncorrelated covariates, and T=100

Summary statistics are averaged across $R^2 = 30\%, 50\%, 70\%$

	DGP-I				**DGP-II**				**DGP-III**				**DGP-IV**				**DGP-V**			
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle*	Lasso	A-Lasso	OCMT
$n = 100$, $T = 100$																				
TPR	1.000	0.909	0.734	0.906	1.000	0.913	0.754	0.890	1.000	0.883	0.768	0.726	1.000	0.843	0.705	0.710	1.000	0.270	0.122	0.197
FPR	0.000	0.054	0.010	0.000	0.000	0.060	0.011	0.008	0.000	0.111	0.017	0.000	0.000	0.095	0.015	0.006	0.000	0.049	0.002	0.000
FDR	0.000	0.478	0.149	0.005	0.000	0.514	0.170	0.145	0.000	0.679	0.238	0.007	0.000	0.641	0.222	0.126	0.000	0.474	0.058	0.006
RMSFE	3.419	3.551	3.570	3.460	3.283	3.425	3.445	3.335	2.105	2.341	2.301	2.265	2.240	2.445	2.426	2.385	1.371	1.362	1.358	1.326
$RMSE_{\tilde{\beta}}$	1.471	1.542	2.879	1.786	0.989	1.499	2.455	1.619	0.404	2.324	2.340	2.145	0.458	2.027	2.235	1.889	0.420	0.183	0.168	0.140
$n = 200$, $T = 100$																				
TPR	1.000	0.900	0.739	0.888	1.000	0.908	0.760	0.874	1.000	0.846	0.740	0.698	1.000	0.812	0.686	0.679	1.000	0.251	0.128	0.190
FPR	0.000	0.040	0.010	0.000	0.000	0.044	0.011	0.004	0.000	0.075	0.017	0.000	0.000	0.065	0.014	0.003	0.000	0.037	0.004	0.000
FDR	0.000	0.562	0.228	0.006	0.000	0.595	0.253	0.140	0.000	0.740	0.355	0.009	0.000	0.708	0.317	0.115	0.000	0.583	0.113	0.007
RMSFE	3.419	3.580	3.602	3.473	3.281	3.455	3.477	3.344	2.106	2.390	2.347	2.287	2.238	2.484	2.460	2.399	1.372	1.371	1.360	1.326
$RMSE_{\tilde{\beta}}$	1.504	1.716	3.258	1.913	0.984	1.657	2.847	1.668	0.414	2.665	2.614	2.326	0.450	2.276	2.461	2.035	0.420	0.208	0.202	0.147
$n = 300$, $T = 100$																				
TPR	1.000	0.898	0.745	0.877	1.000	0.904	0.765	0.859	1.000	0.829	0.724	0.670	1.000	0.794	0.672	0.662	1.000	0.244	0.131	0.185
FPR	0.000	0.034	0.010	0.000	0.000	0.037	0.011	0.003	0.000	0.060	0.016	0.000	0.000	0.051	0.013	0.002	0.000	0.031	0.004	0.000
FDR	0.000	0.615	0.289	0.006	0.000	0.643	0.314	0.135	0.000	0.778	0.435	0.008	0.000	0.740	0.380	0.111	0.000	0.635	0.159	0.008
RMSFE	3.418	3.602	3.635	3.480	3.284	3.478	3.513	3.355	2.113	2.431	2.393	2.309	2.238	2.514	2.495	2.413	1.372	1.380	1.366	1.328
$RMSE_{\tilde{\beta}}$	1.466	1.831	3.548	1.924	0.979	1.773	3.162	1.726	0.410	2.878	2.854	2.448	0.444	2.437	2.665	2.100	0.424	0.227	0.233	0.149

Notes: DGPs I-IV are given by (40) and DGP V is given by (41). See notes to Table 2 for a brief summary of the reported statistics and a description of methods. In DGP V, TPR is computed assuming that covariates $i = 1, 2, ..., 11$ are the signal variables and FPR and FDR are computed assuming covariates $i > 11$ are the noise variables. Oracle* method in DGP V assumes the first 11 covariates are signal variables. The average number of OCMT stages is 1.001, 1.000, 1.561, 1.373, and 1.000, in DGPs I to V, respectively. See Section 5 for further details. The complete set of findings is reported in an online Supplement.

Table 3: Summary of Monte Carlo results for experiments with Gaussian innovations, serially uncorrelated covariates, and T=300

Summary statistics are averaged across $R^2 = 30\%, 50\%, 70\%$

	DGP-I				DGP-II				DGP-III				DGP-IV				DGP-V			
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle*	Lasso	A-Lasso	OCMT
$n = 100$, $T = 300$																				
TPR	1.000	0.987	0.916	0.999	1.000	0.993	0.944	1.000	1.000	0.998	0.984	0.997	1.000	0.983	0.942	0.957	1.000	0.359	0.147	0.300
FPR	0.000	0.054	0.005	0.000	0.000	0.062	0.005	0.014	0.000	0.146	0.010	0.000	0.000	0.126	0.010	0.014	0.000	0.052	0.001	0.000
FDR	0.000	0.461	0.065	0.003	0.000	0.508	0.077	0.225	0.000	0.740	0.127	0.002	0.000	0.695	0.121	0.244	0.000	0.435	0.016	0.003
RMSFE	3.362	3.408	3.407	3.363	3.226	3.274	3.269	3.234	2.072	2.141	2.103	2.074	2.198	2.266	2.241	2.219	1.311	1.314	1.328	1.299
$RMSE_{\tilde{\beta}}$	0.473	0.605	1.195	0.489	0.308	0.563	0.877	0.475	0.128	0.717	0.568	0.222	0.143	0.743	0.727	0.440	0.128	0.068	0.086	0.050
$n = 200$, $T = 300$																				
TPR	1.000	0.986	0.917	0.999	1.000	0.993	0.942	1.000	1.000	0.995	0.978	0.994	1.000	0.974	0.932	0.949	1.000	0.335	0.153	0.292
FPR	0.000	0.034	0.004	0.000	0.000	0.038	0.005	0.006	0.000	0.093	0.012	0.000	0.000	0.080	0.010	0.007	0.000	0.031	0.002	0.000
FDR	0.000	0.509	0.101	0.003	0.000	0.552	0.115	0.219	0.000	0.781	0.212	0.003	0.000	0.740	0.193	0.238	0.000	0.495	0.036	0.003
RMSFE	3.361	3.415	3.409	3.362	3.227	3.287	3.278	3.235	2.072	2.163	2.112	2.077	2.198	2.286	2.251	2.222	1.314	1.320	1.329	1.301
$RMSE_{\tilde{\beta}}$	0.469	0.637	1.222	0.486	0.308	0.601	0.933	0.475	0.128	0.921	0.648	0.290	0.142	0.893	0.801	0.478	0.128	0.071	0.087	0.048
$n = 300$, $T = 300$																				
TPR	1.000	0.986	0.919	0.999	1.000	0.991	0.944	0.999	1.000	0.992	0.972	0.991	1.000	0.968	0.926	0.945	1.000	0.321	0.156	0.285
FPR	0.000	0.026	0.004	0.000	0.000	0.029	0.005	0.004	0.000	0.071	0.011	0.000	0.000	0.061	0.010	0.004	0.000	0.024	0.002	0.000
FDR	0.000	0.539	0.129	0.003	0.000	0.578	0.142	0.217	0.000	0.799	0.263	0.003	0.000	0.759	0.238	0.235	0.000	0.535	0.053	0.004
RMSFE	3.362	3.421	3.414	3.363	3.226	3.290	3.279	3.234	2.070	2.174	2.121	2.077	2.199	2.298	2.260	2.224	1.313	1.321	1.328	1.300
$RMSE_{\tilde{\beta}}$	0.464	0.654	1.256	0.483	0.320	0.626	0.994	0.487	0.128	1.041	0.721	0.330	0.145	0.985	0.855	0.510	0.127	0.076	0.089	0.049

See notes to Table 2. The average number of OCMT stages is 1.000, 1.000, 1.983, 1.809, and 1.000, in DGPs I to V, respectively.

Table 4: Summary of Monte Carlo results for experiments with Gaussian innovations, serially uncorrelated covariates, and T=500

Summary statistics are averaged across $R^2 = 30\%, 50\%, 70\%$

	DGP-I				DGP-II				DGP-III				DGP-IV				DGP-V			
	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle	Lasso	A-Lasso	OCMT	Oracle*	Lasso	A-Lasso	OCMT
	$n = 100$, $T = 500$																			
TPR	1.000	0.996	0.959	1.000	1.000	0.999	0.982	1.000	1.000	1.000	0.999	1.000	1.000	0.998	0.983	0.983	1.000	0.408	0.166	0.350
FPR	0.000	0.054	0.003	0.000	0.000	0.060	0.004	0.015	0.000	0.153	0.008	0.000	0.000	0.134	0.008	0.016	0.000	0.056	0.001	0.000
FDR	0.000	0.459	0.044	0.002	0.000	0.503	0.052	0.259	0.000	0.755	0.096	0.002	0.000	0.715	0.093	0.277	0.000	0.428	0.016	0.002
RMSFE	3.349	3.377	3.373	3.350	3.218	3.246	3.240	3.223	2.069	2.108	2.079	2.069	2.193	2.232	2.211	2.202	1.300	1.304	1.320	1.294
RMSE$_{\tilde{\beta}}$	0.277	0.372	0.724	0.283	0.186	0.338	0.497	0.291	0.077	0.394	0.221	0.079	0.086	0.420	0.353	0.219	0.074	0.043	0.070	0.033
	$n = 200$, $T = 500$																			
TPR	1.000	0.996	0.960	1.000	1.000	0.999	0.983	1.000	1.000	1.000	0.999	1.000	1.000	0.995	0.980	0.978	1.000	0.383	0.173	0.340
FPR	0.000	0.033	0.004	0.000	0.000	0.038	0.004	0.007	0.000	0.097	0.011	0.000	0.000	0.084	0.010	0.008	0.000	0.033	0.002	0.000
FDR	0.000	0.503	0.075	0.003	0.000	0.550	0.085	0.254	0.000	0.793	0.182	0.002	0.000	0.757	0.169	0.275	0.000	0.484	0.032	0.002
RMSFE	3.349	3.381	3.374	3.349	3.214	3.249	3.238	3.219	2.060	2.111	2.076	2.060	2.194	2.245	2.217	2.204	1.302	1.308	1.321	1.296
RMSE$_{\tilde{\beta}}$	0.281	0.397	0.751	0.289	0.184	0.361	0.519	0.289	0.078	0.497	0.259	0.080	0.086	0.514	0.388	0.236	0.075	0.047	0.071	0.034
	$n = 300$, $T = 500$																			
TPR	1.000	0.996	0.961	1.000	1.000	0.999	0.984	1.000	1.000	1.000	0.999	1.000	1.000	0.993	0.977	0.974	1.000	0.370	0.179	0.335
FPR	0.000	0.025	0.003	0.000	0.000	0.028	0.004	0.005	0.000	0.074	0.011	0.000	0.000	0.064	0.011	0.005	0.000	0.025	0.002	0.000
FDR	0.000	0.526	0.096	0.003	0.000	0.571	0.115	0.250	0.000	0.814	0.233	0.002	0.000	0.778	0.222	0.272	0.000	0.516	0.049	0.003
RMSFE	3.349	3.385	3.377	3.350	3.217	3.256	3.243	3.222	2.068	2.127	2.087	2.068	2.190	2.248	2.219	2.202	1.301	1.308	1.318	1.296
RMSE$_{\tilde{\beta}}$	0.279	0.404	0.769	0.287	0.181	0.370	0.544	0.285	0.077	0.570	0.280	0.079	0.084	0.574	0.427	0.247	0.075	0.049	0.071	0.033

See notes to Table 2. The average number of OCMT stages is 1.000, 1.000, 2.000, 1.914, and 1.000, in DGPs I to V, respectively.

Table 5: RMSFE performance of the AR, factor-augmented AR, Lasso and OCMT methods

Evaluation sample:	Full 1990Q3-2008Q4		Pre-crisis 1990Q3-2007Q2	
	RMSFE ($\times 100$)	Relative RMSFE	RMSFE ($\times 100$)	Relative RMSFE
	Real output growth			
$AR(1)$ benchmark	0.560	1.000	0.504	1.000
Factor-augmented $AR(1)$	0.488	0.870	0.467	0.927
Lasso	0.507	0.905	0.463	0.918
Adaptive Lasso	0.576	1.028	0.515	1.021
OCMT	0.487	0.869	0.464	0.920
	Inflation			
$AR(1)$ benchmark	0.655	1.000	0.469	1.000
Factor-augmented $AR(1)$	0.621	0.949	0.452	0.965
Lasso	0.655	1.001	0.488	1.040
Adaptive Lasso	0.715	1.093	0.518	1.105
OCMT	0.626	0.957	0.477	1.017

Notes: RMSFE is computed using a rolling forecasting scheme with a rolling window of 120 observations. We use the smaller dataset considered in Stock and Watson (2012) which contains 109 series. The series are transformed by taking logarithms and/or differencing following Stock and Watson (2012). The transformed series span 1960Q3 to 2008Q4 and are collected in the vector $\boldsymbol{z}_t$. Set of regressors in Lasso and adaptive-Lasso contains y_{t-1} (lagged target variable), $\boldsymbol{z}_{t-1}$, and a lagged set of principal components obtained from the large dataset given by $(y_t, \boldsymbol{z}_t')'$. OCMT procedure is applied to regressions of y_t conditional on lagged principal components, with y_{t-1}, and elements of $\boldsymbol{z}_{t-1}$ considered one at a time. OCMT is reported $p = 0.01$ and for $\delta = 1$ in the first stage, and $\delta^* = 2$ in the subsequent stages of the OCMT procedure, similarly to the MC section. The number of principal components in the factor-augmented $AR(1)$, Lasso, adaptive-Lasso, and OCMT methods is determined in a rolling scheme by using criterion PC_{p_1} of Bai and Ng (2002) (with the maximum number of PCs set to 5). See Section 5 and the Supplement for further details.

A Appendix

For further use throughout this appendix we define the following events. The event of choosing the pseudo true model, $\mathcal{A}_0$ defined in (26), will be written as

$$\mathcal{A}_0 = \mathcal{H} \cap \mathcal{G}, \tag{A.1}$$

where

$$\mathcal{H} = \left\{ \sum_{i=1}^{k} I\left(\widehat{\beta_i \neq 0}\right) = k \right\}, \tag{A.2}$$

is the event that all signals are selected, and

$$\mathcal{G} = \left\{ \sum_{i=k+k^*+1}^{n} I\left(\widehat{\beta_i \neq 0}\right) = 0 \right\}, \tag{A.3}$$

is the event that no noise variable is selected. We also denote the event that exactly j noise variables are selected by $\mathcal{G}_j$

$$\mathcal{G}_j = \left\{ \sum_{i=k+k^*+1}^{n} I\left(\widehat{\beta_i \neq 0}\right) = j \right\}, \text{ for } j = 0, 1, ..., n-k-k^*, \tag{A.4}$$

with $\mathcal{G} \equiv \mathcal{G}_0$. For the analysis of different stages of OCMT, we also introduce the event $\mathcal{B}_{i,s}$, which is the event that variable i is selected at the s^{th} stage of the OCMT procedure.

$$\mathcal{L}_{i,s} = \cup_{h=1}^{s} \mathcal{B}_{i,h}, \tag{A.5}$$

$\mathcal{L}_{i,s}$ is the event that variable i is selected up to and including stage s, namely in any of the stages $j = 1, 2, ..., s$ of the OCMT procedure.

$$\mathcal{L}_s = \cap_{i=1}^{k} \mathcal{L}_{i,s}, \tag{A.6}$$

$\mathcal{L}_s$ is the event that all signal variables are selected up to and including stage s of the OCMT procedure. $\mathcal{T}_s$ is the event that the OCMT procedure stops after s stages or less.

$$\mathcal{D}_{s,T} = \left\{ \hat{k}_{n,T,(j)} \leq l_T, j = 1, 2, ..., s \right\}, \tag{A.7}$$

$\mathcal{D}_{s,T}$ is the event that the number of variables selected in the first s stages of OCMT ($\hat{k}_{n,T,(j)}$, $j = 1, 2, ..., s$) is smaller than or equal to l_T, where $l_T = \ominus\left(n^{\nu}\right)$ and ν satisfies $\epsilon < \nu < \kappa_1/3$. Note that when $T = \ominus\left(n^{\kappa_1}\right)$ then, under this definition of l_T, we have $l_T = \ominus\left(T^{\nu/\kappa_1}\right) = o\left(T^{1/3}\right)$ for $\nu < \kappa_1/3$.

A.1 Proof of Proposition 1

We recall that P_0 is a population quantity. This formally means that, to determine P_0, OCMT is carried out assuming $\Pr\left[\left|t_{\hat{\phi}_{i,(j)}}\right| > c_p(n,\delta) \,|\theta_{i,(j)} \neq 0\right] = 1$ and $\Pr\left[\left|t_{\hat{\phi}_{i,(j)}}\right| > c_p(n,\delta) \,|\theta_{i,(j)} = 0\right] = 0$ for all i, j. So, if $\theta_{i,(1)} \neq 0$, for all i for which $\beta_i \neq 0$, it obviously follows that $P_0 = 1$. Next, assume that the subset of signals in $\mathbf{X}_k$, such that for each element of this subset, $\theta_{i,(1)} = 0$, is not empty. Then, these signals will not be selected in the first stage of OCMT. By Lemma 1, it follows that the subset of signals for which $\theta_{i,(1)} = 0$ is smaller than the set of signals and therefore at least one signal will be picked up in the first OCMT stage. It then follows, by Lemma 1, that in the second OCMT stage, at least one signal, for which $\theta_{i,(1)} = 0$ will have $\theta_{i,(2)} \neq 0$. Therefore, such signal(s) will be picked up in the second stage. Proceeding recursively using Lemma 1, then follows that all signals for which $\theta_{i,(1)} = 0$, will satisfy $\theta_{i,(j)} \neq 0$ for some $j \leq k$, proving the proposition.[10]

A.2 Proofs of theorems and corollaries

This subsection contains the proofs of the main theorems and their corollaries. All theorems are proven based on the set of lemmas presented and proven in Section A.3. In particular, Lemmas 1-9 are auxiliary ones, mostly providing supporting results for the main lemma of the paper which is Lemma 10. This provides the basic exponential inequalities that underlie most of our results.

A.2.1 Proof of Theorem 1

Noting that $\mathcal{T}_k$ is the event that the OCMT procedure stops after k stages or less, we have

$$\Pr\left(\hat{P}_{n,T} > k\right) = \Pr\left(\mathcal{T}_k^c\right) = 1 - \Pr\left(\mathcal{T}_k\right),$$

where $\hat{P}_{n,T}$ is defined by (11). Substituting (A.120) of Lemma 12 for $\Pr\left(\mathcal{T}_k\right)$, we obtain,

$$\Pr\left(\hat{P}_{n,T} > k\right) = O\left(n^{1-\nu-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right],$$

for some $C_0, C_1 > 0$, any $\varkappa$ in $0 < \varkappa < 1$, and any ν in $0 \leq \epsilon < \nu < \kappa_1/3$, where $\kappa_1 > 0$ is a positive constant that defines the rate for $T = \ominus\left(n^{\kappa_1}\right)$ and ϵ in $0 \leq \epsilon < \min\{1, \kappa_1/3\}$ is a positive constant that defines the rate for $k^* = \ominus\left(n^{\epsilon}\right)$. But note that $O\left(n^{1-\nu-\varkappa\delta}\right)$ can be written equivalently as $O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)$. This follows since $1 - \kappa_1/3 - \varkappa\delta = 1 - (\kappa_1/3 - \varepsilon\delta) - (\varkappa + \varepsilon)\,\delta = 1 - \tilde{\nu} - \tilde{\varkappa}\delta$, where $\tilde{\nu} = \kappa_1/3 - \varepsilon\delta$ and $\tilde{\varkappa} = \varkappa + \varepsilon$, for $\varepsilon > 0$ sufficiently small.

[10] Note that in the proposition we have allowed for net effects that depend on T and can therefore be small, in line with Assumption 6 as long as they are not exactly zero. This is possible since Lemma 1 also allows for such net effects.

Specifically, setting $\varepsilon < \min\{1-\varkappa, (\kappa_1/3-\epsilon)/\delta\}$, it follows that $\tilde{\varkappa}$ and $\tilde{\nu}$ satisfy $0 < \tilde{\varkappa} < 1$ and $\epsilon < \tilde{\nu} < \kappa_1/3$, respectively, as required. Hence

$$\Pr\left(\hat{P}_{n,T} > k\right) = \Pr\left(\mathcal{T}_k^c\right) = O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right], \quad \text{(A.8)}$$

for some $C_0, C_1 > 0$ and any $\varkappa$ in $0 < \varkappa < 1$. Noting that $O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right] = O\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$ for any $0 < C_2 < C_1$, we have

$$\Pr\left(\hat{P}_{n,T} > k\right) = O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[\exp\left(-n^{C_2\kappa_1}\right)\right],$$

for some $C_2 > 0$, which establishes (27). Similarly, by (A.123) and noting that $n \geq n^{1-\nu}$ for $\nu \geq 0$, we also have (which is required subsequently)

$$\Pr\left(\mathcal{D}_{k,T}^c\right) = O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 T^{C_1\kappa_1}\right)\right], \quad \text{(A.9)}$$

for some $C_0, C_1 > 0$ and any $\varkappa$ in $0 < \varkappa < 1$.

Consider now (28), and note that

$$\Pr(\mathcal{A}_0^c) = \Pr(\mathcal{A}_0^c|\mathcal{D}_{k,T})\Pr(\mathcal{D}_{k,T}) + \Pr(\mathcal{A}_0^c|\mathcal{D}_{k,T}^c)\Pr(\mathcal{D}_{k,T}^c) \leq \Pr(\mathcal{A}_0^c|\mathcal{D}_{k,T}) + \Pr(\mathcal{D}_{k,T}^c), \quad \text{(A.10)}$$

where $\Pr(\mathcal{D}_{k,T}^c)$ is given by (A.9). Also using (A.1) we have $\mathcal{A}_0^c = \mathcal{H}^c \cup \mathcal{G}^c$, and hence

$$\begin{aligned}\Pr(\mathcal{A}_0^c|\mathcal{D}_{k,T}) &\leq \Pr\left(\mathcal{H}^c|\,\mathcal{D}_{k,T}\right) + \Pr\left(\mathcal{G}^c|\,\mathcal{D}_{k,T}\right) \\ &= A_{n,T} + B_{n,T},\end{aligned} \quad \text{(A.11)}$$

where $\mathcal{H}$ and $\mathcal{G}$ are given by (A.2) and (A.3), respectively. Therefore

$$\mathcal{H}^c = \left\{\sum_{i=1}^{k} I\left(\widehat{\beta_i \neq 0}\right) < k\right\}, \text{ and } \mathcal{G}^c = \left\{\sum_{i=k+k^*+1}^{n} I\left(\widehat{\beta_i \neq 0}\right) > 0\right\}. \quad \text{(A.12)}$$

Consider the terms $A_{n,T}$ and $B_{n,T}$, in turn:

$$A_{n,T} = \Pr\left(\mathcal{H}^c|\,\mathcal{D}_{k,T}\right) \leq \sum_{i=1}^{k}\Pr\left(I\left(\widehat{\beta_i \neq 0}\right) = 0\Big|\,\mathcal{D}_{k,T}\right). \quad \text{(A.13)}$$

But, the event $\left\{I\left(\widehat{\beta_i \neq 0}\right) = 0\Big|\,\mathcal{D}_{k,T}\right\}$ can occur only if $\left\{\cap_{j=1}^{k}\mathcal{B}_{i,j}^c\big|\,\mathcal{D}_{k,T}\right\}$ occurs, while $\left\{\cap_{j=1}^{k}\mathcal{B}_{i,j}^c\big|\,\mathcal{D}_{k,T}\right\}$ can occur without $\left\{I\left(\widehat{\beta_i \neq 0}\right) = 0\Big|\,\mathcal{D}_{k,T}\right\}$ occurring. Therefore,

$$\Pr\left(I\left(\widehat{\beta_i \neq 0}\right) = 0\Big|\,\mathcal{D}_{k,T}\right) \leq \Pr\left(\cap_{j=1}^{k}\mathcal{B}_{i,j}^c\big|\,\mathcal{D}_{k,T}\right). \quad \text{(A.14)}$$

Then,

$$\begin{aligned}\Pr\left(\cap_{j=1}^{k}\mathcal{B}_{i,j}^c\big|\,\mathcal{D}_{k,T}\right) = &\Pr\left(\mathcal{B}_{i,1}^c\big|\,\mathcal{D}_{k,T}\right) \times \Pr\left(\mathcal{B}_{i,2}^c\big|\,\mathcal{B}_{i,1}^c, \mathcal{D}_{k,T}\right) \\ &\times \Pr\left(\mathcal{B}_{i,3}^c\big|\,\mathcal{B}_{i,2}^c \cap \mathcal{B}_{i,1}^c, \mathcal{D}_{k,T}\right) \\ &\times \ldots \times \Pr\left(\mathcal{B}_{i,k}^c\big|\,\mathcal{B}_{i,k-1}^c \cap \ldots \cap \mathcal{B}_{i,1}^c, \mathcal{D}_{k,T}\right).\end{aligned} \quad \text{(A.15)}$$

But, by Proposition 1 we are guaranteed that for some j in $1 \leq j \leq k$, $\theta_{i,(j)} \neq 0$, $i = 1, 2, ..., k$. Therefore, for some j in $1 \leq j \leq k$

$$\Pr\left(\mathcal{B}_{i,j}^c \middle| \mathcal{B}_{i,j-1}^c \cap ... \cap \mathcal{B}_{i,1}^c, \mathcal{D}_{k,T}\right) = \Pr\left(\mathcal{B}_{i,j}^c \middle| \mathcal{B}_{i,j-1}^c \cap ... \cap \mathcal{B}_{i,1}^c, \theta_{i,(j)} \neq 0, \mathcal{D}_{k,T}\right),$$

and by (A.108) of Lemma 10,

$$\Pr\left(\mathcal{B}_{i,j}^c \middle| \mathcal{B}_{i,j-1}^c \cap ... \cap \mathcal{B}_{i,1}^c, \theta_{i,(j)} \neq 0, \mathcal{D}_{k,T}\right) = O\left[\exp\left(-C_0 T^{C_1}\right)\right], \text{ for } i = 1, 2, ..., k,$$

for some $C_0, C_1 > 0$. Therefore,

$$\Pr\left(I\left(\widehat{\beta_i \neq 0}\right) = 0 \middle| \mathcal{D}_{k,T}\right) = O\left[\exp\left(-C_0 T^{C_1}\right)\right], \text{ for } i = 1, 2, ..., k. \tag{A.16}$$

Substituting this result in (A.13), we have

$$A_{n,T} = \Pr\left(\mathcal{H}^c \middle| \mathcal{D}_{k,T}\right) \leq k \exp\left(-C_0 T^{C_1}\right). \tag{A.17}$$

Similarly, for $B_{n,T}$ we first note that

$$\begin{aligned} B_{n,T} =& \Pr\left(\sum_{i=k+k^*+1}^{n} I\left(\widehat{\beta_i \neq 0}\right) > 0 \middle| \mathcal{D}_{k,T}\right) = \Pr\left\{ \cup_{i=k+k^*+1}^{n} \left[I\left(\widehat{\beta_i \neq 0}\right) > 0\right] \middle| \mathcal{D}_{k,T}\right\} \\ \leq& \sum_{i=k+k^*+1}^{n} E\left[I\left(\widehat{\beta_i \neq 0}\right) | \mathcal{D}_{k,T}\right]. \end{aligned} \tag{A.18}$$

Also,

$$\begin{aligned} E\left[I\left(\widehat{\beta_i \neq 0}\right) | \mathcal{D}_{k,T}\right] =& E\left[I\left(\widehat{\beta_i \neq 0}\right) | \mathcal{D}_{k,T}, \mathcal{T}_k\right] \Pr\left(\mathcal{T}_k | \mathcal{D}_{k,T}\right) + E\left[I\left(\widehat{\beta_i \neq 0}\right) | \mathcal{D}_{k,T}, \mathcal{T}_k^c\right] \Pr\left(\mathcal{T}_k^c | \mathcal{D}_{k,T}\right) \\ \leq& E\left[I\left(\widehat{\beta_i \neq 0}\right) | \mathcal{D}_{k,T}, \mathcal{T}_k\right] + \Pr\left(\mathcal{T}_k^c | \mathcal{D}_{k,T}\right), \end{aligned}$$

since $E\left[I\left(\widehat{\beta_i \neq 0}\right) | \mathcal{D}_{k,T}, \mathcal{T}_k^c\right] \leq 1$. Hence

$$B_{n,T} \leq \sum_{i=k+k^*+1}^{n} E\left[I\left(\widehat{\beta_i \neq 0}\right) | \mathcal{D}_{k,T}, \mathcal{T}_k\right] + (n - k - k^*) \Pr\left(\mathcal{T}_k^c | \mathcal{D}_{k,T}\right).$$

Consider now the first term of the above and note that

$$\begin{aligned} \sum_{i=k+k^*+1}^{n} E\left[I\left(\widehat{\beta_i \neq 0}\right) | \mathcal{D}_{k,T}, \mathcal{T}_k\right] =& \sum_{i=k+k^*+1}^{n} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n, \delta) \left| \theta_{i,(1)} = 0, \mathcal{D}_{k,T}, \mathcal{T}_k\right.\right] \\ &+ \sum_{i=k+k^*+1}^{n} \sum_{j=2}^{k} \Pr\left[\left|t_{\hat{\phi}_{i,(j)}}\right| > c_p(n, \delta^*) \left| \theta_{i,(j)} = 0, \mathcal{D}_{k,T}, \mathcal{T}_k\right.\right], \end{aligned}$$

where we have made use of the fact that the net effect coefficients, $\theta_{i,(j)}$, of noise variables are zero for $i = k + k^* + 1, k + k^* + 2, ..., n$ and all j. Also by (A.107) of Lemma 10 and result (ii) of Lemma 2, we have

$$\sum_{i=k+k^*+1}^{n} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,\delta)\left|\theta_{i,(1)} = 0, \mathcal{D}_{k,T}, \mathcal{T}_k\right.\right] + \sum_{i=k+k^*+1}^{n}\sum_{s=2}^{k} \Pr\left[\left|t_{\hat{\phi}_{i,(s)}}\right| > c_p(n,\delta^*)\left|\theta_{i,(s)} = 0, \mathcal{D}_{k,T}, \mathcal{T}_k\right.\right]$$
$$\leq (n-k-k^*)\exp\left[\frac{-\varkappa c_p^2(n,\delta)}{2}\right] + (k-1)(n-k-k^*)\exp\left[\frac{-\varkappa c_p^2(n,\delta^*)}{2}\right] + O\left[n\exp\left(-C_0 T^{C_1}\right)\right]$$
$$= O\left(n^{1-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 T^{C_1}\right)\right].$$

Further, by (A.129),

$$n\Pr\left(\mathcal{T}_k^c|\mathcal{D}_{k,T}\right) = O\left(n^{2-\varkappa\delta^*}\right) + O\left[n^2\exp\left(-C_0 T^{C_1}\right)\right],$$

giving, overall,

$$\begin{aligned} B_{n,T} &= O\left(n^{1-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 T^{C_1}\right)\right] + O\left(n^{2-\varkappa\delta^*}\right) + O\left[n^2\exp\left(-C_0 T^{C_1}\right)\right] \\ &= O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left[n^2\exp\left(-C_0 T^{C_1}\right)\right], \end{aligned} \quad \text{(A.19)}$$

where the second equality follows by noting that $O\left[n\exp\left(-C_0 T^{C_1}\right)\right]$ is dominated by $O\left[n^2\exp\left(-C_0 T^{C_1}\right)\right]$, and $O\left(n^{1-\varkappa\delta^*}\right)$ is dominated by $O\left(n^{1-\varkappa\delta}\right)$ for $\delta^* > \delta > 0$. Substituting for $A_{n,T}$ and $B_{n,T}$ from (A.17) and (A.19) in (A.11) and using (A.10) we obtain

$$\Pr(\mathcal{A}_0^c) \leq O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left[n^2\exp\left(-C_0 T^{C_1}\right)\right] + \Pr(\mathcal{D}_{k,T}^c),$$

where $\Pr(\mathcal{D}_{k,T}^c)$ is already given by (A.9), and $k\exp\left(-C_0 T^{C_1}\right)$ is dominated by $O\left[n^2\exp\left(-C_0 T^{C_1}\right)\right]$. Hence, noting that $\Pr\left(\mathcal{A}_0\right) = 1 - \Pr(\mathcal{A}_0^c)$, then

$$\Pr\left(\mathcal{A}_0\right) = 1 + O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[n^2\exp\left(-C_0 T^{C_1}\right)\right], \quad \text{(A.20)}$$

since $O\left[n\exp\left(-C_0 T^{C_1}\right)\right]$ is dominated by $O\left[n^2\exp\left(-C_0 T^{C_1}\right)\right]$, and $O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right)$ is dominated by $O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)$, for $\delta^* > \delta > 0$. Result (28) now follows noting that $T = \ominus\left(n^{\kappa_1}\right)$ and that $O\left[n^2\exp\left(-C_0 n^{C_1\kappa_1}\right)\right] = O\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$ for some C_2 in $0 < C_2 < C_1$. If, in addition, $\delta > 1$, and $\delta^* > 2$, then $\Pr\left(\mathcal{A}_0\right) \to 1$, as $n,T \to \infty$, for any $\kappa_1 > 0$.

We establish result (30) next, before establishing results (29) and (31). Consider $FPR_{n,T}$ defined by (14), and note that the probability of noise or pseudo-signal variable i being selected in any stages of the OCMT procedure is given by $\Pr\left(\mathcal{L}_{i,n}\right)$, for $i = k+1, k+2, ..., n$. Then

$$\begin{aligned} E\left|FPR_{n,T}\right| &= \frac{\sum_{i=k+1}^{n}\Pr\left(\mathcal{L}_{i,n}\right)}{n-k} \\ &= \frac{\sum_{i=k+1}^{k+k^*}\Pr\left(\mathcal{L}_{i,n}\right)}{n-k} + \frac{\sum_{i=k+k^*+1}^{n}\Pr\left(\mathcal{L}_{i,n}\right)}{n-k}. \end{aligned} \quad \text{(A.21)}$$

Since $\sum_{i=k+1}^{k+k^*} \Pr(\mathcal{L}_{i,n}) \leq k^*$ then

$$E\left|FPR_{n,T}\right| \leq \frac{k^*}{n-k} + \frac{\sum_{i=k+k^*+1}^{n} \Pr(\mathcal{L}_{i,n})}{n-k}. \tag{A.22}$$

Note that

$$\begin{aligned}\frac{\sum_{i=k+k^*+1}^{n} \Pr(\mathcal{L}_{i,n})}{n-k} =& \frac{\sum_{i=k+k^*+1}^{n} \Pr(\mathcal{L}_{i,n}|\mathcal{D}_{k,T})}{n-k} \Pr(\mathcal{D}_{k,T}) \\ &+ \frac{\sum_{i=k+k^*+1}^{n} \Pr\left(\mathcal{L}_{i,n}|\mathcal{D}_{k,T}^c\right)}{n-k} \Pr\left(\mathcal{D}_{k,T}^c\right) \\ \leq& \frac{\sum_{i=k+k^*+1}^{n} \Pr(\mathcal{L}_{i,n}|\mathcal{D}_{k,T})}{n-k} + \Pr\left(\mathcal{D}_{k,T}^c\right). \end{aligned} \tag{A.23}$$

Furthermore

$$\begin{aligned}\Pr(\mathcal{L}_{i,n}|\mathcal{D}_{k,T}) &= \Pr(\mathcal{L}_{i,n}|\mathcal{D}_{k,T},\mathcal{T}_k)\Pr(\mathcal{T}_k) + \Pr(\mathcal{L}_{i,n}|\mathcal{D}_{k,T},\mathcal{T}_k^c)\Pr(\mathcal{T}_k^c) \\ &\leq \Pr(\mathcal{L}_{i,n}|\mathcal{D}_{k,T},\mathcal{T}_k) + \Pr(\mathcal{T}_k^c). \end{aligned} \tag{A.24}$$

An upper bound on $\Pr(\mathcal{T}_k^c) = \Pr\left(\hat{P}_{n,T} > k\right)$ is established in the first part of this proof, see (A.8). We focus on $\Pr(\mathcal{L}_{i,n}|\mathcal{D}_{k,T},\mathcal{T}_k)$ next. Due to the conditioning on the event $\mathcal{T}_k$, we have $\Pr(\mathcal{L}_{i,n}|\mathcal{D}_{k,T},\mathcal{T}_k) = \Pr(\mathcal{L}_{i,k}|,\mathcal{D}_{k,T},\mathcal{T}_k)$, and in view of (A.5) we obtain

$$\Pr\left[\mathcal{L}_{i,k}|\mathcal{D}_{k,T},\mathcal{T}_k\right] \leq \sum_{s=1}^{k} \Pr\left(\mathcal{B}_{i,s}|\theta_{i,(s)}=0,\mathcal{D}_{k,T},\mathcal{T}_k\right), \text{ for } i > k+k^*, \tag{A.25}$$

where we note that $\Pr(\mathcal{B}_{i,s}|\mathcal{D}_{k,T},\mathcal{T}_k) = \Pr\left(\mathcal{B}_{i,s}|\theta_{i,(s)}=0,\mathcal{D}_{k,T},\mathcal{T}_k\right)$, for $i > k+k^*$ since the net effect coefficients of the noise variables at any stage of OCMT are zero. Further, using (A.107) of Lemma 10, for $i = k+k^*+1, k+k^*+2, ..., n$, we have

$$\Pr\left(\mathcal{B}_{i,s}|\theta_{i,(s)}=0,\mathcal{D}_{k,T},\mathcal{T}_k\right) = \left\{\begin{array}{l} O\left\{\exp\left[\frac{-\varkappa c_p^2(n,\delta)}{2}\right]\right\} + O\left[\exp(-C_0 T^{C_1})\right],\ s=1 \\ O\left\{\exp\left[\frac{-\varkappa c_p^2(n,\delta^*)}{2}\right]\right\} + O\left[\exp(-C_0 T^{C_1})\right],\ s>1 \end{array}\right., \tag{A.26}$$

where $\varkappa = \left[(1-\pi)/(1+d_T)\right]^2$. Clearly $0 < \varkappa < 1$, since $0 < \pi < 1$, and d_T is a bounded positive sequence. Hence, given result (ii) of Lemma 2, for $i = k+k^*+1, k+k^*+2, ..., n$, we have

$$\sum_{s=1}^{k} \Pr\left(\mathcal{B}_{i,s}|\theta_{i,(s)}=0,\mathcal{D}_{k,T},\mathcal{T}_k\right) = O\left(n^{-\delta\varkappa}\right) + O\left(n^{-\delta^*\varkappa}\right) + O\left[\exp(-C_0 T^{C_1})\right].$$

Using this result in (A.25) and averaging across $i = k+k^*+1, k+k^*+2, ..., n$, we obtain

$$\frac{\sum_{i=k+k^*+1}^{n} \Pr(\mathcal{L}_{i,k}|\mathcal{D}_{k,T},\mathcal{T}_k)}{n-k} = O\left(n^{-\varkappa\delta}\right) + O\left(n^{-\varkappa\delta^*}\right) + O\left[\exp(-C_0 T^{C_1})\right]. \tag{A.27}$$

Overall, with $\delta^* > \delta$, and using $T = \ominus(n^{\kappa_1})$, $k^* = \ominus(n^{\epsilon})$, (A.8),(A.9), (A.22)-(A.24) and (A.27), we have

$$E\left|FPR_{n,T}\right| = \frac{k^*}{n-k} + O\left(n^{-\varkappa\delta}\right) + O\left(n^{-\varkappa\delta^*}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[\exp(-C_0 n^{C_1\kappa_1})\right] + O\left(n^{\epsilon-1}\right) + O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right].$$

But $O\left[\exp(-C_0 n^{C_1\kappa_1})\right]$ and $O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right]$ are dominated by $\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$ for some $0 < C_2 < C_1$. In addition, since $\delta^* > \delta$ and $\varkappa$ is positive, the terms $O\left(n^{-\varkappa\delta^*}\right)$ and $O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right)$ are dominated by $O\left(n^{-\varkappa\delta}\right)$ and $O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)$, respectively. Hence,

$$E\left|FPR_{n,T}\right| = \frac{k^*}{n-k} + O\left(n^{-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{\epsilon-1}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[\exp\left(-n^{C_2\kappa_1}\right)\right],$$

for some $C_2 > 0$, which completes the proof of (30).

To establish (29) we note from (13) that

$$E\left|TPR_{n,T}\right| = \frac{\sum_{i=1}^{k}\Pr\left[I\left(\widehat{\beta_i \neq 0}\right) = 1\right]}{k}. \tag{A.28}$$

But

$$\Pr\left[I\left(\widehat{\beta_i \neq 0}\right) = 1\right] = 1 - \Pr\left[I\left(\widehat{\beta_i \neq 0}\right) = 0\right],$$

and

$$\begin{aligned}\Pr\left[I\left(\widehat{\beta_i \neq 0}\right) = 0\right] &= \Pr\left[I\left(\widehat{\beta_i \neq 0}\right) = 0 \middle| \mathcal{D}_{k,T}\right]\Pr\left(\mathcal{D}_{k,T}\right) \\ &\quad + \Pr\left[I\left(\widehat{\beta_i \neq 0}\right) = 0 \middle| \mathcal{D}_{k,T}^c\right]\Pr\left(\mathcal{D}_{k,T}^c\right) \\ &\leq \Pr\left[I\left(\widehat{\beta_i \neq 0}\right) = 0 \middle| \mathcal{D}_{k,T}\right] + \Pr\left(\mathcal{D}_{k,T}^c\right).\end{aligned}$$

Using (A.16) and (A.9), and dropping the terms $O\left[\exp\left(-C_0 T^{C_1}\right)\right]$ and $O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right)$ that are dominated by $O\left[n\exp\left(-C_0 T^{C_1}\right)\right]$ and $O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)$, respectively (noting that $\delta^* > \delta > 0$) we obtain

$$\Pr\left[I\left(\widehat{\beta_i \neq 0}\right) = 0\right] = O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[n\exp\left(-C_0 T^{C_1}\right)\right], \text{ for } i = 1, 2, ..., k. \tag{A.29}$$

Hence,

$$\sum_{i=1}^{k}\Pr\left[I\left(\widehat{\beta_i \neq 0}\right) = 1\right] = k + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[n\exp\left(-C_0 T^{C_1}\right)\right],$$

which, after substituting this expression in (A.28) and noting that $T = \ominus(n^{\kappa_1})$ and $O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right] = O\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$ for some C_2 in $0 < C_2 < C_1$ yields

$$E\left|TPR_{n,T}\right| = 1 + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[\exp\left(-n^{C_2\kappa_1}\right)\right], \tag{A.30}$$

for some $C_2 > 0$, as required.

To establish (31) we note from (15) that

$$FDR_{n,T} = \frac{(n-k)\, FPR_{n,T}}{(n-k)\, FPR_{n,T} + kTPR_{n,T}}, \tag{A.31}$$

for $\sum_{i=1}^{n} I\left(\widehat{\beta_i \neq 0}\right) > 0$. Using (A.30) and Markov's inequality, we have

$$kTPR_{n,T} \to_p k, \tag{A.32}$$

if $\delta > 1 - \kappa_1/3$. Using (A.21), we have

$$(n-k)\, E\left(FPR_{n,T}\right) = \sum_{i=k+1}^{k+k^*} \Pr\left(\mathcal{L}_{i,n}\right) + \sum_{i=k+k^*+1}^{n} \Pr\left(\mathcal{L}_{i,n}\right). \tag{A.33}$$

Using the same arguments as in the derivation of (A.17), we have

$$\lim_{n,T\to\infty} \sum_{i=k+1}^{k+k*} \Pr\left(\mathcal{L}_{i,n}\right) = k^*.$$

Moreover, using (A.8),(A.9),(A.23), and (A.24), and noting $T = \ominus\left(n^{\kappa_1}\right)$, we also have

$$\sum_{i=k+k^*+1}^{n} \Pr\left(\mathcal{L}_{i,n}\right) = O\left(n^{1-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left(n^{2-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{2-\kappa_1/3-\varkappa\delta^*}\right)$$
$$+ O\left(n^{2-\varkappa\delta^*}\right) + O\left[n \exp(-C_0 n^{C_1\kappa_1})\right] + O\left[n^2 \exp\left(-C_0 n^{C_1\kappa_1}\right)\right],$$

for some $C_0, C_1 > 0$. Using the above results in (A.33), it then follows that,

$$\lim_{n,T\to\infty} (n-k)\, E\left(FPR_{n,T}\right) = k^*, \tag{A.34}$$

if $\delta > \max\{1, 2-\kappa_1/3\}$, $\delta^* > 2$, and so using again Markov's inequality, we have

$$(n-k)\, FPR_{n,T} \to_p k^*. \tag{A.35}$$

Using (A.32) and (A.35) we establish (31).

To prove (32), first note that regardless of the number of selected regressors, $\hat{k}_{n,T}$, $0 \leq \hat{k}_{n,T} \leq n$, and the orthogonal projection theorem can be used to show that the following upper bound applies

$$\left\|\tilde{\boldsymbol{u}}\right\|^2 \leq \left\|\boldsymbol{y}\right\|^2,$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_T)'$. In particular, this is a direct implication of the fact that that for any $K \geq 0$, we have

$$\min_{\beta_i, i=1,2,...,K} \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i x_{it}\right)^2 \leq \sum_{t=1}^{T} y_t^2.$$

We also note that if for two random variables $x, y > 0$ defined on a probability space, Ω,

$$\sup_{\omega \in \Omega} [y(\omega) - x(\omega)] \geq 0,$$

then $E(x) \leq E(y)$. The above results imply that $E \|\tilde{\boldsymbol{u}}\|^2 \leq E \|\boldsymbol{y}\|^2$. Also, by Assumptions 2 and 3, $E(y_t^2)$ is bounded, and so we have $E \|\boldsymbol{y}\|^2 = O(T)$, and therefore $E \|\tilde{\boldsymbol{u}}\|^2 = O(T)$.

Now let $\mathcal{A}_0$ be the set of pseudo-true models as defined in (26) and let $\mathcal{A}_0^c$ be its complement. Then

$$\frac{1}{T} E \|\tilde{\boldsymbol{u}}\|^2 = P(\mathcal{A}_0) \frac{1}{T} E\left(\|\tilde{\boldsymbol{u}}\|^2 \middle| \mathcal{A}_0 \right) + [1 - P(\mathcal{A}_0)] \frac{1}{T} E\left(\|\tilde{\boldsymbol{u}}\|^2 \middle| \mathcal{A}_0^c \right).$$

Noting that $E\left(\|\tilde{\boldsymbol{u}}\|^2 \middle| \mathcal{A}_0^c \right) \leq E \|\boldsymbol{y}\|^2 = O(T)$, we have

$$\begin{aligned} \frac{1}{T} E \|\tilde{\boldsymbol{u}}\|^2 &\leq P(\mathcal{A}_0) \frac{1}{T} E\left(\|\tilde{\boldsymbol{u}}\|^2 \middle| \mathcal{A}_0 \right) + [1 - P(\mathcal{A}_0)] \frac{E \|\boldsymbol{y}\|^2}{T} \\ &\leq P(\mathcal{A}_0) \frac{1}{T} E\left(\|\tilde{\boldsymbol{u}}\|^2 \middle| \mathcal{A}_0 \right) + [1 - P(\mathcal{A}_0)] C_0, \end{aligned} \tag{A.36}$$

where C_0 is a finite constant that does not depend on n and/or T. Now, using that $P(\mathcal{A}_0) \to 1$ for $\delta > 1$ and $\delta^* > 2$, and that

$$\frac{1}{T} E\left(\|\tilde{\boldsymbol{u}}\|^2 \middle| \mathcal{A}_0 \right) = \sigma^2 + O\left(\frac{1}{\sqrt{T}} \right),$$

in (A.36), we obtain

$$E\left(\frac{1}{T} \sum_{i=1}^{T} \tilde{u}_t^2 \right) \to \sigma^2, \tag{A.37}$$

as required. This completes the proof.

A.2.2 Proof of Theorem 2

Using (A.1) we have $\mathcal{A}_0^c = \mathcal{H}^c \cup \mathcal{G}^c$, where $\mathcal{H}^c$ and $\mathcal{G}^c$ are defined by (A.12). Further, since $\mathcal{H}^c \cup \mathcal{G}^c = \mathcal{H}^c \cup \mathcal{G}^c \cap (\mathcal{H}^c \cup \mathcal{H})$, then using the distributive law given by $(\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{A} \cup \mathcal{C}) = \mathcal{A} \cup (\mathcal{B} \cap \mathcal{C})$, for some events $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, we have $\mathcal{H}^c \cup \mathcal{G}^c = (\mathcal{H}^c \cup \mathcal{G}^c) \cap (\mathcal{H}^c \cup \mathcal{H}) = \mathcal{H}^c \cup (\mathcal{G}^c \cap \mathcal{H}) = \mathcal{H}^c \cup (\mathcal{H} \cap \mathcal{G}^c)$. Therefore,

$$\begin{aligned} \mathcal{A}_0^c &= \left\{ \sum_{i=1}^{k} I\left(\widehat{\beta_i \neq 0}\right) < k \right\} \cup \left\{ \left[\sum_{i=1}^{k} I\left(\widehat{\beta_i \neq 0}\right) = k \right] \cap \left[\sum_{i=k+k^*+1}^{n} I\left(\widehat{\beta_i \neq 0}\right) > 0 \right] \right\} \\ &= \mathcal{H}^c \cup (\mathcal{H} \cap \mathcal{G}^c). \end{aligned}$$

Further, $\mathcal{H}^c \cup (\mathcal{H} \cap \mathcal{G}^c) = \mathcal{H}^c \cup \left\{ \mathcal{H} \cap \left[\cup_{j=1}^{n-k-k^*} \mathcal{G}_j \right] \right\}$, where $\mathcal{G}_j$ is defined by (A.4). Moreover, note that

$$\mathcal{H} \cap \left[\cup_{j=1}^{n-k-k^*} \mathcal{G}_j \right] = \cup_{j=1}^{n-k-k^*} (\mathcal{H} \cap \mathcal{G}_j) = \left[\cup_{j=1}^{l_{\max}-k-k^*-1} (\mathcal{H} \cap \mathcal{G}_j) \right] \cup \left[\mathcal{H} \cap \left(\cup_{j=l_{\max}-k-k^*}^{n-k-k^*} \mathcal{G}_j \right) \right],$$

and that the events $\mathcal{A}_0$, $\mathcal{H}^c$, $\mathcal{H}\cap\mathcal{G}_1, \mathcal{H}\cap\mathcal{G}_2, ..., \mathcal{H}\cap\mathcal{G}_{l_{\max}-k-k^*-1}$, and $\left[\mathcal{H}\cap\left(\cup_{j=l_{\max}-k-k^*}^{n-k-k^*}\mathcal{G}_j\right)\right]$ are mutually exclusive and exhaustive. Therefore,

$$E\left(\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\right) = C_{n,T} + D_{n,T} + E_{n,T} + F_{n,T}, \tag{A.38}$$

where

$$C_{n,T} = E\left(\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\middle|\,\mathcal{A}_0\right)\Pr\left(\mathcal{A}_0\right), \tag{A.39}$$

$$D_{n,T} = E\left(\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\middle|\,\mathcal{H}^c\right)\Pr\left(\mathcal{H}^c\right), \tag{A.40}$$

$$E_{n,T} = \sum_{j=1}^{l_{\max}-k-k^*-1} E\left(\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\middle|\,\mathcal{H}\cap\mathcal{G}_j\right)\Pr\left(\mathcal{H}\cap\mathcal{G}_j\right), \tag{A.41}$$

and

$$F_{n,T} = E\left[\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\middle|\,\mathcal{H}\cap\left(\cup_{j=l_{\max}-k-k^*}^{n-k-k^*}\mathcal{G}_j\right)\right]\Pr\left[\mathcal{H}\cap\left(\cup_{j=l_{\max}-k-k^*}^{n-k-k^*}\mathcal{G}_j\right)\right]. \tag{A.42}$$

We consider the terms $C_{n,T}, D_{n,T}, E_{n,T}$ and $F_{n,T}$ in turn, starting with $C_{n,T}$. By (A.20) we have

$$\Pr\left(\mathcal{A}_0\right) = 1 + O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[n^2\exp\left(-C_0T^{C_1}\right)\right].$$

Also, since $\mathcal{A}_0$ contains k signal variables, at most k^* pseudo signal variables, and no noise variables, then using (A.146) from Lemma 15, with $l_T = k + k^* + 1$, it follows that

$$E\left(\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\middle|\,\mathcal{A}_0\right) = O\left(\frac{(k+k^*+1)^2}{\sqrt{T}}\right),$$

and hence

$$C_{n,T} = O\left[\frac{(k+k^*+1)^2}{\sqrt{T}}\right]\left\{1 + O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]\right\}. \tag{A.43}$$

Next, consider $D_{n,T}$ given by (A.40) and note that by applying (A.147) of Lemma 15 to the regression of y_t on the $\hat{k}_{n,T} \leq l_{\max} - 1$ selected variables and a constant term, for some finite positive constant C_0, we have

$$E\left(\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\middle|\,\mathcal{H}^c\right) \leq C_0\left(\frac{l_{\max}^2}{\sqrt{T}} + l_{\max}\right).$$

where $l_{\max}$ denotes the imposed upper bound on the number of regressors including the constant term $(\hat{k}_{n,T}+1)$. Further,

$$\begin{aligned}\Pr(\mathcal{H}^c) &= \Pr(\mathcal{H}^c|\mathcal{D}_{k,T})\Pr(\mathcal{D}_{k,T}) + \Pr(\mathcal{H}^c|\mathcal{D}_{k,T}^c)\Pr(\mathcal{D}_{k,T}^c)\\ &\leq \Pr(\mathcal{H}^c|\mathcal{D}_{k,T}) + \Pr(\mathcal{D}_{k,T}^c),\end{aligned}$$

and using (A.9) and (A.17) we have

$$\Pr\left(\mathcal{H}^c\right) = O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0T^{C_1}\right)\right] + O\left[\exp\left(-C_0T^{C_1}\right)\right],$$

for some $C_0, C_1 > 0$. Therefore, noting that $O\left[\exp\left(-C_0T^{C_1}\right)\right]$ is dominated by $O\left[n\exp\left(-C_0T^{C_1}\right)\right]$, we have

$$D_{n,T} = O\left(\frac{l_{\max}^2}{\sqrt{T}} + l_{\max}\right)\left\{O\left(n^{1-\kappa_1/3-\varkappa\delta}\right) + O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0T^{C_1}\right)\right]\right\}. \tag{A.44}$$

Consider $E_{n,T}$ given by (A.41) next.

$$\begin{aligned} E_{n,T} &= \sum_{j=1}^{l_{\max}-k-k^*-1} E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{H}\cap\mathcal{G}_j\right)\Pr\left(\mathcal{H}\cap\mathcal{G}_j\right) \\ &\leq \left[\max_{j=1,2,\ldots,l_{\max}-k-k^*-1} E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{H}\cap\mathcal{G}_j\right)\right]\sum_{j=1}^{l_{\max}-k-k^*-1}\Pr\left(\mathcal{H}\cap\mathcal{G}_j\right). \end{aligned}$$

But, $\mathcal{H}\cap\mathcal{G}_j$, for $j = 1, 2, ..., l_{\max} - k - k^* - 1$ are mutually exclusive, and therefore

$$\sum_{j=1}^{l_{\max}-k-k^*-1}\Pr\left(\mathcal{H}\cap\mathcal{G}_j\right) = \Pr\left[\mathcal{H}\cap\left(\cup_{j=1}^{l_{\max}-k-k^*-1}\mathcal{G}_j\right)\right],$$

and

$$E_{n,T} \leq \left[\max_{j=1,2,\ldots,l_{\max}-k-k^*-1} E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{H}\cap\mathcal{G}_j\right)\right]\Pr\left[\mathcal{H}\cap\left(\cup_{j=1}^{l_{\max}-k-k^*-1}\mathcal{G}_j\right)\right]$$

By (A.146) of Lemma 15, (with $l_T = k + k^* + j + 1$, since the event $\mathcal{H}\cap\mathcal{G}_j$ means that k signal variables, at most k^* pseudo signal variables and j noise variables are selected by OCMT)

$$E\left[\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{H}\cap\mathcal{G}_j\right] = O\left(\frac{(k+k^*+j+1)^2}{\sqrt{T}}\right), \text{ for } j = 1, 2, ..., l_{\max} - k - k^* - 1,$$

which leads to

$$\max_{j=1,2,\ldots,l_{\max}-k-k^*-1} E\left(\left\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right\| \middle| \mathcal{H}\cap\mathcal{G}_j\right) = O\left(\frac{l_{\max}^2}{\sqrt{T}}\right).$$

In addition,

$$\Pr\left[\left(\mathcal{H}\cap\left(\cup_{j=1}^{l_{\max}-k-k^*-1}\mathcal{G}_j\right)\right)\right] \leq \Pr\left(\cup_{j=1}^{l_{\max}-k-k^*-1}\mathcal{G}_j\right) \leq \Pr\left(\mathcal{G}^c\right) \leq \Pr\left(\mathcal{G}^c|\mathcal{D}_{k,T}\right) + \Pr\left(\mathcal{D}_{k,T}^c\right),$$

and using (A.9), (A.18) and (A.19), we have

$$\begin{aligned} \Pr\left[\mathcal{H}\cap\left(\cup_{j=1}^{l_{\max}-k-k^*-1}\mathcal{G}_j\right)\right] =& O\left(n^{1-\delta\varkappa}\right) + O\left(n^{2-\delta^*\varkappa}\right) + O\left(n^{1-\kappa_1/3-\delta\varkappa}\right) + O\left(n^{1-\kappa_1/3-\delta^*\varkappa}\right) \\ &+ O\left[n\exp\left(-C_0T^{C_1}\right)\right] + O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]. \end{aligned}$$

Since the terms $O\left(n^{1-\kappa_1/3-\delta\varkappa}\right)$, $O\left(n^{1-\kappa_1/3-\delta^*\varkappa}\right)$ and $O\left[n\exp\left(-C_0T^{C_1}\right)\right]$, are dominated by the terms $O\left(n^{1-\delta\varkappa}\right)$, $O\left(n^{2-\delta^*\varkappa}\right)$ and $O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]$, respectively, we obtain

$$\Pr\left[\mathcal{H}\cap\left(\cup_{j=1}^{l_{\max}-k-k^*-1}\mathcal{G}_j\right)\right]=O\left(n^{1-\delta\varkappa}\right)+O\left(n^{2-\delta^*\varkappa}\right)+O\left[n^2\exp\left(-C_0T^{C_1}\right)\right].$$

So overall,

$$E_{n,T}=O\left(\frac{l_{\max}^2}{\sqrt{T}}\right)\left\{O\left(n^{1-\delta\varkappa}\right)+O\left(n^{2-\delta^*\varkappa}\right)+O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]\right\}. \tag{A.45}$$

Consider the last term $F_{n,T}$ given by (A.41) next. In the case of the event $\mathcal{H}\cap\left(\cup_{j=l_{\max}-k-k^*}^{n-k-k^*}\mathcal{G}_j\right)$ the restriction on the number of regressors ($\leq l_{\max}$) that are allowed to enter the final regression for $\tilde{\boldsymbol{\beta}}_n$ can be binding, and regardless how this restriction is implemented, result (A.147) of Lemma 15 always applies, and therefore

$$E\left[\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\middle|\mathcal{H}\cap\left(\cup_{j=l_{\max}-k-k^*}^{n-k-k^*}\mathcal{G}_j\right)\right]=O\left(\frac{l_{\max}^2}{\sqrt{T}}\right)+O\left(l_{\max}\right).$$

The event $\mathcal{H}\cap\left(\cup_{j=l_{\max}-k-k^*}^{n-k-k^*}\mathcal{G}_j\right)$ can only occur if k signal variables, j noise variables, for some $j\geq l_{\max}-k-k^*$, and any subset of the pseudo-signal variables are selected. In other words, the event $\Pr\left[\mathcal{H}\cap\left(\cup_{j=l_{\max}-k-k^*}^{n-k-k^*}\mathcal{G}_j\right)\right]$ can only occur if, at least, $j+k\geq l_{\max}-k^*$ variables are selected or, equivalently, if $\hat{k}_{n,T}>l_{\max}-k^*-1$. Therefore,

$$\Pr\left\{\mathcal{H}\cap\left[\cup_{j=l_{\max}-k-k^*}^{n-k-k^*}\mathcal{G}_j\right]\right\}\leq\Pr\left(\hat{k}_{n,T}>l_{\max}-k^*-1\right). \tag{A.46}$$

Using (A.132), we have

$$\begin{aligned}\Pr\left(\hat{k}_{n,T}>l_{\max}-k^*-1\right)=&\Pr\left(\hat{k}_{n,T}-k-k^*>l_{\max}-k-2k^*-1\right)\\ =&O\left(\frac{n^{1-\delta\varkappa}}{l_{\max}-k-2k^*-1}\right)+O\left(\frac{n^{2-\delta^*\varkappa}}{l_{\max}-k-2k^*-1}\right)+O\left(n^{1-\nu-\delta\varkappa}\right)\\ &+O\left(n^{1-\nu-\delta^*\varkappa}\right)+O\left[\frac{n^2}{l_{\max}-k-2k^*-1}\exp\left(-C_0n^{C_1\kappa_1}\right)\right].\end{aligned}$$

Combining the above results gives

$$F_{n,T}=\left[O\left(\frac{l_{\max}^2}{\sqrt{T}}\right)+O\left(l_{\max}\right)\right]\left[\begin{array}{c}O\left(\frac{n^{1-\delta\varkappa}}{l_{\max}-k-2k^*-1}\right)+O\left(\frac{n^{2-\delta^*\varkappa}}{l_{\max}-k-2k^*-1}\right)+O\left(n^{1-\nu-\delta\varkappa}\right)\\ +O\left(n^{1-\nu-\delta^*\varkappa}\right)+O\left[\frac{n^2}{l_{\max}-k-2k^*-1}\exp\left(-C_0n^{C_1\kappa_1}\right)\right]\end{array}\right]. \tag{A.47}$$

Using (A.43), (A.44), (A.45), and (A.47) in (A.38), we obtain

$$
\begin{aligned}
E\left(\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\right)=&O\left(\frac{(k+k^*+1)^2}{\sqrt{T}}\right)\left\{\begin{array}{c} 1+O\left(n^{1-\delta\varkappa}\right)+O\left(n^{2-\delta^*\varkappa}\right) \\ +O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)+O\left[n^2\exp\left(-C_0T^{C_1}\right)\right] \end{array}\right\} \\
&+O\left(\frac{l_{\max}^2}{\sqrt{T}}+l_{\max}\right)\left\{O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)+O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right)+O\left[n\exp\left(-C_0T^{C_1}\right)\right]\right\} \\
&+O\left(\frac{l_{\max}^2}{\sqrt{T}}\right)\left\{O\left(n^{1-\delta\varkappa}\right)+O\left(n^{2-\delta^*\varkappa}\right)+O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]\right\} \\
&+\left[O\left(\frac{l_{\max}^2}{\sqrt{T}}\right)+O\left(l_{\max}\right)\right]\left[\begin{array}{c} O\left(\frac{n^{1-\delta\varkappa}}{l_{\max}-k-2k^*-1}\right)+O\left(\frac{n^{2-\delta^*\varkappa}}{l_{\max}-k-2k^*-1}\right)+O\left(n^{1-\nu-\delta\varkappa}\right) \\ +O\left(n^{1-\nu-\delta^*\varkappa}\right)+O\left[\frac{n^2}{l_{\max}-k-2k^*-1}\exp\left(-C_0n^{C_1\kappa_1}\right)\right] \end{array}\right].
\end{aligned}
$$

This expression can be simplified by noting that k is finite, $l_{\max}=\ominus\left(n^{\kappa_2}\right)$ with $\kappa_2>0$, $k^*=\ominus\left(T^{\epsilon}\right)$ with $\kappa_2>\epsilon>0$, $0<\delta<\delta^*$, $T=\ominus\left(n^{\kappa_1}\right)$ with $\kappa_1>0$. In addition, using similar arguments as in the derivation of (A.8), the term $O\left(n^{1-\nu-\delta\varkappa}\right)$ and $O\left(n^{1-\nu-\delta^*\varkappa}\right)$ can be replaced with $O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)$ and $O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right)$, respectively. Hence, we have

$$
\begin{aligned}
E\left(\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\right)=&O\left(n^{2\epsilon-\kappa_1/2}\right)\left\{\begin{array}{c} 1+O\left(n^{1-\delta\varkappa}\right)+O\left(n^{2-\delta^*\varkappa}\right) \\ +O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)+O\left[n^2\exp\left(-C_0n^{\kappa_1C_1}\right)\right] \end{array}\right\} \\
&+O\left(n^{2\kappa_2-\kappa_1/2}+n^{\kappa_2}\right)\left\{O\left(n^{1-\kappa_1/3-\varkappa\delta}\right)+O\left(n^{1-\kappa_1/3-\varkappa\delta^*}\right)+O\left[n\exp\left(-C_0n^{\kappa_1C_1}\right)\right]\right\} \\
&+O\left(n^{2\kappa_2-\kappa_1/2}\right)\left\{O\left(n^{1-\delta\varkappa}\right)+O\left(n^{2-\delta^*\varkappa}\right)+O\left[n^2\exp\left(-C_0n^{\kappa_1C_1}\right)\right]\right\} \\
&+\left[O\left(n^{2\kappa_2-\kappa_1/2}\right)+O\left(n^{\kappa_2}\right)\right]\left[\begin{array}{c} O\left(n^{1-\kappa_2-\delta\varkappa}\right)+O\left(n^{2-\kappa_2-\delta^*\varkappa}\right)+O\left(n^{1-\kappa_1/3-\delta\varkappa}\right) \\ +O\left(n^{1-\kappa_1/3-\delta^*\varkappa}\right)+O\left[n^{2-\kappa_2}\exp\left(-C_0n^{C_1\kappa_1}\right)\right] \end{array}\right].
\end{aligned}
$$

The terms of the form $O\left[n^a\exp\left(-C_0n^{\kappa_1C_1}\right)\right]$ for some a and some $C_0,C_1>0$ are dominated by a single term $O\left[\exp\left(-n^{C_2\kappa_1}\right)\right]$ for some C_2 in $0<C_2<C_1$. Simplifying the expression above and removing some of the terms that are dominated, we obtain

$$
\begin{aligned}
E\left(\left\|\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_n\right\|\right)=&O\left(n^{2\epsilon-\kappa_1/2}\right)+O\left(n^{1-\delta\varkappa}\right)+O\left(n^{2-\delta^*\varkappa}\right)+O\left(n^{1-\delta\varkappa+2\kappa_2-\kappa_1/2}\right) \\
&+O\left(n^{2-\delta^*\varkappa+2\kappa_2-\kappa_1/2}\right)+O\left[\exp\left(-n^{C_2\kappa_1}\right)\right],
\end{aligned}
$$

for some $C_2>0$, as required. This completes the proof.

A.2.3 Proof of Theorem 3

A proof of Theorem 3 is provided in Section A of the online theory supplement.

A.2.4 Proofs of the results for the case when $P_0=1$

Result (35) follows from (23), and (36) follows from the analysis preceding Theorem 1, using (24) and (25). Result on $FDR_{n,T}$ continues to hold using the same arguments as in the proof of (31).

To obtain $\Pr(\mathcal{A}_0)$ we follow the derivations in the proof of the multi-stage version of OCMT provided in Section A.2.1, but note that we only need to consider the terms from the first stage of OCMT. Similarly to (A.11) and without the need to condition on $\mathcal{D}_{k,T}$, we have

$$\begin{aligned}\Pr(\mathcal{A}_0^c) &\leq \Pr\left(\sum_{i=1}^{k} I\left(\widehat{\beta_i \neq 0}\right) < k\right) + \Pr\left(\sum_{i=k+k^*+1}^{n} I\left(\widehat{\beta_i \neq 0}\right) > 0\right) \\ &= A_{n,T} + B_{n,T}.\end{aligned}$$

noting that $I\left(\widehat{\beta_i \neq 0}\right) = I_{(1)}\left(\widehat{\beta_i \neq 0}\right)$. Also, as with (A.17) and (A.18), we have

$$A_{n,T} \leq k \exp\left(-C_1 T^{C_2}\right).$$

Similarly, for $B_{n,T}$ we first note that

$$B_{n,T} \leq \sum_{i=k+k^*+1}^{n} E\left[I_{(1)}\left(\widehat{\beta_i \neq 0}\right) | \beta_i = 0\right] = \sum_{i=k+k^*+1}^{n} \Pr\left[\left|t_{\hat{\phi}_{i,(1)}}\right| > c_p(n,\delta) | \theta_i = 0\right],$$

which, by (A.107) of Lemma 10, yields

$$B_{n,T} \leq (n-k-k^*) \exp\left[\frac{-\varkappa c_p^2(n,\delta)}{2}\right] + O\left[n \exp\left(-C_0 T^{C_1}\right)\right].$$

or upon using result (ii) of Lemma 2,

$$\Pr\left(\mathcal{A}_0^c\right) \leq A_{n,T} + B_{n,T} \leq O\left(n^{1-\delta\varkappa}\right) + O\left[n \exp\left(-C_0 T^{C_1}\right)\right],$$

and hence

$$\Pr\left(\mathcal{A}_0\right) = O\left(n^{1-\delta\varkappa}\right) + O\left[\exp\left(-n^{C_2}\right)\right].$$

for some $C_2 > 0$. If, in addition, $\delta > 1$, then $\Pr(\mathcal{A}_0) \to 1$, as $n,T \to \infty$ such that $T = O\left(n^{\kappa_1}\right)$ for some $\kappa_1 > 0$, as required. The result on the residual norm of the selected model (32) continues to hold using the same arguments as in Section A.2.2 of the Appendix.

To establish (37), we recall (A.38), and noting that we do not need to condition on $\mathcal{D}_{k,T}$ and can drop terms relating to any stage of OCMT after the first, we replace (A.43), (A.44), (A.45), and (A.47) with

$$C_{n,T} = O\left(\frac{(k+k^*+1)^2}{\sqrt{T}}\right)\left\{1 + O\left(n^{1-\delta\varkappa}\right) + O\left[\exp\left(-C_0 T^{C_1}\right)\right]\right\}, \tag{A.48}$$

$$D_{n,T} = O\left(\frac{l_{\max}^2}{\sqrt{T}} + l_{\max}\right)\left\{O\left[\exp\left(-C_0 T^{C_1}\right)\right]\right\}, \tag{A.49}$$

$$E_{n,T} = O\left(\frac{l_{\max}^2}{\sqrt{T}}\right)\left\{O\left(n^{1-\delta\varkappa}\right) + O\left[\exp\left(-C_0 T^{C_1}\right)\right]\right\}. \tag{A.50}$$

and

$$F_{n,T}=\left[O\left(\frac{l_{\max}^{2}}{\sqrt{T}}\right)+O\left(l_{\max}\right)\right]\left[\begin{array}{c}O\left(\frac{n^{1-\delta\varkappa}}{l_{\max}-k-2k^{*}-1}\right)+\\ O\left[\frac{n^{2}}{l_{\max}-k-2k^{*}-1}\exp\left(-C_{0}n^{C_{1}\kappa_{1}}\right)\right]\end{array}\right], \tag{A.51}$$

respectively, where, for (A.51), we have used (A.130) in (A.46), rather than (A.132). Combining the above results, we obtain

$$E\left(\left\|\tilde{\boldsymbol{\beta}}_{n}-\boldsymbol{\beta}_{n}\right\|\right)=O\left(n^{2\epsilon-\kappa_{1}/2}\right)+O\left(n^{1+2\kappa_{2}-\kappa_{1}/2-\varkappa\delta}\right)+O\left(n^{1-\varkappa\delta}\right)+O\left[\exp\left(-n^{C_{2}}\right)\right],$$

which completes the proof.

A.3 Lemmas

Lemma 1 *Let y_t, for $t=1,2,...,T$, be given by DGP (1) and define $\mathbf{x}_i=(x_{i1},x_{i2},...,x_{iT})'$, for $i=1,2,...,k$, and $\mathbf{X}_k=(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_k)$, and suppose that Assumption 1 holds. Moreover, let $\mathbf{q}_i=(q_{i1},q_{i2},....,q_{iT})'$, for $i=1,2,...,l_T$, $\mathbf{Q}=(\mathbf{q}_1,\mathbf{q}_2,...,\mathbf{q}_{l_T})'$, and assume $\mathbf{M}_q=\mathbf{I}_T-\mathbf{Q}\left(\mathbf{Q}'\mathbf{Q}\right)^{-1}\mathbf{Q}'$ exists. Further, assume that the column vector $\boldsymbol{\tau}_T=(1,1,...,1)'$ belongs to $\mathbf{Q}$, $0\leq a<k$ column vectors in $\mathbf{X}_k$ belong to $\mathbf{Q}$, and the remaining $b=k-1>0$ columns of $\mathbf{X}_k$ that do not belong in $\mathbf{Q}$ are collected in $T\times b$ matrix $\mathbf{X}_b$. The slope coefficients that correspond to regressors in $\mathbf{X}_b$ are collected in $b\times 1$ vector $\boldsymbol{\beta}_{b,T}$. Define*

$$\boldsymbol{\theta}_{b,T}=\boldsymbol{\Omega}_{b,T}\boldsymbol{\beta}_{b,T},$$

where $\boldsymbol{\Omega}_{b,T}=E\left(T^{-1}\mathbf{X}_b'\mathbf{M}_q\mathbf{X}_b\right)$. If $\boldsymbol{\Omega}_{b,T}$ is nonsingular, and $\boldsymbol{\beta}_{k,T}=(\beta_{1,T},\beta_{2,T},...,\beta_{k,T})'\neq\mathbf{0}$, then at least one element of the $b\times 1$ vector $\boldsymbol{\theta}_{b,T}$ is nonzero.

Proof. Since $\boldsymbol{\Omega}_{b,T}$ is nonsingular and $\boldsymbol{\beta}_{b,T}\neq\mathbf{0}$, it follows that $\boldsymbol{\theta}_{b,T}\neq\mathbf{0}$; otherwise $\boldsymbol{\beta}_{b,T}=\boldsymbol{\Omega}_{b,T}^{-1}\boldsymbol{\theta}_{b,T}=\mathbf{0}$, which contradicts the assumption that $\boldsymbol{\beta}_{b,T}\neq\mathbf{0}$. ■

Lemma 2 *Consider the critical value function $c_p(n,\delta)$ defined by (8), with $0<p<1$ and $f(n,\delta)=cn^{\delta}$, for some $c,\delta>0$. Moreover, let $a>0$ and $0<b\leq 1$. Then:*

(i) $c_p(n,\delta)=O\left(\left[\delta\ln(n)\right]^{1/2}\right)$,

(ii) $n^a\exp\left[-bc_p^2(n,\delta)\right]=\ominus\left(n^{a-2b\delta}\right)$.

Proof. Results follow from Lemma 3 of supplementary Appendix A of Bailey, Pesaran, and Smith (2016). ■

Lemma 3 *Let z_t be a martingale difference sequence with respect to the filtration $\mathcal{F}_{t-1}^{z}=\sigma\left(\{z_s\}_{s=1}^{t-1}\right)$, and suppose that there exist finite positive constants C_0 and C_1, and $s>0$ such that $\sup_t\Pr\left(|z_t|>\alpha\right)\leq C_0\exp\left(-C_1\alpha^s\right)$, for all $\alpha>0$. Let $\sigma_{zt}^2=E(z_t^2\left|\mathcal{F}_{t-1}^{z}\right.)$ and*

$\sigma_z^2 = \frac{1}{T}\sum_{t=1}^{T}\sigma_{zt}^2$. *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*

$$\Pr\left(\left|\sum_{t=1}^{T} z_t\right| > \zeta_T\right) \leq \exp\left[\frac{-(1-\pi)^2 \zeta_T^2}{2T\sigma_z^2}\right]. \tag{A.52}$$

If $\lambda > (s+1)/(s+2)$, *then for some finite positive constant* C_3,

$$\Pr\left(\left|\sum_{t=1}^{T} z_t\right| > \zeta_T\right) \leq \exp\left[-C_3 \zeta_T^{s/(s+1)}\right]. \tag{A.53}$$

Proof. We proceed to prove (A.52) first and then prove (A.53). Decompose z_t as $z_t = w_t + v_t$, where $w_t = z_t I(|z_t| \leq D_T)$ and $v_t = z_t I(|z_t| > D_T)$, and note that

$$\Pr\left(\left|\sum_{t=1}^{T} [z_t - E(z_t)]\right| > \zeta_T\right) \leq \Pr\left(\left|\sum_{t=1}^{T} [w_t - E(w_t)]\right| > (1-\pi)\zeta_T\right) + \Pr\left(\left|\sum_{t=1}^{T} [v_t - E(v_t)]\right| > \pi\zeta_T\right), \tag{A.54}$$

for any $0 < \pi < 1$.[11] Further, it is easily verified that $w_t - E(w_t)$ is a martingale difference process, and since $|w_t| \leq D_T$ then by setting $b = T\sigma_z^2$ and $a = (1-\pi)\zeta_T$ in Proposition 2.1 of Freedman (1975), for the first term on the RHS of (A.54) we obtain

$$\Pr\left(\left|\sum_{t=1}^{T} [w_t - E(w_t)]\right| > (1-\pi)\zeta_T\right) \leq \exp\left[\frac{-(1-\pi)^2 \zeta_T^2}{2\left[T\sigma_z^2 + (1-\pi) D_T \zeta_T\right]}\right].$$

Consider now the second term on the RHS of (A.54) and first note that

$$\Pr\left(\left|\sum_{t=1}^{T} [v_t - E(v_t)]\right| > \pi\zeta_T\right) \leq \Pr\left[\sum_{t=1}^{T} |v_t - E(v_t)| > \pi\zeta_T\right], \tag{A.55}$$

and by Markov's inequality,

$$\Pr\left(\sum_{t=1}^{T} |[v_t - E(v_t)]| > \pi\zeta_T\right) \leq \left(\frac{1}{\pi\zeta_T}\right)\sum_{t=1}^{T} E|v_t - E(v_t)| \leq \left(\frac{2}{\pi\zeta_T}\right)\sum_{t=1}^{T} E|v_t|. \tag{A.56}$$

[11] Let $A_T = \sum_{t=1}^{T}[z_t - E(z_t)] = B_{1,T} + B_{2,T}$, where $B_{1,T} = \sum_{t=1}^{T}[w_t - E(w_t)]$ and $B_{2,T} = \sum_{t=1}^{T}[v_t - E(v_t)]$. We have $|A_T| \leq |B_{1,T}| + |B_{2,T}|$ and, therefore, $\Pr(|A_T| > \zeta_T) \leq \Pr(|B_{1,T}| + |B_{2,T}| > \zeta_T)$. Equation (A.54) now readily follows using the same steps as in the proof of (B.1).

But by Holder's inequality, for any finite $p, q \geq 1$ such that $p^{-1} + q^{-1} = 1$ we have

$$\begin{aligned} E\left|v_t\right| &= E\left(\left|z_t I\left[\left|z_t\right| > D_T\right]\right|\right) \\ &\leq \left(E\left|z_t\right|^p\right)^{1/p} \left\{E\left[\left|I\left(\left|z_t\right| > D_T\right)\right|^q\right]\right\}^{1/q} \\ &= \left(E\left|z_t\right|^p\right)^{1/p} \left\{E\left[I\left(\left|z_t\right| > D_T\right)\right]\right\}^{1/q} \\ &= \left(E\left|z_t\right|^p\right)^{1/p} \left[\Pr\left(\left|z_t\right| > D_T\right)\right]^{1/q}. \end{aligned} \tag{A.57}$$

Also, for any finite $p \geq 1$ there exists a finite positive constant C_2 such that $E\left|z_t\right|^p \leq C_2 < \infty$, by Lemma A5. Further, by assumption

$$\sup_t \Pr\left(\left|z_t\right| > D_T\right) \leq C_0 \exp\left(-C_1 D_T^s\right).$$

Using this upper bound in (A.57) together with the upper bound on $E\left|z_t\right|^p$, we have

$$\sup_t E\left|v_t\right| \leq C_2^{1/p} C_0^{1/q} \left[\exp\left(-C_1 D_T^s\right)\right]^{1/q}.$$

Therefore, using (A.55)-(A.56),

$$\Pr\left(\left|\sum_{t=1}^{T}\left[v_t - E(v_t)\right]\right| > \pi\zeta_T\right) \leq (2/\pi)\, C_2^{1/p} C_0^{1/q} \zeta_T^{-1} T \left[\exp\left(-C_1 D_T^s\right)\right]^{1/q}.$$

We need to determine D_T such that

$$(2/\pi)\, C_2^{1/p} C_0^{1/q} \zeta_T^{-1} T \left[\exp\left(-C_1 D_T^s\right)\right]^{1/q} \leq \exp\left[\frac{-\left(1-\pi\right)^2 \zeta_T^2}{2\left[T\sigma_z^2 + \left(1-\pi\right) D_T \zeta_T\right]}\right]. \tag{A.58}$$

Taking logs, we have

$$\ln\left[(2/\pi)\, C_2^{1/p} C_0^{1/q}\right] + \ln\left(\zeta_T^{-1} T\right) - \left(\frac{C_1}{q}\right) D_T^s \leq \frac{-\left(1-\pi\right)^2 \zeta_T^2}{2\left[T\sigma_z^2 + \left(1-\pi\right) D_T \zeta_T\right]},$$

or

$$C_1 q^{-1} D_T^s \geq \ln\left[(2/\pi)\, C_2^{1/p} C_0^{1/q}\right] + \ln\left(\zeta_T^{-1} T\right) + \frac{\left(1-\pi\right)^2 \zeta_T^2}{2\left[T\sigma_z^2 + \left(1-\pi\right) D_T \zeta_T\right]}.$$

Post-multiplying by $2\left[T\sigma_z^2 + \left(1-\pi\right) D_T \zeta_T\right] > 0$ we have

$$\begin{aligned} &\left(2\sigma_z^2 C_1 q^{-1}\right) T D_T^s + \left(2C_1 q^{-1}\right)\left(1-\pi\right) D_T^{s+1} \zeta_T - 2\left(1-\pi\right) D_T \zeta_T \ln\left(\zeta_T^{-1} T\right) - \\ &2\left(1-\pi\right) D_T \zeta_T \ln\left[(2/\pi)\, C_2^{1/p} C_0^{1/q}\right] \\ &\geq 2\sigma_z^2 T \ln\left[(2/\pi)\, C_2^{1/p} C_0^{1/q}\right] + 2\sigma_z^2 T \ln\left(\zeta_T^{-1} T\right) + \left(1-\pi\right)^2 \zeta_T^2. \end{aligned} \tag{A.59}$$

The above expression can now be simplified for values of $T \rightarrow \infty$, by dropping the constants and terms that are asymptotically dominated by other terms on the same side of the inequality.[12] Since $\zeta_T = \ominus\left(T^{\lambda}\right)$, for some $0 < \lambda \leq (s+1)/(s+2)$, and considering values of D_T such

[12] A term A is said to be asymptotically dominant compared to a term B if both tend to infinity and $A/B \rightarrow \infty$.

that $D_T = \ominus\left(T^{\psi}\right)$, for some $\psi > 0$, implies that the third and fourth term on the LHS of (A.59), which have the orders $\ominus\left[\ln(T)T^{\lambda+\psi}\right]$ and $\ominus\left(T^{\lambda+\psi}\right)$, respectively, are dominated by the second term on the LHS of (A.59) which is of order $\ominus\left(T^{\lambda+\psi+s\psi}\right)$. Further the first term on the RHS of (A.59) is dominated by the second term. Note that for $\zeta_T = \ominus\left(T^{\lambda}\right)$, we have $T\ln\left(\zeta_T^{-1}T\right) = \ominus\left[T\ln(T)\right]$, whilst the order of the first term on the RHS of (A.59) is $\ominus(T)$. Result (A.58) follows if we show that there exists D_T such that

$$\left(C_1q^{-1}\right)\left[2\sigma_z^2TD_T^s + 2\left(1-\pi\right)D_T^{s+1}\zeta_T\right] \geq 2\sigma_z^2T\ln\left(\zeta_T^{-1}T\right) + \left(1-\pi\right)^2\zeta_T^2. \tag{A.60}$$

Set

$$\left(C_1q^{-1}\right)D_T^{s+1} = \frac{1}{2}\left(1-\pi\right)\zeta_T, \text{ or } D_T = \left(\frac{1}{2}C_1^{-1}q\left(1-\pi\right)\zeta_T\right)^{1/(s+1)}$$

and note that (A.60) can be written as

$$2\sigma_z^2\left(C_1q^{-1}\right)T\left(\frac{1}{2}C_1^{-1}q\left(1-\pi\right)\zeta_T\right)^{s/(s+1)} + \left(1-\pi\right)^2\zeta_T^2 \geq 2\sigma_z^2T\ln\left(\zeta_T^{-1}T\right) + \left(1-\pi\right)^2\zeta_T^2.$$

Hence, the required condition is met if

$$\lim_{T\to\infty}\left[\left(C_1q^{-1}\right)\left(\frac{1}{2}C_1^{-1}q\left(1-\pi\right)\zeta_T\right)^{s/(s+1)} - \ln\left(\zeta_T^{-1}T\right)\right] \geq 0.$$

This condition is clearly satisfied noting that for values of $\zeta_T = \ominus\left(T^{\lambda}\right)$, $q > 0$, $C_1 > 0$ and $0 < \pi < 1$

$$\left(C_1q^{-1}\right)\left(\frac{1}{2}C_1^{-1}q\left(1-\pi\right)\zeta_T\right)^{s/(s+1)} - \ln\left(\zeta_T^{-1}T\right) = \ominus\left(T^{\frac{\lambda s}{1+s}}\right) - \ominus\left[\ln\left(T\right)\right],$$

since $s > 0$ and $\lambda > 0$, the first term on the RHS, which is positive, dominates the second term. Finally, we require that $D_T\zeta_T = o(T)$, since then the denominator of the fraction inside the exponential on the RHS of (A.58) is dominated by T which takes us back to the Exponential inequality with bounded random variables and proves (A.52). Consider

$$T^{-1}D_T\zeta_T = \left(\frac{1}{2}C_1^{-1}q\left(1-\pi\right)\right)^{1/(s+1)}T^{-1}\zeta_T^{\frac{2+s}{1+s}},$$

and since $\zeta_T = \ominus(T^{\lambda})$ then $D_T\zeta_T = o(T)$, as long as $\lambda < (s+1)/(s+2)$, as required.

If $\lambda > (s+1)/(s+2)$, it follows that $D_T\zeta_T$ dominates T in the denominator of the fraction inside the exponential on the RHS of (A.58). So the bound takes the form $\exp\left[\frac{-(1-\pi)\zeta_T^2}{C_4D_T\zeta_T}\right]$, for some finite positive constant C_4. Noting that $D_T = \ominus\left(\zeta_T^{1/(s+1)}\right)$, gives a bound of the form $\exp\left[-C_3\zeta_T^{s/(s+1)}\right]$ proving (A.53). ■

Remark 5 *We conclude that for all random variables that satisfy a probability exponential tail with any positive rate, removing the bound in the Exponential inequality has no effect on the relevant rate at least for the case under consideration.*

Lemma 4 *Let x_t and u_t be sequences of random variables and suppose that there exist $C_0, C_1 > 0$, and $s > 0$ such that $\sup_t \Pr(|x_t| > \alpha) \leq C_0 \exp(-C_1 \alpha^s)$ and $\sup_t \Pr(|u_t| > \alpha) \leq C_0 \exp(-C_1 \alpha^s)$, for all $\alpha > 0$. Let $\mathcal{F}_{t-1}^{(1)} = \sigma\left(\{u_s\}_{s=1}^{t-1}, \{x_s\}_{s=1}^{t-1}\right)$ and $\mathcal{F}_t^{(2)} = \sigma\left(\{u_s\}_{s=1}^{t-1}, \{x_s\}_{s=1}^{t}\right)$. Then, assume either that (i) $E\left(u_t|\mathcal{F}_t^{(2)}\right) = 0$ or (ii) $E\left(x_t u_t - \mu_t|\mathcal{F}_{t-1}^{(1)}\right) = 0$, where $\mu_t = E(x_t u_t)$. Let $\zeta_T = \ominus\left(T^\lambda\right)$, for some λ such that $0 < \lambda \leq (s/2+1)/(s/2+2)$. Then, for any π in the range $0 < \pi < 1$ we have*

$$\Pr\left(\left|\sum_{t=1}^{T}(x_t u_t - \mu_t)\right| > \zeta_T\right) \leq \exp\left[\frac{-(1-\pi)^2 \zeta_T^2}{2T\sigma_{(T)}^2}\right], \tag{A.61}$$

where $\sigma_{(T)}^2 = \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2$ and $\sigma_t^2 = E\left[(x_t u_t - \mu_t)^2 |\mathcal{F}_{t-1}^{(1)}\right]$. If $\lambda > (s/2+1)/(s/2+2)$, then for some finite positive constant C_2,

$$\Pr\left(\left|\sum_{t=1}^{T}(x_t u_t - \mu_t)\right| > \zeta_T\right) \leq \exp\left[-C_2 \zeta_T^{s/(s+2)}\right]. \tag{A.62}$$

Proof. Let $\tilde{\mathcal{F}}_{t-1} = \sigma\left(\{x_s u_s\}_{s=1}^{t-1}\right)$ and note that under (i)

$$E(x_t u_t|\tilde{\mathcal{F}}_{t-1}) = E\left[E\left(u_t|\mathcal{F}_t^{(2)}\right) x_t|\tilde{\mathcal{F}}_{t-1}\right] = 0.$$

Therefore, $x_t u_t$ is a martingale difference process. Under (ii) we simply note that $x_t u_t - \mu_t$ is a martingale difference process by assumption. Next, for any $\alpha > 0$ we have (using (B.2) with C_0 set equal to α and C_1 set equal to $\sqrt{\alpha}$)

$$\Pr[|x_t u_t| > \alpha] \leq \Pr\left[|x_t| > \alpha^{1/2}\right] + \Pr\left[|u_t|^2 > \alpha^{1/2}\right]. \tag{A.63}$$

But, under the assumptions of the lemma,

$$\sup_t \Pr\left[|x_t| > \alpha^{1/2}\right] \leq C_0 e^{-C_1 \alpha^{s/2}},$$

and

$$\sup_t \Pr\left[|u_t| > \alpha^{1/2}\right] \leq C_0 e^{-C_1 \alpha^{s/2}}.$$

Hence

$$\sup_t \Pr[|x_t u_t| > \alpha] \leq 2C_0 e^{-C_1 \alpha^{s/2}}.$$

Therefore, the process $x_t u_t$ satisfies the conditions of Lemma 3 and the results of the lemma apply. ■

Lemma 5 *Let $\boldsymbol{x} = (x_1, x_2, ..., x_T)'$ and $\boldsymbol{q}_{\cdot t} = (q_{1,t}, q_{2,t}, ..., q_{l_T,t})'$ be sequences of random variables and suppose that there exist finite positive constants C_0 and C_1, and $s > 0$ such that*

$\sup_t \Pr(|x_t| > \alpha) \leq C_0 \exp(-C_1\alpha^s)$ *and* $\sup_{i,t} \Pr(|q_{i,t}| > \alpha) \leq C_0 \exp(-C_1\alpha^s)$, *for all* $a > 0$. *Consider the linear projection*

$$x_t = \sum_{j=1}^{l_T} \beta_j q_{jt} + u_{x,t}, \tag{A.64}$$

and assume that only a finite number of slope coefficients β'*s are nonzero and bounded, and the remaining* β*'s are zero. Then, there exist finite positive constants* C_2 *and* C_3*, such that*

$$\sup_t \Pr(|u_{x,t}| > \alpha) \leq C_2 \exp(-C_3\alpha^s).$$

Proof. We assume without any loss of generality that the $|\beta_i| < C_0$ for $i = 1, 2, ..., M$, M is a finite positive integer and $\beta_i = 0$ for $i = M+1, M+2, ..., l_T$. Note that for some $0 < \pi < 1$,

$$\begin{aligned}
\sup_t \Pr(|u_{x,t}| > \alpha) &\leq \sup_t \Pr\left(\left|x_t - \sum_{j=1}^{M}\beta_j q_{jt}\right| > \alpha\right) \\
&\leq \sup_t \Pr(|x_t| > (1-\pi)\alpha) + \sup_t \Pr\left(\left|\sum_{j=1}^{M}\beta_j q_{jt}\right| > \pi\alpha\right) \\
&\leq \sup_t \Pr(|x_t| > (1-\pi)\alpha) + \sup_t \sum_{j=1}^{M} \Pr\left(|\beta_j q_{jt}| > \frac{\pi\alpha}{M}\right),
\end{aligned}$$

and since $|\beta_j| > 0$, then

$$\sup_t \Pr(|u_{x,t}| > \alpha) \leq \sup_t \Pr(|x_t| > (1-\pi)\alpha) + M \sup_{j,t} \Pr\left(|q_{jt}| > \frac{\pi\alpha}{M|\beta_j|}\right).$$

But $\sup_{j,t} \Pr\left(|q_{jt}| > \frac{\pi\alpha}{M|\beta_j|}\right) \leq \sup_{j,t} \Pr\left(|q_{jt}| > \frac{\pi\alpha}{M\beta_{\max}}\right) \leq C_0 \exp\left[-C_1\left(\frac{\pi\alpha}{M\beta_{\max}}\right)^s\right]$, and, for fixed M, the probability bound condition is clearly met. ■

Lemma 6 *Let* x_{it}, $i = 1, 2, ..., n$, $t = 1, 2, ..., T$, *and* η_t *be martingale difference processes that satisfy exponential tail probability bounds of the form (18) and (19), with tail exponents* s_x *and* s_η*, where* $s = \min(s_x, s_\eta) > 0$. *Let* $\boldsymbol{q}_{\cdot t} = (q_{1,t}, q_{2,t}, ..., q_{l_T,t})'$ *contain a constant and a subset of* $\boldsymbol{x}_t = (x_{1t}, x_{2t}, ..., x_{nt})'$. *Let* $\boldsymbol{\Sigma}_{qq} = T^{-1}\sum_{t=1}^{T} E(\boldsymbol{q}_{\cdot t}\boldsymbol{q}'_{\cdot t})$ *and* $\hat{\boldsymbol{\Sigma}}_{qq} = \mathbf{Q}'\mathbf{Q}/T$ *be both invertible, where* $\mathbf{Q} = (\boldsymbol{q}_{1\cdot}, \boldsymbol{q}_{2\cdot}, ..., \boldsymbol{q}_{l_T\cdot})$ *and* $\boldsymbol{q}_{i\cdot} = (q_{i1}, q_{i2}, ..., q_{iT})'$, *for* $i = 1, 2, ..., l_T$. *Suppose that Assumption 5 holds for all the pairs* x_{it} *and* $\boldsymbol{q}_{\cdot t}$*, and* η_t *and* $\boldsymbol{q}_{\cdot t}$*, and denote the corresponding projection residuals defined by (20) as* $u_{x_i,t} = x_{it} - \boldsymbol{\gamma}'_{qx_i,T}\boldsymbol{q}_{\cdot t}$ *and* $u_{\eta,t} = \eta_t - \boldsymbol{\gamma}'_{q\eta,T}\boldsymbol{q}_{\cdot t}$*, respectively. Let* $\hat{\boldsymbol{u}}_{x_i} = (\hat{u}_{x_i,1}, \hat{u}_{x_i,2}, ..., \hat{u}_{x_i,T})' = \mathbf{M}_q\boldsymbol{x}_i$, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iT})'$, $\hat{\boldsymbol{u}}_\eta = (\hat{u}_{\eta,1}, \hat{u}_{\eta,2}, ..., \hat{u}_{\eta,T})' = \mathbf{M}_q\boldsymbol{\eta}$, $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_T)'$, $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}$, $\mathcal{F}_t = \mathcal{F}_t^\eta \cup \mathcal{F}_t^x$, $\mu_{x_i\eta,t} = E(u_{x_i,t}u_{\eta,t} |\mathcal{F}_{t-1})$, $\omega^2_{x_i\eta,1,T} = \frac{1}{T}\sum_{t=1}^{T} E\left[(x_{it}\eta_t - E(x_{it}\eta_t |\mathcal{F}_{t-1}))^2\right]$, *and* $\omega^2_{x_i\eta,T} = \frac{1}{T}\sum_{t=1}^{T} E\left[(u_{x_i,t}u_{\eta,t} - \mu_{x_i\eta,t})^2\right]$. *Let* $\zeta_T = \ominus(T^\lambda)$. *Then, for any* π *in the range* $0 < \pi < 1$, *we have,*

$$\Pr\left(\left|\sum_{t=1}^{T} x_{it}\eta_t - E(x_{it}\eta_t |\mathcal{F}_{t-1})\right| > \zeta_T\right) \leq \exp\left[\frac{-(1-\pi)^2\zeta_T^2}{2T\omega^2_{x_i\eta,1,T}}\right], \tag{A.65}$$

if $0 < \lambda \leq (s/2+1)/(s/2+2)$*. Further, if* $\lambda > (s/2+1)/(s/2+2)$*, we have,*

$$\Pr\left(\left|\sum_{t=1}^{T} x_{it}\eta_t - E\left(x_{it}\eta_t \left|\mathcal{F}_{t-1}\right.\right)\right| > \zeta_T\right) \leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right], \tag{A.66}$$

for some finite positive constant C_0*. If it is further assumed that* $l_T = \ominus\left(T^d\right)$*, such that* $0 \leq d < 1/3$*, then, if* $3d/2 < \lambda \leq (s/2+1)/(s/2+2)$*,*

$$\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}\hat{u}_{\eta,t} - \mu_{x_i\eta,t}\right)\right| > \zeta_T\right) \leq C_0\exp\left[\frac{-\left(1-\pi\right)^2\zeta_T^2}{2T\omega_{x_i\eta,T}^2}\right] + \exp\left[-C_1T^{C_2}\right]. \tag{A.67}$$

for some finite positive constants C_0*,* C_1 *and* C_2*, and, if* $\lambda > (s/2+1)/(s/2+2)$ *we have*

$$\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}\hat{u}_{\eta,t} - \mu_{x_i\eta,t}\right)\right| > \zeta_T\right) \leq C_0\exp\left[-C_3\zeta_T^{s/(s+2)}\right] + \exp\left[-C_1T^{C_2}\right], \tag{A.68}$$

for some finite positive constants C_0*,* C_1*,* C_2 *and* C_3*.*

Proof. Note that all the results in the proofs below hold both for sequences and for triangular arrays of random variables. If $\boldsymbol{q}_{\cdot t}$ contains x_{it}, all results follow trivially, so, without loss of generality, we assume that, if this is the case, the relevant column of $\mathbf{Q}$ is removed. (A.65) and (A.66) follow immediately given our assumptions and Lemma 4. We proceed to prove the rest of the lemma. Let $\boldsymbol{u}_{x_i} = (u_{x_i,1}, u_{x_i,2}, ..., u_{x_i,T})'$ and $\boldsymbol{u}_{\eta} = (u_{\eta,1}, u_{\eta,2}, ..., u_{\eta,T})'$. We first note that

$$\begin{aligned}\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}\hat{u}_{\eta,t} - \mu_{x_i\eta,t}\right) &= \hat{\boldsymbol{u}}_{x_i}'\hat{\boldsymbol{u}}_{\eta} - \sum_{t=1}^{T}\mu_{x_i\eta,t} = \boldsymbol{u}_{x_i}'\mathbf{M}_q\boldsymbol{u}_{\eta} - \sum_{t=1}^{T}\mu_{x_i\eta,t} \\ &= \sum_{t=1}^{T}\left(u_{x_i,t}u_{\eta,t} - \mu_{x_i\eta,t}\right) - \left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\hat{\boldsymbol{\Sigma}}_{qq}^{-1}\left(\mathbf{Q}'\boldsymbol{u}_{\eta}\right),\end{aligned} \tag{A.69}$$

where $\hat{\boldsymbol{\Sigma}}_{qq} = T^{-1}\left(\mathbf{Q}'\mathbf{Q}\right)$. The second term of the above expression can now be decomposed as

$$\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\hat{\boldsymbol{\Sigma}}_{qq}^{-1}\left(\mathbf{Q}'\boldsymbol{u}_{\eta}\right) = \left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\left(\hat{\boldsymbol{\Sigma}}_{qq}^{-1} - \boldsymbol{\Sigma}_{qq}^{-1}\right)\left(\mathbf{Q}'\boldsymbol{u}_{\eta}\right) + \left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\boldsymbol{\Sigma}_{qq}^{-1}\left(\mathbf{Q}'\boldsymbol{u}_{\eta}\right). \tag{A.70}$$

By (B.1) and for any $0 < \pi_1, \pi_2, \pi_3 < 1$ such that $\sum_{i=1}^{3}\pi_i = 1$, we have

$$\begin{aligned}\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}\hat{u}_{\eta,t} - \mu_{x_i\eta,t}\right)\right| > \zeta_T\right) \leq &\Pr\left(\left|\sum_{t=1}^{T}\left(u_{x_i,t}u_{\eta,t} - \mu_{x_i\eta,t}\right)\right| > \pi_1\zeta_T\right) \\ &+ \Pr\left(\left|\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\left(\hat{\boldsymbol{\Sigma}}_{qq}^{-1} - \boldsymbol{\Sigma}_{qq}^{-1}\right)\left(\mathbf{Q}'\boldsymbol{u}_{\eta}\right)\right| > \pi_2\zeta_T\right) \\ &+ \Pr\left(\left|\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\boldsymbol{\Sigma}_{qq}^{-1}\left(\mathbf{Q}'\boldsymbol{u}_{\eta}\right)\right| > \pi_3\zeta_T\right).\end{aligned}$$

Also applying (B.2) to the last two terms of the above we obtain

$$
\begin{aligned}
&\Pr\left(\left|\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\left(\hat{\boldsymbol{\Sigma}}_{qq}^{-1}-\boldsymbol{\Sigma}_{qq}^{-1}\right)\left(\mathbf{Q}'\boldsymbol{u}_{\eta}\right)\right|>\pi_2\zeta_T\right)\\
&\leq \Pr\left(\left\|\hat{\boldsymbol{\Sigma}}_{qq}^{-1}-\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F\left\|T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right\|_F\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F>\pi_2\zeta_T\right)\\
&\leq \Pr\left(\left\|\hat{\boldsymbol{\Sigma}}_{qq}^{-1}-\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F>\frac{\zeta_T}{\delta_T}\right)+\Pr\left(T^{-1}\left\|\boldsymbol{u}_{x_i}'\mathbf{Q}\right\|_F\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F>\pi_2\delta_T\right)\\
&\leq \Pr\left(\left\|\hat{\boldsymbol{\Sigma}}_{qq}^{-1}-\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F>\frac{\zeta_T}{\delta_T}\right)+\Pr\left(\left\|\boldsymbol{u}_{x_i}'\mathbf{Q}\right\|_F>\left(\pi_2\delta_T T\right)^{1/2}\right)\\
&\quad+\Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F>\left(\pi_2\delta_T T\right)^{1/2}\right),
\end{aligned}
$$

where $\delta_T > 0$ is a deterministic sequence. In what follows, we set $\delta_T = \ominus\left(\zeta_T^{\alpha}\right)$, for some $\alpha > 0$. Similarly

$$
\begin{aligned}
&\Pr\left(\left|\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\boldsymbol{\Sigma}_{qq}^{-1}\left(\mathbf{Q}'\boldsymbol{u}_{\eta}\right)\right|>\pi_3\zeta_T\right)\\
&\leq \Pr\left(\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F\left\|T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right\|_F\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F>\pi_3\zeta_T\right)\\
&\leq \Pr\left(\left\|\boldsymbol{u}_{x_i}'\mathbf{Q}\right\|_F\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F>\frac{\pi_3\zeta_T T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F}\right)\\
&\leq \Pr\left(\left\|\boldsymbol{u}_{x_i}'\mathbf{Q}\right\|_F>\frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right)+\Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F>\frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right).
\end{aligned}
$$

Overall

$$
\begin{aligned}
&\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}_{x,t}\hat{u}_{\eta,t}-\mu_{x\eta,t}\right)\right|>\zeta_T\right)\\
&\leq \Pr\left(\left|\sum_{t=1}^{T}\left(u_{x,t}u_{\eta,t}-\mu_{x\eta,t}\right)\right|>\pi_1\zeta_T\right)+\Pr\left(\left\|\hat{\boldsymbol{\Sigma}}_{qq}^{-1}-\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F>\frac{\zeta_T}{\delta_T}\right)\\
&\quad+\Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F>\left(\pi_2\delta_T T\right)^{1/2}\right)+\Pr\left(\left\|\boldsymbol{u}_{x}'\mathbf{Q}\right\|_F>\left(\pi_2\delta_T T\right)^{1/2}\right),\\
&\quad+\Pr\left(\left\|\boldsymbol{u}_{x}'\mathbf{Q}\right\|_F>\frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right)+\Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F>\frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right). \qquad (A.71)
\end{aligned}
$$

First, since $u_{x,t}u_{\eta,t}-\mu_{x\eta,t}$ is a martingale difference process with respect to $\sigma\left(\{\eta_s\}_{s=1}^{t-1},\{x_s\}_{s=1}^{t-1},\{q_s\}_{s=1}^{t-1}\right)$, by Lemma 4, we have, for any π in the range $0<\pi<1$,

$$
\Pr\left(\left|\sum_{t=1}^{T}\left(u_{x_i,t}u_{\eta,t}-\mu_{x_i\eta,t}\right)\right|>\pi_1\zeta_T\right)\leq\exp\left[\frac{-(1-\pi)^2\zeta_T^2}{2T\omega_{x\eta,T}^2}\right], \qquad (A.72)
$$

if $0<\lambda\leq(s/2+1)/(s/2+2)$, and

$$
\Pr\left(\left|\sum_{t=1}^{T}\left(u_{x_i,t}u_{\eta,t}-\mu_{x_i\eta,t}\right)\right|>\pi_1\zeta_T\right)\leq\exp\left[-C_0\zeta_T^{s/(s+1)}\right], \qquad (A.73)
$$

if $\lambda > (s/2+1)/(s/2+2)$, for some finite positive constant C_0. We now show that the last five terms on the RHS of (A.71) are of order $\exp\left[-C_1 T^{C_2}\right]$, for some finite positive constants C_1 and C_2. We will make use of Lemma 4 since by assumption $\{q_{it}u_{\eta,t}\}$ and $\{q_{it}u_{x_i,t}\}$ are martingale difference sequences. We note that some of the bounds of the last five terms exceed, in order, $T^{1/2}$. Since we do not know the value of s, we need to consider the possibility that either (A.61) or (A.62) of Lemma 4, apply. We start with (A.61). Then, for some finite positive constant C_0, we have[13]

$$\sup_i \Pr\left(\|\mathbf{q}_i' \boldsymbol{u}_\eta\| > (\pi_2 \delta_T T)^{1/2}\right) \leq \exp\left(-C_0 \delta_T\right). \tag{A.74}$$

Also, using $\|\mathbf{Q}'\boldsymbol{u}_\eta\|_F^2 = \sum_{j=1}^{l_T}\left(\sum_{t=1}^{T} q_{jt}u_t\right)^2$ and (B.1),

$$\begin{aligned} &\Pr\left(\|\mathbf{Q}'\boldsymbol{u}_\eta\|_F > (\pi_2\delta_T T)^{1/2}\right) = \Pr\left(\|\mathbf{Q}'\boldsymbol{u}_\eta\|_F^2 > \pi_2\delta_T T\right) \\ &\leq \sum_{j=1}^{l_T} \Pr\left[\left(\sum_{t=1}^{T} q_{jt}u_{\eta,t}\right)^2 > \frac{\pi_2\delta_T T}{l_T}\right] \\ &= \sum_{j=1}^{l_T} \Pr\left[\left|\sum_{t=1}^{T} q_{jt}u_{\eta,t}\right| > \left(\frac{\pi_2\delta_T T}{l_T}\right)^{1/2}\right], \end{aligned}$$

which upon using (A.74) yields (for some finite positive constant C_0)

$$\Pr\left(\|\mathbf{Q}'\boldsymbol{u}_\eta\|_F > (\pi_2\delta_T T)^{1/2}\right) \leq l_T \exp\left(-\frac{C_0\delta_T}{l_T}\right),\ \Pr\left(\|\mathbf{Q}'\boldsymbol{u}_x\| > (\pi_2\delta_T T)^{1/2}\right) \leq l_T \exp\left(-\frac{C_0\delta_T}{l_T}\right). \tag{A.75}$$

Similarly,

$$\begin{aligned} \Pr\left(\|\mathbf{Q}'\boldsymbol{u}_\eta\|_F > \frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right) &\leq l_T \exp\left(\frac{-C_0\zeta_T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}\right), \\ \Pr\left(\|\mathbf{Q}'\boldsymbol{u}_x\| > \frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right) &\leq l_T \exp\left(\frac{-C_0\zeta_T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}\right). \end{aligned} \tag{A.76}$$

Turning to the second term of (A.71), since for all i and j, $\{q_{it}q_{jt} - E(q_{it}q_{jt})\}$ is a martingale difference process and q_{it} satisfy the required probability bound then

$$\sup_{ij} \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left[q_{it}q_{jt} - E(q_{it}q_{jt})\right]\right| > \frac{\pi_2\zeta_T}{\delta_T}\right) \leq \exp\left(\frac{-C_0 T\zeta_T^2}{\delta_T^2}\right). \tag{A.77}$$

[13] The required probability bound on u_{xt} follows from the probability bound assumptions on x_t and on q_{it}, for $i = 1, 2, ..., l_T$, even if $l_T \to \infty$. See also Lemma 5.

Therefore, by Lemma A6, for some finite positive constant C_0, we have

$$\Pr\left(\left\|\hat{\mathbf{\Sigma}}_{qq}^{-1}-\mathbf{\Sigma}_{qq}^{-1}\right\|>\frac{\zeta_T}{\delta_T}\right)\leq l_T^2\exp\left[\frac{-C_0T\zeta_T^2}{\delta_T^2 l_T^2\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F^2\left(\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F+\delta_T^{-1}\zeta_T\right)^2}\right]+ \tag{A.78}$$

$$l_T^2\exp\left(\frac{-C_0T}{\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F^2 l_T^2}\right).$$

Further by Lemma A4, $\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F=\ominus\left(l_T^{1/2}\right)$, and

$$\frac{T\zeta_T^2}{\delta_T^2 l_T^2\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F^2\left(\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F+\delta_T^{-1}\zeta_T\right)^2}=\frac{T\zeta_T^2}{\delta_T^{-2}\zeta_T^2\delta_T^2 l_T^2\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F^2\left(\delta_T\zeta_T^{-1}\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F+1\right)^2}$$

$$=\frac{T}{l_T^2\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F^2\left(\delta_T\zeta_T^{-1}\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F+1\right)^2}$$

Consider now the different terms in the above expression and let

$$P_{11}=\frac{\delta_T}{l_T},\ P_{12}=\frac{\zeta_T}{\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F l_T},$$

$$P_{13}=\frac{T}{l_T^2\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F^2\left[\delta_T\zeta_T^{-1}\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F+1\right]^2},\text{ and }P_{14}=\frac{T}{\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F^2 l_T^2}.$$

Under $\delta_T=\ominus\left(\zeta_T^{\alpha}\right)$, $l_T=\ominus(T^d)$, and $\zeta_T=\ominus(T^{\lambda})$, we have

$$P_{11}=\frac{\delta_T}{l_T}=\ominus\left(T^{\alpha-d}\right), \tag{A.79}$$

$$P_{12}=\frac{\zeta_T}{\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F l_T}=\ominus\left(T^{\lambda-3d/2}\right), \tag{A.80}$$

$$P_{13}=\frac{T}{l_T^2\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F^2\left[\delta_T\zeta_T^{-1}\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F+1\right]^2}=\ominus\left(T^{\max\{1-3d-(2\alpha-2\lambda+d),1-3d-(\alpha-\lambda+d/2),1-3d\}}\right)$$

$$=\ominus\left(T^{\max\{1+2\lambda-4d-2\alpha,1+\lambda-7d/2-\alpha,1-3d\}}\right), \tag{A.81}$$

and

$$P_{14}=\frac{T}{\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F^2 l_T^2}=\ominus\left(T^{1-3d}\right). \tag{A.82}$$

Suppose that $d<1/3$, and by (A.80) note that $\lambda\geq 3d/2$. Then, setting $\alpha=1/3$, ensures that all the above four terms tend to infinity polynomially with T. Therefore, it also follows that they can be represented as terms of order $\exp\left[-C_1T^{C_2}\right]$, for some finite positive constants

C_1 and C_2, and (A.67) follows. The same analysis can be repeated under (A.62). In this case, (A.75), (A.76), (A.77) and (A.78) are replaced by

$$\Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F > (\pi_2\delta_T T)^{1/2}\right) \le l_T \exp\left(-\frac{C_0\delta_T^{s/2(s+2)}T^{s/2(s+2)}}{l_T^{s/2(s+2)}}\right) = l_T \exp\left[-C_0\left(\frac{\delta_T T}{l_T}\right)^{s/2(s+2)}\right],$$

$$\Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{x}\right\| > (\pi_2\delta_T T)^{1/2}\right) \le l_T \exp\left(-\frac{C_0\delta_T^{s/2(s+2)}T^{s/2(s+2)}}{l_T^{s/2(s+2)}}\right) = l_T \exp\left[-C_0\left(\frac{\delta_T T}{l_T}\right)^{s/2(s+2)}\right],$$

$$\Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{\eta}\right\|_F > \frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right) \le l_T \exp\left(\frac{-C_0\zeta_T^{s/2(s+2)}T^{s/2(s+2)}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{s/2(s+2)} l_T^{s/2(s+2)}}\right) = l_T \exp\left[-C_0\left(\frac{\zeta_T T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}\right)^{s/2(s+2)}\right],$$

$$\Pr\left(\left\|\mathbf{Q}'\boldsymbol{u}_{x}\right\| > \frac{\pi_3^{1/2}\zeta_T^{1/2}T^{1/2}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{1/2}}\right) \le l_T \exp\left(\frac{-C_0\zeta_T^{s/2(s+2)}T^{s/2(s+2)}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{s/2(s+2)} l_T^{s/2(s+2)}}\right) = l_T \exp\left[-C_0\left(\frac{\zeta_T T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}\right)^{s/2(s+2)}\right],$$

$$\sup_{ij}\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left[q_{it}q_{jt} - E(q_{it}q_{jt})\right]\right| > \frac{\pi_2\zeta_T}{\delta_T}\right) \le \exp\left[\frac{-C_0 T^{s/(s+2)}\zeta_T^{s/(s+2)}}{\delta_T^{s/(s+2)}}\right],$$

and, using Lemma A7,

$$\Pr\left(\left\|\left(\hat{\boldsymbol{\Sigma}}_{qq}^{-1} - \boldsymbol{\Sigma}_{qq}^{-1}\right)\right\| > \frac{\pi_2\zeta_T}{\delta_T}\right) \le l_T^2 \exp\left[\frac{-C_0 T^{s/(s+2)}\zeta_T^{s/(s+2)}}{\delta_T^{s/(s+2)} l_T^{s/(s+2)} \left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{s/(s+2)}\left(\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F + \delta_T^{-1}\zeta_T\right)^{s/(s+2)}}\right] +$$

$$l_T^2 \exp\left[\frac{-C_0 T^{s/(s+2)}}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F^{s/(s+2)} l_T^{s/(s+2)}}\right] =$$

$$l_T^2 \exp\left[-C_0\left(\frac{T\zeta_T}{\delta_T l_T\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F\left(\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F + \delta_T^{-1}\zeta_T\right)}\right)^{s/(s+2)}\right] +$$

$$l_T^2 \exp\left[-C_0\left(\frac{T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}\right)^{s/(s+2)}\right].$$

respectively. Once again, we need to derive conditions that imply that $P_{21} = \frac{\delta_T T}{l_T}$, $P_{22} = \frac{\zeta_T T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}$, $P_{23} = \frac{T\zeta_T}{\delta_T l_T\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F\left(\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F + \delta_T^{-1}\zeta_T\right)}$ and $P_{24} = \frac{T}{\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F l_T}$ are terms that tend to infinity polynomially with T. If that is the case then, as before, the relevant terms are of order $\exp\left[-C_1 T^{C_2}\right]$, for some finite positive constants C_1 and C_2, and (A.68) follows, completing the proof of the lemma. P_{22} dominates P_{23} so we focus on P_{21}, P_{23} and P_{24}. We have

$$\frac{\delta_T T}{l_T} = \ominus\left(T^{1+\alpha-d/2}\right),$$

$$\frac{T\zeta_T}{\delta_T l_T\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F\left(\left\|\boldsymbol{\Sigma}_{qq}^{-1}\right\|_F + \delta_T^{-1}\zeta_T\right)} = \ominus\left[T^{\max(1+\lambda-\alpha-2d,1-3d/2)}\right],$$

and

$$\frac{T}{\left\|\mathbf{\Sigma}_{qq}^{-1}\right\|_F l_T} = \ominus\left(T^{1-3d/2}\right)$$

It immediately follows that under the conditions set when using (A.61), which were that $\alpha = 1/3$, $d < 1/3$ and $\lambda > 3d/2$, and as long as $s > 0$, P_{21} to P_{24} tend to infinity polynomially with T, proving the lemma.[14] ■

Lemma 7 *Let x_{it}, $i = 1, 2, ..., n$, be martingale difference processes that satisfy exponential tail probability bounds of the form (18), with positive tail exponent s. Let $\boldsymbol{q}_{\cdot t} = (q_{1,t}, q_{2,t}, ..., q_{l_T,t})'$ contain a constant and a subset of $\boldsymbol{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$. Suppose that Assumption 5 holds for all the pairs x_{it} and $\boldsymbol{q}_{\cdot t}$, and denote the corresponding projection residuals defined by (20) as $u_{x_it} = x_{it} - \boldsymbol{\gamma}'_{qx_i,T}\boldsymbol{q}_{\cdot t}$. Let $\mathbf{\Sigma}_{qq} = T^{-1}\sum_{t=1}^{T} E\left(\boldsymbol{q}_{\cdot t}\boldsymbol{q}'_{\cdot t}\right)$ and $\hat{\mathbf{\Sigma}}_{qq} = \mathbf{Q}'\mathbf{Q}/T$ be both invertible, where $\mathbf{Q} = (\boldsymbol{q}_{1\cdot}, \boldsymbol{q}_{2\cdot}, ..., \boldsymbol{q}_{l_T\cdot})$ and $\boldsymbol{q}_{i\cdot} = (q_{i1}, q_{i2}, .., q_{iT})'$, for $i = 1, 2, ..., l_T$. Let $\hat{\mathbf{u}}_{x_i} = (\hat{u}_{x_i,1}, \hat{u}_{x_i,2}, ..., \hat{u}_{x_i,T})' = \mathbf{M}_q\mathbf{x}_{i\cdot}$, where $\mathbf{x}_{i\cdot} = (x_{i1}, x_{i2}, ..., x_{iT})'$ and $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}\left(\mathbf{Q}'\mathbf{Q}\right)^{-1}\mathbf{Q}$. Moreover, suppose that $E\left(u_{x_i,t}^2 - \sigma_{x_it}^2|\mathcal{F}_{t-1}\right) = 0$, where $\mathcal{F}_t = \mathcal{F}_t^x$ and $\sigma_{x_it}^2 = E(u_{x_i,t}^2)$. Let $\zeta_T = \ominus(T^\lambda)$. Then, if $0 < \lambda \leq (s/2+1)/(s/2+2)$, for any π in the range $0 < \pi < 1$, and some finite positive constant C_0, we have,*

$$\Pr\left[\left|\sum_{t=1}^{T}\left(x_{it}^2 - \sigma_{x_it}^2\right)\right| > \zeta_T\right] \leq C_0 \exp\left[\frac{-\left(1-\pi\right)^2\zeta_T^2}{2T\omega_{i,1,T}^2}\right]. \tag{A.83}$$

Otherwise, if $\lambda > (s/2+1)/(s/2+2)$, for some finite positive constant C_0, we have

$$\Pr\left[\left|\sum_{t=1}^{T}\left(x_{it}^2 - \sigma_{x_it}^2\right)\right| > \zeta_T\right] \leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right]. \tag{A.84}$$

If it is further assumed that $l_T = \ominus\left(T^d\right)$, such that $0 \leq d < 1/3$, then, if $3d/2 < \lambda \leq (s/2+1)/(s/2+2)$,

$$\Pr\left[\left|\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}^2 - \sigma_{x_it}^2\right)\right| > \zeta_T\right] \leq C_0 \exp\left[\frac{-\left(1-\pi\right)^2\zeta_T^2}{2T\omega_{i,T}^2}\right] + \exp\left[-C_1T^{C_2}\right], \tag{A.85}$$

for some finite positive constants C_0, C_1 and C_2, and, if $\lambda > (s/2+1)/(s/2+2)$,

$$\Pr\left[\left|\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}^2 - \sigma_{x_it}^2\right)\right| > \zeta_T\right] \leq C_0 \exp\left[-C_3\zeta_T^{s/(s+2)}\right] + \exp\left[-C_1T^{C_2}\right], \tag{A.86}$$

[14] It is important to highlight one particular feature of the above proof. In (A.75), $q_{it}u_{x,t}$ needs to be a martingale difference process. Note that if q_{it} is a martingale difference process distributed independently of $u_{x,t}$, then $q_{it}u_{x,t}$ is also a martingale difference process irrespective of the nature of $u_{x,t}$. This implies that one may not need to impose a martingale difference assumption on $u_{x,t}$ if x_{it} is a noise variable. Unfortunately, a leading case for which this lemma is used is one where $q_{it} = 1$. It is then clear that one needs to impose a martingale difference assumption on $u_{x,t}$, to deal with covariates that cannot be represented as martingale difference processes. We relax this assumption in Section 4, where we allow noise variables to be mixing processes.

for some finite positive constants C_0, C_1, C_2 and C_3, where $\omega_{i,1,T}^2 = \frac{1}{T}\sum_{t=1}^T E\left[\left(x_{it}^2 - \sigma_{x_it}^2\right)^2\right]$ *and* $\omega_{i,T}^2 = \frac{1}{T}\sum_{t=1}^T E\left[\left(u_{x_i,t}^2 - \sigma_{x_it}^2\right)^2\right]$.

Proof. If $\boldsymbol{q}_{\cdot t}$ contains x_{it}, all results follow trivially, so, without loss of generality, we assume that, if this is the case, the relevant column of $\mathbf{Q}$ is removed. (A.83) and (A.84) follow similarly to (A.65) and (A.66). For (A.85) and (A.86), we first note that

$$\left|\sum_{t=1}^{T}\left(\hat{u}_{x_i,t}^2 - \sigma_{x_it}^2\right)\right| \leq \left|\sum_{t=1}^{T}\left(u_{x_i,t}^2 - \sigma_{x_it}^2\right)\right| + \left|\left(T^{-1}\boldsymbol{u}_{x_i}'\mathbf{Q}\right)\left(T^{-1}\mathbf{Q}'\mathbf{Q}\right)^{-1}\left(\mathbf{Q}'\boldsymbol{u}_{x_i}\right)\right|.$$

Since $\left\{u_{x_i,t}^2 - \sigma_{x_it}^2\right\}$ is a martingale difference process and for $\alpha > 0$ and $s > 0$

$$\sup_t \Pr\left(\left|u_{x_i,t}^2\right| > \alpha^2\right) = \sup_t \Pr\left(\left|u_{x_i,t}\right| > \alpha\right) \leq C_0\exp\left(-C_1\alpha^s\right),$$

by Lemma 5, then the conditions of Lemma 3 are met and we have

$$\Pr\left[\left|\sum_{t=1}^{T}\left(u_{x_i,t}^2 - \sigma_{x_it}^2\right)\right| > \zeta_T\right] \leq \exp\left[\frac{-\left(1-\pi\right)^2\zeta_T^2}{2T\omega_{i,T}^2}\right]. \tag{A.87}$$

if $0 < \lambda \leq (s/2+1)/(s/2+2)$ and

$$\Pr\left[\left|\sum_{t=1}^{T}\left(u_{x_i,t}^2 - \sigma_{x_it}^2\right)\right| > \zeta_T\right] \leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right],$$

if $\lambda > (s/2+1)/(s/2+2)$. Then, using the same line of reasoning as in the proof of Lemma 6 we establish the desired result. ■

Lemma 8 *Let y_t, for $t = 1, 2, ..., T$, be given by the data generating process (1) and suppose that u_t and $\boldsymbol{x}_t = (x_{1t}, x_{2t}, ..., x_{nt})'$ satisfy Assumptions 2-3, with $s = \min(s_x, s_u) > 0$. Let $\boldsymbol{q}_{\cdot t} = (q_{1,t}, q_{2,t}, ..., q_{l_T,t})'$ contain a constant and a subset of $\boldsymbol{x}_t = (x_{1t}, x_{2t}, ..., x_{nt})'$. Assume that $\boldsymbol{\Sigma}_{qq} = \frac{1}{T}\sum_{t=1}^T E\left(\boldsymbol{q}_{\cdot t}\boldsymbol{q}_{\cdot t}'\right)$ and $\hat{\boldsymbol{\Sigma}}_{qq} = \mathbf{Q}'\mathbf{Q}/T$ are both invertible, where $\mathbf{Q} = (\boldsymbol{q}_{1\cdot}, \boldsymbol{q}_{2\cdot}, ..., \boldsymbol{q}_{l_T\cdot})$ and $\boldsymbol{q}_{i\cdot} = (q_{i1}, q_{i2}, ..., q_{iT})'$, for $i = 1, 2, ..., l_T$. Moreover, suppose that Assumption 5 holds for all the pairs x_t and $\boldsymbol{q}_{\cdot t}$, and y_t and $(\boldsymbol{q}_{\cdot t}', x_t)'$, where x_t is a generic element of $\{x_{1t}, x_{2t}, ..., x_{nt}\}$ that does not belong to $\boldsymbol{q}_{\cdot t}$, and denote the corresponding projection residuals defined by (20) as $u_{x,t} = x_t - \boldsymbol{\gamma}_{qx,T}'\boldsymbol{q}_{\cdot t}$ and $e_t = y_t - \boldsymbol{\gamma}_{yqx,T}'(\boldsymbol{q}_{\cdot t}', x_t)'$. Define $\boldsymbol{x} = (x_1, x_2, ..., x_T)'$, and $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$, and let $a_T = \ominus\left(T^{\lambda-1}\right)$. Then, for any π in the range $0 < \pi < 1$, and as long as $l_T = \ominus\left(T^d\right)$, such that $0 \leq d < 1/3$, we have, that, if $3d/2 < \lambda \leq (s/2+1)/(s/2+2)$,*

$$\Pr\left(\left|\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma_{x,(T)}^2} - 1\right| > a_T\right) \leq \exp\left[\frac{-\sigma_{x,(T)}^4\left(1-\pi\right)^2 T a_T^2}{2\omega_{x,(T)}^2}\right] + \exp\left[-C_0T^{C_1}\right], \tag{A.88}$$

and

$$\Pr\left[\left|\left(\frac{\sigma^2_{x,(T)}}{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}\right)^{1/2}-1\right|>a_T\right]\leq\exp\left[\frac{-\sigma^4_{x,(T)}\left(1-\pi\right)^2Ta_T^2}{2\omega^2_{x,(T)}}\right]+\exp\left[-C_0T^{C_1}\right],\quad (A.89)$$

where

$$\sigma^2_{x,(T)}=\frac{1}{T}\sum_{t=1}^{T}E\left(u^2_{x,t}\right),\ \omega^2_{x,(T)}=\frac{1}{T}\sum_{t=1}^{T}E\left[\left(u^2_{x,t}-\sigma^2_{xt}\right)^2\right].\quad (A.90)$$

If $\lambda>(s/2+1)/(s/2+2)$,

$$\Pr\left(\left|\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}-1\right|>a_T\right)\leq\exp\left[-C_0\left(Ta_T\right)^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right],\quad (A.91)$$

and

$$\Pr\left[\left|\left(\frac{\sigma^2_{x,(T)}}{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}\right)^{1/2}-1\right|>a_T\right]\leq\exp\left[-C_0\left(Ta_T\right)^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right].\quad (A.92)$$

Also, if $3d/2<\lambda\leq(s/2+1)/(s/2+2)$,

$$\Pr\left(\left|\frac{T^{-1}\boldsymbol{e}'\boldsymbol{e}}{\sigma^2_{u,(T)}}-1\right|>a_T\right)\leq\exp\left[\frac{-\sigma^4_{u,(T)}\left(1-\pi\right)^2Ta_T^2}{2\omega^2_{u,(T)}}\right]+\exp\left[-C_0T^{C_1}\right],\quad (A.93)$$

and

$$\Pr\left[\left|\left(\frac{\sigma^2_{u,(T)}}{\mathbf{e}'\mathbf{e}/T}\right)^{1/2}-1\right|>a_T\right]\leq\exp\left[\frac{-\sigma^4_{u,(T)}\left(1-\pi\right)^2Ta_T^2}{2\omega^2_{u,T}}\right]+\exp\left[-C_0T^{C_1}\right],\quad (A.94)$$

where $\mathbf{e}=(e_1,e_2,...,e_T)'$

$$\sigma^2_{u,(T)}=\frac{1}{T}\sum_{t=1}^{T}\sigma^2_t,\ \textit{and}\ \omega^2_{u,T}=\frac{1}{T}\sum_{t=1}^{T}E\left[\left(u^2_t-\sigma^2_t\right)^2\right].\quad (A.95)$$

If $\lambda>(s/2+1)/(s/2+2)$,

$$\Pr\left(\left|\frac{T^{-1}\boldsymbol{e}'\boldsymbol{e}}{\sigma^2_{u,(T)}}-1\right|>a_T\right)\leq\exp\left[-C_0\left(Ta_T\right)^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right],\quad (A.96)$$

and

$$\Pr\left[\left|\left(\frac{\sigma^2_{u,(T)}}{\mathbf{e}'\mathbf{e}/T}\right)^{1/2}-1\right|>a_T\right]\leq\exp\left[-C_0\left(Ta_T\right)^{s/(s+2)}\right]+\exp\left[-C_1T^{C_2}\right],\quad (A.97)$$

Proof. First note that

$$\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T} - \sigma^2_{x,(T)} = T^{-1}\sum_{t=1}^{T}\left(\hat{u}^2_{x,t} - \sigma^2_{xt}\right),$$

where $\hat{u}_{x,t}$, for $t = 1, 2, ..., T,$. is the t-th element of $\hat{\mathbf{u}}_x = \mathbf{M}_q\mathbf{x}$. Now applying Lemma 7 to $\sum_{t=1}^{T}\left(\hat{u}^2_{x,t} - \sigma^2_{xt}\right)$ with $\zeta_T = Ta_T$ we have

$$\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}^2_{x,t} - \sigma^2_{xt}\right)\right| > \zeta_T\right) \le \exp\left[\frac{-(1-\pi)^2\zeta_T^2}{2\omega^2_{x,(T)}T}\right] + \exp\left[-C_0 T^{C_1}\right],$$

if $3d/2 < \lambda \le (s/2+1)/(s/2+2)$, and

$$\Pr\left(\left|\sum_{t=1}^{T}\left(\hat{u}^2_{x,t} - \sigma^2_{xt}\right)\right| > \zeta_T\right) \le \exp\left[-C_0\zeta_T^{s/(s+2)}\right] + \exp\left[-C_1 T^{C_2}\right],$$

if $\lambda > (s/2+1)/(s/2+2)$, where $\omega^2_{x,(T)}$ is defined by (A.90). Also

$$\Pr\left[\left|\frac{T^{-1}\sum_{t=1}^{T}\left(\hat{u}^2_{x,t} - \sigma^2_{xt}\right)}{\sigma^2_{x,(T)}}\right| > \frac{\zeta_T}{T\sigma^2_{x,(T)}}\right] \le \exp\left[\frac{-(1-\pi)^2\zeta_T^2}{2\omega^2_{x,(T)}T}\right] + \exp\left[-C_0 T^{C_1}\right],$$

if $3d/2 < \lambda \le (s/2+1)/(s/2+2)$, and

$$\Pr\left[\left|\frac{T^{-1}\sum_{t=1}^{T}\left(\hat{u}^2_{x,t} - \sigma^2_{xt}\right)}{\sigma^2_{x,(T)}}\right| > \frac{\zeta_T}{T\sigma^2_{x,(T)}}\right] \le \exp\left[-C_0\zeta_T^{s/(s+2)}\right] + \exp\left[-C_1 T^{C_2}\right],$$

if $\lambda > (s/2+1)/(s/2+2)$. Therefore, setting $a_T = \zeta_T/T\sigma^2_{x,(T)} = \ominus\left(T^{\lambda-1}\right)$, we have

$$\Pr\left(\left|\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}} - 1\right| > a_T\right) \le \exp\left[\frac{-\sigma^4_{x,(T)}(1-\pi)^2 Ta_T^2}{2\omega^2_{x,(T)}}\right] + \exp\left[-C_0 T^{C_1}\right], \tag{A.98}$$

if $3d/2 < \lambda \le (s/2+1)/(s/2+2)$, and

$$\Pr\left(\left|\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}} - 1\right| > a_T\right) \le \exp\left[-C_0\zeta_T^{s/(s+2)}\right] + \exp\left[-C_1 T^{C_2}\right],$$

if $\lambda > (s/2+1)/(s/2+2)$, as required. Now setting $\omega_T = \frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}}$, and using Lemma A3, we have

$$\Pr\left(\left|\frac{1}{\sqrt{\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}}}} - 1\right| > a_T\right) \le \Pr\left(\left|\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}} - 1\right| > a_T\right),$$

and hence

$$\Pr\left[\left|\left(\frac{\sigma^2_{u,(T)}}{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}\right)^{1/2} - 1\right| > a_T\right] \le \exp\left[\frac{-\sigma^4_{x,(T)}(1-\pi)^2 Ta_T^2}{\omega^2_{x,(T)}}\right] + \exp\left[-C_0 T^{C_1}\right], \tag{A.99}$$

if $3d/2 < \lambda \leq (s/2+1)/(s/2+2)$, and

$$\Pr\left[\left|\left(\frac{\sigma^2_{u,(T)}}{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}\right)^{1/2} - 1\right| > a_T\right] \leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right] + \exp\left[-C_1 T^{C_2}\right],$$

if $\lambda > (s/2+1)/(s/2+2)$. Furthermore

$$\Pr\left(\left|\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2} - 1\right| > a_T\right) = \Pr\left[\frac{\left|\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right) - 1\right|}{\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2} + 1} > a_T\right],$$

and using Lemma A1 for some finite positive constant C, we have

$$\begin{aligned}\Pr\left[\left|\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2} - 1\right| > a_T\right] &\leq \Pr\left[\left|\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}}\right) - 1\right| > \frac{a_T}{C}\right] + \Pr\left[\frac{1}{\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}}\right)^{1/2} + 1} > C\right] \\ &= \Pr\left[\left|\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}}\right) - 1\right| > \frac{a_T}{C}\right] + \Pr\left[\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T\sigma^2_{x,(T)}}\right)^{1/2} + 1 < C^{-1}\right].\end{aligned}$$

Let $C = 1$, and note that for this choice of C

$$\Pr\left[\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2} + 1 < C^{-1}\right] = \Pr\left[\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2} < 0\right] = 0.$$

Hence

$$\Pr\left[\left|\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2} - 1\right| > a_T\right] \leq \Pr\left[\left|\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right) - 1\right| > a_T\right],$$

and using (A.98),

$$\Pr\left[\left|\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2} - 1\right| > a_T\right] \leq \exp\left[\frac{-\sigma^4_{x,(T)}(1-\pi)^2 T a_T^2}{2\omega^2_{x,(T)}}\right] + \exp\left[-C_0 T^{C_1}\right], \quad \text{(A.100)}$$

if $3d/2 < \lambda \leq (s/2+1)/(s/2+2)$, and

$$\Pr\left[\left|\left(\frac{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}{\sigma^2_{x,(T)}}\right)^{1/2} - 1\right| > a_T\right] \leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right] + \exp\left[-C_1 T^{C_2}\right],$$

if $\lambda > (s/2+1)/(s/2+2)$. Consider now $\mathbf{e}'\mathbf{e} = \sum_{t=1}^T e_t^2$ and note that

$$\left|\sum_{t=1}^T \left(e_t^2 - \sigma_t^2\right)\right| \leq \left|\sum_{t=1}^T \left(u_t^2 - \sigma_t^2\right)\right| + \left|\left(T^{-1}\boldsymbol{u}'\mathbf{W}\right)\left(T^{-1}\mathbf{W}'\mathbf{W}\right)^{-1}\left(\mathbf{W}'\boldsymbol{u}\right)\right|,$$

where $\mathbf{W} = (\mathbf{Q}, \mathbf{x})$. As before, applying Lemma 7 to $\sum_{t=1}^{T}(e_t^2 - \sigma_t^2)$, and following similar lines of reasoning we have

$$\Pr\left[\left|\sum_{t=1}^{T}\left(e_t^2 - \sigma_t^2\right)\right| > \zeta_T\right] \leq \exp\left[\frac{-(1-\pi)^2\zeta_T^2}{2\omega_{u,(T)}^2 T}\right] + \exp\left[-C_0 T^{C_1}\right],$$

if $3d/2 < \lambda \leq (s/2+1)/(s/2+2)$, and

$$\Pr\left[\left|\sum_{t=1}^{T}\left(e_t^2 - \sigma_t^2\right)\right| > \zeta_T\right] \leq \exp\left[-C_0\zeta_T^{s/(s+2)}\right] + \exp\left[-C_1 T^{C_2}\right],$$

if $\lambda > (s/2+1)/(s/2+2)$, which yield (A.93) and (A.96). Result (A.94) also follows along similar lines as used above to prove (A.89). ■

Lemma 9 *Let y_t, for $t = 1, 2, ..., T$, be given by the data generating process (1) and suppose that u_t and $\boldsymbol{x}_t = (x_{1t}, x_{2t}, ..., x_{nt})'$ satisfy Assumptions 2-3. Let $\boldsymbol{q}_{\cdot t} = (q_{1,t}, q_{2,t}, ..., q_{l_T,t})'$ contain a constant and a subset of $\boldsymbol{x}_t = (x_{1t}, x_{2t}, ..., x_{nt})'$, and $l_T = o(T^{1/3})$. Assume that $\boldsymbol{\Sigma}_{qq} = \frac{1}{T}\sum_{t=1}^{T} E(\boldsymbol{q}_{\cdot t}\boldsymbol{q}'_{\cdot t})$ and $\hat{\boldsymbol{\Sigma}}_{qq} = \mathbf{Q}'\mathbf{Q}/T$ are both invertible, where $\mathbf{Q} = (\boldsymbol{q}_{1\cdot}, \boldsymbol{q}_{2\cdot}, ..., \boldsymbol{q}_{l_T\cdot})$ and $\boldsymbol{q}_{i\cdot} = (q_{i1}, q_{i2}, .., q_{iT})'$, for $i = 1, 2, ..., l_T$. Suppose that Assumption 5 holds for the pair y_t and $(\boldsymbol{q}'_{\cdot t}, x_t)'$, where x_t is a generic element of $\{x_{1t}, x_{2t}, ..., x_{nt}\}$ that does not belong to $\boldsymbol{q}_{\cdot t}$, and denote the corresponding projection residuals defined by (20) as $e_t = y_t - \boldsymbol{\gamma}'_{yqx,T}(\boldsymbol{q}'_{\cdot t}, x_t)'$. Define $\boldsymbol{x} = (x_1, x_2, ..., x_T)'$, $\mathbf{e} = (e_1, e_2, ..., e_T)'$, and $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$. Moreover, let $E(\mathbf{e}'\mathbf{e}/T) = \sigma_{e,(T)}^2$ and $E(\mathbf{x}'\mathbf{M}_q\mathbf{x}/T) = \sigma_{x,(T)}^2$. Then*

$$\Pr\left[\left|\frac{a_T}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}\right| > c_p(n,\delta)\right] \leq \Pr\left(\left|\frac{a_T}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{c_p(n,\delta)}{1+d_T}\right) + \exp\left[-C_0 T^{C_1}\right] \quad \text{(A.101)}$$

for any random variable a_T, some finite positive constants C_0 and C_1, and some bounded sequence $d_T > 0$, where $c_p(n,\delta)$ is defined in (8). Similarly,

$$\Pr\left[\left|\frac{a_T}{\sqrt{(\mathbf{e}'\mathbf{e}/T)}}\right| > c_p(n,\delta)\right] \leq \Pr\left(\left|\frac{a_T}{\sigma_{e,(T)}}\right| > \frac{c_p(n,\delta)}{1+d_T}\right) + \exp\left[-C_0 T^{C_1}\right]. \quad \text{(A.102)}$$

Proof. We prove (A.101). (A.102) follows similarly. Define

$$g_T = \left(\frac{\sigma_{e,(T)}^2}{T^{-1}\mathbf{e}'\mathbf{e}}\right)^{1/2} - 1, \; h_T = \left(\frac{\sigma_{x,(T)}^2}{T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}}\right)^{1/2} - 1.$$

Using results in Lemma A1, note that for any $d_T > 0$ bounded in T,

$$\Pr\left[\left|\frac{a_T}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}\right| > c_p(n,\delta)\,|\theta = 0\right] \leq \Pr\left(\left|\frac{a_T}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{c_p(n,\delta)}{1+d_T}\right) + \Pr\left(\left|(1+g_T)(1+h_T)\right| > 1+d_T\right).$$

Since $(1+g_T)(1+h_T) > 0$, then

$$\begin{aligned}\Pr\left(\left|(1+g_T)(1+h_T)\right| > 1+d_T\right) &= \Pr\left[(1+g_T)(1+h_T) > 1+d_T\right] \\ &= \Pr\left(g_T h_T + g_T + h_T\right) > d_T\right).\end{aligned}$$

Using (A.89), (A.92), (A.94) and (A.97),

$$\begin{aligned}&\Pr\left[|h_T| > d_T\right] \leq \exp\left[-C_0 T^{C_1}\right],\ \ \Pr\left[|h_T| > c\right] \leq \exp\left[-C_0 T^{C_1}\right], \\ &\Pr\left[|g_T| > d_T\right] \leq \exp\left[-C_0 T^{C_1}\right],\ \ \Pr\left[|g_T| > d_T/c\right] \leq \exp\left[-C_0 T^{C_1}\right],\end{aligned}$$

for some finite positive constants C_0 and C_1. Using the above results, for some finite positive constants C_0 and C_1, we have,

$$\Pr\left[\left|\frac{a_T}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}\right| > c_p(n,\delta)\,|\theta = 0\right] \leq \Pr\left(\left|\frac{a_T}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{c_p(n,\delta)}{1+d_T}\right) + \exp\left[-C_0 T^{C_1}\right],$$

which establishes the desired the result. ■

Lemma 10 *Let y_t, for $t = 1, 2, ..., T$, be given by the data generating process (1) and suppose that u_t and $\boldsymbol{x}_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})'$ satisfy Assumptions 2-3, with $s = \min(s_x, s_u) > 0$. Let $\boldsymbol{q}_{\cdot t} = (q_{1,t}, q_{2,t}, ..., q_{l_T,t})'$ contain a constant and a subset of $\boldsymbol{x}_{nt}$, and let $\eta_t = \boldsymbol{x}'_{b,t}\boldsymbol{\beta}_b + u_t$, where $\boldsymbol{x}_{b,t}$ is $k_b \times 1$ dimensional vector of signal variables that do not belong to $\boldsymbol{q}_{\cdot t}$, with the associated coefficients, $\boldsymbol{\beta}_b$. Assume that $\boldsymbol{\Sigma}_{qq} = \frac{1}{T}\sum_{t=1}^{T} E\left(\boldsymbol{q}_{\cdot t}\boldsymbol{q}'_{\cdot t}\right)$ and $\hat{\boldsymbol{\Sigma}}_{qq} = \mathbf{Q}'\mathbf{Q}/T$ are both invertible, where $\mathbf{Q} = (\boldsymbol{q}_{1\cdot}, \boldsymbol{q}_{2\cdot}, ..., \boldsymbol{q}_{l_T\cdot})$ and $\boldsymbol{q}_{i\cdot} = (q_{i1}, q_{i2}, ..., q_{iT})'$, for $i = 1, 2, ..., l_T$. Moreover, let $l_T = o(T^{1/3})$ and suppose that Assumption 5 holds for all the pairs x_{it} and $\boldsymbol{q}_{\cdot t}$, and y_t and $(\boldsymbol{q}'_{\cdot t}, x_t)'$, where x_t is a generic element of $\{x_{1t}, x_{2t}, ..., x_{nt}\}$ that does not belong to $\boldsymbol{q}_{\cdot t}$, and denote the corresponding projection residuals defined by (20) as $u_{x,t} = x_t - \boldsymbol{\gamma}'_{qx,T}\boldsymbol{q}_{\cdot t}$ and $e_t = y_t - \boldsymbol{\gamma}'_{yqx,T}(\boldsymbol{q}'_{\cdot t}, x_t)'$. Define $\boldsymbol{x} = (x_1, x_2, ..., x_T)'$, $\mathbf{y} = (y_1, y_2, ..., y_T)'$, $\boldsymbol{e} = (e_1, e_2, ..., e_T)'$, $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$, and $\theta_T = E\left(T^{-1}\boldsymbol{x}'\mathbf{M}_q\mathbf{X}_b\right)\boldsymbol{\beta}_b$, where $\mathbf{X}_b$ is $T \times k_b$ matrix of observations on $\boldsymbol{x}_{b,t}$. Finally, $c_p(n,\delta)$ is given by (8) with $0 < p < 1$ and $f(n,\delta) = cn^{\delta}$, for some $c, \delta > 0$,* *and there exists $\kappa_1 > 0$ such that $T = \ominus(n^{\kappa_1})$. Then, for any π in the range $0 < \pi < 1$, any $d_T > 0$ and bounded in T, and for some finite positive constants C_0 and C_1,*

$$\Pr\left[|t_x| > c_p(n,\delta) | \theta_T = 0\right] \leq \exp\left[\frac{-(1-\pi)^2 \sigma_{e,(T)}^2 \sigma_{x,(T)}^2 c_p^2(n,\delta)}{2(1+d_T)^2 \omega_{xe,T}^2}\right] + \exp\left[-C_0 T^{C_1}\right], \tag{A.103}$$

where

$$t_x = \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{y}}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}, \tag{A.104}$$

$$\sigma_{e,(T)}^2 = E\left(T^{-1}\mathbf{e}'\mathbf{e}\right), \ \sigma_{x,(T)}^2 = E\left(T^{-1}\mathbf{x}'\mathbf{M}_q\mathbf{x}\right), \tag{A.105}$$

and

$$\omega_{xe,T}^2 = \frac{1}{T}\sum_{t=1}^{T} E\left[(u_{x,t}\eta_t)^2\right]. \tag{A.106}$$

Under $\sigma_t^2 = \sigma^2$ and/or $E\left(u_{x,t}^2\right) = \sigma_{xt}^2 = \sigma_x^2$, for all $t = 1, 2, ..., T$,

$$\Pr\left[|t_x| > c_p(n,\delta) | \theta_T = 0\right] \leq \exp\left[\frac{-(1-\pi)^2 c_p^2(n,\delta)}{2(1+d_T)^2}\right] + \exp\left(-C_0 T^{C_1}\right). \tag{A.107}$$

In the case where $\theta_T \neq 0$, let $\theta_T = \ominus\left(T^{-\vartheta}\right)$, for some $0 \leq \vartheta < 1/2$, where $c_p(n,\delta) = O\left(T^{1/2-\vartheta-C_8}\right)$, for some positive C_8. Then, for some bounded positive sequence d_T, and for some $C_2, C_3 > 0$, we have

$$\Pr\left[|t_x| > c_p(n,\delta) | \theta_T \neq 0\right] > 1 - \exp\left(-C_2 T^{C_3}\right). \tag{A.108}$$

Proof. The DGP, given by (2), can be written as

$$\mathbf{y} = a\boldsymbol{\tau}_T + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = a\boldsymbol{\tau}_T + \mathbf{X}_a\boldsymbol{\beta}_a + \mathbf{X}_b\boldsymbol{\beta}_b + \mathbf{u}$$

where $\mathbf{X}_a$ is a subset of $\mathbf{Q}$. Let $\mathbf{Q}_x = (\mathbf{Q}, \boldsymbol{x})$, $\mathbf{M}_q = \mathbf{I}_T - \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'$, $\mathbf{M}_{qx} = \mathbf{I}_T - \mathbf{Q}_x(\mathbf{Q}_x'\mathbf{Q}_x)^{-1}\mathbf{Q}_x'$. Then, $\mathbf{M}_q\mathbf{X}_a = \mathbf{0}$, and let $\mathbf{M}_q\mathbf{X}_b = (\boldsymbol{x}_{bq,1}, \boldsymbol{x}_{bq,2}, ..., \boldsymbol{x}_{bq,T})'$. Then,

$$t_x = \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{y}}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}} = \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{X}_b\boldsymbol{\beta}_b}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}} + \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{u}}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}. \tag{A.109}$$

Let $\theta_T = E\left(T^{-1}\boldsymbol{x}'\mathbf{M}_q\mathbf{X}_b\right)\boldsymbol{\beta}_b$, $\boldsymbol{\eta} = \mathbf{X}_b\boldsymbol{\beta}_b + \mathbf{u}$, $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_T)'$, and write (A.109) as

$$t_x = \frac{\sqrt{T}\theta_T}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}} + \frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta_T\right)}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}. \tag{A.110}$$

First, consider the case where $\theta_T = 0$ and note that in this case

$$t_x = \frac{\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)^{-1/2} \frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{\sqrt{T}}}{\sqrt{(\mathbf{e}'\mathbf{e}/T)}}.$$

Now by Lemma 9, we have

$$\begin{aligned}
\Pr\left[|t_x| > c_p(n,\delta)\,|\theta_T = 0\right] &= \Pr\left[\left|\frac{\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)^{-1/2} \frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{\sqrt{T}}}{\sqrt{(\mathbf{e}'\mathbf{e}/T)}}\right| > c_p(n,\delta)\,|\theta_T = 0\right] \\
&\leq \Pr\left(\left|\frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{c_p(n,\delta)}{1+d_T}\right) + \exp\left(-C_0 T^{C_1}\right).
\end{aligned}$$

where $\sigma^2_{e,(T)}$ and $\sigma^2_{x,(T)}$ are defined by (A.105). Hence, noting that $c_p(n,\delta) = o(T^{C_0})$, for all $C_0 > 0$, under Assumption 3, and by Lemma 6, we have

$$\begin{aligned}
\Pr\left[|t_x| > c_p(n,\delta)\,|\theta_T = 0\right] \leq & \exp\left[\frac{-(1-\pi)^2 \sigma^2_{e,(T)}\sigma^2_{x,(T)} c_p^2(n,\delta)}{2(1+d_T)^2 \omega^2_{xe,T}}\right] \\
& + \exp\left(-C_0 T^{C_1}\right),
\end{aligned}$$

where

$$\omega^2_{xe,T} = \frac{1}{T}\sum_{t=1}^{T} E\left[(u_{x,t}\eta_t)^2\right] = \frac{1}{T}\sum_{t=1}^{T} E\left[u^2_{x,t}\left(\boldsymbol{x}'_{b,t}\boldsymbol{\beta}_b + u_t\right)^2\right],$$

and $u_{x,t}$, being the error in the regression of x_t on $\mathbf{Q}$, is defined by (20). Since by assumption u_t are distributed independently of $u_{x,t}$ and $\boldsymbol{x}_{b,t}$, then

$$\omega^2_{xe,T} = \frac{1}{T}\sum_{t=1}^{T} E\left[u^2_{x,t}\left(\boldsymbol{x}'_{bq,t}\boldsymbol{\beta}_b\right)^2\right] + \frac{1}{T}\sum_{t=1}^{T} E\left(u^2_{xt}\right) E\left(u^2_t\right),$$

where $\boldsymbol{x}'_{bq,t}\boldsymbol{\beta}_b$ is the t-th element of $\mathbf{M}_q\mathbf{X}_b\boldsymbol{\beta}_b$. Furthermore, $E\left[u^2_{x,t}\left(\boldsymbol{x}'_{bq,t}\boldsymbol{\beta}_b\right)^2\right] = E\left(u^2_{x,t}\right) E\left(\boldsymbol{x}'_{bq,t}\boldsymbol{\beta}_b\right)^2 = E\left(u^2_{x,t}\right)\boldsymbol{\beta}'_b E\left(\boldsymbol{x}_{bq,t}\boldsymbol{x}'_{bq,t}\right)\boldsymbol{\beta}_b$, noting that under $\theta = 0$, $u_{x,t}$ and $\mathbf{x}_{b,t}$ are independently distributed. Hence

$$\omega^2_{xe,T} = \frac{1}{T}\sum_{t=1}^{T} E\left(u^2_{x,t}\right)\boldsymbol{\beta}'_b E\left(\boldsymbol{x}_{bq,t}\boldsymbol{x}'_{bq,t}\right)\boldsymbol{\beta}_b + \frac{1}{T}\sum_{t=1}^{T} E\left(u^2_{xt}\right) E\left(u^2_t\right). \tag{A.111}$$

Similarly

$$\begin{aligned}
\sigma^2_{e,(T)} &= E\left(T^{-1}\mathbf{e}'\mathbf{e}\right) = E\left(T^{-1}\boldsymbol{\eta}'\mathbf{M}_{qx}\boldsymbol{\eta}\right) = E\left[T^{-1}\left(\mathbf{X}_b\boldsymbol{\beta}_b + \mathbf{u}\right)'\mathbf{M}_{qx}\left(\mathbf{X}_b\boldsymbol{\beta}_b + \mathbf{u}\right)\right] \\
&= \boldsymbol{\beta}'_b E\left(T^{-1}\mathbf{X}'_b\mathbf{M}_{qx}\mathbf{X}_b\right)\boldsymbol{\beta}_b + T^{-1}\sum_{t=1}^{T} E\left(u^2_t\right),
\end{aligned}$$

and since under $\theta = 0$, $\mathbf{x}$ being a pure noise variable will be distributed independently of $\mathbf{X}_b$, then $E\left(T^{-1}\mathbf{X}_b'\mathbf{M}_{qx}\mathbf{X}_b\right) = E\left(T^{-1}\mathbf{X}_b'\mathbf{M}_q\mathbf{X}_b\right)$, and we have

$$\begin{aligned}\sigma_{e,(T)}^2 &= \boldsymbol{\beta}_b' E\left(T^{-1}\mathbf{X}_b'\mathbf{M}_q\mathbf{X}_b\right)\boldsymbol{\beta}_b + T^{-1}\sum_{t=1}^{T} E\left(u_t^2\right) \\ &= \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\beta}_b' E\left(\boldsymbol{x}_{bq,t}\boldsymbol{x}_{bq,t}'\right)\boldsymbol{\beta}_b + T^{-1}\sum_{t=1}^{T} E\left(u_t^2\right). \end{aligned} \tag{A.112}$$

Using (A.111) and (A.112), it is now easily seen that if either $E\left(u_{x,t}^2\right) = \sigma_{ux}^2$ or $E\left(u_t^2\right) = \sigma^2$, for all t, then we have $\omega_{xe,T}^2 = \sigma_{e,(T)}^2\sigma_{x,(T)}^2$, and hence

$$\Pr\left[|t_x| > c_p(n,\delta)\,|\theta_T = 0\right] \leq \exp\left[\frac{-(1-\pi)^2 c_p^2(n,\delta)}{2(1+d_T)^2}\right] + \exp\left(-C_0T^{C_1}\right),$$

giving a rate that does not depend on error variances. Next, we consider $\theta_T \neq 0$. By (A.101) of Lemma 9, for $d_T > 0$ and bounded in T,

$$\Pr\left[\left|\frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{y}}{\sqrt{(\mathbf{e}'\mathbf{e}/T)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}\right| > c_p(n,\delta)\right] \leq \Pr\left(\left|\frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{y}}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{c_p(n,\delta)}{1+d_T}\right) + \exp\left(-C_0T^{C_1}\right).$$

We then have

$$\begin{aligned}\frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{y}}{\sigma_{e,(T)}\sigma_{x,(T)}} &= \frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\mathbf{X}_b\boldsymbol{\beta}_b}{T} - \theta_T\right)}{\sigma_{e,(T)}\sigma_{x,(T)}} + \frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{u}}{\sigma_{e,(T)}\sigma_{x,(T)}} + \frac{T^{1/2}\theta_T}{\sigma_{e,(T)}\sigma_{x,(T)}} \\ &= \frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta_T\right)}{\sigma_{e,(T)}\sigma_{x,(T)}} + \frac{T^{1/2}\theta_T}{\sigma_{e,(T)}\sigma_{x,(T)}}.\end{aligned}$$

Then

$$\begin{aligned}&\Pr\left[\left|\frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta_T\right)}{\sigma_{e,(T)}\sigma_{x,(T)}} + \frac{T^{1/2}\theta_T}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{c_p(n,\delta)}{1+d_T}\right] \\ &= 1 - \Pr\left[\left|\frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta_T\right)}{\sigma_{e,(T)}\sigma_{x,(T)}} + \frac{T^{1/2}\theta_T}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| \leq \frac{c_p(n,\delta)}{1+d_T}\right].\end{aligned}$$

Note that since $c_p(n,\delta)$ is given by (8), then, $\frac{T^{1/2}|\theta_T|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{c_p(n,\delta)}{1+d_T} > 0$. Then by Lemma A2,

$$\begin{aligned}&\Pr\left[\left|\frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta_T\right)}{\sigma_{e,(T)}\sigma_{x,(T)}} + \frac{T^{1/2}\theta_T}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| \leq \frac{c_p(n,\delta)}{1+d_T}\right] \\ &\leq \Pr\left[\left|\frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta_T\right)}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{T^{1/2}|\theta_T|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{c_p(n,\delta)}{1+d_T}\right].\end{aligned}$$

But, setting $\zeta_T = T^{1/2}\left[\frac{T^{1/2}|\theta_T|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{c_p(n,\delta)}{1+d_T}\right]$ and noting that $\theta_T = O(T^{-\vartheta})$, $0 \leq \vartheta < 1/2$, implies that this choice of ζ_T satisfies $\zeta_T = \ominus\left(T^{\lambda}\right)$ with $\lambda = 1-\vartheta$, (A.68) of Lemma 6 applies regardless of $s > 0$, which gives us

$$\Pr\left[\left|\frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta_T\right)}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{T^{1/2}\left|\theta_T\right|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{c_p\left(n,\delta\right)}{1+d_T}\right] \tag{A.113}$$
$$\leq C_4 \exp\left\{-C_5\left[T^{1/2}\left(\frac{T^{1/2}\left|\theta_T\right|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{c_p\left(n,\delta\right)}{1+d_T}\right)\right]^{s/(s+2)}\right\}$$
$$+\exp\left(-C_6T^{C_7}\right),$$

for some C_4, C_5, C_6 and $C_7 > 0$. Hence, as long as the assumption that $c_p\left(n,\delta\right) = O\left(T^{1/2-\vartheta-C_8}\right)$ holds, for some positive C_8, there must exist positive finite constants C_2 and C_3, such that

$$\Pr\left[\left|\frac{T^{1/2}\left(\frac{\boldsymbol{x}'\mathbf{M}_q\boldsymbol{\eta}}{T} - \theta\right)}{\sigma_{e,(T)}\sigma_{x,(T)}}\right| > \frac{T^{1/2}\left|\theta_T\right|}{\sigma_{e,(T)}\sigma_{x,(T)}} - \frac{c_p\left(n,\delta\right)}{1+d_T}\right] \leq \exp\left(-C_2T^{C_3}\right) \tag{A.114}$$

for any $s > 0$. So overall

$$\Pr\left[\left|\frac{T^{-1/2}\boldsymbol{x}'\mathbf{M}_q\mathbf{y}}{\sqrt{\left(\mathbf{e}'\mathbf{e}/T\right)\left(\frac{\mathbf{x}'\mathbf{M}_q\mathbf{x}}{T}\right)}}\right| > c_p\left(n,\delta\right)\right] > 1-\exp\left(-C_2T^{C_3}\right).$$

■

Lemma 11 *Let* $\mathbf{S}_a$ *and* $\mathbf{S}_b$*, respectively, be* $T \times l_{a,T}$ *and* $T \times l_{b,T}$ *matrices of observations on* $s_{a,it}$*, and* $s_{b,it}$*, for* $i = 1,2,...,l_T$*,* $t = 1,2,...,T$*, and suppose that* $\{s_{a,it}, s_{b,it}\}$ *are either non-stochastic and bounded, or random with finite* 8^{th} *order moments. Consider the sample covariance matrix* $\hat{\boldsymbol{\Sigma}}_{ab} = T^{-1}\mathbf{S}_a'\mathbf{S}_b$ *and denote its expectations by* $\boldsymbol{\Sigma}_{ab} = T^{-1}E\left(\mathbf{S}_a'\mathbf{S}_b\right)$*. Let*

$$z_{ij,t} = s_{a,it}s_{b,jt} - E\left(s_{a,it}s_{b,jt}\right),$$

and suppose that

$$\sup_{i,j}\left[\sum_{t=1}^{T}\sum_{t'=1}^{T}E(z_{ij,t}z_{ij,t'})\right] = O\left(T\right). \tag{A.115}$$

Then,

$$E\left\|\hat{\boldsymbol{\Sigma}}_{ab} - \boldsymbol{\Sigma}_{ab}\right\|_F^2 = O\left(\frac{l_{a,T}l_{b,T}}{T}\right). \tag{A.116}$$

If, in addition,

$$\sup_{i,j,i',j'}\left[\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T}E(z_{ij,t}z_{ij,t'}z_{i'j',s}z_{i'j',s'})\right] = O\left(T^2\right), \tag{A.117}$$

then

$$E\left\|\hat{\mathbf{\Sigma}}_{ab}-\mathbf{\Sigma}_{ab}\right\|_F^4 = O\left(\frac{l_{a,T}^2 l_{b,T}^2}{T^2}\right). \tag{A.118}$$

Proof. We first note that $E(z_{ij,t}z_{ij,t'})$ and $E\left(z_{ij,t}z_{ij,t'}z_{i'j',s}z_{i'j',s'}\right)$ exist since by assumption $\{s_{a,it}, s_{b,it}\}$ have finite 8^{th} order moments. The (i,j) element of $\hat{\mathbf{\Sigma}}_{ab}-\mathbf{\Sigma}_{ab}$ is given by

$$a_{ij,T} = T^{-1}\sum_{t=1}^{T} z_{ij,t}, \tag{A.119}$$

and hence

$$\begin{aligned} E\left\|\hat{\mathbf{\Sigma}}_{ab}-\mathbf{\Sigma}_{ab}\right\|_F^2 &= \sum_{i=1}^{l_{a,T}}\sum_{j=1}^{l_{b,T}} E\left(a_{ij,T}^2\right) = T^{-2}\sum_{i=1}^{l_{a,T}}\sum_{j=1}^{l_{b,T}}\sum_{t=1}^{T}\sum_{t'=1}^{T} E\left(z_{ij,t}z_{ij,t'}\right) \\ &\leq \frac{l_{a,T}l_{b,T}}{T^2}\sup_{i,j}\left[\sum_{t=1}^{T}\sum_{t'=1}^{T} E(z_{ij,t}z_{ij,t'})\right], \end{aligned}$$

and (A.116) follows from (A.115). Similarly,

$$\begin{aligned} \left\|\hat{\mathbf{\Sigma}}_{ab}-\mathbf{\Sigma}_{ab}\right\|_F^4 &= \left(\sum_{i=1}^{l_{a,T}}\sum_{j=1}^{l_{b,T}} a_{ij,T}^2\right)^2 \\ &= \sum_{i=1}^{l_{a,T}}\sum_{j=1}^{l_{b,T}}\sum_{i'=1}^{l_{a,T}}\sum_{j'=1}^{l_{b,T}} a_{ij,T}^2 a_{i'j',T}^2. \end{aligned}$$

But using (A.119) we have

$$\begin{aligned} a_{ij,T}^2 a_{i'j',T}^2 &= T^{-4}\left(\sum_{t=1}^{T}\sum_{t'=1}^{T} z_{ij,t}z_{ij,t'}\right)\left(\sum_{s=1}^{T}\sum_{s'=1}^{T} z_{i'j',s}z_{i'j',s'}\right) \\ &= T^{-4}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} z_{ij,t}z_{ij,t'}z_{i'j',s}z_{i'j',s'}, \end{aligned}$$

and

$$\begin{aligned} E\left\|\hat{\mathbf{\Sigma}}_{ab}-\mathbf{\Sigma}_{ab}\right\|_F^4 &= T^{-4}\sum_{i=1}^{l_{a,T}}\sum_{j=1}^{l_{b,T}}\sum_{i'=1}^{l_{a,T}}\sum_{j'=1}^{l_{b,T}}\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} E\left(z_{ij,t}z_{ij,t'}z_{i'j',s}z_{i'j',s'}\right) \\ &\leq \frac{l_{a,T}^2 l_{b,T}^2}{T^4}\sup_{i,j,i',j'}\left[\sum_{t=1}^{T}\sum_{t'=1}^{T}\sum_{s=1}^{T}\sum_{s'=1}^{T} E\left(z_{ij,t}z_{ij,t'}z_{i'j',s}z_{i'j',s'}\right)\right]. \end{aligned}$$

Result (A.118) now follows from (A.117). ■

Remark 6 *It is clear that conditions (A.115) and (A.117) are met under Assumption 3 that requires z_{it} to be a martingale difference process. But it is easily seen that condition (A.115)* *also follows if we assume that $s_{a,it}$ and $s_{b,jt}$ are stationary processes with finite 8-th moments, since the product of stationary processes is also a stationary process under a certain additional cross-moment conditions (Wecker (1978)). The results of the lemma also follow readily if we assume that $s_{a,it}$ and $s_{b,jt'}$ are independently distributed for all $i \neq j$ and all t and t'.*

Lemma 12 *Consider the data generating process (1) with k signal, k^* pseudo-signal, and $n-k-k^*$ noise variables. Let $\mathcal{T}_k$ be the event that the OCMT procedure stops after k stages or less, and suppose that conditions of Lemma 10 hold. Let $k^* = \ominus(n^\epsilon)$ for some $0 \leq \epsilon < \min\{1, \kappa_1/3\}$, where κ_1 is the positive constant that defines the rate for $T = \ominus(n^{\kappa_1})$ in Lemma 10. Moreover, let $\delta > 0$ and $\delta^* > 0$ denote the critical value exponents for stage 1 and subsequent stages of the OCMT procedure, respectively. Then,*

$$\Pr(\mathcal{T}_k) = 1 + O\left(n^{1-\nu-\varkappa\delta}\right) + O\left(n^{1-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 n^{C_1\kappa_1}\right)\right], \tag{A.120}$$

for some $C_0, C_1 > 0$, any $\varkappa$ in $0 < \varkappa < 1$, and any ν in $\epsilon < \nu < \kappa_1/3$.

Proof. Consider the event $\mathcal{D}_{k,T}$, defined in (A.7), for $s = k \geq 1$, which is the event that the number of variables selected in the first k stages of OCMT is smaller than or equal to $l_T = \ominus(n^\nu)$, where ν lies in the interval $\epsilon < \nu < \kappa_1/3$. Such a ν exists since by assumption $0 \leq \epsilon < \min\{1, \kappa_1/3\}$. We have $\Pr(\mathcal{T}_k) = 1 - \Pr(\mathcal{T}_k^c)$, and

$$\begin{aligned}\Pr(\mathcal{T}_k^c) &= \Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T})\Pr(\mathcal{D}_{k,T}) + \Pr\left(\mathcal{T}_k^c|\mathcal{D}_{k,T}^c\right)\Pr\left(\mathcal{D}_{k,T}^c\right) \\ &\leq \Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}) + \Pr\left(\mathcal{D}_{k,T}^c\right),\end{aligned}$$

Therefore,

$$\Pr(\mathcal{T}_k) \geq 1 - \Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}) - \Pr\left(\mathcal{D}_{k,T}^c\right). \tag{A.121}$$

We note that

$$\Pr(\mathcal{D}_{k,T}) \geq \Pr\left[\left(\hat{k}^o_{n,T,(1)} \leq \frac{l_T}{k}\right) \cap \left(\hat{k}^o_{n,T,(2)} \leq \frac{l_T}{k}\middle| \mathcal{D}_{1,T}\right) \cap ... \cap \left(\hat{k}^o_{n,T,(k)} \leq \frac{l_T}{k}\middle| \mathcal{D}_{k-1,T}\right)\right],$$

where $\hat{k}^o_{n,T,(s)}$ is the number of variables selected in the s-th stage of OCMT and $\mathcal{D}_{s,T}$ for $s = 1, 2, ..., k$ is defined in (A.7). Hence

$$\Pr\left(\mathcal{D}_{k,T}^c\right) \leq \Pr\left\{\left[\begin{array}{c}\left(\hat{k}^o_{n,T,(1)} \leq \frac{l_T}{k}\right) \cap \left(\hat{k}^o_{n,T,(2)} \leq \frac{l_T}{k}\middle| \mathcal{D}_{1,T}\right) \cap ... \\ \cap \left(\hat{k}^o_{n,T,(k)} \leq \frac{l_T}{k}\middle| \mathcal{D}_{k-1,T}\right)\end{array}\right]^c\right\}.$$

Furthermore

$$\Pr\left\{\left[\begin{array}{c}\left(\hat{k}^o_{n,T,(1)} \le \frac{l_T}{k}\right) \cap \left(\hat{k}^o_{n,T,(2)} \le \frac{l_T}{k}\middle| \mathcal{D}_{1,T}\right) \cap ... \\ \cap \left(\hat{k}^o_{n,T,(k)} \le \frac{l_T}{k}\middle| \mathcal{D}_{k-1,T}\right)\end{array}\right]^c\right\}$$

$$= \Pr\left\{\left[\begin{array}{c}\left(\hat{k}^o_{n,T,(1)} > \frac{l_T}{k}\right) \cup \left(\hat{k}^o_{n,T,(2)} > \frac{l_T}{k}\middle| \mathcal{D}_{1,T}\right) \cup ... \\ \cup \left(\hat{k}^o_{n,T,(k)} > \frac{l_T}{k}\middle| \mathcal{D}_{k-1,T}\right)\end{array}\right]\right\}$$

$$\le \Pr\left(\hat{k}^o_{n,T,(1)} > \frac{l_T}{k}\right) + \sum_{s=2}^{k} \Pr\left(\hat{k}^o_{n,T,(s)} > \frac{l_T}{k}\middle| \mathcal{D}_{s-1,T}\right).$$

Since k is finite and $0 \le \epsilon < \nu$, there exists T_0 such that for all $T > T_0$ we have $l_T/k > k + k^*$, and we can apply (A.130) of Lemma 13 (for $j = l_T/k - k - k^* > 0$), to obtain

$$\Pr\left(\hat{k}^o_{n,T,(1)} > \frac{l_T}{k}\right) = \Pr\left(\hat{k}^o_{n,T,(1)} - k - k^* > \frac{l_T}{k} - k - k^*\right)$$
$$\le \frac{n-k-k^*}{\frac{l_T}{k} - k - k^*}\left\{\exp\left[-\frac{\varkappa c_p^2(n,\delta)}{2}\right] + \exp(-C_0 T^{C_1})\right\},$$

for some $C_0, C_1 > 0$ and any $0 < \varkappa < 1$. Noting that for $0 \le \epsilon < \nu$,

$$\frac{n-k-k^*}{\frac{l_T}{k} - k - k^*} = \ominus\left(n^{1-\nu}\right), \tag{A.122}$$

and using also result (ii) of Lemma 2, we obtain

$$\Pr\left(\hat{k}^o_{n,T,(1)} > \frac{l_T}{k}\right) = O\left(n^{1-\nu-\varkappa\delta}\right) + O\left[n^{1-\nu}\exp\left(-C_0 T^{C_1}\right)\right].$$

Similarly,

$$\Pr\left(\hat{k}^o_{n,T,(s)} > \frac{l_T}{k}\middle| \mathcal{D}_{s-1,T}\right) = \Pr\left(\hat{k}^o_{n,T,(s)} - k - k^* > \frac{l_T}{k} - k - k^*\middle| \mathcal{D}_{s-1,T}\right)$$
$$\le \frac{n-k-k^*}{\frac{l_T}{k} - k - k^*}\left\{\exp\left[-\frac{\varkappa c_p^2(n,\delta^*)}{2}\right] + \exp(-C_0 T^{C_1})\right\}$$
$$= O\left(n^{1-\nu-\varkappa\delta^*}\right) + O\left[n^{1-\nu}\exp\left(-C_0 T^{C_1}\right)\right],$$

where the critical value exponent in the higher stages ($s > 1$) of OCMT (δ^*) could differ from the one in the first stage (δ). So, overall

$$\Pr\left(\mathcal{D}^c_{k,T}\right) \le \Pr\left(\hat{k}^o_{n,T,(1)} > \frac{l_T}{k}\right) + \sum_{s=2}^{k} \Pr\left(\hat{k}^o_{n,T,(s)} > \frac{l_T}{k}\middle| \mathcal{D}_{s-1,T}\right)$$
$$= O\left(n^{1-\nu-\varkappa\delta}\right) + O\left(n^{1-\nu-\varkappa\delta^*}\right) + O\left[n^{1-\nu}\exp\left(-C_0 T^{C_1}\right)\right], \tag{A.123}$$

for some $C_0, C_1 > 0$, any $\varkappa$ in $0 < \varkappa < 1$, and any ν in $\epsilon < \nu < \kappa_1/3$. Next, consider $\Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T})$, and note that

$$\begin{aligned}\Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}) &= \Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}, \mathcal{L}_k)\Pr(\mathcal{L}_k|\mathcal{D}_{k,T}) + \Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}, \mathcal{L}_k^c)\Pr(\mathcal{L}_k^c|\mathcal{D}_{k,T}) \\ &\leq \Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}, \mathcal{L}_k) + \Pr(\mathcal{L}_k^c|\mathcal{D}_{k,T}), \qquad (A.124)\end{aligned}$$

where $\Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}, \mathcal{L}_k)$ is the probability that a noise variable will be selected in a stage of OCMT that includes as regressors all signals, conditional on the event that fewer than l_T variables are selected in the first k steps of OCMT. Note that the event $\mathcal{T}_k^c|\mathcal{D}_{k,T}, \mathcal{L}_k$ can only occur if OCMT selects some pseudo signals and/or some noise variables in stage $k+1$. But the net effect coefficient of pseudo signal variables in stage $k+1$ must be zero when all signal variables were selected in earlier stages ($s = 1, 2, ..., k$), namely $\theta_{i,(k+1)} = 0$ for $i = k+1, k+2, ..., k+k^*$. Moreover, $\theta_{i,(k+1)} = 0$ also for $i = k+k^*+1, k+k^*+2, ..., n$, since the net effect coefficient of noise variables is always zero (in any stage). Therefore, we have

$$\Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}, \mathcal{L}_k) \leq \sum_{i=k+1}^{n} \Pr\left[\left|t_{\hat{\phi}_{i,(k+1)}}\right| > c_p(n, \delta^*)\,|\theta_{i,(k+1)} = 0, \mathcal{D}_{k,T}\right].$$

Note that the number of regressors in the regressions involving the t statistics $t_{\hat{\theta}_{i,(k+1)}}$, does not exceed $l_T = \ominus(n^\nu)$, for ν in the interval $0 \leq \epsilon < \nu < \kappa_1/3$ and hence $l_T = o(T^{1/3})$ as required by the conditions of Lemma 10. Using (A.107) of Lemma 10, we have

$$\begin{aligned}\Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}, \mathcal{L}_k) \leq (n-k)\exp\left[\frac{-\varkappa c_p^2(n, \delta^*)}{2}\right] \\ + (n-k)\exp\left(-C_0 T^{C_1}\right). \qquad (A.125)\end{aligned}$$

for some $C_0, C_1 > 0$ and any $0 < \varkappa < 1$. By Lemma 2, $\exp\left[-\varkappa c_p^2(n, \delta^*)/2\right] = \ominus\left(n^{-\varkappa\delta^*}\right)$, for any $0 < \varkappa < 1$, and noting that $n - k \leq n$ we obtain

$$\Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}, \mathcal{L}_k) = O\left(n^{1-\varkappa\delta^*}\right) + O\left[n\exp\left(-C_0 T^{C_1}\right)\right]. \qquad (A.126)$$

Consider next the second term of (A.124), $\Pr(\mathcal{L}_k^c|\mathcal{D}_{k,T})$, and recall from (A.6) that $\mathcal{L}_k = \cap_{i=1}^k \mathcal{L}_{i,k}$ where $\mathcal{L}_{i,k}$, defined by (A.5), is $\mathcal{L}_{i,k} = \cup_{j=1}^k \mathcal{B}_{i,j}$, $i = 1, 2, ..., k$. Hence $\mathcal{L}_{i,k}^c = \cap_{j=1}^k \mathcal{B}_{i,j}^c$, and

$$\begin{aligned}\Pr\left(\mathcal{L}_{i,k}^c\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right) = \Pr\left(\cap_{j=1}^k \mathcal{B}_{i,j}^c\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right) = \\ \Pr\left(\mathcal{B}_{i,1}^c\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right)\Pr\left(\mathcal{B}_{i,2}^c\middle|\mathcal{B}_{i,1}^c, \mathcal{T}_k, \mathcal{D}_{k,T}\right) \\ \Pr\left(\mathcal{B}_{i,3}^c\middle|\mathcal{B}_{i,2}^c \cap \mathcal{B}_{i,1}^c, \mathcal{T}_k, \mathcal{D}_{k,T}\right) \times ... \times \\ \Pr\left(\mathcal{B}_{i,k}^c\middle|\mathcal{B}_{i,k-1}^c \cap ... \cap \mathcal{B}_{i,1}^c, \mathcal{T}_k, \mathcal{D}_{k,T}\right).\end{aligned}$$

But by Proposition 1 we are guaranteed that for some $1 \leq j \leq k$, $\theta_{i,(j)} \neq 0$. Therefore,

$$\Pr\left(\mathcal{B}_{i,j}^c\middle|\mathcal{B}_{i,j-1}^c \cap ... \cap \mathcal{B}_{i,1}^c, \mathcal{T}_k, \mathcal{D}_{k,T}\right) = \Pr\left(\mathcal{B}_{i,j}^c\middle|\mathcal{B}_{i,j-1}^c \cap ... \cap \mathcal{B}_{i,1}^c, \theta_{i,(j)} \neq 0, \mathcal{T}_k, \mathcal{D}_{k,T}\right),$$

and by (A.108) of Lemma 10,

$$\Pr\left(\mathcal{B}_{i,j}^{c}\middle|\,\mathcal{B}_{i,j-1}^{c}\cap...\cap\mathcal{B}_{i,1}^{c},\theta_{i,(j)}\neq 0,\mathcal{T}_{k},\mathcal{D}_{k,T}\right)=O\left[\exp\left(-C_{0}T^{C_{1}}\right)\right],$$

for some $C_0, C_1 > 0$. Therefore, for some $j \in \{1, 2, ..., k\}$ and $C_0, C_1 > 0$,

$$\begin{aligned}\Pr\left(\mathcal{L}_{i,k}^{c}\middle|\,\mathcal{T}_{k},\mathcal{D}_{k,T}\right)&\leq\Pr\left(\mathcal{B}_{i,j}^{c}\middle|\,\mathcal{B}_{i,j-1}^{c}\cap...\cap\mathcal{B}_{i,1}^{c},\theta_{i,(j)}\neq 0,\mathcal{T}_{k},\mathcal{D}_{k,T}\right)\\&=O\left[\exp\left(-C_{0}T^{C_{1}}\right)\right].\end{aligned}\tag{A.127}$$

Noting that k is finite and

$$\begin{aligned}\Pr\left(\mathcal{L}_{k}^{c}\middle|\,\mathcal{T}_{k},\mathcal{D}_{k,T}\right)&=\Pr\left(\cup_{i=1}^{k}\mathcal{L}_{ik}^{c}\middle|\,\mathcal{T}_{k},\mathcal{D}_{k,T}\right)\\&\leq\sum_{i=1}^{k}\Pr\left(\mathcal{L}_{ik}^{c}\middle|\,\mathcal{T}_{k},\mathcal{D}_{k,T}\right),\end{aligned}$$

it follows, using (A.127), that

$$\Pr\left(\mathcal{L}_{k}^{c}\middle|\,\mathcal{T}_{k},\mathcal{D}_{k,T}\right)=O\left[\exp\left(-C_{0}T^{C_{1}}\right)\right],\tag{A.128}$$

for some $C_0, C_1 > 0$. Using (A.126) and (A.128) in (A.124) now gives[15]

$$\Pr\left(\mathcal{T}_{k}^{c}|\mathcal{D}_{k,T}\right)=O\left(n^{1-\varkappa\delta^{*}}\right)+O\left[n\exp\left(-C_{0}T^{C_{1}}\right)\right].\tag{A.129}$$

Using (A.123) and (A.129) in (A.121), yields

$$\begin{aligned}\Pr\left(\mathcal{T}_{k}\right)=&1+O\left(n^{1-\nu-\varkappa\delta}\right)+O\left(n^{1-\nu-\varkappa\delta^{*}}\right)+O\left[n^{1-\nu}\exp\left(-C_{0}T^{C_{1}}\right)\right]\\&+O\left(n^{1-\varkappa\delta^{*}}\right)+O\left[n\exp\left(-C_{2}T^{C_{3}}\right)\right],\end{aligned}$$

for some $C_0, C_1, C_2, C_3 > 0$ and any $\varkappa$ in $0 < \varkappa < 1$, and any ν in $\epsilon < \nu < \kappa_1/3$. But $O\left(n^{1-\nu-\varkappa\delta^*}\right)$ is dominated by $O\left(n^{1-\varkappa\delta^*}\right)$, and $O\left[n^{1-\nu}\exp\left(-C_0T^{C_1}\right)\right]$ is dominated by $O\left[n\exp\left(-C_2T^{C_3}\right)\right]$, since $\nu > \epsilon \geq 0$. Hence,

$$\Pr\left(\mathcal{T}_{k}\right)=1+O\left(n^{1-\nu-\varkappa\delta}\right)+O\left(n^{1-\varkappa\delta^{*}}\right)+O\left[n\exp\left(-C_{0}T^{C_{1}}\right)\right],$$

for some $C_0, C_1 > 0$, any $\varkappa$ in $0 < \varkappa < 1$, and any ν in $\epsilon < \nu < \kappa_1/3$. This result in turn establishes (A.120), noting that $T = \ominus(n^{\kappa_1})$. ■

Lemma 13 *Consider the data generating process (1) with k signal variables, k^* pseudo-signal variables, and $n - k - k^*$ noise variables. Let $\hat{k}^o_{n,T,(s)}$ be the number of variables selected at the stage s of the OCMT procedure and suppose that conditions of Lemma 10 hold. Let $k^* = \ominus(n^\epsilon)$ for some $0 \leq \epsilon < \min\{1, \kappa_1/3\}$, where κ_1 is the positive constant that defines the rate for $T = \ominus(n^{\kappa_1})$ in Lemma 10. Let $\mathcal{D}_{s,T}$, be the event that the number of variables*

[15] We have dropped the term $O\left[\exp\left(-C_0T^{C_1}\right)\right]$, which is dominated by $O\left[n\exp\left(-C_0T^{C_1}\right)\right]$.

selected in the first s stages of OCMT is smaller than or equal to l_T, where $l_T = \ominus(n^{\nu})$ and ν satisfies $\epsilon < \nu < \kappa_1/3$. Then there exist constants $C_0, C_1 > 0$ such that for any $0 < \varkappa < 1$, any $\delta_s > 0$, and any $j > 0$, it follows that

$$\Pr\left(\hat{k}^o_{n,T,(s)} - k - k^* > j | \mathcal{D}_{s-1,T}\right) \leq \frac{n-k-k^*}{j}\left\{\exp\left[-\frac{\varkappa c_p^2(n,\delta_s)}{2}\right] + \exp(-C_0 T^{C_1})\right\}, \tag{A.130}$$

for $s = 1, 2, ..., k$.

Proof. By convention, the number of variables selected at the stage zero of OCMT is zero. Conditioning on $\mathcal{D}_{s-1,T}$ allows the application of Lemma 10. We drop the conditioning notation in the rest of the proof to simplify notations. Then, by Markov's inequality

$$\Pr\left(\hat{k}^o_{n,T,(s)} - k - k^* > j\right) \leq \frac{E\left(\hat{k}^o_{n,T,(s)} - k - k^*\right)}{j}. \tag{A.131}$$

But

$$\begin{aligned} E\left(\hat{k}^o_{n,T,(s)}\right) &= \sum_{i=1}^{n} E\left[I_{(s)}\widehat{(\beta_i \neq 0)}\right] \\ &= \sum_{i=1}^{k+k^*} E\left[I_{(s)}\widehat{(\beta_i \neq 0)}\right] + \sum_{i=k+k^*+1}^{n} E\left[I_{(s)}\widehat{(\beta_i \neq 0)}\left|\theta_{i,(s)} = 0\right.\right]. \\ &\leq k + k^* + \sum_{i=k+k^*+1}^{n} E\left[I_{(s)}\widehat{(\beta_i \neq 0)}\left|\theta_{i,(s)} = 0\right.\right], \end{aligned}$$

where we have used $I_{(s)}\widehat{(\beta_i \neq 0)} \leq 1$. Moreover,

$$E\left[I_{(s)}\widehat{(\beta_i \neq 0)}\left|\theta_{i,(s)} = 0\right.\right] = \Pr\left(\left|t_{\hat{\phi}_{T,i,(s)}}\right| > c_p(n,\delta_s)\left|\theta_{i,(s)} = 0\right.\right),$$

for $i = k+k^*+1, k+k^*+2, ..., n$, and using (A.107) of Lemma 10, we have (for some $0 < \varkappa < 1$ and $C_0, C_1 > 0$)

$$\sup_{i > k+k^*} \Pr\left(\left|t_{\hat{\phi}_{T,i,(s)}}\right| > c_p(n,\delta_s)\left|\theta_{i,(s)} = 0\right.\right) \leq \exp\left[-\frac{\varkappa c_p^2(n,\delta_s)}{2}\right] + \exp(-C_0 T^{C_1}).$$

Hence,

$$E\left(\hat{k}^o_{n,T,(s)}\right) - k - k^* \leq (n-k-k^*)\left\{\exp\left[-\frac{\varkappa c_p^2(n,\delta_s)}{2}\right] + \exp(-C_0 T^{C_1})\right\},$$

and therefore (using this result in (A.131))

$$\Pr\left(\hat{k}^o_{n,T,(s)} - k - k^* > j\right) \leq \frac{n-k-k^*}{j}\left\{\exp\left[-\frac{\varkappa c_p^2(n,\delta_s)}{2}\right] + \exp(-C_0 T^{C_1})\right\},$$

as desired. ■

Lemma 14 *Consider the data generating process (1) with k signal, k^* pseudo-signal, and $n-k-k^*$ noise variables. Let $\hat{k}_{n,T}$ be the number of variables selected by the OCMT procedure, and suppose that conditions of Lemma 10 hold. Let $k^* = \ominus(n^{\epsilon})$ for some $0 \leq \epsilon < \min\{1, \kappa_1/3\}$, where $\kappa_1 > 0$ is the positive constant that defines the rate for $T = \ominus(n^{\kappa_1})$ in Lemma 10. Moreover, let $\delta > 0$ and $\delta^* > 0$ denote the critical value exponents for stage 1 and subsequent stages of OCMT, respectively. Then for some C_0,$C_1 > 0$, any $\varkappa$ in $0 < \varkappa < 1$, and any ν in $\epsilon < \nu < \kappa_1/3$, we have*

$$\Pr\left(\hat{k}_{n,T} - k - k^* > j\right) = O\left(j^{-1}n^{1-\varkappa\delta}\right) + O\left(j^{-1}n^{2-\varkappa\delta^*}\right) + O\left[\frac{n^2}{j}\exp\left(-C_0 n^{C_1\kappa_1}\right)\right] + O\left(n^{1-\nu-\varkappa\delta}\right) + O\left(n^{1-\nu-\varkappa\delta^*}\right), \quad \text{(A.132)}$$

for $j = 1, 2, ..., n-k-k^*$.

Proof. Consider the event $\mathcal{D}_{k,T}$, defined in (A.7), for $s = k \geq 1$, and recall that this is the event that the number of variables selected in the first k stages of OCMT is smaller than or equal to $l_T = \ominus(n^{\nu})$, where ν satisfies $\epsilon < \nu < \kappa_1/3$, noting that by assumption $0 \leq \epsilon < \min\{1, \kappa_1/3\}$. We have

$$\begin{aligned}\Pr\left(\hat{k}_{n,T} - k - k^* > j\right) =& \Pr\left(\hat{k}_{n,T} - k - k^* > j|\mathcal{D}_{k,T}\right)\Pr\left(\mathcal{D}_{k,T}\right) \\ &+ \Pr\left(\hat{k}_{n,T} - k - k^* > j|\mathcal{D}^c_{k,T}\right)\Pr\left(\mathcal{D}^c_{k,T}\right) \\ \leq& \Pr\left(\hat{k}_{n,T} - k - k^* > j|\mathcal{D}_{k,T}\right) + \Pr\left(\mathcal{D}^c_{k,T}\right). \quad \text{(A.133)}\end{aligned}$$

An upper bound to $\Pr\left(\mathcal{D}^c_{k,T}\right)$ is established in (A.123). For the rest of the proof we focus on $\Pr\left(\hat{k}_{n,T} - k - k^* > j|\mathcal{D}_{k,T}\right)$. We first note that by Markov's inequality

$$\Pr\left(\hat{k}_{n,T} - k - k^* > j|\mathcal{D}_{k,T}\right) \leq \frac{E\left(\hat{k}_{n,T} - k - k^*|\mathcal{D}_{k,T}\right)}{j}. \quad \text{(A.134)}$$

But,

$$\begin{aligned}E\left(\hat{k}_{n,T}|\mathcal{D}_{k,T}\right) &= E\left(\hat{k}_{n,T}\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right)\Pr\left(\mathcal{T}_k|\mathcal{D}_{k,T}\right) + E\left(\hat{k}_{n,T}\middle|\mathcal{T}^c_k, \mathcal{D}_{k,T}\right)\Pr\left(\mathcal{T}^c_k|\mathcal{D}_{k,T}\right) \\ &\leq E\left(\hat{k}_{n,T}\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right) + E\left(\hat{k}_{n,T}\middle|\mathcal{T}^c_k, \mathcal{D}_{k,T}\right)\Pr\left(\mathcal{T}^c_k|\mathcal{D}_{k,T}\right). \quad \text{(A.135)}\end{aligned}$$

An upper bound on $\Pr\left(\mathcal{T}^c_k|\mathcal{D}_{k,T}\right)$ is derived in (A.129). We consider $E\left(\hat{k}_{n,T}\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right)$ next, and note that

$$\begin{aligned}E\left(\hat{k}_{n,T}\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right) =& \sum_{i=1}^{k}\Pr\left(\mathcal{L}_{i,k}\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right) + \sum_{i=k+1}^{k+k^*}\Pr\left(\mathcal{L}_{i,k}\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right) \\ &+ \sum_{i=k+k^*+1}^{n}\Pr\left(\mathcal{L}_{i,k}\middle|\mathcal{T}_k, \mathcal{D}_{k,T}\right), \quad \text{(A.136)}\end{aligned}$$

and it must also be that

$$E\left(\hat{k}_{n,T}\right| \mathcal{T}_k^c, \mathcal{D}_{k,T}\Big) \leq n. \tag{A.137}$$

(A.137) is a very loose upper bound (since $\hat{k}_{n,T}$ cannot exceed n by definition), but this bound will be sufficient for the purpose of this proof. Note that by (A.127) $\Pr\left(\mathcal{L}_{i,k}^c\right| \mathcal{T}_k, \mathcal{D}_{k,T}\Big) = O\left[\exp\left(-C_0 T^{C_1}\right)\right]$ for $i = 1, 2, ..., k$, and it follows that

$$\sum_{i=1}^{k} \Pr\left(\mathcal{L}_{i,k}\right| \mathcal{T}_k, \mathcal{D}_{k,T}\Big) = \sum_{i=1}^{k} \left[1 - \Pr\left(\mathcal{L}_{i,k}^c\right| \mathcal{T}_k, \mathcal{D}_{k,T}\Big)\right] = k + O\left[\exp\left(-C_0 T^{C_1}\right)\right], \tag{A.138}$$

for some $C_0, C_1 > 0$. Next, we have

$$\sum_{i=k+1}^{k+k^*} \Pr\left(\mathcal{L}_{i,k}\right| \mathcal{T}_k, \mathcal{D}_{k,T}\Big) \leq k^*, \tag{A.139}$$

since $0 \leq \Pr\left(\mathcal{L}_{i,k}\right| \mathcal{T}_k, \mathcal{D}_{k,T}\Big) \leq 1$. Now consider the last term on the right side of (A.136). Recalling that $\mathcal{L}_{i,k} = \cup_{s=1}^{k} \mathcal{B}_{i,s}$, then, given that $\theta_{i,(s)} = 0$ for all $i = k + k^* + 1, k + k^* + 2, ..., n$ and all $s = 1, 2, ..., k$, we have

$$\Pr\left[\mathcal{L}_{i,k} | \mathcal{T}_k, \mathcal{D}_{k,T}\right] \leq \sum_{s=1}^{k} \Pr\left(\mathcal{B}_{i,s} | \theta_{i,(s)} = 0, \mathcal{T}_k, \mathcal{D}_{k,T}\right), \text{ for } i > k + k^* + 1,$$

and hence

$$\begin{aligned}\sum_{i=k+k^*+1}^{n} \Pr\left(\mathcal{L}_{i,k}\right| \mathcal{T}_k, \mathcal{D}_{k,T}\Big) \leq &\sum_{i=k+k^*+1}^{n} \Pr\left(\mathcal{B}_{i,1}\right| \theta_{i,(1)} = 0, \mathcal{T}_k, \mathcal{D}_{k,T}\Big) \\ &+ \sum_{i=k+k^*+1}^{n} \sum_{s=2}^{k} \Pr\left(\mathcal{B}_{i,s}\right| \theta_{i,(s)} = 0, \mathcal{T}_k, \mathcal{D}_{k,T}\Big).\end{aligned}$$

Now using (A.26) for $\Pr\left(\mathcal{B}_{i,s}\right| \theta_{i,(s)} = 0, \mathcal{T}_k, \mathcal{D}_{k,T}\Big)$, $i = k+k^*+1, k+k^*+2, ..., n$, $s = 1, 2, ..., k$, it readily follows that (noting k is fixed and $n - k - k^* < n$)

$$\begin{aligned}\sum_{i=k+k^*+1}^{n} \Pr\left(\mathcal{L}_{i,k}\right| \mathcal{T}_k, \mathcal{D}_{k,T}\Big) =& O\left\{n \exp\left[-\frac{\varkappa c_p^2(n,\delta)}{2}\right]\right\} + O\left\{n \exp\left[-\frac{\varkappa c_p^2(n,\delta^*)}{2}\right]\right\} \\ &+ O\left[n \exp\left(-C_0 T^{C_1}\right)\right].\end{aligned} \tag{A.140}$$

Using (A.138)-(A.140) in (A.136), we obtain

$$\begin{aligned}E\left(\hat{k}_{n,T}\right| \mathcal{T}_k, \mathcal{D}_{k,T}\Big) \leq\ & k + k^* + C_0 \exp\left(-C_1 T^{C_2}\right) + C_3 n \exp\left[-\frac{\varkappa c_p^2(n,\delta)}{2}\right] \\ &+ C_4 n \exp\left[-\frac{\varkappa c_p^2(n,\delta^*)}{2}\right] + C_5 n \exp\left(-C_6 T^{C_7}\right),\end{aligned} \tag{A.141}$$

for some $C_0, C_1, ..., C_7 > 0$. (A.141) provides an upper bound on the first term on the right side of (A.135). Consider next the second term on the right side of (A.135). Using (A.129) for $\Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T})$ and (A.137) for $E\left(\hat{k}_{n,T}\middle|\mathcal{T}_k^c,\mathcal{D}_{k,T}\right)$ yields,

$$E\left(\hat{k}_{n,T}\middle|\mathcal{T}_k^c,\mathcal{D}_{k,T}\right)\Pr(\mathcal{T}_k^c|\mathcal{D}_{k,T}) = O\left(n^{2-\varkappa\delta^*}\right) + O\left[n^2\exp\left(-C_0T^{C_1}\right)\right]. \tag{A.142}$$

Using (A.141) and (A.142) gives an upper bound for $E\left(\hat{k}_{n,T}-k-k^*|\mathcal{D}_{k,T}\right)$, which when used in (A.134) yields the following bound on $\Pr\left(\hat{k}_{n,T}-k-k^*>j|\mathcal{D}_{k,T}\right)$,

$$\begin{aligned}\Pr\left(\hat{k}_{n,T}-k-k^*>j|\mathcal{D}_{k,T}\right) &= O\left[j^{-1}\exp\left(-C_0T^{C_1}\right)\right] + O\left\{\frac{n}{j}\exp\left[-\frac{\varkappa c_p^2(n,\delta)}{2}\right]\right\}\\ &+ O\left\{\frac{n}{j}\exp\left[-\frac{\varkappa c_p^2(n,\delta^*)}{2}\right]\right\} + O\left[\frac{n}{j}\exp\left(-C_2T^{C_3}\right)\right]\\ &+ O\left(j^{-1}n^{2-\varkappa\delta^*}\right) + O\left[\frac{n^2}{j}\exp\left(-C_4T^{C_5}\right)\right],\end{aligned}$$

for some $C_0, C_1, ..., C_5 > 0$. Noting that $O\left[j^{-1}\exp\left(-C_0T^{C_1}\right)\right]$ and $O\left[nj^{-1}\exp\left(-C_2T^{C_3}\right)\right]$ are both dominated by $O\left[n^2j^{-1}\exp\left(-C_4T^{C_5}\right)\right]$, and using result *(ii)* of Lemma 2 for the terms involving $c_p^2(n,\delta)$ and $c_p^2(n,\delta^*)$, we obtain

$$\begin{aligned}\Pr\left(\hat{k}_{n,T}-k-k^*>j|\mathcal{D}_{k,T}\right) &= O\left(j^{-1}n^{1-\varkappa\delta}\right) + O\left(j^{-1}n^{1-\varkappa\delta^*}\right)\\ &+ O\left(j^{-1}n^{2-\varkappa\delta^*}\right) + O\left[\frac{n^2}{j}\exp\left(-C_0T^{C_1}\right)\right].\end{aligned}$$

But $O\left(j^{-1}n^{1-\varkappa\delta^*}\right)$ is dominated by $O\left(j^{-1}n^{2-\varkappa\delta^*}\right)$, hence

$$\begin{aligned}\Pr\left(\hat{k}_{n,T}-k-k^*>j|\mathcal{D}_{k,T}\right) &= O\left(j^{-1}n^{1-\varkappa\delta}\right) + O\left(j^{-1}n^{2-\varkappa\delta^*}\right)\\ &+ O\left[\frac{n^2}{j}\exp\left(-C_0T^{C_1}\right)\right].\end{aligned} \tag{A.143}$$

Finally using (A.123) and (A.143) in (A.133), we have

$$\begin{aligned}\Pr\left(\hat{k}_{n,T}-k-k^*>j\right) =& O\left(j^{-1}n^{1-\varkappa\delta}\right) + O\left(j^{-1}n^{2-\varkappa\delta^*}\right) + O\left[\frac{n^2}{j}\exp\left(-C_0T^{C_1}\right)\right]\\ &+ O\left(n^{1-\nu-\varkappa\delta}\right) + O\left(n^{1-\nu-\varkappa\delta^*}\right) + O\left[n^{1-\nu}\exp\left(-C_2T^{C_3}\right)\right],\end{aligned}$$

for some $C_0, C_1, C_2, C_3 > 0$. Recalling that $0 \leq \epsilon < \nu < \kappa_1/3$ and $j < n$, the term $O\left[n^{1-\nu}\exp\left(-C_2T^{C_3}\right)\right]$ is always dominated by $O\left[\frac{n^2}{j}\exp\left(-C_0T^{C_1}\right)\right]$, and noting that $T = \ominus(n^{\kappa_1})$, establishes (A.132). ■

Lemma 15 *Suppose that the data generating process (DGP) is given by*

$$\underset{T\times 1}{\mathbf{y}} = \underset{T\times k+1}{\mathbf{X}} \cdot \underset{k+1\times 1}{\boldsymbol{\beta}} + \underset{T\times 1}{\mathbf{u}}, \tag{A.144}$$

where $\mathbf{X} = (\boldsymbol{\tau}_T, \mathbf{X}_k)$ *includes a column of ones,* $\boldsymbol{\tau}_T$*, and consider the regression model*

$$\underset{T\times 1}{\mathbf{y}} = \underset{T\times l_T}{\mathbf{S}} \cdot \underset{l_T\times 1}{\boldsymbol{\delta}} + \underset{T\times 1}{\varepsilon}. \tag{A.145}$$

where $\mathbf{u} = (u_1, u_2, ..., u_T)'$ *is independently distributed of* $\mathbf{X}$ *and* $\mathbf{S}$*,* $E(\mathbf{u}) = \mathbf{0}$*,* $E(\mathbf{uu}') = \sigma^2 \mathbf{I}_T$*,* $0 < \sigma^2 < \infty$*,* $\mathbf{I}_T$ *is a* $T \times T$ *identity matrix, and elements of* $\boldsymbol{\beta}$ *are bounded. In addition, it is assumed that the following conditions hold:*

i. Let $\boldsymbol{\Sigma}_{ss} = E(\mathbf{S}'\mathbf{S}/T)$ *with eigenvalues denoted by* $\mu_1 \le \mu_2 \le ... \le \mu_{l_T}$*. Let* $\mu_i = O(l_T)$*,* $i = l_T - M + 1,\ l_T - M + 2, ..., l_T$*, for some finite* M*, and* $\sup_{1\le i\le l_T - M} \mu_i < C_0 < \infty$*, for some* $C_0 > 0$*. In addition,* $\inf_{1\le i< l_T} \mu_i > C_1 > 0$*, for some* $C_1 > 0$*.*

ii. $E\left[\left(1 - \|\boldsymbol{\Sigma}_{ss}^{-1}\|_F \left\|\hat{\boldsymbol{\Sigma}}_{ss} - \boldsymbol{\Sigma}_{ss}\right\|_F\right)^{-4}\right] = O(1)$*, where* $\hat{\boldsymbol{\Sigma}}_{ss} = T^{-1}\mathbf{S}'\mathbf{S}$*.*

iii. Regressors in $\mathbf{S} = (s_{it})$ *have finite* 8^{th} *moments and* $z_{ij,t} = s_{it}s_{jt} - E(s_{it}s_{jt})$ *satisfies conditions (A.115) and (A.117) of Lemma 11. Moreover,* $z^*_{ij,t} = s_{it}x_{jt} - E(s_{it}x_{jt})$ *satisfies condition (A.115) of Lemma 11, and* $\|\boldsymbol{\Sigma}_{sx}\|_F = \|E(T^{-1}\mathbf{S}'\mathbf{X})\|_F = O(1)$*.*

Then, if $\mathbf{S} = (\mathbf{X}, \mathbf{W})$ *for some* $T \times k_w$ *matrix* $\mathbf{W}$*,*

$$E\left\|\hat{\boldsymbol{\delta}} - \boldsymbol{\beta}_0\right\| = O\left(\frac{l_T^2}{\sqrt{T}}\right), \tag{A.146}$$

where $\hat{\boldsymbol{\delta}}$ *is the least square estimator of* $\boldsymbol{\delta}$ *in the regression model (A.145) and* $\boldsymbol{\beta}_0 = \left(\boldsymbol{\beta}', \mathbf{0}'_{k_w}\right)'$*. Further, if some column vectors of* $\mathbf{X}$ *are not contained in* $\mathbf{S}$*, then*

$$E\left\|\hat{\boldsymbol{\delta}} - \boldsymbol{\beta}_0\right\| = O(l_T) + O\left(\frac{l_T^2}{\sqrt{T}}\right). \tag{A.147}$$

Proof. The least squares estimator of $\boldsymbol{\delta}$ is

$$\hat{\boldsymbol{\delta}} = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{y} = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}).$$

In addition to $\hat{\boldsymbol{\Sigma}}_{ss} = \mathbf{S}'\mathbf{S}/T$, $\boldsymbol{\Sigma}_{ss} = E(\mathbf{S}'\mathbf{S}/T)$ and $\boldsymbol{\Sigma}_{sx} = E(\mathbf{S}'\mathbf{X}/T)$, define

$$\hat{\boldsymbol{\Sigma}}_{sx} = \frac{\mathbf{S}'\mathbf{X}}{T},\ \boldsymbol{\delta}_* = \boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Sigma}_{sx}\boldsymbol{\beta},$$

and

$$\boldsymbol{\delta} = E\left(\hat{\boldsymbol{\delta}}\right) = E\left[(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}\boldsymbol{\beta}\right].$$

Note that

$$(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{X} = \hat{\boldsymbol{\Delta}}_{ss}\hat{\boldsymbol{\Delta}}_{sx} + \hat{\boldsymbol{\Delta}}_{ss}\boldsymbol{\Sigma}_{sx} + \boldsymbol{\Sigma}_{ss}^{-1}\hat{\boldsymbol{\Delta}}_{sx} + \boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Sigma}_{sx},$$

where

$$\hat{\boldsymbol{\Delta}}_{ss} = \hat{\boldsymbol{\Sigma}}_{ss}^{-1} - \boldsymbol{\Sigma}_{ss}^{-1},\ \hat{\boldsymbol{\Delta}}_{sx} = \hat{\boldsymbol{\Sigma}}_{sx} - \boldsymbol{\Sigma}_{sx}.$$

Hence

$$\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_{*}=\hat{\boldsymbol{\Delta}}_{ss}\hat{\boldsymbol{\Delta}}_{sx}\beta+\hat{\boldsymbol{\Delta}}_{ss}\boldsymbol{\Sigma}_{sx}\boldsymbol{\beta}+\boldsymbol{\Sigma}_{ss}^{-1}\hat{\boldsymbol{\Delta}}_{sx}\beta+\hat{\boldsymbol{\Sigma}}_{ss}^{-1}\left(\frac{\mathbf{S}'\mathbf{u}}{T}\right).$$

Using (2.15) of Berk (1974),

$$\left\|\hat{\boldsymbol{\Delta}}_{ss}\right\|_F\leq\frac{\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F^2\left\|\hat{\boldsymbol{\Sigma}}_{ss}-\boldsymbol{\Sigma}_{ss}\right\|_F}{1-\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F\left\|\hat{\boldsymbol{\Sigma}}_{ss}-\boldsymbol{\Sigma}_{ss}\right\|_F},$$

and using Cauchy-Schwarz inequality,

$$E\left\|\hat{\boldsymbol{\Delta}}_{ss}\right\|_F\leq\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F^2\left[E\left(\left\|\hat{\boldsymbol{\Sigma}}_{ss}-\boldsymbol{\Sigma}_{ss}\right\|_F^2\right)\right]^{1/2}$$
$$\cdot\left\{E\left[\frac{1}{\left(1-\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F\left\|\hat{\boldsymbol{\Sigma}}_{ss}-\boldsymbol{\Sigma}_{ss}\right\|_F\right)^2}\right]\right\}^{1/2}. \tag{A.148}$$

We focus on the individual terms on the right side of (A.148) to establish an upper bound for $E\left\|\hat{\boldsymbol{\Delta}}_{ss}\right\|_F$. The assumptions on eigenvalues of $\boldsymbol{\Sigma}_{ss}$ in this lemma are the same as in Lemma A4 with the only exception that $O(.)$ terms are used instead of $\ominus(.)$. Using the same arguments as in the proof of Lemma A4, it readily follows that

$$\|\boldsymbol{\Sigma}_{ss}\|_F=O(l_T),$$

and

$$\left\|\boldsymbol{\Sigma}_{ss}^{-1}\right\|_F=O\left(\sqrt{l_T}\right). \tag{A.149}$$

Moreover, note that (i,j)-th element of $\left(\hat{\boldsymbol{\Sigma}}_{ss}-\boldsymbol{\Sigma}_{ss}\right)$, $z_{ijt}=s_{it}s_{jt}-E(s_{it}s_{jt})$, satisfies the conditions of Lemma 11, which establishes

$$E\left(\left\|\hat{\boldsymbol{\Sigma}}_{ss}-\boldsymbol{\Sigma}_{ss}\right\|_F^2\right)=O\left(\frac{l_T^2}{T}\right). \tag{A.150}$$

Noting that $E(a^2)\leq\sqrt{E(a^4)}$, Assumption (ii) of this lemma implies that the last term on the right side of (A.148) is bounded, namely

$$E\left[\frac{1}{\left(1-\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F\left\|\hat{\boldsymbol{\Sigma}}_{ss}-\boldsymbol{\Sigma}_{ss}\right\|_F\right)^2}\right]=O(1), \tag{A.151}$$

Using (A.149), (A.150), and (A.151) in (A.148),

$$E\left\|\hat{\boldsymbol{\Delta}}_{ss}\right\|_F=O(l_T)\sqrt{O\left(\frac{l_T^2}{T}\right)}O(1)=O\left(\frac{l_T^2}{\sqrt{T}}\right). \tag{A.152}$$

It is also possible to derive an upper bound for $E\left(\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F^2\right)$, using similar arguments. In particular, we have

$$\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F^2 \leq \frac{\|\mathbf{\Sigma}_{ss}^{-1}\|_F^4 \left\|\hat{\mathbf{\Sigma}}_{ss}-\mathbf{\Sigma}_{ss}\right\|_F^2}{\left(1-\|\mathbf{\Sigma}_{ss}^{-1}\|_F \left\|\hat{\mathbf{\Sigma}}_{ss}-\mathbf{\Sigma}_{ss}\right\|_F\right)^2},$$

and using Cauchy-Schwarz inequality yields

$$E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F^2 \leq \|\mathbf{\Sigma}_{ss}^{-1}\|_F^4 \left[E\left(\left\|\hat{\mathbf{\Sigma}}_{ss}-\mathbf{\Sigma}_{ss}\right\|_F^4\right)\right]^{1/2} \cdot \left\{E\left[\frac{1}{\left(1-\|\mathbf{\Sigma}_{ss}^{-1}\|_F \left\|\hat{\mathbf{\Sigma}}_{ss}-\mathbf{\Sigma}_{ss}\right\|_F\right)^4}\right]\right\}^{1/2},$$

where $\|\mathbf{\Sigma}_{ss}^{-1}\|_F^4 = O\left(l_T^2\right)$ by (A.149), $E\left(\left\|\hat{\mathbf{\Sigma}}_{ss}-\mathbf{\Sigma}_{ss}\right\|_F^4\right) = O\left(l_T^4/T^2\right)$ by (A.118) of Lemma 11, and $E\left[\left(1-\|\mathbf{\Sigma}_{ss}^{-1}\|_F \left\|\hat{\mathbf{\Sigma}}_{ss}-\mathbf{\Sigma}_{ss}\right\|_F\right)^{-4}\right] = O(1)$ by Assumption *ii* of this lemma. Hence,

$$E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F^2 = O\left(l_T^2\right)\sqrt{O\left(\frac{l_T^4}{T^2}\right)}O(1) = O\left(\frac{l_T^4}{T}\right). \tag{A.153}$$

Using Lemma 11 by setting $\mathbf{S}_a = \mathbf{S}$ ($l_{a,T} = l_T$) and $\mathbf{S}_b = \mathbf{X}$ ($l_{b,T} = k < \infty$), we have, by (A.116),

$$E\left(\left\|\hat{\mathbf{\Sigma}}_{sx}-\mathbf{\Sigma}_{sx}\right\|_F^2\right) = O\left(\frac{l_T}{T}\right). \tag{A.154}$$

We use the above results to derive an upper bound for

$$\begin{aligned} E\left\|\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_*\right\| \leq\ & E\left[\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F \left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F\right]\|\boldsymbol{\beta}\| \\ & + E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F \|\mathbf{\Sigma}_{sx}\|_F \|\boldsymbol{\beta}\| \\ & + \left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F \|\boldsymbol{\beta}\| \\ & + E\left\|\hat{\mathbf{\Sigma}}_{ss}^{-1}\left(\frac{\mathbf{S}'\mathbf{u}}{T}\right)\right\|_F. \end{aligned} \tag{A.155}$$

First, note that $\|\boldsymbol{\beta}\| = O(1)$, and (using Cauchy-Schwarz inequality)

$$E\left[\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F \left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F\right]\|\boldsymbol{\beta}\| \leq \left(E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F^2\right)^{1/2}\left(E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F^2\right)^{1/2}\|\boldsymbol{\beta}\|.$$

But $E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F^2 = O\left(l_T^4/T\right)$ by (A.153), and $E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F^2 = O\left(l_T/T\right)$ by (A.154), and therefore

$$E\left[\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F \left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F\right]\|\boldsymbol{\beta}\| = \left[O\left(\frac{l_T^4}{T}\right)\right]^{1/2}\left[O\left(\frac{l_T}{T}\right)\right]^{1/2}$$
$$= O\left(\frac{l_T^{5/2}}{T}\right). \quad \text{(A.156)}$$

Next, note that $E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F = O\left(l_T^2/\sqrt{T}\right)$ by (A.153), $\|\mathbf{\Sigma}_{sx}\|_F = O(1)$ by Assumption *iii* of this lemma (and $\|\boldsymbol{\beta}\| = O(1)$), and we obtain

$$E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F \|\mathbf{\Sigma}_{sx}\|_F \|\boldsymbol{\beta}\| = O\left(\frac{l_T^2}{\sqrt{T}}\right). \quad \text{(A.157)}$$

Moreover, using (A.149), and noting that $E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F = O\left(\sqrt{l_T/T}\right)$by (A.154),[16]

$$\left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F = O\left(\sqrt{l_T}\right)O\left(\frac{\sqrt{l_T}}{\sqrt{T}}\right) = O\left(\frac{l_T}{\sqrt{T}}\right),$$

and hence

$$\left\|\mathbf{\Sigma}_{ss}^{-1}\right\|_F E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F \|\boldsymbol{\beta}\| = O\left(\frac{l_T}{\sqrt{T}}\right). \quad \text{(A.158)}$$

Finally, consider

$$E\left\|\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}'\mathbf{u}\right\|_F^2 = E\left\{Tr\left[\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}'\mathbf{u}\mathbf{u}'\mathbf{S}\left(\mathbf{S}'\mathbf{S}\right)^{-1}\right]\right\}$$
$$= \frac{\sigma^2}{T}E\left\{Tr\left[\left(\frac{\mathbf{S}'\mathbf{S}}{T}\right)^{-1}\right]\right\},$$

where $E\left(\mathbf{u}\mathbf{u}'/T\right) = \sigma^2\mathbf{I}_T$, and we have also used the independence of $\mathbf{S}$ and $\mathbf{u}$. Hence

$$E\left\|\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}'\mathbf{u}\right\|_F^2 = \frac{\sigma^2}{T}E\left[Tr\left(\hat{\mathbf{\Sigma}}_{ss}^{-1}\right)\right]$$
$$= \frac{\sigma^2}{T}Tr\left(\mathbf{\Sigma}_{ss}^{-1}\right) + \frac{\sigma^2}{T}E\left[Tr\left(\hat{\mathbf{\Sigma}}_{ss}^{-1} - \mathbf{\Sigma}_{ss}^{-1}\right)\right].$$

But $Tr\left(\mathbf{\Sigma}_{ss}^{-1}\right) = O\left(l_T\right)$, and using (A.152), we have

$$E\left|Tr\left(\hat{\mathbf{\Sigma}}_{ss}^{-1} - \mathbf{\Sigma}_{ss}^{-1}\right)\right| \leq l_T E\left\|\hat{\mathbf{\Sigma}}_{ss}^{-1} - \mathbf{\Sigma}_{ss}^{-1}\right\|_F$$
$$= l_T E\left\|\hat{\mathbf{\Delta}}_{ss}\right\|_F = O\left(\frac{l_T^3}{\sqrt{T}}\right).$$

[16] $E\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F \leq \left[E\left(\left\|\hat{\mathbf{\Delta}}_{sx}\right\|_F^2\right)\right]^{1/2} = \sqrt{O\left(K_T/T\right)} = O\left(\sqrt{K_T/T}\right)$.

It follows that,

$$E\left\|\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}'\mathbf{u}\right\|_F^2 = O\left(\frac{l_T}{T}\right) + O\left(\frac{l_T^2}{\sqrt{T}}\right)\frac{1}{T}$$
$$= O\left(\frac{l_T}{T}\right) + O\left(\frac{l_T^3}{T^{3/2}}\right). \tag{A.159}$$

Overall, using (A.156), (A.157), (A.158), and (A.159) in (A.155),

$$E\left\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_*\right\| = O\left(\frac{l_T^{5/2}}{T}\right) + O\left(\frac{l_T^2}{\sqrt{T}}\right) + O\left(\frac{l_T}{\sqrt{T}}\right)$$
$$+ O\left(\frac{l_T}{T}\right) + O\left(\frac{l_T^3}{T^{3/2}}\right).$$

Therefore

$$E\left\|\boldsymbol{\delta} - \boldsymbol{\delta}_*\right\| \to 0 \text{ when } l_T^4/T \to 0,$$

regardless whether $\mathbf{X}$ is included in $\mathbf{S}$ or not. Consider now

$$E\left\|\hat{\boldsymbol{\delta}} - \boldsymbol{\beta}_0\right\| = E\left\|\boldsymbol{\delta} - \boldsymbol{\delta}_* + \boldsymbol{\delta}_* - \boldsymbol{\beta}_0\right\|$$
$$\leq E\left\|\boldsymbol{\delta} - \boldsymbol{\delta}_*\right\| + E\left\|\boldsymbol{\delta}_* - \boldsymbol{\beta}_0\right\|.$$

But when $\mathbf{S} = (\mathbf{X}, \mathbf{W})$, then

$$\boldsymbol{\Sigma}_{ss} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xw} \\ \boldsymbol{\Sigma}_{wx} & \boldsymbol{\Sigma}_{ww} \end{pmatrix}, \boldsymbol{\Sigma}_{sx} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} \\ \boldsymbol{\Sigma}_{wx} \end{pmatrix},$$

and therefore $\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Sigma}_{ss} = \mathbf{I}_{l_T}$. This implies $\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Sigma}_{sx} = (\mathbf{I}_k, \mathbf{0}_{k\times kw})$ and $\boldsymbol{\delta}_* = \boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Sigma}_{sx}\boldsymbol{\beta} = \boldsymbol{\beta}_0$ when $\mathbf{S} = (\mathbf{X}, \mathbf{W})$. Result (A.146) now readily follows. When at least one of the columns of $\mathbf{X}$ does not belong to $\mathbf{S}$, then $\boldsymbol{\delta}_* \neq \boldsymbol{\beta}_0$. But

$$\left\|\boldsymbol{\delta}_* - \boldsymbol{\beta}_0\right\| \leq \left\|\boldsymbol{\delta}_*\right\| + \left\|\boldsymbol{\beta}_0\right\|,$$

where $\left\|\boldsymbol{\beta}_0\right\| = O(1)$, since $\boldsymbol{\beta}_0$ contains finite (k) number of bounded nonzero elements, and

$$\left\|\boldsymbol{\delta}_*\right\| = \left\|\boldsymbol{\Sigma}_{ss}^{-1}\boldsymbol{\Sigma}_{sx}\right\|_F$$
$$\leq \left\|\boldsymbol{\Sigma}_{ss}^{-1}\right\|_F \left\|\boldsymbol{\Sigma}_{sx}\right\|_F.$$

$\|\boldsymbol{\Sigma}_{ss}^{-1}\|_F = O\left(\sqrt{l_T}\right)$ by (A.149), and $\|\boldsymbol{\Sigma}_{sx}\|_F = O(1)$ by Assumption *iii* of this lemma. Hence, when at least one of the columns of $\mathbf{X}$ does not belong to $\mathbf{S}$,

$$\left\|\boldsymbol{\delta}_* - \boldsymbol{\beta}_0\right\| = O(l_T),$$

which completes the proof of (A.147). ■

References

ANTONIADIS, A., AND J. FAN (2001): "Regularization of Wavelets Approximations," *Journal of the American Statistical Association*, 96, 939–967.

BAI, J., AND S. NG (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

——— (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74, 1133–1150.

BAILEY, N., M. H. PESARAN, AND L. V. SMITH (2016): "A Multiple Testing Approach to the Regularisation of Large Sample Correlation Matrices," *CAFE Research Paper No. 14.05*.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014a): "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspectives*, 28, 29–50.

——— (2014b): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

BERK, K. N. (1974): "Consistent Autoregressive Spectral Estimates," *Annals of Statistics*, 2, 489–502.

BICKEL, J. P., Y. RITOV, AND A. TSYBAKOV (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *Annals of Statistics*, 37, 1705–1732.

BUHLMANN, P. (2006): "Boosting for High-Dimensional Linear Models," *The Annals of Statistics*, 34(2), 599–583.

BUHLMANN, P., AND S. VAN DE GEER (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

CANDES, E., AND T. TAO (2007): "The Dantzig Selector: Statistical Estimation When p is Much Larger Than n," *Annals of Statistics*, 35, 2313–2404.

DENDRAMIS, Y., L. GIRAITIS, AND G. KAPETANIOS (2015): "Estimation of random coefficient time varying covariance matrices for large datasets," *Mimeo*.

DONOHO, D., AND M. ELAD (2003): "Optimally Sparse Representation in General (Nonorthogonal) Dictionaries via l_1 Minimization," *Proceedings of the National Academy of Sciences*, 100, 2197–2202.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004): "Least Angle Regression," *Annals of Statistics*, 32, 407–499.

Fan, H. L. J., and B. M. Pötscher (2006): "Can one estimate the conditional distribution of post-model-selection estimators?," *Annals of Statistics*, 142, 2554–2591.

——— (2008): "Can one estimate the unconditional distribution of post-model-selection estimators?," *Econometric Theory*, 24, 338–376.

Fan, J., and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., and J. Lv (2008): "Sure Independence Screening for Ultra-High Dimensional Feature Space," *Journal of Royal Statistical Society B*, 70, 849–911.

——— (2013): "Asymptotic equivalence of regularization methods in thresholded parameter space," *Journal of the American Statistical Association*, 108, 1044–1061.

Fan, J., R. Samworth, and Y. Wu (2009): "Ultra High Dimensional Variable Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 1829–1853.

Fan, J., and R. Song (2010): "Sure Independence Screening in Generalized Linear Models with NP-Dimensionality," *Annals of Statistics*, 38, 3567–3604.

Fan, J., and C. Tang (2013): "Tuning parameter selection in high dimensional penalized likelihood," *Journal of the Royal Statistical Society Series B*, 75, 531–552.

Freedman, D. A. (1975): "On Tail Probabilities for Martingales," *Annals of Probability*, 3, 100–118.

Friedman, J. (2001): "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 29, 1189–1232.

Friedman, J., T. Hastie, and R. Tibshirani (2000): "Additive Logistic Regression: A Statistical View of Boosting," *Annals of Statistics*, 28, 337–374.

Huang, J., J. J. Horowitz, and S. Ma (2008): "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *Annals of Statistics*, 36, 587–613.

Lv, J., and Y. Fan (2009): "A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares," *Annals of Statistics*, 37, 3498–3528.

Pesaran, M. H., and R. P. Smith (2014): "Signs of Impact Effects in Time Series Regression Models," *Economics Letters*, 122, 150–153.

ROUSSAS, G. (1996): "Exponential Probability Inequalities with Some Applications," *Statistica, Probability and Game Theory, IMS Lecture Notes - Monograph Series*, 30, 303–319.

STOCK, J. H., AND M. W. WATSON (2012): "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business and Economic Statistics*, 30, 481–493.

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

WECKER, W. E. (1978): "A Note on the Time Series which Is the Product of Two Stationary Time Series," *Stochastic Processes and their Applications*, 8, 153–157.

WHITE, H., AND J. M. WOOLDRIDGE (1991): "Some Results on Sieve Estimation with Dependent Observations," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. J. Barnett, J. Powell, and G. Tauchen, pp. 459–493. Cambridge University Press, New York.

ZHANG, C. H. (2010): "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *Annals of Statistics*, 38, 894–942.

ZHENG, Z., Y. FAN, AND J. LV (2014): "High Dimensional Thresholded Regression and Shrinkage Effect," *Journal of the Royal Statistical Society B*, 76, 627–649.

ZHOU, H., AND T. HASTIE (2005): "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society B*, 67, 301–320.